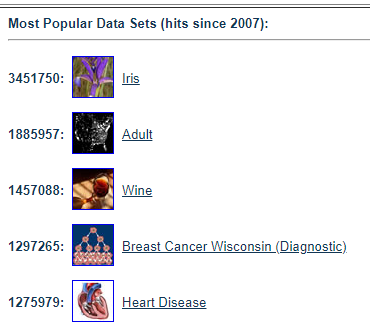
**Scope and Purpose:** This project examines the University of California Irvine (UCI) Machine Learning Repository (MLR) to identify and explore and analyze information about the data sets within. Throughout the project, we will transition from information to analysis of machine learning data, based 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.

**Source:** 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.

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. Figure 1 graphicly displays the number of data sets for all categorized areas of interest. 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. Figure 2 displays these results. The 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; Breast Cancer Wisconsin, Life, 1995; and 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, this has not been realized as well as data scientists would like. In fact, 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.

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. The second largest is “Physical” with 591881 average rows of data; “Life” ranks fourth behind “Other” with 310184.7 average rows. Reference Figure 3 for the average size of data sets by area of interest. 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. For many “Physical” or “Life” data may rely 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.



Iris – Classification – Life (1988)

Adult – Classification – Social (1996)

Wine – Classification – Physical (1991)

Breast Cancer Wisconsin – Classification – Life (1995)

Heart Disease – Classification – Life (1988)

Iris Data Set: This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Creator:  
  
R.A. Fisher -1936  
  
Donor:  
  
Michael Marshall   
  
Predicted attribute: class of iris plant.

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