

UC-Irvine ML's Meta-Dataset

An analysis of the datasets made available through UC Irvine's ML repository website.



Introduction

This project examines the University of California Irvine (UCI) Machine Learning Repository (MLR) to identify and explore and analyze information about the data sets within. Throughout the project, we will transition from information to analysis of machine learning data, based on observable changes over time in the contents and uses of the MLR. From this we can create knowledge on recent trends in machine learning research, and assess some predictions about future activity within this domain.

The Dataset(s)

The UCI MLR defines itself as “a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms” (Dua and Graff). First created as an FTP archive in 1987, the repository is a widely-used resource for researchers, particularly in academic pursuits toward machine learning. In 2007, the repository transitioned to its current version as a website, and has grown considerably in size of data since then.

To access the data, we used the Python package BeautifulSoup to scrape the data from the web. We saved the metadata for over 500 datasets into a comma separated file for analysis. First the “View All Datasets” page was scraped for hyperlinks and basic summary information, and then the individual page for each dataset was scraped to provide more fine detail and context for each dataset. Finally the dataset file links were checked and HTTP file_size was collected for every dataset. By combining all this data together, we constructed one robust and complete dataset to represent all the metadata describing each individual dataset provided by UCI MLR.

Preprocessing

Upon retrieving the data, it was immediately clear that in order to extract contextually useful information, or ideally functional knowledge, we needed to implement a range of data cleaning

Most Popular Data Sets (hits since 2007):

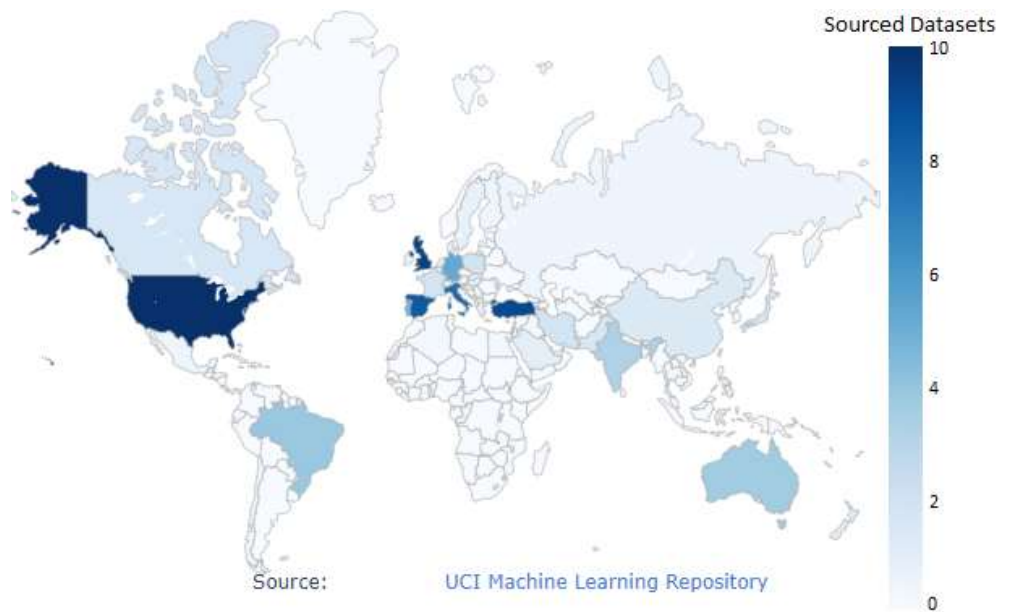
3457536:		Iris
1889000:		Adult
1459154:		Wine
1299338:		Breast Cancer Wisconsin
1278485:		Heart Disease

procedures. This preprocessing involved filling in missing and incomplete values with appropriate replacements, denormalizing columns of multiple variables into separate indicator columns, and converting data to the appropriate type. In cases where there was no way to fill in missing values in critical columns, we had to eliminate the data point from our analysis - this was approximately 5% of the originally scraped data. We were also able to extract location and contributor data from the citations in each dataset. Our efforts resulted in a dataframe with 45 attributes describing 472 data sets.

Data Structure

The metadata we collected includes the size of each dataset, including both the number of observations and the number of attributes for each observation, the type of data, such as image, time series, text, or spatial, and the type of attributes such as categorical or integers. Additionally, the metadata contains the machine learning task associated with each dataset, such as regression or classification. We transformed the date that each dataset was given to the site into year categories, as well as an “age” for each dataset. Regarding origin, we carefully extracted data from the “Source” section for each dataset, including the Institution of origin (e.g. Aarhus University) and then meticulously researched and attached relevant City, Country, and Country Code (Aarhus, Denmark, DNK). Finally, we scraped the pages where the final real datasets and associated metadata were provided by UCI ML. In our scraped dataset, this is represented by relative URL suffixes, a list of dataset filenames and sizes, a “small” indicator assessment, and a “shortname” uniquely identifying the dataset for user input if needed. The “shortname” field required some manual massaging to ensure that the unique names were unique, short, and contextually meaningful.

To test that our data cleaning was performing as intended, we tested every function with a small manually created dataframe that would allow us to review the results and check outputs of specific lines against their expected values.



Data Exploration and Analysis

Geographically, the datasets are very diverse, but not all areas are equally represented. Over 40% of the datasets are from the United States and approximately 55% is from english speaking countries, heavily skewing our data towards the US and other english speaking countries. However,

six continents are represented in the MLR by 48 countries. While conclusions we draw based on this data will primarily apply to the US, we will assume that they may be broadly applied to global trends in machine learning.

The MLR is categorized by areas of interest, which we may use to draw conclusions about the applications of machine learning. If students, educators, and researchers exhibit particular interest in a subject, it likely follows that we will see machine learning develop towards these subjects. Perhaps unsurprisingly, “Computer” subjects make up the majority, with 176 unique data sets currently stored in the repository, followed by “Life”, with 108 data sets (See Figure 1). Interestingly, the “Life” area of interest is the most popular category in terms of web hits since the MLR’s inception. Despite a greater number of “Computer” data sets available, “Life” data sets have garnered 19,495,730 web hits compared to 15,417,400 web hits for the “Computer” sets (See Figure 2).

The top 5 most popular data sets in the MLR since 2007 are:

1. Iris, a “*Life*” data set donated in 1988
2. Adult, *Social*, 1996
3. Wine, *Physical*, 1991
4. Breast Cancer Wisconsin, *Life*, 1995
5. Heart Disease, *Life*, 1988

With 3 of the top 5 most popular sets representing “Life”, one could expect a significant development of machine learning within medical and biological fields. Although the interest and potential may exist, however, a recent assessment states that “There is a stark contrast between the lack of concrete penetration of AI in medical practice and the expectations set by the presence of AI in our daily life.” (Cosgriff, Stone, Weissman, et al). The reasons for this ineffective implementation go beyond the scope of this project. It is assessed that due to their cleanliness and accessibility, several of the most popular data sets have

Figure 1

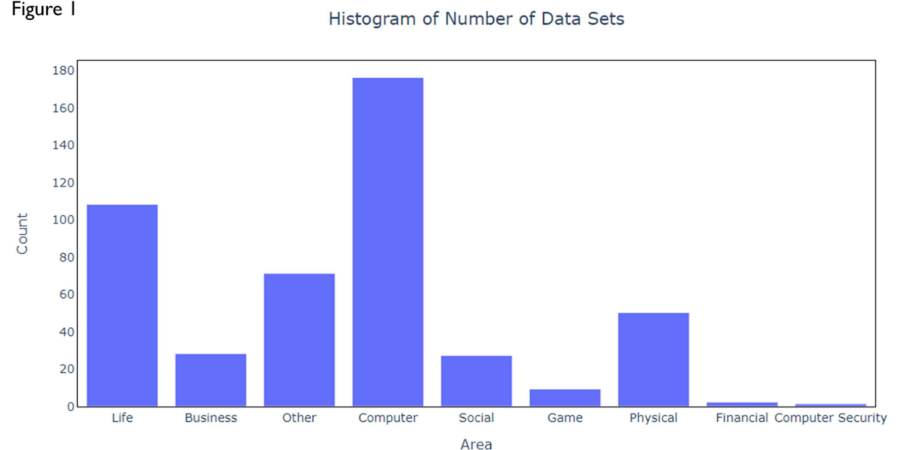


Figure 2

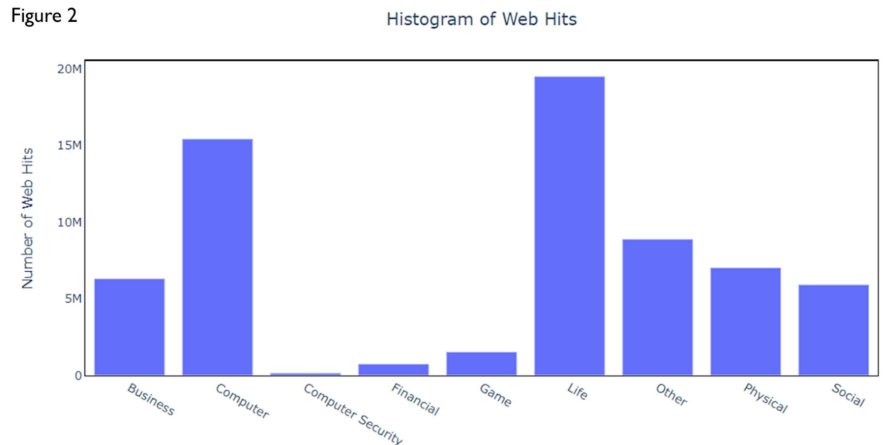
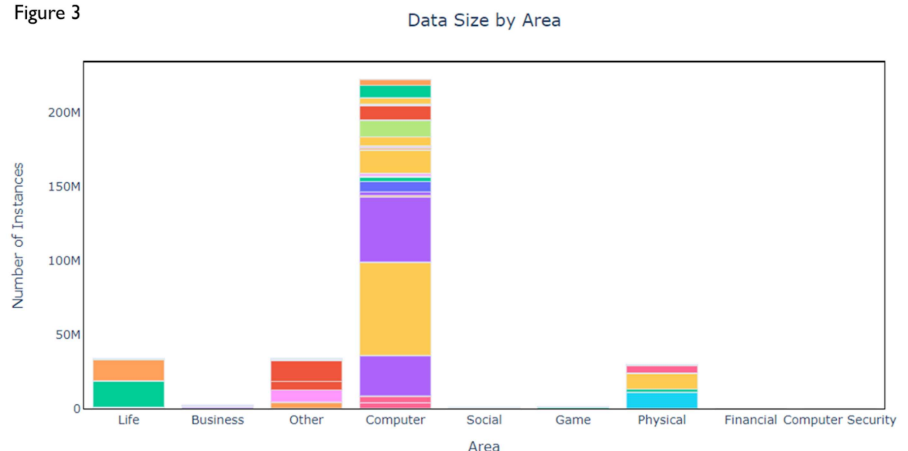


Figure 3



been used as examples for academic research, so their popularity may be due more to convenience of instruction than to intrinsic interest in the subject matter.

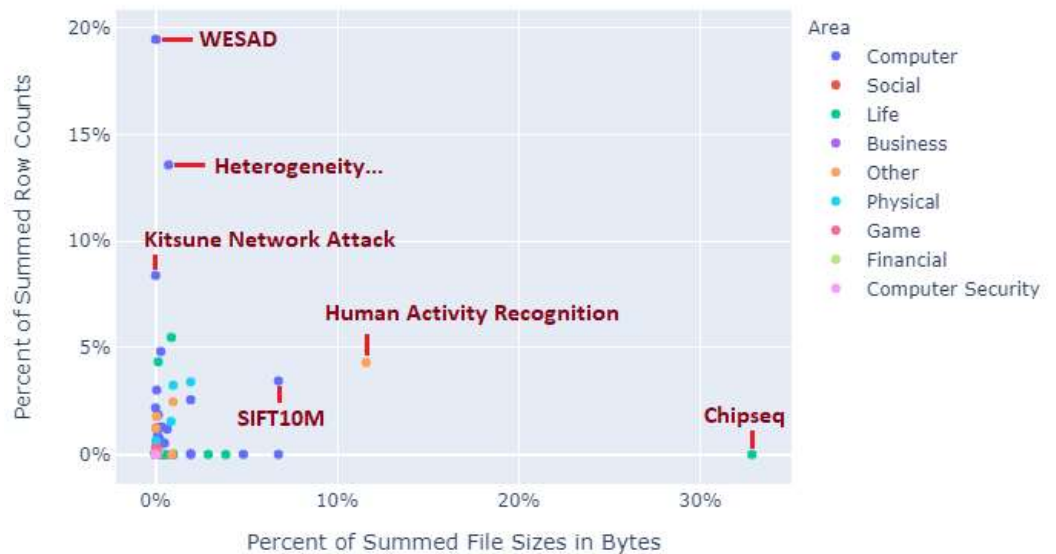
In terms of data size, “Computer” data collectively averages 1,265,119.8 rows of data per data set, making it by far the largest category in terms of row size. The second largest is “Physical” with 591881 average rows (number of instances) of data; “Life” ranks fourth behind “Other” with 310184.7 average rows (See Figure 3). One may interpret this information as partly owing to the ease of collection of computer data, as it generally exists in forms of system or application outputs and is easily ingested for analysis.

“Physical” or “Life” data often relies on historical documentation about the natural world, which may need to be digitized and formatted before it can be readily imported for analysis. In fact, some of these data sets are based on decades-old scientific studies, which have been digitized and subsequently donated to the repository. It is no surprise that more recent data sets, particularly those focused on computers and automated processes, are much deeper in terms of available information.

Two datasets within the “Computer” domain have notably large row counts: the WESAD (Wearable Stress and Affect Detection) and Heterogeneity Activity Recognition data sets. Each of these data sets focuses on recording and analyzing human physical activity and response through the use of worn sensors. The WESAD

dataset comprises 19.5% of all UCI MLD instances, and the Heterogeneity dataset makes up 13.6%. These two data sets have more rows than any entire category besides their own. The WESAD data set has 63 million rows of data, and the Heterogeneity Activity Recognition set has over 43 million rows. Although these two sets account for a large portion of the overall instances, the “Computer” category would still dwarf the other categories without them.

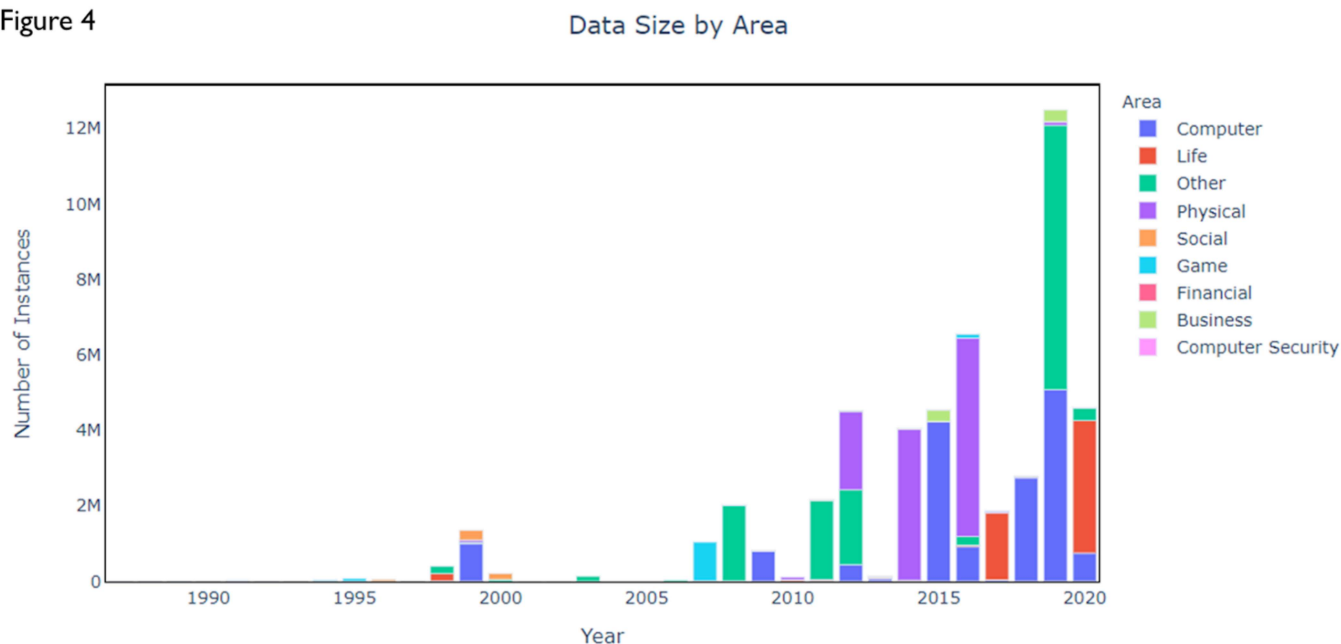
In terms of raw data, in bytes, the “Chipseq” dataset from McGill University dwarfs all others at 34GB, which out of about 103GB total (from all datasets, including metadata) is fully a third of all of UCI ML’s available data. The next highest amount of data is the SIFT10M dataset. Per Wikipedia: “The scale-invariant feature transform (SIFT) is a feature detection algorithm in computer vision to detect and describe local features in images.” Incredibly, 34GB is the compressed version of the data, as “each SIFT feature is a 128D column, and the corresponding patch is saved in 41*41 png format. The png files are compressed into 307 tar files for downloading.” The magnitude of these freely available datasets is staggering.



Results/Development Over Time

The UCI MLR has grown significantly in terms of data sets by size since its transition to the current website version in 2007. This closely follows Geoffrey Hinton's creation of the term "deep learning" in 2006, a concept of algorithms associated with complex Artificial Intelligence development (Roboticsbiz). However, this growth became notably greater from 2015 until the present (See Figure 4). This growth in data size correlates to major technological developments within industry and

Figure 4

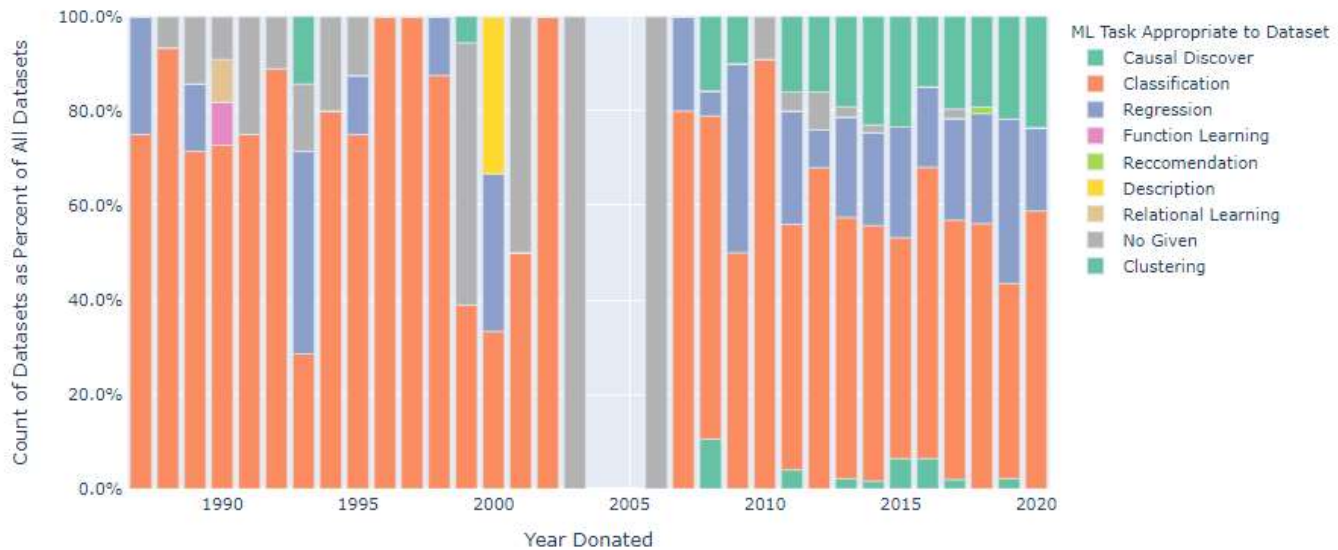


research fields. In 2015-2016, Many important open source tools and frameworks, including TensorFlow, Keras, OpenAI, and PyTorch were released (Leapfrog Technology). Amazon's AWS Machine Learning services launched, along with Microsoft's Distributed Machine Learning Toolkit (Roboticsbiz).

These advancements in cloud-based distributed computing have made learning about and acting on Machine Learning (ML) and Artificial Intelligence (AI) more accessible and affordable. This ground-level growth in inexpensive and easily available capabilities and skills means that anyone could use publicly available data sources for complicated distributed computing and analysis. We believe this growth in public capabilities contributed to the increased sizes of data sets being donated to the UCI repository. Other technological advances have had similar effects: the previously mentioned Heterogeneity Activity Recognition and WESAD data sets were enabled by growth in motion sensing technology, building on the release of Microsoft Kinect in 2010.

Growth is not a one-way street. How do new tools and technologies become established in industries and serious government projects? We believe the answer is what we are doing right now: students being exposed to easily available datasets and technologies and being taught how they are used. Soon we will be embedded in many industries as the knowledgeable professionals on these subjects, and this will come with us. In this way, another possible relationship is between dataset growth and how that impacts students who become professionals over time.

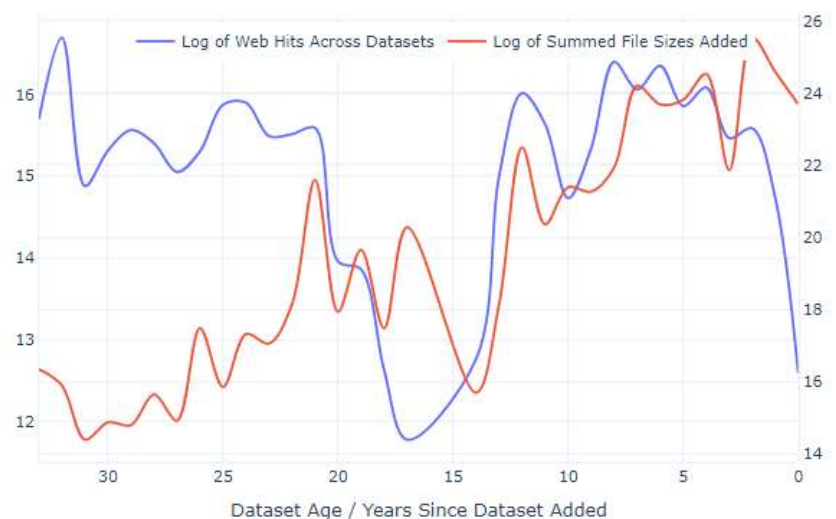
Dataset Count by ML Research Area by Year Submitted



Looking at dataset counts over time as percentages of all datasets submitted in that year, we can see in particular that **Causal Discovery** tasks have grown steadily since 2007. Causal Structure Discovery (CSD) is the problem of identifying causal relationships from large quantities of data through computational methods. This followed the previously mentioned “deep learning revolution” in 2006. Regression and classification, both established standards for analysis in any field, have stayed steady since the UCI MLR re-opened.

The least established dataset task types available on the UCI MLR are **Recommendation** and **Clustering**. We would like to see if these data analysis types will grow over time as we become exposed to them. However, it’s hard to tell exactly how much students are looking at these websites because we don’t know how their popularity has grown over time. The website only provides the total number of webhits for the entire lifespan of these datasets. However, as we can see in the graph, even as the log sum of file sizes donated per year has increased steadily and dramatically over time, the amount of visitors to these datasets is not descending dramatically over time as we would expect if people were equally visiting datasets over time. Instead, the log of web hits across datasets has stayed more or less even with the exception of when the website was down. This tells us that visitors are always checking out what’s new, and for that reason we feel safe in thinking that as data type donation has changed over time, so too have the interests of those who visit.

Webhits vs Count of Datapoints Added by Dataset Age on Website



UC_Irvine_datasets Class

With some intuition and familiarity, there are many possibilities with these datasets. However, most users are mostly concerned with perusing the datasets, and want to access them as quickly as possible. For that reason, we created the class of *UC_Irvine_datasets()*. This class allows the user to access, restrict, clearly print, and ultimately automatically load some datasets that are available (if they are not big datasets!).

```
# Create an instance of the class, which loads the dataframe of UC Irvine datasets
ucid = UC_Irvine_datasets()
```

```
ucid.list_all_datasets()
```

string representation

The string property allows users to understand the current state of the class object.

```
In [2]: print(ucid)
```

```
count of datasets: 471
avg age: 10.7 yrs
median # datapoints: 43890.0
```

```
## Dataset content Areas: ##
```

Area	Business	Computer	Computer Security	Financial	Game	Life	Other	\
Index	28	175		1	2	9	108	71

Area	Physical	Social
Index	50	27

```
## Here are the number of rows in each category: ##
```

multivariate_data	372
time_series_data	100
data_generator_data	5
domain_theory_data	18

```
there are ## 471 ## datasets returned
```

ID	Title
00480	2.4 GHZ Indoor Channel Measurements
00246	3D Road Network (North Jutland, Denmark)
00512	A study of Asian Religious and Biblical Texts Dat...
00314	AAAI 2013 Accepted Papers
00307	AAAI 2014 Accepted Papers
abalone	Abalone
abscisic-acid	Abscisic Acid Signaling Network
00445	Absenteeism at work

The *UC_Irvine_datasets* class loads the dataset we have accumulated and gives the user tools to access all dataset information, and if the dataset is small and simple enough to parse, we can automatically load the dataset into a dataframe for instant analysis. The string representation shown above introduces the data types available, and basic summary information about all current datasets loaded. This makes it easy to understand how many datasets are available in each domain available and what general volume of data is contained.

Available Methods:

- **object.list_all_datasets()**: returns text output of ID values and Titles of datasets in the object.
- **object.to_df()**: output the underlying dataframe of the object.
- **object.small_datasets_only()**: returns a new object containing only datasets of size less than 1MB and in a format pandas can easily read.
- **object.load_small_dataset_df(dataset_ID)**: returns a new dataframe containing the dataset requested.
- **object.show_me_dataset(dataset_ID)**: returns a text summary of a single dataset's metadata.
- **object.limit(field name, restrict to this value)**: changes the underlying data to restrict the given column to a value provided.
- **object.print_distribution(field name)**: prints out a plotly graph histogram of the metadata field.
- **object.print_barplot(xcol, ycol, colorcol="")**: prints out a barplot of the given field names, optionally a third field which distinguishes the dots by color.

Just as the easy availability of tools and data has enabled students across the years to grow and bring their expertise to many domains, perhaps this ability to directly import datasets may enable others to explore datasets they might not have considered and through that exploration they might expand what is possible.

Testing

The process of this project has included:

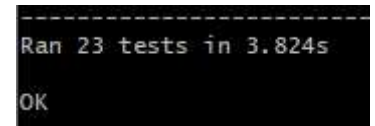
- Extensive webscraping: 2000 pages < $\Sigma(1 \text{ parent page} + 3x \text{ per dataset (502)} + 2x \text{ location (295)})$
- Thorough content research on the history of Machine Learning
- Careful analysis and extraction of scraped data

For the full extent of this work, we have approached testing in several major ways:

Visual Inspection: Due to the number of graphs involved, we succeeded by reviewing the output repeatedly and fixing issues until they were right. When data was scraped from pages, the files were carefully inspected to ensure they are complete and reflect what we are expecting.

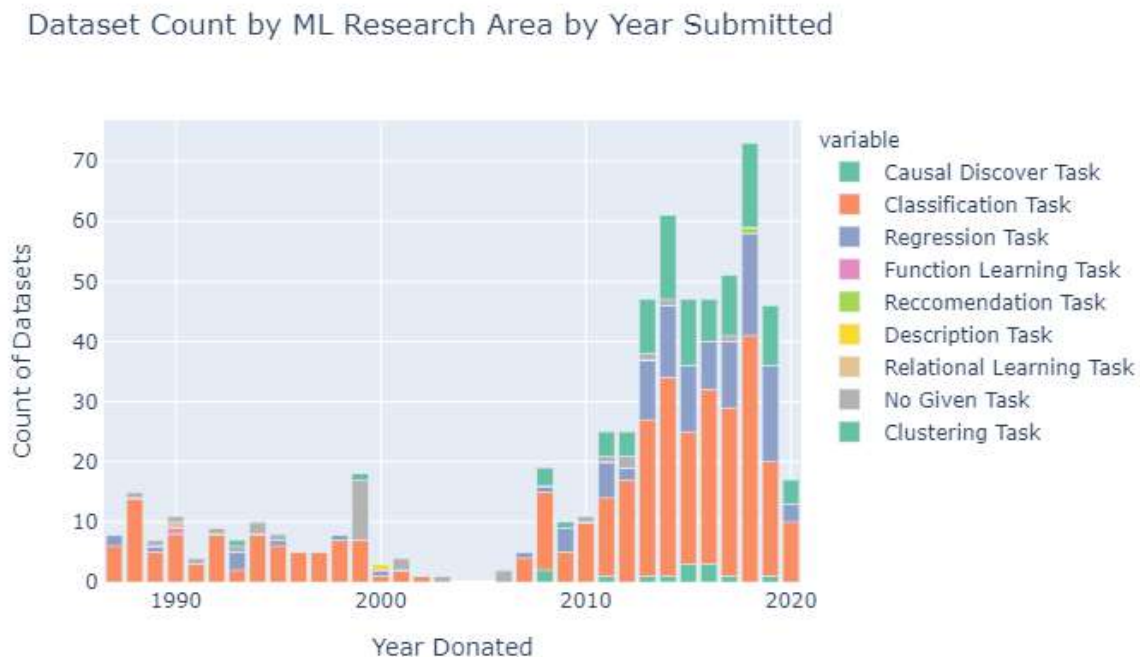
Unit Testing: Each data cleaning, class method, and visualization function has its own unit test. Each of these tests ensures that the contents of the dataset meets our expectations and returns the data that we are expecting in the right format. Furthermore this testing ensures that we are quickly notified of any failures.

One example of testing is testing the `UC_Irvine_datasets` class method `limit(field, value)`. To test this field, like in all tests, we imported `unittest` and the class being tested. We created an instance of the class, and then used the method `limit` on the field `"Area"`, restricting to where the value = `"Computer"`. First we tested that the `value_counts` of `"Area"` from the resulting dataset were 1. Then we tested `assertIsNone` by sending false values to the `"notreal"` column, and restricting to `"madeup"`. Both of these tests ran successfully.



```
-----  
Ran 23 tests in 3.824s  
OK
```


Conclusions



Based on analysis of the MLR, it appears likely that the repository will continue to grow as ML and AI continue proliferation in the near term. This data exploration will remain useful to data scientists and researchers. In the best case, the MLR will prove to be both reflective and predictive of the trajectory of ML and AI. That is, the MLR will be a cutting edge predictor and recipient of developments in this domain. Certainly, the MLR will remain a valuable source of sample datasets to empower education and development within the academic world. Based on our findings, however, causal discovery and clustering are two ML activities that are gaining more popularity within the MLR and the scientific community at large. This information will likely inform at least near-term developments. Although it is impossible to predict how industry and research will develop in the far future, some experts purport that in the next 25-50 years, we may see AI perform surgery, write a New York Times Best-Seller, and conduct high-level research (MIT Technology Review). It is clear that Machine Learning and Artificial Intelligence will play a crucial role for humankind for the foreseeable future, and it is likely that the UCI MLR, possibly with our tool's help, will play an important role in information sharing, education, and research in this field.

Sources:

- 2011 Harvard calls data science the sexist job of the 21st century
<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- Cosgriff CV, Stone DJ, Weissman G, et al The clinical artificial intelligence department: a prerequisite for success BMJ Health & Care Informatics 2020;27:e100183. doi: 10.1136/bmjhci-2020-100183
- Roboticsbiz. Machine Learning: The complete history in a timeline. (December 7, 2019).
<https://roboticsbiz.com/machine-learning-the-complete-history-in-a-timeline/>. Accessed July 23, 2020.
- Leapfrog Technology. Evolution of AI. <https://www.lftechnology.com/blog/ai/ai-evolution/>. Accessed July 22, 2020.
- MIT Technology Review. Experts Predict When Artificial Intelligence Will Exceed Human Performance.
<https://www.technologyreview.com/2017/05/31/151461/experts-predict-when-artificial-intelligence-will-exceed-human-performance/>. Accessed July 22, 2020.
- Time line of data science / big data <https://www.verdict.co.uk/big-data-timeline/>
- Background on UC irvine