Harvard Data Science Review • Issue 4.2, Spring 2022

Data Inventories for the Modern Age? Using Data Science to Open Government Data

Julia Lane[1] Ernesto Gimeno[2] Ekaterina Levitskaya[2] Zheyuan Zhang[2] Alberto Zigoni[3]

[1]New York University, New York City, New York, United States of America, [2]Coleridge Initiative, United States of America, [3]Research Intelligence Solutions, Elsevier

Published on: Apr 28, 2022

DOI: https://doi.org/10.1162/99608f92.8a3f2336

ABSTRACT

This article describes how data science techniques—machine learning and natural language processing—can be used to open the black box of government data. It then describes how an incentive structure can be established—using human–computer interaction techniques —to create a new and sustainable data ecosystem. The particular focus is on the United States and on scientific researchers, who are major users of government data. However, the approach can be deployed to other use cases, such as data mentions in newspapers and government reports, and many other countries

Keywords: data inventories, democratizing data, Kaggle competition, machine learning, natural language processing

Media Summary

Scientists now have access to previously unimaginable amounts of data—including government data. As a result, empirical research can be conducted at a scale that would not have been possible a generation or two ago. Yet the current approach to finding what data sets are used to answer scientific questions—the foundation of empirical research—is largely manual and ad hoc. Better information on the use of data is likely to have at least two results: (i) government agencies might use the information to describe the return on investment in data sets and (ii) scientists might use the information to reduce the time taken to search and discover other empirical research.

This article describes how data science techniques—machine learning and natural language processing—can be used to open the black box of government data. It then describes how an incentive structure can be established—using human–computer interaction techniques—to create a new and sustainable data ecosystem. The particular focus is on the United States and on scientific researchers, who are major users of government data. However, the approach can be deployed to other use cases, such as data mentions in newspapers and government reports, and many other countries.

1. Introduction

This article describes how data science techniques—machine learning and natural language processing—can be used to open the black box of government data. It then describes how an incentive structure can be established—using human-computer interaction techniques—to create a new and sustainable data ecosystem. The particular focus is on the United States and on scientific researchers, who are major users of government data. However, the approach can be deployed to other use cases, such as data mentions in newspapers and government reports, and many other countries.

The United States impetus is the Foundations for Evidence-Based Policymaking Act (2018), signed into law in January 2019, which requires the federal government to modernize its data management practices. In particular, it requires agency data to be accessible and agencies to use data and evidence to inform their work, to build measures to inform the public of data use, and to provide ways for the public to request that specific data assets be prioritized (Brown, 2021). As part of the act, the Government Accountability Office (GAO) is required to report to Congress identifying (1) the value of information made available to the public, (2) whether publishing information that has not yet been published would be valuable to the public, and (3) the completeness of each comprehensive data inventory developed.[1]

The impact of these requirements could be far-reaching. Lew Platt, the former CEO of Hewlett-Packard, once said “If HP knew what HP knows, it would be three times more profitable” (Coates, 2001). If our country knew what data it had, it would be many times more productive —indeed, as Peled has argued, the result will be the federal government’s third-most important computer innovation in history, following the American government’s contribution to the inventions of the computer and the internet (Peled, 2011, 2013; Peled & Nahon, 2015). Indeed, if the information were readily available through automated, 21st-century tools, the reproducibility of empirical science would be vastly enhanced (Yarkoni et al., 2021), agencies could more easily respond to executive orders and congressional mandates, and the time cost for analysts to find out about data could be reduced by at least a third.

However, the impact will only be far-reaching if the implementation is well-designed. The initial attempt, in 2009, to make data open was largely supply driven. Like the most recent law, it was an unfunded mandate. It required agencies to make their data available through open data portals (data.gov) and was criticized for being badly designed, flawed in execution, costly, with few benefits, and resulting in the system being gamed (Peled, 2011, 2013). More than a decade later, the lessons learned from the previous effort, combined with new tools and technologies, can inform a new, more demand-driven approach in which incentives and feedback are keys to success.

The critical element is creating incentives for federal agencies. The approach must be grounded in the provision of reporting tools that enable them to provide data to the public with minimal cost and burden, that helps them do their jobs better, and are rewarded for their efforts. Users need to be similarly incentivized. They need to be presented with tools that minimize the cost and burden and maximize the rewards associated with documenting their use of government data. Finally, intermediaries, such as publishers, that make it easier to find how data are being used must be rewarded for their investment of time and money.

The first step is to identify the value proposition for the government. Data-driven private sector companies like Amazon.com, Yelp, and Airbnb have clear mission statements (Amazon's, for example, is "to be Earth's most customer-centric company where people can find and discover anything they want to buy online" (Brandt, 2011, p. 1). We worked with selected informants at three separate agencies[2] as well as the Federal Chief Data Officers' Council to find their view of the equivalent mission statement.

The second step is to provide the foundation for identifying the use of data with minimum cost and burden for both agencies and researchers. The key insight here is that when data are used, and publicly acknowledged, there is likely to be a digital trace that data science tools can identify. For example, major users of government data —empirical scientists—will typically publicly mention the use of data in their scientific publications.[3] Data science tools—such as machine learning (ML) and natural language processing (NLP) techniques can be deployed to find data sets that are publicly acknowledged in publications. A major barrier is that researchers do not routinely cite a data set that is used in empirical research, even when they publicly acknowledge having used it, in a way that the information can readily be retrieved. This step was achieved through a Kaggle competition[4] that resulted in over 1,600 data science teams competing to develop the best ways to do so. Since this task is both the heaviest technical lift, and the most challenging in terms of technical tools, most of the article and appendices is devoted to describing the approach.

The third is to characterize the use with measures that can convey the value of the data set to decision makers and to the public at large. While Section 202(c) of Title 2 of the Foundations for Evidence-Based Policymaking Act (2018) is clear in terms of the charge to agencies, the specifics are up to agency interpretation. There are hundreds of ways in which use can be categorized in the abstract; in practice, we applied the best techniques from the Kaggle competition to a rich corpus of publications to identify the topics, authors, and visibility of the work. The results can be provided in an application programming interface (API) and drawn on to create a variety of usage metrics in a manner that is open and transparent.

The fourth step is to ensure that the results are not a one-off application. The approach must create incentives for publishers, government agencies, the public and academic researchers to continuously update and validate the data ecosystem and associated publications. The goal is to create a self-reinforcing ‘information marketplace’—an Amazon.com for data—so that knowledge about how data that are publicly acknowledged as being used can be shared by all stakeholders.

The article is structured as follows. It begins with describing the context that makes the information marketplace possible: both the interest of Congress and the current administration in evidence-based policymaking, as well as the important infrastructure investments that have been made by publishers, libraries, and data repositories. Subsequent sections describe the progress that has been made in each of the four steps outlined above. It concludes with an outline of a future research agenda.

2. Context

Although scientists now have access to previously unimaginable amounts of data, and as a result, empirical research can be conducted at a scale that would not have been possible a generation or two ago, the current approach to finding what data sets are used to answer scientific questions is largely manual and ad hoc (Yarkoni et al., 2021). While at least part of this situation is due to historically limited incentives for researchers to publicize the data sets they use, the incentive structure seems to be changing—although only a fraction of data sets are identified in scientific research, those publications that do cite data are cited up to 25% more than those that do not. Researchers who conduct their work in an ‘open’ fashion have more traction in obtaining funding and positions (Colavizza et al., 2020). And the National Academies of Science, Engineering, and Medicine (NASEM) have recently launched an effort to improve upon the reporting of scientific methods.

Government agencies are also realizing that they have massive amounts of administrative and programmatic data to curate and disseminate—and that they are charged with doing a better job of doing so. The Federal Data Strategy charges agencies with leveraging their data as a strategic asset and producing inventories of their data. The Information Quality Act requires agencies to consider the appropriate level of quality (utility, integrity, and objectivity) for each of the products that it disseminates based on the likely use of that information. The Foundations for Evidence-Based Policymaking Act (2018) requires the federal government to modernize its data management practices and agencies' strategic plans to contain an assessment of the coverage, quality, methods, effectiveness, and independence of the agency's statistics, evaluation, research, and analysis efforts (Lane, 2020). More generally, the importance of data and data reuse, while grounded in academics, is not purely of academic interest. The reliability of data for policy purposes is threatened in many areas, ranging from public health to the enumeration of our population to estimating unemployment (Lane, 2020).

The infrastructure is also in place. The General Services Administration is modernizing Data.gov and has substantial funds available in the Technology Modernization Fund. Publishers have established DataCite's metadata schema (Robinson-García et al., 2017; Starr & Gastl, 2011). The American Geophysical Union is working with data repositories and publishers to develop Findability, Accessibility, Interoperability, and Reuse (FAIR) best practices (Stall et al., 2016). However, as Allen has noted, data sets are still not identified as first-class scientific assets in the same way as are publications, and as a result, researchers are still much less likely to cite data sets than publications (Allen, 2020). Notably, one very recent hand-curated analysis of over 12,000 full-text articles studying COVID found that about 128 data set links were mentioned, and in about 28% of the articles—and found that citation patterns of the data sets varied substantially (Zuo et al., 2021).

There is ample precedent for building platforms to search for and discover scientific assets. GitHub is a platform for collaboration and for sharing of existing scripts, visualizations, and even workflows (Ieong et al., 2014). Code sharing and collaboration can facilitate reusability, reproducibility, and credibility. Code sharing is associated with metrics of research impact in several fields (Vandewalle, 2012). GitHub facilitates 'open research,' which has been shown to increase not only metrics associated with research impact, but even those in job and funding opportunities. Workflow marketplaces/exchanges are exchanges in which scientists can share workflows that they have configured for a specific experiment, such as myExperiment, the Library Workflow Exchange, the CWL (Common Workflow Language) Workflow Collection, and the SHIWA (SHaring Interoperable *Workflows)* Workflow Portal. Ontologies can be used to describe workflows and sections of it; describing research objects with ontologies enables easier sharing and interoperability of objects like workflows; for example, the e-Science lab offers the RightField tool, which can be used to annotate data sets in Excel spreadsheets (Wolstencroft et al., 2011). Similarly, Figshare, Jupyter Notebook, R Markdown, and Binder are tools that enable sharing of visualizations, communication of a scientific process, and replication of results and are important tools in creating reproducible research (Forde et al., 2018; Kluyver et al., 2016). There have been some efforts to create research recommendation systems (Luan, 2018), but they are limited to search engines such as Google Scholar or Semantic Scholar. Although these engines manage to provide the publications related to a specific Word query, it is challenging to filter and organize all the suggested publications associated with data sets. Indeed, recent work has suggested the establishment of data market platforms (Fernandez et al., 2020).

2.1. Identifying the Value Proposition

A set of interviews was conducted with agency heads, data stewards, and librarians from each of the agencies to get input about the appropriate measures for data use. Although the interviews revealed that agencies would like to close the gaps in their knowledge about dataset use, as the conversation snippets highlighted in Figure 1 suggest, there is no real consensus on the Knowledge Performance Indicators (KPIs) and there is limited ability to track usage.

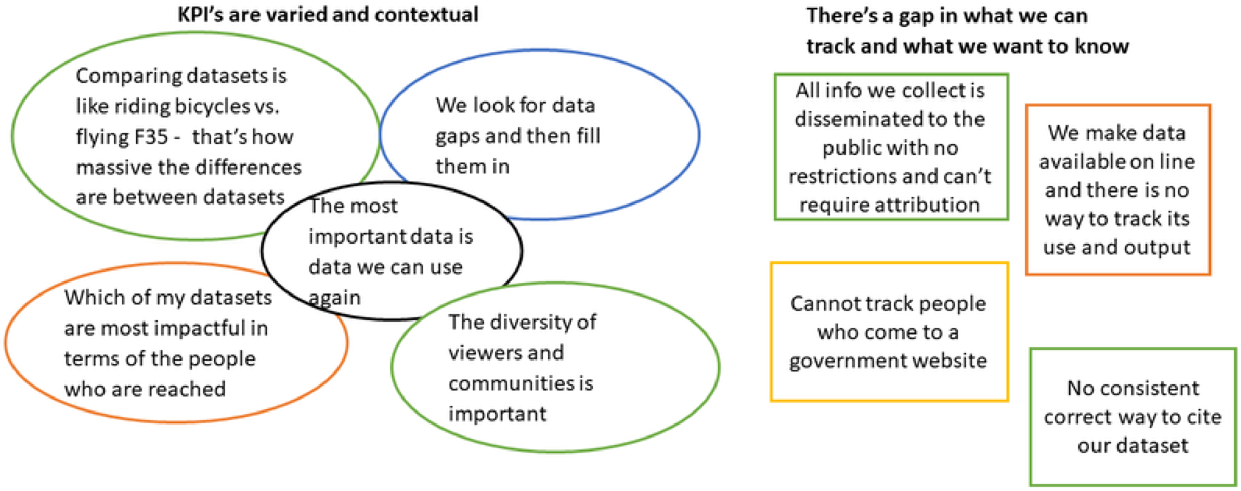

KPI's are varied and contextual
Comparing datasets is like riding bicycles vs. flying F35 - that's how massive the differences are between datasets
We look for data gaps and then fill them in
The most important data is data we can use again
Which of my datasets are most impactful in terms of the people who are reached
The diversity of viewers and communities is important
There's a gap in what we can track and what we want to know
All info we collect is disseminated to the public with no restrictions and can't require attribution
We make data available on line and there is no way to track its use and output
Cannot track people who come to a government website
No consistent correct way to cite our dataset

Figure 1. Summary of interviews with agencies.

A subsequent set of meetings, held with agency chief data officers (CDOs), uncovered needs such as agency priority setting, the democratization of data discovery, and increasing the usability of clearing houses like data.gov. The CDOs were also particularly interested in using data inventories to find data related to their work and understanding the rate of return to investment in data.

2.2. Building the Foundation

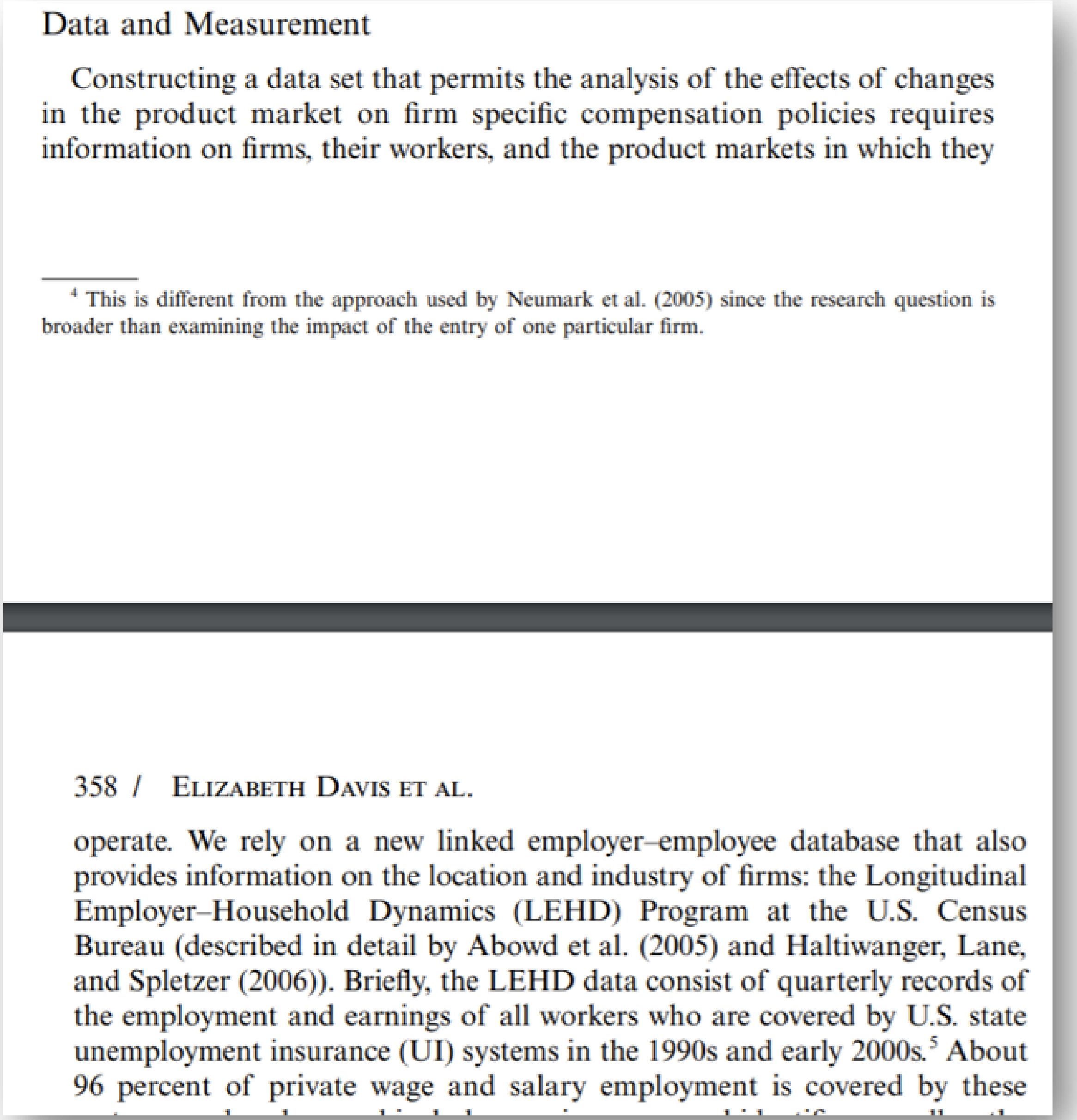
The empirical problem is substantial. In the corpora described following, where we have manually curated the public acknowledgements of the data sets used, they are buried in the text of the publication itself. Fewer than 5% actually cite a data set in the references, despite active exhortations to do so (Vannan et al., 2020). Searching the text of the references themselves using strings representing data set names results in a serious undercount (Appendix A).

The data science challenge is to find how a data set is publicly named (data set name [DSN]) in a publication *p*, by an individual author, *I*, in a field *f*. Once a data set-publication dyad is identified, it is possible to develop a foundation for step 3—a rich set of usage measures—like the number of times a data set is mentioned, the research topics that it is used to answer (from text analysis of the article or from publisher keywords), the other data sets with which it is used (from the same publication), the authors/experts who use the data sets, and the relative take-up of the work (from the citations).

In simple terms, the issue is to find, in every empirical publication, the DSN as a function of individual, field, and publication characteristics, as well as the semantic context (*X*) within which the data set is mentioned.

$$DSN_{i,f,p} = \theta_i + \vartheta_f + \varphi_p + \alpha X + \varepsilon_{i,f,p}$$

Consider the use case of a publication one of us authored (Davis et al., 2009) that referred to a data set (Longitudinal Employer–Household Dynamics [LEHD]) that she had created (Figure 2). A human reading the text would know that the data set referenced is the LEHD data (Abowd et al., 2004)—because the semantic flags make it clear to the human—but since the dataset is not cited, the Census Bureau would not know of the particular use of LEHD.



Data and Measurement

Constructing a data set that permits the analysis of the effects of changes in the product market on firm specific compensation policies requires information on firms, their workers, and the product markets in which they

[4] This is different from the approach used by Neumark et al. (2005) since the research question is broader than examining the impact of the entry of one particular firm.

358 / Elizabeth Davis et al.

operate. We rely on a new linked employer–employee database that also provides information on the location and industry of firms: the Longitudinal Employer–Household Dynamics (LEHD) Program at the U.S. Census Bureau (described in detail by Abowd et al. (2005) and Haltiwanger, Lane, and Spletzer (2006)). Briefly, the LEHD data consist of quarterly records of the employment and earnings of all workers who are covered by U.S. state unemployment insurance (UI) systems in the 1990s and early 2000s.[5] About 96 percent of private wage and salary employment is covered by these

Figure 2. Example of data publicly mentioned in text but not in references.

The Kaggle competition was designed to identify the best data science tools to find data sets in a set of publications. We worked with Kaggle to define the competition; about 1,600 data science teams entered. The details of the competition are important, because they determine the success of the effort, and are described in Appendix B; we only provide a brief overview here.

Competition Specifics: We used key data sets from multiple agencies.[5] Simple string search methods were applied to find about 22,000 open access publications that could be used for both the training and test data sets. A number of steps, also described in Appendix B, was necessary to label each data set in the publication corpus. A major conceptual challenge that emerged as a result of the labeling exercise—that was also highlighted by Kaggle competitors—was how to define a data set name. A data set is defined in the W3C-DCAT (W3C Data Catalog Vocabulary) as “a collection of data, published or curated by a single agent” such as a statistical agency (Allen, 2020). But as Peled notes, the GAO recently pointed out “the concepts of ‘record’ and ‘database’ in governments are hopelessly arcane in the age where agencies acquire data from different sources” (Peled, 2011, p. 2092).

The challenge with defining a data set correctly also had implications for developing metrics for the Kaggle competition. If we required the prediction to be too precise, we would understate usage. If we were too lax in terms of accepting loosely defined data sets as correct, we would overstate usage. Empirically, the challenge was evidenced by the fact that the data set labels could be long and contain several words, which led to having to decide how close a set of words had to be to classify a prediction as true. The metric chosen to evaluate similarities between texts was a micro Jaccard-based FBeta score between prediction texts and target texts to evaluate the submissions from the Kaggle competition.[6]

Techniques Used: The techniques used by the Kaggle competitors provided a useful insight into the current state of the art. One of simplest and most straightforward approaches, and adopted by a number of competitors, was a string-matching method. The basic idea of this method is to record all the data set names that appear in the training labels and create a lookup table. The test data set was then searched for the data set name in the lookup table. This method is very intuitive, but only finds the data sets known a priori, and is hence fundamentally flawed if the goal is to identify all data sets used in the corpus (Lane et al., 2020). Some competitors improved on the naïve approach by dynamically updating the lookup table during training and testing. The solution methods include searching candidates in certain formats, filtering data set candidates based on Keywords and searching based on the frequency of data set appearance. These methods can achieve a satisfactory performance if delicately designed.

Another approach was to use state-of-the-art deep learning techniques, using the named entity recognition (NER) model from Hugging Face (Wolf et al., 2019) or SpaCy

(Schmitt et al., 2019)—very popular libraries in the NLP community—combined with three different types of rule-based processing methods. The first method is to take the embeddings of the entire sentence as input, the second method is to take only the data set candidates as input, and the third method is to use the MASK token to hide data set candidates and focus solely on its surrounding context. Of those who decided to use deep learning, the majority used transformer-based pretrained language models, among which Bidirectional Encoder Representations from Transformers (BERT) and its variations, such as RoBERTa and DistilBert were the most popular. A few competitors used models designed to handle long sequences, such as Big Bird—a sparse-attention-based transformer specially designed by Google (Zaheer et al., 2020).

Table 1 summarizes each main method mentioned above, including both rule-based methods and deep learning methods, and includes both examples and a summary of each method’s pros and cons. Table 2 is a summary of the approaches that seven leading winners used in this competition. The successful NLP models all required a combination of those methods, and in fact, the solutions that integrated language models and rule-based filters showed the best performance (Ghani, 2021).

Table 1. Summary of methods.

Method	**Example**	**Pros**	**Cons**
Searching Candidates in Certain Format	Abbreviation Detector from scispacy library; regular expression search based on custom rules (e.g., mixed cap words followed by an acronym in parentheses, such as “Baltimore Longitudinal Study of Aging (BLSA)”)	Easier to include connecting words or punctuations inside data set names (such as “and” in data set “Deep-ocean Assessment and Reporting of Tsunamis”)	Not possible to find other data sets beyond the rules; other entities (such as organizations) that are not data sets could lead to false positives

Filtering Based on Keywords	Keyword check process (data set-indicative words, e.g., “survey,” “database,” versus non–data set words, such as “center,” “committee”)	Good performance results (many data set names indeed include the data set-indicative keywords, such as “Survey of Earned Doctorates”; noisy outputs, such as organization names, are filtered out)	Not all data sets follow this rule (e.g., “NOAA Tide Gauge” or “Ocean Heat Content”)
Searching Based on the Frequency of Data Set Appearance	Searching data set names based on its frequency in the corpus (a data set is likely to appear multiple times in multiple publications)	Can find data sets in publications where language models could fail, because the context was not enough to predict	Can be misleading, as in some contexts the same string can have a different meaning (e.g., “Program for International Student Assessment” can be mentioned as a data set, but in some contexts, it is an assessment)
Data Augmentation	Using external/additional data sources for training the models to achieve better performance (e.g., an external catalogue of U.S. government’s data set names)	More data improves the performance	Time-consuming and expensive; additional data is not easily collected or cleaned
Models Learning From Whole Sentences	Training a named entity recognition model based on parsed out sentences from the corpus (using pretrained transformers such as BERT or RoBERTa)	Using parsed out sentences as opposed to the full-length text allows the implementation of the end-to-end deep learning models	Because of the scarcity of training labels and the intrinsic ambiguity of data sets, even the state-of-art language models could hardly outperform the rule-based approaches

Models Learning From Candidate Strings	Using approaches to select data set candidates first and inputting only these candidate strings into a named entity recognition or text classification model	By limiting the input to a list of data set candidate strings, the performance of the model is greatly improved	Requires efforts of preprocessing to extract candidates for input; doesn’t take context into account
Masked Language Model	Implementing a <MASK> on the candidates and inputting the sentence to the model to force the model to learn from the context	Good performance results (the winning method of the competition)	Requires efforts of preprocessing to extract candidates for the masking

Table 2. Summary of methods.

Competition position	**Approach Name**	**Methods used**	**F0.5 score**
1	Context Similarity via Deep Metric Learning	Searching candidates in certain format + Filtering out prediction based on keywords + Searching based on the frequency of data set appearance + Masked language model	0.576
2	Transformer-Enhanced Heuristic Search	Searching candidates in certain format + Searching based on the frequency of data set appearance + Models learning from candidate strings	0.575

3	Simple and Strong Baseline	Searching candidates in certain format + Filtering out prediction based on keywords + Searching based on the frequency of data set appearance	0.558
4	Acronym Detection and Variation Detection by Named Entity Recognition	Searching candidates in certain format + Filtering out prediction based on keywords + Searching based on the frequency of data set appearance + Models learning from whole sentences + Data augmentation	0.513
5	Regular Expressions	Searching candidates in certain format + Filtering out prediction based on keywords + Searching based on the frequency of data set appearance	0.486
6	A Hybrid Approach Using Different Natural Language Processing Techniques and SpaCy	Searching candidates in a certain format + Filtering out prediction based on keywords + Models learning from whole sentences + Models learning from candidate strings + Data augmentation	0.478
7	Schwartz and Hearst Propose, BERT Decides	Searching candidates in certain format + Masked language model	0.478

Laying the Foundation: The winning ML models were then applied to three separate ScienceDirect corpora provided by Elsevier. Since the goal was to produce information about the use for the three agencies, each corpus was developed sequentially. The research topics of Kaggle competition corpora were identified. The topics were used to filter ScienceDirect publications, as was the author affiliation (at least one author needed to be based in the United States). Then the ML model was applied.

The results were fascinating. Not surprisingly, a simple string search approach worked well if the data set name and aliases were identified. But the ML models were able to pick up a wider variety of ways in which authors refer to the same data sets.

This is especially true for the models using the method ‘masked language modeling,’ which forced the models to learn from the context and not from the way in which the aliases were written in the training set. On a randomly selected 20,000 corpus of publications from ScienceDirect, the first-place model correctly identified more data sets—1,000 data set hits and 600 unique data sets—than did either the second-place model (transformer-enhanced heuristic search)—around 400 hits and 300 unique data sets—or the third-place model (simple and strong baseline—based on regular expressions)—around 600 hits and 300 unique data sets. It was simpler and faster to run approaches not relying on deep learning methods, and such approaches should not be neglected as part of a bigger pipeline. The second and third place models performed well if applied to publication domains similar to the one used on the Kaggle competition. Interestingly, there was little overlap in the data sets identified by the different models, meaning there is still great potential for learning and improvement.

There were a number of lessons learned. First, it is difficult to build a machine learning model that has a good performance without introducing any rule-based approach. Even the models trying to predict data sets based on semantic context relied heavily on a rule-based approach as preprocessing or postprocessing. Second, pure rule-based approaches are insufficient: integrating machine learning models with rule-based learning approaches generates more correct prediction. Third, ML models appear to perform better on publications that cite more data sets. Finally, and not surprisingly, the ML models appear to have better performance retrieving data sets and their aliases if the data set exists in the training set.

2.3. Measuring Use and Creating Value

What is clear is that there is an almost overwhelming number of ways to measure use. A knowledge graph could be built to create indicators based on bibliometrics on the connected publications, data set use, and data set impact, initially by publication and data set citation. Some additional information could include the data set size (e.g., number of samples), and other attributes that could be important to know from the beginning, before deciding on a particular data set (e.g., image resolution, number of classes, age range of survey participants, etc.). The task of collection of such attributes can be treated as slot filling or knowledge graph construction (for each data set, trying to find as many standardized properties [slots] as possible).

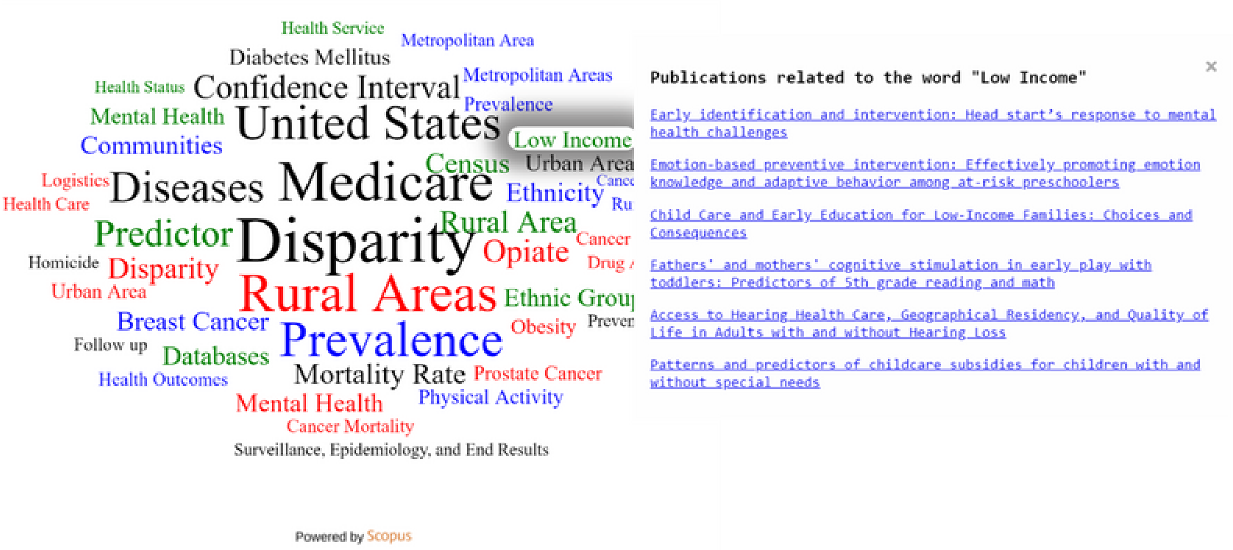
In essence, the state of the art for establishing the value proposition is similar to that 60 years ago, when the Science Citation Index began to get used to show the impact of publications (Cronin & Sugimoto, 2014). There are several lessons to be learned from the resultant emergence of bibliometrics (Sugimoto & Larivière, 2018)—not least of which is the importance of multiple open and transparent ways of developing measures from the publication–data set dyads that are based on a conceptual framework.

A particularly attractive framework identified by a NOAA researcher (Tyler Christensen) can be drawn from forestry, which has developed measures of ‘high conservation value’ and applied them to other natural landscapes. The idea is to rank forested areas in order of importance for conservation. The areas with high conservation value should be preserved as natural areas, and forestry activities should focus instead in areas with lower conservation value.[7] Christensen suggested a set of data usage measures. Inspired by the forestry value measures, namely:

> **Diversity**: Simply count how often the data are used, with special consideration of data sets that are the sole source of information in nearly every study within a research discipline.
>
> **Landscape-level ecosystems**: Data that are often used in combination with federal data sets from other agencies.
>
> **Ecosystems and habitats**: Clusters of data that are often used together in a research discipline, so that if one is lost then the others would lose value.
>
> **Ecosystem services**: Data used in research topics aimed at protecting life and property in critical situations, for example, floods, pandemics, war.
>
> **Community needs**: Data used in research topics aimed at supporting basic community needs: for example, health, food, housing, livelihoods.

> **Cultural values**: Data used in research topics aimed at protecting historical, ecological, sacred, or intangible values.

Another approach is simply to show the distribution of topics as way to describe data use. Figure 3 provides an example of a wordcloud that was developed from the topics of publications that cited the United States Department of Agriculture's (USDA) Rural-Urban Continuum Codes.



Publications related to the word "Low Income"

Early identification and intervention: Head start's response to mental health challenges

Emotion-based preventive intervention: Effectively promoting emotion knowledge and adaptive behavior among at-risk preschoolers

Child Care and Early Education for Low-Income Families: Choices and Consequences

Fathers' and mothers' cognitive stimulation in early play with toddlers: Predictors of 5th grade reading and math

Access to Hearing Health Care, Geographical Residency, and Quality of Life in Adults with and without Hearing Loss

Patterns and predictors of childcare subsidies for children with and without special needs

Figure 3. Example of how data usage information can be visualized.

2.4. Building an Ecosystem

The long-term vision would be for the federal government to create a new data ecosystem, joint with publishers, agencies, the public, and the academic community, while ensuring that all privacy issues are addressed and that all stakeholders get credit for participating.

The practical approach might be as follows.

Expand the universe: Take the ML model resulting from the Kaggle competition and apply it to a full corpus of documents—starting with scientific publications, but eventually expanding to other text, such as government documents, grey literature, *Federal Register* notices, and the popular press. This will require agreement from multiple sources, so a legal infrastructure will need to be put in place.

Expand the ways in which researchers can contribute to publications that were 'missed,' in working paper format or in preprint status.

Expand the value: Build an API that will expose continuously updated data and that provides an interactive way of working with the API's endpoints—authors, publications, and topics. The API can be used by agencies to document the usage of agency data, and by prospective agency researchers to get more information about their data. It can also be used to create production-level measures with an ongoing workflow that is fed from the publications. Enhance the API to include research that is done inside research data centers and joint with data repositories to identify emerging and prepublication research fields. Build a knowledge graph based on that API, and facilitate the automation of validation and feedback loops by working with the research community to produce open source tools.

Expand the engagement with publishers to feed the validated data set references to data repositories—particularly federally funded data repositories

3. Future Agenda

The future agenda will be to emulate the success of the private sector in integrating the demand side (incentives) with the supply side (such as data citation standards).

We are working with the three agencies to identify practical use cases that could serve as a pilot. One part of the pilot is to build an engaged community that shares knowledge and information frequently. In those use cases, we will create usage measures and build a feedback loop from agencies and users by adding feedback utility inside our usage application. We will give the users an option to submit new publications or validate the use of a data set in a publication. Part of the agenda will also be to develop a recommendation system to improve data set search and discovery so that data set use can be dynamically generated by the agencies and users themselves. The core idea is that the most relevant data sets, authors, topics, and publications that they may be interested in would be recommended based on prior use. Data producers could use a recommender system to predict who used what data, and who is more likely to use future releases of data. In either case, the access to and use of data would be broadly enhanced, meeting some of the fundamental goals of the Foundations for Evidence-based Policymaking Act (2018).

Regardless of the future, the approach should help inform agencies in responding to Congress's request to report on the (1) the value of information made available to the public, (2) whether publishing information that has not yet been published would be valuable to the public, and (3) the completeness of each comprehensive data inventory developed. And we hope that the result will be the federal government's third-most important computer innovation in history (Peled, 2013; Peled & Nahon, 2015)—with a minimum of cost, effort, and burden on the taxpayer.

Acknowledgments

This work draws on the input from many colleagues. We are particularly grateful to our agency colleagues Mark Denbaly, Kelly Maguire, and Cindy Parr (USDA), Tyler Christensen, Eugene Burger and Deirdre Clarkin (NOAA), Vipin Arora and WanYing Chang (NSF), Haritz Puerto Roman, José Ramón Romero Pineda, Paco Nathan, Mike Stebbins, Maggie Demkin, Phil Culliton, Howard Ratner, Ian Mulvany, Ismail Coskun, Jordan Boyd Graber, Rayid Ghani, Ophir Frieder, Alex Wade, Frauke Kreuter, Alice Grantham, Christina Acheson, Claire Morris and CHORUS members. Mikhail Arkhipov, one of the Kaggle winners, was particularly insightful with his comments.

Disclosure Statement

The authors were funded by the Alfred P. Sloan Foundation, the McGovern Foundation, the National Science Foundation, the Overdeck Family Foundation, Schmidt Futures, and the U.S. Department of Agriculture.

References

Abowd, J. J., Haltiwanger, J., & Lane, J. (2004). Integrated longitudinal employer-employee data for the United States. *American Economic Review*, *94*(2), 224–229. http://doi.org/10.1257/0002828041301812

Allen, R. B. (2020). Metadata for social science datasets. In J. I. Lane, I. Mulvany, & P. Nathan (Eds.), *Rich search and discovery for research datasets* (pp. 40–52). Sage Publishing.

Brandt, R. L. (2011). *One click: Jeff Bezos and the rise of Amazon.com*. Penguin.

Brown, O. (2021) *Scientific data and the evidence act*. Bureau of Economic Analysis.

Chadegani, A. A., Salehi, H., Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi, M., & Ebrahim, N. A. (2013). A comparison between two main academic literature collections: Web of Science and Scopus databases. *Asian Social Science, 9*(5), 18–26. http://doi.org/10.5539/ass.v9n5p18

Coates, J. F. (2001). One point of view: Knowledge management is a person-to-person enterprise. *Research-Technology Management*, *44*(3), 9–13. https://doi.org/10.1080/08956308.2001.11671423

Colavizza, G., Hrynasziewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data*. PloS one*, 15(4), Article e0230416. https://doi.org/10.1371/journal.pone.0230416

Cronin, B., & Sugimoto, C. R. (2014). *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*. MIT Press.

Davis, E., Freedman, M., Lane, J., & Mccall, B. P. (2009). Product market competition and human resource practices in the retail food sector. *Industrial Relations: A Journal of Economy and Society*, *48*(2), 350–371. http://doi.org/10.1111/j.1468-232X.2009.00561.x

Drew, L. W., (2011). Are We Losing the Science of Taxonomy? As need grows, numbers and training are failing to keep up*. BioScience, 61*(12), 942–946. https://doi.org/10.1525/bio.2011.61.12.4

Fernandez, R. C., Subramaniam, P., & Franklin, M. J. (2020). Data market platforms: trading data assets to solve data problems*. Proceedings of the VLDB Endowment*, *13*(12), 1933–1947. https://doi.org/10.14778/3407790.3407800

Forde, J., Bussonnier, M., Fortin, F., Granger, B., Head, T., Holdgraf, C., Ivanov, P., Kelley, K., Pacer, M., Panda, Y., Pérez, F., Nalvarte, G., Ragan-Kelley, B., Sailer, Z. R., Silvester, S., Sundell, E., & Willing, C. (2018). *Reproducing machine learning research on Binder.* Paper presented at the 32nd Conference on Neural Information Processing Systems (NIPS 2018 Workshop MLOSS). https://openreview.net/pdf?id=BJepbQkJ5Q

Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2018).

Ghani, R. (2021).*The winning methods*, in *show US the data*. Coleridge Initiative. https://coleridgeinitiative.org/wp-content/uploads/2021/11/Coleridge-Conference-Deck-rayid-ghani.pdf

Godfray, H. C. J. (2007). Linnaeus in the information age*. Nature, 446*(7133), 259–260. https://doi.org/10.1038/446259a

Gormley, C., & Tong, Z. (2015). *Elasticsearch: The definitive guide: a distributed real-time search and analytics engine*. O'Reilly Media.

Ieong, P. U., Sørensen, J., Vemu, P. L., Wong, C. W., Demir, Ö., Williams, N. P., Wang, J., Crawl, D., Swift, R. V., Malmstrom, R. D., Altintas, I., & Amaro, R. E. (2014). Progress towards automated Kepler scientific workflows for computer-aided drug discovery and molecular simulations*. Procedia Computer Science*, *29*, 1745–1755. https://doi.org/10.1016/j.procs.2014.05.159

Kang, Y. M., Choi, J. E., Komakech, R., Park, J. H., Kim, D. W., Cho, K. M., Kang, S. M., Choi, S. H., Song, K. C., Ryu, C. M., Lee, K. C., & Lee, J.-S. (2017). Characterization of a novel yeast species Metschnikowia persimmonesis KCTC 12991BP (KIOM G15050 type strain) isolated from a medicinal plant, Korean persimmon calyx (Diospyros kaki Thumb). *AMB Express*, *7*(1), 1–12. https://doi.org/10.1186/s13568-017-0503-1

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., Ivanov, P., Avila, S., Abdalla, S., Willing, C., & Jupyter Development Team. (2016). Jupyter Notebooks—A publishing format for reproducible computational workflows. In *ELPUB 2016*. https://doi.org/10.3233/978-1-61499-649-1-87

Lane, J. (2020). *Democratizing our data: A manifesto*. MIT Press.

Lane, J., Mulvany, I., & Nathan, P. (2020). *Rich search and discovery for research datasets: Building the next generation of scholarly infrastructure*. Sage.

Luan, Y., (2018). Information extraction from scientific literature for method recommendation*. arXiv*. https://doi.org/10.48550/arXiv.1901.00401

Peled, A. (2011).When transparency and collaboration collide: The USA open data program. *Journal of the American Society for Information Science and Technology, 62*(11), 2085–2094. https://doi.org/10.1002/asi.21622

Peled, A. (2013). Re-designing open data 2.0. In *Conference for E-Democracy and Open Government*. http://doi.org/10.13140/2.1.3485.7929

Peled, A., & Nahon, K. (2015). Towards open data for public accountability: Examining the US and the UK models. In *Proceedings of iConference 2015*.

Robinson-García, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations*. Journal of Informetrics*, *11*(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019). A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).* IEEE. https://doi.org/10.1109/SNAMS.2019.8931850

Stall, S., Hanson, B., & Wyborn, L. (2016). The American Geophysical Union Data Management Maturity Program*. Geological Society of America Abstracts with Programs*, *48*(7), Article 156-1. http://dx.doi.org/10.1130/abs/2016AM-284514

Starr, J., & A. Gastl, (2011). *isCitedBy:* A metadata scheme for DataCite*. D-Lib Magazine*, *17*(1/2). https://doi.org/10.1045/january2011-starr

Sugimoto, C. R., & Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford University Press.

Vandewalle, P., (2012). Code sharing is associated with research impact in image processing. *Computing in Science & Engineering*, *14*(4), 42–47. http://doi.org/10.1109/MCSE.2012.63

Vannan, S. S., Downs, R. R., Meier, W., Wilson, B. E., & Gerasimov I. V. (2020). Data sets are foundational to research. Why don’t we cite them? *Eos, 101*. https://doi.org/10.1029/2020EO151665

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., . . . Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv*. https://doi.org/10.48550/arXiv.1910.03771

Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J. L., du Preez, F., & Goble, C. (2011). RightField: Embedding ontology annotation in spreadsheets*. Bioinformatics*, 27(14), 2021–2022. https://doi.org/10.1093/bioinformatics/btr312

Wright, G. E., Koornhof, P. G. J., Adeyemo, A. A., & Tiffin, N.(2013). Ethical and legal implications of whole genome and whole exome sequencing in African populations. *BMC Medical Ethics, 14*(1), 1–15. https://doi.org/10.1186/1472-6939-14-21

Yarkoni, T., Eckles, D., Heathers, J. A. J., Levenstein, M. C., Smaldino, P. E., & Lane, J. (2021). Enhancing and accelerating social science via automation: Challenges and opportunities.*Harvard Data Science Review, 3*(2). https://doi.org/10.1162/99608f92.df2262f5

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed A. (2020). Big Bird: Transformers for longer sequences. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

Zhang, Q., Cheng, Q., Huang, Y., & Luet, W. (2016). A bootstrapping-based method to automatically identify data-usage statements in publications. *Journal of Data and Information Science, 1*(1), 69–85. https://doi.org/10.20309/jdis.201606

Zhang, Q., Lu, W., Yang, Y., Chen, H., & Chen, J. (2018). Automatic identification of research articles containing data usage statements. In *Knowledge Discovery and Data Design Innovation: Proceedings of the International Conference on Knowledge Management (ICKM 2017)* (pp. 67–87). World Scientific. https://doi.org/10.1142/9789813234482_0004

Zuo, X., Chen, Y., Ohno-Machado, L., & Xu, H. (2021). How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed central articles. *Briefings in Bioinformatics*, *22*(2), 800–811. https://doi.org/10.1093/bib/bbaa331

Appendices

Appendix A: Finding datasets

The fundamental challenge that is addressed in this article is that data sets are typically not referenced in the bibliography, and must be found in the full text of scientific publications (Vannan et al., 2020). One alternative approach to the machine learning approach described would be to simply do a string search for the text of known data sets in the references. Such an approach would have the advantage of being quick and relatively cheap; the disadvantage is that it would not find data sets that did not match the string exactly, would not find data sets that were not cited in the references, and would not find other cited data sets. In order to understand the potential for string search methods, the team examined Scopus, Elsevier's abstract and citation database launched in 2004 (Chadegani et al., 2013). We find that string search is very precise in finding data sets (precision exceeds 90%), but that it misses the vast majority of known data sets: only about 19% of known data sets can be found using string search on references. The proportions variy dramatically by field. This appendix describes the methodology.

Approach

Precision. The first step is determining the precision of a string search approach. This was estimated by applying string search methods to the entire Scopus corpus (using the REF) call in the Scopus application programming interface) to find the 45 data sets in the Kaggle training set (including 133 data set aliases) and determine how often that approach correctly identifies the data set of interest. Table A1 provides an example of the output of such a search.[8]

Table A1. Example of search output.

Agency	Data Set Name	Data Set Alias	Citing Scopus Record ID	Reference Text
NOAA	Sea, Lake, and Overland Surges from Hurricanes	SLOSH Model	84858649869	National Oceanic and Atmospheric Administration (NOAA), 2008a, SLOSH Model, http://www.nhc.noaa.gov/HAW2/english/surge/slosh.shtml , http://www.nhc.noaa.gov/HAW2/english/surge/slosh.shtml
NOAA	Sea, Lake, and Overland Surges from Hurricanes	Sea, Lake, and Overland Surges from Hurricanes	85061774659	NOAA National Weather Service Sea, Lake, and Overland Surges from Hurricanes (SLOSH): https://www.nhc.noaa.gov/surge/slosh.php

The precision of this approach was estimated by manually checking each matched reference and classifying them as either True Positive (TP)[9] or False Positive (FP). The test was performed on two data sets—Survey of Earned Doctorates (SED) from National Science Foundation (NSF) and Rural-Urban Continuum Codes (RUCC) from the United States Department of Agriculture (USDA). False Positives can occur in cases where data sets with the same name or alias have been produced by other institutions, for example, there is a Canadian Survey of Earned Doctorates. The precision of the string search for the SED was 91.5%; for the RUCC, 99.6%.

Recall. The second step, calculating the recall, was substantially more complex. Recall was estimated by documenting how many publications belonging to the training set were retrieved by searching for data set aliases within the references. Since the Kaggle competition corpus included publications from a variety of sources, recall had to be calculated on the intersection of the Scopus corpus and the Kaggle competition corpus. Eighty-two percent of the training corpus publications show up in SCOPUS after searching on first publication DOI and then publication title if DOI was not available.[10] Of the 2,055 unmatched publications, 1,891 have a publication DOI, meaning that, if they haven't been matched by that field, they are not indexed in Scopus.[11] Figure A1 provides a visual overview of the connections between the three sets of publications, as well as the formula to calculate the overall recall.

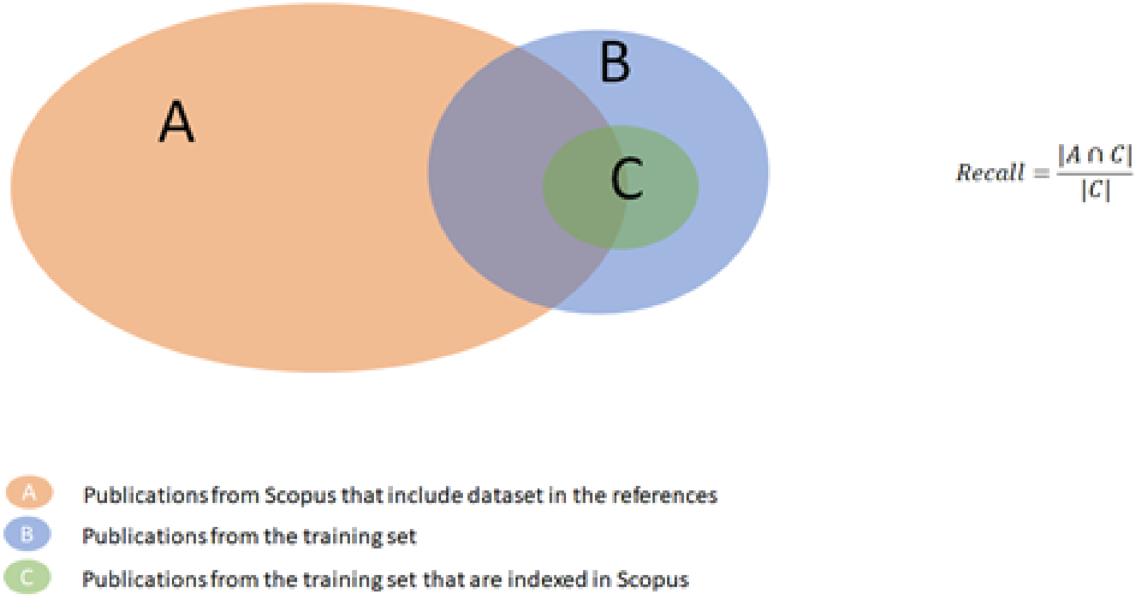

A
B
C
$Recall = \frac{|A \cap C|}{|C|}$
A Publications from Scopus that include dataset in the references
B Publications from the training set
C Publications from the training set that are indexed in Scopus

Figure A1. Measuring recall.

Cases such as this pose the question of whether technical reports, white papers, and so on, that are directly derived from the data set should be counted as data citations. It was agreed to consider these matches as True Positives, at this stage in the investigation.

The overall recall was 19%; the number of publications with data sets in references |A| was 24,927; the number of publications in the training set |B| was 11,340; the number of publications in the training set indexed in Scopus |C| was 9,285; and the number of publications with data sets in references included in the training set indexed in Scopus |A∩C| was 1,762.

Recall was much higher for clearly named data sets and varied by field. For example, for NOAA's SLOSH data, recall was 61%, but it was only 27% for the Survey of Earned Doctorates and 32% for the Agricultural Resource Management Survey.

The low recall seems to confirm the starting hypothesis that researchers typically don't include data set citations in the references. A manual check of a sample of 10 Elsevier articles classified as False Negatives, shown in Table A2, illustrates the finding.

Table A2. Examples of false negatives.

Data Set	Article DOI	Notes
Rural-Urban Continuum Codes	10.1016/j.pmedr.2019.100859	**Data set not included in the references**
Rural-Urban Continuum Codes	10.1016/j.addbeh.2019.01.023	**Data set included in the references** U.S. Department of Agriculture, E.R.S., 2016 U.S. Department of Agriculture, E.R.S. Rural-Urban Continuum Codes https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/ (2016) (Accessed April 23, 2018)

Rural-Urban Continuum Codes	10.1016/S2468-2667(18)30208-1	**Data set included in the references** United States Department of Agriculture Economic Research Service Rural–Urban Continuum Codes https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx (October, 2016), Accessed December 4, 2018
Rural-Urban Continuum Codes	10.1016/j.jsat.2014.07.010	**Data set not included in the references**
Rural-Urban Continuum Codes	10.1016/j.envres.2018.06.020	**Data set included in the references** United States Department of Agriculture, 2017 Rural-Urban Continuum Codes, 2010. https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/ Updated April 28, 2004. Accessed September 1.
Rural-Urban Continuum Codes	10.1016/j.amepre.2017.10.021	**Data set not included in the references**

Rural-Urban Continuum Codes	10.1016/j.ssmph.2017.07.013	**Data set included in the references** Cromartie & Parker, 2016 Cromartie, J., & Parker, T. (2016). Rural classifications. Economic Research Service, U.S. Department of Agriculture. Retrieved June 21, 2016. http://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx
Survey of Earned Doctorates	10.1016/j.neuron.2017.03.049	**Data set included in the references** NCSES and NSF, 2015 NCSES (National Center for Science and Engineering Statistics) and NSF (National Science Foundation). (2015). Doctorate Recipients from U.S. Universities: 2013 (NSF 15-304). https://www.nsf.gov/statistics/sed/2013/
Survey of Earned Doctorates	10.1016/j.econedurev.2016.04.005	**Data set included in the references** National Science Foundation, 2013 National Science Foundation (2013). Survey of earned doctorates. Available at http://www.nsf.gov/statistics/srvydoctorates/#sd Accessed April 4, 2014.
Survey of Earned Doctorates	10.1016/j.respol.2014.12.013	**Data set not included in the references**

Six out of 10 data sets are False Negatives: this suggests that the recall could be significantly increased by including more data set aliases and improving the full-text scanning of the references. An interesting possibility that could improve the results of the machine learning models applied to the full-text would be the presence of a cross-reference in the full-text that might be a signal to consider that would decrease the rate of False Positives.

Appendix B: The Kaggle Competition

In order to develop models that would find such data sets, we hosted a competition using the Kaggle platform—called “Show US the Data” which attracted over 1,600 data science teams. This appendix describes the technical approach that was used.

Defining a Mention

A major challenge was defining a data set mention in developing both the test and training data sets. Then, it is necessary to have a corpus of reliable used data sets, as large as possible and related to the publications that used them. The manual construction of this corpus seems impossible due to the number of existing publications. This leads us to think about developing an automated strategy for identifying data sets in scientific texts.

Three challenges need to be addressed:

1. Multiple data set categories.

The challenge is that there are multiple ways of referring to data. Christian Zimmerman, from RePEc (Research Papers in Economics), has identified four typographies: (1) raw data, (2) events of interest, (3) more processing, and (4) even more processing. Another approach identifies two types. The first is named data sets, which are well-defined, usually large-scale, and publicized data sets, typically referred to with different versions, known abbreviations, and often containing the name of an institution or commercial data vendor (such as the National Health and Nutrition Examination Survey conducted by the National Center for Health Statistics). The second is 'created data sets,' which are usually collected or created by researchers to answer a specific question (small surveys, interviews, or randomized controlled trials). Unlike named data sets, created data sets do not typically have a particular name. Instead, they are often referenced with a description of their content and structure, making it difficult to extract them correctly with traditional text-mining methods.

2. The absence of a standardized data set mention process.

Especially for 'named data sets,' researchers can refer to the same data set in different ways. For example, they can use additional years, samples of the same data set, different versions of the name spelling, and acronyms, among others. For example, the National Health and Nutrition Examination Survey can be referenced as NHANES 1999–2000 Questionnaire Data. Moreover, even within a publication the same data set can be mentioned in different ways: the researcher might start by using “Longitudinal Employer-Household Dynamics (LEHD) data,” and then later just call it “LEHD.” In the same way, the 'created data sets' can be mentioned differently even throughout the same text and depending on the document, the exact text string can be a data set in one publication, and in another, it can simply be the use of English to express an idea. This lack of standardization suggests that the problem can be divided into two steps.

The first step is the data set citation string. In this step, the goal is to identify which text snippets may be referring to a data set. However, solving this problem from a traditional named entity recognition (NER) approach would not adapt well to the existing challenges. The second step is positive and negative case-based learning and identification. It seeks to improve NER performance through simultaneously learning two opposite cases, selecting linguistic characteristics of both cases (NER and non-NER).

Despite several controlled vocabularies, empirical strategies to identify data sets in publications, overcoming nomenclature problems, synonyms, noun phrases, and acronyms are still challenging. For this reason, others compared two vocabulary-based methods (a keyword-based and a Wikipedia-based) and three model-based methods (conditional random fields [CRF], CRF with keyword-based dictionary, and CRF with Wikipedia-based dictionary) to identify data sets in computer science publications. The results show that the model-based methods outperform the vocabulary-based ones.

3. Data set mentions are domain-specific.

Each field of science has a different way of referring to data sets. They have differences in the distribution of data set types (named data sets, created data sets, events of interest, etc.). Depending on the context, names can give more or less information about whether or not it is a data set. For example, named data sets in the social sciences often describe the data set's content (Longitudinal Employer-Household Dynamics–LEHD), unlike those found in genetic tests where the name of the data set can refer to a particular encoding, unnoticed for a nonexpert eye (CYP7A1 association study). This presented a challenge for the manual tagging strategy we used to create the test set for the Kaggle competition.

This heterogeneity in the data set mentions translates into a difference in the performance of the empirical strategies used to solve the problem, according to the scientific field in which it is used. Zhang et al. (2016, 2018) describe implementing a bootstrapping-based unsupervised training strategy based on previous work to distinguish articles with data use and reuse from those without data usage.

4. Data sets mentions could indicate data sets used for analysis or cited.

Based on the first Rich Context competition (Lane et al., 2020) researchers found that data sets could appear in the text in two different ways. They could be mentioned because it was 'used data,' meaning it was analyzed to obtain the results of the publication. However, they can also be 'cited data sets,' meaning they are just mentioned as a reference but not used. Mentioned data sets can be usually located in the literature review or references sections to talk about other results that used that data set or about the structure of the database or its metadata. Due to the absence of a standardized identification of data sets, it is hard for algorithms to distinguish between these data set categories. Although it is an important distinction, the Kaggle competition focused on identifying all data set appearances

Building a Labeled Corpus for the Machine Learning Competition

As discussed, there is substantial field-specific variation in how data sets are named and how authors refer to them. To be able to produce machine learning (ML) models that generalize across different disciplines, the corpus of publications used to train such models needs to include a diversity of fields of study. Since it is not infrequent to see data sets that share their names with the program or organization that produces it, the focus should not be on models 'learning' how data sets are named, but instead on learning the semantic context that signals the author is mentioning a data set. Thus, it is not required to create a corpus with a high diversity of data sets, but instead it is necessary to provide many instances of data set mentions.

The Kaggle team set a minimum threshold of 20,000 publication–data set dyads to provide enough training data to develop generalizable models and also a diverse testing set to be able to assure the correct identification of the best ones. The goal was to have a disjointed split between the training set and the test set.

In order to build the publication-data set dyads two elements were required: the known named data sets and the publication corpus. Given that the full text of the scientific publications needed to be shared publicly as part of the training set, the corpus had to be built from open access publications. A detailed description of the publications' sources used is in the next section.

There were two options for selecting a list of known named data sets: to use an existing catalog or create one specifically for this purpose. Using existing catalogs such as the data.gov portal has the advantage of leveraging a huge amount of data set names, which in turn could make it easy to find a large quantity of publications mentioning them. On the other hand, it presents the risk of having the Kaggle competitors to figure out the data set catalog used as a source, which in turn would make it easier to produce an overfitted solution just to win the competition but without solving the real challenge.

Creating a custom data set list has the opposite tradeoff: while reducing the risk of Kagglers gaming the competition, a custom list of data sets—since it is a shorter list—makes it more challenging to find the required number of publications. Another advantage of this approach is that the selected data sets could be tailored to contain relevant data sets and not just random ones. Because identifying many different data sets did not add much value to this use case, the second approach was the one selected.

At the end of the process, the publication corpus built for the Kaggle competition was composed of publications that mention at least one key data set as defined by our collaborating U.S. federal agencies. The publications were selected through the exact string search of the data set in the publication's full text, as described in the following sections.

Each instance of the training set is a pair of `<Publication, Dataset label>` where the data set label is a text span taken from the publication and acts as the target that must be predicted by the ML models, and the full text of the publication is provided in JSON (JavaScript Object Notation) format.

Labeling Known Data Sets

The definition of the list of known data sets used for labeling of the corpus was part of an iterative process. The initial list of known data sets was provided by subject matter experts (SME) from the partnering U.S. federal agencies. A first pilot phase was done with the collaboration of NOAA’s SMEs who provided an initial list of 14 key data sets related to the coastal inundation topic.

Using the list of key NOAA data sets and aliases, two machine learning models were applied to a corpus of publications (20,000 documents) from the NOAA Institutional Repository. Several issues with the data set names were identified. Some of the data sets’ names and aliases were too general and were used in the sentences that do not refer to data sets. For example, in this sentence a provided NOAA data set name, “water level,” is not used in relation to data: “We established a temporary water level station at each site in accordance with the criteria of the National Oceanic and Atmospheric Administration’s Center for Operational Oceanographic Products and Services (NOAA 2007).” In another example, an acronym related to a data set title “Tsunami Forecast Models,” MOST, didn’t match the actual data set in this context: “In addition, they used LES results to formulate the mixing length beyond the surface layer where Monin-Obukhov similarity theory (MOST) is considered valid.”

Because of such instances, the list of originally provided data set names was revised (Appendix Table B1). An agency name was appended to the data set name, and the acronyms were removed from the search. Some of the very general data set names were also completely removed from the list (such as “Global Sea Level Rise” and “Sea level rise rates”).

Table B1. Example of the initial and revised data set names.

Initial			**Revised**		
Data set title	**Aliases**	**Acronym**	**Data set title**	**Aliases**	**Acronym**
Tide gauges / water levels	water level, tide gauge	N/A	NOAA tide gauge	NOAA tide station, NOAA tidal station	N/A
Tsunami Forecast Models	Method of Splitting Tsunami model, tsunami inundation model	MOST, SIM(S), SIFT, TFS	Tsunami Forecast Models	Method of Splitting Tsunami model, tsunami inundation model	Removed

Note. The agency name (NOAA) is appended to the beginning of the data set name to make it more specific. Acronyms are removed in order to avoid noisy results.

After the NOAA implementation, a list of 10–15 key data sets was obtained from other partnering agencies. Based on the lessons learned with the NOAA data set names, the data set names from the other agencies were revised as well. For example, the “Food Expenditure Series” data set from USDA was used in the string search using two conditions: the agency name was appended to the beginning of the data set name (“USDA Food Expenditure Series”), or the search was performed with the condition that both “Food Expenditure Series” and “USDA” co-occur in the same text, in order to prevent noisy results. More about the way that the string search was performed using the finalized data set list is described in the next section.

Labeling Using String Search

Data set labels must be prepared in order to serve as the ground truth for models in the Kaggle competition to learn how to detect them from the corpus. The first step is to label the known data sets provided by the agencies. One intuitive option is to directly search the entire corpus sequentially and label when the text matches the data set names. However, it is too time-consuming, because the option requires iteration on the entire corpus, and when it comes to processing data over 180 gigabytes, this option becomes impractical. Thus, it is necessary to apply a methodology that searches data set matches on millions of academic publications efficiently. To meet this need, the Elasticsearch (ES) engine was used (Gormley & Tong, 2015). ES is a search engine based on the Lucene library, which is highly efficient for doing string searches on full texts because of its indexing mechanism. It transforms text into inverted indices, therefore dramatically speeding up the searching process. As mentioned in the previous section, all the publications were preprocessed into the same format, preserving the section structure of the original document (section title—section full text). Then we index them and search for text occurrences of the data set name and aliases using the ES engine.

Additionally, we manually tag those data set names and aliases that could potentially lead to false positives and perform a more constrained search that requires the data provider names show up close to the query search to ensure that the hits refer to data sets.

The string search is performed using two approaches: (i) if a data set title or alias are specific enough, the exact string match is performed; (ii) if a data set title or alias are too general, the string search is performed with additional constraints. For example, the “Census of Agriculture” data set title could be too general, therefore, in the string search additional terms are added with “AND” condition, that is, both of these terms should appear within the publication: “Census of Agriculture” AND “U.S. Department of Agriculture,” sand so on (Table B2). Two data sets from NSF/NCSES, "Survey of Earned Doctorates" and "Business Enterprise Research and Development Survey,” required additional terms in the search, as there are surveys with the same names in other countries (there is a Survey of Earned Doctorates that is conducted in Canada, and there is a Business Enterprise Research and Development Survey that is conducted in the United Kingdom).

Table B2. Example of a data set entry for the string search.

Data set title	**Aliases**	**Agency**	**General name flag**	**Additional terms, if general name flag is True**

Census of Agriculture	USDA Census of Agriculture;
	NASS Census of Agriculture
		USDA	True	U.S. Department of Agriculture;
				National Agricultural Statistics Service;
				NASS

Early Childhood Longitudinal Study	N/A	NCES	False	NCES; National Center for Education Statistics
Survey of Earned Doctorates	NSF Survey of Earned Doctorates, NCSES Survey of Earned Doctorates	NSF/NCSES	True	National Science Foundation, NSF, NCSES, National Center for Science and Engineering Statistics

Business Enterprise Research and Development Survey	NSF Business Enterprise Research and Development Survey; NCSES Business Enterprise Research and Development Survey	NSF/NCSES	True	National Science Foundation, NSF, NCSES, National Center for Science and Engineering Statistics

Note. Because the “Census of Agriculture” data set title could be too general, additional terms are added to the string search: “Census of Agriculture” AND “U.S. Department of Agriculture,” and so on. Note that for the “Early Childhood Longitudinal Study” title we do not use an alias “ECLS,” as in the string search this match can become noisy (i.e., not necessarily refer to this data set). The data set name itself is specific enough in this case and doesn’t require additional terms (the general name flag is False). Two data sets from the National Science Foundation (NSF)/National Academies of Science, Engineering, and Medicine (NCSES), “Survey of Earned Doctorates” and “Business Enterprise Research and Development Survey,” required additional terms in the search, as there are surveys with the same names in other countries.

After the string search was performed, we took random samples of 20 text snippets (each one from different publications) where the alias was present and we reviewed to check if it contained noise. If some noise was detected, a larger sample would be reviewed. If the noisy mentions were 10% or more, the alias would be marked as noisy and removed from the corpus.

For example, a data set title from the United States Geological Survey (USGS), Integrated Taxonomic Information System (ITIS), is in some cases clearly used as a data set, but in more than 20% of cases, was referred to as ‘standard’ to follow, or the organization maintaining it (Table B3).

Table B3. Example of correct and noisy data set mentions.

Correct data set mention	Noisy data set mention

"Each group of identified microorganisms is presented according to the ***Integrated Taxonomic Information System*** *(ITIS) taxonomic counting heat maps. Microsoft Excel (Redmond, WA, USA) was used for data analysis."* (Kang et al., 2017)	*"The* ***Integrated Taxonomic Information System*** *is "a partnership of federal agencies formed to satisfy their mutual needs for scientifically credible taxonomic information" (www.itis.gov)."* (Drew, 2011)

Note. In the first case, ITIS is used in the context of a data set reference, in the second case, ITIS is referred to as an organization/partnership.

Splitting the Corpus Into Training and Test Sets

Once the labeling of known data sets was finalized, the training and test sets were created. The split requirements were: 1) to be disjoint and 2) to keep the fields of study diversity on both sides. The split was disjoint in both publications and data sets to discourage competitors from approaching the challenge by implementing lookup tables. Both the training and testing sets contained publications from all the fields of study present in the publication corpus to enable the creation of ML models that generalize to a variety of different writing styles. This distribution was achieved using the data sets as a proxy, more precisely, the partnering agency that recommended each data set.

Labeling of Other Data Sets

Once the full corpus of publications was labeled with the known data sets, the resulting <publication-dataset label> dyads were split to create the training and the testing sets, as previously described. In the case of the testing set, all the other data sets that appear in the publications needed to be labeled. Otherwise, if a model predicts a correct but not labeled data set, it will be incorrectly penalized by considering the prediction is false (a false positive).

However, it was impractical to read through the full content of all the publications manually, so we exploited the fact that when authors mentioned one data source that they used, they tended to mention other sources in the surrounding context as well—creating a 'hotspot.' A named entity recognition (NER) model was used to identify such hotspots. A tagging spreadsheet was created where each row had a publication ID, a predicted data set label, and the specific publication text snippet where the data set label was found. This way, the task was reduced to a manual review of specific text chunks across the raw corpus identifying positive hits of data set names and aliases. Ten percent of the NER results were manually validated and completed, then the string search with the newly found data set names from the manual review was rerun on the entire full-text corpus. This allowed the capture of as many data set labels as possible from the corpus and generated more text snippets for further review. This work had two goals: (i) to validate whether the predicted label is correct and (ii) to label other data sets that escaped the NER model.

In order to standardize the labeling process among the team of manual reviewers, certain protocols were followed in order to obtain reliable results and aligned with the main purpose of the competition: to find data sets in the corpus without using a lookup-table approach based on known data sets. One of the most important rules was to label each text snippet based on the context and not based on prior knowledge. The manual reviewer would only label the string as a data set label, if they could infer from the context that it was used as a data set, rather than if they personally knew about that particular data set or identified that data set in other snippets. For example, consider the “1000 Genomes Project” mention in the following text snippet: “In general, the informed consent form templates used for the genomic studies in Africa are not available to external researchers, except for those used by large-scale projects of human genetic variation (i.e., the HapMap Project and the 1000 Genomes Project) as well as the Malaria Genomic Epidemiology Network (MalariaGEN)” (Wright et al., 2013). Even though “1000 Genomes Project” is mentioned as a data source by authors in many cases, a reviewer would not label this instance, as there is no surrounding semantic context that would identify this entity as a specific data source used in this publication (such as “the data was obtained from,” etc.) As another example, mentions like “Trends in International Mathematics and Science Studies” and “Program for International Student Assessment” sometimes could refer to data sets, but they could also refer to tests, or standards in general, depending on the context. It is crucial to make sure the data sets are labeled based on the semantic context and not the ‘mental lookup table’ a person starts to build after reading so many documents.

The labeling and validation protocols were very time intensive. New labelers were trained prior to the start of their review work. During this period, a senior member reviewed all the labeling work and iterated with the labeler to be sure that the labeling protocols were understood and implemented. If any member of the reviewing team was not sure whether the string could be inferred as a data set, they were encouraged to flag and leave the record for the senior members of the team to review and discuss. After all the labeling work was done, there was a final review by the senior team to check if the labeled strings indeed appeared in the full text of the corresponding publication. That review also corrected typos or mistakes during the manual labeling work. Due to time constraints, no inter-labeler correlation, or kappa, was calculated, however.

Labeling Issues and Possible Additional Approaches[12]

A major issue with labeling a corpus is that it is fundamentally difficult to agree on a common definition of a data set mention, both because authors differ in naming data sets and because there are multiple nomenclatures both within and across fields. In addition, data set definitions need to be determined in advance and be immutable, since any change in the definition might diminish all previous labeling efforts.

Another issue is that labeling has some gray areas, particularly the generalizable context-based labeling approach adopted for the Kaggle competition. The Kaggle competition was intended to find data sets across multiple fields of science. As such, the labels did not incorporate field-specific knowledge, such as sequence of words specific to a field.

Finally, the subtle boundary between ‘possible and impossible to infer from the context’ was not captured—future work might include ‘impossible to infer from context mentions’ as a separate label.

There are several possible alternative approaches. For example, a multiclass or multilabel markup scheme that could cover all cases (named and created data sets could be two distinct categories). Furthermore, as mentioned, distinction between labeled inferable/noninferable from context can be also treated as a classification problem. It is obvious that classification-like markup requires more labeling effort. From this perspective, construction of the minimal acceptable set of classes is a cornerstone problem.

Evaluation Metrics

In this section, we will discuss each part of the evaluation metrics including why we choose the metric and the pluses and minuses of the choice.

Jaccard Similarity

Jaccard similarity is a commonly used proximity measurement for computing similarities between two text objects. It is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Jaccard similarity has obvious advantages over simple exact-string-match methods. It does not penalize a prediction simply because the prediction is missing or added one insignificant word. For example, the “National Health and Nutrition Examination Survey” is one of the data sets that appears several times in the corpus. If the prediction is ‘National Health and Nutrition Examination’ or ‘the National Health and Nutrition Examination Survey,’ it is fairly close to the target text, so in this specific case, we should consider it a valid prediction. It certainly should have a different score if another prediction is entirely irrelevant, such as ‘food policies.’ Another reason why we need to use string similarity metrics is that the corpus sometimes has typos or parsing errors when converted from PDF to plain text, such as parsing ‘database’ into ‘data base’ or simply having typos in the words. Although there are other popular string similarity metrics for validation, such as Hamming distance, Levenshtein distance and cosine similarity, we chose to use the Jaccard score, as the Jaccard evaluation model was already implemented in the Kaggle system.

The particular Jaccard method we applied is a tokenized-based metric, which doesn’t suffer from the huge impact on different string lengths. But it also has its intrinsic shortcomings, one of which is that it cannot capture similarity within the words. Therefore, if the prediction contains a word slightly different from the word in the target, for example, if the prediction word is ‘surveys’ in the last example, it will be regarded as a mismatched word although it is, in fact, similar to the word ‘survey’ and should be treated differently if another prediction word is ‘policy.’

Another shortcoming of the Jaccard method is that it is not sequence-sensitive, which means ‘National Health Survey’ and ‘Survey Health National’ are regarded as the same string. To solve this, we added another hidden function to ensure that the word sequence is correct before a prediction can be marked true.

FBeta Score

F-score is one of the standard and most widely used metrics for recognizing named entities, such as data sets. It is a measure of a test's accuracy, which is calculated from the precision and recall of the test. The precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. The F1 score is the harmonic mean of the precision and recall, while the more generic Fβ score applies additional weights, valuing one of precision or recall more than the other. For example, the more the beta decreases, the more the model will be penalized by precision than recall.

Here we use Beta = 0.5, which means the model is required to be more sensitive to precision than to recall. To put it in a simple way, we prefer to have each predicted data set to be correct rather than having as many data sets to be predicted as possible. This decision made it more likely that the data sets that are detected from the corpus and presented to the federal agencies and used by researchers are correct. In future work, we expect both agencies and researchers to contribute to the corpus.

Calculating the Score

For each publication's set of predictions, a token-based Jaccard score is calculated for each potential prediction/target pair. The prediction with the highest score for a given target is matched with that target. Predicted strings for each publication are sorted alphabetically and processed in that order. Any scoring ties are resolved based on that sort. To make it a classification task, we created a threshold for the Jaccard similarity.

Any matched predictions where the Jaccard score meets or exceeds the threshold of 0.5 are counted as true positives (TP), the remainder as false positives (FP). Any unmatched predictions are counted as false positives (FP). Any targets with no nearest predictions are counted as false negatives (FN).

A simplified example is helpful. Suppose that a target is ‘ABC|DEF|XYZ.’ Here each letter represents a word, and ABC, DEF, and XYZ are valid data set labels (indicating there are three data sets in the publication). If the corresponding prediction is ‘AB|DGH’ (two data sets), then ABC and AB make a pair, and the Jaccard score is 0.67, thus AB is considered a TP; DGH and DEF also make a pair, but the score is 0.33, which is below 0.5, hence it is an FP; and for XYZ, there is no matching prediction, so it is an FN. All TP, FP, and FN across all samples are used to calculate a final micro F0.5 score.

Note that for micro F0.5 score, each prediction/target pair is a unit and used to calculate a score for the entire set of predictions, whereas the macro score will first decide if a record is true based on the prediction/target pairs in that record, then the total score will be calculated based on the results of each record instead of each pair. The micro score was chosen because multiple targets are common in the corpus and it is desirable that the score precisely reflects how capable the model can capture each data set.

Future Research Agenda/Discussion

We believe this article and competition represent the first steps in what should be an important new research agenda. As noted by the winners and by the Kaggle competitors, there is a great deal of work to be done. Most obviously, much work needs to be done to improve natural language processing (NLP) models, given that the overlap across the three winning models was so small. Another major focus should be to build models that can identify whether a data set was actually used in the publication or just mentioned (such as being a part of the literature review) based on the semantic context.

The structures of the training and test data sets also require much work. Although the data science community always makes clear the importance of carefully labeled training data, there are at least two implementation challenges. One of these is defining a data set, which is, as noted above, not clear-cut.[13] Of course, this is a fundamental challenge in all science, going back to Linnaeus, and the goal should be to develop data set definitions that are “backed by the community and firmly linked to end-user demand” (Godfray, 2007 p. 260x).

Another implementation challenge is labeling data sets. There are pros and cons to the context-based labeling approach. The advantage of this approach is in the fact that it allows measuring generalization instead of knowledge. It also helps to make the task easier and improve quality by not trying to fit what is impossible to fit. Some of the limitations of this approach include the fact that it makes it hard to evaluate models that have knowledge (e.g., an ensemble of a known data set list and some neural model). Also, a model pretrained on a large collection of scientific texts may obtain some intrinsic knowledge about a particular sequence of words. For example, if only in half of publications it is possible to understand that the sequence of words is a data set, the model can still obtain this knowledge and use it even when the context is not showing that this is a data set. Finally, there is a subtle boundary between possible/impossible to infer from the context. During the labeling, it could be beneficial to leave ‘impossible to infer from context mentions’ as a separate label, which can be dropped in construction of the train set.

A third is to incorporate more metadata into the labeling. It is intuitively obvious that data sets are more likely to be mentioned in certain parts of a paper than others. It would be helpful to incorporate the position in the text of the data set mentions as features, for example, the section titles (Introduction, Methodology, etc.). Adding more features could help generalization by allowing the models to adjust to different fields. Features could include: field of study (for example, based on the journal) or research topic, section title names where each data set is mentioned, and publish time.

Supplementary Files

The GitHub repository with the winning models and their documentation: https://github.com/Coleridge-Initiative/rc-kaggle-models

The training set data is published in the Kaggle competition: https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data/data

Footnotes

1. https://www.congress.gov/bill/115th-congress/house-bill/4174. ↩
2. The National Science Foundation's National Center for Science and Engineering Statistics (NCSES), the Economic Research Service at the U.S. Department of Agriculture, and the National Oceanic and Atmospheric Administration at the U.S. Department of Commerce. ↩
3. Future work could include government and media reports are similarly likely to either directly mention the data sources or refer to scientific reports. ↩
4. https://www.Kaggle.com/c/coleridgeinitiative-show-us-the-data. ↩
5. U.S. Department of Agriculture, National Oceanic and Atmospheric Agency (NOAA), NCSES at the National Science Foundation (NSF), National Center for Education Statistics (NCES), U.S. Geological Survey, and National Institutes of Health. ↩
6. The F-beta score is the weighted harmonic mean of precision (the proportion of predicted positives that are correct) and recall (the proportion of all positives that

are correct). The highest possible value is 1 and the worst possible value is 0. More details on the metric choice and rationale are provided in Appendix B. ↩

7. https://fsc.org/en/for-forests/high-conservation-values#the-hcv-approach. ↩

8. Sometimes the string search pointed to a publication that included the data set name; these searches were discarded. The frequency with which this occurred was typically less than 10%, but in the case of the SLOSH (Sea, Lake, and Overland Surges from Hurricanes) data, it was over 40%. ↩

9.

This manual classification exercise highlighted an interesting problem related to data set citations: in many cases, the reference to the data set is indirect, meaning that the reference points to a report, or other derivative work, that is based on the data set itself. For example, the article

Costabile-Heming, C. A. (2011). Responding to the MLA report: Re-contextualizing the study of German for the 21st century. *German Quarterly*, *84*(4), 403–413. https://doi.org/10.1111/j.1756-1183.2011.00123.x

cites a report based on the Survey of Earned Doctorates (SED) in the full text as follows:

“In 2009, the number of doctoral degrees awarded in German represented a decrease of 23.7% over 1995 (*Report on the Survey of Earned Doctorates 3*)”

and the corresponding reference in the bibliography is the following:

Modern Language Association (MLA). Report on the Survey of Earned Doctorates 2008-09. April 2011. Web. 8 Sept. 2011. ↩

10. All instances of multiple matches, either via DOI (which are due to errors in the Scopus metadata) or by title (which happens especially with very short titles such as 'dementia' that was matched 1,940 times) were discarded. All other cases (i.e., individual matches either via DOI or title) have been considered valid. On the other hand, we have noticed that in many instances of valid matching by DOI, the title of the article in the training set and that of the record in Scopus are quite different, because of truncations, translations, and so on. ↩

11. For example, the 161 publications in the training set published in SSRN (a preprint server) are surely not indexed in Scopus, because preprints are not part of

the core Scopus index. This leaves 164 publications without a DOI that could have been matched by title with fuzzy matching approaches (e.g., Levenshtein distance), but this would have in the best case contributed to a 1% increase in recall, so we preferred to leave them out to avoid increasing False Positives. ↩

12. This section draws heavily on contributions made by Mikhail Arkhipov ↩

13. Training sets could include multiple records with each record including only the text snippets around one dataset mentions, potentially classified as ‘data set used,’ ‘data set mentioned,’ and ‘not a dataset,’ which could be served as the corpus for next NLP model. ↩