AI on the Edge: Fusing Artificial Intelligence and IoT Will Catalyze New Digital Value Creation

Published 5 June 2017 - ID G00319614 - 21 min read

By Analysts Mark Hung, Tom Austin

Supporting Key Initiative is Artificial Intelligence

Artificial Intelligence and the Internet of Things are symbiotic technologies that will be the foundation of a new platform for digital business value creation. CIOs engaged in IoT initiatives should leverage these capabilities for strategic advantage.

More on This Topic
This is part of two in-depth collections of research. See the collections:

- IoT Technology Disruptions: A Gartner Trend Insight Report
- Applying Artificial Intelligence to Drive Business Transformation: A Gartner Trend Insight Report

Overview

Key Findings

- IoT and AI are impacting digital business from both a scalability and algorithmic perspective. Together as a new business platform, they will support more complex analytics and decision making in the face of accelerating data growth from IoT.
- Three architectural styles provide the blueprint for IoT-AI deployments: IoT data as the input to an AI system, AI as the application on an IoT endpoint, and IoT and AI as a two-way system.

- Adoption of AI in IoT will progress in different phases, depending on the maturity level of the deep learning paradigm and the organization's target application.
- The current embryonic stage of the IoT-AI platform will bias it toward experimentation and incubation rather than development and implementation.

Recommendations

CIOs responsible for strategic IoT initiatives should approach the emerging IoT-AI platform as follows:

- Determine whether the new AI breakthroughs in deep learning are applicable to your IoT deployments, or whether traditional data analytics and AI methodologies are adequate.
- Map the IoT solution and AI system to the most suitable architectural style (or combination of styles) by examining the project's functional requirements and data characteristics.
- Pilot visual and/or audio-related applications that have proven to be most effective with deep learning. After some initial success, follow with the exploration of more unproven (but potentially high-reward) approaches.
- Define the business problem first, in the face of the nascent nature of IoT and AI. With the two technologies still rapidly evolving, create a "sandbox" where education, evaluation, partnership and experimentation can mature until the IoT-AI platform "hatches."

Strategic Planning Assumption

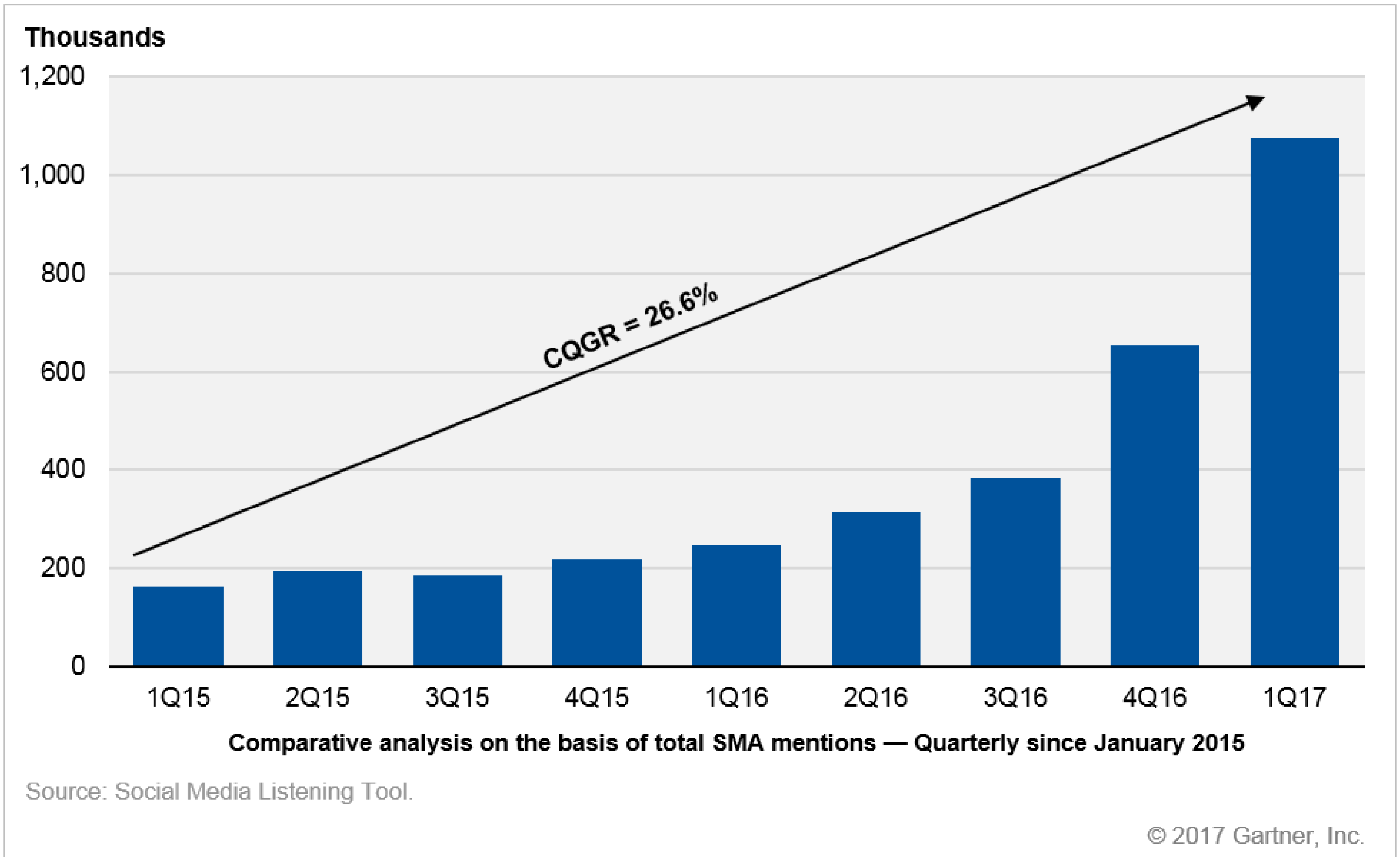
By 2022, more than 80% of enterprise IoT projects will have an AI component, up from less than 10% today.

Analysis

Introduction

Two of today's most hyped technologies are AI and IoT. Hardly a week goes by without some "breaking news" about either or both of them, and social media conversation volume around AI has grown by more than a 26% CQGR in the last two years across a diverse set of industries (see Figure 1). Coincidentally, the two technologies strongly complement each other, with each technology's strength addressing the other's weakness. Forward-thinking CIOs responsible for strategic IoT programs can bring these technologies together to create a powerful new platform for digital business value.

Figure 1. Social Media Conversations on AI Have Grown



Source: Adapted from Social Media Listening Tool

For IoT, the hype builds on the capability of instrumenting a wide range of devices with sensors, and then networking the devices to collect and process data. Gartner's most recent forecast projects that the number of IoT endpoints will grow at a 33% CAGR from 2015 through 2020, reaching an installed base of 20.4 billion units (see "Forecast Analysis: Internet of Things — Endpoints, Worldwide, 2016 Update" (https://www.gartner.com/document/code/302435?ref=grbody&refval=3738018)). The cross-industry category, in which devices are used across multiple vertical industries, will have the highest CAGR, at 39%, driven by energy management and automotive applications. Consumer IoT will grow nearly as fast, with a CAGR of 34%, driven by a mix of products in consumer electronics, energy management, home security, automotive and emerging connected "everyday" items.

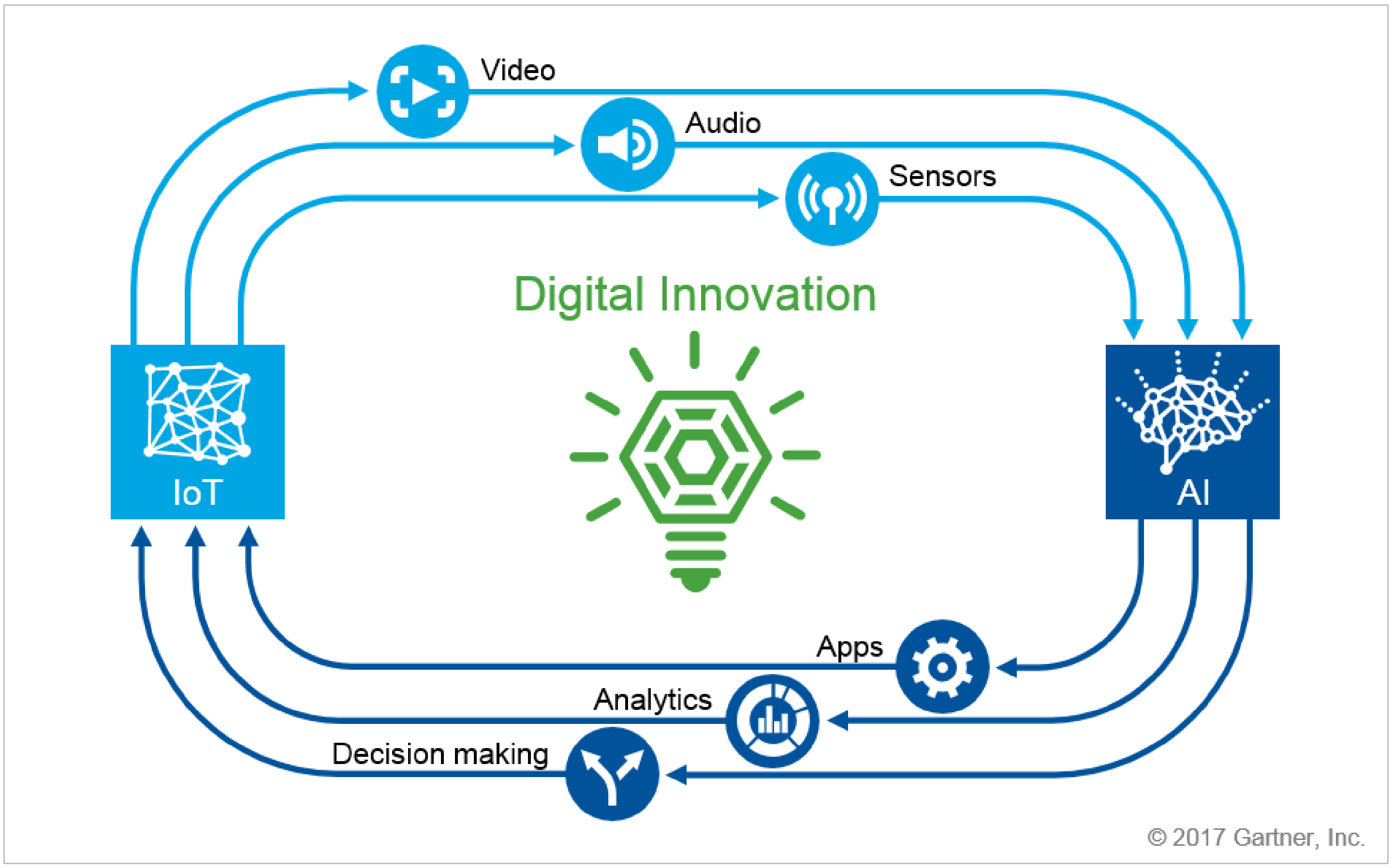
For "AI," which is a term that encompasses a range of fast-emerging technologies, the breakthroughs often are highly dramatic. In March 2016, AlphaGo, a computer program created by Alphabet's Google DeepMind AI project, beat the world's second-ranked Go champion in a five-game match. Unlike hand-coded strategies used unsuccessfully by other systems, AlphaGo was trained with data describing hundreds of thousands of online Go games played between humans, and it used the sequences of moves as data for a machine-learning algorithm. For this achievement, Science picked AlphaGo as one of its "Breakthrough of the Year" runners-up for 2016 (http://www.sciencemag.org/news/2016/12/ai-protein-folding-our-breakthrough-runners) .

These examples show the algorithmic and scalability impact of the two technologies, and both are attracting massive public- and private-sector investments. Today, AI models are largely trained in vast data centers, powered by thousands of specialized GPUs. But the need to transfer the data generated by the vastly increasing number of endpoints to these data centers, or to provide the decision making from the data centers to the endpoints, will not easily scale. Therefore, AI must branch out to the IoT endpoints: that is, "AI on the edge." When the two technologies are brought together, each supplies something that the other is lacking. Together as a new business platform, their impact will be multiplied.

How IoT and AI Create a Virtuous Cycle

IoT's missing puzzle piece is a "killer app" — some application, capability or feature that is considered to be indispensable or manifestly superior to alternative products. Conversely, AI's missing puzzle piece is data — the vast volumes of labeled data needed to train the ML algorithms that enable AI systems to "learn on their own" without explicit programming (see Figure 2).

Figure 2. IoT and AI Create a Virtuous Cycle for Digital Innovation



Source: Gartner (June 2017)

Despite the hype, most of today's IoT benefits are incremental, whether in consumer or commercial applications. Many commercial applications, such as building management systems in enterprises or predictive maintenance in manufacturing, are seen as desirable, even important, but not yet essential for most organizations. Most consumer applications' benefits, such as smart homes and wearables, are even more doubtful, as shown by the slow adoption of these device categories. But what the proliferation of IoT endpoints is creating is an abundance of data. IoT generates such large volumes of both structured and unstructured data that many organizations are in a quandary about how to make sense of it and unlock its value. As with big data years ago, organizations have often put technology selection ahead of business problem identification.

AI's biggest advance in the last few years has been in the area of ML, made possible by recent breakthroughs in readily available and affordable computational power and in vast datasets for certain applications, such as image classification. This is especially true for a branch of ML called deep learning. Deep learning relies on algorithms and models that can be trained to discover, on their own, novel representations from data in order to apply classification and prediction to an array of business problems. The most common type of deep learning today is deep neural networks (DNNs), which use numerous layers (hence, "deep") of interconnected nodes to sift vast amounts of data. The algorithms "learn" from their experience with the data and become more capable over time. For more information, see "Innovation Insight for Deep Learning." (https://www.gartner.com/document/code/319191?ref=grbody&refval=3738018)

The first widely accepted demonstration that deep learning significantly surpassed all earlier feature engineering approaches to image classification occurred in the ImageNet contest of 2012. This "big-bang" effect, a unique combination of three elements — deep neural nets (algorithms behind deep learning), big data and massive amounts of parallel computing hardware — enabled this big-bang effect (see "Smart Machines See Major Breakthroughs After Decades of Failure" (https://www.gartner.com/document/code/291251?ref=grbody&refval=3738018)). The same approach has been successfully applied in many areas, including image captioning, facial recognition, language translation and speech to text. The accelerating improvements in VPAs, such as Apple's Siri and Google's Assistant, are due to deep learning.

As a result, the AI applications that have advanced most quickly benefited from visual and audio data, both of which are core outputs for many IoT endpoints, such as surveillance cameras and drones. The flood of data made possible by IoT endpoints will feed AI-based applications that learn from the data, uncover new behavior patterns or relationships, and recommend and take action. This virtuous cycle between IoT and AI at first is having the greatest impact on applications that exploit image and audio recognition, but deep learning will eventually spread to other applications as well.

Combining AI and IoT creates a platform that moves beyond simple IFTTT automation use cases or rule-based expert systems that struggle with complex, dynamic problems. The emerging IoT-AI platform will enable applications that can support much more complex data analysis and decision making.

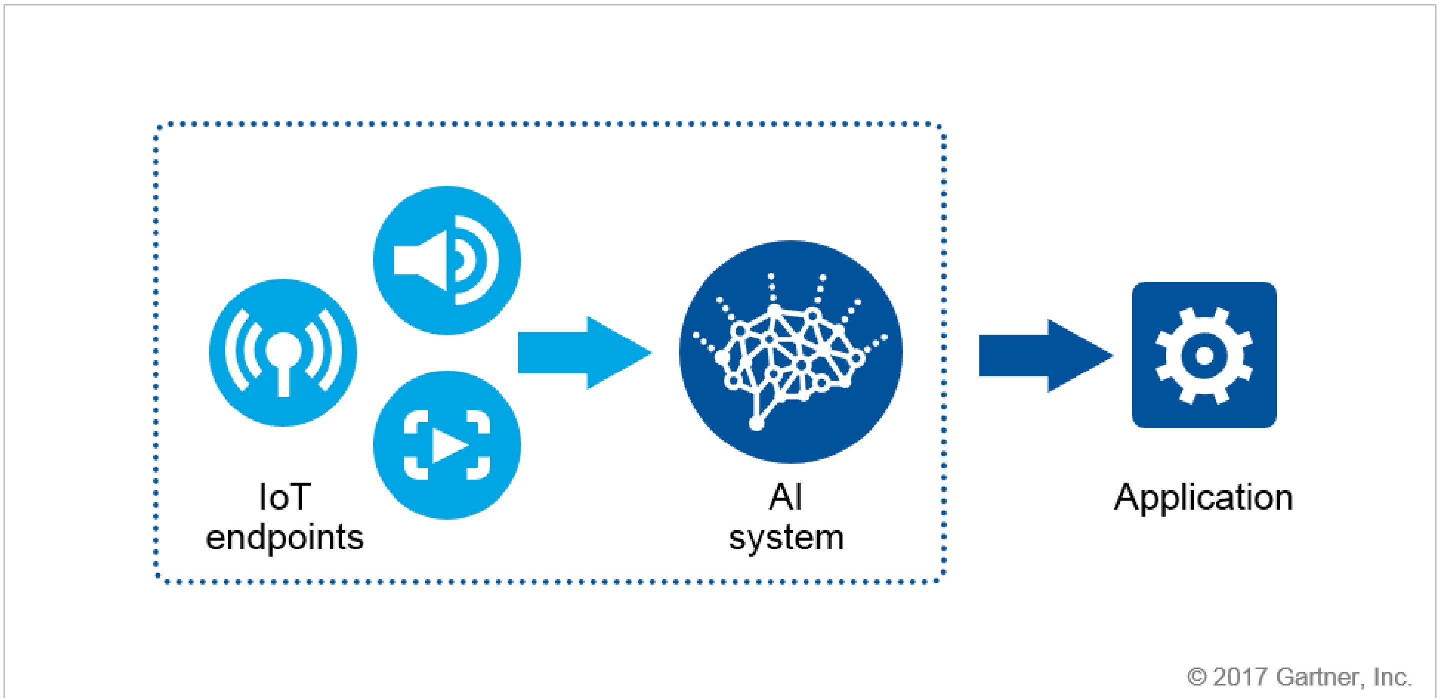
What the Emerging IoT-AI Platform Will Look Like

As complementary technologies, IoT and AI can work together in different ways. In some cases, IoT will be the "dominant" element in the platform, with AI adding supplemental capabilities. In other cases, AI will be the dominant element, "fed" by IoT data. Finally, the two technologies can work together symbiotically, each leveraging the other in a mutually beneficial relationship. The result is three architectural styles for the emerging IoT-AI platform. The use case or end application determines which architecture is the best fit. One implication of this is that an organization might make use of any or all of these architectures. There will be many use cases and applications that blur these architectural boundaries, but these three serve as a starting point for CIOs in conceptualizing and developing the IoT-AI platform.

Architecture 1: IoT Data as the Input to the AI System

In this architecture, the IoT system is a peripheral to the AI system. IoT acts as a data gatherer that then feeds this data to the AI system, as shown in Figure 3.

Figure 3. IoT as the Input to the AI System



Source: Gartner (June 2017)

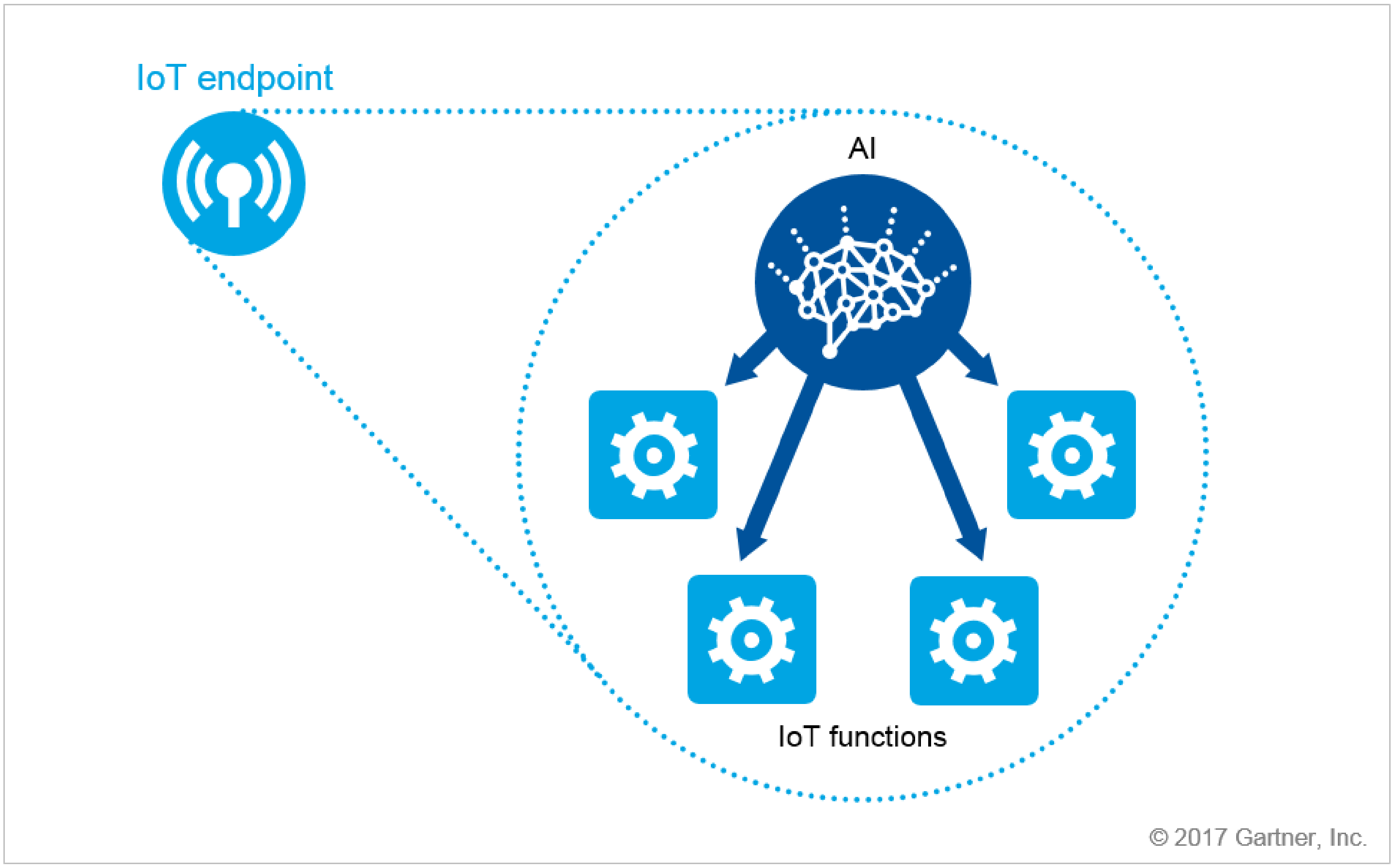
One key function in this relationship is using the IoT dataset to train the ML model used in the AI system. The training process uses data to enable the model to learn what to look for, such as predicting the demographics of a retail customer via the in-store video camera (an emerging IoT-AI application area called "intelligent video analytics [IVA]"). Ongoing training makes the model more accurate, more "intelligent."

An example of a use case for this architecture is improving construction industry safety levels. Safety is a leading concern for the industry. But construction firms rely generally on human safety experts for safety inspections and prevention. These experts typically go through five to seven years of training. An IoT approach will use remote cameras mounted throughout the construction site or on airborne drones, as well as other relevant sensor data. This image data feeds an AI system trained to analyze images to identify unsafe or potentially unsafe activities or situations. One vendor in this space is Smartvid.io, which uses ML for industrial media management, collaboration and analytics: Construction site safety is one of the vendor's target use cases.

Architecture 2: AI as the Application in the IoT System

In this architecture, AI is one of (many) applications in the IoT system, as shown in Figure 4. Specifically, the AI capability serves as the IoT system's inference engine, interpreting the data generated by the IoT endpoints and driving some of the endpoints' functions.

Figure 4. AI as the Application in the IoT System



Source: Gartner (June 2017)

This architecture is suited to use cases that have stringent requirements or constraints, such as:

- Limited network availability and/or bandwidth
- Strict (low) latency requirements
- Data privacy restrictions

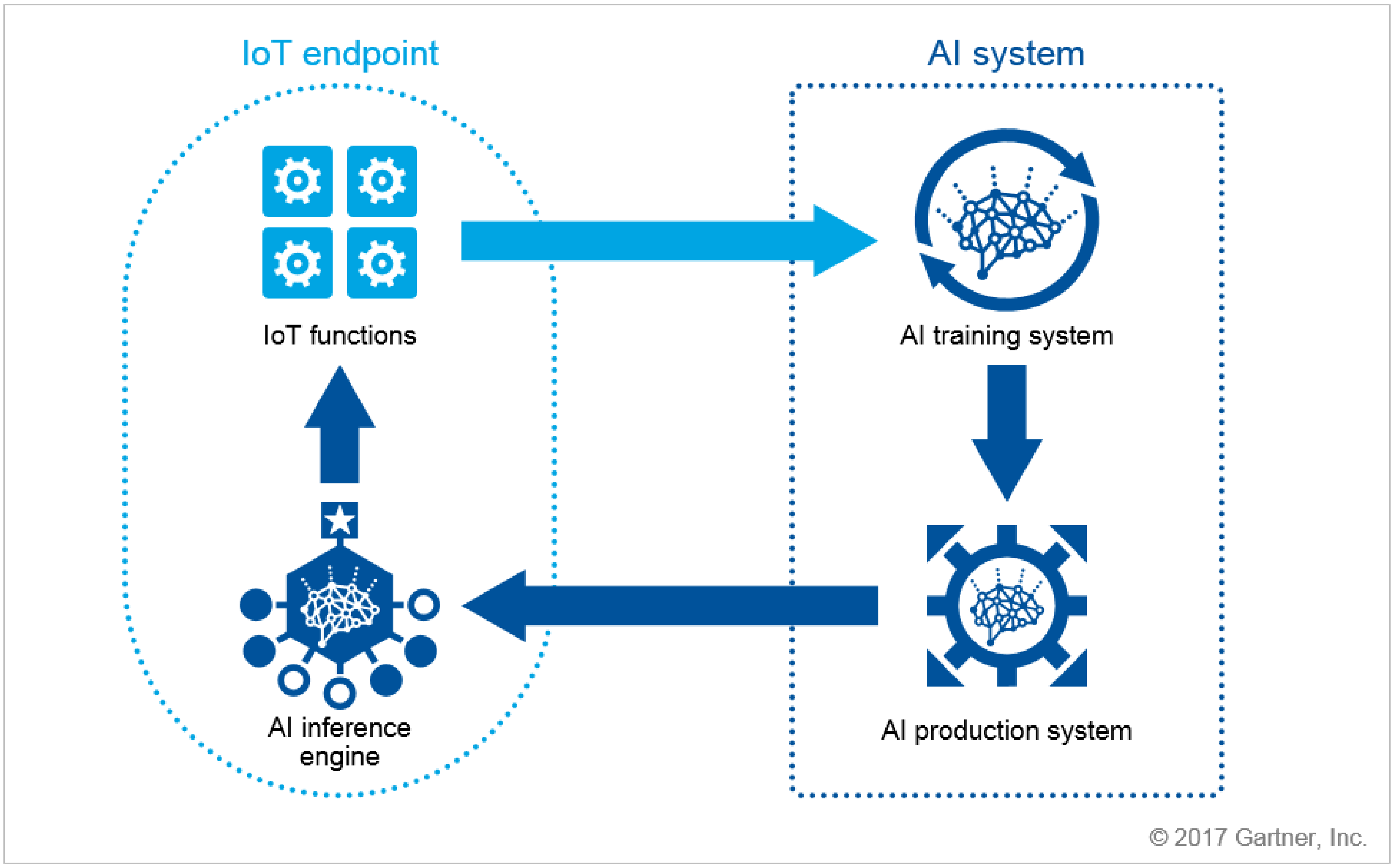
An example of this architecture can be found in medical wearables that leverage sensor data and AI to help visually impaired people navigate the world in their daily lives. Horus is one vendor in this space. The Horus device can "see" the world for its wearer via its headset/camera, which links with a pocket unit that has the AI inference engine integrated. The product can handle text, facial and object recognition, alert the user to obstacles in their path, and to caption images — in essence, this wearable device converts visual data into audible information for its user, empowering the visually impaired.

Computational resources can be a challenge, depending on the type of IoT endpoint and its application. For example, an autonomous vehicle has plenty of computing power, while an agricultural moisture sensor has very little. However, the technology has advanced to the point in which a full image classification model can now fit on a low-end IoT design, such as the Raspberry Pi.

Architecture 3: IoT and AI as a Two-Way System

In this architecture, both the IoT and the AI systems interact with each other to each other's benefit, as shown in Figure 5. The IoT system provides data to the AI system on an ongoing basis. This data is then used to train the AI system on a periodic basis. As the AI system improves over time, eventually a new production system is created and deployed to the IoT system, which in turn improves the IoT system's decision making (inferencing).

Figure 5. A Bidirectional IoT-AI System



Source: Gartner (June 2017)

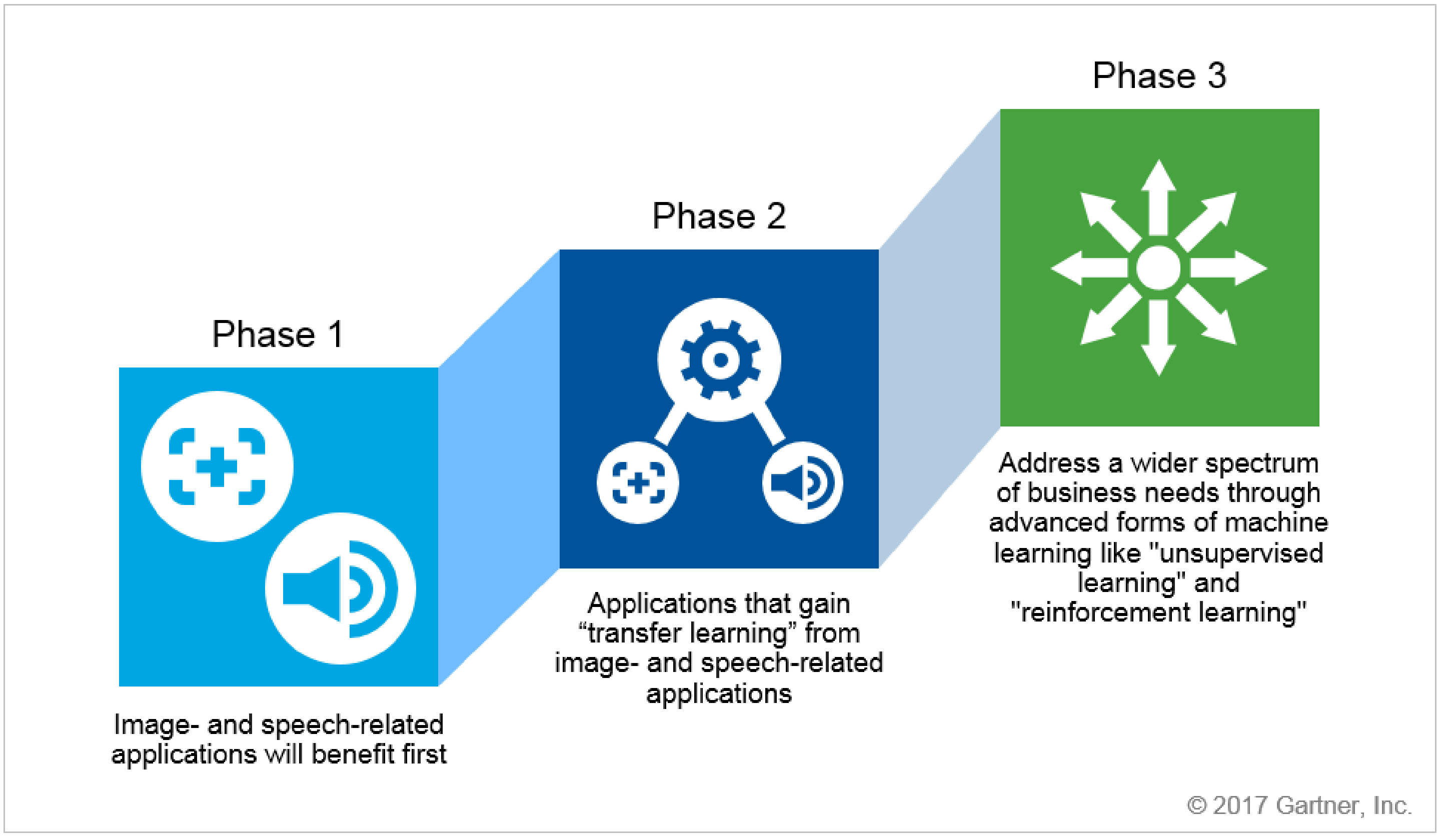
One key use case for this approach is autonomous vehicles. The physical world in which autonomous vehicles navigate is always changing, due to variables such as road construction or repair, vehicle accidents and pedestrian flows. Therefore, autonomous vehicles must be continually updated about new environmental characteristics and new behaviors to respond appropriately to them.

Generally, autonomous vehicles leverage the various onboard sensors, such as cameras, radars and lidar, which is a remote sensing method that uses a pulsed laser to measure variable distances. The sensor data feeds into the AI system's training model. In turn, the training system periodically creates new production models that are deployed to the vehicles.

How AI Will Be Adopted for IoT Projects

Gartner expects the IoT-AI platform to go through three phases of adoption, distinguished by the maturity of the respective AI technology (see Figure 6). CIOs should be aware that these phases may be implemented by different entities: the internal IT organization, business stakeholders, application vendors, system integrators and service providers. The implementation decision will rest on a combination of business needs and the organization's technical competence.

Figure 6. The Three Phases of Adoption of AI for IoT Projects



Source: Gartner (June 2017)

Phase 1

Image- and speech-related applications will benefit first in leveraging AI capabilities, by exploiting the recent advances in deep learning. This is because AI has had its greatest advances so far in these particular fields. In addition to the autonomous vehicle example mentioned earlier, myriad use cases have been explored in such verticals as manufacturing, energy exploration and smart city applications.

Most of these deep learning applications will rely on "supervised learning," wherein tags or labels are ascribed typically to various sections of each image by people before the images are fed to the DNN model. Once the deep learning process is finished, the model will be tested with a new dataset to determine if the training worked properly. If not, both the dataset and the DNN model will be re-evaluated and the process begins anew. If testing is successful, the trained model is ready for production use (often referred to as "inferencing").

Phase 2

Applications that can take advantage of"transfer learning" from image- and speech-related applications will be the next ones to see the greatest commercial adoption. While supervised learning has grown by leaps and bounds with the increases in computational power and labeled datasets, it usually only works well when the training and test data are drawn from a domain with similar characteristics. For example, if an AI system is trained to recognize daytime pictures of cats, it may not perform well on nighttime images of cats. This is where transfer learning comes in. In cases in which there may be an insufficient pool of labeled training data, transfer learning may be applied to "transfer" the knowledge gained from one system to another in a new domain.

One example is image captioning, which combines image classification and NLU and applies them in a completely new domain: The deep learning system's models not only recognize the image, but also create a descriptive sentence (caption) for that image. For example, in a picture with a shark swimming in the ocean, instead of just outputting "shark" and "ocean," the system will caption the picture as "a shark is swimming in the ocean." Captioning systems such as this can greatly aid security monitoring applications in retail and smart city applications, where this type of automation can help the human monitors from being overwhelmed by the sheer number of IoT endpoints.

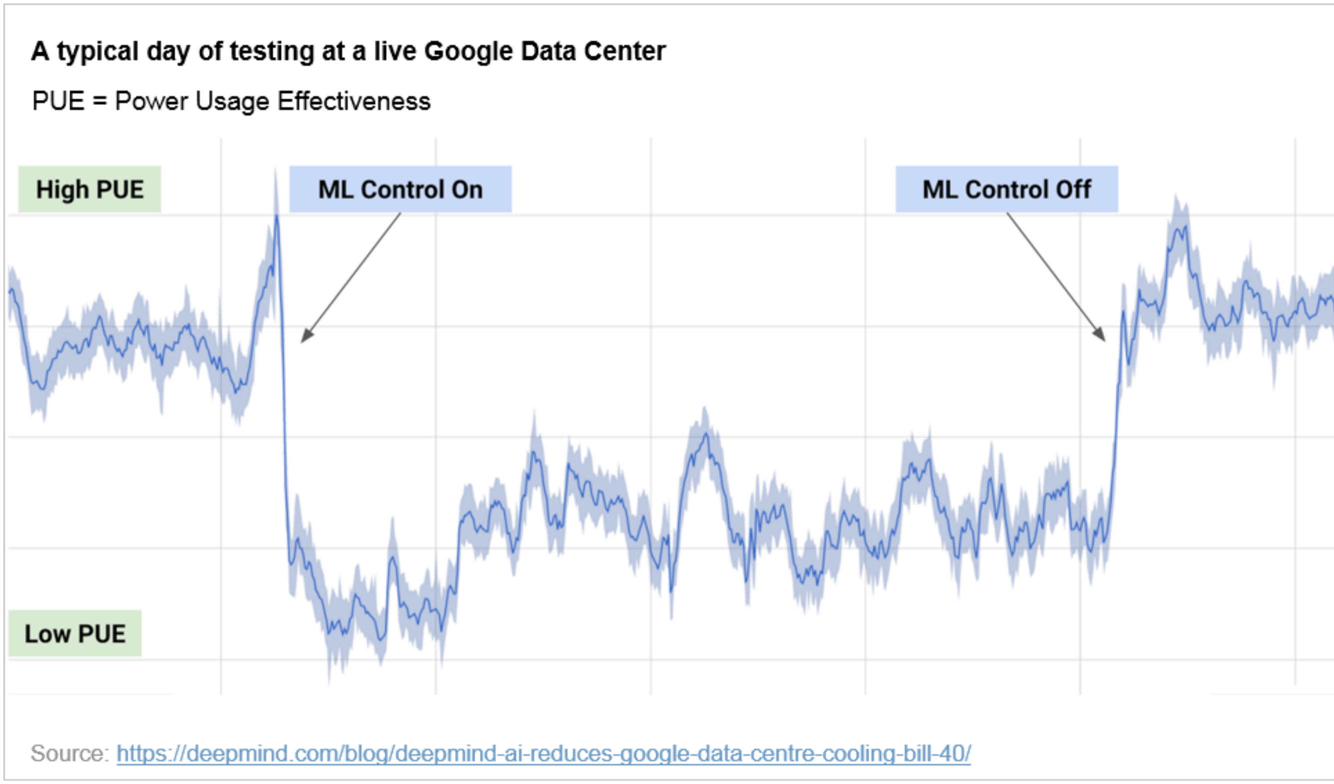
The bleeding edge of transfer learning is in semisupervised learning, where companies have seen success with text-based applications. In semisupervised learning, initial training of the DNN is done using labeled datasets (that is, traditional supervised learning). After the initial training has been completed, the model is then greatly extended with similar unlabeled data (that is, unsupervised learning). Companies like Google and Microsoft have used this method to build "knowledge graphs" encompassing huge swaths of the web. But those systems may not be appropriate for more targeted enterprise use cases. The size of the training datasets matter (the larger the better), and the engineering is inexact for now.

Phase 3

Finally, as more advanced forms of ML improve, they hold the promise for organizations to address a wider spectrum of business needs. Increased interest in GANs (which are a form of unsupervised learning) and reinforcement learning (used in Google's AlphaGo) has led to increased investments in both by the top technology companies.

One of the key findings from Gartner's 2016 IoT Backbone Survey was that internally focused IoT projects, particularly those that address operational efficiencies, topped the list of priorities for organizations. An example of reinforcement learning that has been successfully applied for such an IoT project is within Google's data centers. Leveraging the "thousands of sensors within the data center — data such as temperatures, power, pump speeds, setpoints, etc.," (https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/) Google was able to reduce its cooling bill by 40%, and the company now gets approximately 3.5 times the computing power out of the same amount of power consumption. As shown in Figure 7, the data center's PUE (defined as the ratio of total building energy usage to the IT energy use — the lower the better) decreased when ML is turned on.

Figure 7. Machine Learning's Impact on Energy Usage in a Google Data Center

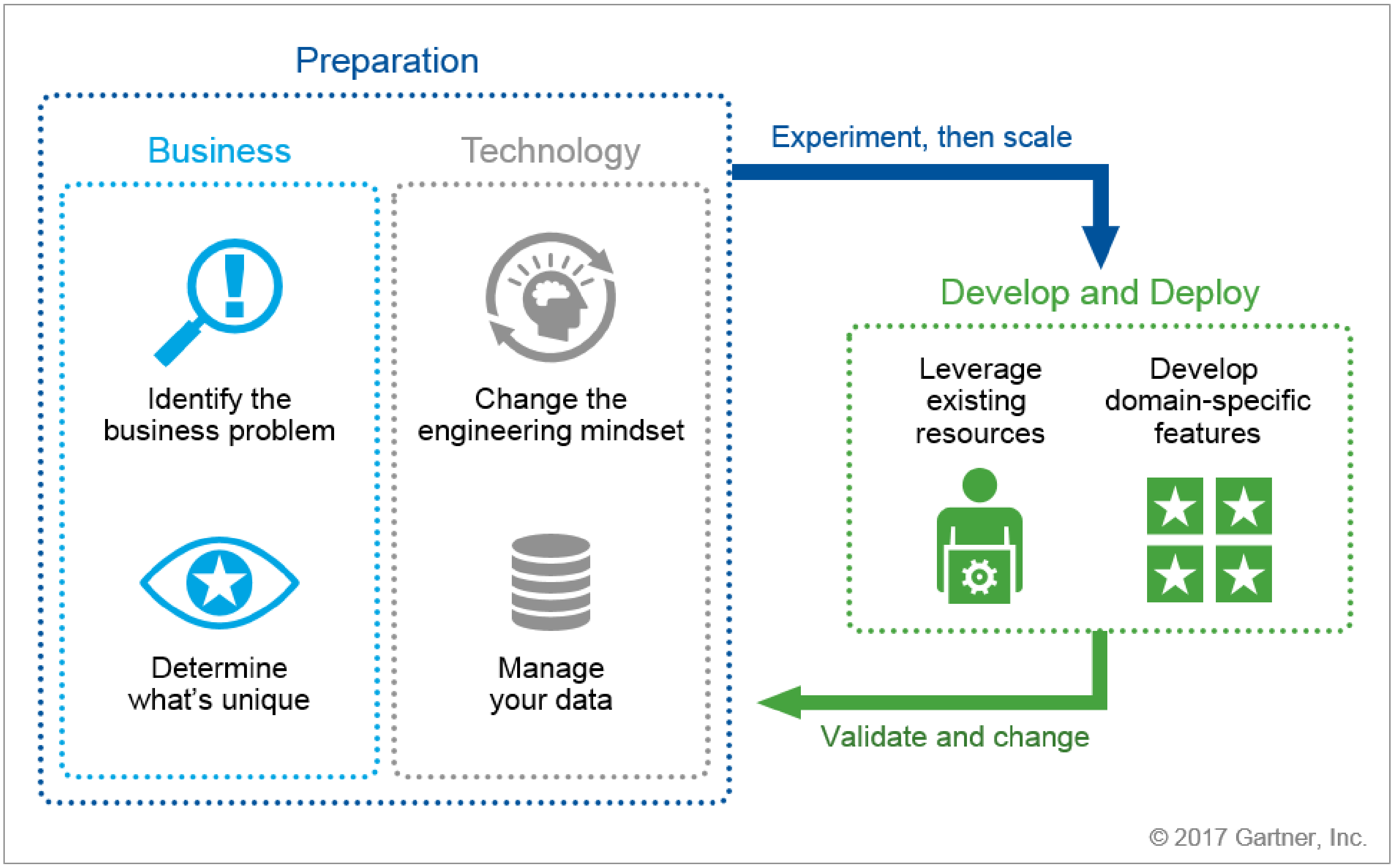


Source: Adapted from Google/DeepMind

How to Incubate the IoT-AI Platform

The IoT-AI platform is still in an embryonic stage in most organizations and industry verticals. More importantly, there is no single recipe for creating it. Instead of seeking a step-by-step process, forward-looking CIOs should incubate the platform: Create a sandbox, whereby the necessary education, evaluation, partnership and experimentation can mature until the platform hatches (see Figure 8).

Figure 8. How to Start With IoT and AI



Source: Gartner (June 2017)

The incubator should cover three areas.

Address the Business Issues

Identify the Business Problem

Too often, CIOs tend to focus on technology first, with the goal of picking the "right" platform, applications and tools. But with IoT and AI still rapidly evolving, it is especially important first to define the business problem. As previously noted, recent advances in AI and IoT are best-suited for particular use cases and data types. In the near term, these highlight the most promising areas for the IoT-AI platform. CIOs should engage with executive leadership and business unit leadership to identify those business problems and opportunities where this emerging platform can have the most impact.

Determine What's Unique

A key part of this work will be specifying that which differentiates the organization (or industry segment) from competitors. In other words, CIOs should figure out what needs to be developed internally to differentiate themselves, and the remainder can use existing third-party solutions. For example, a visual inspection system for a manufacturing line should not reinvent the wheel and develop an image classification engine from scratch. Instead, the organization should leverage existing classification engines and build its organizational know-how on top of them via algorithmic enhancements and/or training using proprietary data.

Establish a Strong Technology Foundation

Change Your Engineering Mindset

CIOs should model, encourage and incentivize a new way of thinking about IoT and AI as complementary technologies with far-reaching implications for the organization. IoT involves physical things instead of just applications that can be virtually instantiated and killed. How these things get installed, onboarded, connected, managed and secured requires a new skill set and, more importantly, a new mindset (see "IoT's Challenges and Opportunities in 2017: A Gartner Trend Insight Report" (https://www.gartner.com/document/code/324746?ref=grbody&refval=3738018)).

ML is unlike previous software paradigms in three areas:

- It's more like research than engineering. There is no clear set of specifications defining how to do it. There is no equivalent to Donald Knuth's "The Art of Computer Programming" on specific algorithms and trade-offs. There are only general guidelines to experiment with. AI projects require lots of experimentation or staff with extensive experience in building the specific type of DNN you need.
- Everyone has to understand the stochastic nature of DNNs. Extensive test data is needed, and humans need to monitor results very closely. A one-shot test dataset or standard QA process isn't enough.
- DNNs are programmed primarily with data. They are programmed in the architecture of the DNN model and in the collection, curation and tagging of training and testing data.

Manage Your Data

Data management strategy can be a lower priority during the proof of concept or trial stage, when the organization is still exploring where the IoT-AI platform might be applicable and which problems to address. But even in these early stages, the organization should take note of what data is being used and how. The IoT-AI platform project should include a strategy for organizing existing data, ingesting new data and combining the two (see "Prepare to Monetize Data From the Internet of Things" (https://www.gartner.com/document/code/309409?ref=grbody&refval=3738018)).

As mentioned, to properly train DNNs, organizations will need labeled data in target domains in sufficient volume. Some IoT endpoints, such as autonomous vehicles, are expected to generate an enormous amount of data (as much as 4TB per day), much of which has little value. Finding, using and maintaining high quality data for this purpose should be a priority in IoT-AI projects.

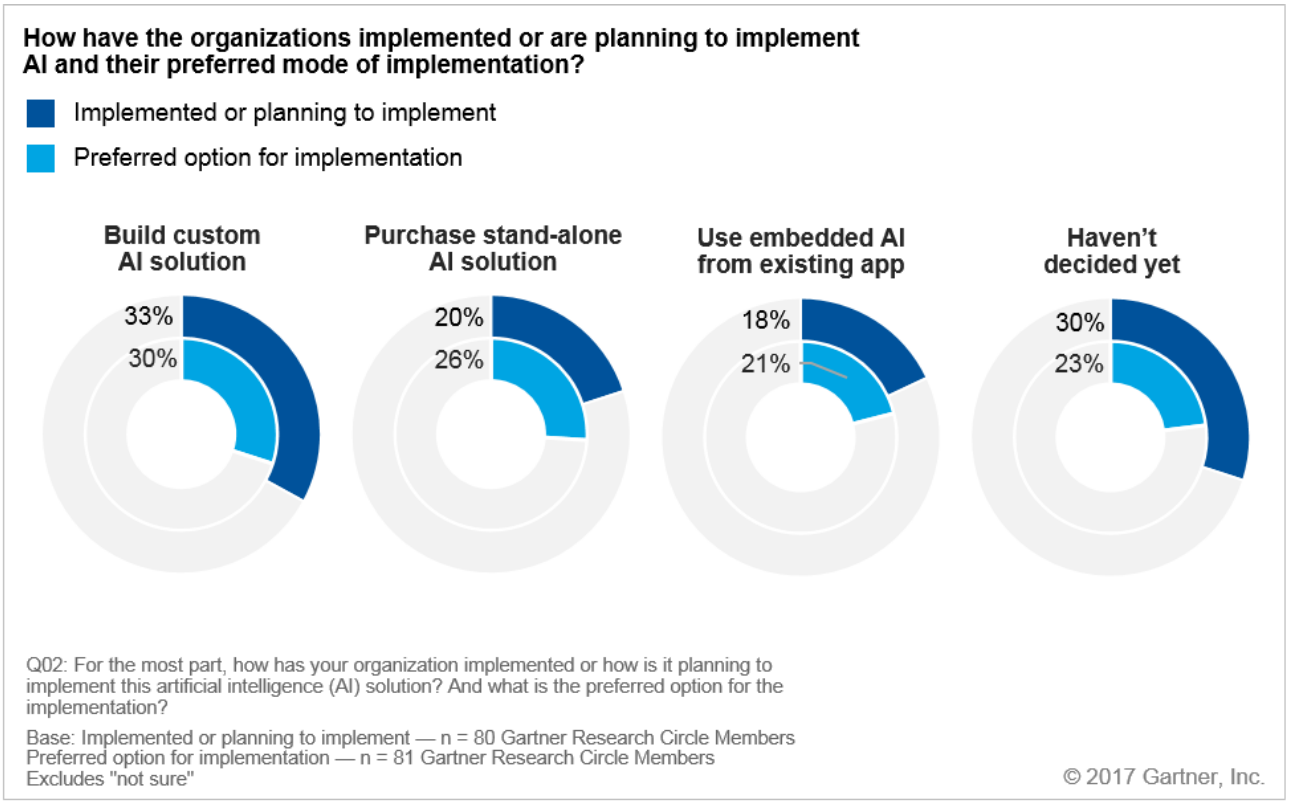
Develop and Deploy

Leverage External Expertise and Capabilities

Commoditized and standardized competencies and capabilities can be leveraged rather than reinvented. For example, there's a plethora of open-source AI frameworks, and the top CSPs all have ML libraries for use in a number of domains. In such cases, organizations should choose the best-fit technologies or product/service offerings (which may not be the largest or most popular) and not waste resources on creating them organically. For example, a conversational platform targeted at CRM applications should build industry-specific dictionaries on top of an existing NLP engine rather than create one from scratch.

Gartner's recent Research Circle survey indicates that there's not a clear winner in terms of how organizations are approaching their AI development (see Figure 9).

Figure 9. Organizations' Considerations for AI Implementation



Source: Gartner (June 2017)

Experiment, Validate and Scale

Starting small has three advantages:

- Developers unfamiliar with ML can gain experience and learning.
- Organizations can experiment with AI in a range of different IoT problems.
- Business value of using AI can be proven to get investment to operationalize.

"Small" applies to both the initial problem and the initial investment. Look for opportunities where the required compute resources are minimal or can be cost-effectively leveraged via a CSP. Consider using off-the-shelf AI algorithms and models, rather than investing in specialized, in-house development of such algorithms and models. Amazon Web Services, Google, Microsoft and IBM are just four providers offering access to such cloud-based AI resources.

Acronym Key and Glossary Terms

AI	artificial intelligence

CAGR	compound annual growth rate
CQGR	compound quarterly growth rate
CSP	cloud service provider
DNN	deep neural network
GAN	generative adversarial networks
IFTTT	if this then that
IoT	Internet of Things
ML	machine learning
NLP	natural-language processing
NLU	natural-language understanding
PUE	power usage effectiveness
QA	quality assurance
VPA	virtual personal assistant

Evidence

Methodology for analysis of social media conversations: We used automated social media listening tools to track users' responses on social media and public discussion forums. The period for the analysis of overall mention count was from 1 January 2015 through 31 March 2017. "Social media mentions" denote the inclusion of a monitored keyword in a textual post on a social media platform. High counts of mentions should not be considered an indication of positive sentiment by default. Social media sources considered for this analysis included Twitter, Facebook (publicly available information only), images (comments only), aggregator websites, blogs, news, mainstream media, forums and videos (comments only). All regions and major world languages were covered for the study. Social media analytics research and contributions by Anjali Grover, Ayush Saxena and Sindhu Jayakumar.

Recommended by the Authors

Forecast Analysis: Internet of Things — Endpoints, Worldwide, 2016 Update (https://www.gartner.com/document/code/302435?ref=ggrec&refval=3738018)

Innovation Insight for Deep Learning (https://www.gartner.com/document/code/319191?ref=ggrec&refval=3738018)

IoT's Challenges and Opportunities in 2017: A Gartner Trend Insight Report (https://www.gartner.com/document/code/324746?ref=ggrec&refval=3738018)

Survey Analysis: 2016 Internet of Things Backbone Survey (https://www.gartner.com/document/code/317142?ref=ggrec&refval=3738018)

Prepare to Monetize Data From the Internet of Things (https://www.gartner.com/document/code/309409?ref=ggrec&refval=3738018)

Develop Your Artificial Intelligence Strategy Expecting These Three Trends to Shape Its Future (https://www.gartner.com/document/code/324590?ref=ggrec&refval=3738018)

Recommended For You

Maverick* Research: What Does Good Artificial Intelligence Look Like? Build an {a}IQ (https://www.gartner.com/document/3819170?ref=ddrec&refval=3738018)

Maverick* Research: The Emergent Machine Society — Where Machines Meet, Talk, Scheme, Fight and Marry (https://www.gartner.com/document/3826267?ref=ddrec&refval=3738018)

Develop Your Artificial Intelligence Strategy Expecting These Three Trends to Shape Its Future (https://www.gartner.com/document/3686117?ref=ddrec&refval=3738018)

Deliver Artificial Intelligence Business Value: A Gartner Trend Insight Report (https://www.gartner.com/document/3872663?ref=ddrec&refval=3738018)

Where You Should Use Artificial Intelligence — and Why (https://www.gartner.com/document/3754164?ref=ddrec&refval=3738018)