TECHNOLOGY BRIEF

Artificial Intelligence
Readying Your Data

The Seven-Step Plan to Preparing Your Data for AI

Sharpen your competitive advantage with this plan for embracing AI

In a 2017 report, McKinsey forecasted that by 2019 100 percent of Internet of Things (IoT) efforts will be supported by artificial intelligence (AI).[1] The same study predicts that some 40 percent of digital transformation initiatives will be enabled by AI by 2019.[1]

81%
The percentage of CEOs who identified AI as a key investment vector for the future.[2]

82%
The percentage of executives with plans to implement AI in the next three years.[3]

USD 47 BILLION
Expected enterprise spending on AI by 2020, up from USD 8 billion in 2016.[4]

It's easy to see why companies are making the move to AI. The potential benefits are profound. The upside includes enhanced efficiency, lower operating costs, and faster time to market. It makes possible new business insights, greater productivity, working capital optimization, and richer customer experiences. Together, these advances can contribute to a sharpened differentiation in the marketplace that drives competitiveness—and innovation-powered growth.

While companies increasingly understand both the promise of and the urgency behind embracing AI, knowing how to begin that process is often another matter. As AI is democratic in its impact, touching industries and businesses of all kinds, many are finding building a path to AI beyond their depth and experience. They need help to get started.

In this guide, we aim to help provide guidance and guardrails as you determine, first, if AI is right you, and second, if it is, how to go about best positioning your company to make the most of this groundbreaking opportunity. Why? Because besides succeeding with digital technologies, companies identified as AI leaders are more likely to be top performers in productivity (55 percent vs. 9 percent of laggards), profitability (51 percent vs. 21 percent), and the ability to adapt to evolving market conditions (55 percent vs. 16 percent).[3]

Data is the difference in AI

At its heart AI runs on data. So just as the fuel in your vehicle must be right for your engine to operate optimally, the same is true for the data with which you hope to ignite your AI effort. It must be right and ready.

Unfortunately, too many organizations are saddled with bad data, meaning that it is incomplete, siloed, compromised by privacy issues, mislabeled, corrupted, or otherwise not ready for prime time. Just consider that today's data scientists spend as much as 85 percent of their time cleaning, labeling, and classifying data—before they ever analyze or mine it.

Seven steps to preparing your data

Because AI starts with data, it is where you must start too. You want to establish the data strategy today that will become the foundation of your training models, machine learning (ML) applications, and business process development in the future.

Those companies that do the work to prepare for AI stand to bring new value and opportunities to their business. Those that fail will find it increasingly challenging to compete. Here are seven steps to help guide you as you prepare for our AI future.

STEP 1: IDENTIFY YOUR AI GOAL

Before you undertake any serious move toward AI, it is crucial to first determine if leadership can agree on a compelling AI use case for the company to pursue. In other words, what problem do you want to solve with AI? The business case should present clear, costed criteria for what constitutes success.

Say, for example, that your aim is to improve employee productivity. Simply stating the objective is only part of the equation. You will want to quantify and qualify that goal as best you can: we want to improve worker productivity by 15 percent by optimizing existing systems and drawing on new data sources such as sensors. Articulating the specific problem creates a direct correlation between the input, its format and shape, and the decision output.

Establishing this baseline is important for company alignment, but it is also critical when it comes to positioning yourself favorably for the steps to follow. As it applies to data, clearly stating your goal will make it easier to determine which data sets you will use and how you need to scrub, transform, or clean them.

STEP 2: DEFINE YOUR DATA STRATEGY

Now that your AI use case is set, it is time to address your data. The first step is identifying what data sets are reasonable to use to achieve your stated goals, including a review of regulatory, legal, ethical, and societal issues that could impact the data you choose. Once you have established the right data sets to use, you will need them in sufficient volumes and with the right variables or features for your purposes. Proceeding without that clarity is not only likely to prove ineffective, it can also be costly. According to IBM estimates, poor-quality data costs the US economy approximately USD 3.1 trillion every year.[3]

As noted above, a great deal of time and energy are devoted to data ingestion and management. To help ensure your organization's effort is contributing to your AI goal, you will want to make data governance procedures a priority. AI relies on sound data to deliver on the promise of greater insights, improved efficiency, predictive planning and response, and better decision-making.

Data scrubbing and formatting is one of the greatest challenges in implementing AI. Data arrives in a range of states and often requires significant preprocessing before it can be successfully analyzed. Missing values are very common in the data acquisition processes. You can either remove those instances or use a probabilistic model to fill in the missing values (i.e., missing value imputation).

Other fixes might include culling duplicate information, correcting any errors or misspellings, and remedying data corruption or noise. You can use data polishing methods to improve the data and/or noise filters to remove noisy instances. If you have large data sets, also be mindful of predictor variables/instances.

You will want to normalize your data because this will make it much easier to use and aggregate across the company later. Call out and set aside any data that conflicts with your models or workflows.

The amount of data you could face initially may seem overwhelming. You could get a bit lost trying to document everything all at once. Try focusing on just the data points specific to your particular business AI use case. Consider forming a cross-business unit task force. This team can head up integrating the different data sets across the company and sorting out inconsistencies to ensure that the data is accurate, rich, and embodies all of the dimensions required for ML.

Prepping your data to ensure it is both relevant and compliant can be difficult and expensive. As a result, many companies are choosing to invest in tools and processes, while others are looking to purchase existing data sets or pretrained models to speed the process. You might also consider setting up repeatable data generation for training and deployment.

The type of data you want to use makes a big difference in the hardware and software tools used to run AI applications. Structured data is where most of us begin today. This includes enterprise resource planning (ERP) systems, billing, financials, customer relationship management (CRM), and sensor data. Unstructured data is becoming much more common but takes more effort to process. Examples of unstructured data include text, web pages, social media, documents, and voice.

STEP 3: CREATE APPROPRIATE STORAGE

Large volumes of data can help build more accurate models. At the beginning of your AI system design, consider adding fast storage that is optimized for data ingest, workflow, and modeling. A data lake can be a useful solution. Still, many sources will be located in distributed environments. That means you will want to make those digitally accessible to help ensure that the data can be consumed at a later date.

Centralizing your data will enable you to better manage it to ensure that this new level of data quality you have achieved can be managed, maintained, and protected. Corporate protocols and processes need to be instituted (if they do not already exist), promoted across the organization, and followed. You will want to label your data as well. Both AI and ML need the labels to accurately analyze the data and produce insights.

The value of data increases when it can be blended together to create new insights. Therefore, moving away from application-centric models where data is bound within applications in prestructured relational databases and moving to open, data-centric environments makes it helpful for interpretation.

Some companies might conclude it is simply unrealistic that they could achieve the necessary level of rich, insightful data, and in the quantities needed to support quality algorithms. If that describes your business, AI-as-a-Service is emerging as a way for companies to combine structured and unstructured data, and external and internal data, into actionable analytics. Intel can also help through its technical experts and consulting partners, and by providing training courses.

STEP 4: TEST YOUR HYPOTHESIS

You now have an established AI goal, developed a data strategy to power it, and considered the storage of that data. At this stage, it can be tempting to simply race into building a model. But making the move to AI is best undertaken methodically. Start simple: use AI incrementally to prove value, collect feedback, and expand accordingly.

Carry out a quick data exploration exercise before you move on. Implement algorithms. Run a set of rules and calculations to make predictions about future outcomes. Look for patterns and behaviors. The aim is to explore and validate the data assumptions and understanding informing your project. It can establish whether the data is actually telling the right story based on your company's expertise and goals.

Taking this important step will also help you get a handle on the important variables or features of the AI use case you have in mind. It should assist you in addressing data categorizations as well, which will be critical as input for potential models.

It is important to define your model-building methodology before getting too far down the road, and you want to engage your business and domain experts as you do so. Their continuous feedback is critical as it will help ensure that all stakeholders are on the same page. Your algorithm will only be as effective as the experts you have involved in this important evaluation.

Again, take your time. Document examples of your test usage and outcomes and show its capabilities. If you share the results across the organization, you can more broadly make the case for AI and inspire others in the company to get involved. And remember, every business is different. Experiment to prove value, and when it fails, start over with a different approach.

STEP 5: DEFINE, TRAIN, AND REFINE YOUR MODEL

You have tested your hypothesis and have outcomes to show. Now it is time to properly define your model and its methodology. Start by defining the performance measures. This will make it easier to evaluate, compare, and analyze the results from multiple algorithms and refine the models later.

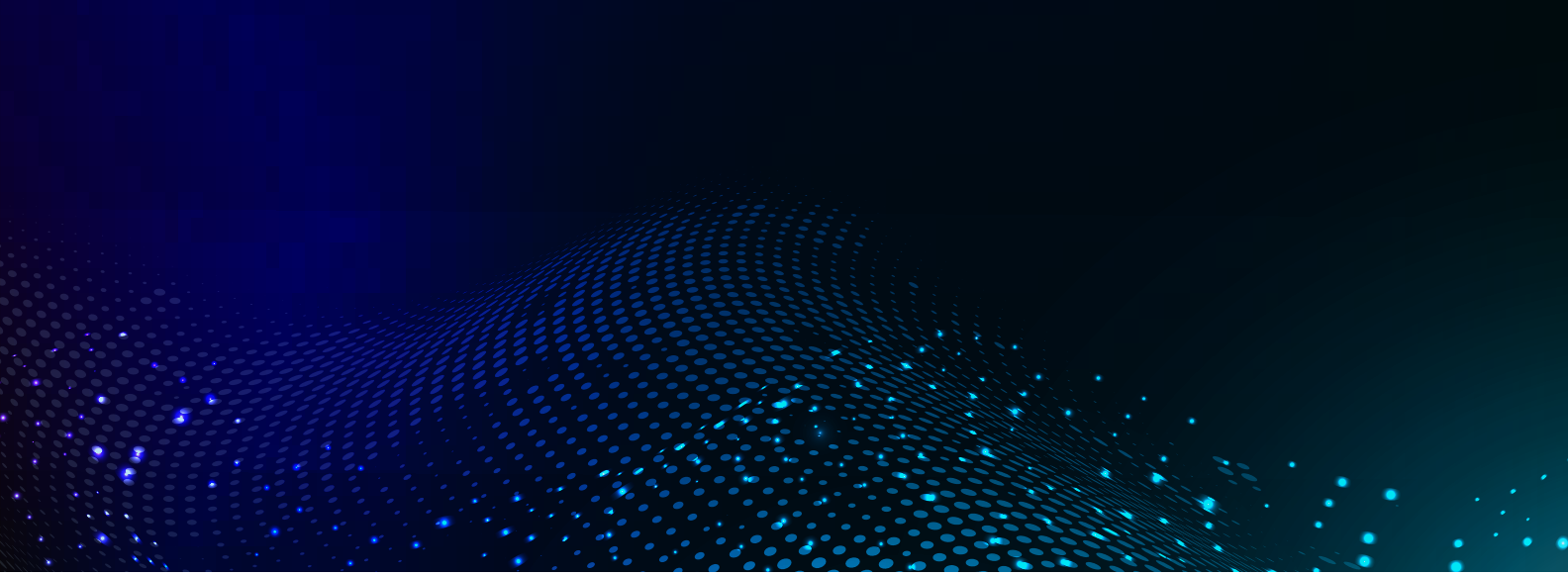
An example of a performance measurement could be classification accuracy. In this case, you find that figure by taking the number of correct predictions made, dividing it by the total number of predictions made, and then multiplying that number by 100.

The best approach is dividing your data into two data sets. The first will be your training set, which will be used to train the algorithm.

The second is your test set against which the training set will be evaluated. You can choose to split the data as you wish (e.g., 60 percent for training and 40 percent for testing), but note that depending on the complexity of your algorithm you may need to undertake more involved sampling processes. And keep your business and domain experts involved here as well.

It is worth noting that a new breed of start-ups is meeting a growing need by selling clean data lakes to help companies develop models. They often also offer pretrained models that you can slowly develop over time as you bring your own data sources online.

Finally, bear in mind that an AI effort tends to be best undertaken by building on your CPU-based systems. That's because most of the world's primary deep learning inference methods already run on CPUs. When focusing on training, the newest CPUs can support much more of the system memory required for complex models. Increasingly, this is the case for common deep learning applications like 2D image detection.



STEP 6: LAUNCH YOUR SOLUTION

After crafting your hypothesis and developing and testing your model, you can now bring the model into the real world of your daily business. That means taking the necessary steps to operationalize and standardize processes. And bear in mind that the new data insights you are making possible will have the greatest impact when they are made available across the company.

A lot of organizations get stuck at this stage. They struggle to determine how to integrate their model into their existing business processes. So, start with a limited rollout. Keep it to a few weeks or months. That should provide sufficient time for users to use, explore, and provide feedback on the model's behavior and outcomes. Take what you learn from that step to adjust the process and educate staff before undertaking a broader rollout.

Also important are the tools and platform you choose for automating data ingestion—and the systems you install for sharing results. Your chosen platform will work best if it offers multiple interfaces that reflect the different degrees of knowledge among your users. For example, some may want to be able to conduct deeper analysis, while others will be perfectly happy to have a dashboard or other kinds of visualizations that offer a high-level picture of status.

STEP 7: REFINE AND REPEAT

After much thoughtful consideration, planning, and testing, your model has been published and deployed. Continuous monitoring is the new objective. As the model gets used, capture any issues or problems you observe or that are shared with you by users. This information will be important when it comes time to update the model.

Time itself is likely to dictate changes to the model as well as it falls out of sync with evolving market dynamics and/or the focus and goals of your company. Because models rely on historical data to predict future outcomes, once the nature of your data changes, the effectiveness of the model is bound to change as well. Therefore, keeping the model as up to date as possible is a key feature of ongoing success.

Finally, keep these seven steps handy. The process can and should be revisited regularly to continue to refine and improve your data. Repeating the steps can also reveal new and potentially beneficial use cases to explore. With your data organized, it will be easier as you begin to incorporate solutions powered by AI—object detection, natural language processing, predictive analytics. That way you can better know how to orient your algorithms for the right outcomes.

The reality is that given its virtually limitless benefits, AI will find its way into a growing number of applications in your organization. As long as you treat your data as the true lifeblood of your business, AI will continue to deliver new insights and new opportunities.

Seize the AI opportunity

AI is no longer part of some distant imagined future. It is now a real and viable tool for empowering businesses to rethink virtually every facet of how they operate. That promise is being made possible by data. In this white paper, we have outlined one strategic path for readying your data to deliver for your organization. The more time you take now to address your data and plan your implementation will only put you in a stronger position to reap the rewards of AI down the road.

A BRIEF GLOSSARY OF AI TERMS

Bad data
An inaccurate set of information, including missing data, inappropriate data (e.g., data entered in the wrong column), nonconforming data, duplicate data, and poorly entered data (e.g., misspellings, typos)

Data categorization (or data classification)
The process of organizing data into categories for its most effective and efficient use to make essential data easy to find and retrieve

Data governance
The overall management of the availability, usability, integrity, and security of data used in an enterprise

Data labeling
The act of taking unlabeled data and augmenting it with meaningful tags that offer details on that data to allow machine learning models to guess or predict the label of other unlabeled data

Data lakes
A system or repository of data stored in its natural format, usually object blobs or files, and usually in a single store that includes raw copies of source system data and transformed data

Data polishing
The act of filtering out the noise in the data set after its detection, or altering it such that it fits into the overall regression of the whole data

Machine learning
The use of statistical techniques to give computer systems the ability to "learn" from data, making data-driven predictions or decisions, without being explicitly programmed

Modeling
The process of documenting a complex software system design as an easily understood diagram, using text and symbols to represent the way data needs to flow

Pretrained models
A source of help for people looking to learn an algorithm or try out an existing framework by using a pretrained model as a benchmark to either improve the existing model or test a new model

Probabilistic model
Models that incorporate random variables and probability distributions into the model of an event or phenomenon, giving a probability distribution as a solution

Learn more about how Intel is powering AI across an exciting set of industry use cases: **ai.intel.com**.

1. McKinsey AI Report, 2017.
2. Murray, Alan, "Fortune 500 CEOs on Trump, the Economy, and Artificial Intelligence," June 8, 2017, http://fortune.com/2017/06/08/fortune-500-companies-ceo-survey/.
3. "Is your business AI-ready?" Genpact, 2017, www.genpact.com/lp/ai-research-c-suite.
4. Ismail, Nick, "Revenue for cognitive/AI systems to top $47 billion by 2020," April 4, 2017, https://www.information-age.com/revenue-ai-systems-top-47-billion-2020-123465508/.

Intel and the Intel logo are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.
© Intel Corporation
1018/RD/CMD/PDF