WEKA insideHPC

December 2020 | WHITEPAPER

10 Things to Know When Starting With AI

CONTENT

INTRODUCTION

Planning for success beyond the initial stages of a project is key.

Artificial intelligence (AI) and machine learning (ML) technologies are disrupting virtually all industries globally, and AI technologies are not just being applied within robotics and vehicle automation. Companies in all sectors are seeing business improvements through insights generated by AI and ML, including financial services, retail, manufacturing, health and life sciences, and more.

Digital leaders are also paying attention to this emerging technology. According to the 2019 Digital Business study by IDG, organizations planned to spend $15.3 million on digital initiatives with AI and ML are high on that list. Despite the enthusiasm around the technologies, however, failure rates on AI and ML projects range anywhere from 50% all the way up to 85%.

Reasons given for these failures include not having a plan ahead of time, not getting executive or business leadership buy-in, or failing to find the proper team to execute the project. Chasing the hot technology trend without having a proper strategy often leads companies down the path of failure.

Fortunately, enough lessons have been learned through these failures to give companies a better game plan for their next AI or ML project.

QUESTIONS TO ASK AND ANSWER

Listed below are 10 questions AI teams should ask themselves when they are beginning new AI projects.

#1: Have we clearly defined a goal and identified the right questions to get us there?

Amazingly, many companies do not have a clear vision of the goal they want to achieve from an AI project. Moreover, they don't have a good sense of the questions they need to ask and answer in order to take the necessary steps on the path toward achieving it.

"A lot of companies will start with 'We know that AI is a game changer, so let's see what we can do with it,'" says Shimon Ben David, the Field CTO at WekaIO, which offers a parallel file system to help companies with large storage problems—very much like those faced by companies starting their AI journeys.

Like an explorer preparing to reach a destination, a project leader needs to establish a final end point and then provide a map that includes specific directions to follow for each step of the journey. The specific outcomes need to be identified for AI projects, and then directions can be formed by asking and answering questions to help reach goals and achieve the desired outcomes.

The key here is to create a good AI team with the ability to ask and answer these initial questions. This team could include software engineers, business leaders, subject matter experts, and possibly even customers within the environment.

For example, let's imagine a financial institution that has the ultimate goal of increasing its bottom line by improving its profit margin. The first question to ask is, "How do we use AI to do this?" One answer could be to consider using AI to help decrease the high rate of defaults on loans, thereby getting a better return on investments.

Therefore, who could ask the right questions to identify customers with the highest risk of defaulting on a loan? In this case, the institution's individual account team members would be good candidates to ask questions and gather the data because they're the ones closest to the sources of data—the customers, themselves. The account team knows the situational problems that customers face and routinely interacts with them, often hearing the reasons why payments come in late, which leads to loan status jeopardy and, sometimes, default.

For good customers, the institution can then provide either rewards or incentives, such as a lower interest rate. For the high-risk customers, the institution can offer programs and monitoring to make sure they stay on track or to get them out of the high-risk category.

Keep in mind that the questions a company creates to get to the final goal can change and evolve as more data is gathered. If you've chosen the right goal, it should stay the same, but the steps to get there might change as you encounter roadblocks and obstacles. If you haven't identified the right goal, asking questions will make that clear so you can pivot in the right direction.

"Companies need to maintain a constant set of questions, and chances are good that they will morph with a project's progress, but it's imperative that you have an initial response to them at the beginning, thereby giving yourselves a place to start," says Ben David.

#2 What data is required to achieve your goal or solve your problem?

After an AI project team has identified the goal or specific problem that AI can solve, the team continues to ask questions to determine the data or variables required to achieve the goal or resolve the specific problem.

In the case of the financial institution, after customers in the at-risk category have been identified, the team has reached only one step toward reaching the goal. Remember that the goal is not only to define those at high risk for defaulting but also to keep them from actually defaulting so that the institution can increase its profit margin.

The team asks more questions to reach the next step: Does everyone in the high-risk category face the same situation that keeps them from paying? Probably not, so how does the team identify and categorize customer subsets that need different forms of help to achieve timely payment? What are the remedies that might help people in each subset and keep them from defaulting?

This point marks the spot where evaluating the full data set comes into play. A bank may have a customer's name, personal information, banking details, social media postings, images, videos, and other records in order to answer those questions. Multiple data points exist, and companies might not need all of them. On the other hand, some information might be missing. In reality, most companies begin an AI project thinking they have enough data to answer the questions, but often a good portion of the data is missing, or the data that they have isn't useful to answer the question. In his experience, Ben David says that he has never encountered a company that has collected too much data.

"Maybe I have bank records, but they don't have a credit score," says Ben David. "Maybe I don't have social media with relevant tags that they posted that could help me understand their financial situation. Understanding the data and what's in the data is important."

Sometimes companies have to come up with their own data to fill in for what's missing. The tools that you would use to extract your data set would vary depending on the type of data you need to collect. For example, Google Analytics provides website visitor data and metrics, but you might also have a customer or contact database through Hubspot, Salesforce, or numerous other services.

Fair warning: Keep everything! Companies tend to acquire a vast amount of data, distill it when creating the AI or ML model, and then either store the raw data somewhere never to be accessed again or, even worse, delete the unused data. Ben David says that data can be critical later when reassessing a particular model where raw data is needed again.

For example, look at how criminologists utilize newer DNA technology and methods to help find or exonerate suspects in a crime that happened years or decades earlier. Because evidence is stored and saved in these cases, criminologists can go back and re-analyze the clues. The same principle applies with AI: you might not think you need all collected data now, but years down the road a better algorithm or new technological advancement may elevate a piece of seemingly useless data into a highly relevant piece of evidence (think DNA sampling of hair strands), and you would be wise have kept that old data available.

#3 Where will I get my data if I don't have it already?

If you find yourself needing more data, the next step would be to determine where you can get the data you need. Do you generate it (as in the case with our aforementioned customer questionnaire), do you buy it, or do you rent it?

For example, a medical company embarking on an AI project involving genetics might look at data in a public genome database, but then the researchers might discover that they do not have the data needed for their particular AI model, in which case they might need to conduct their own experiments. Alternately, perhaps they need only a single piece of data in an image versus looking at a complete set of labeled data.

"You want to make sure you know where you will acquire the data at the starting point of the journey, but also with the understanding that this could change along the way," Ben David says.

In another example, imagine a farmer that sends drones out in the field to take multiple pictures and collect data through sensors for crop tracking or soil moisture. Even if the farmer conducts this data discovery for a month, conditions change on a regular basis (weather, crop growth, wildlife, etc.) to the extent that the data collection truly is never finished. Data acquisition is not a one-and-done proposition. "You need to plan ahead for when and where you will get your next batch of data and take the steps to acquire it, often in parallel with your other work" says Ben David.

Several websites let you search for free data sets to use in your machine learning training models:

- Google Dataset Search (https://datasetsearch.research.google.com/)
- Kaggle (https://www.kaggle.com/datasets)
- VisualData Discovery (https://www.visualdata.io/discovery) for computer vision datasets

Here are two great articles that provide other locations to search:

- "Best Public Datasets for Machine Learning and Data Science" (https://medium.com/towards-artificial-intelligence/best-datasets-for-machine-learning-data-science-computer-vision-nlp-ai-c9541058cf4f)
- "Top Sources for Machine Learning Datasets" (https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b)

#4 What is our organizational compute strategy: on-premises, cloud, or hybrid?

A big way to get in trouble with an AI project is to have it run on a computing platform that is not aligned with the organization's overall digital compute strategy. Knowing current and future plans can help an AI team properly plan for the best way to approach which platform to use for AI or ML models.

"You want to take the most effective way that aligns with your organization's strategy. It could be that your organization is heavily invested in an on-premises environment with multiple GPUs," says Ben David. "You might as well leverage that because it would be your fastest path to success."

AI and ML projects can find success with on-premises, cloud, or hybrid platforms as long as they align with a company's overall strategy and won't conflict with changes or modifications down the road. Smaller companies might start with a cloud environment because they think it's faster and less expensive, yet they may find that the costs become larger as the projects grow, making more sense to move to on-premises environments.

#5 What is my plan to move and store the data?

As they work to process the AI models, companies often discover they did not plan for how they would store and move their data. Imagine a company with divisions all over the world, generating petabytes worth of data in multiple locations across different continents. "Do I try to process it where it was created, or do I try to move petabytes of data somehow between sites worldwide?" Ben David asks. "It's one of the critical things that sometimes is not considered in AI projects."

Another option is to centralize the data in a single data center, but moving data includes the possible need to compress data or physically ship it instead of transferring it across the cloud, which can get expensive quickly. Moreover, making sure the data is secured is also an issue, as some data cannot be moved due to local or federal regulations. Finally, by the time the data arrives at the site of AI processing, you might find that it's already obsolete.

"Each organization has a different answer, and most often they're all correct," says Ben David. "But if you do not think about this on day one, then you are more than likely to have a problem."

Additionally, companies need to consider having a strategy for retaining data for future use. In many cases a company cannot generate data from experiments over and over again. Data from these experiments needs to be saved, stored, and secured, yet also be available for quick retrieval if needed. As mentioned earlier, this includes raw data that may seem irrelevant now but be needed later as the AI model grows and the ability to analyze evolves. Ben David stresses the idea that raw data should not be deleted nor ignored. "This notion cannot exist in an AI project," Ben David says.

#6 How will we validate our model's results as well as remove bias?

After data is collected and in place, make sure you know how to validate the results that the AI or ML model is generating. One such way is to run it against a known data set and look at the results to make sure that you have a high level of accuracy on the expected results.

For example, if your AI algorithm is identifying a batch of photographs and determining which ones are apples and which are oranges, will your model accurately identify the correct fruit? Ben David says humans can often validate the answers on a simple level, but this ability doesn't scale well when the data set includes hundreds or thousands of images. In this case, AI experts often run validations through a simulator, which can verify the AI models on a larger scale.

Furthermore, validating the results is an important step in determining whether the AI has any inherent biases built into the model. One well-known example was when Amazon discovered that a resume-screening application was not rating candidates for software developer jobs and other technical positions in a gender-neutral way. Because the models were trained to choose applicants by observing patterns in resumes submitted over a 10-year pattern, most of the resumes were from men (who, at the time, dominated the field).

When evaluating your AI models, be sure to have a strategy for spotting and eliminating bias, or the results you end up with could be skewed and affect the project's credibility.

#7 How often will we fine-tune the models?

Because much of AI and ML is based in software, developers often adopt a "set it and forget it" approach, which can be disastrous for this technology. Fine-tuning not only involves being ready to change the model regularly but also understanding how practitioners can change different variables within the model to achieve different results.

Some AI models, for example, will provide results based on your data and also explain how they achieved those results. Others, however, simply spit out results and leave it up to the data scientists to figure out why, causing what many data scientists refer to as "explainable AI." "Any AI project is always a work in progress," says Ben David. Creating and executing on a model that can provide good reasons for its decisions is an important step in building trust in the model.

Fine-tuning (and whether to deploy a new model, discussed in Question #8) can often be a result of discovering that you have "bad data." In general, bad data is data that has not been "cleaned up," or it contains missing fields, duplications, or the data type is not in the correct format, such as dates written in text instead of the date format.

But even clean data can still be considered bad if it is too specific or presents biases, such as problems generated in facial recognition or the gender-bias that was discovered in Amazon's resume scanning application. The data may have appeared to be good initially but turned out to be bad after the algorithm kept eliminating female resumes because the model didn't account for fewer female resumes in the historical data. This mistaken elimination indicated that the historical data was not broad enough.

The best way to determine whether your data is good or bad is to first make sure it is clean, then check that it is broad enough to produce unbiased results.

#8 How do we deploy a new model?

With a model that is fine-tuned on a regular basis, companies then need to have a strategy around the possibility of deploying new AI models that can better answer the original questions or the possibility of generating new questions based on the results they are seeing.

For example, at some point data scientists may decide to move to a different neural network for their AI model or algorithm, which might require creating something new rather than fine-tuning or modifying an older model. Many of these decisions are dependent on the specific algorithms that companies use or on the goals that they are aiming to achieve, but an AI team's radar screen should include the question of how to deploy a new model should the need arise at a later date.

Some may think that acquiring more data is a way to fine-tune or create better outcomes, but this can be a trap for many companies. If the data is not good to start with, adding more of it will not suddenly solve the problem. When people suggest that getting more data will help, they are often implying the need to acquire a broader data set that meets high quality standards.

In a 2018 article for Harvard Business Review, Thomas C. Redman, president of Data Quality Solutions, said good data must be right in two ways: First, it must be correct, properly labeled, de-duplicated, etc. Second, make sure it's the right data for your project.

> "But you must also have the right data — lots of unbiased data, over the entire range of inputs for which one aims to develop the predictive model. Most data quality work focuses on one criterion or the other, but for machine learning, you must work on both simultaneously."

Earlier this year, Redman also spoke about how companies often waste critical resources in dealing with bad data in an MIT Sloan Management Review article. "Bad data, in turn, breeds mistrust in the data, further slowing efforts to create advantage," he said.

#9 How does my infrastructure look on day 3 vs. day 300?

AI projects are constantly changing and evolving. The algorithm or software could change, as could the computing infrastructure, meaning that the model could start to run on company-owned servers and then convert to running in a public cloud or a hybrid platform. If a company has aligned its AI data strategy with the organization's overall compute strategy (see question #3), this is not much of a problem.

"For example, today a company might be running on premises, with one or two data scientists running from their laptops with an external GPU," says Ben David. "I know that if everything works out in a year, then I'll have 20 data scientists, and then I'll need a heavier infrastructure. You want to plan for that. Again, the notion is that if you know it on day one, two and three, etc., then you can plan ahead for it."

As data volumes scale and the models become more complex, so does the need for more robust compute; otherwise, the fact that you have 20x the volumes of data means that your models will take 20x longer, reducing productivity and agility. Compute needs pipes that can saturate it, so you want to make sure that you can expand your pipes, (i.e., your network) accordingly.

One frequent and expensive mistake companies make is not planning for the significant data growth over the course of the project. Amassing 20x more data means a significant increase in storage costs and additional delays, often due to storing more data in cold tiers and moving them back and forth to hot/fast tiers. Those reads and writes are time consuming. Some companies tier some data in the cloud for economies of scale and flexible capacity, which introduces management overhead with multiple name servers and different operational models.

Newer file systems, such as WekaFS, manage the different tiers under a single name server with throughput that is comparable to local storage. Using a modern file system can dramatically alleviate the cost and management burden, helping you to keep

productivity high as data increases. Most modern file systems are designed from the ground up to support exabytes of data and AI and ML workloads.

#10 How do we future-proof the project?

Ben David says he sees many companies kicking off AI projects with high hopes for success, but the team has not taken a holistic view of the entire project, so down the line they run into trouble when it comes to growth. "We see projects that are starting with some environments that are adequate for one to five data scientists, but then the environment expands and suddenly they need additional infrastructure," he says. "More often than not, you see customers trying to extend their existing infrastructure instead of re-architecting it."

For example, a data scientist might start to work on a single laptop, and then additional data scientists are brought in, and suddenly the team needs to work on a network-attached storage appliance. On the other hand, a project might start in the cloud, but then the team suddenly has 10 to 50 data scientists contributing to the project, so business leaders determine that it is more cost-effective to buy on-premise equipment for the computing, network, and storage environment. Having a strategy around how to effectively manage the growth and to scale the project can help future-proof a company's AI project.

KNOWLEDGE IS THE KEY

It is possible for many AI projects to succeed without having all of the answers or without following the strategies laid out here. Nevertheless, the long-term success of a project must have an AI team willing to be flexible on infrastructure changes, willing to fine-tune their model, and forward-thinking enough to have a plan to move and store data safely and efficiently.

With these plans in place, your chances for success can go beyond the 15% to 50% success rates that many of today's AI projects experience today.

Additional Resources and Links

How to Manage Data in the AI Era - Accelerated DataOps
https://www.weka.io/blog/accelerated-dataops-with-weka-aiedge-to-core-to-cloud-pipelines-part-1

Weka AI and NVIDIA Accelerate AI Data Pipelines
https://www.weka.io/blog/together-weka-ai-and-nvidia-accelerate-streamline-and-protect-edge-to-core-to-cloud-data-pipelines/

Data management in the Age of AI
https://www.forbes.com/sites/forbestechcouncil/2020/07/13/data-management-in-the-age-of-ai/#6c277485161f

6 Reasons to Re-Architect Your SAS Analytics for Large Datasets
https://www.weka.io/blog/architect-your-sas-analytics/

Accelerated DataOps Paves the Pathway to Explainable AI
https://www.weka.io/blog/accelerated-dataops-with-weka-aiedge-to-core-to-cloud-pipelines-part-3/

WekaFS Architecture White Paper
https://www.weka.io/wp-content/uploads/files/2017/12/Architectural_WhitePaper-W02R6WP201812-1.pdf

The One Thing Companies Neglect to Consider When Starting With AI (archived webinar)
https://hs.weka.io/webinars/optimizing-ai

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io www.weka.io

W07WP202012