**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. 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. 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. 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. Two data sets of note within “Computer” are extremely large: the Heterogeneity Activity Recognition and the WESAD (Wearable Stress and Affect Detection) data sets. Each of these data sets is larger than most entire categories, and focus on recording and analyzing human physical activity and response through the use of worn sensors.

**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. This growth in data size correlates to major technological developments within industry and 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). This growth opened vast opportunities for information distribution across systems and platforms, and likely contributed to the increased sizes of data sets being donated to the UCI repository. Cloud advancements have also made knowledge on Machine Learning (ML) and Artificial Intelligence (AI) more accessible and affordable. The types of studies have also evolved with technology as well. 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. These two studies also fall into clustering and regression activities, both of which saw contributions significantly increase from 2015 on.

**Future Implications:** 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. Although it is impossible to predict how industry and research will develop in the 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 will play an important role in information sharing, education, and research in this field.