newData.ai
data-driven success starts here

What is Artificial Intelligence?

by Taymour | March 12, 2019

May 2023 Update

Since the original article was written in 2017, there have been several significant updates in the field of machine learning and statistics. One important development is the increasing use of deep learning, a subfield of machine learning that uses artificial neural networks to model complex data relationships. Deep learning has enabled breakthroughs in fields such as computer vision, natural language processing, and speech recognition, and has opened up new applications in areas like autonomous driving and personalized medicine.

Another trend has been the growing emphasis on interpretability and fairness in machine learning models. As machine learning is used in high-stakes applications like healthcare and criminal justice, there is a growing need to understand how decisions are being made and to ensure that they are not biased against certain groups. To address this, researchers have developed new techniques for visualizing and explaining the inner workings of machine learning models, as well as methods for auditing them for fairness and bias.

In addition, there has been a growing interest in combining machine learning and statistics in what is sometimes called "statistical learning." This approach seeks to combine the strengths of both fields, using statistical models to make predictions and machine learning techniques to improve their accuracy and scalability. Some researchers have also explored ways to incorporate uncertainty into machine learning models, drawing on probabilistic modeling techniques from statistics to better handle situations where the data is noisy or incomplete.

Overall, the use of machine learning and statistics continues to grow and evolve, with new applications and techniques emerging all the time. As the amount of data being generated continues to increase, and as the need for accurate and interpretable predictions becomes more urgent, it

seems likely that both fields will continue to play important roles in the future of data analysis.

Both business and scientific communities have learned to successfully use machine learning and statistics to provide predictive analysis, yet machine learning has increasingly become the preferred analyzation method. Before looking at why, it's important to understand the difference between machine learning and statistics. To distinguish between the two, it helps to understand why businesses and scientific communities favor machine learning over statistics. The prevailing view is that their purposes are different: Statistics makes inferences, whereas machine learning makes predictions. This difference is evident in each word's Latin roots. In Latin, **prediction** derives from **praedicere**, which means "to make known beforehand" and **inference** stems from **inferentem**, or "to bring into; conclude, deduce." A statistical inference deals with how two or more variables are related. In other words, its purpose is descriptive, in that it quantitatively explains some type of a relationship. Machine learning, on the other hand, is primarily focused on prediction.

However, a quantitatively-defined description is often successfully used to make predictions. To make a head-to-head comparison between machine learning and statistics, it is essential to keep this common purpose in mind. We're seeking to highlight some of the distinctions regarding *how* predictions are made, employed, and interpreted. This article provides several examples of why machine learning is gaining favor in business and scientific applications.

*****Fun fact, the above update was written by Chat GPT*****

New Technology Put Statistics on the Map

The rise of statistical thinking is a result of numerous new technologies that appeared on the scene in the first decade of the 1900s. As desk calculators replaced the early tabulation machines at the beginning of the twentieth century, they were able to solve more complex calculations like *Ordinary Least Squares (OLS)* equations. Throughout the century, statistical thinking based on the mathematics of drawing projectable inferences from a smaller sample continued and expanded rapidly. In turn, improved technology made it possible to process increasingly large volumes of data faster.

Fast forward a century. Modern-day data storage and blazingly fast CPUs/GPUs can process massive amounts of data using statistical methodologies. However, while such horsepower can process samples that approach the population (n N), the fundamental small-to-large deductive principles that underlie statistics remain unchanged from earlier days. While statistics' predictive capacity has

improved with access to more data and processing power, its predictions do not incorporate data it has not previously encountered; it must rely on how well the sample fits a hypothetical, unknown population. The model's "fit" is manifested by its "parameter estimates," which are literally guesses of how the predictive dataset is expected to look. In other words, while the model estimates a hypothetical and unknown population's parameter, we assume the dataset used in the prediction literally refers to this theoretically unknown population. Machine learning, however, doesn't require any assumptions. Instead, it starts with a training dataset and then applies the patterns it learned to a predictive dataset. Unlike the statistical approach, machine learning refines its prediction by learning from the new data.

Whether this approach results in superior prediction depends largely on the scenario at hand. Understandably, either approach can go awry. In the case of statistics, the sample data may not represent the population to be predicted. Similarly, a machine learning training dataset may not resemble the predictive dataset. In these scenarios, the respective results are inadequate predictions. In the world of big data, however, machine learning generally maintains an advantage in overall predictive accuracy and precision, as it can process more information and deal with greater complexity.

What Are the Differences in How Predictions Are Made?

Statistics makes predictions (really inferences used for predictive purposes) about the large (a lot of data) from the small (sample data).. Machine learning makes predictions about the large from the large. It's important to note that both types of predictions can be delivered at the individual or population levels. Statistics draws inferences from a sample using probability theory, whereas machine learning uses mathematics as a "brute force" means to make its predictions. As one might expect, because machine learning processes more data iteratively, it tends to be far more computationally demanding than statistics. However, this limitation is increasingly becoming less of an obstacle with the recent explosion of processing power and increased storage capacity.

On the surface, both machine learning and statistics are numerically based, which begs the question: What is the difference between mathematics and statistics? While statistical methods may employ mathematics, their conclusions employ non-mathematical concepts. Because statistics is grounded in probability, uncertainty is rooted in its conclusions, whereas mathematics is precise and axiomatic. Statistics is empirically-based, deductive logic; mathematics uses formal, inductive logic.

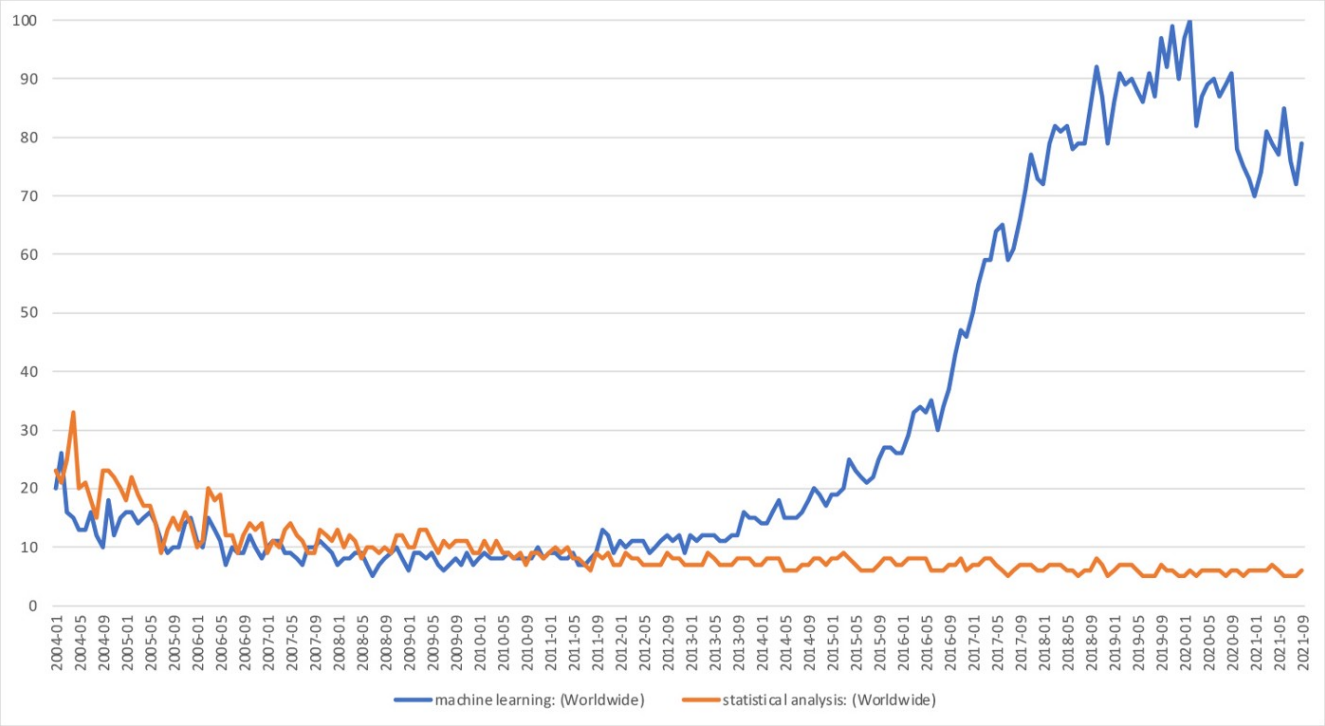
Compared to statistically-based prediction methods, machine learning doesn't make any assumptions about the data. Statistics require sample distribution assumptions to be satisfied, which is not always possible or easy to do. In addition, in statistical analysis, the sample data must be clean and pristine for its estimates to be accurate and precise. Machine learning, on the other hand, is less fussy. It can utilize structured, unstructured, or even messy data. While inaccurate or "noisy" data may slip into the machine learning process, using larger datasets has the potential to reveal patterns that may otherwise have been lost. A larger data pool generally improves the machine learning model's overall predictive power.

Interpretability

Statistics is typically more interpretable (it answers *what* and some *why* questions) than machine learning, which answers primarily *what* questions. For example, a statistics regression model can give insights into why certain variables are included, such as whether headaches are normally associated with the flu. Statistics tries to prove that headaches are a flu symptom by testing this hypothesis on other flu datasets. Machine learning, however, can plow through large amounts of data to uncover correlations between the flu and other features that happen to be correlated with it in the training dataset. In this example, machine learning may confirm headaches as a common symptom of the flu, but it may also uncover other correlations, such as the lack of sunlight exposure or something less obvious like the per capita mass transit usage. Of course, mass transit usage is not a flu symptom, but it could be a factor that helps explain the flu incidence in a certain region during the winter season. Or, it may find a factor that isn't open to explanation but nevertheless helps its prediction. Using marginal and conditional probability distributions, statistics can delve deeper into *why* questions, which is currently not possible with machine learning. However, raw predictive power may be more valuable than the ability to delve deeper into a subject when correlations also lead to actionable strategies.

Machine Learning Takes Center Stage

Both statistics and machine learning are used today, and both continue to evolve. However, Google searches for these terms show that machine learning began surpassing statistical analysis in popularity in early 2011 (see Figure 1).



Statistical Analysis vs. Machine learning January 2004 to May 2021

Faster and cheaper technology can harness data proliferation for both greater profits and social improvements. Though both statistical modeling and machine learning benefit from these advances, machine learning takes greater advantage, as it can utilize all the data it can access. The implication is that the prediction gap will widen as technology continues to improve.

Machine learning and other artificial intelligence techniques allow for proactive rather than reactive decision-making. In healthcare, this approach translates to significant savings through cost avoidance and reduction. Across Decode Health's use cases, we know that it also saves lives by leading to better decisions and earlier interventions.

A (non-exhaustive) list of differences in vernacular are presented below:

Statistical Predction	Machine Learning
Random Variable, Covariate, Predictor Variable, Independent Variable	Feature, Input, Side Information
Target, Response, Dependent, or Outcome Variable	Target, Label or Output
Estimation	Learning
Weight	Parameter

A (non-exhaustive) list of differences in methods are presented below:

Statistical Prediction	Machine Learning
Predictive Model Validation: Holdout Sample (cannot adapt to data it hasn't seen before)	Predictive Model Validation: Training Data (can adapt to data it hasn't seen before)
Structural Equations	Bayesian Networks
Sequential Experimental Design, Optimal Experimental Design	Hyper Parameter Optimization, Reinforcement Learning, Active Learning, Stream-Based Pool-Based Sampling, Membership Query Synthesis
Discrete Distribution: Maximize Likelihood	Discrete Distribution: Minimize Entropy

Demystifying the Move to Machine Learning

While it isn't exactly right to say that machine learning is superior to statistical analysis, several factors have made it more reliable in terms of making predictions. Perhaps the greatest shift drivers are the recent advances in processing power, which make it possible for machine learning to churn larger datasets iteratively, yielding better predictions than statistics. The fact that machine learning models don't make assumptions about the data means they tend to be more reliable as well.

One of machine learning's greatest strengths is that it can adapt to data it has not encountered, which means it can make predictions about something new. This outcome isn't possible with statistical analysis. Additionally, advances in machine learning are making the models more interpretable, a factor that was once one of the statically-based model's advantages. Since machine learning can handle and exploit an increasing amount of unstructured, "messy" data, it requires less work to prepare for analysis.

Ultimately, machine learning provides a much faster, more accurate way of working with large amounts of data, a feature that's more in demand with the rise of big data. Sometimes, statistical analysis remains the better option, but machine learning can make predictions with less data cleanup, so businesses can make decisions faster and scientific communities can start understanding the data and its patterns sooner.