RESEARCH
Influence and insight through social media

August 2018

The Success of

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Requires an Architectural Approach to Infrastructure

WHITE PAPER

Prepared by
Zeus Kerravala

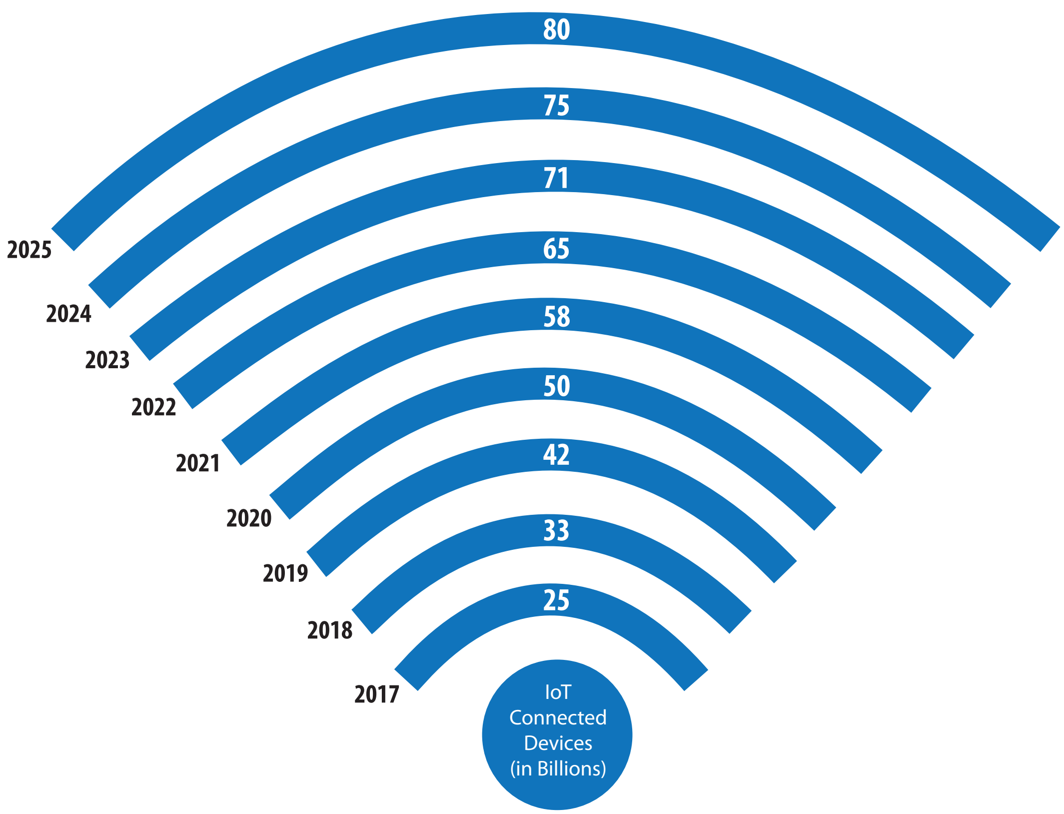
ZK Research
A Division of
Kerravala Consulting

INTRODUCTION: THE MACHINE LEARNING ERA IS HERE

Artificial intelligence (AI) and machine learning (ML) are emerging technologies that will transform organizations faster than ever before. In the digital transformation era, success will be based on using analytics to discover the insights locked in the massive volume of data being generated today. Historically, these insights were discovered through manually intensive data analytics—but the amount of data continues to grow, as does the complexity of data. AI and ML are the latest tools for data scientists, enabling them to refine the data into value faster.

In the past, businesses worked with a finite set of data generated from large systems of record. Today, there are so many more endpoints connected to a business, each generating its own set of data that needs be analyzed. For example, a decade ago, the concept of the Internet of Things (IoT) did not exist. Now, businesses are connecting new devices at a furious rate. ZK Research forecasts that by 2025, there will be 80 billion connected endpoints (Exhibit 1), each generating significant volumes of data. IoT isn't needed for AI and ML, as many other data sources exist, but the addition of IoT accelerates the need for AI and ML. Given the difficulty companies have analyzing today's volumes, it's impossible to see how organizations will adapt to the upcoming explosion of ingested data. The only way to compete effectively is by using AI and ML.

Exhibit 1: IoT Is One of Many Data Sources That Demand AI and ML for Deeper Analysis



ZK Research 2018 Global IoT Forecast

ABOUT THE AUTHOR

Zeus Kerravala is the founder and principal analyst with ZK Research. Kerravala provides tactical advice and strategic guidance to help his clients in both the current business climate and the long term. He delivers research and insight to the following constituents: end-user IT and network managers; vendors of IT hardware, software and services; and members of the financial community looking to invest in the companies that he covers.

AI and ML are applicable across all verticals—hence there is no single "killer application."

The terms "machine learning" and "artificial intelligence" are often used interchangeably, which is incorrect. AI is a broad term used to describe the process in which computers mimic human intelligence. Machine learning is a set of algorithms that can create models that mimic the exhibited behavior based on a data set. Because there are also other ways for algorithms to mimic human intelligence, machine learning is generally considered a subset of artificial intelligence.

AI and ML are applicable across all verticals—hence there is no single "killer application." Each enterprise has different business challenges and access to different data sets. Therefore, the approach and application of AI and ML will differ. Despite the differing uses of AI and ML, here are some of their more common use cases:

Anomaly detection: Based on the set of training data, machine learning–based systems can identify things that are anomalistic in nature. A common use case for this is in the healthcare industry, where AI can locate bleeds, tumors, or other problems in brain MRI scans that are typically indiscernible during human inspection.

Classification: An AI system learns from a set of training data and can then classify new inputs into specific groupings. An example of this is when a self-driving car sees an object and can then categorize it as a tree, a person, a sign, or another object.

Predictions: AI is used to estimate or predict the next value in a specific sequence. Human prediction has obviously existed for decades, but AI can incorporate a wider set of data. For example, a retailer can more accurately forecast future sales by including weather information.

Recommendations: An AI system can make specific suggestions regarding responses to questions or comments. These systems are gaining traction in contact centers to help agents respond to common customer complaints. Also, chatbots are now being used to suggest products to buy based on the patterns of similar individuals.

Categorization: Data often needs to be stored in specific clusters. An AI system can analyze large data sets and group data instances by common traits. For example, when studying shopping habits, AI can determine specific age demographics such as age and income. Another emerging use case is using AI for voice analytics to determine if an audio stream contains a person's voice or background noise, the latter of which should be automatically muted.

Translation of information: Machine learning–based systems can be used to quickly translate data from one form to another. The best example of this is AI combined with natural language processing to help people on a video call each speak in his or her native language. AI can understand what the language is and translate it in real time so everyone can participate in the conversation.

Success with AI and ML depends on having the right infrastructure to process the data fast enough to keep up with the business.

"Deep learning" is another term that is commonly used in AI circles to describe the utilization of deep layers of neural networks. Traditional machine learning algorithms are linear or "shallow" in nature, whereas deep learning algorithms use neural networks to handle the varying complexity and abstractions of the incoming data. Consider the following example. A parent tells a child what a cat is by pointing to a cat and then confirming it is indeed a cat. The parent then can point out what is not a cat as well. Over time, the child becomes more aware of which features define a cat and which ones do not and is eventually able to quickly identify them. The child is actually doing complex abstraction (i.e., cat identification) by building a hierarchy in which each level of abstraction is created using the knowledge stored in preceding layers.

Computer deep learning goes through roughly the same process, where each algorithm in the hierarchy applies a non-linear transformation to the data to create a model for inferencing. The stacked processing layers can be very deep—hence the term "deep learning." However, deep learning requires a massive amount of data.

Shallow learning is much more common, and it uses and requires less data than deep learning. Shallow learning techniques include linear and logistic regression and decision trees. They also require more data scientists to supervise the learning process, especially when identifying the subset of input that leads to the best results. This selection of subsets is often referred to as feature engineering.

Because deep learning requires less feature engineering, many businesses have come to prefer this approach. However, deep learning also requires more data, leading to the need for more hardware to make sense of the information.

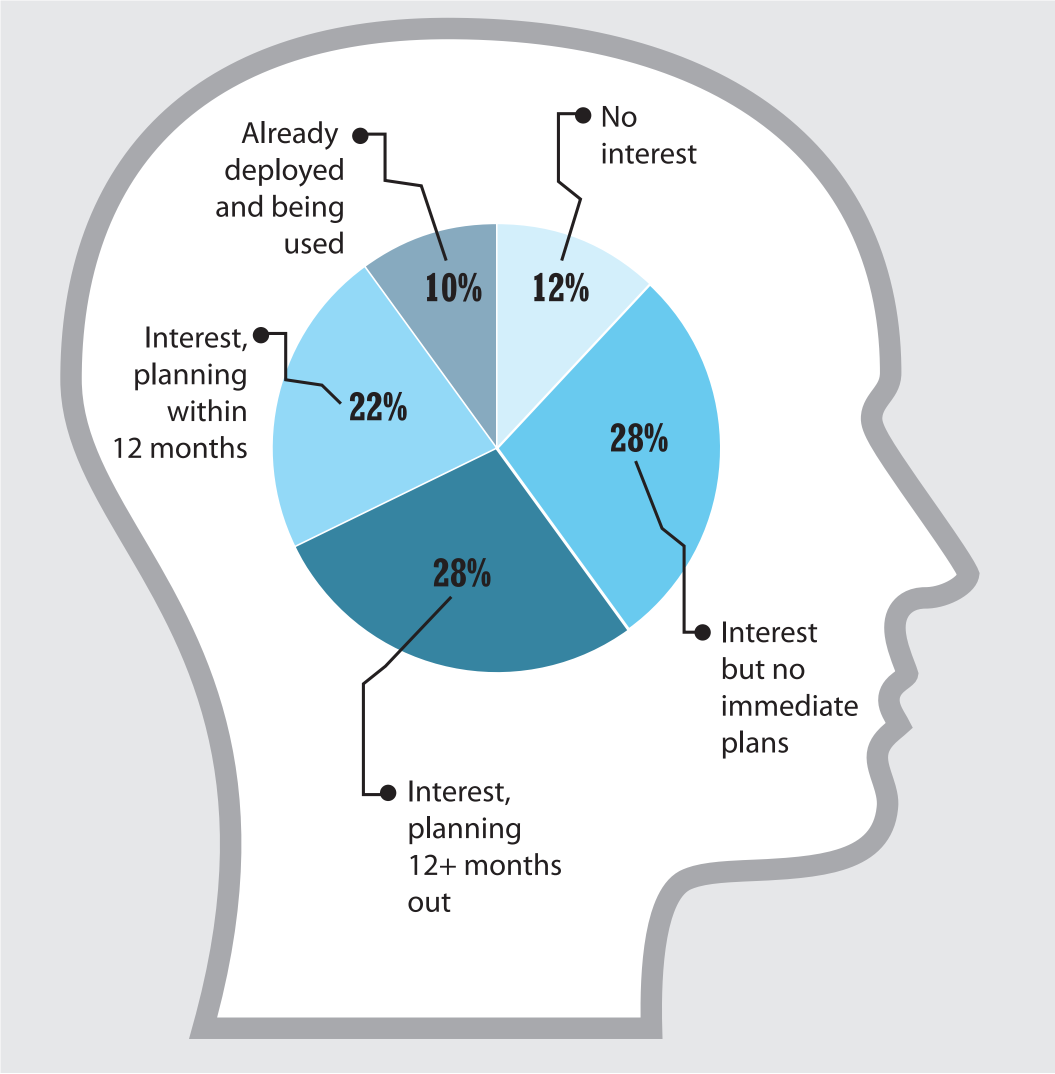
The rise of AI and ML will transform every business in every industry vertical. It will streamline operations, save a significant amount of money and enable companies to interact with employees and customers in entirely new ways. However, success with AI and ML depends on having the right infrastructure to process the data fast enough to keep up with the business.

SECTION II: ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING ADOPTION TRENDS

The technology industry sits on the precipice of the biggest change in history. Many businesses are already deploying AI and ML. The rise of AI- and ML-based systems will forever change the business landscape, and therefore CIOs need to be ready. To the IT team, AI and ML may seem like something futuristic, but it's happening right now. In the ZK Research 2018 IT Priorities Study, we asked more than 800 IT decision makers, including CIOs, about the state of AI in their organization, and only 10% said it had already been deployed (Exhibit 2). However, the survey also showed that another 22% have interest in and plans to implement AI within 12 months, and 28% have plans to do so in 12 or more months, indicating there is a swell of implementations coming.

One challenge for CIOs is that, unlike most IT projects, AI and ML initiatives are not solely IT driven. Line-of-business managers, data scientists, application developers, customer care managers, marketing teams and other departments all have a vested interest in AI initiatives. The level of

Exhibit 2: AI Is on the Near-Term Horizon



ZK Research 2018 IT Priorities Study

AI savviness will obviously vary from company to company, but ZK Research has found that most customers will generally fall into one of these two categories:

Experienced: These organizations have skilled developers and have deployed AI and ML systems. In these organizations, the data sciences teams would lead the initiatives. However, as infrastructure requirements change, it's likely that the IT organization will have more influence. Experienced organizations will be familiar with the specific infrastructure needs at every point in the AI journey and will likely buy best of breed for the different phases such as testing and development, training and inferencing.

AI and ML success is largely based on the quality of data fed into the systems.

Little to no experience: As Exhibit 2 showed, a good chunk of the population is just starting the AI journey with few or no skilled developers. These companies will likely turn to the IT organization for their infrastructure needs. Because AI initiatives are in their infancy, it will be hard for IT to know what infrastructure to buy—meaning an end-to-end solution will be best. Also, it's likely that the company will need some professional services assistance to ensure the deployment goes smoothly.

Regardless of how much experience businesses have, the infrastructure choices they make will largely dictate the success of their AI and ML initiatives. Making the wrong choice increases complexity and costs—and means highly paid data scientists will experience significant amounts of delay.

SECTION III: MAKING THE RIGHT INFRASTRUCTURE CHOICE

One key to making the best infrastructure choice is understanding the role of data. AI and ML success is largely based on the quality of data fed into the systems. There's an axiom in the AI industry stating that "bad data leads to bad inferences"—meaning businesses should pay particular attention to how they manage their data.

Data plays a key role in every use case of AI, although the type of data used can vary. For example, innovation can be fueled by having machine learning find insights in the large data lakes being generated by businesses. In fact, it's possible for businesses to cultivate new thinking inside their organization based on data sciences. The key is to understand the role that data plays at every step in the AI/ML workflow (Exhibit 3).

One of the most significant challenges with data is that building a data pipeline in real time is very difficult. Data scientists who are conducting exploratory and discovery work with new data sources need to collect, prepare, model and infer. Therefore, IT requires change during each phase and as more data is gathered from more sources.

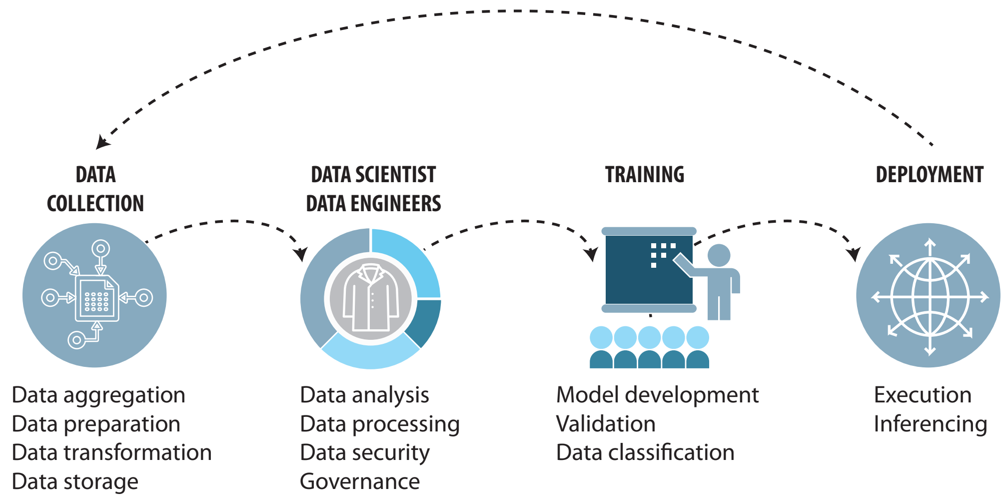
It's also important to note that the workflow is an iterative cycle in which the output of the deployment phase becomes an input to data collection and improves the model. The success of moving data through these phases is largely dependent on having the right infrastructure.

Key considerations for infrastructure are as follows:

Location: AI and ML initiatives are not solely conducted in the cloud nor are they handled on premises. These initiatives should be executed in the location that makes the most sense given the output. For example, a facial recognition system at an airport should conduct the analysis locally, as the time taken to send the information to the cloud and back adds much latency to the process. It's critical to ensure that infrastructure is deployed in the cloud, in the on-premises data center and at the edge so the performance of AI initiatives is optimized.

Breadth of high-performance infrastructure: As mentioned earlier, AI performance is highly dependent on the underlying infrastructure. For example, graphical processing units (GPUs) can

Exhibit 3: The Role of Data in the AI/ML Workflow



ZK Research, 2018

accelerate deep learning by 100 times compared to traditional central processing units (CPUs). Underpowering the server will cause delays in the process, while overpowering wastes money. Whether the strategy is end to end or best of breed, ensure the compute hardware has the right mix of processing capabilities and high-speed storage. This requires choosing a vendor with a broad portfolio that can address any phase in the AI process.

Validated design: Infrastructure is clearly important, but so is the software that runs on it. Once the software is installed, it can take several months to tune and optimize it to fit the underlying hardware. Choose a vendor that has pre-installed the software and has a validated design in order to shorten the deployment time and ensure the performance is optimized.

Extension of the data center: AI infrastructure does not live in isolation and should be considered an extension of the current data center. Ideally, businesses should look for a solution that can be managed with their existing tools.

End-to-end management: There's no single "AI in a box" that can be dropped in and turned on to begin the AI process. It's composed of several moving parts, including servers, storage, networks and software, with multiple choices at each position. The best solution would be a holistic one that includes all or at least most of the components that could be managed through a single interface.

Network infrastructure: When deploying AI, an emphasis is put on GPU-enabled servers, flash storage and other compute infrastructure. This makes sense, as AI is very processor and storage intensive. However, the storage systems and servers must be fed data that traverses a network.

No single AI vendor can provide all technology everywhere.

Infrastructure for AI should be considered a "three-legged stool" where the legs are the network, servers and storage. Each must be equally fast to keep up with each other. A lag in any one of these components can impair performance. The same level of due diligence given to servers and storage should be given to the network.

Security: AI often involves extremely sensitive data such as patient records, financial information and personal data. Having this data breached could be disastrous for the organization. Also, the infusion of bad data could cause the AI system to make incorrect inferences, leading to flawed decisions. The AI infrastructure must be secured from end to end with state-of-the-art technology.

Professional services: Although services are not technically considered infrastructure, they should be part of the infrastructure decision. Most organizations, particularly inexperienced ones, won't have the necessary skills in house to make AI successful. A services partner can deliver the necessary training, advisory, implementation and optimization services across the AI lifecycle and should be a core component of the deployment.

Broad ecosystem: No single AI vendor can provide all technology everywhere. It's crucial to use a vendor that has a broad ecosystem and can bring together all of the components of AI to deliver a full, turnkey, end-to-end solution. Having to cobble together the components will likely lead to delays and even failures. Choosing a vendor with a strong ecosystem provides a fast path to success.

San Jose–based Cisco is best known as the market leader in networking. Over the past decade, the company has also built a strong server and hyperconverged infrastructure (HCI) portfolio with a wide range of options for different use cases. Recently, it has upgraded its portfolio to meet the needs of the different phases of AI. It is the opinion of ZK Research that Cisco has a broad range of solutions and products that fulfill the criteria listed above. Cisco can help customers meet their AI and ML goals regardless of where they are in the deployment cycle.

SECTION IV: CONCLUSION AND RECOMMENDATIONS

The AI and ML era is here, and businesses are on the precipice of the biggest technology change in the history of business. Massive amounts of data are being generated today, and businesses need to analyze it and find new insights faster than the competition. Those that can do that will gain a significant and sustainable competitive advantage, while those that cannot will quickly fall behind and struggle to survive.

Historically, AI and ML projects have been run by data science specialists, but that is quickly transitioning to IT professionals as these technologies move into the mainstream. As this transition happens and AI initiatives become more widespread, IT organizations should think more broadly about the infrastructure that enables AI. Instead of purchasing servers, network infrastructure and other components for specific projects, the goal should be to think more broadly about the business's needs both today and tomorrow, similar to the way data centers are run today.

Over the next several years, businesses will find more use cases for AI and ML. These technologies will become a core component of every organization's IT strategy, and IT leaders must be ready. To help prepare for this transition, ZK Research makes the following recommendations:

Partner with lines of business. AI and ML are not solely owned by the IT department, as the output can have significant business value. IT leaders need to understand this and work with line-of-business data scientists to meet the demands of a constantly changing environment.

Start with small concurrent projects in different areas. Inexperienced companies should experiment initially with small projects, using AI projects for educational purposes instead of focusing on ROI. Learn from these projects and apply the learnings to future projects to refine the skills.

Think broad when making infrastructure decisions. The infrastructure requirements at each phase of the AI/ML workflow will vary greatly. For instance, the needs in the learning phase are different from those in the inferencing phase. However, before any decision is made, take a step back, look at AI and ML holistically, and choose a vendor that has a purpose-built product for each phase but also a common management tool. This delivers the benefits of best of breed and marries them with the advantages of an end-to-end solution.

Augment the skill set with professional services. Even experienced firms will have gaps in knowledge. Leverage a services partner to complement the current in-house engineering team with skills that are lacking. This includes every phase of deployment such as business case analysis, implementation services, architectural and design support, and optimization services.

CONTACT

zeus@zkresearch.com

Cell: 301-775-7447

Office: 978-252-5314