Machine Learning Visionary

Quarterly ML Trend Compendium

Partners Report

2022 Q1

Table of Contents

Authors of 2022 Q1 edition:
Aramide Kehinde, Dylan Tong, Kris Skrinak, Kristof Schum, Mario Bourgoin, Victor Rojo, Zhang Qiong

Introduction: The ML Visionary Report

~~ what ~~

Welcome to the first edition of the AWS Machine Learning (ML) Visionaries Report. This report is an initiative of the Machine Learning Horizontal Technologies team, part of the AWS Partner Organization.

The ML Visionaries Report tracks, selects, collates an distributes horizontal technology capabilities enabled by machine learning in areas the AWS ML partner team expects to be transformative in a 1-3 years horizon.

~~ why ~~

The purpose of this publication is to share our insights with our partners and to collect their interest, expertise and insights in co-building along the prioritized themes. Therefore, if you and your organization find parts of the report relevant for your business and your future customers, do not hesitate to reach out to your AWS partner development manager and reference this publication for a deeper dive.

~~ when~~

We hope that the highlighted themes will help you and your customers prioritize investment decisions along emerging technologies and capabilities. We will distribute quarterly updates of this report. At each year-end, our ML partner team consolidates trend signals and selects themes for the upcoming year. Then, the first quarter's edition (this Q1 release) covers the topics from an introductory, definition, opportunity and challenges perspective. In our Q2 and Q3 releases, we will dive deeper in each theme with a more technical lens through selected examples. In our final release, we will include all our partner solutions in a white-paper format theme by theme.

In conclusion, if you see partnership opportunity, overlap and synergy between the outlined topics and your capabilities, please reach our team at ml-visionary-partners@amazon.com.

Key Take-Aways from Key Themes

Federated Learning Operations

Federated learning trains on distributed datasets without moving large amounts of raw data. Federated learning has benefits of preserving privacy, simplifying data management, and potentially saving costs. We discuss the key challenges of federated learning and propose to build an end-to-end federated learning operations platform that can span regions, accounts/organizations, and clouds.

ML Ops on Kubernetes

Kubernetes is a leader in the deployment of end-to-end machine learning (ML) services because it provides for the flexible and robust allocation and management of compute resources. Anticipating what would disrupt Kubernetes's position will provide us and our partners with useful insights on the coming changes in the ML compute landscape. Current main long-standing challenges for Kubernetes that affect their use for ML are intra-cluster security, the management of multiple clusters, and the sharing of hardware accelerators.

ML using High Performance Computing

The convergence of NLP and High Performance Computing (HPC) on the cloud brings time-reduction, flexibility, and cost-effective benefits for companies trying to leverage *model-parallel distributed training* techniques in their current workloads. The distributed training of Large Language Models (LLMs) for general and domain-specific purposes is cost prohibitive, but generates a niche for few large companies. We recommend to the readers to explore the use of pre-trained models (PTMs) and/or existing NLP solutions in their projects and products.

Trusted AI

Earning trust with software written by machine learning, presents unique challenges. Outcomes may be worse than the problems you endeavored to solve in the first place. Automated algorithmic decisions can reflect and amplify undesirable patterns. AI will do as the data used to train models reflects. It will do so without respect to the social boundaries which humans instinctively work within. All stakeholders need to communicate clearly about machine learning. Effective communication about Trusted AI begins with knowing the terminology, concepts, and tools.

Decision Intelligence

Decision Intelligence (DI) operationalizes AI in various forms to ask questions about what's likely to happen if you act on a decision today. It empowers users with AI capabilities to facilitate the entire decision-making value chain from insights to automating actions at scale. Among the challenges in DI are designing tools that are suited for business decision-makers to use and operationalize machine learning models that facilitate a broad range of decision-making scenarios that are unique to their business. Furthermore, Trusted AI capabilities need to be ingrained into DI tools. If the toolchain is more complex than mainstream decision support systems and decision makers don't trust or misuse the models, DI tools will have limited success like the current state of predictive analytics.

Generative AI

Generative AI autonomously explore a problem space and create solutions with optimized outcomes. They augment the innovation process for human co-creators and enable speed to market, enhanced discovery, and higher quality outcomes. As the Metaverse grows in importance, Generative AI will be instrumental in the automatic and speedy delivery of the compelling experiences that it demands. When coupled with the physics-based methods of photogrammetry, generative algorithms will not only create immersive virtual models of the real world; but they will also use dynamic real world data to create novel worlds in the Metaverse.

Theme 1: Federated Learning Operations

What is Federated Learning?

privacy / prahy-*vuh*-see/ *noun.* the condition of being concealed or hidden.

Federated learning (FL) trains machine learning models on distributed datasets, without sending sensitive data to a central repository. The goal of federated learning is to improve the accuracy of machine learning (ML) models by using more data, while preserving the privacy of distributed datasets. FL increases the amount of data available for training ML models, especially data associated with rare and new events, resulting in a more general ML model.

During FL training, a central FL server collaborates with a group of clients to train a global model. Clients can access local datasets and perform local training. In a FL training round, the Clients train the global model with local data, then provide local models back to the server. The server aggregates the local models into a new global model, then starts a new training round. There may be hundreds of training rounds until the global model converges or until the number of training rounds reaches a threshold. Hence, FL exchanges ML models between the central FL server and clients, without moving training data to a central repository.

Why is Federated Learning growing in importance?

There is a growing number of machine learning use cases in which learning on distributed datasets is a requirement. The machine learning at the edge is an abundant field for federated learning: imagine all the information recorded with camera edge devices or information provided on mobile phones used for training without the need for giving up the privacy of the people behind them. Health care and life sciences for similar reasons are also seeking privacy-focused ML training methods. Moreover, the recipe works across institutions as well: an example is fighting financial crime with better fraud detection capabilities. With the advent of federated learning banks can collaborate with each other and create a richer, centralized machine learning model the improves all customers and all banks in a way that no institution or customer has to give up ownership and privacy of their data.

For privacy preservation, one straightforward approach is **anonymization** where personal identification and sensitive information (e.g., names, credit card numbers) are removed or encrypted. However, linkage attacks are able to re-identify individuals in an anonymized dataset by combining different data sources.

Privacy can be preserved by keeping data locally without sending data out. Google first introduced FL in 2016 to learn a prediction model by keeping all the training data locally on mobile phones. For this type of **cross-device** FL [1, 2], datasets are located at a large number of devices (e.g., mobile phones or

IoTs) with limited and unstable network connections. Thus, the design of cross-device FL needs to consider frequent dropping of FL clients.

When datasets are prohibited from moving out of organizations and datacenter regions, due to data protection regulations, operational challenges (e.g., data duplication, synchronization), or high costs, FL can be used to trains general models on datasets distributed at different organizations and geo-distributed datacenters. In this type of **cross-silo** FL, organizations or datacenters usually have reliable network connections, powerful computing resources, and addressable datasets.

FL has been applied to various vertical industries, such as finance, healthcare and telecommunications, where data is highly sensitive. For instance, as the aforementioned example, FL has been used to train a global model for financial crime detection among multiple financial institution [3]. The global model outperforms models trained with local datasets by 20%. In healthcare, FL has been used to train a global model across 20 healthcare institutions in order to predict future oxygen requirements of symptomatic patients with COVID-19 [4]. The global model improves 38% in average area under the curve (AUC) measured at all FL clients, compared to local models trained with local datasets. FL has also been used for brain tumor segmentation [5]. The global model for brain tumor segmentation performs similarly to the model trained by collecting distributed datasets at a central repository. In telecommunications, FL can be applied to edge computing, wireless spectrum management, and 5G core networks [6].

What are the challenges and opportunities?

We highlight the key challenges of FL as follows. These challenges can be addressed by algorithms running at FL servers and clients in a common FL architecture.

- **Data heterogeneity**: data heterogeneity among FL clients is common and may be due to particular geographic locations, organizations, or time windows. Data heterogeneity impacts the accuracy of global models, leading to more training iterations and longer training time. Many solutions have been proposed to mitigate the impact of data heterogeneity, such as optimization algorithms [1], partial data sharing among clients [7], and domain adaptation [8].
- **Privacy preservation:** Local and global models may leak private information via reverse engineering. Many privacy preservation approaches have been proposed for FL. A secure aggregation approach [9] can be used to preserve the privacy of local models exchanged between FL servers and clients. Local and global differential privacy approaches [10] bound the privacy loss by adding noise to local or global models, which provides a controlled trade-off between privacy and model accuracy. Depending on the privacy requirements, combinations of different privacy preservation approaches can be used.
- **Federated analytics:** Federated analytics provides statistical measurements of distributed datasets without violating privacy requirements. Federated analytics are very

important not only for data analysis across distributed datasets before training, such as private histogram computation [11], but also for model monitoring at inference, for example, balancing the model interpretability and data privacy [12].

Despite the above challenges of FL algorithms, it is critical to **build a secure architecture that provides end-to-end FL operations**. The architecture must coordinate FL servers and clients for FL model building, training, and deployment, including continuous integration/continuous development (CI/CD) among clients, traceability, authentication of servers and clients, and access control of FL servers and clients. These features are similar to centralized machine learning operations (MLOps), but they are more challenging to implement. This is because that FL involves more parties and they (FL clients) can be on premises or AWS Cloud.

We propose a **F**ederated **L**earning **Arch**itecture on AWS (**FLArch**), which is a cloud-native platform that provides secure end-to-end MLOps for FL using AWS services and supports customized FL algorithms. The objective of FLArch is to simplify initial configurations, model building, training, and deployment for FL.

FLArch on AWS unlocks end-to-end transparency of architectures and operations for FL by leveraging AWS services with proven security and operational excellence. For example, the traceability of FL servers and clients can take advantage of AWS CloudTrail and AWS Model Registry. Secure network connections between FL servers and clients can be supported by AWS services for encryption in-transit. The access control of FL servers and the authentication of FL clients can leverage Amazon Cognito and AWS Identity and Access Management. FL clients can also take advantage of AWS Outposts for local training and inference on premises and at edge locations.

Partnering with AWS

New privacy laws continue to be implemented worldwide, and technology infrastructures are rapidly expanding across multiple regions and extending to network edges. Federated learning can help cloud customers leverage distributed datasets to train more accurate ML models. Federated learning can also preserve data privacy, simplify data management, and potentially save costs, since large amounts of raw data do not need to be moved.

We would like hear from our partners about the challenges experienced in data management when developing AI and ML solutions. We would like to offer our thoughts on federated learning and jointly build federated learning platforms, which can cross regions, accounts/organizations, and cloud providers. We would like to develop go-to-market joint solutions with our partners.

We also would love to work with SI Partners who are capable of solving federated learning use cases and building federated learning blue prints. We're also looking for ISV partners who have ML Ops platforms and plan to offer federated learning as a new capability on their ML Ops platforms. Together, we can deliver unified and fly-wheel solutions for applications in different vertical industries.

Theme 2: ML OPS on Kubernetes

What is ML Ops on Kubernetes?

Orchestrate [awr-k*uh*-streyt] verb. plan and direct a complex undertaking

Kubernetes (K8s) is an open-source container orchestration platform. It is a common approach to the deployment of end-to-end machine learning (ML) services because it provides for the flexible and robust allocation and management of compute resources. At AWS, K8s takes the form of the Elastic Kubernetes Service (EKS), our managed open-source K8s service.

ML Ops is an orchestration of the production, deployment, monitoring, and maintenance of the machine learning (ML) components of applications. This uses a combination of automated tooling and human intervention for the cleaning and featurization of operational data, the training and evaluation of AI models, their deployment and monitoring on infrastructure, and their governance. Kubernetes is not only growing in importance for application modernization, but also gaining traction in the field of ML Ops as an orchestration platform.

Why is ML Ops on K8s growing in importance?

K8s has long been an incumbent leader in running and managing ML workloads on clusters. As such, it profits from network effects, and also from the promise of liberating ML systems from being attached to a particular infrastructure, be it cloud, on-prem, on laptop, at the edge, or hybrid. Anticipating what would disrupt K8s's position will provide us and our partners with useful insights on the coming changes in the ML compute landscape.

Many K8s features support ML. K8s clusters support the manipulation and transformation of the terabytes of data needed for training ML models. By using these clusters, data parallel and model parallel training jobs can also be run on shared resources. With K8s's monitoring and maintenance of pods and its ability to restart failed pods, parallel jobs are more likely to run to completion, saving both time and resources. And by using auto-scaled endpoints, ML models can be robustly served at scale.

Both our Containers and ML teams have identified ML as one of their fastest trending workloads on EKS. Also, in the ML segment, we heard customers asking for clear guidance on when and how K8s can be used for ML.

The 2021 estimated compute revenue for DIY ML on AWS infrastructure is 6x that of ML workloads with AI/ML services (internal source). Although the service teams are working hard to increase their share of the market (198 announcements in 2020, 253 announcements in 2021, and 51 announcements in Q1

2022 according to What's New with Machine Learning?), and while some customers prefer managed services so they can focus on their core competencies, this ratio remains unchanged.

What are the challenges and opportunities?

The complexity of setting up and operating K8s clusters is a common challenge. (For examples of failures, see https://k8s.af.) This complexity is multiplied by adding the use of Spark to the clusters to implement parallel distributed algorithms. This is not a blocker for customers with K8s experience who are looking to add ML-based capabilities to their products. But customers in need of people with K8s experience face a strong demand for them. (K8s job openings: more than 100,000 on LinkedIn, more than 72,000 on Indeed, more than 85,000 on ZipRecruiter.) And these customers may welcome guidance on how to use K8s for ML. (K8s and machine learning job openings: more than 14,000 on LinkedIn, more than 9,000 on Indeed, more than 16,000 on ZipRecruiter.)

Machine Learning and Machine Learning Operations (ML Ops) are rapidly developing areas that are much in demand as companies in all verticals are integrating machine learning technologies in their operations and products. Because ML Ops is about the entire ML model life cycle, from data acquisition and labeling, through the use of data to create models, to the deployment and monitoring of the model in production, multiple offerings exist that only address part of this end-to-end process. The combination of these offerings on K8s is complex, and is a challenge.

The size of ML workloads is growing with the size of data used, and the number of model parameters. This affects the total training time, which affects both the cost of models and limits the speed at which new models can be created. This can be partly addressed by the use of both data parallel and model parallel learning, and this is supported by the use of K8s clusters for ML. K8s features of particular interest here is the scheduling of jobs by their resource needs, and the ability to restart failed pods so that a long ML job is more likely to complete.

Another tool to reduce training time is the use of accelerators such as GPUs. However, the K8s pod scheduler currently allocates node accelerators exclusively to individual -pods. This limits the ability of K8s to promote the sharing of GPU resources between multiple tasks.

Customers may like to use open source frameworks such as K8s because of the perception that otherwise, they will be locked-in to a vendor or cloud. However, the use and configuration of open-source components for secure deployment also poses a challenge, in particular for companies in highly regulated verticals such as healthcare and finance. By offering customers paths to securely using managed open source tooling such as EKS for their ML, we may be anticipating their needs. For example, EKS uses separate networks for the control plane components and nodes as is recommended in the NSA-CISA 2022 report, “Kubernetes Hardening Guide.”

A common approach for addressing the security concerns of multi-user K8s clusters is to provide separate teams with their own clusters. As a consequence, the sharing of infrastructure is decreased, with

both good and bad results. This is also one factor that pushes the rapid rise of systems to manage hundreds of K8s clusters from one pane of glass.

Containerized microservices is a growing approach for deploying applications. K8s is an open source container orchestration platform that is often used to deploy, manage, and monitor microservice-based applications. While serverless applications are a challenge to containers, they are not a challenge to K8s, as is demonstrated by the existence of serverless K8s-based frameworks.

Partnering with AWS

The AWS machine learning teams see the growing number of end-to-end machine learning cycle incorporating Kubernetes. and the AWS Containers teams see the growing number and share of ML workload among Kubernetes implementations. However, what we also hear is the challenges around the complexity our customers are facing. At a growing pace across all verticals the “How?” is heard loud and clear.

That is at the core of the partnership we seek with our consulting and technology partners to bring simplicity in K8s for ML for customers.

Theme 3: Natural Language Processing (NLP) on High Performance Computing (HPC) - NLP on HPC

What is Natural Language Processing on High Performance Computing?

speed /ˈspēd / *verb.* To move, work, or take place faster.

The amount of data generated and ingested each millisecond into enterprise, academic, end-consumers, and public sector environments has fueled the development and evolution of Natural Language Processing (NLP). It has also introduced text-based language models trained with billions of parameters (also known as Large Language Models - LLMs). Despite their shortcomings and functionality, these models have shown extraordinary capabilities (ex., text summarization, chatbots, search, and code generation) with low to zero supervision. LLMs have also outperformed models trained for a specific task or purpose and are in use by day-to-day search services.

LLMs have also raised the need for techniques and infrastructure that *speed up* the training of NLP models with tens or hundreds of billions of parameters. Today, researchers distribute model training across HPC nodes with multiple Graphical Process Units (GPUs) to accelerate the LLMs training. They also use different techniques such as *data parallelism (process mini-batches of data in nodes that contain a replica of the model), model parallelism (split a model across multiple processing nodes that contain subset of the model), pipeline parallelism (break the layers of a model into parallel-process stages)* or *model-parallel distributed training (partition a model over a cluster of multiple nodes).* To overcome the challenges posed by computing parallelism and distributed training (for example, latency, synchronization of data, or under-utilization of GPUs), researchers also leverage network devices that provide high-performance inter-node communication, including a high performance file system, and distributed training libraries that support parallelism strategies, inter-GPU and inter-node communications.

Why is NLP on HPC important?

To quickly hit better outcomes, meet higher efficiency and quality bars in today's world, and differentiate from their competitors, customers need to accelerate and improve their product development lifecycles and customer experiences or to provide faster and more accurate text predictions, enhanced speech transcription analysis, or better insights from historical documents. As a result, the demand for LLMs that help customers achieve these goals is rising: companies like NVIDIA, Cohere, A21 Labs, and Primer have started to train horizontal LLMs that companies can apply to their industry or domain. In addition, other startups began to create a new generation of platforms that help synthesize data from non-structured

sources, assist customers in effective writing, perform effective language translation, and help improve sales performance. Finally, other startups are exploring conversational AI interfaces for self-service, contact center agent assistance, patients companionship, content moderation automation, or the development of new RNA-based therapeutics.

Historically, industries like automotive, healthcare, insurance, oil and energy, and financial services have benefited from HPC. They have reported time reduction to train models from days to hours, reduced costs for training models, faster decisions to improve patient outcomes, improved and faster insights from historical data, reduced times in the delivery of improved product designs, improved delivery of more accurate risk management insights, and better and quicker radiology diagnosis. In the NLP field, HPC is a strategic asset for companies looking to explore the development of LLM-based solutions, or build functionality using pre-trained models (PTM).

What are the challenges with NLP on HPC?

Although the benefits of HPC are clear, moving existing HPC infrastructure from the ground to the cloud is one of the companies' main challenges. In addition, the complexity of migrating and transforming to the cloud existing businesses with their models and workloads running on traditional HPC could also be intimidating. Resiliency, business continuity, and disaster recovery plans are other challenges for companies training LLMs on top of cloud-based HPC infrastructure.

Sustainability and financial feasibility are other critical elements when using HPC infrastructure to train LLMs. HPC is an undeniable approach that provides faster results, but also it is an energy and cost-intensive infrastructure that needs to be considered by companies in their strategic plans. Finally, the inherent biases identified in LLMs impact societies' social, ethical, and equity dimensions. Suppose companies base their businesses on this type of model: they need to create the initial "rules of use" to drive the correct use of LLMs in real-life production workloads, including the existing or future regulations established by the industry and governments.

Partnering with AWS

In figure 1, we show a map with the areas where AWS partners can find opportunities to leverage Large Language Models. We have identified three areas:

- Partners that provide hardware and software solutions for training LLMs in distributed and high-performance environments
- Partners that train generic and custom LLMs on top of HPC and partners that leverage pre-trained LLMs to build different NLP tasks
- Partners that leverage NLP Tasks to build platforms, solutions and services offerings to AWS customers.

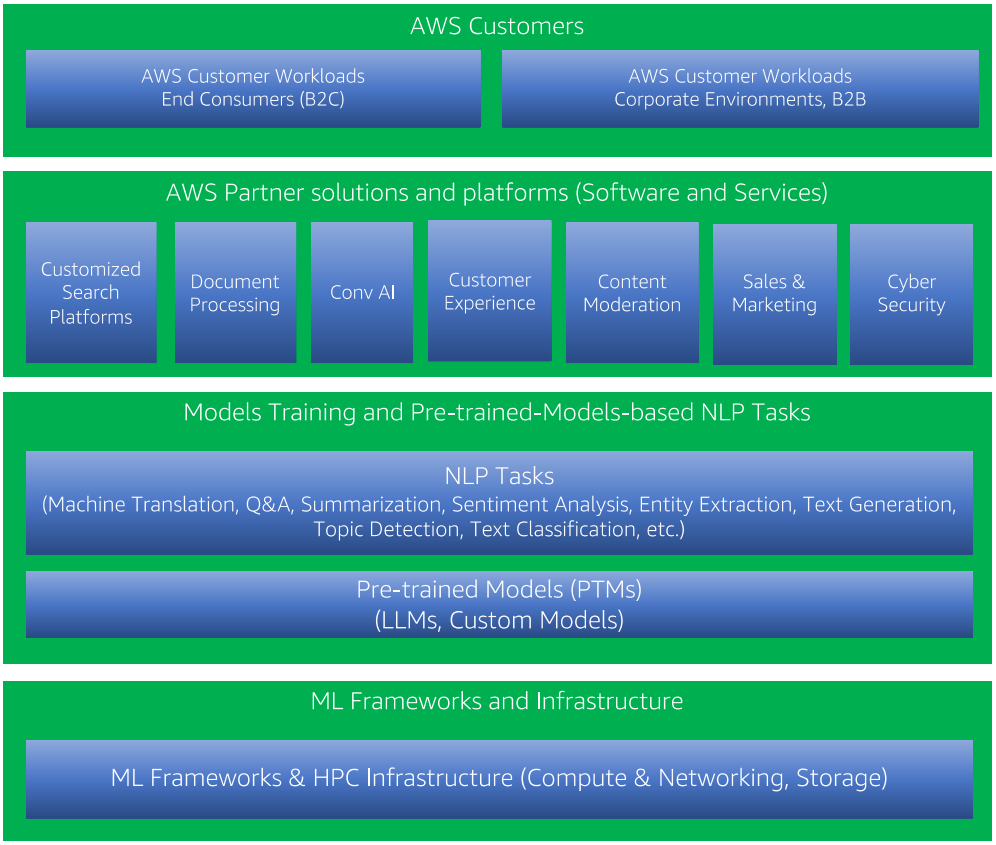


Figure 1 LLMs on HPC Value Chain Analysis

Theme 4: Trusted AI

What is Trusted AI?

trust [truhst] noun. reliance on the integrity, strength, ability, surety, etc., of a person or thing.

Trust is the foundation of harmonious community activity. All successful software products have earned an element of trust that ranges from ensuring secure financial transactions to sharing private moments with photos and video. When it comes to software that's written by data, the essence of all machine learning ("ML"), earning trust takes on new dimensions and responsibilities many of which are not currently well defined. It's not just the code. Data sources for ML may include personally-identifying information, out-of-date or inaccurate entries, poorly-defined sources, and other anomalies that produce inaccurate outcomes.

Trusted AI is a fast-evolving field that has yet to settle on a single name. It may be referred to by other names including and not limited to:

- Trustworthy AI
- Responsible AI
- Human-Centered AI
- Accountable AI
- Explainable / Interpretable AI

Most of these names are interchangeable and generally refer to creating socially responsible outcomes to ML models in production.

Gartner defines "Responsible AI" as an umbrella term for many aspects of making the right business and ethical choices when adopting AI that organizations often address independently [1]. These include business and societal value, risk, trust, transparency, and accountability. Trusted AI extends this definition to include any aspect of successful machine learning deployment where human interests are the principal success driver. Deloitte's Trustworthy AI™ framework requires governance and regulatory compliance throughout the AI lifecycle and requires each phase of the lifecycle to be transparent and explainable, fair and impartial, robust and reliable, respectful of privacy, safe and secure, and responsible and accountable. [2]

The era of designing products that see, hear, and understand what humans and the machines themselves are doing is just beginning. The goal of Trusted AI is to embrace these capabilities by way of guided, human-informed technology. Trusted AI enables people in business to make reliable decisions, frequently automated decisions. They develop clearer strategies and solutions through Trusted AI's combination machine learning with human input and values.

By applying the principles of behavioral science [3] to technology through Trusted AI, developers and product designers are able to tap into user behavior and subconscious patterns. This results in products and services that follow more satisfying, enriching, and addictively rewarding user experiences.

Examples of Trusted AI are so embedded it our daily experience that we rarely notice them. These include unlocking your phone through facial recognition, real-time credit card fraud detection, voice-driven customer service, accurate delivery time prediction, and others.

Why is Trusted AI Important?

Trusting the output of models has a lot to do with why the models were invented in the first place. What business problem are you setting out to solve? What are the expected outcomes? If these are not well known before you start it's easy to get lost in the experimental development process. Outcomes may not only be wrong: their unintended side effects may be worse than the problems you were trying to solve in the first place.

The resurgence of ML in enterprise software is a direct outcome of innovations in deep learning. That is, software-driven neural networks that form the basis of computer vision and natural language processing. It's what the excitement of ML is all about.

Statistical methods and classic machine learning are interpretable: to understand how their predictions are made you just look at the code. It's clear and meaningful. Often you can export the weights output from these models and use them explicitly with the factors that create their predictions. No code. It's clear. You don't need blind trust.

Here's the problem: ML with deep neural networks must be explained: models are opaque, built upon layers of hidden values. The ability to explain models is essential to understanding their use in the real world. Deep neural nets have grown in popularity exponentially and promise to continue doing so. It's a modern miracle, but it comes at it an expense: we can't always explain how they come to their predictions.

Advancements have been made in neural net explainabilty. We can correlate input features with the output predictions to glean insights on what's happening inside. It's essential that executives, policy makers, and a range of stakeholders understand the fundamentals of Trusted AI to maximize opportunities and minimize risk.

Alternatively, Trusted AI is defined by its failures. In Kathy O'Neill's book, "Weapons of Math Destruction", she outlines a case where machine learning algorithms were implemented to score the effectiveness of teachers in a public school district. In practice, the ratings that were produced were slightly more than random scores. Once implemented, and teachers had a chance to review their scores, good teachers left disadvantaged school districts in order to pursue a better career where these methods were not used. Consequently, the opposite outcome was achieved: good teachers left, leaving poor teachers behind.

Machine learning is prediction. The misuse, or misunderstanding, of accuracy metrics can lead to highly deleterious effects. A 2018 test by the American Civil Liberties Union (ACLU) of Amazon's Rekognition image-recognition tool matched 28 photos of members of Congress to criminal mugshots. Amazon recommends law enforcement users set a minimum 95% threshold, and reported that its own test comparing congressional photos to a common data set of 850,000 faces had a 0% false positive rate, when set at the higher threshold. The ACLU has said it used the 80% threshold because it's the program's default.

There are other case studies where racial bias was effectively charged. Upon examination the algorithms were benign. In addition to misuse of accuracy metrics, models can be hacked, the data used to train the data may be bad, littered with typos, or there can be simply insufficient data to make valid predictions.

What are the Challenges for Trusted AI?

With traditional programming languages outcomes are deterministic. When a product manager discovers a solution to a problem and writes up a feature specification, it's up to the software developer then to code solutions which are then handed over to quality assurance. QA then determines if the solution meets the specifications. If not, it's sent back to the developer. The developer can then either fix the code or accumulate technical debt. This is the way software has been written for more than fifty years.

With machine learning things are different: data writes the code. Specifically, what happens is you aggregate data around some event which you would like to predict. That data is then fit to an algorithm to produce a model. The actual name of the function that creates the model is fit(): you fit the data to an algorithm. The outcome is a model, specifically model artifacts that are integrated into code. There are significant consequences to this reality. The most important of which is that your outcome is not binary. You come close to meeting the specifications you were given, but not 100%. Not only are all models approximate, they are also obsolete on production. This is true because all models are built with historical data. Upon exposure to the real world they're faced with new challenges.

Unreliable and insufficient data causes poor and sometimes dangerous outcomes, often outcomes that are the exact opposite of our intentions. Amazon SageMaker, and AWS ML Partners, provide tools necessary to detect and circumvent these data issues. Importantly, executives and technical teams both own essential roles in understanding these problems and need to communicate well with each other on a continuous basis.

AI results in "artificial stupidity" if poorly designed, implemented, or adapted without regard to the human social context. Human, organizational, and societal factors are crucial. Decision makers need to be engaged. They don't need to do the Data Scientists job just as Data Scientists don't need to do theirs. However, executives and Product Managers in particular, need to know the terminology and communicate well about the probabilistic nature of ML applications. Opportunities abound, so does risk.

Automated algorithmic decisions can reflect and amplify undesirable patterns in the data they are trained on, because AI will do as it is trained, without a sense that it has gone off the rails by misinterpreting user behavior, or being unaware of social boundaries which humans instinctively work within.

Human behavior, the principal data input to ML algorithms, changes when that code is put into production. The psychological impact of these changes must be considered throughout the processing, especially in production.

When you build predictive models you need comprehensive data auditability. It's essential to track data lineage, versioning as necessary, starting with the raw data sources. In the traditional software development lifecycle only the code is tracked. Tracking data is a major departure for the lifecycle and ML practitioners still don't have their arms around it completely.

Researchers at Amazon are developing a single metric that scores responsibility. Though the AI Responsibility Index is not yet in production customer demand for simple, understandable guidance on trust has led to the release in 2021 of SageMaker Clarify. Clarify detects potential bias during data preparation, in model training, and in deployed models by examining attributes you specify. For instance, you can check for bias related to age in your initial dataset or in your trained model and receive a detailed report that quantifies different types of possible bias. SageMaker Clarify also includes feature importance graphs that help you explain model predictions and produces reports which can be used to support internal presentations or to identify issues with your model that you can take steps to correct.

AWS Clarify is an open-source project that is accessible to any data scientist in the popular python language and integrates easily into Jupyter notebooks, the most common tools for data science.

Trusted AI Best Practices

It doesn't matter which role you own in the ML process, you must stay engaged. This is not a "set it and forget it" operation. Unlike deterministic software development of the past, ML requires the constant care and feeding of data. Models are obsolete at version 1 and must be retrained with new data on exposure to the real world.

- Ensure all stakeholders know the terminology. Decision makers such as C-Class and VPs, Product Managers, Data Scientists, Programmers, Auditors and Operations must know the concepts covered here to communicate effectively about their products and avoid being blindsided. Understand that it's predictive and there's an accuracy measurement involved below which human intervention is required.
- Use interpretable models when possible. Many problems can be solved with traditional statistical techniques rather than machine learning. OpenCV and other classic tools such as scikit-learn frequently achieve the same objectives as deep learning tools and have the

benefit of clear interpretable code. Some results are so clear their coefficients can be listed as footnotes in a Word or PowerPoint document.

- Trust the domain expert, not the dataset. Look for typos, misrepresentations such as out-of-bounds observations listed as 999. Spend time with the experts who may know nothing about the ML development process but have years of intuition built into their judgements which are reflected in the data.
- Default to using ML as a decision-making aid, not the agent. It's rare when a model can run unsupervised. Plan for built-in or at least scheduled human review.
- Be skeptical of your own results and demanding of outcomes. If you get a high outcome on your first run, say 97%, you might think you're done. When you see 100% accuracy you know something is wrong. Engineer more features into the dataset. Seek to augment your data when possible.
- Actively look out for side effects. People react differently when they know a machine is making decisions for them. They game the system. We all do it naturally with voice-controlled customer support systems, search engine optimization, video games, and others. It's just what we do. Expect behavior to change after models are introduced which changes the data that wrote your code. Retrain and adjust for those side effects and be prepared to pull the code entirely when things go wrong.
- Use the tools: SageMaker's GroundTruth, Data Wrangler, Clarify, Augmented AI, Model Monitor, Feature Store, and Pipelines provide unprecedented capacity to reduce risk and achieve optimal outcomes in your applications.

Partnering with AWS

In the coming quarters we'll be documenting how partner solutions address these challenges. We expect a certain set of baseline issues to be addressed by all partners. We also expect highly-differentiated solutions. This information will inform the entire organization on matching customer challenges to partner solutions. It will fill gaps in AWS service capabilities and raise the bar for the entire industry.

We'll also be discovering unique problems with no clear guidance. This will form the basis of bespoke solutions developed in tandem with partners. We'll be providing thought leadership in the form of reference architectures, blogs, videos, and other mechanisms for driving ML on AWS.

ML Competency Consulting Partners

- Accenture
- Cognizant
- Deloitte
- Inawisdom
- Onica by Rackspace Technology
- Provectus
- Quantiphi
- Reply
- Slalom
- SoftServe

- TensorIoT

ML Competency Independent Software Vendors

- Alteryx
- Anodot
- C3 IoT
- Databricks
- Dataiku
- DataRobot
- Domino Data Lab
- Fiddler.ai
- H2O.ai

Theme 5: Decision intelligence

What is Decision Intelligence?

decision [dih-sizh-*uh*n] noun. a conclusion or resolution reached after consideration.

Decision Intelligence (DI) is an emerging field that brings together methodology and advanced technologies to help organizations make better decisions and elevate outcomes. In October 2020, Gartner heralded Decision Intelligence as a top trend in analytics [1] and that "by 2023, more than 33% of large organizations will have analysts practicing decision intelligence including decision modeling." Gartner defines DI as a field that [2]:

"…brings together a number of disciplines, including decision management and decision support. It encompasses applications in the field of complex adaptive systems that bring together multiple traditional and advanced disciplines. It provides a framework to help data and analytics leaders design, compose, model, align, execute, monitor and tune decision models and processes in the context of business outcomes and behavior."

DI seeks to rethink and reinvent how we currently do decision support. Today, many organizations practice business intelligence (BI), which is designed around a data-centric approach to decision support. The approach is oriented around acquiring data to power insights, which inspire action.

The first difference in DI is that it augments this paradigm by embracing decision science methodologies [3] that prescribe a decision-centric approach that works backwards from business objectives and its associated constraints and requirements. Decision science methods also account for multi-party decision processes and considers different decision-making models appropriate for the situation.

For instance, in the webinar, Decision Science for Data Scientists, Carl Spetzler, CEO of the Strategic Decisions Group (SDG) [4] prescribes different approaches appropriate for the degree of organizational versus analytical complexity.

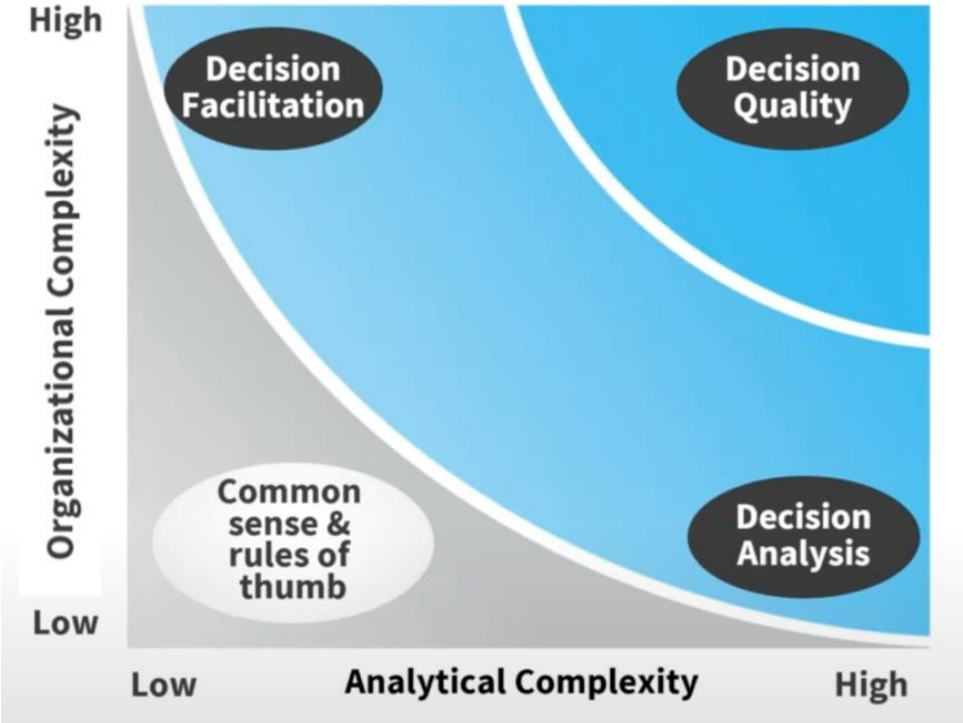


Exhibit. 1 Decision methodology: Organizational versus Analytical Complexity *Source: Strategic Decision Group*

The second difference is that DI raises the bar in analytics capability. Most companies that have a BI practice are typically limited to "descriptive analytics" that are only designed to answer questions about "what happed in the past?" Gartner reports [5] that "87 percent of organizations have low business intelligence (BI) and analytics maturity." Only a small number of companies are empowered with predictive and prescriptive insights that allow them to answer "what is likely to happen" and "what actions should I take?" TDWI describes this problem as "The Chasm" in their Analytics Maturity Model [6].

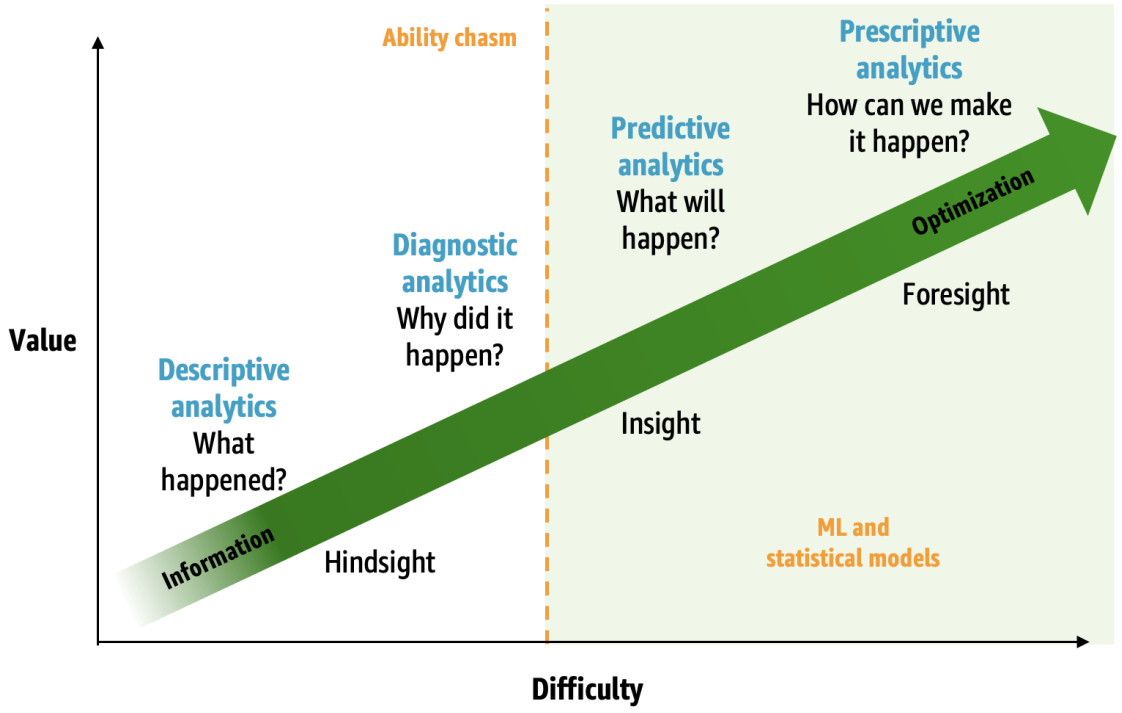


Exhibit 2. Value vs difficulty of analytics implementations Source: Adapted from Gartner's Analytics Maturity Model

So, where predictive and prescriptive analytics is aspirational in BI, the DI charter requires them for high decision quality. DI calls for advanced analytics to enable you to answer these types of questions:

Category	Usage	Question Type	Example
Predictive	Enables higher likelihood of successful outcomes and mitigates biases from human dominant decision processes.	What is likely to happen?	Is this customer likely to churn?
Prescriptive	Enables optimized decision making at scale by prescribing high quality direction to your workforce.	What should I do?	Reduce your pricing by 5% because it will reduce churn by 10% and optimize profit.
Counterfactual	Allows the decision maker to explore options and understand the likely outcomes, trade-offs and risks.	What if I did this, what will be the likely outcome?	What if this customer had a support plan, will they still be likely to churn?

Causal	Assists the decision maker with identifying cause versus correlation to ensure the right corrective action is made. Uses emerging Casuality AI algorithms that go beyond the capability of human-driven diagnostics and inspection.	What is the likely cause of this outcome?	Why did this customer churn?
Explainable	Provides decision makers with insights into opaque models with the intent to improve their trust in the model.	Why is this model predicting this outcome?	What characteristics of this customer are influencing the model to predict churn?
Responsible	Provides the decision maker with information on how the model accounts for sensitive attributes. The intent is to facilitate trust in the model and help the decision maker judge when the model can be safely used.	Does this decision model present bias risks?	Are my benefits programs likely rejecting individuals based on their demographic?
Scenario Simulation	Accelerates complex decision-making situations by automatically simulating different scenarios for the decision maker to consider.	What are my best options?	What are my top 3 options for reducing customer churn, and what is the cost-benefit and risk analysis between options?

Exhibit 3. Types of Advanced Analytics to power the Decision Intelligence Toolchain

So, for example, a Customer 360 solution may provide you with a single view of all your customer's historic data, but what if you had a goal to improve the quality of your customer engagements—how would you decide which initiatives to prioritize? You might aggregate engagement metrics and capture a snapshot of recent and seasonal customer trends to get a sense of what's working and what isn't to provide you with some intuition on what to do next, but there will be a lot of uncertainty in what is the best course of action.

A DI solution could instead work backwards from goals like reducing customer churn and empower business users to build machine learning (ML) models that are optimized for this purpose. These models go beyond human limits by uncovering latent patterns on the entire data corpuses that learn how to optimally predict outcomes—in this case, predicting customer churn. Your decision makers can use these models to ask the aforementioned questions and gain more certainty that effective decisions are made.

Third difference for DI is that it goes beyond insights. When is a decision a decision? Is it when you decide to commit to an action, or is it when you actually execute on it? Decision professionals say it is the latter. Thus, DI needs to go beyond insights and automate action. It should automate business processes by tracking actions, outcomes and collect feedback. The process should empower continuous improvement by tracking actuals versus the expected outcomes to improve future decisioning.

DI is centered on AI. Cassie Kozyrkov, Chief Decision Scientist at Google, says "Decision intelligence is the discipline of turning information into better actions at any scale," and AI as a means to decision making at scale [7].

ML models and algorithms power the predictive insights that help you answer the aforementioned questions—these are questions that human decision-makers can't answer effectively due to large data volumes or latent patterns. Additionally, AI-assisted decisions facilitated by high-quality ML models also help you mitigate human biases that could sabotage outcomes. By using ML enhanced decision models, you will be able to act more quickly and efficiently at scale. You can streamline business process workflows with minimum human intervention through Intelligent Automation—the use of ML-powered Robotic Process Automation (RPA) and Business Process Management (BPM).

Why is Decision Intelligence growing in importance?

Decision Intelligence aims to elevate customer value by disrupting the existing decision support and automation markets using AI. These markets have significant customer demand. For instance, in this 2020 report [8], Gartner estimated that the BI market was $24.8B growing at a rate of 10.4%. In this 2021, Gartner estimated [9] that the RPA market grew 22.6% to $13.8B.

DI benefits customers by uplifting their ROI on decision support and automation investments by using AI to improve decision quality—mitigating biases and errors—and enabling execution at greater speed and scale.

According to Gartner's AI Hype Cycle [10], Decision Intelligence is among the top emerging themes in AI set to mature over the next 2-5 years and that "by 2023, more than 33% of large organizations will have analysts practicing decision intelligence including decision modeling." DI will cross and bridge existing technology domains some of which are mature and ripe for disruption and others are among the highest-growth tech domains. These domains included advanced analytics, specialized ML (e.g. Causal AI, Graph ML, AutoML), Intelligent Automation and Knowledge Graphs.

What are the challenges?

The biggest challenge to DI will be translating advanced machine learning and analytics tooling into something that is practical for business users.

A lot of required capabilities for DI already exist today in some form, but solutions involve large investments in lengthy projects that require scarce experts to build, maintain and operate highly customized implementations. What TDWI calls "The Chasm" in their Analytics Maturity Model still exists [11]. Among the problems are organizations inability to hire or afford enough data scientists and engineers required to deliver on business expectations—this skill gap challenge is affirmed by numerous studies include those from Gartner [12], Deloitte [13], and McKinsey & Company [14].

The core users of decision support systems today like Tableau, Qlik and Domo are business users. Business analysts storyboard reports and dashboards and business decision makers consume them. What capability they can't easily use will have limited success. Business users with no data science expertise need to be empowered with some self-service model building capabilities. Time-to-value needs to be on the scale of days to weeks and not months to years.

Furthermore, DI technology will need to address various human factors. Decision makers are unlikely to trust models they can't understand especially if they contradict their unknowingly biased inclinations. Trusted AI capabilities need to be further improved and prioritized as first-class features within DI products.

Lastly, decision support is a mature industry and it's often difficult to motivate customers to take action when they believe what they have is good enough. Disruption requires "10X Transformations". Emerging capabilities like Causal AI and Graph ML are among the potential catalysts. These technologies can empower business with root-cause-analysis within complex systems such as supply chain networks and predict how change effects participants and cascades throughout the network. Nonetheless, these technologies need to be further developed and made consumable by mainstream business.

Partnering with AWS

The time is now to act on this market opportunity, and the largest share of cloud customers are sitting on a gold mine of data in AWS. Customers will expect a broad set of capabilities within a DI toolchain, and partners will have the opportunity to provide options and fill-in emerging capability gaps within AWS.

Today, the AI for Data Analytics (AIDA) [15] ecosystem represents the beginnings of DI on AWS. Currently, there are three core categories:

- AWS AutoML-enabled BI and Database partners designed to empower business experts to craft custom predictive analytics solutions. Partners in this category include Qlik, Domo, Tableau and Snowflake.
- AWS ML-enabled business applications that provide turn-key predictive analytics for specialized use cases. Partners in this category include Anaplan, Causality Link and Amplitude.

- AWS ML-enabled Intelligent Automation partners designed to provide low-code experiences for automating business workflows that require minimal human intervention. Partners in this category include Pega Systems and Workato.

AWS will continue to invest in marketing for our AIDA ecosystem and drive co-sell motions. Our doors are open to partners who believe in a "better together" DI story. We're looking for partners that can help solve the aforementioned challenges and deliver capabilities in advanced analytics, specialized ML (e.g. Causal AI, Graph ML, AutoML), Intelligent Automation and Knowledge Graphs.

The vision is to jointly build a Decision Intelligence tech stack that will empower customers and consulting partners to deliver specialized DI solutions with a low time-to-value. Together, we can deliver a holistic solution through a better together story that reduces friction by providing customers and partners with choice.

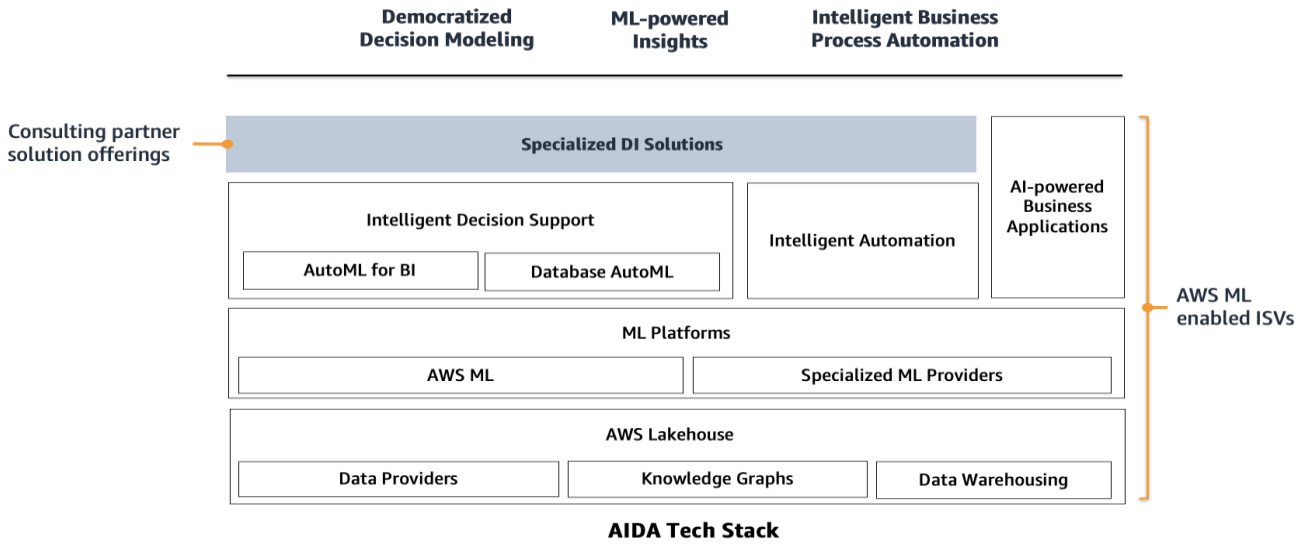


Exhibit 4. A joint technology stack for powering Decision Intelligence solutions

Theme 6: Generative AI

What is Generative AI?

To capture the promise of the augmented age, businesses are investing in systems that magnify humans' capability to think and create. In collaboration with human co-creators, these systems autonomously explore infinite options in a problem's solution space while adhering to constraints. What underpins this technology is Generative AI.

While much of Artificial Intelligence makes predictions based on interpreting existing data, Generative AI algorithms use existing data comprised of audio files, text, video, and images to create novel artifacts. The value of Generative AI is in its ability to create new artifacts with improved outcomes. For example, a generative algorithm can propose a new design for a product with increased durability and decreased weight. Generative AI can also automate repetitive tasks in the development process which allows human co-creators to focus on more value-added work.

Why is Generative AI important?

In recent years, synthetic media has gained notoriety as a primary example of what Generative AI can do. In addition to synthetic media, Generative AI capabilities lend themselves well to a dynamic set of business challenges across industries. While their context of use can vary, teams that adopt Generative AI are able to gain the following benefits:

1. **Speed to market:** With generative AI systems in place, teams can eliminate time intensive or cost prohibitive tasks in their process, which accelerates ideation to launch of a new product or feature
2. I**mproved content or product monetization:** For digital content (including product displays), Generative AI can automatically create enhanced views for end users, which in turn improves their engagement and influences their purchasing decisions
3. **Enhanced discovery:** For R&D, generative systems can autonomously explore near-infinite design options while adhering to required constraints. In this manner, they enable a divergent ideation process that helps teams to innovate
4. **Higher quality outputs:** Generative AI can build robust training datasets for machine learning systems, which enables more accurate prediction. A key example of this is in facial perception systems that need ethnically diverse facial databases in order to make unbiased predictions.

Here, we focus on four industry use cases where generative AI have enabled these results. These use cases are exemplary of the types of challenges that generative algorithms can solve for.

Data Generation

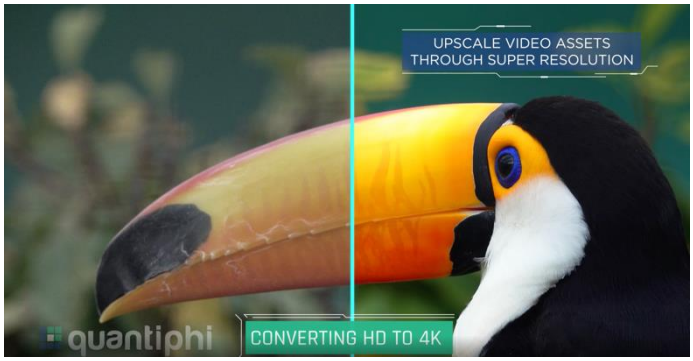
According to Gartner, Generative AI will account for 10% of all data produced by 2025, up from less than 1% today. This is particularly useful for computer vision, a technique that requires machines to learn from large image datasets that are often hard to access. In some cases, the required data is highly specialized to the use case, and therefore not available in the size and quality needed.

To solve this problem, a generative algorithm can learn from a set of sample images and then synthesize new images that teams can use to train their computer vision models. The sample images need not be classified as big data. Rather, it can be a smart-sized dataset that adequately represents the problem space. Generative AI algorithms have learned from large image datasets and are built on deep neural networks. Teams can therefore fine-tune the algorithms to generate new images for a specific use cases using a smaller sample dataset. The result is a computer vision model that operates with higher accuracy.

Media & Entertainment

To date, media production teams have used Generative AI to create content by way of special effects production and content enhancement. These applications eliminate manual and time intensive tasks that are typical in the production workflow. They also provide capabilities that were previously impossible to achieve. This enables media companies to meet the complex demands for content and viewing experiences.

For example, the prevalence of HD devices and OTT content have increased quality standards for media. This creates a need for content owners with archives of old TV shows, films, and documentaries to re-master their inventory in order to monetize it. The remastering process typically requires a painstaking effort to manually enhance content frame by frame. It can take weeks of labor from a multi-disciplinary team to re-master a 90 minute film.



Quantiphi's Super Resolution solution for Media & Entertainment

Generative algorithms for super resolution can alleviate the weight of this process. For example, Quantiphi has developed a super resolution service that generates high-resolution video frames using existing low-resolution frames. The service captures existing content and intelligently generates aligned video frames with higher quality. A test to upscale a 1 minute and 40 second video from 144px to HD took 8 minutes, while upscaling from 720px to 5K took 42 minutes. The result is faster and cost effect content enhancement that accelerates the process of monetizing old content.

Generative AI also provides a solution for globalizing content at scale. The topic of content globalization usually focuses on creating text transcripts of videos and translating them to multiple languages. This is

particularly valuable for viewers with hearing limitations. Synthetic media created using generative algorithms take content globalization one step further by enabling subjects in a video to convincingly "speak" in multiple languages. In their video campaign with David Beckham, Malaria Must Die featured the star speaking in nine languages using recorded testimonies from malaria survivors around the world. To create the synthetic speech, a generative algorithm learned from survivors' testimonials as well as Beckham's image in order to synthesize his voice and facial movements for the 9 languages. The result was a video campaign that had more than 400 million impressions globally.

Retail

For the Retail industry, using Generative AI can enrich customers' online experiences and streamline the resources needed to display product inventories. Consider digital product displays. As assets that influence customers' purchasing decisions, product images should ideally provide multiple views of the product. To provide these views, a vendor typically needs to ship their products for professional 360-degree photography, which is a time consuming process. As of 2020, the price per one product photo could range from $10 to $400 depending on the item. Extrapolate this to all products in a brand's inventory, and this becomes cost prohibitive.

Retailers are applying generative algorithms to this problem. In fashion, for example, they are using generative algorithms to create fashion models that display clothing, skipping the model scouting and photography process all together. With available algorithms, teams can both generate a realistic human frame and present the frame in different poses. Additionally, generative algorithms can animate static photographs of the models, enabling videos that provide customers with a dynamic view of the product inventory.

Engineering Design and Manufacturing

Engineering design is a fundamental process that gets needed products in the hands of society at large. It also cuts across all product-centric industries. Traditionally this process from ideation to product launch is linear, requiring specialized expertise and manual steps at each phase. An error in one phase passes on to the next, with the end results being a product recall and redesign

For this context, AI-driven software enables generative design. Generative design software helps to streamline time intensive engineering design processes with an intuitive interaction between designer or engineer and computer. For example, in ideation, the software accelerates the manual process of searching for ideas and renders them for review amongst global or interdisciplinary team members. This is done while adhering to constraints defined for the system. Generative algorithms can learn to optimize designs for specific engineering and commercial parameters such material weight, cost or even aesthetic qualities.

Photo from Autodesk: Consolidated seat bracket designed using Fusion 360

When coupled with other capabilities such as additive manufacturing or 3D printing, Generative design can enable marginal improvements that create a meaningful competitive advantage for business. Exemplary of this is Autodesk's generative-design software, Fusion 360. To address a design challenge for car seat brackets, General Motors used Fusion 360 to create more than 150 design options that optimized the design for mass. Typically, the team would consider just 2 design options that were bounded by the existing bracket design. Generative designed allowed the team to re-envision the bracket design, resulting in a seat bracket design that was 40% lighter and 20% stronger.

A future in the Metaverse

Viewing Generative AI with a lens of what the future holds is crucial to maximize its potential. As the Metaverse grows, brands are racing to define what their presence in a 3D virtual world centered around social interaction can look like. Given its capacity for autonomous creation, Generative AI can be a key enabler for these brands.

There are many questions a brand needs to answer in order to define their presence in the Metaverse. For example, what type of experiences appropriately extends the brand's persona in the Metaverse? What type of interactions and artifacts do these experiences require? Additionally, how often do these interactions and artifacts need a refresh? And how does a brand rapidly generate them as its persona evolves, or to simply keep the experience interesting? Depending on the brand at hand, it may even need to enable its users to generate their own artifacts in the metaverse.

Considering generative algorithms developed to date, it's clear that they are a bridge to developing the Metaverse further – both from the lens of what the Metaverse itself can be, and what a brand can be in the Metaverse. 3D generative models continue to be a hot topic with large investment from both academia and industry. For example, NVIDIA has used generative neural networks to render 3D environments using existing videos. Likewise, generative language models can be fundamental to enabling compelling social interactions in the Metaverse. This and other technology show that 3D environments coupled with the novel interactions they require can indeed be created on the fly.

When coupled with a technique known as photogrammetry, Generative AI has the potential to create compelling experiences that seamlessly blend the physical and virtual worlds. Photogrammetry uses photography and physics-based methods to obtain credible data about physical objects and environments. Consider then the potential of a generative system that is trained on the enriched visual

and spatial data that photogrammetry provides. This system can not only recreate existing 3D environments in the Metaverse, but it can also use its data to generate new environments that bend the physical laws of the real world. There is opportunity here to define how this works, and to display the art of the possible in order to generate more interest and exploration.

What are the challenges with Generative AI?

A challenge to adoption of generative AI lies in the current understanding of what the technology can do. It is still the case that when Generative AI is mentioned, synthetic media is largely what comes to mind. However, synthetic media is a discrete example what Generative AI can do. There needs to be deeper understanding of what Generative AI are capable of and likewise frameworks that help business leaders identify problems to use the technology for.

There are potential ethics issues that come with generative AI, particularly with synthetic media. For example, who can claim ownership of an image or video of a real human being that is generated as synthetic media. Likewise, should the owner of this media have the rights to use it to purport whatever message they desire with the subject in the image? This can lead to increased cases of fraud and identify theft, which synthetic media have already enabled to date.

Additionally, generative AI can exhibit unstable behavior in terms of the quality and consistency of their outcomes. A prime example of this is the webpage, thispersondoesnotexist.com. Hitting refresh on the page generates a realistic human face created by a generative algorithm called styleGAN. While results are largely impressive, there can be results with glitches that trigger the uncanny valley for a human viewer. In this sense, there is a slight control limitation with Generative AI. Nonetheless, there are ongoing efforts to develop mechanisms that stabilize outcomes, especially when targeted toward specific or discrete use cases. Additionally in many cases, business processes enabled by generative systems can and should involve human-in-the-loop decision making to validate and maximize outcomes.

Lastly, there is an inherent cost factor that teams experience when working with generative AI solutions. Generative algorithms are typically deep learning frameworks that require a high amount of computational resources to efficiently process the millions of parameters in training datasets. For visual cases specifically, generative AI requires GPUs to train from visual data as well as generate visual data. This introduces some level of variability to economic factors for businesses that invest in generative systems.

Partnering with AWS

In recent years, leading technology companies have publicized generative AI software or projects targeted towards specific domains. Most recently, Open AI announced DALL.E, a neural network that generates images using text input, in built with image editing capabilities. As these first movers continue to improve upon their current capabilities, they will eventually have the ability to expand their offerings

to multiple industry verticals. This will result in higher adoption of generative technology with the natural evolution being ubiquitous demand for software and applications with generative capabilities embedded.

The present moment is an opportunity for partners to strategically position themselves as Generative AI technology providers and thought leaders. As Generative AI comes to the forefront, customers will need a deeper understanding of what the technology can do, and likewise frameworks to easily embed it in their existing tech stacks. They will need commercialized products that accelerate their adoption of generative AI systems to capture its benefits for their specific industry contexts.

We are looking for partners to jointly explore the described industry opportunities for Generative AI. The goal is to solve for gaps in adoption in each context, and likewise capitalize on market opportunities with solutions that address customers' needs.

APPENDIX A: Federated learning

Reference list for "Federated Learning Operations"

[1] P. Kairouz, et al., "Advances and Open Problems in Federated Learning," *Journal of Foundations and Trends in Machine Learning*, June 2021.

[2] J. Ding, et al., "Federated Learning Challenges and Opportunities: An Outlook," *In Proceedings of the 47th International Conference on Acoustics, Speech, & Signal Processing*, 2022.

[3] T. Suzumura, et al., "Towards Federated Graph Learning for Collaborative Financial Crimes Detection," https://arxiv.org/pdf/1909.12946.pdf
[4] I. Dayan, et al., "Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19," *Journal of Nature Medicine*, September 2021.

[5] W. Li, et al., "Privacy-preserving Federated Brain Tumour Segmentation," https://arxiv.org/pdf/1910.00962.pdf

[6] S. Niknam, et al., "Federated Learning for Wireless Communications: Motivation, Opportunities and Challenges," https://arxiv.org/pdf/1908.06847.pdf

[7] Y. Zhao, et al., "Federated Learning with Non-iid Data," https://arxiv.org/abs/1806.00582

[8] X. Li, et al., "Multi-site fMRI Analysis Using Privacy-preserving Federated Learning and Domain Adaptation: ABIDE Results," https://arxiv.org/abs/2001.05647

[9] K. Bonawitz, et al., "Practical Secure Aggregation for Federated Learning on User-Held Data," https://arxiv.org/pdf/1611.04482.pdf

[10] M. Naseri, et al., "Local and Central Differential Privacy for Robustness and Privacy in Federated Learning," https://arxiv.org/pdf/1710.06963.pdf

[11] A. Bharadwaj, et al., "Sample-and-threshold Differential Privacy: Histograms and Applications," NeurIIPS 2021 Workshop PRIML, 2021.

[12] G. Wang, "Interpret Federated Learning with Shapley Values," https://arxiv.org/ftp/arxiv/papers/1905/1905.04519.pdf

APPENDIX B: ML Ops on EKS

References

NSA-CISA. 2022. "Kubernetes Hardening Guide." Technical Report.

https://media.defense.gov/2021/Aug/03/2002820425/-1/-1/0/CTR_Kubernetes_Hardening_Guidance_1.1_20220315.PDF.

Self-Managed Kubernetes Alternatives

Alternatives to using self-managed Kubernetes clusters include:

- **Container as a Service (CaaS):** manage containers at scale without the complex orchestration capabilities provided by Kubernetes.
  - AWS Fargate/Elastic Container Service (ECS) and Lambda container images
  - Azure Container Instances
  - Google Cloud Run
- **Managed Kubernetes services:** run managed, hosted Kubernetes clusters without the complexity of deploying, upgrading and maintaining Kubernetes
  - Amazon Elastic Kubernetes Service (EKS)
  - Azure Kubernetes Service
  - Google Kubernetes Engine (GKE)
- **PaaS using Kubernetes:** complete cloud computing platforms, which have Kubernetes at their core, but offer simpler operation and built-in capabilities like security and networking.
  - OpenShift Container Platform
  - Rancher
- **Lightweight container orchestrators:** mature orchestrators, that are less complex to use and maintain than Kubernetes.
  - Docker Swarm
  - Nomad

Sources: Aqua. 2022. "10 Kubernetes Alternatives and Why You Need Them." 2022. https://www.aquasec.com/cloud-native-academy/kubernetes-101/kubernetes-alternatives/.

Kubernetes-Specific Challenges

Challenge	Details
Meeting isolation standards	Isolation of applications on the same server is vital for security.
Internal networking challenges	The latency between services needs to be accounted for in large clusters.
Limited visibility into resource utilization	In large clusters with many applications and services running, it is difficult to know what level of resources systems are consuming.
Optimizing resource usage	As the cluster grows in size, cost management and resource optimization become more critical and more complex.
Deployment failovers and downtime mitigation	Large-scale application rollouts must have fail-safes to avoid downtime risk.
Reduced visibility into data flows	As a cluster grows in size and across regions, maintaining visibility into data flows becomes increasingly complex.
Patch management	Applying K8s security patches and updates at scale without interrupting service or impacting performance is an operational challenge.
Access management and tenant isolation	For large clusters, isolation between different teams or customers using the cluster must be configured.
On-prem challenges	Running clusters on your own data center infrastructure adds additional management complexities, including: avoiding network partitions, and upgrading servers

Source: OpsRamp. 2022. "The Challenges With Kubernetes." OpsRamp. 2022. https://www.opsramp.com/why-kubernetes/challenges-with-kubernetes/.

Kubernetes and Serverless

K8s is often associated with containers with good reason. Serverless computing has arisen as a challenge to containers, but this is not a challenge for K8s. The following six projects run serverless applications on K8s clusters.

1. OpenFaas
2. OpenWhisk
3. Kubeless
4. Knative
5. Fission
6. Fn

Source: Whatmuff, Daniel. 2019. "6 Serverless Frameworks on Kubernetes You Need to Know." Appvia.Io. July 19, 2019. https://www.appvia.io/blog/serverless-on-kubernetes.

Improving Kubernetes Security

NSA-CISA. 2022. "Kubernetes Hardening Guide." Technical Report. https://media.defense.gov/2021/Aug/03/2002820425/-1/-1/0/CTR_Kubernetes_Hardening_Guidance_1.1_20220315.PDF.

GPUs on Kubernetes Clusters

The scheduling of GPU-requiring workloads on K8s has been in beta since release 1.10 on March 2018, while the current release of K8s is 1.23 on December 2021. The current API allows the specification of the GPU requirements of a pod's containers by make, either NVIDIA or AMD. There are restrictions on the allocation of GPUs, the current being:

1. Containers and Pods do not share GPUs.
2. It is not possible to request a fraction of a GPU.

Clusters may contain different types of GPUs. Their selection is done through the addition of labels to nodes, and their use for node selection in container specifications.

Although some hardware makers support virtual GPUs to enhance hardware sharing, and there are ways of virtualizing GPUs for use on clusters, it is not clear at the time of this writing when this will be directly supported by the Kubernetes releases.

Sources:

Kubernetes Team. 2022. "Schedule GPUs." Kubernetes. 2022. https://kubernetes.io/docs/tasks/manage-gpus/scheduling-gpus/.

NSA-CISA. 2022. “Kubernetes Hardening Guide.” Technical Report. https://media.defense.gov/2021/Aug/03/2002820425/-1/-1/0/CTR_Kubernetes_Hardening_Guidance_1.1_20220315.PDF.

Murray, Justin. 2022. “How Virtual GPUs Enhance Sharing in Kubernetes for Machine Learning on VMware VSphere | VMware.” The Cloud Platform Tech Zone. March 10, 2022. https://core.vmware.com/blog/how-virtual-gpus-enhance-sharing-kubernetes-machine-learning-vmware-vsphere.

“NVIDIA Quadro Solutions for Multiple VGPUs.” 2022. NVIDIA. 2022. https://www.nvidia.com/en-us/design-visualization/solutions/multi-virtual-gpus/.

Limits to the Size of Kubernetes Clusters

Usually K8s is limited to thousands of nodes, but some are trying to push that.

Amazon's EKS is limited to 13,500 nodes per cluster by default: 30 node groups per region and 450 nodes per node group. However, these are not hard limits.
Source: https://docs.aws.amazon.com/eks/latest/userguide/service-quotas.html

Azure's AKS is limited to 1,000 nodes per cluster. It is not clear whether if this number can be adjusted. Source: https://docs.microsoft.com/bs-latn-ba/azure/aks/quotas-skus-regions

Google's GKE 1.18 is limited to 15,000 nodes per cluster. This is because they can have at most 15,000 connected VMs per VPC. Source: https://cloud.google.com/kubernetes-engine/docs/best-practices/scalability

SUSE Rancher's Fleet architecture claims to be able to support 1 million nodes: "While most K3s edge deployments are less than 10,000 nodes, it is not unreasonable that one could get to 1 million." (K3s is lightweight K8s.) Source: https://www.suse.com/c/rancher_blog/scaling-fleet-and-kubernetes-to-a-million-clusters/

JD.com is reputed to have the largest K8s clusters in the world. "We run Kubernetes and container clusters on roughly tens of thousands of physical bare metal nodes." Source: https://kubernetes.io/case-studies/jd-com/

Open Source ML Ops Tools

See https://mlops.toys for a partial list of open-source ML Ops solutions.

Kubeflow and Airflow are common tools for orchestrating data and ML pipelines.

MLf\Flow is used for experiment tracking, model packaging, and model registry.

ML Ops Sources and Communities

- https://mlops.community
- https://ai-infrastructure.org
- https://ml-ops.org

APPENDIX C: Natural Language Processing (NLP) on High Performance Computing (HPC) - NLP References ON HPC

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021, February 4). *Understanding the capabilities, limitations, and societal impact of large language models*. arXiv.org. Retrieved April 5, 2022, from https://arxiv.org/abs/2102.02503

Language models are few-shot learners - neurips. (n.d.). Retrieved April 6, 2022, from https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Toews, R. (2022, March 30). *A wave of billion-dollar language AI startups is coming*. Forbes. Retrieved April 5, 2022, from https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/?sh=569b58242b14

Simon, J. (n.d.). *Large language models: A new moore's law?* Hugging Face – The AI community building the future. Retrieved April 13, 2022, from https://huggingface.co/blog/large-language-models

APPENDIX D: Trusted AI

References

1. **Create a Responsible AI Strategy**, https://www.gartner.com/en/webinars/4003778/create-a-responsible-ai-strategy
2. **Trustworthy AITM, Bridging the ethics gap surrounding AI**, https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html
3. **Center for Advanced Study in the Behavioral Sciences**, https://casbs.stanford.edu

APPENDIX E: Decision Intelligence

References

1 https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020nalytics

2 https://www.gartner.com/en/information-technology/glossary/decision-intelligence

3 https://www.insead.edu/faculty-research/academic-areas/decision-sciences

4 https://www.youtube.com/watch?v=ICpa19HZoZs

5 https://www.gartner.com/en/newsroom/press-releases/2018-12-06-gartner-data-shows-87-percent-of-organizations-have-low-bi-and-analytics-maturity

6 https://go.tdwi.org/rs/626-EMC-557/images/TDWI_Analytics-Maturity-Model-Assessment-Guide_2020.pdf

7 https://www.youtube.com/watch?v=bCjMhZZYlP4

8 https://www.gartner.com/en/documents/3985421

9 https://www.gartner.com/en/newsroom/press-releases/2021-02-15-gartner-forecasts-worldwide-low-code-development-technologies-market-to-grow-23-percent-in-2021

10 https://www.gartner.com/en/newsroom/press-releases/2021-09-07-gartner-identifies-four-trends-driving-near-term-artificial-intelligence-innovation

11 https://go.tdwi.org/rs/626-EMC-557/images/TDWI_Analytics-Maturity-Model-Assessment-Guide_2020.pdf

12 https://www.gartner.com/smarterwithgartner/3-barriers-to-ai-adoption/

13 https://www2.deloitte.com/content/dam/insights/us/articles/6546_talent-and-workforce-effects-in-the-age-of-ai/DI_Talent-and-workforce-effects-in-the-age-of-AI.pdf

14 https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain

15 https://aws.amazon.com/blogs/apn/ai-for-data-analytics-aida-partner-solutions-will-empower-business-experts-with-predictive-analytics/