The lecture notes used to make the slides can be found here: https://docs.google.com/presentation/d/1nbhPh8BjHHT_74rWYQKbyathfuqt5tF_z0z7KILpI8Q/edit?usp=sharing

Machine learning has been a revolution in modern prediction and clustering. Machine learning has become an expansive field involving computer science, statistics and engineering. Some of the algorithms have their roots in artificial intelligence (like neural networks and deep learning).

For data scientists, we decompose two main activities of machine learning. (Of course, this list is non-exhaustive.) These are are

1. Unsupervised learning - trying to uncover unobserved factors in the data. It is called "unsupervised" as there is no gold standard outcome to judge against. Some example algorithms including hierarchical clustering, principal components analysis, factor analysis and k-means.
2. Supervised learning - using a collection of predictors, and some observed outcomes, to build an algorithm to predict the outcome when it is not observed. Some examples include: neural networks, random forests, boosting and support vector machines.

We give a famous early example of unsupervised clustering in the computation of the g-factor. This was postulated to be a measure of intrinsic intelligence. Early factor analytic models were used to cluster scores on psychometric questions to create the g-factor. Notice the lack of a gold standard outcome. There was no true measure of intrinsic intelligence to train an algorithm to predict it.

For supervised learning, we give an early example, the development of regression. In this, Francis Galton wanted to predict children's heights from their parents. He developed linear regression in the process. Notice that having several children with known adult heights along with their parents allows one to build the model, then apply it to parents who are expecting.

It is worth contrasting modern machine learning and prediction with more traditional statistics. Traditional statistics has a great deal of overlap with machine learning, including models that produce very good predictions and methods for clustering. However, there is much more of an emphasis in traditional statistics on modeling and inference, the problem of extending results to a population. Modern machine learning was somewhat of a revolution in statistics not only because of the performance of the algorithms for supervised and unsupervised problems, but also from a paradigm shift away from a focus on models and inference. Below we characterize some of these differences.

For this discussion, I would summarize (focusing on supervised learning) some characteristics of ML as:

- the emphasis on predictions;
- evaluating results via prediction performance;
- having concern for overfitting but not model complexity per se;
- emphasis on performance;
- obtaining generalizability through performance on novel datasets;
- usually no superpopulation model specified;
- concern over performance and robustness.

In contrast, I would characterize the typical characteristics of traditional statistics as:

- emphasizing superpopulation inference;

- focusing on a-priori hypotheses;
- preferring simpler models over complex ones (parsimony), even if the more complex models perform slightly better;
- emphasizing parameter interpretability;
- having statistical modeling or sampling assumptions that connect data to a population of interest;
- having concern over assumptions and robustness.

In recent years, the distinction between both fields have substantially faded. ML researchers have worked tirelessly to improve interpretations while statistical researchers have improved the prediction performance of their algorithms.