WHITE PAPER

Data Management for Artificial Intelligence

sas
THE POWER TO KNOW.

Contents

Introduction

Think back to the grade-school game where one person whispers something in the ear of the next person, and the phrase is then whispered from person to person around a circle.

The last person reveals what he or she heard, and it is always something wildly and hilariously different from the statement that started the cycle.

Working with bad data can be like that.

And if your process is some form of artificial intelligence (AI) – where the machine actually adapts the underlying algorithms based on what it learns from the data – bad data can really get you into trouble.

The results will be wildly skewed from the input that went into the cycle, but it's not hilarious at all.

Artificial intelligence is the science of training systems to emulate human tasks through learning and automation. With AI, machines can learn from experience, adjust to new inputs and accomplish tasks without manual intervention.

The explosion in market hype around the term is closely tied to advances in deep learning and cognitive science, but AI spans a variety of algorithms and methods. It doesn't require the flashiest new technologies to still be considered an AI application.

As a topic of interest for years – from science fiction plots to futurists' prophecies – the promise of AI has always been at the forefront of our minds. But what was once a distant vision is becoming reality as organizations embrace the value of AI now:

- By 2025, the artificial intelligence market will surpass $100 billion. (Source: Constellation Research)
- Seventy-two percent of business leaders believe AI will be fundamental in the future. (Source: PwC)
- In the immediate future, execs are looking for AI to alleviate repetitive, menial tasks such as paperwork (82 percent), scheduling (79 percent) and timesheets (78 percent). (Source: PwC)

Machine learning (a subset of artificial intelligence) automatically creates analytic models that adapt to what they find in the data. Over time, the algorithm "learns" how to deliver more accurate results, whether the goal is to make smarter credit decisions, retail offers, medical diagnoses or fraud detection.

Why AI is a hot topic

C-level executives are taking a close look at AI for a host of good reasons:

- **AI automates repetitive learning and discovery through data**. Unlike robotics, which automate manual tasks, AI automates high-volume computing tasks such as search and classification – reliably and tirelessly.
- **AI adds intelligence to existing products**. Think about how Siri added new value to Apple products. Or how AI can make online chat with a bot feel like talking with a human. AI enhances technologies in the home or workplace, from consumer marketing and security intelligence to investment analysis.
- **AI adapts through progressive learning**. In essence, the data does the programming. An algorithm can teach itself how to play chess or what product to recommend next online. Through back-propagation, AI models adapt to new data they are given or what they learned from experience.
- **AI analyzes more and deeper data** using neural networks that have many hidden layers. For example, building a fraud detection system with five hidden layers was almost impossible a few years ago. It's achievable now, thanks to incredible computer power and big data.
- **AI achieves accuracy that was previously impossible**. For example, our interactions with Alexa, Google Search and Google Photos (all based on deep learning) keep getting more accurate the more we use them. In the medical field, AI techniques can find cancer on MRIs as well as highly trained radiologists.
- **AI gets the most out of data**. When algorithms are self-learning, the data itself can become intellectual property and a competitive differentiator. The answers are in the data; you just have to apply AI to ferret them out.

For the first time, companies have access to the full set of building blocks to begin embedding machine intelligence in their business processes. Almost every industry is already seeing the effects, from agriculture to transportation, health care to financial services. Machine learning empowers people to be more productive with tools that exist today.

While you may think your company is way behind regarding AI, Gartner says only 4 percent of all companies are currently using it.[1] The vast majority of others are researching, taking a wait-and-see approach, and determining how to be successful when they are ready to act.

Read on to discover the biggest weakness in AI projects – the cause of failure for 40 percent of analytics projects in general – and how to get it right.

1 Gartner news release, "Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence," February 13, 2018, https://www.gartner.com/newsroom/id/3856163

Confusing and often interchangeably used AI terminology

Artificial intelligence is the umbrella concept of machines carrying out tasks in ways we would deem smart or even human-like.

Machine learning is an applied form of AI based on giving machines access to data and letting them learn for themselves.

Neural networks are computer systems that can classify information in much the way a human brain does, making decisions or predictions based on data fed to it.

Deep learning is a type of machine learning that uses many layers of processing to perform human-like tasks, such as recognizing speech, identifying images or making predictions.

Why data management should be an equally hot topic

Here's the reality: Machine learning systems don't just extract insights from the data they are fed, as traditional analytics do. They actually change the underlying algorithm based on what they see in the data. The more data they are fed, the more tightly they define the algorithm and the more confidently they make classifications or predictions.

So the "garbage in, garbage out" truism that applies to all analytic pursuits is truer than ever. If the data that feeds machine learning algorithms is not well managed, the results could be like the end result of the whisper game – wrong statements where errors have multiplied upon themselves. The dangers in that are obvious: inconsistency, inaccurate insights, loss of trust and AI results becoming stale.

On the other hand, since data carries more weight than ever before, data management can become a real competitive advantage. Even if everyone is applying similar techniques in a competitive industry, the one with the best data management program will win. So it's no surprise that 95 percent of C-level executives believe data is an integral part of forming their business strategy.[2] This has always been true, but machine learning magnifies the possibilities.

In short, machine learning is only as good as the data that goes in it – and it provides the best return when it is supported by a well-governed data management program.

The machine learning algorithm:

1. Ingests massive amounts of data.
2. Determines the best way to analyze the data given different parameters.
3. Learns automatically by discerning patterns in the data.
4. Produces an intelligent output.

2 *An Effective Data Management Program Starts With the C-Suite*, Experian Information Solutions, 2016 (http://images.go.experian.com/Web/ExperianInformationSolutionsInc/%7B285d1efe-e7f4-46ed-992c-8b90547dbf67%7D_c-suite-data-management-white-paper.pdf)

However, data management has been problematic for analytics teams long before machine learning started going mainstream. Traditionally, data management is a burden that consumes the vast majority of a data scientist's time. With so little time left over for analysis, it may be tempting to scale back data management, even if it yields poorer results.

Furthermore, there is a general assumption that AI's automation handles much of the legwork of data management. Wrong. Data management is very much needed at the outset. While AI can support data management processes – such as determining what data to keep or discard, or classifying data for optimal storage – it doesn't by default manage data for its own consumption.

AI calls for chief data officers to take a holistic approach to data integration, data quality and governance. Underlying all these areas is cloud, where we're likely to see the most growth in the analytics data management market in general.

> You need lots of data to train deep learning models because they learn directly from the data. The more data you feed them, the more accurate they become – if it's good data.

Five data management best practices for machine learning

These best practices provide the carefully managed data foundation AI initiatives require:

- Simplify access to traditional and emerging data.
- Drive smarter integration with statistical AI.
- Scrub data to build quality into existing processes.
- Shape data using flexible manipulation techniques.
- Share metadata across data management and analytics domains.

1. Simplify access to traditional and emerging data

The more accurate and managed data you present to machine learning programs, the more accurate they become, whether the data is from unstructured sources such as images and text, SAS data sets, streaming or from traditional data warehouses.

But accessing all that data is challenging. The diversity of data calls for native data access capabilities that make it easier to work with a variety of data from disparate sources, formats and structures.

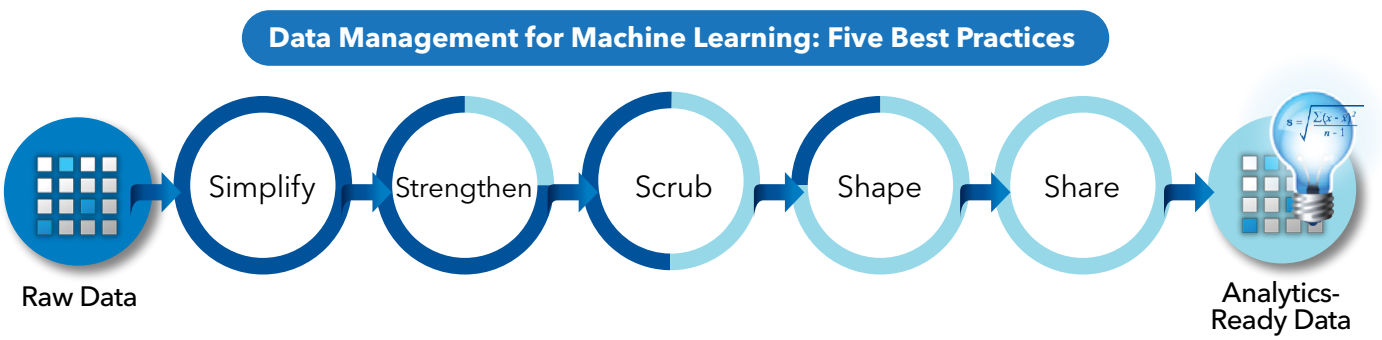

Figure 1: Five best-practice steps in preparing data for machine learning

To support the speed and agility expected of machine learning, consider the following capabilities:

- **Simplified access to multiple data sources.** From an Excel spreadsheet to a relational database table to Hadoop and cloud-based sources, automatic conversion removes the complexity of reconciling data types.
- **Minimal data movement**. Pushing data processing down to the data source improves governance and dramatically boosts performance.
- **Self-service data preparation with intuitive user interfaces**. By making data accessible to more users, with less training, you can free IT personnel from iterative data provisioning and preparation tasks.
- **Agile, secure techniques for managing data**. For example, data virtualization creates quick views of the data without moving it. Dynamic data masking protects sensitive data.

To get full return on investment in AI, start by putting a data management program in place to support it.

2. Drive smarter data integration with statistical AI

Think beyond traditional data integration approaches by using statistical AI capabilities. For example:

- **Frequency analysis** helps identify outliers and missing values that can skew other measures such as mean, average and median.
- **Summary statistics** help analysts understand the distribution and variance – because data isn't always normally distributed, as many statistical methods assume.
- **Correlation** shows which variables or combination of variables will be most useful based on predictive capability strength – in light of which variables may influence one another, and to what degree.

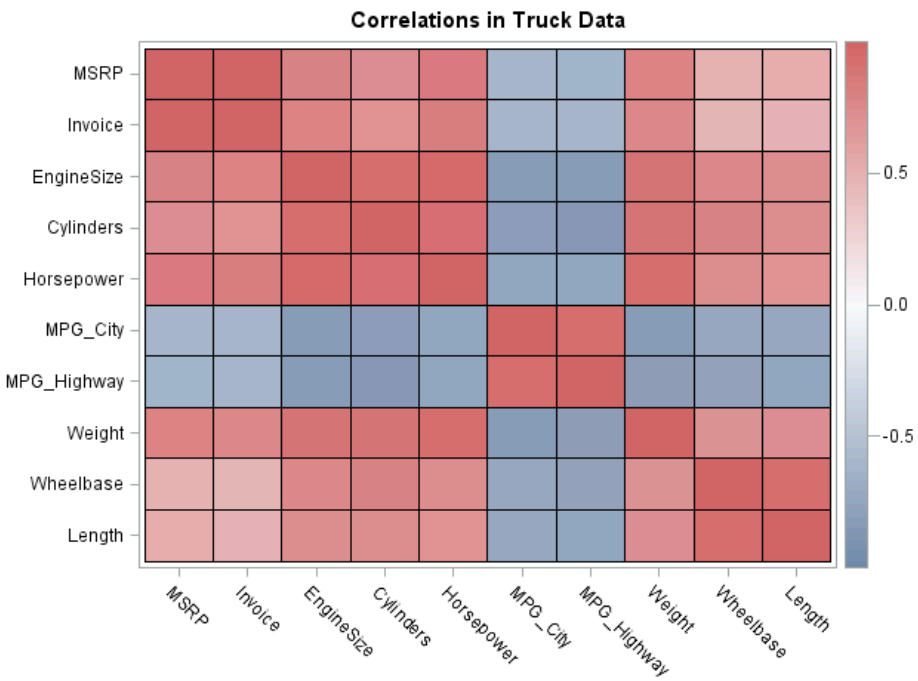

Figure 2: A correlation heat map shows measures of association – that is, the strength of the relationship – between variables.

3. Scrub data to build quality into existing processes

Data cleansing begins with understanding the data through profiling, correcting data values, adding missing data values, finding and dealing with duplicate data, and standardizing data formats (dates, monetary values, units of measure).

The strict data demands of machine learning require a data quality platform that:

- Incorporates the cleansing capability into the data integration flow.
- Pushes data quality processing down to the database to improve performance.
- Removes invalid data – such as outliers, missing or redundant data – based on the machine learning method being used.
- Enriches data via binning – that is, grouping data to create a smaller number of relevant data points from a high number of data points, such as binning individual ages into age groups ("between 35 and 45") or "cholesterol level" into a group of patients with cholesterol higher than 190.

4. Shape data using flexible manipulation techniques

Organizations have hundreds of data sources, such as enterprise and desktop applications, cloud services and third-party data – each with its own structure, codes and calculations. Preparing data for machine learning requires merging, transforming, de-normalizing and sometimes aggregating data from these disparate sources into one very wide table, often called an analytic base table. This is data shaping.

This data transposition can be an arduous data manipulation task. If done with programming, it can involve hundreds of lines of code.

SAS simplifies data shaping with intuitive, graphical interfaces for transformations. Plus it lets you use other reshaping transformations such as frequency analysis, appending data, partitioning and combining data, and multiple summarization techniques.

AI makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks. Most AI examples that you hear about today – from chess-playing computers to self-driving cars to smart personal assistants – rely heavily on deep learning and natural language processing.

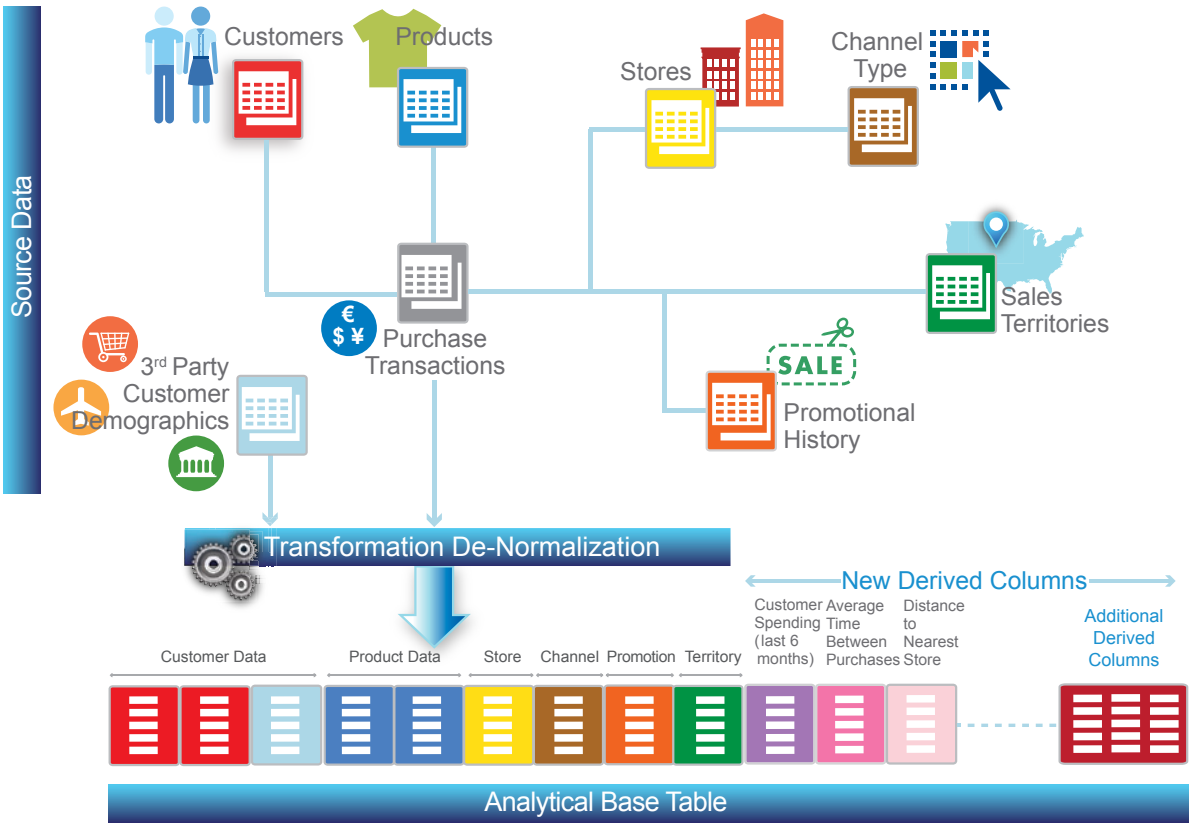

Figure 3. Data shaping transforms data from multiple sources into consistent, analysis-ready form.

5. Share metadata across data management and analytics domains

A common metadata layer lets you consistently repeat your data preparation processes. Common metadata also provides lineage information so you can answer such questions as:

- Where did the data come from?
- What was its quality?
- What data was used, and where else has it been used?
- How was the data transformed?
- What additional reports or information products are developed using this data?
- What did the machine consider important, and what did the algorithm focus on?

Applying metadata across the analytics life cycle delivers savings on multiple levels. When a common metadata layer serves as the foundation for the model development process, it eases the intensely iterative nature of data preparation, the burden of the model creation process and the challenge of deployment.

You'll notice more efficient collaboration, better productivity, more accurate models, faster cycle times, more flexibility and auditable, transparent data.

Use AI to help manage the organization's digital assets

Machine learning has a fitting role in managing the organization's digital resources. For example:

File classification. A machine learning algorithm can sort through vast troves of stored emails, documents, images and more; make recommendations for how to classify it all; and present those recommendations to a human for review and action.

Data retention. Machine learning can handle the time-consuming task of sifting through archives to find files that are potentially obsolete – say, haven't been accessed in a given period of time, are no longer relevant or aren't needed for compliance or e-discovery. Then all a human has to do is decide if there's a reason to keep it.

Managing updates. Machine learning can provide staff with real-time updates and notification of appending updates.

Storage optimization. Machine learning can determine which resources are rarely needed and which are used more often – and make smart decisions about assigning them either to lower-cost, slow storage or higher-cost, fast storage.

If you only remember four things

Remember these top takeaways:

1. AI holds much promise for its ability to autonomously analyze vast data volumes and uncover insights that hypothesis-driven techniques might miss.
2. The promise of AI can only become reality if the data that feeds it is curated by a well-governed platform managed by the chief data officer.
3. Companies that forgo data management or just leave it to the data scientist to figure out are taking a stopgap approach that will cost more money in the long run and yield inferior results.
4. Our recommendation: A holistic approach to data management that brings data integration, data quality and governance together in a platform that supports a structured, five-step process under the auspices of the chief data officer.

Data management best practices position your organization to get the full value from machine learning. You'll have access to all types of raw data that you can cleanse, transform and shape for any analytical purpose. As you glean continually deeper insights from your data, you can embed that knowledge into analytical models and machine learning algorithms, share your new discoveries and automate decision-making processes across the organization.

One last thought, a comfort to anyone who sees a future of robot rebellions and the takeover of the planet away from humans: For the foreseeable future, AI should be seen as an adjunct to human wisdom and experience.

AI is not about replacing human wisdom and values. It's about freeing us from the rote, high-volume tasks that computers can do better, such as scanning thousands of documents or millions of data points looking for patterns and predictive insights. Human intelligence is still indispensable, although at an increasingly higher level, reserved to tap the judgment that only the human brain can offer.

About SAS

SAS is the leader in analytics. Through innovative software and services, SAS empowers and inspires customers around the world to transform data into intelligence. SAS gives you THE POWER TO KNOW®.

SAS has been providing AI solutions for years and continues to push the boundaries in disciplines such as machine learning and deep learning. Today, we are helping our customers capitalize on the growth opportunities AI presents. Moving forward, we will continue to embed AI solutions across the SAS portfolio to help bring the transformational benefits of machine-assisted decision making into every arena.

Learn more

Find out more about how SAS, an analyst-recognized leader in data quality, data integration and advanced analytics, supports data management best practices to help you get the most from AI: sas.com/data

To contact your local SAS office, please visit: sas.com/offices