

**Viewpoint**  
**David Meninger**  
 SVP &  
 Research Director

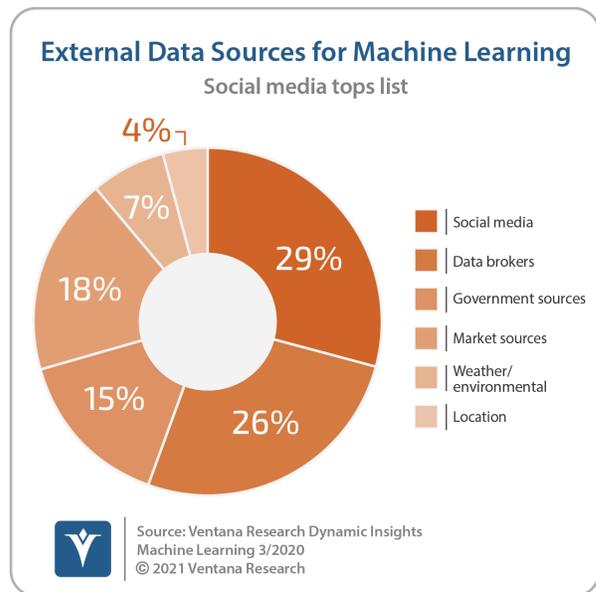
## Better Insights with External Data

External data can be valuable to many organizations for a variety of reasons. It can be used by planning and operations teams to benchmark an organization’s financial or operational performance. It can enrich the data an organization collects and analyzes about its customers, prospects and partners. It can augment machine learning analyses to produce more accurate models. For the CPG and retail sectors, demographic data is essential to understanding customers’ buying behaviors. In financial services, market data and economic indicators are essential to formulating financial strategies. For insurers, weather and geospatial data can be essential to minimizing claims from damaging storms.

Our research shows that more than three-quarters (77%) of participants consider external data to be an important part of their machine learning efforts. The most important external data source identified is social media, followed by demographic data from data brokers. Organizations also identified government data, market data, environmental data and location data as important external data sources. External data is not just part of machine learning analyses though. Our research shows that external data sources are also a routine part of data preparation processes with 80% of organizations incorporating one or more external data source. And a similar proportion of participants in our research (84%) include external data in their data lakes.

Organizations face several key challenges when working with external data. First, they must decide which data to purchase, how to maximize the value of their spend and how to manage the ongoing budget associated with external data. Second, they must help individuals within the organization understand and find the data that has been acquired. Third, organizations need a strategy for storing and managing external data. And finally, as with any data asset, organizations must maintain the external data to ensure it is up to date. Let’s examine these challenges in turn.

Some external data sources are free, such as data published by governments and their agencies. But data from commercial sources, such as demographic data, are often licensed by the volume of data requested so budget concerns typically dictate that you know which data is needed. The process of selecting data to purchase can slow down the business since there can be lag time between when the data is selected, the purchase is approved and the data is made available. In addition, departments can





end up with runaway budgets as more and more external data is requested. Alternatively, organizations may buy all the data they anticipate will be needed instead of requiring departments to specify their data selections. This approach can, of course, be expensive as organizations may buy more data than they need.

If external data sources are not understood or if individuals cannot find the information they need for their analyses, data will go unused and the analyses will be less accurate. Organizations can reduce this risk by deploying ways to catalog and search the information. With machine learning, organizations can go further. Introspection of the specific analysis being conducted, combined with information about the external data available can result in recommendations of which data to use and even recommendations on the type of analysis to perform. We assert that by 2024, more than 60% of all data processes will use artificial intelligence and machine learning to boost the value that can be derived from the data.

One of the questions organizations face is how to manage this information. As noted above, many choose to load this information into their data lakes or data warehouses. This may seem like the obvious approach, but it also presents some challenges. The process of loading information may be slow, especially if incremental data purchases are made. Loading the incremental purchases should be reviewed and approved, even if an automated routine has been created to handle the incremental data. External data can often contain sensitive information subject to privacy or other policy restrictions and the governance of that data should be considered as incremental purchases are made.

Once the decision has been reached to utilize external data, this decision also implies putting processes in place to maintain the data. Market data is constantly in motion, for example as customer demographics change or as transaction histories grow. Various economic indicators fluctuate with expanding and contracting cycles. Maintaining up-to-date information is critical to ensuring accurate analyses and supporting optimal operational decisions in organizations.

One way to address these issues is with an external data platform or “data as a service.” The external data platform provider manages the information, keeps it current and makes it available as necessary. In some cases, the platform can assist with analyses by automatically discovering features of the data and their potential impacts on machine learning models. The platform can also help match and validate multiple external data sources with each other and with internal data sources. Where data needs to be loaded into internal systems, the data vendor can provide various ways to export data or analyses such as machine learning models based on the data.

Organizations that are not using external data should strongly consider how this data could enhance their financial, operational and analytical processes. As organizations utilize external data, they should address the issues associated with purchasing, storing, accessing and maintaining this data. Each organization will need to address these issues with approaches appropriate for its situation, but you should ensure your strategy encourages rather than discourages the use of external data.



**Dave Menninger – SVP and Research Director, Ventana Research**

David Menninger is responsible for the overall direction of research on data and analytics technologies at Ventana Research. He covers major areas including artificial learning and machine learning, big data, business intelligence, collaboration, data science and information management along with the additional specific research categories including blockchain, data governance, data lakes, data preparation, embedded analytics, natural language processing (NLP) and IoT.