ComputerWeekly.com | TechTarget

E-guide

CW+ Content

The ultimate guide to data preparation

In this e-guide

In this e-guide:

Machine learning is all the rage in the tech industry today, which is not surprising given its promise to transform how organisations operate and people work.

But machine learning is only as good as the training data that is used to build a data model to enable an AI application to make its own decisions, rather than rely entirely on a human decision maker. Data must be in the right format and of a sufficiently high quality to support machine learning applications. Poor data equates to wrong decisions, bias in the AI algorithm and flawed decision-making. What's more, data that personally identifies an individual can only be used in very specific ways, as stipulated by GDPR.

Read about data preparation in this e-guide. This is the process of gathering, combining, structuring and organising data so it

In this e-guide

can be analysed as part of data visualisation, analytics and machine learning.

Karl Flinders, EMEA content editor

In this e-guide

Definition: Data preparation

Margaret Rouse, WhatIs.com

Data preparation is the process of gathering, combining, structuring and organizing data so it can be analyzed as part of data visualization, analytics and machine learning applications.

The components of data preparation include pre-processing, profiling, cleansing, validation and transformation; it often also involves pulling together data from different internal systems and external sources.

Data preparation work is done by information technology (IT) and business intelligence (BI) teams as they integrate data sets to load into a data warehouse, NoSQL database or Hadoop data lake repository. In addition, data analysts can use self-service data preparation tools to collect and prepare data for analysis when using data visualization tools such as Tableau.

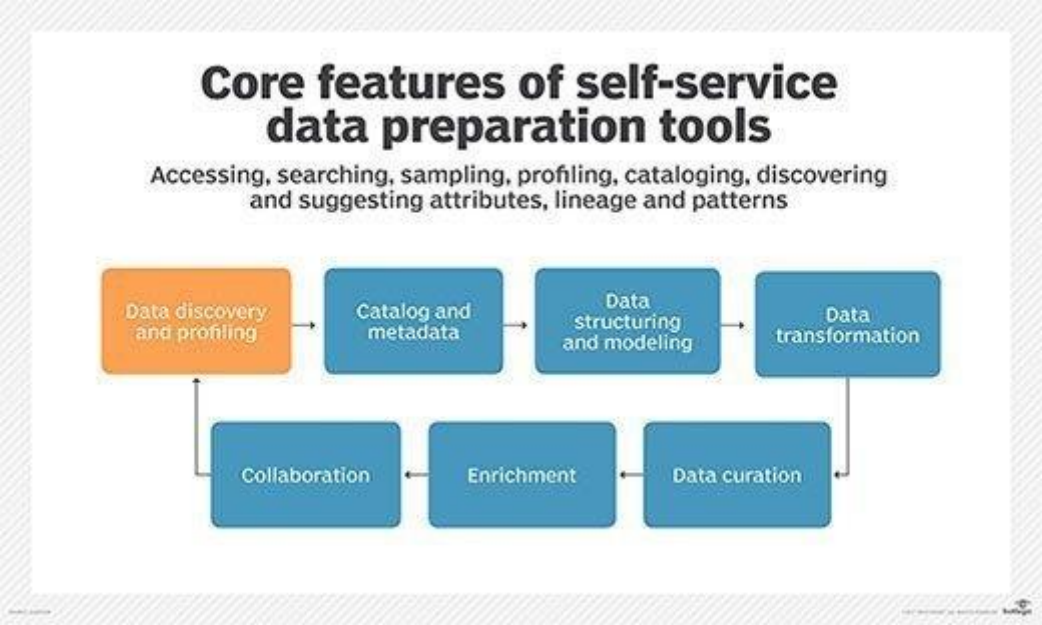
Purposes of data preparation

One of the primary purposes of data preparation is to ensure that information being readied for analysis is accurate and consistent, so the results of BI and

In this e-guide

analytics applications will be valid. Data is often created with missing values, inaccuracies or other errors. Additionally, data sets stored in separate files or databases often have different formats that need to be reconciled. The process of correcting inaccuracies, performing verification and joining data sets constitutes a big part of the data preparation process.

In big data applications, data preparation is largely an automated task, since it could take years of work by IT staffers or data analysts to manually correct every field in every file that's due to be used in an analysis. Machine learning algorithms can speed things up by examining data fields and automatically filling in blank values or renaming certain fields to ensure consistency when data files are being joined.



In this e-guide

Data preparation process

After data has been validated and reconciled, data preparation software runs files through a workflow, during which specific operations are applied to files. For example, this step may involve creating a new field in the data file that aggregates counts from preexisting fields, or applying a statistical formula -- such as a linear or logistic regression model -- to the data. After going through the workflow, data is output into a finalized file that can be loaded into a database or other data store, where it is available to be analyzed.

Even though data preparation methods have become highly automated, it can still take up significant amounts of time -- especially as the volume of data used in analyses continues to grow. Data scientists often complain that they spend a majority of their time locating and cleansing data rather than actually analyzing it.

Partly for that reason, there has been an increase in the number of software vendors attempting to tackle the data preparation problem, and many organizations are putting more resources toward automating data preparation. In 2017, data visualization vendor Tableau added self-service data preparation as part of its software, using machine learning methods to simplify the data preparation process.

In this e-guide

Benefits of data preparation

One of the biggest benefits of instituting a formal data preparation process is that users can spend less time finding and structuring their data.

Many enterprises have implemented data lakes, often built around Hadoop data stores, where they store large amounts of semistructured and unstructured data. When a data scientist needs a data set for an analysis, they have to hunt down the data first. With a formal data preparation process in place, repetitive analyses can be fed data automatically, rather than requiring users to locate and cleanse their data each time.

In this e-guide

The evolution of the data preparation process and market

Andy Hayler, CEO

Ever since people started to enter data into computers, there has been a desire to analyze that data. Yet, the reality has proved problematic. Computers are great at taking tables of numbers and performing calculations on them at great speed. The problem has always been in deciding which numbers to apply those calculations to, and how to approach the data preparation process.

When I started working in technology in the 1980s, I first came across the elegant notion of a data warehouse, a database designed to hold data specifically for analytics purposes. There were good reasons why it was tricky to apply analytical programming directly to the source transaction systems where data was entered. First, the databases of that time were optimized for transaction processing and were not built for mixed workloads. Accessing large swaths of data -- which was what you wanted for analysis -- would slow down the operation of the critical transaction systems.

In this e-guide

Challenges to the data preparation process

There were bigger problems, too. Companies had set up their transaction systems as separate data silos, often run by different departments. This meant the information needed for the context of a transaction -- such as which retail outlet it occurred at, which product was involved and which customer purchased it -- was typically stored locally with that system. As systems proliferated, the data about customers, products, locations and assets was duplicated and became inconsistent across the enterprise.

Inconsistent data wasn't a huge problem for each individual system on its own, but it mattered a great deal when you needed to take a view of data across the enterprise, such as calculating the total sales by product or customer. One huge data warehouse project I became involved with in the late 1990s was justified because the new CEO of a global consumer goods company had decided to rationalize the company's sprawling brand portfolio and focus on only the most profitable brands. Once the organization had gone through the data preparation process and analyzed figures, he was shocked to find no one knew which brands were actually profitable on a global basis in that decentralized company.

Furthermore, the quality of the data itself was often dubious. Human beings typing in customer names and addresses, product codes and prices inevitably made mistakes, despite the best efforts made by the designers of the data entry systems to validate at the source. I recall a project in which a subsidiary of a large organization discovered after a major data clean-up exercise that once all

the duplicate entries were removed, the organization had just one-fifth the number of corporate customers it was thought to have had. Combining corporate data with external data from third parties just added to the complications.

Data warehouses were built by IT departments to solve issues related to the data preparation process. These took feeds from multiple source systems, including ERP systems, and often applied data quality tools to clean up the data loaded into the warehouse. The idea was to construct a reliable and authoritative set of data that could be used to satisfy the analytics and reporting needs of the company. Such projects were typically large-scale and ambitious, and some that I worked on at various companies were in the hundred-million-dollar range. Some of the data warehouses worked well; many did not, stumbling across the difficulty of maintaining the integrity of the data as the organizations went through changes, mergers and acquisitions.

Traditional databases ran well when they were set up, but changing their schema structure once running was a major exercise. Adapting the data feeds into the data warehouse to accommodate a new corporate acquisition might take months, during which time the warehouse data was not fully up to date. If there were enough such changes, which could happen with acquisitive corporations, then the data warehouse quickly resembled a cartoon cat chasing its own tail as the staff was unable to keep the warehouse current enough.

At this point, business analysts lost faith and would start to download their own copies of data to spreadsheets for reports. Once that happened, it was a vicious circle as the data warehouse became less and less trusted and the analysts spent more and more time maintaining their shadow copies of data outside the reach of the IT department. Needless to say, such shadow copies did not stay up to date, nor were data quality tools applied to check the accuracy of this data. Such analysts regarded themselves as data freedom fighters, but were regarded by the central IT department as data terrorists.

The current state of data preparation

This is pretty much the stage that many organizations have reached today, something that can be seen by the emergence of tools devoted to the data preparation process. A range of tools, some by new vendors and some by existing ones, has emerged that allow business analysts to prepare their own data for analytical purposes from different internal and external sources. This market had reached $1.78 billion by 2017 and is growing rapidly. Every such tool sold is an indictment of the state of data in corporate data warehouses, because if the latter were working properly then there would be no need for such tools. The rise of big data from nontraditional sources like Hadoop files, in volumes that traditional databases struggle to store and process, has been a further nail in the data warehouse coffin.

In large companies, though, systems are like old soldiers: They never die -- they just fade away. The corporate data warehouses are still there and are being

In this e-guide

supplemented by Hadoop data lakes and other big data systems, with teams of puzzled analysts struggling to make sense of it all. There are ever more clever reporting and analytics tools to produce impressive charts, but who is to say whether the data that underlies those attractive graphics is truly correct?

For most companies, the original dream of a single, authoritative copy of data for an enterprise has broken down under the sheer weight of expectations that have been placed on it. The rise of the data preparation tools market is proof of that: The technology offers a treatment for a symptom of a fundamental underlying data disease in large organizations.

In this e-guide

Users find data preparation tools vital to BI strategies

Brian Holak, Site Editor

As data preparation tools become increasingly self-service-oriented and cloud-based, more enterprise users are employing them -- and recognizing their importance in their organizations' business intelligence strategies.

That's according to Dresner Advisory Services' 2019 Data Preparation Market Study. In the survey, 63% of all 350 respondents said data preparation is either "critical" or "very important." About 87% said data preparation is at least "important."

"Data prep has long been identified as an essential process of analysis," said Jim Ericson, vice president and research director at Dresner Advisory Services and co-author of the study, released Feb. 28. "Although it's not really glamorous, it has gotten a lot of prioritization. It's a really high priority among users."

It's such a high priority that respondents -- mainly professionals in an IT, BI or executive management function -- ranked data preparation tools as more

important than more familiar categories, like cloud computing, big data and the internet of things.

Still, user enthusiasm about data preparation tools and processes declined slightly year over year, according to the study. Although Ericson doesn't think that means data preparation is becoming less important, he said it could signal a watershed moment for data preparation.

"Other priorities have come to the fore, as this has been digested and has become more mature," Ericson said. "So, in that sense, you could say that data preparation might be becoming commoditized, but it's still evolving. As new data architectures and data types come onto the market, everybody is going to have to keep their eye on the ball for evolving needs."

Evolution of data preparation

Progression in the data preparation market has already begun, according to Ericson. Respondents to the Dresner survey preferred on-premises deployment of data preparation capabilities to private or public cloud deployment, but that's changing, Ericson said.

"This year is the first year that we're seeing more cloud-based services than on-prem services," he said. "Most of the users that we hear from still have their data preparation tools on-prem, but they're migrating toward cloud."

The options that are coming to the data preparation market more and more are cloud-based, so both the industry as a whole and the users Dresner surveyed are projecting that they're going to be using data preparation tools off premises, Ericson said.

A growing number of vendors offer both cloud-based and on-premises data preparation tools and capabilities.

Notable among these vendors is Tableau, which recently released its self-service, desktop-based Prep Conductor tool. With its release of Prep Conductor, Tableau moved further into the data preparation market, competing with the likes of Trifacta, Tibco Software, Power BI and Unifi Software -- all of which offer both cloud-based and on-premises self-service capabilities -- as well as Datameer, which provides cloud-based data prep software.

A wider variety of end users -- particularly savvy business users -- are also using data preparation tools, according to Forrester analyst Cinny Little. "More and more, business user-friendly data preparation at the enterprise level is a priority that firms are reaching for," she said.

Ericson agreed, saying that data preparation tools -- as they're being sold and marketed now -- provide a lot more distributed self-service capabilities for less technically inclined users than before. The survey respondents who said they used data preparation tools most frequently were research and development, finance, marketing and sales professionals.

"You find a lot of intense use and innovation in places like sales and marketing," Ericson said. "They use data preparation because they want to be able to ask and answer questions without having to go to IT."

That doesn't mean IT and data professionals are any less needed in data preparation, Ericson said.

"There's always going to be a requirement for IT to do complex data manipulations and queries that are not in the realm of tools," he said. "There may never be a time when [data preparation] is completely distinct from IT, so IT professionals need to keep their eye on the ball as some of these emerging architectures come to market."

In this e-guide

Data preparation for machine learning still requires humans

Kathleen Walch, Principal analyst

Data is at the core of AI and machine learning projects. Even more so than application code, data is crucial in training, testing, validating and supporting the machine learning algorithms at the heart of AI systems. Part of the reason why AI has surged again in popularity is due to the combination of almost limitless cloud computing, the availability of big data to train machine learning models, and the evolution of deep learning algorithms. The last two of those three reasons are data dependent. In fact, the more data you can feed AI algorithms, the better they perform and the more significant the machine learning results.

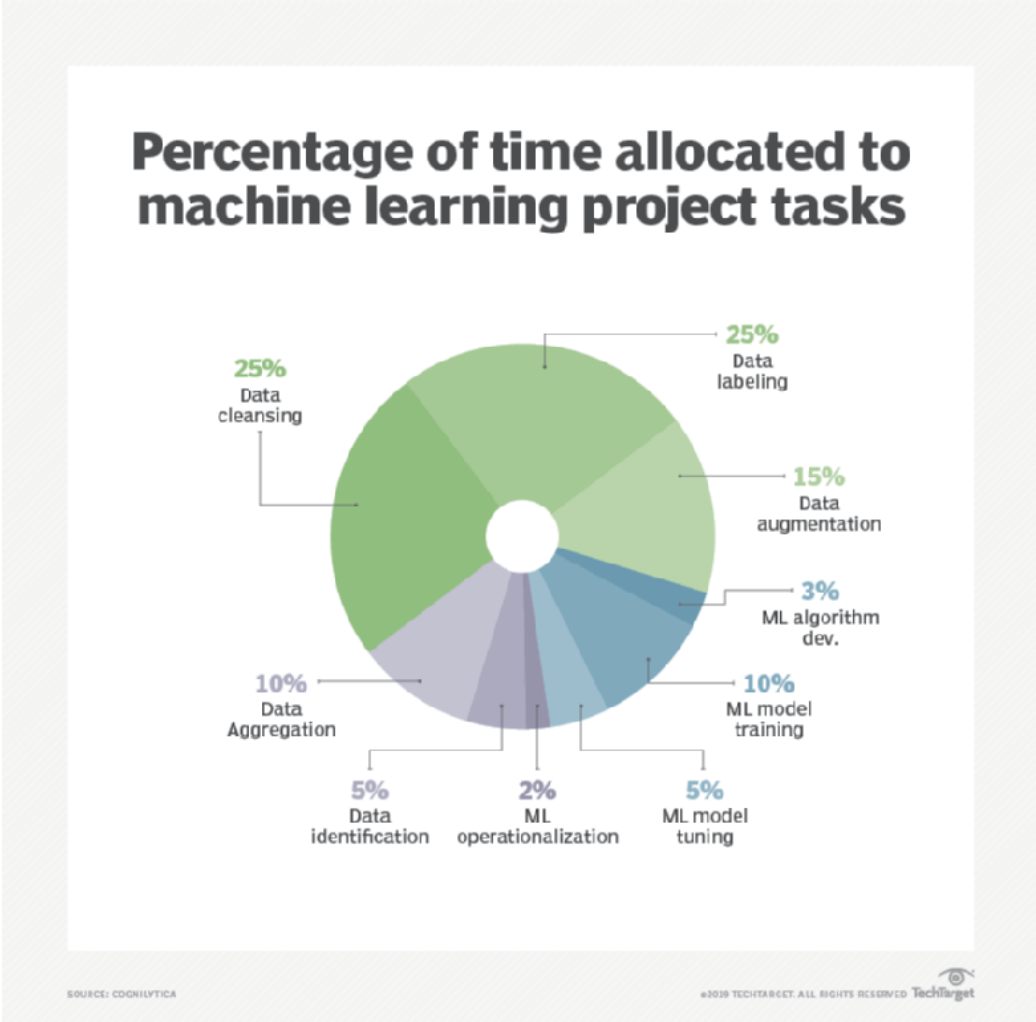
However, it's not enough to have a lot of data. Without good quality data, AI systems fail. The root of many machine learning project failures has little to do with the machine learning algorithms or code, or even any particular technology vendor choice. Problems almost always come back to the quality of data. In order for machine learning models to be properly trained and provide the expected accurate results, the data used needs to be clean, accurate, complete and well-labeled. Data preparation for machine learning is a crucial step.

Because of this, the majority of time that companies spend on AI projects goes toward data collection, cleaning, preparation and labeling. Enterprises are finding that they need to invest more in these data prep steps than on the data science, model training and operationalization parts. This has led to a substantial growth in demand for tools and services to assist with data preparation and labeling.

The many steps of data preparation for AI

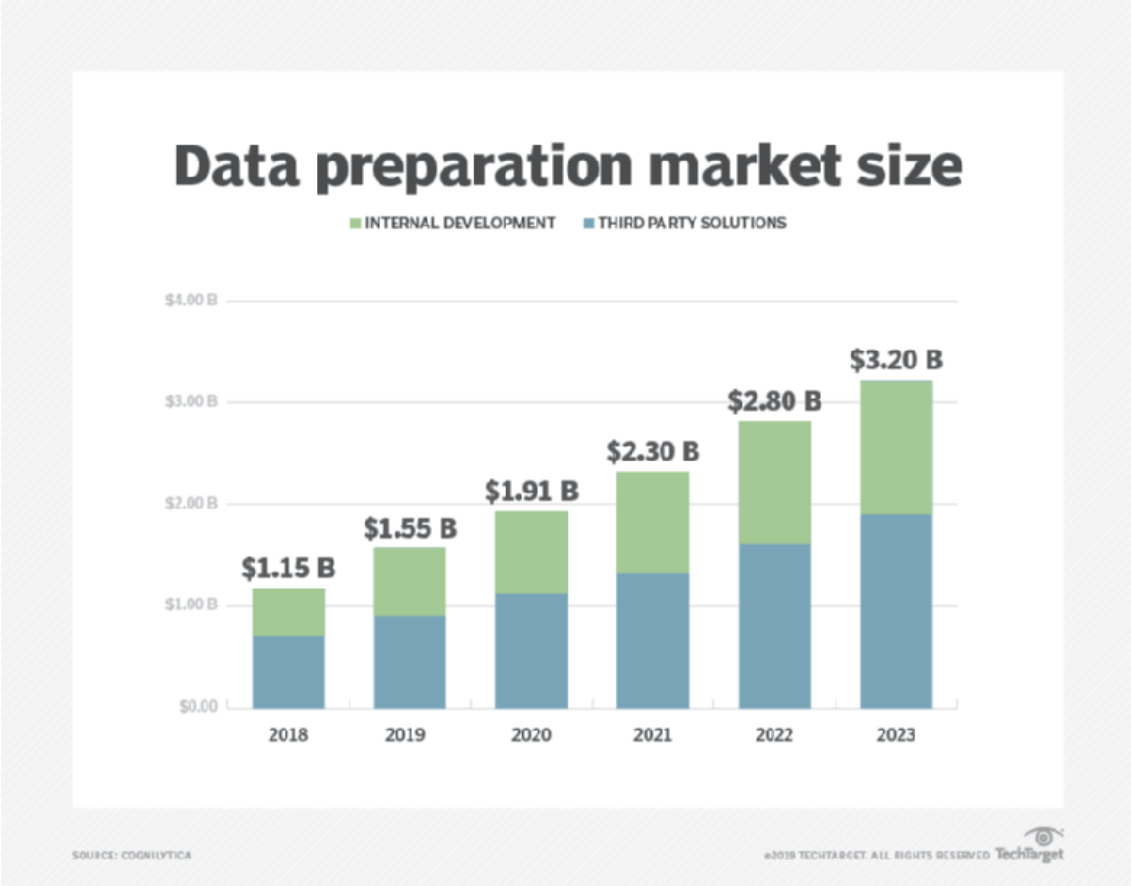
A recent report from AI research and advisory firm Cognilytica finds that over 80% of the time enterprises spend on AI projects goes toward preparing, cleaning and labeling data. Specifically, the report finds that the many steps involved in data collection, aggregation, filtering, cleaning, deduping, enhancing, selecting and labeling data far outnumber the steps on the data science, model building, and deployment sides.

In this e-guide



In this e-guide

A new class of data preparation tools has emerged into the market, built to manage big data sets and optimized to address the problems of machine learning projects. According to the report, the market for AI-focused data preparation tools is currently valued at over $500 million and expected to more than double to $1.2 billion by the end of 2023.



In this e-guide

Most enterprise data is not ready to be used by machine learning applications and requires significant effort in preparation. Tools offering data preparation for machine learning need to be able to perform a long list of tasks, including standardize formats across different data sources, remove or replace invalid and duplicate data, confirm data is accurate and up to date, help enhance and augment data as needed, reduce data noise, anonymize data, normalize data, allow for proper data sampling especially when working with large volumes of data, and allow for feature engineering and extraction.

Cognilytica's report finds that the AI-relevant data preparation tools provide iterative and interactive ways to allow people to quickly view the impact of data prep activities on big data. Some of the key features of these tools allow you to quickly spot anomalies in the data, identify and remove duplicates, resolve data conflicts, normalize data formats, establish pipelines for extracting and collating data from multiple sources, enhance data with additional features required for models, and anonymize data as might be necessary for certain applications.

In the past, enterprises relied on a category of tools known as extract, transform, load (ETL) to move data into and out of big data warehouses to facilitate reporting, analytics, business intelligence and other operations. However, in the new cloud-based, big data-oriented environment, moving data into and out of warehouses with ETL is falling out of favor. In its place, companies are looking to work with data in whatever location it currently sits. Some refer to this as "sipping from the data lake." Instead of ETL, companies

are looking at tools that can draw information on demand from the data source and transform them once extracted and loaded. This is more like ELT than ETL, and many of the data preparation tools on the market, including offerings from Melissa Data, Trifacta, and Paxata, work from the perspective of assuming data is located in different formats throughout the organization.

Data labeling and AI's secret

For supervised machine learning to work, the algorithms need to be trained on data that has been labeled with whatever information the model needs. For example, image recognition models need to be trained on accurate, well-labeled data that represents what the system will recognize. If you're trying to identify cats then you need many images of cats for a cat-recognition model.

It might come as a surprise, especially to those who don't deal with machine learning models on a daily basis, just how human-intensive much of this data labeling work is. Supervised machine learning projects form the bulk of AI projects. AI projects relating to object and image recognition, autonomous vehicles, audio analysis and text and image annotation are the most common workloads for data labeling efforts. Indeed, human-powered data labeling is a necessary component for any machine learning model that needs to be trained on data that hasn't already been labeled. One of AI's little secrets is that humans are still needed for manually labeling data and performing AI quality control.

Many companies resort to using internal labor or contracting general labor pools for this labeling work. According to Cognilytica's report, companies spent over $750 million in 2018 on internal labeling efforts and this number is projected to grow to over $2 billion by the end of 2023.

In the past few years, a new class of vendor has emerged to provide third-party labeling. Vendors such as Figure Eight, iMerit, and CloudFactory provide dedicated data labeling labor pools that are able to offload much of this work to remote workers who operate at better scales and cost of operation. The report cites that the market for third-party data labeling services was $150 million in 2018, growing to over $1 billion by 2023.

Yet, despite the use of third-party data labeling services, companies using those third-party offerings must still spend twice as much supporting those efforts than the cost of the actual data labor work. Part of the reason why it's so expensive to handle this portion of the machine learning project is there is just no way to entirely take the human out of the loop. This is where AI is running into the chicken and egg problem. In order to train machine learning algorithms you need lots of clean, accurate, well-labeled data, but to get that data you need humans to do the hard work to clean and manually label that data. Obviously if machines could do it, you wouldn't need the humans. But to get machines to be able to do it, you need the humans.

How AI can play a bigger role in data prep

Fortunately, as AI models become more intelligent and better trained, they can actually help in some of these activities related to data preparation for machine learning. In fact, the report highlights the fact that most of the tools on the market are adding AI to their systems to assist with data preparation activities, handle repetitive tasks autonomously and provide assistance to guide humans on prep activities. Increasingly, data prep and data labeling providers are applying machine learning to their own labeling efforts to provide some autonomous quality control and, to some extent, autonomous labeling.

Some of these companies use AI to help detect anomalies, patterns, matches and other aspects of data cleansing. Other companies use inferencing to identify data types and things that don't match the structure of a data column. This helps spot potential data quality or formatting issues and provides recommendations on how to clean the data. The report claims that all the leading data prep tools on the market will have AI at their core by 2021.

Similarly, Cognilytica's report sees data labeling efforts as increasingly being augmented by AI and machine learning capabilities. The use of pre-trained models, transfer learning, and AI-enhanced labeling tools will reduce the amount of human labor needed to build new models. That in turn will accelerate AI efforts and further increase efficiency on the more human-intensive side of AI.

In this e-guide

Since data is at the heart of AI and machine learning, the need for companies to have good, clean, well-labeled data will only increase. At some point in the near future, there will be pre-trained neural networks available for organizations to use. Until then, companies need to invest in software that performs data preparation for machine learning.

In this e-guide

Getting more CW+ exclusive content

As a CW+ member, you have access to TechTarget's entire portfolio of 140+ websites. CW+ access directs you to previously unavailable "platinum members-only resources" that are guaranteed to save you the time and effort of having to track such premium content down on your own, ultimately helping you to solve your toughest IT challenges more effectively—and faster—than ever before.

Take full advantage of your membership by visiting www.computerweekly.com/eproducts

Images; stock.adobe.com