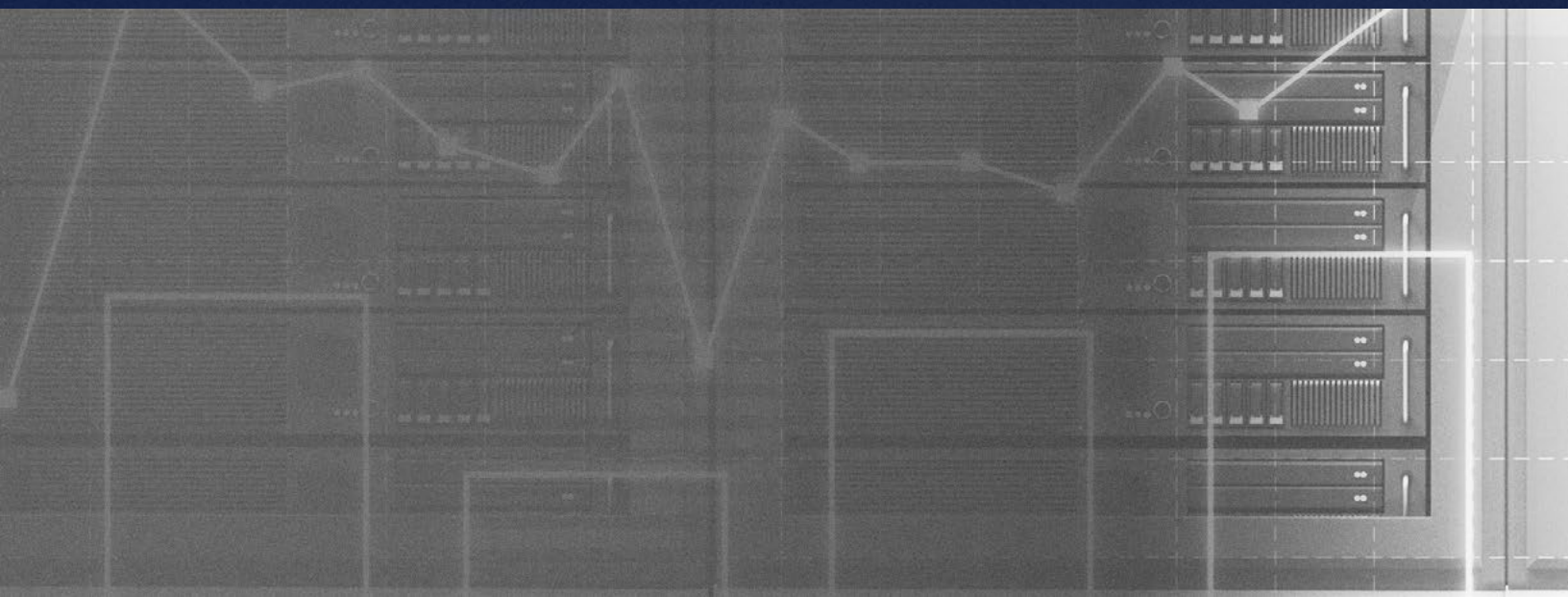
An Intro to Machine Learning for Big Data

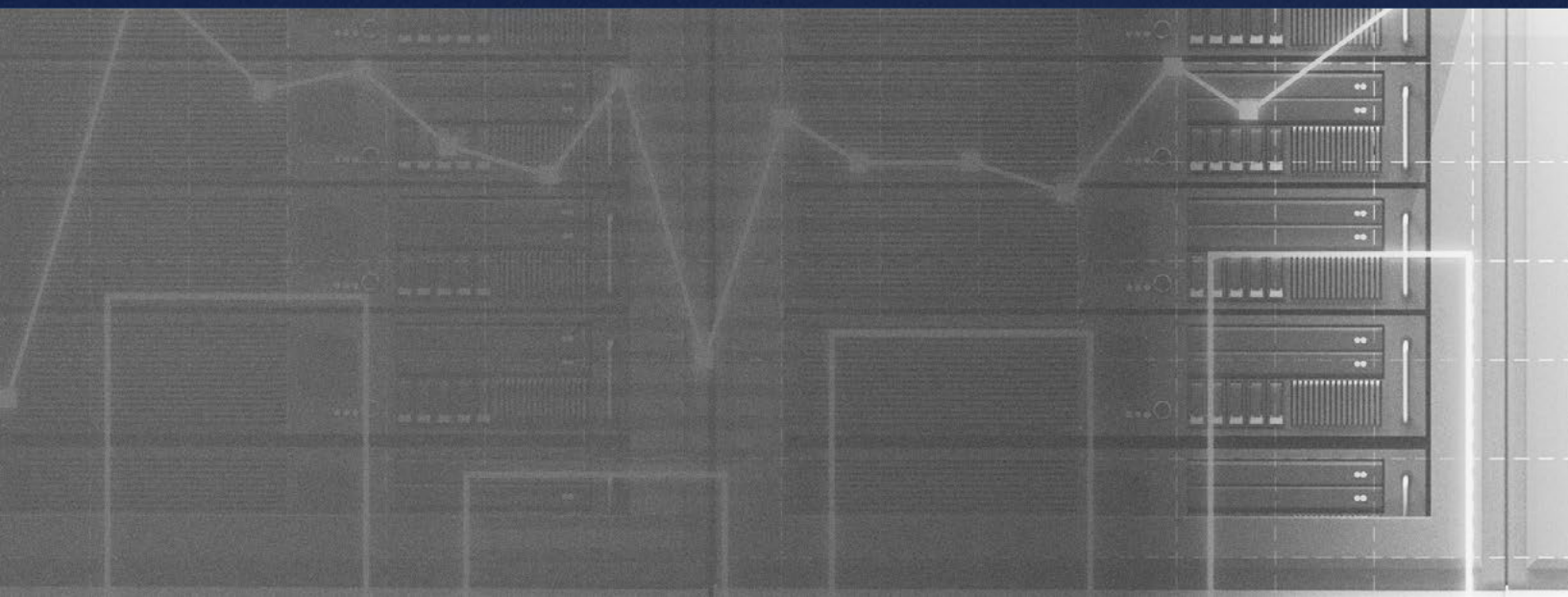
What is machine learning and how does it help solve Big Data challenges?

tamr
DATA UNITED



tamr
DATA UNITED

Table of Contents



Introduction

Machine learning, although not a new technology, is a topic that has taken news cycles by storm over the course of the past few years, spanning multiple industries and applications. And it's with good reason, as artificial intelligence (AI) and machine learning are changing the way we—as consumers, workers, and human beings—perform tasks and even interact with one another. It's an extremely useful technology that, when applied correctly, can be used to answer big questions and solve business critical challenges for large enterprises.

But what exactly is machine learning? And, more importantly, what are its applications for common challenges that enterprises encounter with Big Data? In this eBook, we're tackling some of the common questions about machine learning, namely:

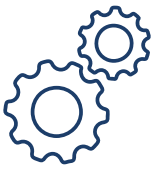
- What is machine learning?
- What are the types of machine learning?
- What role do humans play in machine learning?
- How does machine learning help solve Big Data challenges?

What is Machine Learning and How Does it Work?

The terms ‘artificial intelligence’ and ‘machine learning’ are sometimes used interchangeably, but it’s important to understand that machine learning is a subset of artificial intelligence. At its core, machine learning is about teaching machines to make data-driven decisions. How does a computer know how to distinguish an orange from an apple, for instance, or to mark certain emails as spam? It’s because the computer is leveraging machine learning to predict certain outcomes based on models that enable it to infer patterns, much the way a human brain would.

Machine learning algorithms use available sample data, or **training data**, to create a model. This model allows the computer to make predictions or decisions about new data that is introduced without someone needing to explicitly train the machine to do the required task.

The best machine learning problems are those where enough data exists for patterns to emerge. Data volumes don’t need to be massive—machine learning can be applied on hundreds of records for simple problems as well—but they do need to be large enough for patterns to exist.



At its core, machine learning is about teaching machines to make data-driven decisions.

An Example of Machine Learning

Consider the example of a data analyst who wants to predict the products a new customer will buy . To start, the analyst wants to understand whether age is a determining factor for purchasing particular product lines. To build a machine learning algorithm, the analyst collects a sample dataset from the customer base that includes customer ages as well as products purchased. This sample dataset will be used as the training data. That training data is then used to build a model that can predict future purchases. As more customer data is fed into the model, it continues to improve and become more accurate over time.

In this example, the data involves only two data fields or **features**: age and purchase history. But in most cases, there would be several additional features, such as income, location, etc. And the analyst could also choose to include publicly available, industry-wide data to expand his or her dataset. As a general rule, the more training data available, the more accurate the machine learning model is likely to be.

Types of Machine Learning

The situation described above is an example of **supervised learning**. There are three general types of machine learning: supervised learning, unsupervised learning, and reinforcement learning.

Supervised Learning

This technique uses a set of human-labeled training data to develop a model. The algorithm learns a set of inputs along with corresponding correct outputs. The training data used to create a machine learning model is assumed to be ground truth, meaning that its validity is not questioned—however, the model must still be tested for accuracy before it can be deployed. This is typically done by asking a subject matter expert to review the model predictions, usually through a set of sample data. If the reviewer corrects the model, then this information can be used as additional training data, and the model can be rerun.

This is known **as active learning**, **a type of semi-supervised learning**, because the machine learning model is improved with each additional correction or piece of information collected. Since one of the goals of machine learning is to reduce the level of human-dependency, it's important to be smart about which predictions to ask a human to correct. Asking a human to review the prediction that the model is most uncertain about will produce the maximum corrective effect.

Since one of the goals of machine learning is to reduce the level of human-dependency, it's important to be smart about which predictions to ask a human to correct.

Unsupervised Learning

In contrast to supervised learning, unsupervised learning infers patterns from unlabeled data to create a machine learning model. While this type of machine learning can be used to uncover previously unknown patterns in data, these are usually poor approximations compared to what can be achieved with supervised learning. The best time to use unsupervised machine learning is when you don't have data on desired outcomes.

Reinforcement Learning

This machine learning technique is based on the underlying idea of learning by doing. As with unsupervised learning, the machine is presented with unlabeled data, but is also given positive or negative feedback depending on the solution it proposes. Over time, the machine learns to choose the desired outcome based on this positive or negative reinforcement.

The Role of Humans in Machine Learning

Despite the fact that machine learning is used as a way to teach computers to make predictions without a human having to manually do the work, accurate and effective machine learning still requires significant human input and assistance. Humans are involved in various stages of machine learning, from labeling and providing training data, to fitting models, to inputting new data that improves the model's accuracy over time.

Human-guided machine learning is a process whereby subject matter experts accelerate the learning process by teaching the technology in real-time. For example, if the machine learning model comes across a piece of data it is uncertain about, a human can be asked to weigh in and give feedback. The model then learns from this input, and uses it to make a more accurate prediction the next time. This means that, inevitably, the amount of time a human needs to spend performing a specific task will decrease as the machine learning accuracy increases — improving efficiency and decreasing the amount of time a human needs to spend working on these kinds of tasks.

With human-guided machine learning, subject matter experts accelerate the learning process by teaching the technology in real-time.

Machine Learning and Big Data

Enterprises need ways to quickly and efficiently make decisions based on hundreds of thousands of datasets stored across different regions and business units. This is where machine learning can help—by providing the scalability needed to tackle the volume, velocity, and variety of Big Data.

Deterministic rules are useful in a number of scenarios for data preparation and analysis. But they're only one part of the solution, and simply aren't sufficient when it comes to large-scale projects. A human can easily write a certain number of rules to classify a small dataset and be confident that the results are accurate. Even a few thousand tables or records is theoretically manageable—although extremely slow and tedious—with a strict rules-based approach. But even then, by the time analysts are finished getting data ready to mine or model, the data is often already out of date.

In today's world, most enterprises are dealing with sets of transactions that number upwards of 20 million. It's therefore become nearly impossible for humans to write enough rules to handle all of the data, and unify or fix this data manually.

Beyond the volume and velocity of modern data, many enterprises also struggle with variety, or having too much data coming from too many places. And at many organizations, this challenge is solved by data analysts. Data analysts are pulling data from a variety of sources—databases, data lakes, data files, and relevant information available on the web—to answer a particular question. Once they collect all this data, they have to perform data integration on the resulting datasets. This means that a data analyst's time is largely spent integrating and cleaning dirty data before they can even begin analysis.

Machine learning helps enterprises manage the mapping, integration and transformation of many datasets into a common data model in a scalable way by:

- Greatly reducing the time to add new sources of data
- Enabling a small team to manage many data sources
- Improving the quality of the data by letting subject matter experts do more

Machine learning is an impressive technology that's capable of automating many tasks, but to perform at its best machine learning is still very dependent on humans: to gather and input data, fit models, and train the models to improve over time.

Conclusion

Machine learning is an important technology that is changing so much about our lives—from everyday tasks to how we work. And it's important to understand, at least at a basic level, what machine learning is and how it benefits us. In particular, it's important to understand the difference between hype and reality when it comes to machine learning. On its own, machine learning is an impressive technology that's capable of automating many tasks, but to perform at its best machine learning is still very dependent on humans: to gather and input data, fit models, and train the models to improve over time. That's why many solutions leverage human-guided machine learning, to expedite the learning process and improve accuracy of the machine learning models over time.

With the right solution, mastering large, diverse datasets through machine learning is significantly easier than creating and managing a network of custom rules and formulas. In fact, building machine learning models may not require any technical or data science knowledge at all—just general knowledge about your data.

When it comes to data analysis in particular, the sheer volume and variety of data enterprises are tasked with managing has now exceeded a level where humans can easily or manually unify data. This is where machine learning can help—by offering an effective solution to the growing challenge of data variety. With machine learning, enterprises can unify datasets as they come in. And when algorithms are constantly matching and connecting incoming data to other available datasets, all business units have broader access to the enterprise-wide data asset. This results in faster, more consistent, and scalable analytics.

About Tamr

Tamr is the enterprise-scale data unification company trusted by industry leaders like Toyota, Society Generale, GE, Thomson Reuters, and GSK. The company's patented software platform uses machine learning supplemented by human expertise to unify and prepare data across myriad silos to deliver previously unavailable business-changing insights. With a co-founding team led by Andy Palmer (founding CEO of Vertica) and Mike Stonebraker (Turing Award winner) and backed by investors including NEA and Google Ventures, Tamr is transforming how companies get value from their data.

To find out more or register for a demo visit tamr.com

© Copyright 2018 Tamr, Inc. All rights reserved. | 12.18