Uptime Institute® INTELLIGENCE

EMERGING TECHNOLOGIES

UI Intelligence report 25

Very smart data centers

How artificial intelligence will power operational decisions

Author
Rhonda Ascierto, Vice President of Research, Uptime Institute

The complexity and scale of modern data centers, coupled with the rapid speed of change in IT environments, is becoming too great for humans to effectively manage. For computers and artificial intelligence, however, it can be relatively simple. A key goal is to predict and prevent incidents and to detect and remediate inefficiencies and capacity shortfalls. Given that many data centers are suboptimally utilized, many operators are likely to benefit from the technology. This report provides an overview of data center artificial intelligence technology and its requirements, use cases, risks and costs.

30-45 minutes to read

UII-25 v1.2 published March 26, 2019 last updated September 3, 2019

This Uptime Institute Intelligence report covers:

Uptime Institute Intelligence and methodology

Uptime Institute Intelligence is an independent unit of Uptime Institute dedicated to identifying, analyzing, and explaining the trends, technologies, operational practices, and changing business models of the mission-critical infrastructure industry. For more about Uptime Institute Intelligence and this report, see the **Appendices** or visit https://uptimeinstitute.com/ui-research.

KEY FINDINGS

• Data center artificial intelligence (AI) is in the early stages of development and adoption. The overall market is currently confused and confusing, but it is clear that most use cases today and in the near future will support narrow areas of operational decision making.

• AI use cases in data centers are typically improvements over existing functions and processes. AI promises better and more accurate analysis. Forecasts may be more precise or project further into the future; recommendations may be more accurate, timely and detailed; and real-time alerts may be activated much earlier.

• This is an area of technology that is ripe with hype and misinformation. Operators should invest with care.

• Human data center domain expertise is as critical for AI as real-time data. But the proliferation of AI-driven decision making in data centers may, over time, in many situations, replace the need for humans who have hands-on domain expertise.

• While there is much discussion about who owns the data with AI-driven cloud services (the data center or the supplier?), there are greater risks that receive less attention, such as the risk of vendor lock-in and other legal and technical complications associated with AI.

• AI-driven data center management as a service (DMaaS) and other AI cloud services have low barriers for adoption (with little or no historic data required), involve limited or no data center automation (and, therefore, low risk) and are priced inexpensively. Adoption is therefore likely to be strong.

• A data center infrastructure management (DCIM) system is not required for AI-driven DMaaS or related cloud services. However, on-premises DCIM software remains a requirement for data centers seeking real-time mission-critical alarming and closed-loop automation.

• Providers of AI services do not always provide transparency into AI-driven decisions nor accurate barometers for success of the technology. The onus is often on the data center customer to ask for greater transparency and for accuracy rates, and to otherwise validate results, which is particularly important for areas of potential automation.

• Many data center AI use cases require data from beyond the physical data center. External service providers are likely to be the best equipped (including with up-to-date application programming interfaces [APIs]) to aggregate and converge different types of data, from different places. Because of this, we believe a fundamental change is underway concerning the management and use of critical operational data, with greater use of external services.

Introduction

AI-driven software, mostly notably from the specialist supplier Vigilent, has been dynamically adjusting cooling equipment and capacity in data centers for the past decade. In the past couple of years, however, a wave of new use cases, as well as products and specialist suppliers, has hit the market. One high-profile example was Google, in 2016, announcing it had reduced its power usage for data center cooling (in a limited trial) by up to 40% by applying recommendations from its AI platform DeepMind. In 2018, it began using machine learning algorithms to automatically adjust its cooling plant settings continuously and in real time. Meanwhile, AI is gaining traction in related IT fields, including supplier Splunk, using analytics and AI to improve network and IT operations. There are numerous other examples.

Implementing basic AI capabilities in the data center can be done using streamlined, cloud-delivered service. Other new tools mean that the development of sophisticated AI is now within reach of almost any data center.

AI promises to improve data center efficiency, reduce facility risk and transform operational processes ranging from anomaly detection and equipment maintenance to white-space optimization and dynamic workload management. The technology can be applied to almost every facet of data center management. Several suppliers are investigating ways to simulate and optimize designs of data center not yet built. The potential of AI to transform data centers is significant.

This potential has excited many marketers yet has also stoked fear in operators and managers wary of the technology. The power of AI to improve management and operations can be viewed as a powerful risk: How can you trust machines to make decisions that cannot be easily understood?

It seems clear, however, that the recent growth in new AI products and services for the data center will continue, driven by:

- Growing demand from operators, especially larger ones, for systems that can help manage complexity and identify patterns that humans are unlikely to see.

• The availability of inexpensive computing and storage from public cloud providers.

• A transfer of AI tools and knowledge to the market from academia and large operators such as Google, IBM and others. Investor interest in the AI field following well-publicized advancements, such as artificial neural networks" enabling of natural language processing in consumer goods.

In data centers (and elsewhere), AI is still at the early stages of development and adoption. However, most in the industry expect that AI will be widely used in data centers in the future, according to our most recent survey of more than 250 data center designers, consultants and vendors (see Figure 1).

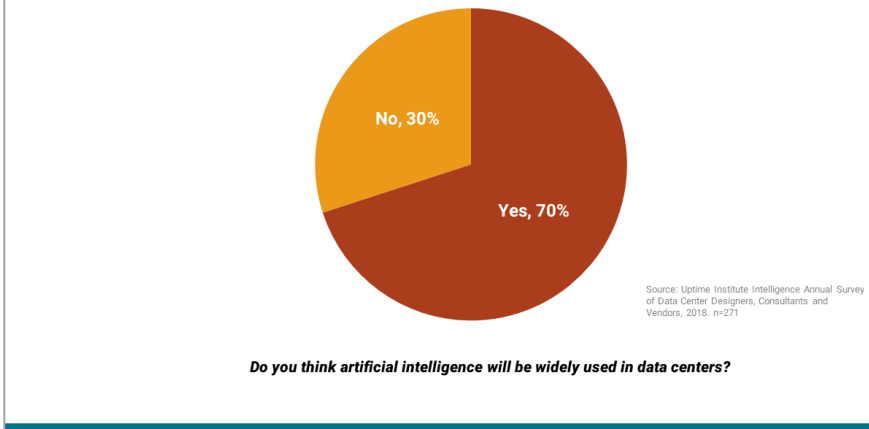

Figure 1. **Most think artificial intelligence will be widely used in data centers**

Despite some impressive deployments, AI is still far from being capable of overtaking our collective human knowledge. Some believe AI will reach the singularity — the point at which computers will be more powerful than all human intelligence combined — as soon as 2045 (for a long time, the year of the singularity had been 2029). However, we and many others believe this inflexion point is significantly further out.

Initial deployments of AI have not been revolutionary or disruptive in their impact; AI is currently employed in narrow use cases to spot patterns. In data centers, as in other industrial environments, the technology is used to improve existing processes rather than create radical new approaches. It is guiding operators and managers to make decisions. A reinvention of designs, operations and management promises to be a long-term future inevitably, but changes will be incremental over a long timeframe.

As is typical with any "new" technology, not everyone believes AI as it stands today is worth the investment. Some, including operators seeking to develop AI for their data centers, say there is a diminishing return in using expensive AI for improvements when more straightforward, cheaper big-data analytics are available. Others have found it difficult to source skilled staff and are wary of introducing a key technology without a full understanding of what is being deployed.

For some operators, however, new AI-driven services delivered via the cloud and paid for as an operating expense are a low-cost, easy way to explore the technology.

Understanding AI

Myths about AI

Marketing hype and misinformation is being fueled by a combination of AI's dazzling complexity, which only specialists can deeply understand, and by its novelty in most data centers. It is a confused and confusing area. For example:

Myth #1: There is a best type of AI for data centers. The best type of AI will depend on the specific task at hand. Simpler big-data approaches can be more suitable than AI. For this reason, new "AI-driven" products such as DMaaS often use a mix of AI and non-AI techniques (see **Adopting data center machine learning**.)

Myth #2: AI replaces the need for human knowledge. Domain expertise is critical to the usefulness of any big-data approach, including AI. Human data center knowledge is needed to train AI to make reasonable decisions/recommendations and, especially in the early stages of a deployment, to ensure that any AI outcome is appropriate for a particular data center.

Myth #3: Data centers need a lot of data to implement AI. While this is true for those developing AI, it is not the case for those looking to buy the technology. DMaaS and some DCIM use prebuilt AI models that can provide limited but potentially useful insights within days.

Basic terminology

The terminology of AI can be confusing. Below is a high-level explanation of the main types of AI being used in data centers (and elsewhere).

Big data explained

Big data is a broad term for ways to analyze and systematically extract information from data sets that are too large or complex to be handled with traditional data-processing software. All AI is big-data analytics, but not all big-data analytics is AI. Some big-data analytics use AI, some do not. Again, one is not necessarily better than the other.

In data centers, examples of non-AI big-data analytic applications or functions include:

- Data visualization (charts, graphs, other visuals).
- Data correlation/trend analysis.
- Anomaly detection.
- Complex “what-if” scenarios (modeling the outcome of changes).

As an example, certain suppliers have amassed large data sets from many different data centers about different uninterruptible power supply (UPS) configurations and performance. By applying big-data statistical comparison and correlation, they can determine best practices for UPS configuration that, for example, prolongs the equipment’s lifetime. The software used does not necessarily require anything that might be classed as AI.

Artificial intelligence, machine learning and deep learning defined

Artificial intelligence, a term that flourished in the 1950s, is an umbrella term that describes software running on machines that is capable of perception and learning in similar ways to humans.

AI is often used interchangeably with machine learning and deep learning, but these are subsets of AI (see Figure 2). Machine learning has been around since the 1980s, while deep learning, a newer development, is a subset of machine learning. (For more detailed information on each, including branches of machine learning in use in data center today, see **Appendix 2. AI Techniques**.)

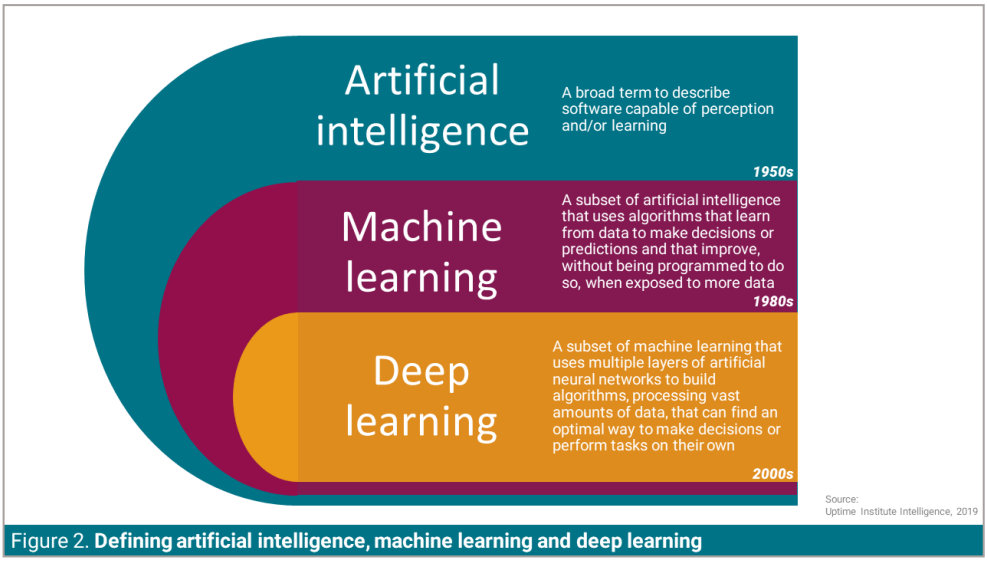

Figure 2. **Defining artificial intelligence, machine learning and deep learning**

Machine learning uses algorithms that learn from data to make decisions or predictions and improves, without being programmed to do so, when exposed to more data.

Deep learning is a type of machine learning that uses multiple layers of artificial neural networks to build algorithms, processing vast amounts of data, that can find an optimal way to make decisions or perform tasks on their own.

Most AI in data centers today is machine learning rather its subset, deep learning, although both are being used. In this report, we use the term AI in reference to all subsets and the term machine learning as shorthand to also include its deep learning subset.

For a deeper dive into the three main types of machine learning being used in data centers (in addition to deep learning) and for their various use cases, see **Appendix 2. AI Techniques**.

Data center AI use cases

AI is initially being used in data centers to improve existing functions and processes. Use examples include alarm suppression/rationalization, spotting new anomalies and predicting risks with greater accuracy than other technologies. AI's promise is it that will perform functions and processes better, faster and more accurately.

As discussed earlier, because AI and non-AI big-data analysis is often used in combination (or interchangeably by different suppliers), we have included both approaches in our use cases. The type of data required and the analytical techniques used will vary by use case and the level of granularity sought.

At a high level, different data center use cases can be grouped into three primary types:

- Industrial automation: Intelligent systems and operations that optimize decision support for processes, logistics and automation.
- Perception automation: The use of deep learning algorithms to automate the replication of human perception, cognition and communication.
- Customer engagement automation: The application of AI to drive commercial initiatives through improved customer engagement and service.

Figure 3 shows some of the AI data center uses cases for each type. Of these, industrial automation is most widely applied/most applicable to data centers, although today it is being used mostly to guide decisions rather than for automation.

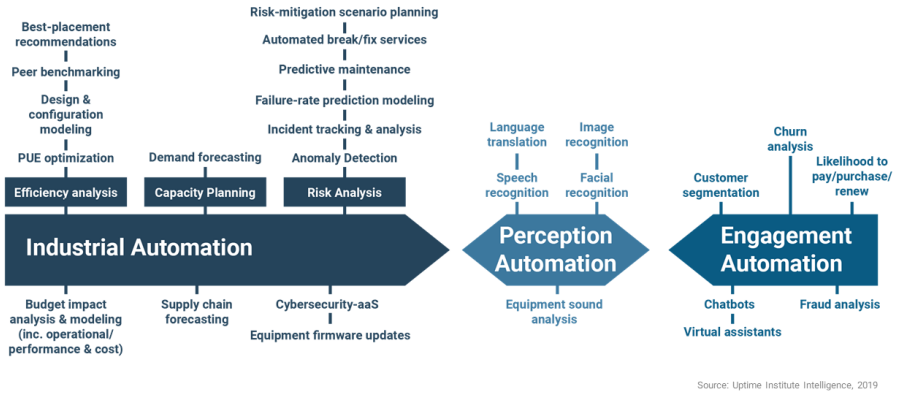

Figure 3. **Use cases: Analytics and machine learning in data centers (by type)**

Most AI in data centers is being used in a semi-automated, incremental way with humans validating or approving recommendations made by AI before a change is made. For example, for predictive maintenance:

1. Machine learning recommends an action, such as predictive maintenance on equipment.
2. A human reviews the recommendation and agrees to the action.
3. A service work order is triggered and a human technician is dispatched to the site to deliver the manual service.

If the same machine learning recommendation is suggested multiple times for similar equipment, a service work order can be automatically generated without the need for a human to review. In this case, the operational process for predictive maintenance is streamlined via automation. This typically happens over a period of time, once humans have a high level of confidence in a specific machine learning recommendation.

Another way to view use cases for data center machine learning is by use case area. Drawing from and expanding on some of the examples above, Figure 4 lists different areas of use cases for machine learning in data centers and provides some specific examples.

Use case area	Specific use cases
Efficiency analysis and optimization	• Cooling utilization • Space utilization • Power utilization • Cabling/port utilization • Server optimization (utilization) • Workload distribution • Transactive energy (utility demand-response)
Risk analysis and mitigation	• Alarms • Cooling performance • Power performance • Root-cause analysis • Scenarios (including 'what-if' modeling, disaster recovery) • Compliance • Predictive maintenance • Faster mean time to recovery
Capacity planning	• Best-placement recommendations • Design and configuration modeling • Best-execution venue analysis
Security	• Image recognition from surveillance cameras • IT network surveillance (cybersecurity)
Budget impact forecasting	• Revenue per data center • Revenue per rack

Source: Uptime Institute Intelligence, 2019

Figure 4. **Use cases: Analytics and machine learning in data centers (by area)**

Some of the use cases we've described require data from beyond the physical data center. This convergence or aggregation of data could be achieved by importing it into a local data lake (via APIs). However, external service providers are likely to be the best equipped (including with up-to-date APIs). We believe that because external providers typically have the skills and tools to integrate a wide range of data for AI, a fundamental change is underway concerning the management and use of critical operational data — a gradual shift from humans in data centers analyzing data to machines analyzing data under a supplier's control, typically in a cloud.

Examples of AI use cases using converged data include:

- Root-cause analysis: Weather data is often analyzed along with data center data to help determine the root cause of an incident.
- Transactive energy/demand-response facilitation: A machine learning algorithm ingests telemetry data from the data center about power consumption, along with data from the utility's transactive energy service.
- Forecasting revenue per rack: An algorithm captures rack-capacity data from the data center, rack-cost data and workload data from the IT service management (ITSM) system, and workload revenue allocations from the financial system. The algorithm can then calculate revenue per rack today and forecast an estimate for, say, a year from now.

One of the most promising use cases is best-execution venue analysis — where is the best venue to execute a workload, taking into account costs,

risks and governance requirements? This would require various types of data, including telemetry/utilization data from the data center, workload information from the ITSM, cost data from the financial system, and cost and availability data from a public cloud service.

Figure 5 outlines some of the types of data that being used to achieve an AI use case or application.

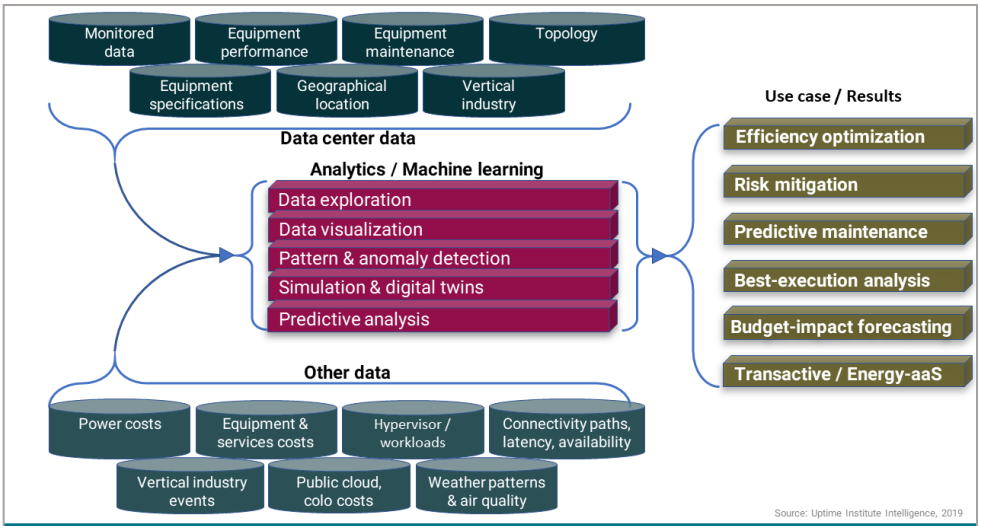

Figure 5. **Data inputs and use cases: Analytics and machine learning in data centers (by type)**

For examples of use cases by different types of machine learning techniques, see **Appendix 2. AI Techniques**.

Adopting data center machine learning

There are different ways to deploy machine learning in a data center, ranging from pay-as-you-go cloud services to building technology in-house. The best approach will depend on the available data, an organization's internal resources (see below) and the overall goals/ strategy.

Some large, global colocation providers, for example, are developing machine learning internally as a component of a broader data-driven strategy. For example, they have hired a team of data scientists to develop different machine learning algorithms and models to help them lower PUE across their portfolio, to automate optimal interconnections for customers, to power chat bots for customers' online queries, and to develop sales approaches for customers based on their purchase and payment history.

However, most data centers using AI are buying commercial in a more embedded form, as capabilities in one of two ways via:

- DMaaS, where DCIM software may or may not be required.
- DCIM software deployed either on-premises or as a service.

Requirements

The barriers for suppliers and others to develop machine learning have been significantly lowered in recent years due to the open-sourcing of techniques and technologies, such as from Google, and the availability of commercial toolsets and platforms, such as IBM's Watson internet of things (IoT) platform. This so-called democratization of technology has led to new products and new specialist suppliers.

Developers of machine learning approaches, whether commercial or homegrown, require all of the following:

- Data scientists.
- Data center domain experts.
- Significant and fast computing capabilities (public cloud is cheapest).
- Significant amounts of data, including for data center telemetry, design, configuration and operation.

Users of machine learning products and services require:

- Budget for an operating expense (although free basic services do exist).
- Willingness to share data about the facility with a supplier (in an anonymized form, usually). This assumes that the user is not also the developer, as may be the case with some very large operators.

Many services will also require sharing encrypted data about a facility via a wide area network, typically the internet. Some services require the user be comfortable with public cloud.

Below are descriptions of and requirements for the two main types of commercially available machine learning offerings: AI-driven DMaaS and AI-driven DCIM.

Algorithms vs. models

AI marketers can use the terms algorithms and models to mean the same thing, although they are not.

An **algorithm** is a sequence of mathematical steps or computational instructions. It is an automated instruction set. An algorithm can be a single instruction, or a sequence of instructions — its complexity depends on how simple or complex each individual instruction is and/or the sheer number of instructions that the algorithm needs to execute.

In AI, a **model** refers to a mathematical model that is able to process data and provide the expected response to or outcome of that data. For example, if an algorithm is applied to a data set, the outcome would be the model. A model changes if the data fed into the algorithm changes, or if the same data is fed through a different algorithm.

AI-driven DMaaS

Suppliers of DCIM, some of which are also suppliers of facility equipment, have been early developers of machine learning in data centers. In late 2016, suppliers began offering big-data cloud services known as DMaaS, which deliver customized analysis via a wide area network and which are paid for on a recurring, as-you-go basis (one supplier charges by the minute).

DMaaS aggregates and analyzes large sets of anonymized monitored data about equipment and operational environments from different facilities (customers). The data is analyzed using different big-data techniques, including machine learning. Results for individual customers may be tailored to their specific data center and delivered via an online dashboard (including mobile), email, text and phone.

DMaaS suppliers include Carbon Relay, Eaton, Nlyte Software, Siemens, Schneider Electric and others (see **Appendix 1. Suppliers** for a longer list of data center AI suppliers).

DMaaS can include prebuilt machine learning models, which means no historical data or an existing DCIM system is required. As new data is fed into a prebuilt model (via the service), the accuracy of analysis improves.

One AI cloud service, for example, can simulate data center environments, which marketers like to call a “digital twin,” by applying a prebuilt model. It can do a basic simulation using as few as two types of data: blueprints of the data center building and mechanical systems.

Using a prebuilt model (and visualization tools), the service simulates the current operations of a data center (i.e., its digital twin today) and a more efficient simulation (i.e., its theoretical digital twin if certain actions are taken). The delta between the two helps identify areas for efficiency improvements or lower risk.

To adopt DMaaS or similar cloud services, the basic requirements are typically two downloads:

- Gateway software to gather and send data from monitored devices to the suppliers' cloud.
- Web-based mobile and/or desktop software that acts as a personalized dashboard for managers to consume analysis, including for alarm notifications (which are typically pushed to mobile devices), for an overview of all sites under management, and for recommendations, assessments and reports.

Another area of data center AI is cloud services or software as a service (SaaS) that make use of AI algorithms embedded into data center equipment, such as cooling units. These services are in development.

DCIM + DMaaS?

Most AI services do not require DCIM, but some do. For example, DCIM specialist Nlyte Software provides DMaaS that utilizes IBM's IoT Watson platform (AI) — but it available only for users of Nlyte's DCIM monitoring software. (An advantage of a platform such as Watson is access to an enterprise-ready tooling environment to customize models and data outcomes — something that is possible with or without DCIM or a DMaaS subscription.)

What is the best practice? It is still early days. In facilities where DCIM is deployed, DCIM can be a rich source of data for any machine learning approach. A major advantage is the historic data stored within DCIM, which can be used to speed up the training and accuracy of machine learning models. However, the same accuracy can be achieved just over a long timeframe, without pre-existing models.

For any data center seeking closed-loop automation or mission-critical alarming, it is advisable to also install DCIM locally, on-premises — to avoid potential latency or service interruptions that can arise when relying solely on a wide area network/the internet.

AI-driven DCIM

Machine learning can also be embedded in on-premises DCIM software. Prebuilt machine learning models in the software are automatically fed historical and real-time data from DCIM — all that is required is the software. Optionally, additional data can be added automatically and manually.

For example, DCIM supplier Maya HTT embeds a machine learning model into its DCIM system. The model uses machine learning algorithms that are applied to the DCIM data. Data from additional systems can also be automatically applied via APIs. Projections for future changes, such as a facility expansion, can be entered manually. In this way, the model can predict performance, capacity demand and costs for a data center for medium-/long-term planning (one year or more into the future).

Here's an example of how this might be used: With data from DCIM asset management, ***plus*** rack-costing data from an ITSM system or financial system (such as SAP), ***plus*** manually entered data about a major new

cloud initiative, the software could calculate annual rack revenue today versus when the new initiative is implemented.

The type, depth and breadth of machine learning — and its accuracy — will ultimately depend on the availability of actual data from a data center (via sensors, meters, etc.). Historic data, such as from DCIM systems, can mean greater breadth and accuracy of results from machine learning. As more data about the data center is collected over time, the more accurate and broad the machine learning analysis will likely be.

Risks

We are unaware of any commercial product or service that does not enable customers to keep their own data should they wish. As such, all providers advertise that their customers own their own data. In reality, customers are co-owners of their data. The supplier typically also keeps a copy of the data, even long after the customer stops using the service (when the paid service stops, the data becomes an anonymous part of a supplier's data lake).

Whether lack of certainty or clarity over data ownership and locality is a risk to data centers is vigorously debated. Some say that if hackers accessed data, it would be of little use as the data is anonymized and, for example, does not include specific location details. Others say hackers could apply techniques, including AI, to piece together sensitive information.

Below are four areas of risk that we have identified with commercial machine learning offerings: commercial, legal and service level agreement (SLA), technical, and interoperability and other "unknown unknowns."

- **Commercial**: AI models and data are (often) stored in the public cloud and outside of immediate control (if using a supplier model) or may be on-site but not understood.
    - Commercial machine learning products and services raise the risk of lock-in because processes and systems may be built on top of models using data that cannot be replicated.
    - Pricing may be increase as adoption grows — at present, prices are low to attract new data (to build up the effectiveness of AI models) or to attract equipment services or sales.
    - A high reliance on AI could change skills requirements or "deskill" staff positions (see **AI changes skills requirements**), which could potentially be an issue.

• **Legal and SLA**: Again, AI models and data are stored outside of immediate control (if using a supplier model) or may be on-site but not understood.

  ◦ This may be unacceptable for some, such as service providers or organizations operating within strict regulatory environments.

  ◦ In theory, it could also shift liability back to an AI service supplier — a particular concern for any automated actions provided by the service.

• **Technical**: While we usually understand what types of data are being used for human actions and recommendations, it is not always possible to understand why and exactly how a machine reached a decision.

  ◦ It may not be possible to easily change or override decisions.

  ◦ As machines guide more decisions, core skills may become outsourced, leaving organizations vulnerable.

• **Interoperability and other "unknown unknowns"**: The risk from the development of "2001" HAL scenarios (i.e., the singularity) are overplayed but there is an unknown, long-term risk.

  ◦ One example is that AI is likely to be embedded in most cases (such as inside individual equipment and management systems). This could lead to situations where two or three or five systems all have some ability to take action according to their own models, leading to potential runaway situation — or conflict with each other. A building management system (BMS), for example, may turn up the cooling, while an IT system moves workload to another location, which turns up cooling elsewhere, for example.

AI changes skills requirements

Will more machine learning decision support in data centers reduce the need for or "deskill" human staff such as operators, managers and technicians?

Deskilling, reskilling

Data center AI is unlikely to eliminate the need for all skilled labor, but it is likely to reduce and change requirements. This is already beginning to happen in facilities that have adopted automation and other technologies. For example:

• Real-time DCIM monitoring can reduce the frequency of rounds carried out by technicians.

- DCIM change management and configuration can reduce the need for manual asset audits.
- Change-management automation in networking eliminates the need to manually assign connections.

According to consensus thinking, AI in data centers will solve repetitive operational tasks ("set and forget") so that staff can focus on specialty and strategic areas. In some data centers, for example, AI-driven robots are racking and stacking IT hardware. But as more strategic and higher-level AI capabilities are developed, it could reduce the reliance on data center human expertise.

Higher-level AI capabilities require domain expertise from humans to implement with confidence — for example, raising server inlet air temperature, reconfiguring the layout of computer rooms and adding new data center capacity in anticipation of future increased demand. As a best practice, AI would make these recommendations and a human would approve or validate before implementing the change. Over time, this could mean:

- Human experts spending significant time assessing recommendations from AI rather than coming up with solutions themselves.
- An inability for data center newcomers to gain hands-on expertise, including in order to effectively assess recommendations from AI.

This could create a future dilemma: The sector already faces growing challenges in attracting human talent; how will it develop human domain experts who can ask the right questions for AI to solve and will know how to best respond to insight from models?

The singularity, it could be argued, will eventually solve for this: humans will not be required to determine the best response. However, the data center sector is facing a staffing shortage in the coming decade — before the singularity promises to become a reality.

Costs

Most (but not all) suppliers of machine learning-driven DMaaS and DCIM price according to number of devices under management. Some charge per kilowatt (kW) under management.

Below are examples of pricing, at a high level, for some AI-driven DMaaS and similar data center cloud services:

- $60 per device per year: mobile device service only for alarms and device status. Does not does not include recommendations or other machine learning results.
- $365 per device per year: mobile and desktop service that includes alarms, predictive analytics, forecasts, recommendations and other machine learning results.
- 25 cents per minute for capacity under coverage plus five percent of energy costs saved. The supplier requires a direct line of sight to customer's data center energy bill.

These prices do not include the costs associated with instrumenting a data center with the sensors and meters necessary to provide machine learning with operational data about a specific data center. Instrumentation can be a significantly greater expense than the service itself when taking into account the purchase cost of the hardware, its implementation and ongoing maintenance.

Recommendations

It is still very early days in terms of AI adoption in the data center — and perhaps too early for clear guideless and best practices to fully emerge. But Uptime does make the following observations/recommendations:

- Develop a broad and long-term plan for data center operations that includes the use of humans (and includes a staff attraction and retention strategy), DCIM software and AI-driven DMaaS. Deploying AI is a process. Set realistic milestones that must be achieved before moving forward, and then do so in an incremental manner.

• Determine your organization's ownership and use of data for commercial AI products and services. Ask suppliers for certifications/ best practices for data security, data use and data storage. Clarify the legal responsibility of the supplier for failures and poor outcomes as a result of AI recommendations.

• Seek to validate results and recommendations from AI products and services. At a minimum, understand the fundamental level of the depth and breadth of the machine learning being applied. Ask the supplier to show the data points in the model and the relationship between those items — in other words, how the machine learning is using the data to make recommendations for action. Also track the results when actions are taken (by the operator).

• Ask suppliers to provide accuracy rates and rates of false positives — that is, how often the system is providing information that is inaccurate or recommendations that are inappropriate.

• For automation, ask to manually intervene — that is, for a human domain expert to authorize the "go/no-go" decision before an action is triggered — until a baseline of machine learning performance is established. Consider automation an iterative, human process until a high level of confidence in the machine is reached.

Appendices

Appendix 1. Suppliers

Suppliers of data center facility AI products and services include:

• California Data Sciences

• Carbon Relay

• Eaton

• IBM Global Services

• Litbit (partnership with CBRE)

• Nlyte Software

• Schneider Electric

• Siemens (machine learning capabilities for its DMaaS are in development)

• Solecular

• Vertiv (machine learning capabilities are in development)

• Vigilent (partnerships include Hitachi Vantara, Schneider and Siemens)

• Wave2Wave

Appendix 2. AI Techniques

Understanding how an AI product or service works is not necessary for operators and managers. It is, however, important to understand the types of data that influence a machine learning outcome/ decision (see **Recommendations**).

Some AI techniques intrinsically require more human intervention than others to develop, such as those that require training of AI models, labeling of data (which over time AI can help with) and ensuring that models don't become inaccurate over time (known as model "drift"). A relatively high level of human intervention required means higher-priced AI products and services, say some suppliers. However, while the development of some AI techniques requires little human intervention, they can also require a greater amount of data and computational power, the cost of which may offset any savings from using fewer humans.

It is too early to tell — most AI products and services for data centers today are priced relatively cheaply, so suppliers can grow their customer base and, importantly, their stores of data from those customers.

There are two main types of AI used in data centers today: Machine learning and deep learning.

Labeled vs. unlabeled data?

In machine learning, **labeled** data refers to a specific way of identifying the data so that it can be used as "training data." This means that data that is labeled as either an "input" or an "output" (responses to the data) for a machine learning model. When algorithms are fed labeled data, they are able to understand and then identify the optimal relationships between the input and the output data (and, therefore, improve a model). Not all machine learning techniques require labeled data.

The process of labeling data can be time consuming and expensive, particularly in the early stages of training a model. Over time, big-data analytics can be used to automatically label some new data.

Machine learning: detailed

Machine learning uses algorithms that learn from data to make decisions or predictions and improve, without being programmed to do so, when exposed to more data.

There are three main types of machine learning techniques (described below). Of these, supervised machine learning is the most commonly used in data centers and across other industries.

- **Supervised learning**: Humans supply a model and training data. Algorithms take the training data and fine-tune the model so that the inputs and outputs/responses are more closely aligned. As more data is added over time, the algorithms further improve the model and can make reasonable predictions for responses to new data.
- **Unsupervised learning**: Algorithms find patterns or intrinsic structures in unlabeled data. In some scenarios, unsupervised machine learning techniques are combined with supervised ones. In effect, the output of unsupervised machine learning can become the training data for supervised machine learning.
- **Reinforcement learning**: Humans supply a model and unlabeled data. When an algorithm determines an optimal outcome for the data, it is reinforced by a positive mathematical "reward." (An open-sourced reinforcement learning model from Google is appropriately called Dopamine.) By providing feedback, it learns through different variations.

Of these, reinforcement learning is the newest machine learning technique.

Deep learning: detailed

As previously discussed, a relatively new subset of machine learning is deep learning. Deep learning is a type of machine learning that enables computers to learn by example. Deep learning uses multiple layers of artificial neural networks to build algorithms, based on vast data, that find an optimal way to make decisions or perform tasks on their own.

Humans supply training data and algorithms, which break down the input data into a hierarchy of very simple concepts. Each concept becomes a mathematical node on the neutral network. Instead of using machine learning models provided by humans, deep learning uses the training data like a neural network, which works like a decision tree. Deep learning builds new models from its own analysis of the training data. Deep learning can be supervised, semisupervised or unsupervised.

Which is best?

Which machine learning technique is best for which use case? It depends on the quality and sophistication of the algorithm, as well as the model and the data being used. If all these things are equal, however, there are certain techniques that are particularly well suited to certain use cases.

Some assert that deep learning can find greater levels of inefficiencies than supervised learning because it is unfettered by known models and the need for humans to retrain models as or if they become stale. Deep learning can, for these reasons, be less expensive to develop. And deep learning systems can spot patterns or relationships that humans did not even know existed.

On the other hand, supervised learning is more transparent than deep learning, making it easier for domain experts to validate results and, it can be argued, quicker to automate.

Figures A2-1 and A2-2 below give examples of some uses cases that can be well-suited to different types of machine learning and for deep learning.

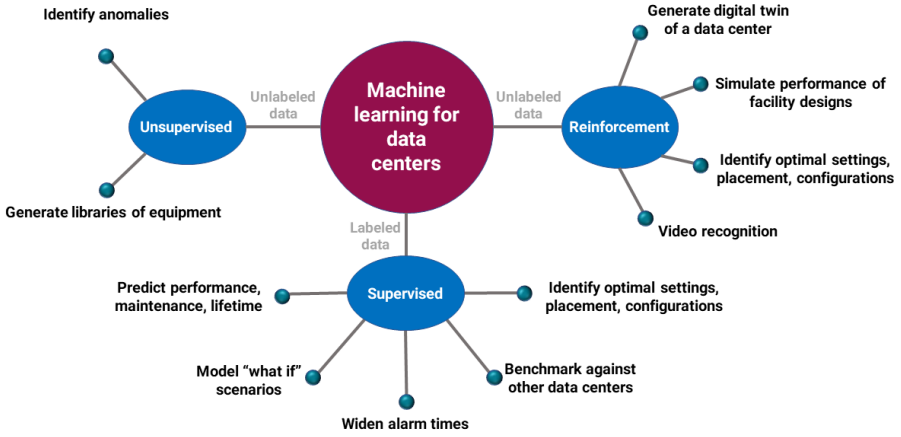

Figure A2-1. **Machine learning: Data center use cases**

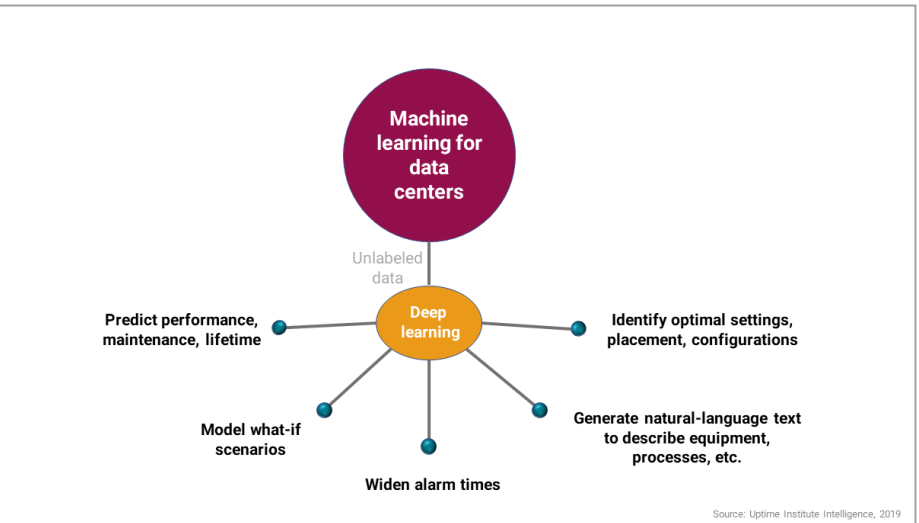

Figure A2-2. **Deep learning: Data center use cases**

It is still early days, but it is likely that specific AI techniques are likely to dominate specific use cases over time.

Uptime Institute® | INTELLIGENCE

ABOUT THE AUTHOR

Rhonda Ascierto is Vice President of Research at the Uptime Institute. She has spent nearly two decades at the crossroads of IT and business as an analyst, speaker, adviser, and editor covering the technology and competitive forces that shape the global IT industry. Contact: rascierto@uptimeinstitute.com

Uptime Institute is an unbiased advisory organization focused on improving the performance, efficiency, and reliability of business critical infrastructure through innovation, collaboration, and independent certifications. Uptime Institute serves all stakeholders responsible for IT service availability through industry leading standards, education, peer-to-peer networking, consulting, and award programs delivered to enterprise organizations and third-party operators, manufacturers, and providers. Uptime Institute is recognized globally for the creation and administration of the Tier Standards and Certifications for Data Center Design, Construction, and Operations, along with its Management & Operations (M&O) Stamp of Approval, FORCSS® methodology, and Efficient IT Stamp of Approval.
Uptime Institute – The Global Data Center Authority®, a division of The 451 Group, has office locations in the U.S., Mexico, Costa Rica, Brazil, U.K., Spain, U.A.E., Russia, Taiwan, Singapore, and Malaysia.
Visit www.uptimeinstitute.com for more information.
All general queries:
Uptime Institute
5470 Shilshole Avenue NW, Suite 500
Seattle, WA 98107
USA
+1 206 783 0510
info@uptimeinstitute.com

