The Industrialization of AI

Upgrading from experimentation to implementation at scale

In partnership with:

Brought to you by Informa Tech

Information Classification: General

Contents

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Introduction

Executive Summary

No broad industry trend, not client/server computing, not affordable hardware, not even the cloud itself, promises to so completely reshape the enterprise than artificial intelligence (AI). Melding decades-old mathematical principles with cutting edge algorithms and readily available, high performance hardware, AI is creating a seismic shift in the way companies across all industries build, maintain, and understand their core and departmental business operations. And owing to an accelerating market investment in the creation of machine learning (ML) lifecycle management tools, enterprise AI practitioners now expect to move rapidly and affordably from AI experimentation to AI in operation.

Whether driving business value within core business such as sales enablement or tackling departmental needs such as marketing campaign evaluation, the use of AI within the enterprise has reached a point of maturity where the overall value proposition of AI no longer dominates budgeting and purchasing discussions. Rather, the conversation surrounding AI has now become more pragmatic in nature, focusing on how to translate early, tentative wins within non-critical processes into more substantial gains within mission-critical business processes. Within this paper, Omdia will analyze this transition, discussing the obstacles, enablers, and best practices that lie along the path between AI experimentation and operational mastery of AI as a critical business substrate and key source of market differentiation.

AI flourishing in times of global disruption

The unprecedented market disruption and societal upheaval that began with the COVID-19 outbreak in early 2020 shows no sign of abating in the near future. The virus has already irrevocably reshaped the relationship between employees and their employers. It has forced companies to radically alter routes to market and supportive business processes. And it has forced enterprise IT buyers to rethink purchasing priorities, trading long-term aspirations for investments capable of yielding more immediate value, be that direct revenue, cost savings or functional resiliency.

And yet, the rapid and growing adoption of AI within the enterprise continues largely unabated by the global COVID-19 crisis. Why is this? It turns out AI outcomes are commonly tied to three general desires, which align with concerns raised by large-scale disruptive events.

- ***Improve*** the efficiency of a process, product, or service often with the objective of automating human tasks disrupted by the pandemic
- ***Reduce*** the cost of performing a task or process, thereby enabling the direct reapportion of funds to better weather temporarily suppressed revenue streams

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

- ***Generate*** wholly new value opportunities by helping companies identify and pivot toward business models better aligned with current market demands

Broadly speaking, these goals hinge on AI's ability to uncover new and more efficient ways of handling or performing a given task, reveal new patterns or correlations in the data, and identify more efficient means of connecting disparate data points. In doing so, AI has the unique ability to improve on human or pre-programmed processes by setting up an algorithm that can learn from past actions and make adjustments in a far faster and more informed (eg. data-driven) manner than is possible with basic process automation or humans-led decisions.

Such capabilities hit at the heart of many problems faced by companies seeking to survive and perhaps even thrive amidst large-scale disruptive events such as a global pandemic. Within numerous industries, Omdia has noted several use cases where AI is driving new business opportunities. Take for example business-to-business (B2B) sales (business services) companies, which are taking AI-driven product recommendation engines typically in use among consumer media services and replicating those in the business services sector as a means of growing B2B e-commerce transaction volume. In this way, distributors and manufacturers are using product recommendations on their own sales websites or on third-party platforms to leverage new sales.

More overtly, AI has shown within this pandemic that it can help many companies reinvent the way they do business. Restaurants within the food and beverage market, for instance, have had to endure both outright closures and highly disruptive reopening requirements. In response, many companies are turning to personalizing drive-through, kiosk, and mobile app menus, all driven by AI. These systems look at factors such as the weather, time, local events, traffic levels, historical sales data, and currently popular items before recommending food choices or suggesting wait times for tables or pickup orders.

And within the food and beverage industry, particularly for companies employing human labor, the use of AI as a means of avoiding downtime due to localized COVID-19 outbreaks has become paramount. To that end, many meat rendering and packaging companies such as Tyson Foods, JBC and Pilgrim's Pride, are accelerating investments in robotic processing. This will help protect against facility closures and thereby strengthen supply chains that have been rendered fragile by the global pandemic. Supply chain customers themselves are hoping to use AI to identify and predict demand for products and components, as well as managing historical data and mixing it with current and future demand figures to create optimized delivery schedules.

Key business drivers to justify AI investment

As mentioned above, AI outcomes center around cost reduction, value generation, and operational efficiency. As we will discuss later in this paper, each distinct AI business outcome, such as increased revenue through B2B recommendation engines, relies on the application of one or more AI technologies. Omdia defines four such technology and solution categories.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

- **Machine learning (ML)**: Computerized predictive or categorical mathematical algorithms
- **Deep learning (DL)**: A subset of ML that uses human-like neural networks.
- **Natural language processing (NLP)**: an application of ML and DL that understands and generates human speech.
- **Computer vision (CV)**: Another ML and DL utilization focused on identifying and classifying images of objects from live camera or captured data.

Of these, ML and DL appear as the primary engines of AI adoption and innovation, given that the other AI technologies rely on the use of ML and DL to make sense of the data captured. The key features of ML and DL are based on the ability to identify patterns in data, connect discrete data elements, and provide faster and more powerful analysis than humans or static analytics programs. Doing so depends not just on algorithms and models but also upon the availability of specialized hardware from vendors capable of training exceptionally large data models in a short amount of time and delivering model results (predictions) with as little latency as possible. Fortunately, market leader NVIDIA$^{\rightarrow}$ along with AMD, Intel, Microsoft, Google, and others have invested heavily in this market opportunity, resulting in an abundance of available hardware acceleration options across both premises and cloud. The most dominant acceleration technology is the graphics processing unit (GPU) popularized by NVIDIA in the late nineties originally as a means of speeding up gaming graphics. As a result, enterprises are able to handle routine tasks more quickly and accurately, thereby increasing productivity and efficiency. Furthermore, ML and DL enable the development of systems that permit more intuitive, human-like processing of information, making it simpler and more instinctual for humans to interact with machines and technology.

The end benefit for enterprises is the ability to augment or replace functions that are time or resource intensive with automated, intelligent technology. This leads to increased productivity and increased efficiency, and often can open up new technological, product, or service offerings that can directly improve a company's bottom line. AI technologies do so in two ways.

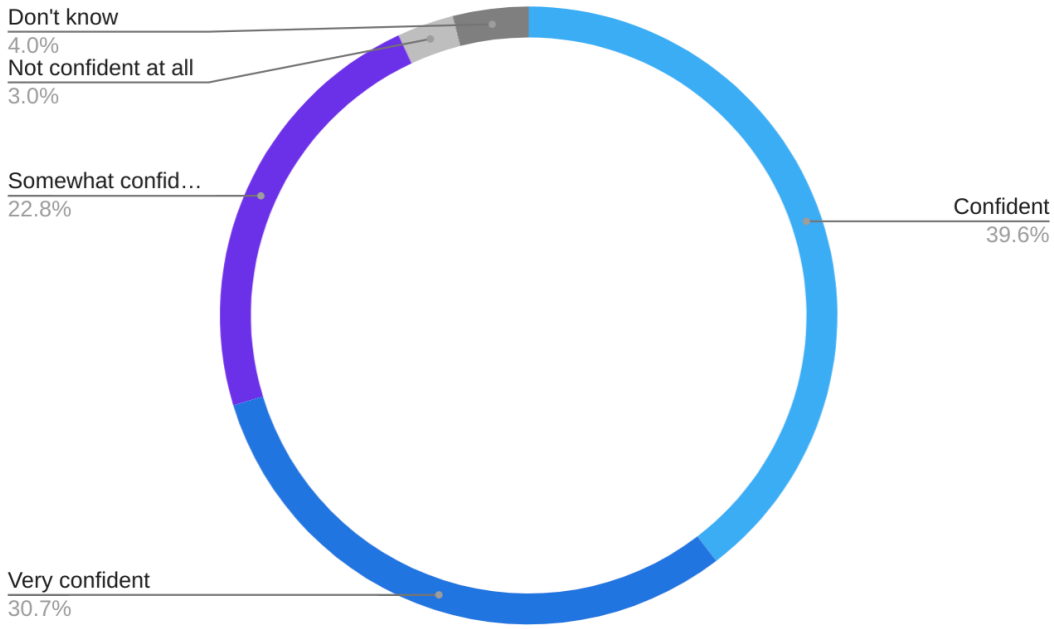
- **Direct revenue**: This revenue represents the income derived from selling an AI-based solution. For example, the prevention of a cybersecurity threat is a direct application of AI wherein AI is used for cybersecurity or emotion analysis and emotion analysis is provided as a service to third parties.
- **Indirect revenue**: This revenue does not necessarily represent the income that is derived from the use case itself, but where AI is essentially seen as a layer or plugin that improves a product or service. For example, web searches or e-commerce product recommendations are both well-known use cases where AI has been used to provide better results. The revenue in both cases cannot be directly linked or attributed to AI, as it is advertising-based revenue or e-commerce sales.

Omdia has found that despite the pandemic, enterprise buyers remain bullish on AI with continued spending in support of both direct and indirect revenue opportunities with a recent survey revealing that the vast majority of enterprise practitioners had faith in near-term positive outcomes (see figure 1). For practitioners, the center of that confidence resides in operational efficiencies and

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

organizational stability, with survey respondents noting that the majority of upcoming business unit deployments (within the next 24 months) will support IT operations and supply chain management use cases.

Figure 1: Near-term confidence in positive AI results despite COVID-19



© 2020 Omdia

n = 146

End-user companies were asked, "How confident are you that AI will deliver positive results towards your business goals within the next 12–24 months?"

Source: Omdia

Achieving competitive differentiation with AI

Underpinning these objectives is AI technology itself. A substantial level of AI competency can serve both core and departmental use cases. It can drive an exceptionally broad array of use cases. And it can deliver an overall competitive differentiator for companies. As a result, organizations are working toward an enterprise-wide, "all in" mentality, believing this to be necessary in fully accessing the operational and economic benefits of AI. But enabling AI in this manner is not a plug-and-play proposition. Significant time, resources, and capital must be invested, and in most cases, internal company teams do not yet possess the necessary operational experience, nor do they have the cutting-edge data science skills, software development expertise, and computing infrastructure available to effectively complete a truly transformational AI implementation journey.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Omdia believes, therefore, that the financial constraints of COVID-19 will ultimately strengthen AI adoption as a differentiator, exposing gaps in investment, skill and experience, thereby separating the “AI haves” from the “AI have nots.” Early adopter companies—those that have made AI investments and commitments prior to the COVID-19 pandemic—are reaping the benefits of their investments. And as mentioned above, the current research indicates that most practitioners will not slow down their adoption of AI because of the COVID-19 crisis. Conversely, companies that did not invest in AI prior to COVID-19 will likely delay or shrink their investments until better economic conditions emerge. This could create further competitive advantages for opportunistic early adopter companies that may be difficult for those lagging behind to overcome.

Further, because the disruptive nature of the pandemic promises to continue well into 2021 and beyond, this crucible will extend beyond early adopters to test the mettle of even mature AI practitioners, separating those who are able to bring a large measure of AI projects to fruition (and keep them there) from those who fail to effectively and consistently move beyond a few, highly specialized, early projects.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

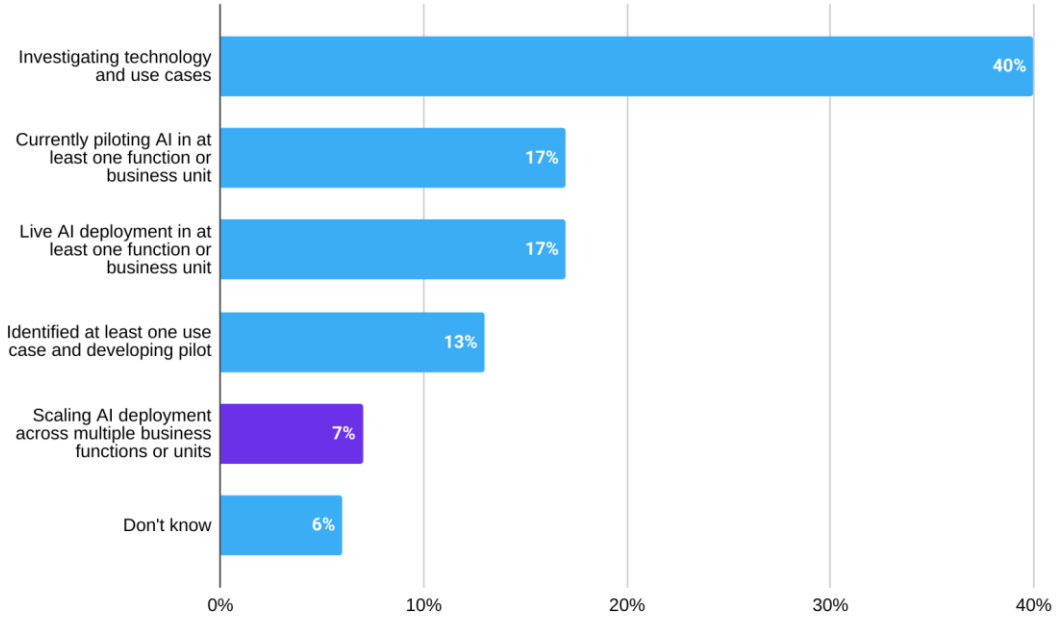
AI market maturity

Moving beyond early pioneers and toward everyday AI

Compared with mature markets such as enterprise cloud compute and storage, from an organizational and revenue-generating standpoint, AI is still in its infancy in terms of how companies build, deploy, and manage AI outcomes. A recent study from Omdia research reveals that 53% of enterprises are investigating use cases or have identified and are currently piloting at least one use case. And yet of those same companies, only 7% of survey respondents are currently running AI at scale across multiple business functions or business units (See figure 2.)

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Figure 2: State of AI deployment



© 2020 Omdia

n = 161

End-user companies were asked, "Q: What is the state of AI deployment at your company today?"

Source: Omdia

Why is there such a sizable gap between isolated experimentation and widespread implementation? At a very fundamental level, AI practitioners must run a complex gauntlet of challenges in creating even the most mundane ML task. Whether training a pre-built chatbot model for customer service or coding a fraud detection DL algorithm from scratch, at each step in the ML lifecycle, a single misstep could slow development, elevate cost, or render the final outcome mute. Worse, a single misstep could deliver misleading or erroneous results that may go undetected even as the solution enters production. Above basic processes, a host of challenges and potential pitfalls await enterprise AI practitioners including use case selection and performance measurement, infrastructure investment, project ownership, skills acquisition, and privacy/security compliance. But it is the mundane, day-to-day workings of ML which pose the greatest, ongoing challenge for companies wishing to become data-and AI-driven.

Process Complexities

What kinds of missteps must enterprise AI practitioners avoid in building ML outcomes? Consider an extremely simple ML project that aims to predict which customers are least likely to renew an

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

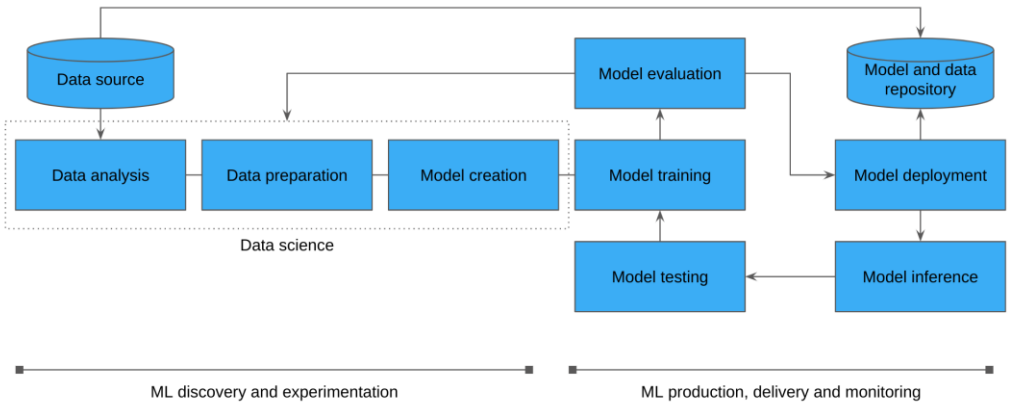
annual service subscription using historic customer data. For this example, there are at the most fundamental level three important phases of development:

- Data acquisition and preparation
- Data science exploration and experimentation
- Predictive model deployment and monitoring

With this simplified customer churn example, an AI practitioner, usually a data engineer, must obtain, clean, transform and label the necessary data for training and testing the predictive model. A data scientist must then explore that data, further engineering data features, and then embark on a series of iterative experiments to find and optimize the most applicable ML model that will yield the desired outcome. And all of these efforts presume a fitting investment in an underlying infrastructure tuned to the needs of this project, providing, for instance, access to an adequately sized compute cluster and supporting set of GPUs. For this example, once a model reaches the desired accuracy, the data scientist may publish his or her findings, helping marketers best decide how to communicate with "at risk" customers.

This model, however, could just as easily run continuously in production, using live data to deliver an ongoing view into the state of the subscriber base. In that situation, provisions must be made to host the running model, deliver predictions programmatically, and monitor the model for accuracy over time (see figure 3).

Figure 3: Enterprise ML workflow



© 2020 Omdia

Source: Omdia

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

A question of repetition, performance, and trust

Given these very fundamental complexities, the relatively low penetration rate of AI as a company-wide endeavor begins to make perfect sense. The challenge for companies looking to move beyond the realm of basic AI proof of concepts (PoCs) and pilot programs rests within three key concepts for any given ML experiment or AI implementation:

- **Repeatability** -- Can it be replicated or iterated reliably?
- **Scalability** -- Can it be trained affordably and run at scale?
- **Surety** -- Can it be trusted to deliver the right predictions?

Unfortunately, there are a wide range of obstacles preventing even the most experienced IT organization from putting these three concepts together in harmony. The notion of data science itself stands as an impediment, owing to its highly experimental and iterative nature. For each general phase in the AI implementation lifecycle, practitioners must face a number of challenges including this small set of examples.

Repeatability -- It is difficult for AI practitioners to repeat what is a highly investigational methodology, one filled with stops and starts, dead-ends and unforeseen avenues of exploration. Unmanaged code, often written in Python and R, lives ungoverned within various Jupyter notebook implementations, making it nearly impossible for anyone but the original author to track and manage that code over time. Add to this the challenge of maintaining versioning across a wide array of libraries and frameworks, which change from project to project, and it's easy to see how lessons learned in one project do not readily carry over into future endeavors.

Scalability -- It can be notoriously difficult for companies to effectively manage resource requirements for both development (training) and deployment (inference), as these tasks are themselves dependent upon a myriad of malleable conditions. The same holds true for all supportive storage and processing resources (database instances, data pipeline processing, inference engine execution, etc.). This high degree of entanglement makes it difficult for IT managers and CTOs to predict and therefore manage costs -- a difficulty that grows exponentially as new AI projects enter development and production.

Surety -- And it can be difficult to trust AI business outcomes altogether, owing to a lack of transparency within many DL predictive models, unchecked biases lurking in both data and model alike, poor code documentation across the project lifecycle, and inadequate testing of models prior to deployment. Even if an organization successfully tackles these (and many other similar) challenges during deployment, maintaining a level of confidence over time demands a high degree of vigilance, monitoring models over time to ensure that their efficacy does not diminish owing to changes in the supporting data or surrounding systems. For highly regulated industries, this kind of monitoring can demand the actual replication of a given model's output at a given time. This can be an impossible task for organizations unable to fully document the entire ML lifecycle: data, data preparation, feature engineering, model selection, parameter tuning, model testing, etc.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

In short, each AI project operates as complex systems, which can be very difficult to predict or control. Any attempt to scale beyond a handful of AI projects only exacerbates these difficulties. It is interesting to note that software development found itself in a similar situation in the 1990s, when success depended heavily upon navigating a complex and often unpredictable landscape of underlying hardware and software systems. Fortunately, the software development market evolved to incorporate highly flexible and manageable innovations such as agile development methodologies, collaborative version control practices, as well as software virtualization and containerization. These innovations, collectively described as operationalized development (DevOps) have enabled developers to radically improve software development, deployment and maintenance outcomes.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Operationalizing AI: moving from experimentation to implementation

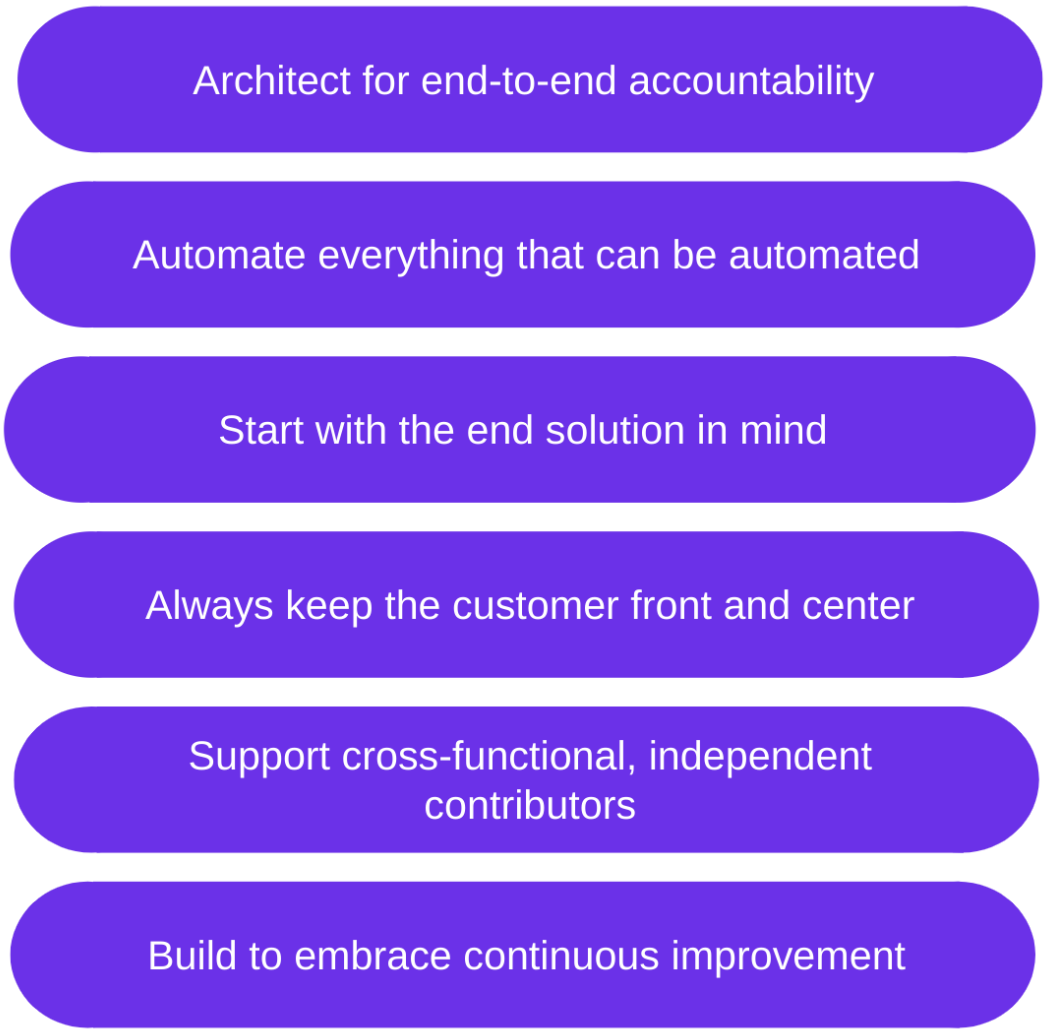
Today, enterprise AI practitioners suffer from the same challenges that software developers faced back in the 1990s. They must tolerate a high degree of system- and organization-level complexity, and they must do so without any reliable means of controlling project dependencies at scale across the enterprise.

ML Ops principles and practices

Fortunately, many of the same DevOps techniques and technologies that modern enterprise software developers have come to rely on are available to AI practitioners. Stated simply, DevOps seeks to operationalize the software lifecycle through the application of several core principles (see figure 4).

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Figure 4: Core DevOps principles



Source: Omdia

Collectively, these DevOps principles make up the operationalization of ML (eg. ML Ops), but each in turn points to just one singular starting point: a rigorously enforced means of managing all of the metadata that feeds into and is created from within the lifecycle of a given ML project. Once established, this metadata can enable the adoption of numerous DevOps toolchain technologies including

- Collaborative source-code management and versioning system
- Centralized, managed artifact repository

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

- Model testing, deployment, and retraining platform
- Monitoring service for running inference engines
- Governance service to track, explain and replicate model outcomes

And yet, ML Ops is not DevOps. Unlike software code, ML models degrade over time, requiring continuous monitoring. For DevOps, an entire project can be versioned as a single unit. With ML Ops, unfortunately, practitioners must version code, predictive models and supporting data...data that changes over time. The idea of continuous testing in DevOps, therefore, doesn't map well to ML Ops, because ML Ops workflows demand continuous training and validation -- two very different mechanisms.

Applying DevOps to data science

Given these differences, mapping DevOps principles and toolchain technologies to ML Ops seems to require a complete makeover of established data science departments and a forced relationship upgrade between data scientists and IT staff members. Fortunately, because the AI marketplace was built on and still continues to run on open source software, the barrier to entry for ML Ops is surprisingly low. Moreover, ML Ops does not demand a top-down, systemic approach. Rather, users can readily begin their ML Ops journey by adopting a few simple tools.

To begin controlling Python and R resources, customers can put Microsoft's freely available GitHub platform to work with very little effort. GitHub enables data scientists to basically package their Jupyter notebooks up into a format that can be versioned and shared collaboratively. To better manage library and language interdependencies within Jupyter notebook, users can adopt Anaconda. And to begin defining repeatable data pipelines that feed into and support Jupyter notebooks for training and inference, customers can employ DVC. Together these three freely available solutions can address many ML Ops concerns such as repeatability and can therefore help enterprises scale data science experience on a project-by-project basis.

Using GitHub for code, DVC for data, and Anaconda for libraries will indeed speed time to market, lower development costs, and mitigate many risks. Moving individual projects to production and growing beyond individual projects with ML Ops, however, demands a more centralized architecture capable of applying these same ideas at scale, over time. Of course, these and other open source projects can be combined and built on in a bespoke manner to accommodate the complete ML Ops lifecycle. However, this kind of undertaking lies well beyond the capacity (existing resources, skills, and experience) of most enterprises that stand on the cusp between pilot experimentation and critical use case implementation.

To close this gap, a large portion of enterprise AI practitioners are increasingly turning to lifecycle complete ML Ops platforms that emphasize repeatability, scalability, and surety. Anaconda, as an early example, now tackles far more than the management of software packages through containerization, evolving into an end-to-end ML Ops platform spanning development, collaboration, deployment, and governance. More recently, vendors, particularly those with experience in delivering supportive infrastructure, have gone a step further, invested in a combined ML Ops and cloud-native workload orchestration layer as a means of driving digital transformation. This is the case with HPE Ezmeral ML Ops and HPE GreenLake, which together seek to operationalize

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

ML solutions built in ML Ops frameworks like Anaconda together with a data fabric, container platform, a management and security layer, edge services, and even services specific to the operationalization of IT itself (AIOps).

Such innovations have led to a highly dynamic market which now holds many platforms similar to Anaconda, HPE Ezmeral ML Ops, and hyperscale cloud platform providers, Microsoft, Amazon, Google, and IBM. Interestingly, a small but influential number of enterprise AI practitioners have also brought to market their own ML Ops platforms, which were initially built internally as bespoke solutions. These include Iguazio from Netflix, Michelangelo from Uber, Metaflow from Netflix, and Flyte from Lyft. These coupled with the numerous pure-play offerings on the market from Databricks, H2O.ai, Datarobot, Cnvrg.io, Cloudera, SAS, Dataiku, and many more make up what is an exceptionally rich set of potential options for enterprise AI practitioners looking to purchase a ready-made ML Ops platform.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Key requirements of AI industrialization

Higher level complexities

Concerning the business of operationalizing ML, organizationally, enterprise AI presents practitioners with a complex tapestry of interwoven and interdependent elements. There are numerous subscription and purchasing structures used by services firms, platform vendors, hardware providers, and other market participants. Even with a growing number of lifecycle-complete AI development platforms available, most enterprise AI practitioners must contend with a wide array of providers and supporting technologies. All of these solutions provide enterprise practitioners with an operationalized framework upon which to build AI outcomes. For example, while some integration and customization service work may be included within a given AI solution, an additional, costly professional services component may be required to ensure a smooth delivery. In fact, many ML Ops providers incorporate professional and managed hosting services into their core software subscription contracts as a means of ensuring that customers realize a marked return on investment.

AI implementations, successful or otherwise, are powered at least in part by open source software. Data scientists and data engineers actively leverage a staggeringly complicated collection of libraries, frameworks, and tools. Each open source technology (and numerous versions of each technology) must be provisioned and supported by enterprise IT practitioners on a project by project basis, creating an uneven patchwork of support contracts and version dependencies that must be managed (see figure 5).

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Figure 5: Important open source projects supporting AI

© 2020 Omdia

Source: Omdia

To combat such complexities, enterprise practitioners are increasingly turning toward fully managed cloud platforms able to demonstrate a significant level of dedication to open source. This dedication often comes through direct partnerships between the cloud provider and vendors built on open source offerings such as Confluent (Kafka), DataStax (Cassandra), and Elastic (Elasticsearch). Often cloud providers will contribute to and offer their own implementations of popular open source projects. They will often allow for direct support of these projects within their own proprietary offerings as is the case with both Microsoft and Google, which support Keras within their own DL frameworks.

This kind of openness extends to computing infrastructure as well. Recognizing the importance of matching a given AI project with the correct AI acceleration hardware regardless, cloud platform providers have been quick to invest in a wide array of accelerated architectures both their own and those offered directly by manufacturers such as NVIDIA.

AI requires more than accelerated hardware. NVIDIA itself has invested heavily in creating a cloud software platform, NVIDIA NGC, a hub for GPU-optimized software and AI models for deep learning, accelerated analytics, scientific computing and high performance computing (HPC). AI practitioners building on top of cloud platforms such as AWS, Microsoft Azure, etc. can orchestrate containerized training and inference workloads via NGC.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Organizational concerns

Organizationally, companies must weave together a similarly complex pattern of human expertise, including software developers, IT architects, data engineers, data analysts, data scientists, business analysts, security specialists, and domain experts. With the application of AI itself, as with AutoML, some of these tasks associated with each role can be augmented, simplified, and in some, limited instances fully automated. Still, to put a proper AI team in place and to enable that team to operate across the business at scale, demands a high degree of cross-departmental coordination and ongoing oversight.

The ownership of AI within an organization is key to how AI projects get implemented and managed on a day-to-day basis. Centralized management structures for AI require a Chief AI Officer in charge of AI projects, spearheading management and implementation of AI within an organization. This role, however, is not yet a common practice. Instead this role is often filled by a chief data officer or chief technology officer. Centralized AI management allows for a "paintbrush" approach to AI where AI can be applied to multiple silos within an organization. But central management risks not clearly understanding the key business metrics or needs within each of the silos. Centralized management is, however, a good approach to attracting the best talent, with AI engineers looking to join a company where the "Head of AI" is someone that has a major standing in the AI community.

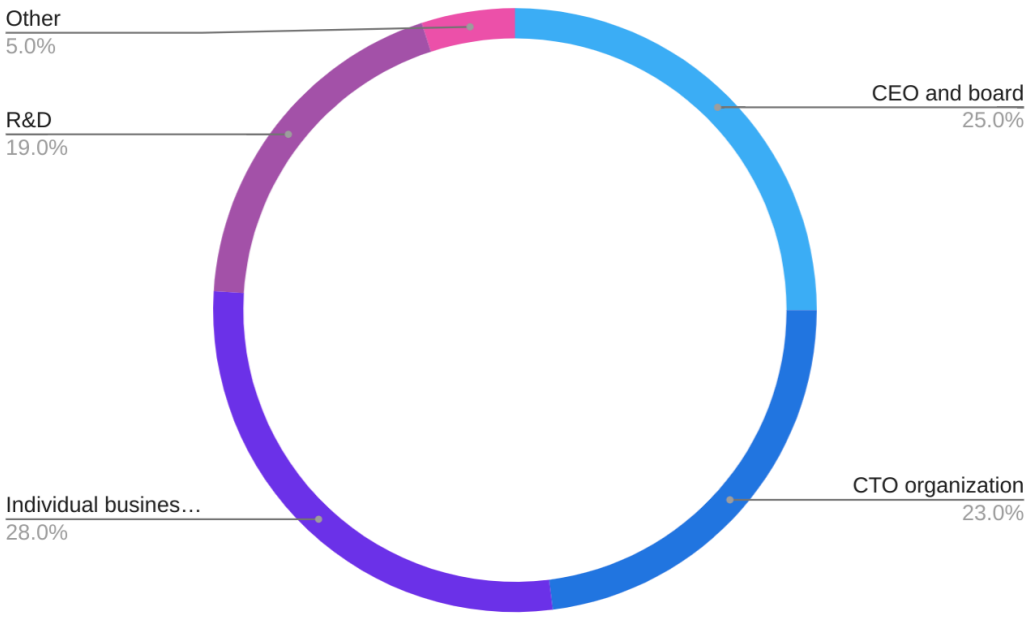
For the most part, partially decentralized AI is a better approach to follow as it retains central authority but grants departmental autonomy. This hybrid management model gives departments the freedom to define and drive their own AI strategy rather than have central management dictate terms. Also, in the long run as AI becomes the default way of running and managing a business, the CEO can serve as the default Chief AI Officer or Head of AI, and therefore create a separate business function for AI in a centralized approach separate from the CEO.

Until that time, Omdia recommends that practitioners establish an AI center of excellence (CoE), staffed with a core set of data scientists, data engineers and other domain experts that falls under the auspices of a Chief AI Officer. This group can oversee without dictating total control over individual departments or endeavors. This encourages business units to employ their own domain experts as needs demand.

According to a recent Omdia survey of enterprise AI practitioners, such coordination is still difficult to come by. Today, there is no absolute consensus on where AI responsibility resides within end-user respondents, as decision-making appears to be evenly distributed across different power centers (see figure 6). It is likely that such decision-making will consolidate over time, perhaps moving completely out of R&D as many companies begin to think about AI expertise and intellectual property (IP) as a core competency. Still, for most organizations, AI will remain diffused across various business stakeholders with overall control passing back to some central CoE or other guiding authority.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Figure 6: Centers of AI responsibility



© 2020 Omdia

n = 161

End-user companies were asked, “Q: Where does AI responsibility and decision-making reside within your company?”

Source: Omdia

Levels of investment and infrastructure considerations

There are no hard-and-fast metrics on how much AI deployments will cost initially or over time. But one thing is clear. Managing cost at scale matters. A single pilot program can generally require an investment ranging from the low to mid-six figure range and take upwards of nine months to reach fruition. That overall investment can be impacted by the length of the pilot program, the complexity of the project, and the number of people working on the engagement.

Even so, to take the successes of a pilot program and expand that success across the organization to attain tangible benefits, additional investments in hardware, software, and support and maintenance will be required.

This is particularly true for supportive AI acceleration infrastructure and services built on the NVIDIA GPU accelerated computing platform, mentioned earlier. Such infrastructure is paramount because training time for an individual AI model or a portfolio of models can be extremely time intensive. For DL workloads such as computer vision and natural language processing, the simple act of training a predictive model can consume system resources over hours and even days. For this reason, AI

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

practitioners must carefully balance time with cost, selecting an AI acceleration configuration (available memory, type of GPU architecture, number of GPUs, etc.) that is aligned with the needs of the project at hand.

The. As an example, fetching real-time clickstream data to power a live inference model running on an edge device, if not correctly provisioned, can drive up both cost and latency. This challenge is often due to the way cloud-based ML Ops solutions generally charge enterprises, which can be a complex, interdependent assembly of numerous utility services organized by compute hour. Pricing and management can become even more complex if the project at hand demands a hybrid or multi-cloud deployment scenario. In these circumstances, many of the benefits of cloud-native niceties such as auto-scaling and fully managed software are simply not available. Paying for the amount of compute time required to train or run complex ML implementations, therefore, may wind up costing significantly more over the long term than simply purchasing the hardware and software outright.

Hardware vendors, of course, will price their offerings based on the power, size, and reliability of their solution. Fortunately, as with other areas of the technology market, the development of new AI computing hardware such as NVIDIA's recently released A100 Tensor Core GPU coupled with GPU-optimized software will reduce the cost of hardware over time. The A100 , for example, provides a six-fold speed increase of NVIDIA's V100 accelerator in training the popular BERT (Bidirectional Encoder Representations from Transformers) NLP model. For both on-premises hardware and cloud-born AI acceleration services, proper discipline in provisioning and active monitoring must be employed. Fortunately, most ML Ops solutions come equipped to manage AI acceleration hardware, often directly for those running on top of public cloud platforms such as Microsoft Azure or Google Cloud.

For some applications, AI solutions can be handled completely in the cloud, with data from an organization being transferred to a remote third party that handles the development, processing, and output for ML workflows. But for many applications and use cases where security, privacy, and performance are paramount, organizations will need to invest in some degree of AI infrastructure on premises or at the edge of corporate boundaries. For these customers, an investment in a cloud-native infrastructure platform such as HPE's GreenLake platform becomes imperative. Such platforms can bring cloud-native technologies and practices down from the public cloud to run at scale next to on-premises data.

Organizations that want to own and manage their own AI infrastructure will need to select a deployment-agnostic ML Ops platform upon which to build and deploy AI solutions. Thankfully, owing to the rise of containerization and the industry-wide commitment to a Kubernetes-based containerized platform such as HPE Ezmeral Container Platform or Red Hat OpenShift, most MLOps solutions can run on nearly any hardware footprint, allowing companies to easily segment functionality according to need.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Case study examples

While AI utilization is far from ubiquitous, there is significant activity occurring within organizations. By leveraging lessons learned during early pilot programs, incorporating internal domain and process knowledge, and then integrating external hardware, software, and AI resources, organizations are actively deploying highly impactful AI solutions at scale across a broad swath of use cases. The following examples taken from three key industries, illustrate how ML technologies and techniques are being deployed successfully, providing organizations with tangible, real-world results.

Financial services: Identifying both threats and opportunities

The financial services market is awash in data: transaction data used for inter-institution and market-making activities, product sales, customer data, and operational data used for managing the day-to-day operations of an institution, managing security, and marketing operations. The ability to harness this data, identify patterns, and create new efficiencies is a prime driver of AI technology in the financial services industry. A recent Omdia study (ICT Enterprise Insights 2019/20 – Global: IoT, Cloud, AI, and 5G) finds that investment in ML has reached ubiquity with approximately 20% of companies surveyed having actively deployed ML. Looking forward, nearly 60% of those surveyed indicated that they were either trialing or planning ML deployments.

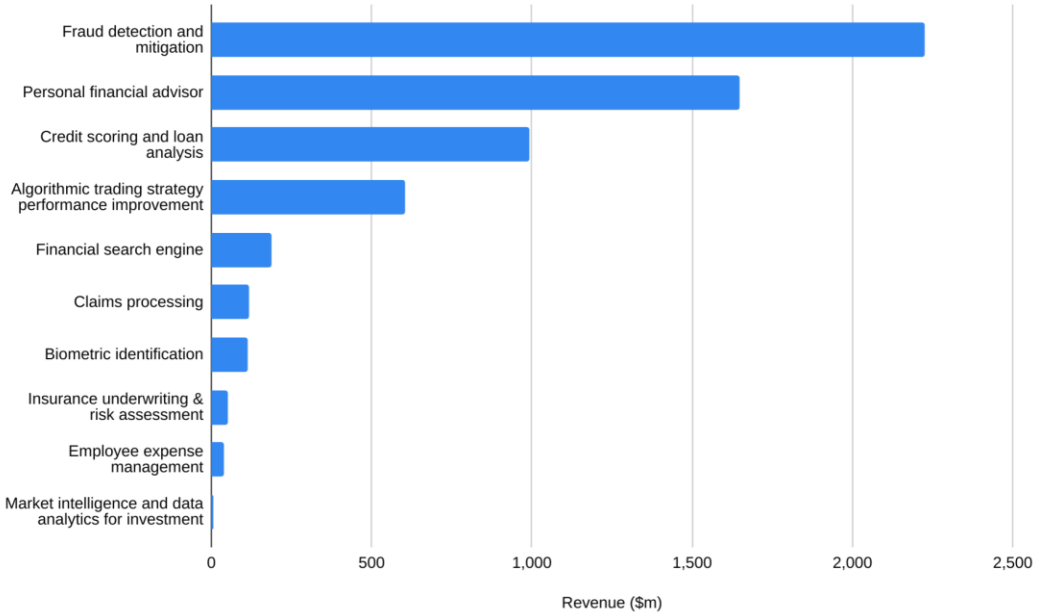
This ability to leverage AI as a means of harnessing data has become paramount for the financial services market, as it is undergoing an incredibly unique transformation. Global markets are actively digitizing business processes. Competitive pressures are pushing customer fees downward. And the barrier of entry and exit for customers is diminishing. Conversely, the financial industry, as an early adopter of AI, has a unique opportunity at hand. The emergence of on-premises, cloud-native technologies and highly scalable AI acceleration hardware aligns perfectly with the ubiquity and scale of data financial services forms need in order to successfully reimagine how they interact with and service their customers.

Using ML predictive models that look across a massive array of interrelated measures, banks can better understand and meet customer needs. This allows them to not just limit but more accurately focus churn rates, thereby retaining higher lifetime value for the bank and higher customer satisfaction for the customer. Building on the same data, predictive models can anticipate which products a given customer will need in the future, thereby improving cost to profit ratios for cross-selling and upselling programs. Conversely, ML predictive models capable of uncovering and limiting exposure to risk have grown and will continue to grow in importance. Banks, for example, can identify cyber threats, track and document fraudulent customer behavior, and better predict risk for

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

new products. The top three use cases as measured by Omdia are fraud detection and mitigation, personal financial advisor services, and credit scoring and loan analysis (see figure 7), reflecting this dual nature of investment with companies emphasizing both internal- and customer-focused outcomes.

Figure 7: Cumulative ML revenue by financial use case, world markets: 2019–25



© 2020 Omdia

Source: Omdia

Retail: thriving amid market disruption

The retail vertical is focused on both e-tailers, in terms of using AI to further connect with customers, and brick-and-mortar retailers, which are using AI to track offline behavior. AI is being applied to both customer-facing applications and behind the scenes to make the retail experience more efficient and personalized.

The fashion industry is increasingly turning to AI to help it cope with changing trends and disparate customer preferences. Fashion designers, apparel and fashion accessory manufacturers, and fashion industry consultants are using AI to connect browsing and shopping history data, sales figures, and even product features to ensure future products meet the ever-changing desires of their customers. Companies are also using AI to offer new and innovative experiences in the store, such as virtual fitting rooms. According to a recent Omdia survey (ICT Enterprise Insights 2019/20 – Global: IoT,

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

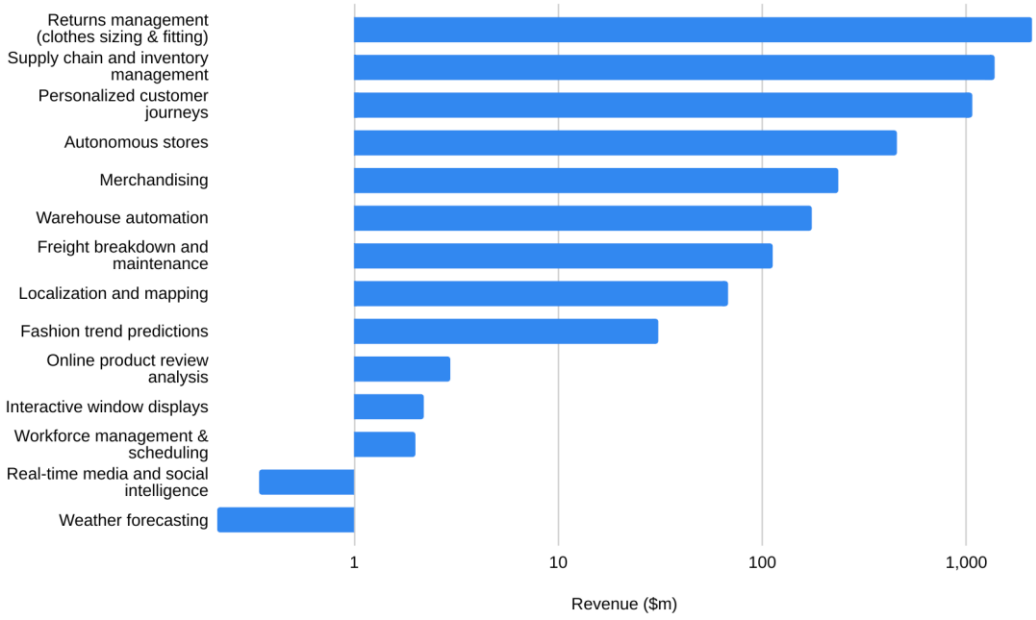
Cloud, AI, and 5G), approximately 36% of retailers have deployed or are trialing image recognition software. For this use case, which is extremely resource hungry in terms of the hardware processing power necessary, retailers need to invest in AI acceleration hardware in order to build and execute extremely large models both affordably and at scale.

Even before the COVID-19 crisis, the retail industry was facing major headwinds spawned by disruptors like Amazon, Alibaba, and Walmart, as well as hundreds of hungry startups, that have built lean, analytics-driven organizations based on scale and efficiency. The goal of these organizations is to drive top-line revenue and reduce operating costs. Overall, revenue and margins are under pressure as these more efficient and scalable disruptors draw more buyers and sales with sharper pricing, personalized customer journeys, and finely tuned assortments. Meanwhile, they are driving down costs through efficiencies in supply chain and inventory management. The pandemic brought another layer of financial pressure as well as supply chain and labor disruptions, some of which AI can help manage.

In response, the retail industry is looking for AI solutions to help leverage contextual personal data and predictive analytics to serve up personalized recommendations, promotions, and marketing. They are also seeking to improve search functionality through sharper meta-tagging and visual search techniques. And retailers are looking to apply predictive analytics and data science to significantly improve price optimization and demand forecasting/inventory management. These needs align with Omdia research findings that show a distinct emphasis on returns management (clothes sizing & fitting), supply chain and inventory management, and personalized customer journeys (see figure 8). In the aftermath of COVID-19, which has curtailed in-person shopping, this ability to meet the customer virtually will serve as a prime differentiator among retailers. Moreover, early notions of how to utilize AI within retail have changed. The inclusion of social media and ancillary information data such as weather, for example, does not show any sign of growing compared with all other use cases tracked by Omdia. This change serves as an indicator of maturity among retailers, which are now focusing their AI efforts on outcomes with direct impact on revenue.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Figure 8: Cumulative ML revenue by Retail use case, world markets: 2019–25



© 2020 Omdia

Source: Omdia

Manufacturing: achieving true agility

The manufacturing industry has embraced automation, with the most notable example being automobile manufacturing. AI is a natural addition to this industry, which includes large and small makers of durable and nondurable goods. AI use cases in this vertical include predictive maintenance of manufacturing equipment, improving product yields, assessing and improving quality, and system-wide use cases, such as enabling the use of digital twins of the manufacturing process or products themselves.

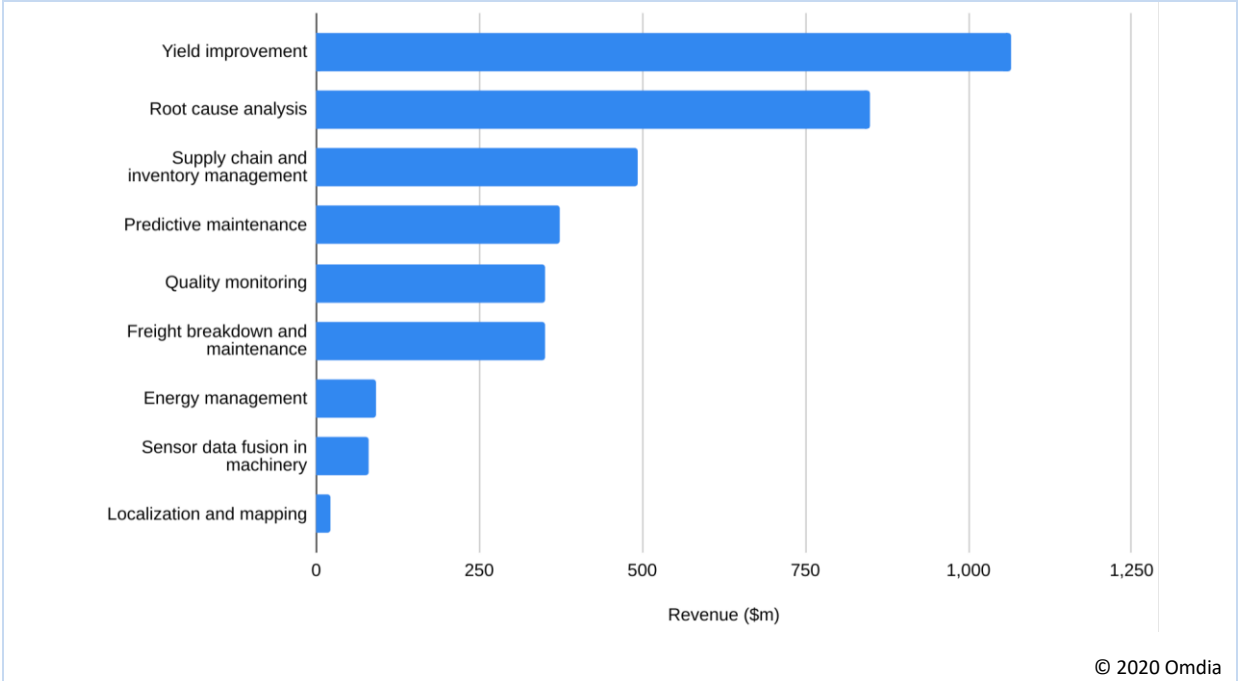
The manufacturing marketplace, as an early adopter of AI and automation in particular, is well-positioned to thrive in the wake of the COVID-19 pandemic, especially those businesses already heavily invested in intelligent robotic process automation (RPA). Companies with a heavy reliance on human labor, however, have struggled and will continue to do so with unplanned closures and staff reductions. Wide-ranging market shifts caused by the pandemic have also presented manufacturers with an interesting challenge. Supply chain disruptions and outright market-shutdowns, have forced many manufacturers to completely retool their infrastructure to accommodate a new line of business or support demand for critical infrastructure. For example, many craft brewers switched from brewing beer to producing hand sanitizer. Likewise, many metal and parts-manufacturers redirected their facilities toward the production of medical equipment.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

With or without disruption, the manufacturing marketplace is reaping the rewards of AI across a wide array of use cases. Using live and historical equipment operating parameters (temperature, revolutions, etc.), predictive models can identify and recommend the replacement of equipment that is likely to fail in the future. But putting this sort of automation into practice at scale demands a careful investment in both specialized video measurement equipment, AI acceleration hardware, and edge processing gear that together can operate harmoniously in a decentralized manner within what is often an inhospitable environment.

By instrumenting both equipment and processes, manufacturers are using ML modeling to re-organize and optimize production in a way that is both responsive to current demand and conscious of future change. The end result is a manufacturing process that is at once agile and resilient. The top three ML use cases identified by Omdia, center on this notion with practitioners investing heavily in ML for yield improvements, root cause analysis, as well as supply chain and inventory management (see figure 9.)

Figure 9: Cumulative ML revenue by manufacturing use case, world markets: 2019–25



Source: Omdia

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Outlook and recommendations

Across all vertical markets and horizontal use cases, enterprise organizations have fully embraced AI as a proven means of driving new business opportunities, realizing financial and operational efficiencies, and maintaining business stability. Whether it's driving business value within core business or tackling departmental needs, the use of AI within the enterprise has reached a point of maturity where the overall value proposition of AI no dominates budgeting and purchasing discussions. Rather, the conversation surrounding AI has become more pragmatic in nature. Enterprise AI practitioners now expect to move rapidly and affordably from AI experimentation to AI in operation. However, doing so at scale presents a number of significant challenges that are procedural, technological and organizational in nature.

Within the AI technology market, the dominant response to this challenge borrows heavily from the software IT operations and software development markets. Plying the popular software development shift toward DevOps as a foundation, many AI technology and service providers have responded to this need by creating ML Ops tools and platforms capable of heavily automating the ML development, deployment and management lifecycle. These solutions, while not yet fully mature, are readily available from a wide range of providers and can help companies more effectively walk the path between AI experimentation and operational mastery as a critical business substrate and key source of market differentiation.

For companies able to crack this code of doing AI at scale, the reward will be well worth any initial investment in ML Ops platforms. But this investment must extend well beyond the boundaries of ML models and algorithms to encompass underlying AI acceleration infrastructure and cloud-native platform resources. Such holistic AI mastery will lead to a widespread adoption of AI within a company, which in turn will drive value exponentially with faster ML model development and deployment, more accurate and more trustworthy model results, a higher degree of technology re-use, and more time to focus on increasingly advanced use cases.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Appendix

Methodology

This report was written drawing on briefings, customer events, and industry events with decision-makers, technology and IT services vendors, and end users across a number of geographies. This is combined with desk research and Omdia's research report, Omdia AI Market Maturity, 2020.

Author

Bradley Shimmin
Chief Analyst, AI platforms, analytics and data management
askananalyst@omdia.com

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Get in touch

www.omdia.com
askananalyst@omdia.com

Omdia consulting

Omdia is a market-leading data, research, and consulting business focused on helping digital service providers, technology companies, and enterprise decision-makers thrive in the connected digital economy. Through our global base of analysts, we offer expert analysis and strategic insight across the IT, telecoms, and media industries.

We create business advantage for our customers by providing actionable insight to support business planning, product development, and go-to-market initiatives.

Our unique combination of authoritative data, market analysis, and vertical industry expertise is designed to empower decision-making, helping our clients profit from new technologies and capitalize on evolving business models.

Omdia is part of Informa Tech, a B2B information services business serving the technology, media, and telecoms sector. The Informa group is listed on the London Stock Exchange.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help your company identify future trends and opportunities.

About HPE

Hewlett Packard Enterprise is the global edge-to-cloud platform as-a-service company that helps organizations accelerate outcomes by unlocking value from all of their data, everywhere. Built In decades of reimagining the future and innovating to advance the way we live and work,, HPE delivers unique, open and intelligent technology solutions, with a consistent experience across all clouds and edges, to help customers develop new business models, engage in new ways, and increase operational performance. For more information, visit: www.hpe.com.

About NVIDIA

NVIDIA'S invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots, and self-driving cars that can perceive and understand the world. For more information, visit: www.nvidia.com

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Copyright notice and disclaimer

The Omdia research, data and information referenced herein (the “Omdia Materials") are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together “Informa Tech”) and represent data, research, opinions or viewpoints published by Informa Tech, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness or correctness of the information, opinions and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees and agents, disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial or other decisions based on or made in reliance of the Omdia Materials.

© 2020 Omdia. All rights reserved. Unauthorized reproduction prohibited.