Hype Cycle for Artificial Intelligence, 2021

Published 29 July 2021 - ID G00747539 - 131 min read

By Analyst(s): Shubhangi Vashisth, Svetlana Sicular

Initiatives: Artificial Intelligence

AI initiatives continue to accelerate as more enterprises embrace digital transformation of their core operations. Data and analytics leaders must leverage this research to successfully navigate AI-specific innovations that are in various phases of maturation, adoption and hype.

Analysis

What You Need to Know

As enterprises accelerate their digital transformation, artificial intelligence (AI) techniques are at the core of their efforts to mechanize data-driven models and insights. Curating sustainable small and wide data foundations is key for successful and value-optimized AI initiatives. Organizations are focusing on operationalization, risk management and responsible AI as they look to scale their AI initiatives.

Data and analytics (D&A) leaders must leverage this research to prioritize and accelerate investments in AI technologies that are at various stages of maturity. Doing so will minimize risk and maximize the value realized from AI initiatives.

In addition, D&A leaders should consult the following adjacent Hype Cycles to form a holistic view of D&A:

- Hype Cycle for Analytics and Business Intelligence, 2021

- Hype Cycle for Data Science and Machine Learning, 2021

- Hype Cycle for Natural Language Technologies, 2021

- Hype Cycle for Customer Experience Analytics, 2021

The Hype Cycle

The AI market remains in an evolutionary state, with more AI innovations appearing on the upward-sloping Innovation Trigger (see Figure 1). This status is indicative of end users seeking specific technology capabilities that are often beyond the capabilities of current AI tools. The field of AI continues to attract high levels of research and development (R&D) and equity investments. Enterprises are looking to innovate with AI, which continues to be viewed as an important differentiator for many organizations.

The following megatrends dominate the AI landscape:

- The urgency and criticality of productizing AI and transforming business are driving the need for **operationalization of AI platforms**. These platforms enable reusability, scalability and governance, which accelerate AI adoption and growth. AI orchestration and automation platforms (AIOAPs) and model operationalization (ModelOps) reflect this trend.

- Innovation in AI manifests itself in **efficient use of all resources** — data, models and compute. The following innovations have gained visibility in the AI sector: multiexperience, composite AI, generative AI and transformers.

- **Responsible AI** includes explainable AI, risk management and AI ethics for increased trust, transparency, fairness and auditability of AI initiatives.

- **Data for AI** forms the foundation for successful AI initiatives. Small and wide data approaches enable more robust analytics and AI, reduce organizations' dependency on big data, and deliver richer, more complete situational awareness.

As these innovations mature, adoption levels will rise, causing some technologies to move rapidly to the currently empty Plateau of Productivity.

Edge AI, decision intelligence and knowledge graphs remain at the Peak of Inflated Expectations as organizations continue to explore more use cases involving these technologies. Widely adopted techniques, such as chatbots and computer vision, have moved deeper into the Trough of Disillusionment. Intelligent applications have progressed rapidly to the trough, due to an increasing number of organizations looking to embed AI and "intelligence" within existing applications.

Figure 1: Hype Cycle for Artificial Intelligence, 2021
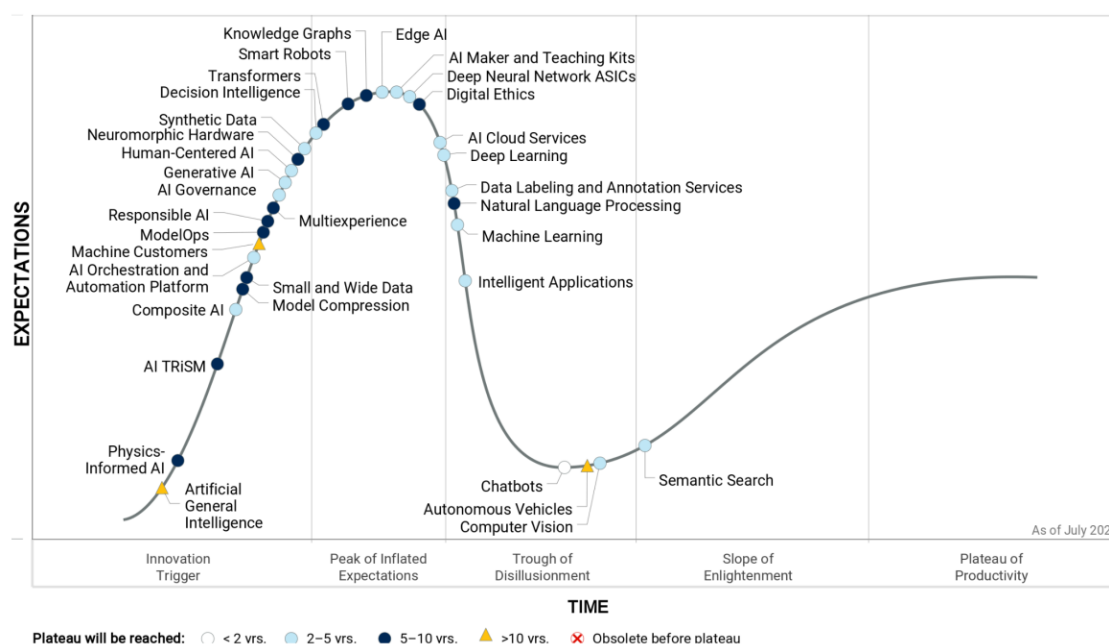
Source: Gartner (July 2021)

Downloadable graphic: Hype Cycle for Artificial Intelligence, 2021

The Priority Matrix

Compared with other Hype Cycles, the AI Hype Cycle is more fast-paced, with an above-average number of innovations reaching mainstream adoption within two to five years. Investments and interest in AI remain high. Through this research, you should be able to develop the mindset, prioritize, and adopt the innovations and tools necessary to meet this time frame. Such preparation will enable your organization to maximize the benefits of many AI innovations once they become mainstream. Manage risk by implementing the emerging practices associated with responsible AI, AI governance and digital ethics. This approach will allow you to better evaluate innovations at the Peak of Inflated Expectations.

We see an increased focus on minimum viable products and accelerated AI development cycles. We hope that this approach will become established as a best practice. We recommend employing it today to adopt and scale technologies, starting with some of those in the Trough of Disillusionment — namely, natural language processing (NLP), machine learning, chatbots, intelligent applications and computer vision.

Remarkably, all benefits but two are either high or transformational. Thus, ignoring AI will put enterprises at a competitive disadvantage. Awareness of the vendors' technology roadmaps is critical: Enterprises must determine whether they should invest in homegrown AI capabilities to fill current gaps, or wait for the vendors to deliver the necessary capabilities.

Table 1: Priority Matrix for Artificial Intelligence, 2021
(Enlarged table in Appendix)

Benefit	Years to Mainstream Adoption			
	Less Than 2 Years	2 - 5 Years	5 - 10 Years	More Than 10 Years
Transformational		Composite AI Computer Vision Decision Intelligence Deep Learning Edge AI Generative AI Human-Centered AI Intelligent Applications Machine Learning	Multiexperience Natural Language Processing Neuromorphic Hardware Physics-Informed AI Responsible AI Transformers	Artificial General Intelligence Autonomous Vehicles
High	Chatbots	AI Cloud Services AI Governance AI Maker and Teaching Kits AI Orchestration and Automation Platform Deep Neural Network ASICs Synthetic Data	AI TRiSM Digital Ethics Knowledge Graphs Model Compression ModelOps Small and Wide Data Smart Robots	Machine Customers
Moderate		Data Labeling and Annotation Services Semantic Search		
Low				

Source: Gartner (July 2021)

Off the Hype Cycle

The following list indicates which innovations were dropped from this year's Hype Cycle and why:

- **AI Marketplaces:** Dropped due to less market traction.

- **Augmented Intelligence:** Dropped due to less market traction.

- **Cognitive Computing:** Became obsolete before plateau.

- **FPGA Accelerators:** Dropped due to less market traction.

- **GPU Accelerators**: Plateaued.

- **Insight Engines**: Covered as part of the Hype Cycle for Natural Language Technologies, 2021.

- **Things as Customers**: Renamed to Machine Customers.

On the Rise

Artificial General Intelligence

Analysis By: Farhan Choudhary

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition

Artificial general intelligence (AGI) is the hypothetical intelligence of a machine that has the capacity to understand or learn any intellectual task that a human being can. AGI is also often attributed to functionalities within computer programs that enable them to exhibit self-preservation, awareness, sentience, sapience and consciousness similar to human beings.

Why This Is Important

As the lines between human-machine interaction blur, AGI for some extremely narrow and controlled use cases could start to emerge, which will allow even better decision-making capabilities in those systems. AGI can be viewed as important for pushing the overall boundaries of learning systems as we know them today.

Business Impact

- AGI is unlikely to emerge in the next 10 years or even more, although research will continue. If it does finally appear, it will probably be the result of a combination of many special-purpose AI technologies.

- The form under which it might emerge could be radically different from what people expect.

- Some of the economic, social and political implications will be disruptive — and highly controversial.

Drivers

- Considerable hype is driven by the applications developed through reinforcement learning, neuro-symbolic reasoning, simulation approaches, evolutionary algorithms, ambient intelligence and knowledge representation.

- Vendors such as Google, IBM, NNAISENSE, OpenAI and Vicarious are also actively researching the field of AGI.

- AGI can be used to experiment with mundane tasks that are highly narrow and constrained use cases. However, that defeats the notion of an all-encompassing AGI and becomes narrow AI.

- The innate desire of mankind to "shoot for the stars" is also a major driver to AGI. At one point in history, humans wanted to fly by mimicking bird flight. Today, airplane travel is a reality. The inquisitiveness of the human mind to take inspiration from nature, from itself, is not going to fizzle out.

Obstacles

- AGI at its core exhibits self-preservation, awareness, sentience, sapience and consciousness. These traits have a moral dimension and, therefore, open up a bevy of legal rights of AI and a social contract for AI (for instance, in Isacc Asimov's principles).

- The current simulation-based and neural-network-based approach is far from adequate to achieve AGI. Significant breakthroughs and revisiting the fundamentals will be required to even think of an AGI future.

- There's a lack of consensus on the definition of intelligence itself. Flamboyant representations through science fiction create a disconnect from reality.

- We have just started to take baby steps in controlled experiments, and we are debating on whether or not we'll be able to sustain them. Furthermore, we are inadvertently creating AI in our own image, without fully understanding our own selves, and expecting AI to be better.

User Recommendations

- Ignore AGI until researchers and advocates demonstrate significant progress. Until then, dismiss any suppliers' claims that their offerings have AGI or artificial human intelligence — these are generally illusions created by programmers.

- Focus on narrow AI, not on AGI. Special-purpose AI will have a huge and disruptive impact on business and personal life. Deliver business results enabled by applications that exploit special-purpose AI technologies, both leading-edge and older.

- Identify business results enabled by applications that exploit a broad range of AI techniques, represented in this Hype Cycle.

- Experiment with less-proven AI technologies that have precedents of success and give you differentiated advantage.

Sample Vendors

Google; IBM; NNAISENSE; OpenAI

Gartner Recommended Reading

Maverick* Research: Creativity Is Dead, Long Live AI Creativity!

Maverick* Research: Artificial Intelligence Will Make Us Dumber Unless We Can Teach It to Teach Us Back

Physics-Informed AI

Analysis By: Erick Brethenoux, Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition

Physics-informed AI (PIAI) incorporates physical and analog principles, governing laws and domain knowledge into AI models. By opposition, purely digital AI models do not necessarily obey the fundamental governing laws of physical systems and first principles — nor generalize well to scenarios on which they have not been trained. PIAI extends AI engineering to complex system engineering and model-based systems.

Why This Is Important

As AI becomes critical, greater demand is placed on AI's ability to abstract problems and better represent its context. Digital-only AI solutions cannot generalize well enough beyond the training data, limiting their adaptability. PIAI instills a more reliable representation of the context and the physical product, yielding more adaptive systems. A better ability to abstract leads to greater physical consistency, reduced training time, improved data efficiency and better generalization.

Business Impact

- Build physically consistent and scientifically sound AI models, significantly improving their applicability.

- Increase data efficiency, i.e., train models with fewer data points.

- Accelerate the training process, i.e., help models converge faster to optimal solutions.

- Improve the generalizability of models to make reliable predictions for unseen scenarios, including applicability to non-stationary systems.

- Enhance transparency and interpretability to make models more trustworthy.

Drivers

- Among many lessons, the pandemic has shown how brittle our traditional business modeling approaches were. That brittleness also comes from the fact that the digital building blocks making up our solutions cannot generalize well enough beyond their initial training data, therefore limiting the adaptability of those solutions. PIAI approaches can instill a more flexible representation of the context and conditions in which our systems operate, allowing developers to build more adaptive systems.

- Traditional AI techniques, particularly in the machine learning family, have been confronted with severe limitations — especially when it comes to causality and dependency analysis, context flexibility and memory retention mechanisms. Increasing demand on those techniques calls for new methods to overcome those limitations. PIAI approaches provide additional physical knowledge presentations, such as partial differential equations or active metadata, to guide or bound AI models. Asset-centric industries have already started leveraging these methods in physical prototyping, predictive maintenance or composite materials analysis also in conjunction with Augmented Reality/Virtual Reality implementations.

- Complex systems like climate and environmental issues, large scale digital-twin modelization and complex health science problems have been particularly challenging to model. Composite AI approaches have helped and provide more concrete answers and manageable solutions to those problems, but their engineering remains a significant challenge. PIAI can provide more immediate answers to some of those problems.

- The need for more robust and adaptable business simulation systems will also promote the adoption of PIAI approaches. With a better range of context modelization and more accurate knowledge representations techniques, simulations will be more reliable and account for a wider range of possible scenarios — all better anchored in reality.

Obstacles

- From a diagnostic perspective, the development of systematic tests and standardized evaluation for these models — across benchmark datasets and problems — could slow down the adoption of PIAI capabilities.

- Computationally, the scaling of the training, testing and deployment of complex PIAI models on large datasets — in an efficient manner — so they perform well in a rapidly changing computational landscape will also be an issue.

- Resource-wise, the collaboration across many diverse communities: physicists, mathematicians, computer scientists, statisticians, AI experts and domain scientists, will also be a challenge.

User Recommendations

- Encourage reproducible and verifiable models by starting with small-scoped problems; complex systems and environments are generally good candidates for this approach.

- Enforce standards for testing accuracy and physical consistency applicable to state-of- the-art physics and first-principles-based models of the relevant domain, while characterizing sources of uncertainty.

- Set realistic development objectives by identifying errors that cannot be reduced and discrepancies that cannot be addressed — including the quality of training or synthetic data.

- Promote model-consistent training for PIAI models and train models with data characteristics representative of the downstream application, such as noise, sparsity and incompleteness.

- Quantify generalizability in terms of how performance degrades with degree of extrapolation to unseen initial conditions, boundary conditions and scenarios.

- Build interpretable models and use semantics and active metadata to inform the context where models operate.

Sample Vendors

Google (Deepmind); MathWorks; NNaisense; NVIDIA

AI TRiSM

Analysis By: Avivah Litan

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

AI trust, risk and security management (AI TRiSM) ensures AI model governance, trustworthiness, fairness, reliability, efficacy, security and data protection. This includes solutions and techniques for model interpretability and explainability, AI data protection, model operations, and adversarial attack resistance.

Why This Is Important

AI poses new trust, risk and security management requirements that conventional controls don't address. Enterprises must implement a disciplined structured AI TRiSM approach to benefit from AI models that improve their business goals. AI models that are deployed in production should be subject to the right checks and balances to ensure sustained value generation. A set of risk and security controls, and trust enablers, must be deployed and continuously used to govern and manage the AI life cycle.

Business Impact

- Organizations who do not manage AI risk are much more likely to experience negative AI outcomes. Models will not perform as intended, because of either benign mistakes and process errors or malicious interference by bad actors.

- Organizations will make suboptimal business decisions due to AI misperformance. Worst case, malicious interference in AI will result in security failures, financial and reputation loss, and social harm from incorrect, manipulated, unethical or biased AI outcomes.

Drivers

- AI poses considerable data compromise risks, as large, sensitive datasets are often used to train AI models and are shared across organizations. Access to confidential data needs to be carefully controlled to avoid adverse regulatory, commercial and reputational consequences.

- AI risk and security management poses new operational requirements that are not well-understood, and which are not addressed by existing management systems.

- AI models drift for many different reasons and if not constantly monitored and validated, they can generate adverse unforeseen results with severe societal, ethical and operational consequences.

- AI models and data must be constantly monitored to ensure compliance, fairness and ethical implementations. Risk management measures can and should identify and eliminate bias from training data and AI algorithms.

- AI model explainability must be constantly tested to ensure original explanations and interpretations of AI models hold up during model operations. If they don't, corrective actions must be taken.

- Detecting and stopping adversarial attacks on AI requires new methods that are not available from most enterprise security systems.

- Regulations for AI risk management, such as SR 11-7: Guidance on Model Risk Management from the Federal Reserve Bank of the U.S., is driving U.S. financial services to institute measures for managing AI model risk.

Obstacles

- AI trust, risk and security management is an afterthought. Organizations generally don't consider it until models are in production.

- Most AI threats are not well-understood and are therefore not effectively addressed.

- Compliance is the main driver for AI TRiSM, but nonregulated industries have the same threats and issues as regulated sectors do. Regulated sectors, such as financial services and healthcare, are most likely to put in concrete measures to address AI TRiSM.

- AI trust, risk and security management requires a cross-functional team to work together on common goals, using common frameworks. This includes staff from legal, compliance, security, privacy, IT, data analytics and AI model development.

- Integrating life cycle controls is difficult but must be done as part of a comprehensive AI TRiSM program.

- Some toolsets, such as open-source tools for model explainability, need customization to work effectively in the enterprise environment.

User Recommendations

- Set up an organizational task force or dedicated unit to manage your AI TRiSM efforts. Include members throughout the organization who have a vested interest in your organization's AI projects — for example, from legal, privacy, security, risk, data analytics and AI development.

- Work across your organization to effectively manage best-of-breed toolsets (assuming your AI platform does not fulfill all requirements) as part of a comprehensive AI TRiSM program.

- Rightsize AI models for explainability or interpretability using open-source tools or vendor solutions that add value.

- Implement solutions that protect data used by AI models and prepare to use different methods for different use cases and components thereof.

- Incorporate risk management into model operations by using solutions that assure both model and data integrity, and that constantly validate reliable operations of both.

- Adopt specific AI security tools to ensure adversarial attack resistance and AI model resilience.

Gartner Recommended Reading

Use Gartner's MOST Framework for AI Trust and Risk Management

Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework

Composite AI

Analysis By: Erick Brethenoux, Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition

Composite AI refers to the combined application of different AI techniques to improve the efficiency of learning to broaden the level of knowledge representations and, ultimately, to solve a wider range of business problems in a more efficient manner.

Why This Is Important

Composite AI is currently mostly about combining "connectionist" AI approaches like machine learning (ML), with "symbolic" and other AI approaches like rule-based reasoning, graph analysis, agent-based modeling or optimization techniques. The ideas behind composite AI are not new. The goal is to enable AI solutions that require less data and energy to learn and which embody more "common sense." Composite AI recognizes that no single AI technique is a silver bullet.

Business Impact

Composite AI offers two main benefits. First, it brings the power of AI to a broader group of organizations that do not have access to large amounts of historical or labeled data but possess significant human expertise. Second, it helps to expand the scope and quality of AI applications (that is, more types of reasoning challenges can be embedded). Other benefits, depending on the techniques applied, include better interpretability and resilience and the support of augmented intelligence.

Drivers

- Limited availability of data, or small data, has pushed organizations to combine multiple AI techniques. Where raw historical data has been more scarce, enterprises have started to complement it using additional AI techniques such as knowledge graphs and generative adversarial networks (GANs) to generate synthetic data.

- Combining AI techniques is much more effective than relying only on heuristics or a fully data-driven approach. A heuristic or rule approach can be combined with a deep learning model (for example, predictive maintenance). Rules coming from human experts or the application of physical/engineering model analysis may specify that certain sensor readings indicate inefficient asset operations, which can be used as a feature to train a neural network to assess and predict the asset's health.

- The democratization of computer vision solutions is also a driver of this technology. In computer vision, (deep) neural networks are used to identify or categorize people or objects in an image. This output can be used to enrich or generate a graph, which represents the image entities and their relationships, answering questions like which object is in front of another, what the speed of an object is and so on. Using an ML approach only, such simple questions are very hard to answer.

- Agent-based modeling is the next wave of composite AI. In supply chain management, for example, a composite AI solution can be composed of multiple agents, each representing an actor in the ecosystem, with its own sensors to monitor local conditions and ML to make predictions. Combining these agents into a "swarm" enables the creation of a common situation awareness, more global planning optimization and more responsive scheduling.

- ML- and analytics-based AI techniques often lead to insights informing actions. In addition, the most appropriate actions can be further determined by combinations of rule-based and optimization models — a combination often referred to as prescriptive analytics.

Obstacles

- The lack of knowledge relevant to leverage multiple AI techniques could prevent organizations from considering the techniques particularly suited in solving specific problem types.

- If the methods, best practices and platforms are starting to adequately address the MLOps domain (that is, the operationalization of ML models), the ModelOps domain (that is, the operationalization of multiple AI models, such as optimization models, rules models and graph models) remains an art much more than a science. A robust ModelOps approach will be necessary to efficiently manage and govern composite AI environments, not to mention its harmonization with other disciplines such as DevOps and DataOps.

- The AI engineering discipline is also starting to take shape, but only mature organizations have started to apply its benefits in operationalizing AI techniques. Security, ethical model behaviors, models autonomy and change management practices will have to be addressed across the combined AI techniques.

User Recommendations

- Identify projects in which a fully data-driven, ML-only approach is inefficient or ill-fitted. For example, this is the case when not enough data is available or when the required type of intelligence is very hard to represent in current artificial neural networks.

- Leverage domain knowledge and human expertise to provide context to and complement data-driven insights by applying decision management with business rules, knowledge graphs or physical models in conjunction with machine learning models.

- Combine the power of machine learning in data science, image recognition or natural language processing with graph analytics to add higher-level, symbolic, spatiotemporal and relational intelligence.

- Extend the skills of ML experts or recruit/upskill additional AI experts, to also cover graph analytics, optimization or other required techniques for composite AI. In the case of rules and heuristics, skills for knowledge elicitation and knowledge engineering should also be available.

Sample Vendors

ACTICO; BlackSwan Technologies; Exponential AI; FICO; IBM; Indico Data Solutions; Petuum; SAS Institute

Gartner Recommended Reading

Top Strategic Technology Trends for 2021: AI Engineering

How to Use Machine Learning, Business Rules and Optimization in Decision Management

How to Use AI to Fight COVID-19 and Beyond

How to Manage the Risks of Decision Automation

When and How to Combine Predictive and Prescriptive Techniques to Solve Business Problems

Model Compression

Analysis By: Martin Reynolds

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition

Model compression is a technique that reduces the size of a trained neural network (the model) for deployment on small devices or increasing capacity of central systems. Model compression can reduce model size by up to 90%, using less memory and simpler computational units, by eliminating or simplifying parameters. This technique enables faster execution and lower latency, and delivers cost-performance benefits for edge inferencing.

Why This Is Important

Model compression enables deep neural networks (DNNs) to be deployed on smaller devices, reducing the computational load on data center and edge systems. Therefore, it expands AI benefits and capabilities to more edge devices by making inferencing systems more cost-effective. Model compression eliminates data compression otherwise required for data transfer to an inferencing system, delivering potentially better results than a centralized system.

Business Impact

Model compression, combined with local inferencing, can favorably shift a project's cost-benefit balance by reducing costs associated with data transfer, and with remote decompression and analysis. These costs can be substantial, particularly if a project scales out to many nodes and involves long-life assets. Local inferencing can also identify details that might otherwise be lost in compression for network transfer.

Drivers

- As IoT projects grow in scale, they will face inferencing obstacles that apply to both data center and edge inferencing servers. These include reduced inferencing performance because inferencing loads pressure servers and networks, and lower inferencing capabilities because compression techniques eliminate smaller features. With no inferencing capability in hardware, projects may face an expensive future upgrade.

- Edge inferencing addresses these issues by offloading inferencing to processors in the edge devices, for example, cameras or machinery vibration sensors. Edge inferencing in devices is richer and more scalable than that from network-based resources. Edge devices can extract more features from local data streams because there is no compression.

- Model compression aids in solving a number of problems. Edge devices with relatively small processors and memory can run relatively large models. Offloading inferencing to edge devices frees capacity at the edge and data center for richer analytics. Model compression may render a project financially viable by decreasing data center and edge operating costs. Using feature transmission communicates high-quality information using little bandwidth.

Obstacles

Projects that require inferencing on complex streams such as audio or video require the planner to address the risks below:

- The capital outlay associated with more capable edge devices may be undesirable to managers.

- Finding the right balance between device edge and central processing may be difficult.

- A compressed, but poor, inferencing solution will still be poor after model compression.

- Future modifications might prove too costly, inhibiting field upgrades

User Recommendations

Edge devices have a long life, and model compression enables inferencing on relatively inexpensive processors. Thus, plan for on-device inferencing for sophisticated devices such as cameras, vibration sensors or process metrology sensors where the model can fit inside an existing controller. When planning a system that requires inferencing on edge data that can scale to drive outsize data center computing needs:

- Include future inferencing needs to properly estimate project cloud computing costs.

- Accommodate future model compression needs by appropriately segmenting your AI pipelines.

- Leverage edge AI silicon as part of the project, even though it adds cost.

- Ensure your DNNs are functioning well before attempting to use compression.

Sample Vendors

AISPeech; CoCoPIE; Deeplite; Kneron; Latent AI; LeapMind

Gartner Recommended Reading

Deploy Leaner AI at the Edge: Comparing Three Architecture Patterns to Enable Edge AI

Emerging Technologies: Top Edge AI Use Cases in Manufacturing Industries

Small and Wide Data

Analysis By: Jim Hare, Pieter den Hamer, Svetlana Sicular

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

Small data is about the application of analytical techniques that require less data but still offer useful insights. Wide data enables the analysis and synergy of a variety of small, large, unstructured and structured data sources. Together these approaches apply a variety of data augmentation techniques such as X analytics, simulation, synthetic data, transfer learning, federated learning, self-supervised learning, few-shot learning and knowledge graphs.

Why This Is Important

Analytics and AI need to be able to use available data more effectively, either by reducing the required volume or by extracting more value from unstructured and diverse data sources. Small and wide data approaches enable more robust analytics, reducing dependency on big data and helping attain a more complete situational awareness or 360-degree view. This enables organizations to make analytics more resilient in the increasingly complex context of disruptions and evolving customer demands.

Business Impact

Small and wide data techniques help make analytical and machine learning models more resilient to disruptions. Small data is helpful for AI problems, where big datasets are not available, by applying less data-hungry techniques, synthetic approaches or by augmenting data through the synergy with unstructured, external or synthetic data sources. Wide data uses a broader variety of data sources to increase context and situational awareness for both human decision makers and AI applications.

Drivers

- Disruptions such as the COVID-19 pandemic have resulted in many production AI models across different industry verticals losing accuracy and relevance because they were trained using past big data that reflected how the world worked before the pandemic hit. Retraining models using the same approach was not feasible, because more recent data, just a few weeks old, was too limited to reflect the patterns of the new market circumstances.

- Organizations continue to struggle when getting started with AI projects if there is not enough volume or variety of data available to find the relevant model features or training datasets for complex models.

- Decision making is also becoming more complex and demanding, requiring a greater variety of data for better situational awareness and/or detecting rare events.

- More mature organizations that already implemented solutions for which they have enough data want to solve unique, differentiating problems where they need to overcome data size and variety limitations.

Obstacles

- Lack of tools and user skills needed to link disparate datasets across different data formats to uncover new insights or add more context to existing business decision making.

- Misperception that AI projects require large datasets before organizations can get started, resulting in lost productivity and delayed deployments.

- Organizations waiting until production AI models encounter accuracy issues rather than proactively incorporating small and wide data techniques early on as part of the model life cycle process.

- Confusion or lack of understanding about which small and wide data techniques are best for specific classes of AI problems.

- Nascent emerging techniques such as zero-, one- and few-shot learning that require specialized skills.

- Small data techniques are fragmented; they address specific challenges, such as only image analysis, only tabular data or only exclusively algorithmic side, rather than addressing the small and wide data issues overall.

User Recommendations

■ Explore small data and algorithmic approaches to increase model resilience and lower the barrier to entry for AI. Data techniques include data and feature enrichment and expansion via synthetic data, external data sources, metadata, graphs, etc. Algorithmic techniques include generative AI (GANs, few-shot learning) and composite AI.

■ Enrich and improve the predictive power of data by incorporating a greater variety of structured and unstructured data sources. These formats include tabular, text, image, video, audio, voice, temperature, or even smell and vibration. Wide data comes from an increasing range of internal and external data sources, such as data marketplaces and exchanges, brokers/aggregators, industry consortia, open data, social media, IoT sensors and digital twins.

Sample Vendors

Diveplane; Google; iOmniscient; Landing AI; MOSTLY AI; MyDataModels; Owkin; Veritone

Gartner Recommended Reading

Top Trends in Data and Analytics for 2021: From Big to Small and Wide Data

Tech Providers 2025: Why Small Data Is the Future of AI

3 Types of Machine Learning for the Enterprise

Understanding When Graph Technologies Are Best for Your Business Use Case

Working With Semistructured and Unstructured Datasets

AI Orchestration and Automation Platform

Analysis By: Chirag Dekate

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

AI orchestration and automation platforms (AIOAPs) enable orchestration, automation and scaling of production-ready and enterprise-grade AI pipelines. AIOAP helps standardize DataOps, ModelOps, MLOps and deployment pipelines as well as deliver enterprise-grade governance including reusability, reproducibility, release management, lineage, risk and compliance management, and security. It also unifies development, delivery (hybrid, multicloud, IoT) and operational (streaming, batch) contexts.

Why This Is Important

AIOAPs can help enterprises to:

- Standardize, govern and automate data, model and deployment pipelines and accelerate productization of AI.

- Eliminate handoff friction and impedance mismatch across DataOps, ModelOps, MLOps, deployment and governance.

- Scale AI initiatives by enabling orchestration across hybrid, multicloud, edge AI or IoT.

- Enable discoverable, composable and reusable AI artifacts (data catalogs, feature stores, model stores) across the enterprise context.

Business Impact

AIOAPs enable end-to-end AI platforms that:

- Systemize analytics and AI model development and management pipelines including ModelOps and MLOps tools.

- Integrate existing data, analytics and DSML platforms.

- Create reusability components including feature and model stores, monitoring, experiment management, model performance and lineage tracking.

- Homogenize governance including compliance, risk, security and cost across deployment (hybrid, multicloud, IoT) and operational (streaming, batch) contexts.

Drivers

- IT leaders need to deliver, manage and govern AI models within enterprise applications deployed across multiple contexts (hybrid, multicloud, edge AI and IoT).

- Traditional siloed approaches of data management and AI engineering create impedance mismatch across the data ingest, processing, model engineering and deployment. The impedance mismatches aggravate challenges from data and concept drifts.

- AIOAP enables enterprises to standardize and automate development, management, deployment, maintenance and governance best practices across end-to-end AI platforms.

- It helps align and automate the data, AI model deployment and governance pipelines.

- These operationalization and automation platforms are a core part of how early enterprise AI pioneers scale productization of AI by leveraging existing data, analytics and governance frameworks.

- Standardizing data pipelines including DataOps toolchains, creating reusability components such as data catalogs and ETL registries, monitoring, security, access control, and lineage tracking.

- For enterprise architects and I&O leaders, the enterprise AIOAP enables unification of two core contexts: deployment context across hybrid, multicloud, edge AI and IoT; and operational context across batch and streaming processing modes that commonly occur as enterprises train and deploy production models.

Obstacles

- Enterprises with low data and AI maturity levels will find the AIOAP intimidating to build, deliver and support. IT leaders in these enterprises should actively leverage their existing data management, DSML, MLOps and ModelOps platforms as building blocks rather than starting from scratch.

- AIOAP requires integration of full-featured solutions with select tools that address portfolio gaps with minimal overlap. These include capability gaps around feature stores, model stores, governance capabilities and more.

- AIOAP today requires a high degree of cloud maturity, or ability to integrate data and model pipelines across deployment contexts. The potential complexity and costs may be a deterrent for organizations that are just starting their AI initiatives.

- Enterprises seeking to deliver AIOAP often seek "unicorn" experts to productize AI. Very few vendors provide AIOAP as a result such skills are hard to come by and enterprises often have to build and support these environments on their own.

User Recommendations

IT leaders seeking to deliver AIOAP should:

- Audit current data and analytics practices and identify people, processes and technologies involved in current DataOps (management and transformation), ModelOps, MLOps (model engineering and validation) and Platform Ops (enterprise architectures, AI deployment and I&O).

- Simplify data and analytic environment and leverage current (simplified subset of) investments in data management, DSML, ModelOps and MLOps tools to build AIOAP.

- Leverage cloud service provider environments as foundational environments to build AIOAP along with rationalizing your data, analytics and AI portfolios as you migrate to the cloud.

- Avoid building patchwork AIOAP that integrates piecemeal functionality from scratch (and add another layer of tool sprawl). Utilize point solutions sparingly and surgically to plug feature/capability gaps in fully featured DataOps, MLOps and ModelOps tools.

Sample Vendors

Algorithmia; Amazon Web Services; Dataiku; DataRobot; Domino; Google Cloud;Hewlett Packard Enterprise (HPE) Ezmeral; IBM; Iguazio; Microsoft

Gartner Recommended Reading

Predicts 2021: Operational AI Infrastructure and Enabling AI Orchestration Platforms

Demystifying XOps: DataOps, MLOps, ModelOps, AIOps and Platform Ops for AI

Innovation Insight for ModelOps

Machine Customers

Analysis By: Don Scheibenreif, Mark Raskino

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition

A machine customer is a nonhuman economic actor that obtains goods or services in exchange for payment. Examples include virtual personal assistants, smart appliances, connected cars and IoT-enabled factory equipment. These machine customers act on behalf of a human customer or organization.

Why This Is Important

Today there are more internet-connected machines with the potential to act as customers than humans on the planet. We expect the number of machines and pervasive artificial intelligence (AI), like virtual personal assistants, with this capability to rise steadily over time. They are increasingly gaining the capacity to buy, sell and request service. Machine customers will advance beyond the role of simple informers to advisors and, ultimately, decision makers.

Business Impact

Over time, trillions of dollars will be in the control of nonhuman customers. This will result in new opportunities for revenue, efficiencies and managing customer relationships. Digital-savvy business leaders seeking new growth horizons will need to reimagine both their operating models and business models to take advantage of this ultimate emerging market, whose numbers will dwarf the number of human customers on (and one day perhaps off) the planet.

Drivers

- According to Gartner research, both CEOs and CIOs agree on the potential of this emerging trend. Seventy six percent of CIOs and 61% of CEOs we surveyed in 2019 believe demand from machine customers will become significant in their industry by 2030. On average, these leaders believe at least 21% of their revenue will come from machine customers by 2030.

- Today, most machines simply inform or make simple recommendations. We do see some examples of machines as more complex customers emerging, such as smart grid technologies. HP Inc. embraced this future when it created Instant Ink — a service that already enables connected printers to automatically order their own ink when supplies run low. Some Tesla cars already order their own spare parts, and Walmart has patented grocery auto reordering based on home Internet of Things (IoT) sensing.

- In B2B, U.S.-based industrial supply company Fastenal uses smart vending machines that proactively place orders when stocks run low. Thinking forward, an autonomous vehicle could determine what parking garage to take its human passengers to based on criteria such as distance from destination, price, online review score, parking space dimensions, valet options, etc. In this case, it is the parking garage marketing to the car, not the humans.

- The rise of machine customers begs some important questions. These include: (1) How do you market to, sell, service and obtain feedback from a machine customer?; (2) What will get a machine customer to buy from you when its decisions are based on algorithms, not emotion?; (3) What does "customer experience" even mean for a machine customer?

- Machine customers have the potential to generate new revenue opportunities, increase productivity and efficiency, improve health/well-being and enhance security of physical assets and people. They will also result in new sources of competition, fraud, legal and taxation challenges, and operational challenges.

Obstacles

- **Trust** — Can the human customer trust the technology to accurately predict and execute? Conversely, can the machine customer trust the organization that offers the service? The complexity involved in developing a machine customer that can learn the depth and breadth of knowledge and preference trade-offs required to act on behalf of a human customer in a variety of situations is staggering.

- **Fear** — Some humans may initially be uneasy about delegating purchasing functions to machines. And, organizations will have to consider what ethical standards, legal issues and risk mitigation are needed to operate in a world of machines as customers.

- **Technology that works** — Other barriers include: complex AI technologies, privacy, security and risk, regulatory compliance issues and data sharing.

- All this will mean that machine customers across industries will not reach the Plateau of Productivity for at least five to 10 years.

User Recommendations

- Create scenarios to explore the market opportunities. Initiate collaboration with your chief digital officer, chief data officer, chief strategy officer, sales leaders, chief customer officers and others to explore the business potential of machines as your customers.

- Identify specific use cases where your products and services can be extended to machine customers; and pilot those ideas to understand the technologies, processes and skills required.

- Build your organization's capabilities around digital commerce and AI over the next five years. First in machine learning, then extending to other facets involved in machine customers processing information, making informed decisions, and performing purchase transactions. Or, join other platforms that already have those capabilities if you don't have the resources to build them yourself.

- Follow examples from organizations like Tesla, Google, Amazon and HP to look for evidence of capabilities and business model impact.

Sample Vendors

Amazon; Google; HP; John Deere; Tesla

Gartner Recommended Reading

[Machine Customers: The Next Massive Emerging Market](#)

[How Customer Experience Changes When Your Customer Is a Thing](#)

[Why Machine Customers May Be Better Than Human Customers](#)

[Meet Your Machine Customers: 10 Machines That Will Drive Business Growth in the 2020s](#)

[IoT-Based Thing Commerce Requires a Differentiated Customer Experience](#)

[The Future of Customer Self-Service: The Digital Future Will Stall Without Customer-Led Automation](#)

ModelOps

Analysis By: Farhan Choudhary, Erick Brethenoux, Soyeb Barot

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

Model operationalization (ModelOps) is primarily focused on the end-to-end governance and life cycle management of all analytics, AI and decision models (including analytical models and models based on machine learning, knowledge graphs, rules, optimization, linguistics, agents and others).

Why This Is Important

As per Gartner's 2019 AI in Organizations survey, machine learning was the most leveraged AI technique, but not the only one. Organizations across all maturity levels rely on a variety of analytics and AI techniques, such as analytical, graph, agent-based, physical, simulation and ML models. This is where ModelOps helps — with operationalization agnosticity. Although MLOps primarily focuses on monitoring and governance of machine learning models, ModelOps assists in the operationalization and governance of all analytics, decision and AI models.

Business Impact

- Lays down the foundation for management of various knowledge representation models, reasoning capabilities and composite integration

- Creates the ability to manage decision models, integrating multiple analytics techniques for robust decision making

- Ensures collaboration among a wider business, development and deployment community, and the ability to correlate analytics model outcomes with business KPIs

- DataOps practices and ModelOps are key to addressing the data and models overlay/dependencies and ensure frictionless transfer of artifacts from one stage to another

Drivers

- As the number of projects at organizations increase, and as projects become more complicated, they will have to manage different types of analytics, AI and decision models that require different operationalization and governance procedures, especially if built from scratch.

- Organizations want to be more agile and responsive to changes within their analytics and AI pipelines, not just with models but also with data, application and infrastructure.

- The operationalization of aspects of ML models is not new, but it is in its early stages. However, with ModelOps, the functionalities provided by MLOps are extended to other non-ML models.

- ModelOps provides an appropriate abstraction layer by separating the responsibilities across various teams for how models (including analytics, machine learning, physical, simulation, symbolic and more) are built, tested, deployed and monitored across different environments (for example, development, test and production). This enables better productivity and lowers failure rates.

- There's a need to create resilient and adaptive systems that use a combination of various analytical techniques for decision support, augmentation and automation.

- There are wider risk management concerns with different models — drift, bias, explainability and integrity — which ModelOps helps address.

Obstacles

- Organizations using different types of models in production often don't realize that, for some kinds of analytics, decision and AI models (rule-based, agent-based, graph or simulation models) end-to-end governance and management capabilities can be expanded further.

- Not all analytical techniques currently benefit from mature operationalization methods. Because the spotlight has been on ML techniques, MLOps benefits from a more mature understanding, but others, such as agent-based modeling, require more attention.

- The lack of knowledge relevant to leveraging multiple analytics and AI techniques could prevent organizations from considering the techniques particularly suited to solving specific problems.

- Organizations that are siloed reinforce the practice and even separate their analytics model development from their AI model development for what is essentially the same process. This leads to redundancy in effort, or can, and reinforces the silos.

User Recommendations

- Leverage different analytics and AI techniques to increase the success rate of data and analytics initiatives.

- Utilize DevOps best practices across data, models and applications to ensure transition, reduce friction and increase value generation (e.g., using agile and lean).

- Extend the skills of ML experts, or recruit/upskill additional AI experts, to also cover graph analytics, optimization or other required techniques for composite AI. In the case of rules and heuristics, skills for knowledge elicitation and knowledge engineering should also be available.

- Establish a culture that encourages collaboration between development and deployment teams and empower them to make decisions to automate, scale and bring stability to the analytics pipeline.

- Optimize the adaptability and efficiency of your AI projects by considering a composite AI approach — integrating various AI techniques to solve business problems.

Sample Vendors

Algorithmia; Hewlett Packard Enterprise (HPE); IBM; ModelOp; Modzy; ONE LOGIC; SAS; Veritone

Gartner Recommended Reading

Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework

Innovation Insight for ModelOps

Demystifying XOps: DataOps, MLOps, ModelOps, AIOps and Platform Ops for AI

Responsible AI

Analysis By: Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

Responsible artificial intelligence is an umbrella term for aspects of making appropriate business and ethical choices when adopting AI that organizations often address independently. These include business and societal value, risk, trust, transparency, fairness, bias mitigation, explainability, accountability, safety, privacy and regulatory compliance. Responsible AI encompasses organizational responsibilities and practices that ensure positive, accountable AI development and exploitation.

Why This Is Important

Responsible AI has emerged as the top AI topic for Gartner clients. The more AI replaces human decisions at scale, the more it amplifies the positive and negative impacts of such decisions. Responsible AI pursues positive outcomes and prevents negative results by resolving dilemmas rooted in delivering value versus tolerating risks. Recently, many jurisdictions globally introduced new and pending AI regulations that challenge data and analytics leaders to respond in meaningful ways.

Business Impact

Responsible AI signifies the move toward accountability for AI development and use at the individual, organizational and societal levels. If AI governance is practiced by designated groups, responsible AI applies to everyone involved in the AI process. Responsible AI helps achieve fairness, even though biases are baked into the data; gain trust, although transparency and explainability methods are evolving; and ensure regulatory compliance, despite the AI's probabilistic nature.

Drivers

Responsible AI means a deliberate approach in many directions at once. Data science's responsibility to deliver unbiased, trusted and ethical AI is just the tip of the iceberg. Responsible AI helps AI participants develop, implement, exploit and resolve the dilemmas they face. Ideally, it enhances both sides at the following levels:

- **Organizational** — Resolving AI's business value versus risk in regulatory, business and ethical constraints. It could also include employee reskilling and intellectual property protection.

- **Societal** — Resolving AI effectiveness for societal well-being versus limiting human freedoms. Existing and pending legal guidelines and regulations, such the EU's Artificial Intelligence Act, make responsible AI a necessity.

- **Customer, citizen** — Resolving privacy versus convenience involves a thin line between customers' readiness to give their data in exchange for goods or benefits and customer/citizen concerns about their privacy. Fairness and ethics are the greatest drivers in this space. Regulations shed light on the necessary steps — for example, the U.S. Federal Trade Committee's "Using Artificial Intelligence and Algorithms" for consumer protection. However, this does not relieve organizations of deliberation specific to their constituents.

- **Workplace** — Resolving work efficiency versus employer "creepiness" includes concerns about AI's effect on jobs and employee morale, as well as change management.

AI affects all ways of life and touches all societal strata; hence, the responsible AI challenges are multifaceted and cannot be easily generalized. New problems constantly arise with rapidly evolving technologies and their uses, such as using generative AI for creating deepfakes. Most organizations combine some of the following drivers under the umbrella of responsible AI:

- Accountability

- Diversity

- Ethics

- Explainability

- Fairness

- Human centricity

- Operational responsibility

- Privacy

- Regulatory compliance

- Risk management

- Safety

- Transparency

- Trustworthiness

Obstacles

- Unawareness of AI's unintended consequences prevails. Many organizations turn to responsible AI only after they hit AI's negative effects, whereas prevention is easier and less stressful.

- Legislative pace, uncertainty and complexity puts responsible AI on hold in many firms. It also leads to one-sided efforts for regulatory compliance, while ignoring other responsible AI drivers.

- Rapidly evolving AI technologies, including tools for explainability, bias detection, privacy protection, and some regulatory compliance, lull organizations into a false sense of responsibility, while mere technology is not enough. A disciplined AI ethics and governance approach that brings together multiple perspectives and diversity of opinions is necessary, in addition to technology.

- Poorly defined accountability and incentives for responsible AI practices make responsible AI look good on paper, but ineffective in reality.

User Recommendations

Data and analytics leaders, take responsibility — it's not AI, it's you who are liable for the results and impacts, either intended or unintended.

- Combine the responsible AI aspects you currently address independently to promulgate consistent approaches across all focus areas. The most typical areas of responsible AI in the enterprise are fairness, bias mitigation, ethics, risk management, privacy and regulatory compliance.

- Designate a champion accountable for the responsible development of AI, for each use case.

- Raise awareness of AI differences from the familiar concepts continuously. Provide training and education on responsible AI, first to most critical personnel, and then to your entire AI audience.

- Establish an AI ethics board to resolve AI dilemmas. Ensure diversity of participants and the ease to voice AI concerns.

- Participate in industry or societal responsible AI groups. Learn best practices and contribute your own, because everybody will benefit from this.

Sample Vendors

Google, H2O.ai, IBM, Microsoft, SAS, Tazi.ai

Gartner Recommended Reading

Predicts 2021: Artificial Intelligence and Its Impact on People and Society

Top Trends in Data and Analytics for 2021: Smarter, More Responsible and Scalable AI

Cool Vendors in AI Governance and Ethical Response

Case Study: Ethical AI With an External Board (Axon)

What Non-Technology Executives Should Do in Support of Responsible AI Initiatives

Financial Services CIOs Must Focus AI Investments on 'Responsible AI' in 2021

Multiexperience

Analysis By: Jason Wong

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

Multiexperience describes interactions that take place across a variety of digital touchpoints (such as web, mobile apps, conversational apps, AR, VR, MR and wearables) using a combination of interaction modalities in support of a seamless and consistent digital user journey. Modalities include no-touch, voice, vision and gesture. Multiexperience is part of a long-term shift from the individual computers we use today to a multidevice, multisensory and multilocation ambient computing experience.

Why This Is Important

Multiexperience (MX) is the new "omnichannel" for a digital-first world. Through 2030, the digital user experience (UX) will undergo a significant shift in terms of how customers, partners, citizens and employees experience their environments. MX is about the shift both in UX perception and in interaction models, which leads to a multisensory, multidevice, multilocation and multitouchpoint digital journey.

Business Impact

To achieve digital business transformation, it is essential to understand and exploit multiexperience. Applying multiexperience design to digital experiences removes friction and effort for the users — both customers and employees. Adopting MX will allow organizations to be more agile — delivering positive business outcomes by serving customers and employees in ways that best suit their needs and expectations.

Drivers

- Organizations are shifting their delivery models from projects to products, but beyond products is the experience — the collection of feelings, emotions and memories. Web and mobile apps are already commonplace, but they are undergoing UX changes driven by new capabilities like progressive web apps, WebXR and artificial intelligence (AI) services. Conversational platforms allow people to interact more naturally and effortlessly with the digital world. Virtual reality (VR), augmented reality (AR) and mixed reality (MR) are changing the way people interact with and perceive the physical-digital world.

- As organizations continue to invest in customer experience (CX) and employee experience (EX), they will need to apply MX front-end architecture and technology strategies to be more agile at serving business needs and user expectations. When MX discipline is applied with great UX in support of CX and EX strategies, total experience (TX) transformation is achieved. TX requires MX to be executed with CX, EX and UX in harmony and synchronicity.

- The long-term manifestation of MX is a composable digital experience that is adaptive, seamless, collaborative, consistent, personalized and ambient. This will happen over the next five years — and has already been accelerated by the COVID-19 pandemic, which has increased reliance on digital touchpoints and no-touch modes of interaction. In this new decade, MX is needed to deliver transformative and memorable experiences for customers, employees and all users of your digital products and services.

Obstacles

- Privacy concerns may dampen the enthusiasm and impact of MX adoption. Users will need to consent to sharing their location, accepting notifications and being tracked across their devices.

- On the technical front, the fragmentation of many consumer devices and the inconsistency of interoperability standards are enormous barriers to seamless MX integration of front-end technologies.

- The skills needed for MX development, such as immersive interaction design, are still lacking in most enterprise software engineering teams.

- Don't expect automatic plug-and-play of off-the-shelf devices, applications and services for MX. Instead, proprietary ecosystems of MX solutions will exist in the near term.

User Recommendations

Application and software engineering leaders should:

- Identify three to five high-value, proof-of-concept projects in which MX design can lead to more effortless, compelling and transformative experiences.

- Use personas and journey mapping to address the requirements of diverse business use cases. Use external-facing and internal-facing scenarios to support a unified digital experience.

- Collaborate with UX design teams to create a design system that spans desired MX touchpoints and modes of interaction. This ensures that MX development teams can accurately and consistently apply visual, behavioral and written guidelines.

- Establish a multidisciplinary fusion team including (but not limited to) IT, product managers, UX designers and business stakeholders.

- Focus on understanding how unified digital experiences impact the business, and use evolving MX technologies to create targeted solutions for customers or internal constituencies.

Gartner Recommended Reading

How to Apply Design and Architecture to Multiexperience Application Development

Transcend Omnichannel Thinking and Embrace Multiexperience for Improved Customer Experience

Multiexperience Will Be the New Normal for Consuming Analytics Content in the Augmented Era

AI Governance

Analysis By: Svetlana Sicular

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

AI governance is the process of creating policies, assigning decision rights and ensuring organizational accountability for risks and investment decisions for the application, and use of artificial intelligence techniques. AI governance is part of adaptive data and analytics governance. It addresses the perceptive, predictive and probabilistic nature of AI.

Why This Is Important

With AI now delivering value in practical enterprise application, data and analytics leaders see that scaling AI without governance is dangerous. When each AI output is replicated millions of times, they ask how to balance the business value promised by AI against the need for appropriate oversight, risk management and investment decisions. AI draws the attention of legislators worldwide. AI regulations proliferate, mandating actions but giving more clarity about AI governance priorities.

Business Impact

AI governance as part of organizational governance structure enacts responsible AI. It gives the common implementation and adherence mechanisms across the business ecosystem when it comes to:

- Ethics, fairness and safety to protect the business and its reputation

- Trust and transparency to support AI adoption via explainability, bias mitigation, model governance, operationalization and collaboration norms and capabilities

- Diversity to ensure the right technology and roles for each AI project

Drivers

- AI governance has moved to the Peak area of the Hype Cycle. Enterprise practitioners are making steps toward establishing AI governance. Leading organizations in the industries establish AI governance by addressing standards for AI development and operations, providing best practices, guidelines for model management and monitoring, data labeling and interpretation, explainability, fairness, bias mitigation, security and legal.

- Regulations around the Globe target AI directly and affect AI practices indirectly, making AI governance goals more concrete. The FTC's law enforcement actions, studies and guidance emphasize that the use of AI tools should be transparent, explainable, fair and empirically sound, while fostering accountability. The Algorithmic Impact Assessment is a mandatory risk assessment tool intended to support the Canada Treasury Board's Directive on Automated Decision Making. AI Governance in Japan provides intermediate, nonbinding guidelines and explanations of legal responsibilities for AI within the scope of the other laws. AI governance prioritizes methods for proactive regulatory compliance.

- Trust and transparency of AI solutions are crucial for AI adoption. The probabilistic and opaque nature of AI are new to the audience who is used to deterministic outcomes. AI governance can minimize misinterpretations of AI results via scrutinizing trust in data sources and explainability of AI decisions. It provides specific testing and validation guidelines, differentiating "life-critical AI."

- AI governance is necessary to establish AI accountability. It is difficult because all use cases differ in terms of their data, solution and outcomes requirements. It outlines reactive responsibilities, actions and procedures in the case of unanticipated and unintended consequences. It ensures ethics is considered for each use case.

Obstacles

- Often, AI governance is a stand-alone initiative, which stalls its progress. The best way is to extend existing governance mechanisms to take advantage of the recognizable policies and methods, such as in data governance. AI governance benefits from a conversation with security, legal and customer experience functions.

- Many governance initiatives assume command and control. Instead, adaptive governance supports freedom and creativity in AI teams, but also protects the organization from reputational and regulatory risks. Little or no governance in AI teams to facilitate freedom and creativity is an acceptable approach if this is a conscious governance decision.

- AI value assurance and model risk management are new in AI. While methods exist, for example, in the financial industry, they are largely unknown to others, and every governance organization is inventing its own.

- Technologies to support AI governance are nascent and fragmented.

User Recommendations

- Apply the framework of trust, transparency and diversity, and to data, algorithms and people: this lets them meet the AI-specific, ever-evolving considerations. This framework should extend and advance existing governance mechanisms, such as risk management or data and analytics governance.

- Decide on the common organizational structure and accountability for propagating responsible AI. For example, what to centralize and what to do locally.

- Establish and refine processes for handling AI-related business decisions.

- Establish processes for AI review and validation. For each AI use case, require an independent AI model validator, a data scientist whose job is to assure model explainability and robustness. Have everyone in the process defend their decisions in front of their peers and validators.

- Gain agreement on AI risk guidelines that are driven by the business risk appetite and regulations.

- Ensure that humans are in the loop to mitigate AI deficiencies.

Sample Vendors

Arthur; Chatterbox Labs; DarwinAI; DreamQuark; Google; IBM Watson OpenScale; KenSci; Prodago; SAS; Weights & Biases

Gartner Recommended Reading

Build AI-Specific Governance on Three Cornerstones: Trust, Transparency and Diversity

Adaptive Data and Analytics Governance to Achieve Digital Business Success

Effective Data Governance for Government AI Projects — What CIOs Need to Know

Cool Vendors in AI Governance and Responsible AI

Reset Your Information Governance Approach by Moving From Truth to Trust

Generative AI

Analysis By: Svetlana Sicular, Brian Burke, Avivah Litan

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition

Generative AI refers to AI techniques that learn a representation of artifacts from the data, and use it to generate brand-new, completely original artifacts that preserve a likeness to original data. Generative AI can produce totally novel media content (including text, image, video and audio), synthetic data and models of physical objects. Generative models also can be used in drug discovery or for the inverse design of materials having specific properties.

Why This Is Important

Exploration of generative AI methods is growing and proving itself in life sciences, healthcare, manufacturing, material science, media, entertainment, automotive, aerospace, defense and energy industries. It is embraced for creative work in marketing, design, architecture and creative media content. Synthetic data that is created using generative AI supports the accuracy and speed of AI delivery. Generative AI is becoming more common and accessible.

Business Impact

- The field of generative AI will progress rapidly in both scientific discovery and technology commercialization.

- It is currently as futuristic as it gets, and, at the same time, successful in a wide range of applications, from creating new molecules to preserving data privacy.

- Negative use of generative AI, such as deepfakes, will increasingly cause problems in the future.

- Technologies that provide AI trust and transparency will become an important complement to the generative AI solutions.

Drivers

- The hype around generative AI is accelerating. The fast progress of transformers is top of mind in the AI community. Notably, GPT-3 by OpenAI and AlphaFold 2 by Google's DeepMind, both of which use transformers, were the main AI news in 2020. Generative adversarial networks (GANs), variational autoencoders, autoregressive models and zero/one/few-shot learning have been rapidly improving generative modeling while reducing the need for training data.

- Machine learning and NLP platforms are introducing generative AI capabilities, along with transfer learning for reusability of generative models, making them accessible to data science teams.

- Industry applications of generative AI are growing. For example, in healthcare, generative AI creates medical images that depict the future development of a disease. In consumer goods, it generates catalogs. In e-commerce, it can help customers "try on" various makeups and outfits. In manufacturing, quality inspection models use synthetic data. A growing number of life sciences companies are examining generative AI to accelerate drug development.

- Content creation and improvement, such as text, images, video and sound, is already penetrating marketing, media, entertainment and more. Examples include personalized copywriting, and noise cancellation and visual effects in videoconferencing. A combination of generative techniques, like audio to video generation, inspires new creative and business applications.

- Creative AI, such as music, produces artwork that typically requires imagination.

- Synthetic data draws enterprises' attention by helping augment scarce data, mitigate bias or preserve data privacy. For example, Duke Health will use synthetic data to protect the privacy of patients and their data in the institution's research efforts.

- Generative AI will disrupt software coding. When combined with existing development automation techniques, it has the potential to automate up to 70% of the work done by programmers.

Obstacles

- Regrettably, generative AI can be used for fraud, malware, disinformation and instigation of social unrest. Full and accurate detection of generated content will remain challenging for years and may not be completely possible.

- Generative AI technologies underpin deepfakes, content that is dangerous in politics, business and society. Technical, institutional and political interventions combined will be necessary to fight deepfakes. We will see unusual collaborations, even among competitors, to solve the problem of deepfakes and other ethical issues rooted in generative AI.

- Reproducibility of generative AI results will be challenging in the near term.

- Fragmented and specialized technology offerings (such as generating only images or only text) currently lead to a combination of tools rather than a single solution.

- Compute resources for training large transformer models are high and are not affordable to most enterprises. It is possible to exploit them, but not develop your own models.

User Recommendations

- Determine how synthetically generated data could accelerate the analytics development cycle, lessen regulatory concerns, facilitate data monetization and lower the cost of data acquisition, especially if you lack data for rare events.

- Investigate how generative AI techniques benefit your industry or sector. Determine initial use cases where you can rely on purchased capabilities or partner with research institutions.

- Examine and quantify the advantages and limitations of generative AI. Supply guidelines in cases where generative AI could bring breakthroughs, as they require skills, funds and caution. Weigh technical capabilities with ethical considerations.

- Prepare to mitigate the impact of deepfakes, which can cause serious disinformation and reputational risk. Methods like algorithmic detection and tracing content provenance to do this are evolving.

- Pay close attention to the generative AI techniques, as we expect their rapid adoption.

Sample Vendors

Adobe (Sensei); Bitext; Dessa; Diveplane; Google (DeepMind); Landing AI; MOSTLY AI; OpenAI; Phrasee; Rosebud; Spectrm; Tanjo; Textio

Gartner Recommended Reading

Innovation Insight for Generative AI

Predicts 2021: Artificial Intelligence and Its Impact on People and Society

How to Benefit From Creative AI — Assisted and Generative Content Creation

Emerging Technologies: Critical Insights Into AI-Augmented Software Development

Human-Centered AI

Analysis By: Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition

Human-centered AI (HCAI) is a common AI design principle calling for AI to benefit people and society. It assumes a partnership model of people and AI working together to enhance cognitive performance, including learning, decision making and new experiences. HCAI is sometimes referred to as "augmented intelligence," "centaur intelligence" or "human in the loop," but in a wider sense, even a fully automated system must have human benefits as a goal.

Why This Is Important

HCAI, when AI serves human and societal goals, continues to emerge as a design approach to deliver the most value from AI. An early idea that AI is simply a neutral technology is passing. Organizations see that HCAI allows them to manage the AI risks, to be ethical and more efficient with automation, while complementing AI with a human touch and with common sense. Many AI vendors have also shifted their positions to the more impactful and responsible HCAI approach.

Business Impact

HCAI compensates for human limitations and expands the possibilities for AI in the following key scenarios:

- Certain job tasks are done by AI, and the rest are done by people.

- People complete the job started with AI when AI reaches the limits of its capabilities or resources.

- Assistive AI develops and expands people's skills and talents.

- Innovation when neither AI nor people can accomplish the task without each other.

- Fully automated or autonomous systems where humans have an override capacity.

Drivers

- AI solutions must support human goals and objectives. This includes nonhuman, beneficial to people, ways of optimizing a process or solving a problem in order to arrive at new and different results, taking advantage of machine learning working differently from human learning.

- People are the ones who adopt AI (they can sabotage it too). It is human nature to use what we like, understand and trust. AI can remove many avoidable limitations, biases and blind spots. However, there are many intrinsically human irrationalities that we admire and want to preserve as a society. People do not want to be treated as robots. These people are your employees and your customers, and they are the key to AI adoption.

- More organizations are turning to the HCAI approach where they lead an ongoing discussion about what's right and wrong to do with AI before and during the progress of AI projects.

- HCAI is an intentional approach that questions and validates AI optimization goals. AI systems that solely focus on optimizing for a single business metric, like making customers click on the next news item or video, lead to dangerous societal outcomes and damage reputation in the eyes of customers, partners and employees.

- AI is probabilistic: It means that AI's mistakes are unavoidable. AI-related opportunities promise to do what only people could do in the past — diagnose diseases, play games and maintain cogent conversations. Some results could be (egregiously) incorrect, although most of them are amazingly accurate. AI mistakes without a human in the loop lead to unintended consequences.

- People's flexibility compensates for automation's limitations. Properly orchestrated autonomy makes AI impactful, for example, when AI substitutes a human in harsh working conditions. But unattended automation may lead to a misappropriation of investment and often presents insurmountable complexity.

Obstacles

- Many data science and AI teams include exclusively technical reviews for AI projects, while the resulting human impact might invalidate the entire project.

- AI systems often make decisions and take actions, but miss a feedback loop or include it as an afterthought. This doesn't mean that a human must validate every single decision, but there must always be a review and override possibility for decisions. For example, autonomous vehicle design is centered on human safety and always includes a possibility of giving control to a human driver.

- It is hard to define what AI solution is socially beneficial and human-centered. Not everything that is socially beneficial is human-centered — for instance, a social credit system.

- Anthropomorphizing AI does not mean it is human-centered. For example, virtual assistants might not give users enough understanding and control over AI-enabled answers, thus impairing AI adoption.

User Recommendations

- Establish HCAI as a key principle and a design approach. Always determine who will benefit from an AI solution. Implement AI to focus human attention where it is most needed in order to accelerate organizational competencies that fulfill your vision for digital transformation.

- Create an AI oversight board that reviews your AI plans from the HCAI position as part of its charter. Make AI goals explicit and a decision process about AI planning and validation transparent. Ensure all people can voice their concerns.

- Ensure human safety — for example, for AI moderation in social media.

- Include user experience design to facilitate HCAI. This design could be abstract (software, services, digital) or in the physical space (physical robots).

- Maximize the effects of AI-augmented roles via ongoing education, experience labs, AI-enabled just-in-time training and other methods, so the company, ecosystem and the entire society can take on more exceptional and forward-looking work.

Gartner Recommended Reading

A Human-Centric Approach to Data and Analytics: Introducing the Homo Analyticus

AI Ethics: Use 5 Common Guidelines as Your Starting Point

Neuromorphic Hardware

Analysis By: Alan Priestley

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition

Neuromorphic hardware comprises semiconductor devices inspired by neurobiological architectures. Neuromorphic processors feature non-von-Neumann architectures and implement spiking neural network execution models that are dramatically different from traditional processors. They are characterized by simple processing elements, but very high interconnectivity.

Why This Is Important

As of 2021, most AI development leverages parallel processing designs based on GPUs. These are high-performance, but at the same time high-power-consuming, devices that are not applicable in many deployments.

Neuromorphic systems utilize asynchronous, event-based designs that have the potential to offer extremely low power operation. This makes them uniquely suitable for edge and endpoint devices, where their ability to support object and pattern recognition can enable image and audio analytics.

Business Impact

AI techniques are rapidly evolving, enabled by radically new hardware designs.

- As of 2021, DNN algorithms require the use of high performance processing devices and vast amounts of data to train these systems, limiting scope of deployment.

- Neuromorphic devices can be implemented using low power devices, bringing the potential to drive the reach of AI techniques out to the edge of the network, accelerating key tasks such as image and sound recognition.

Drivers

- Gartner expects that many different AI-related hardware technologies and architectures will come to market before a stable state of widespread, high-volume deployment is reached.

- Currently, much of the work on developing AI-based systems is focused on the use of DNNs that leverage highly parallel data-intensive processing techniques to simulate the operation of a biological brain.

- These AI developments typically use semiconductor devices, such as GPUs or custom designed ASICs. While meeting today's demands in terms of performance, the power consumption of many of these designs limits their ability to scale to meet the challenges of large scale AI implementations.

- Neuromorphic computing leverages the concept of spiking neural networks (SNNs) to model a biological brain.

- Different design approaches are being taken to implement neuromorphic computing designs — large scale devices for use in data centres, and smaller scale devices for edge computing and endpoint designs. Both these paths implement asynchronous designs that have the benefit of being extremely low power when compared to current DNN-based designs.

- Semiconductor vendors are developing chips that utilize SNNs to implement AI-based solutions.

- Neuromorphic computing architectures have the potential to deliver extreme performance for use cases such as deep neural networks and signal analysis at very low power.

- Neuromorphic systems can be simpler to train than DNNs, with the potential of in-situ training.

Obstacles

- **Accessibility:** As of 2021, GPUs are more accessible and easier to program than neuromorphic hardware. However, this could change when neuromorphic hardware and the supporting ecosystems mature.

- **Knowledge gaps:** Programming neuromorphic hardware will require new programming models, tools and training methodologies.

- **Scalability:** The complexity of interconnection challenges the ability of semiconductor manufacturers to create viable neuromorphic devices.

- Significant advances in architecture and implementation are required to compete with other DNN-based architectures. Rapid developments in DNN architectures may slow advances in neuromorphic hardware but there are likely to be major leaps forward in the next decade.

User Recommendations

- Prepare for future utilization as neuromorphic architectures have the potential to become viable over the next five years.

- Create a roadmap plan by identifying key applications that could benefit from neuromorphic computing.

- Partner with key industry leaders in neuromorphic computing to develop proof of concept projects.

- Identify new skill sets required to be nurtured for successful development of neuromorphic initiatives.

Sample Vendors

AnotherBrain; BrainChip; IBM; Intel; NeuroBlade; SynSense

Gartner Recommended Reading

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Emerging Technologies: Using Neuromorphic Neural Networks to Advance IoT Vision Projects

Emerging Technologies and Trends Impact Radar: Artificial Intelligence

Synthetic Data

Analysis By: Anthony Mullen, Alexander Linden

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

Synthetic data is a class of data that is artificially generated, i.e., not obtained from direct observations of the real world. Data can be generated using different methods such as statistically rigorous sampling from real data, semantic approaches, generative adversarial networks or by creating simulation scenarios where models and processes interact to create completely new datasets of events.

Why This Is Important

One of the major problems with AI development today is the burden in obtaining real-world data and labeling it so AI models can be trained effectively. This is remedied by synthetic data. Furthermore, synthetic data is critical in removing personally identifiable information (PII).

Business Impact

Adoption is increasing across various industries, along with use in natural language processing (NLP) applications. We predict massive increase in adoption as synthetic data:

■ Avoids using PII when training machine learning (ML) models via synthetic variations of original data or synthetic replacement of parts of data.

■ Reduces cost and saves time in ML development as it is cheaper and faster to obtain.

■ Improves ML performance as more training data leads to better training outcomes.

Drivers

- In healthcare and finance, buyers' interest is growing as synthetic data can be used to preserve privacy in AI training data.

- To meet increasing demand for synthetic data for natural language automation training, especially chatbots and speech applications, new and existing vendors are bringing offerings to market. This is expanding the vendor landscape and driving synthetic data adoption.

- Synthetic data applications have expanded beyond automotive and computer vision use cases to include data monetization, external analytics support, platform evaluation and the development of test data.

- Increasing adoption of simulation techniques is accelerating synthetic data.

- While row/record, image/video, text and speech applications are common, R&D labs are expanding the concept of synthetic data to graphs. Synthetically generated graphs will resemble but not overlap the original. As organizations begin to use graph technology more, we expect this method to mature and drive adoption.

Obstacles

- Synthetic data still has significant flaws. It can have bias problems, miss natural anomalies, be complicated to develop or may not contribute any new information to existing, real-world data.

- Data quality is tied to the model that develops the data.

- Buyers are still confused over when and how to use the technology with other data pipeline tools. As the number of techniques in data and model pipeline increases, buyers struggle to determine what techniques to use to achieve their aims (e.g., synthetic data, federated learning, differential privacy) and how to use them together.

- Synthetic data can still reveal a lot of sensitive details about an organization so security is a concern. An ML model could be reverse-engineered via active learning. With active learning, a learning algorithm can interactively query a user (or other information sources) to label new data points with the desired outputs, meaning learning algorithms can actively query the user/teacher for labels.

User Recommendations

- Identify areas in your organization where data is missing, incomplete or expensive to obtain, and thus, currently blocking AI initiatives. In regulated industries, such as pharma or finance, exercise caution and adhere to rules.

- Use synthetic variations of the original data or synthetic replacement of parts of data, when personal data is required but data privacy is a requirement.

- Begin with a sampling approach and leverage data scientists to ensure statistical validity of the sample and distribution of the synthetic data.

- Leverage specialist vendors while the technology matures.

- Mature toward the simulation-driven approach, emphasizing creating agents and processes within a simulation framework to generate permutations of interactions that result in synthetic data.

Sample Vendors

AI.Reverie; Bitext; MOSTLY AI; Neuromation; Tonic; Twenty Billion Neurons (TwentyBN)

Gartner Recommended Reading

Top Trends in Data and Analytics for 2021: From Big to Small and Wide Data

2021 Strategic Roadmap for Enterprise AI: Natural Language Architecture

Cool Vendors in AI Core Technologies

At the Peak

Decision Intelligence

Analysis By: Erick Brethenoux

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition

Decision intelligence (DI) is a practical discipline used to improve decision making by explicitly understanding and engineering how decisions are made and how outcomes are evaluated, managed and improved via feedback.

Why This Is Important

The current hype around automated decision making and augmented intelligence, fueled by AI techniques in decision making, is pushing DI toward the Peak of Inflated Expectations. The COVID-19 pandemic has revealed the brittleness of decision models; rebuilding those models to be resilient, adaptable and flexible will require the discipline brought by DI methods and techniques. A fast-emerging market around various software disciplines is starting to provide sensible answers for decision makers.

Business Impact

Decision intelligence helps organizations:

- Improve the impact of business processes by materially enhancing the sustainability of organizations' decision models based on the power of their relevance, the quality of their transparency and the strength of their resilience, thus making decisions more transparent and auditable.

- Reduce the unpredictability of decision outcomes by properly capturing and accounting for the uncertainty factors in the business context.

Drivers

- A dynamic and complex business environment, with an increasingly unpredictable and uncertain pace of business. The combination of AI techniques (such as NLP, knowledge graphs, machine learning), and the confluence of several technology clusters around composite AI, smart business process, decision management and advanced personalization platforms, are creating a new market around decision systems platforms supporting the DI discipline.

- Need to curtail unstructured, ad hoc decisions that are siloed and disjointed. Often uncoordinated, such decisions promote local optimizations at the expense of global efficiency.

- Expanding collaboration between humans and machines, supplemented by a lack of trust in technologies (such as AI) increasingly replacing tasks and promoting uneasiness from a human perspective. DI practices promote transparency, interpretability, fairness, reliability and accountability of decision models critical for the adoption of business-differentiating techniques.

- Tighter regulations that are making risk management more prevalent. From privacy and ethical guidelines to new laws and government mandates, it is becoming difficult for organizations to fully understand the risk impacts of their decisions. DI enables an explicit representation of decision models, reducing this risk.

- Uncertainty regarding decision consistency across the organization. Lack of explicit representation of decisions prevents proper harmonization of collective decision outcomes. This is remedied by DI.

Obstacles

- Decision-making silos have created data, competencies and technology clusters that are difficult to reconcile and could slow down the implementation of decision models.

- An inadequate organizational structure around advanced techniques, such as the lack of an AI center of excellence, could impair DI progress.

- Reputation-damaging outcomes from autonomous decision models (from embedded analytical assets to self-contained machine agents) and the failure to understand their collective impact impede DI adoption.

- Lack of proper coordination between business units and inability to impartially reconsider critical decision flows within and across departments diminish the effectiveness of early DI efforts.

User Recommendations

- Improve the outcomes of decision models and accommodate uncertainty factors by evaluating the contributing decision-modeling techniques.

- Promote the sustainability of cross-organizational decisions by building models using principles aimed at enhancing traceability, replicability, pertinence and trustworthiness.

- Improve the predictability and alignment of decision agents, by simulating their collective behavior while also estimating their global contribution versus local optimization.

- Develop staff expertise in traditional and emerging decision augmentation and decision automation techniques, including descriptive, diagnostic (interactive data exploration tools), predictive (machine learning) and prescriptive (optimization, business rule processing and simulation) analytics.

- Tailor the choice of decision-making technique to the particular requirements of each decision situation by collaborating with subject matter experts, AI experts and business process analysts.

Gartner Recommended Reading

Improve Decision Making Using Decision Intelligence Models

How to Manage the Risks of Decision Automation

How to Choose Your Best-Fit Decision Management Suite Vendor

AI Security: How to Make AI Trustworthy

Top Strategic Technology Trends for 2021: AI Engineering

Transformers

Analysis By: Martin Reynolds

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition

Transformer-based language models are a type of deep neural network architecture that evaluates words as sequences in a sentence. This sequence-based approach significantly improves transcription accuracy, improves processing across languages and can synthesize complex well-constructed sentences. These models can be tuned to different text domains with minimal customization. Transformers are also finding use in domains such as imaging and biochemistry.

Why This Is Important

Transformers deliver a significant improvement in text prediction accuracy. These predictions, for example, can deliver superior translations between languages, create large, well-formed sections of text from a few key phrases, create music, or identify hidden relationships across bodies of text. Transformer solutions result in many successful AI deployments and are extending into images and biochemistry.

Business Impact

Transformers deliver material improvement in text processing, sufficient that many language-based applications advanced from potentially useful to generally effective. This advance is further supported by relatively simple techniques that adapt transformers to different text domains, be it languages or specialist fields such as biochemistry or law.

Transformers deliver superior speech classifications for automated voice response systems, improving overall response accuracy.

Drivers

- **Transformers require only limited model customization to deliver effective results.** Clients report to us that they are able to use Google's open-source BERT implementation without training the entire network. Rather, they retrain only the top few layers to customize for their language domain.

- **Transformers deliver superior text classifications.** The difference between transformer-based models and prior DNN solutions is stark. Transformers model patterns in relatively large blocks of text, as opposed to predicting the next word based on the preceding snippet. These improvements have materially advanced speech and text applications. A notable example is the improvement in Google Translate.

- **Transformers can create well-formed text passages from minimal inputs.** A variant on transformer models is GPT-3, developed by OpenAI and licensed by Microsoft. This transformer model, which incorporates 175 billion parameters, is designed to create paragraphs or pages of text from small snippets. GPT-3 does this based on predicting the most likely next word in a sentence, based on its absorbed accumulated training.

Obstacles

- **Transformers do not deliver perfect results.** Although a significant advance, transformers still require careful training, and can deliver unacceptable results. GPT-3 delivers near-perfect prose, but the output may also be perfectly inaccurate. As with all AI systems, constraints are important to keep the program aligned with business needs.

- **Transformer configuration requires appropriate skills and talent.** As with all AI solutions, the end result is dependent on the skills, knowledge and talent of the trainers. Although transformers appear to be somewhat easier to configure than other systems, better training will deliver better results.

- **Newer transformer models are expanding to impractical sizes.** Large models are up to a trillion parameters, with relatively small improvements. These models are impractically large to train for most organizations because of the necessary compute resources, and their large scale drives relatively heavy inferencing workloads.

User Recommendations

- **Immediately slipstream transformers into existing speech or text programs.** If you have any older language processing systems, deployed or in development, moving to a transformer-based model could significantly improve performance. One example might be a voice interface, where transformers can interpret multiple ideas in a single utterance (e.g., "I want a pizza with thick crust, extra cheese and sausage," as opposed to selecting each item from a voice-based menu). This shift in approach can significantly advance language interfaces by reducing the number of interactions.

- **Start with Google's BERT open-source model.** Although GPT-3, Microsoft DeepSpeed and Google Switch Transformers promise superior results, the extra complexity may not justify advancing from BERT. BERT is supported by many DSML platforms.

- **Leverage existing trained transformer networks as the base of language projects.** Many BERT models are available as open source. A good place to find these models is this list.

Sample Vendors

Amazon; Google; Microsoft; OpenAI

Smart Robots

Analysis By: Annette Jump

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

A smart robot is an AI-powered, often-mobile machine designed to autonomously execute one or more physical tasks (create predictable outcomes and learn within a range of defined parameters or unpredictable outcomes but within a specific range of parameters). Smart robots can be split into different types based on the tasks/use cases, such as personal, logistics and industrial.

Why This Is Important

Smart robotics is an AI use case, while robotics in general does not certainly imply AI. Smart robots have had less adoption compared with industrial counterparts but received great hype in the marketplace, which is why smart robots are still climbing the Peak of Inflated Expectations. The pandemic increased interest in smart robots, as they can provide logistic support and automation, and support social distancing, while enterprises can demonstrate ROI, without significant capital expenditure.

Business Impact

Smart robots will make their initial business impact across a wide spectrum of asset-centric, product-centric and service-centric industries. Their ability to reduce physical risk to humans from doing specific tasks, as well as do work with greater reliability, lower costs and higher productivity, is common across these industries. Smart robots are already being deployed among humans to work in logistics and social venues, as well as safety applications.

Drivers

- The market is becoming more dynamic with technical developments of the last two years, enabling a host of new use cases that have changed how smart robots are perceived and how they can deliver value.

- The physical building blocks of smart robots — motors, actuators, chassis and wheels — have incrementally improved over time. However, areas such as Internet of Things (IoT) integration, edge AI and conversational capabilities have seen fundamental breakthroughs. This changes the paradigm for robot deployments.

- Specialization also is very important to success, as no smart robot can address all industry-specific use cases.

- The COVID-19 pandemic has accelerated smart robot adoption, and typical use cases include:

- Logistics and warehousing — Medical/healthcare: Patient care, medical materials handling, interdepartment deliveries and sanitization; Goods delivery due to social distancing and quarantine with COVID-19; Manufacturing: Product assembly, stock replenishment, support of remote operations and quality control (QC) check; Last-mile delivery; Inspection of industrial objects or equipment; Surgical robots; Agriculture: Harvesting and processing crops; Reception/concierge in hospitality, retail, hospitals, airports and so forth.

Obstacles

- Despite some advancements in AI, product and material experimentation in 2020, the progress beyond proofs of concept (POCs) is relatively slow. Companies are still trying to identify valuable business use cases and assess ROI for robots. Therefore, the position of "smart robots" still remains on the Innovation Trigger curve.

- Hype and expectations will continue to build around smart robots during the next few years, as providers expand their offerings and explore new technologies, like reinforcement learning to drive continuous loop of learning for robots and swarm management.

- Lack of ubiquitous wireless connectivity solutions outside of smart spaces and immaturity of edge AI technologies can inhibit the pace at which smart robots become semiautomated and mobile.

- The need to offload computation to the cloud will decrease from 2024, as robots will make more autonomous decisions.

- The continuous evolution of pricing models, like buy, monthly lease or hourly charge versus robot as a service for robotic solutions, can create some uncertainty for organizations.

User Recommendations

- Evaluate smart robots as both substitutes and complements to their human workforce in manufacturing, distribution, logistics, retail, healthcare or defense.

- Begin pilots designed to assess product capability and quantify benefits, especially as ROI is possible even with small-scale deployments.

- Examine current business processes into which smart robots can be deployed now and in three to five years for large-scale deployment.

- Consider different purchase models for smart robots.

- Dissolve the reluctance from staff by developing training resources to introduce robots alongside humans, as an assistant.

- Ensure there are sufficient cloud computing resources to support high-speed and low-latency connectivity in the next two years.

- Evaluate multiple global and regional providers due to fragmentation within the robot landscape.

Sample Vendors

Amazon; Ava Robotics; Geek+; iRobot; Locus Robotics; Rethink Robotics; SoftBank Robotics; Symbotic; Temi; UBTECH

Gartner Recommended Reading

[Emerging Technologies Venture Capital Growth Insights: Robots](#)

[Emerging Technologies: Smart Robots Will Augment Human Workers, Not Replace Them](#)

[Market Trends: 4 Technologies That Will Revolutionize Drones and Robots](#)

Knowledge Graphs

Analysis By: Afraz Jaffri

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition

Knowledge graphs are data structures, representing knowledge of the real world including entities (people, companies, digital assets, etc.) and their relationships, which adheres to a graph data model — a network of nodes (vertices) and links (edges/arcs). The knowledge within the graph can be explicitly stated or implicitly inferred using rules that are defined in an ontology for classes of entities and relationships. Further knowledge can be derived using graph analytics and machine learning.

Why This Is Important

Knowledge graphs capture information about the world in an intuitive way that is often easier to understand, manipulate and use than other types of data models. Google, Facebook, Amazon and other tech companies use graphs as the backbone of a number of products and services due to their ability to encode and interrelate disparate data at source. They support collaboration and sharing, search and discovery, and the extraction of insights through analysis.

Business Impact

Knowledge graphs can drive business impact in a variety of different settings including:

- Digital workplace (e.g., collaboration, sharing and insight).

- Automation (e.g., ingestion of data from content to RPA).

- Machine learning (e.g., augmenting training data).

- Investigative analysis (e.g., law enforcement, cybersecurity or financial transactions).

- Digital commerce (e.g., product information management and recommendations).

- Data management (e.g., metadata management, data cataloging and data fabric).

Drivers

- Ongoing digitization and globalization initiatives lead to growing levels of complexity and dynamics, creating a need for more adaptive and integral approaches, as offered by knowledge graphs, replacing more static and siloed approaches.

- Increasing awareness of the use of knowledge graphs in consumer products and services such as smart devices and voice assistants, chatbots, search engines, recommendation engines and route planning.

- Improvements in graph DBMS technology that can handle the storage and manipulation of graph data structures at scale. This includes PaaS offerings that take away the complexity of provisioning and optimizing hardware and infrastructure.

- Knowledge graph platform providers are entering the market that provide a suite of tools for creating, managing and using knowledge graphs. Low/no-code tools are developing and expanding the use of knowledge graphs to business and nontechnical users.

- The desire to make better use of unstructured data held in documents, images and videos using standardized metadata that can be related and managed.

- The need to manage the increasing number of data silos where data is often duplicated and usage and consumption cannot be controlled.

- The increasing use of graph algorithms and machine learning to identify influencers, customer segments, suspicious activity and critical bottlenecks in complex networks.

- Service providers are specializing in knowledge graph implementation and building offerings based on the technology.

Obstacles

- Awareness of knowledge graph use cases are increasing but business value is difficult to capture in the early stages of implementation making them low priority initiatives.

- Moving knowledge graph models from prototypes to production requires engineering and system integration expertise. Methods to maintain knowledge graphs as their size increases to ensure reliable performance and handle duplication and data quality remain immature.

- Fragmentation of the graph DBMS market across the types of knowledge graph data models (RDF or property), implementation architectures (native or multimodal) and differences in optimal workloads (operational or analytical) continue to cause confusion and hesitancy among adopters.

- Key to the long-term success of knowledge graphs is enabling data within organizations to be interoperable with external knowledge graphs to enable the ingestion, validation and sharing of ontologies and data relating to entities e.g., geography, people, events, etc.

User Recommendations

- Identify use cases where there is a need for custom-made knowledge graphs through the use of a pilot project that delivers tangible value for the business, but also learning and development for data and analytics staff.

- Take an agile approach to knowledge graph development to decrease time to value. Assess the data that is needed to feed a knowledge graph, both structured and unstructured, creating a minimum viable subset that can be used to capture the information of a business domain.

- Utilize vendor and service provider expertise to validate use cases, educate stakeholders and provide an initial knowledge graph implementation.

- Include knowledge graphs within the scope of data and analytics governance and management. To ward against perpetuating data silos, investigate and establish ways for multiple knowledge graphs to interoperate. This is likely to extend to third party data knowledge graphs.

Sample Vendors

Cambridge Semantics; Diffbot; eccenca; Ontotext; Semantic Web Company; TopQuadrant

Gartner Recommended Reading

How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications

Top Trends in Data and Analytics for 2021: Graph Relates Everything

Graph Steps Onto the Main Stage of Data and Analytics: A Gartner Trend Insight Report

Working With Graph Data Stores

Financial Data Strategy and Knowledge Graphs

Edge AI

Analysis By: Alan Priestley

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition

Edge AI refers to the use of AI techniques embedded in IoT endpoints, gateways and edge servers, in applications ranging from autonomous vehicles to streaming analytics. While predominantly focused on AI inference, more sophisticated systems may include a local training capability to provide in-situ optimization of the AI models.

Why This Is Important

An increasing number of edge computing use cases are latency sensitive (autonomous navigation), data intensive (video analytics), and require an increasing amount of autonomy for local decision making. This creates a need for AI-based applications in a wide range of edge computing and endpoint solutions. Examples include video analytics which, driven by the rapid growth in use of surveillance cameras and the need for real-time interpretation of captured video, is starting to see adoption.

Business Impact

The business benefits of deploying edge AI include:

- Improved operational efficiency, such as manufacturing visual inspection systems.

- Enhanced customer experience.

- Reduced latency in decision making, with the use of local analytics.

- Communication cost reduction, with less data traffic between the edge and the cloud.

- Increased availability even when the edge is disconnected from the network.

- Reduced storage demand through a more reactive exploitation of the data.

- Preserved data privacy at the endpoint.

Drivers

- Increasing demand for the deployment of DNN-based data analytics close to or at the point of data capture, either in edge computers or endpoint devices.

- Edge AI implementations are impacted by application and design constraints of the equipment being deployed; this includes form factor, power budget (i.e., battery powered versus mains powered), data volume, decision latency, location, and security requirements.

- AI systems can be hosted within an edge computer, gateway or aggregation point and data captured at an IoT endpoint may need to be transferred. In this architecture, the IoT endpoint is a peripheral to the AI system. The endpoint acts as a data gatherer that feeds this data to the AI system. An example of this is environmental sensors deployed for a smart agriculture application.

- AI embedded in the IoT endpoint. In this architecture, the IoT endpoint is capable of running AI models to interpret data captured by the endpoint and drives some of the endpoints' functions. In this case, the AI model (e.g., a machine learning model) is trained and updated on a central system and deployed to the IoT endpoint. An example is a medical wearable that leverages sensor data and AI to help visually impaired people navigate the world in their daily lives.

- R&D into training AI models at the edge for decentralized machine learning.

Obstacles

- Systems deploying AI techniques can be nondeterministic. This can limit the ability to control and replicate analysis results, and may impact applicability in certain use cases, especially where safety and security requirements are important.

- The autonomy implicit in an AI deployment can lead to questions of trust, especially where the operation of the AI models is not transparent.

- The deployment of edge AI solutions can raise governance and privacy concerns. While analyzing data at or close to its point of capture can alleviate some privacy concerns, it may not mitigate them completely.

- Training DNNs is a compute-intensive task, often requiring the use of high performance chips with corresponding high power budgets. This can limit deployment locations, especially where small form factors and lower power requirements are paramount.

User Recommendations

- Determine whether the new AI developments are applicable to their IoT deployments, or whether traditional centralized data analytics and AI methodologies are adequate.

- Evaluate when to consider AI at the edge versus a centralized solution. Applications that have high communications costs are sensitive to latency or ingest high volumes of data at the edge are good candidates for AI.

- Assess the different technologies available to support edge AI and the viability of the vendors offering them. Many potential vendors are startups, which may have interesting products but limited support capabilities.

- Use edge gateways and servers as the aggregation and filtering point to perform most of the edge analytics functions. Make an exception for compute-intensive endpoints, where AI-based analytics can be performed on the devices themselves.

Sample Vendors

Baidu; Google; Intel; Microsoft; NVIDIA; Qualcomm

Gartner Recommended Reading

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Forecast: AI Semiconductors, Worldwide, 2019-2025, 1Q21 Update

Emerging Technologies and Trends Impact Radar: Artificial Intelligence

AI Maker and Teaching Kits

Analysis By: Eric Hunter, Annette Jump

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition

Artificial intelligence (AI) maker and teaching kits are applications and software development kits (SDKs) that abstract data science platforms, frameworks, analytic libraries and devices to enable software engineers to incorporate AI into new or existing applications. Maker kits also emphasize teaching new skills and integration best practices between software and devices for engineers — some of which also include hardware devices.

Why This Is Important

AI maker kits leverage custom hardware devices (such as cameras, musical instruments, speakers or vehicles) with developer-friendly APIs and SDKs to encourage platform developer adoption while educating developers around new AI capabilities. These kits have also driven new vendor innovations targeting mainstream enterprise use cases — the Google Coral initiative is a good example of this.

Business Impact

The demand for AI is significant and is increasing at a rate beyond what experienced data scientists can meet alone. These offerings will equip software developers to become a key contingent for AI development and implementation. AI maker kits will also continue to reduce adoption barriers in the deployment of AI capabilities for software engineers and citizen data scientists.

Drivers

- As the demand for more proficient data scientists rises, the adoption of AI maker and teaching kits will continue to increase.

- Within many kits, developers can deploy prebuilt models and, optionally, update those models from cloud services at model runtime.

- Convergence in deployment for language, translation, vision, machine learning (ML) and cognitive search kits will support new AI use cases in the digital enterprise.

- Continued demand for organizational outcomes focusing on computer vision, natural language and other AI-aligned capabilities fit well with AI maker kit offerings.

- Maker kit paths to production-caliber offerings will continue to mature as new capabilities are integrated into market offerings.

Obstacles

- Across all categories, vendor offerings require distinct deployment considerations and have varied feature coverage differences, but we expect greater consistency in the future.

- AI maker and teaching kits support a limited set of native use cases, such as computer vision, image recognition, image labeling, natural language processing and text analytics.

- There has been an increase in user gravity and stickiness to broader, vendor-based cloud and platform offerings, including platform as a service (PaaS), which drives associated maker and teacher kit adoption.

- Market offerings do not follow a consistent set of standards, which can impact portability and introduce challenges when promoting content across environments.

- AI maker and teaching kits have an inconsistent level of support/capabilities for production-ready use cases. Some support scaling development concepts to full-scale production use cases while others offer no path from development-only scenarios.

User Recommendations

Vendor offerings are being released at a rapid pace in the market with a desire to attract new development communities. Application development leaders adopting these offerings to incorporate AI capabilities and features into applications should:

- Leverage maker kits to upskill developer knowledge and skills, which can translate to present and future enterprise needs that may directly or indirectly relate to kit-specific use cases.

- Carefully evaluate and stress-test employed maker kit offerings, along with fully understanding the going concern support for each specific offering.

- Abstract adopted vendor development kit offerings where possible to minimize portability constraints and lock-in.

- Ensure deployed capabilities are aligned to direct end-user benefits that cannot be easily achieved without AI.

- Adopt offerings in alignment with larger organizational development standards and strategies.

Sample Vendors

Amazon Web Services; Google; Intel; Microsoft; Nvidia; Pantech Solutions

Gartner Recommended Reading

How to Fast-Track Your Product Roadmap With Cloud Vendors' AI Development Accelerators

Democratization of Computer Vision Presents New Opportunities for Differentiating Personal Devices

A Framework for Applying AI in the Enterprise

How to Move Beyond AI Trials, to AI in Production

Technology Insight for Cloud AI Developer Services

Deep Neural Network ASICs

Analysis By: Alan Priestley

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition

A deep neural network (DNN) application-specific integrated circuit (ASIC) is a purpose-specific chip designed to execute the DNN computations utilized in a wide range of artificial intelligence applications. These chips can be deployed in either data center servers, edge computing systems or endpoint devices.

Why This Is Important

An increasing range of applications require the use of DNN-based AI techniques to analyze captured data. These include object detection and classification in images and video streams, natural language processing, social media recommendation engines, autonomous vehicles and pharmaceutical analytics. To effectively execute many of these applications requires the use of data center and edge computing systems, and endpoint devices that include DNN ASICS optimized for specific workloads.

Business Impact

Leveraging DNN ASIC-based systems enables:

- Efficient analysis of high-volume complex datasets, such as videos, images, audio streams enabling video analytics, object detection and classification, image recognition, natural language processing and recommendation systems.

- Edge computers and endpoint devices capable of sophisticated local automated decision making, and delivering enhanced user experience.

- Better performance and power efficiency than solutions based on GPUs of general purpose CPUs.

Drivers

■ Increasing volume of complex unstructured data requires the use of AI techniques that leverage DNN models to analyze and enable business decisions to be made based on the data content.

■ Executing DNN-based AI applications typically requires the use of computer systems that are capable of executing high volumes of highly parallel math operations.

■ Many DNN models require training using large sets of known good data. GPUs can be used for this task but high performance DNN ASICs designed for data center deployments can deliver a better solution to this problem.

■ DNN ASICs can offer significantly better performance, at lower power, than many existing CPU or GPU-based solutions available to execute AI-based workloads.

■ Often trained AI applications are deployed in locations, such as edge computing or endpoint devices, where power or form factor constraints prevent the use of many high-power AI devices. Many DNN ASICs are designed specifically for these deployments.

Obstacles

■ Today, GPUs are still the device of choice for many companies developing DNN-based AI applications.

■ Most of the open-source software frameworks used by AI developers have native support for GPUs and require dedicated software tools and workflows to support DNN ASICS.

■ Many companies developing DNN ASICs are startups, and while they often have the funding to develop a DNN ASIC and supporting software, they lack the size to scale and grow their business, having limited resources to support a broad range of AI developers.

■ There is no standardization in DNN ASIC hardware design, with every vendor offering their own unique design and requiring specific software implementation to support each DNN ASIC.

■ The large hyperscale cloud service providers are developing ASICs optimized for their specific DNN-based workloads, examples include Google's Tensor Processing Units (TPUs) optimized for its TensorFlow-based applications.

User Recommendations

Application and software engineering leaders planning an effective long-term strategy for the use of DNN-based applications and hardware must:

- Use CPUs or cloud when DNN workloads are light enough to fit in conventional CPU-based infrastructure.

- Use GPUs or dedicated AI servers with DNN ASICS when DNN workloads would otherwise consume excessive server resources.

- Select DNN ASICs and vendors that offer or support the broadest set of DNN frameworks and toolsets.

- Specify edge computing and endpoint devices that integrate low-cost DNN ASICs to support edge inferencing and local decision making in locations where power, formfactor and communications cost are critical.

Sample Vendors

Amazon Web Services (AWS); Google; Graphcore; Groq; Intel; NVIDIA; SambaNova Systems; Syntiant

Gartner Recommended Reading

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

 Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Forecast: AI Semiconductors, Worldwide, 2019-2025, 1Q21 Update

Emerging Technologies and Trends Impact Radar: Artificial Intelligence

Predicts 2021: Artificial Intelligence Core Technologies

Digital Ethics

Analysis By: Pieter den Hamer, Frank Buytendijk, Svetlana Sicular, Bart Willemsen

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition

Digital ethics comprise the systems of values and moral principles for the conduct of electronic interactions among people, organizations and things.

Why This Is Important

Digital ethics, and in particular privacy and bias, remain a growing concern. The voice of society and AI-specific ethical considerations are rapidly coming into focus for individuals, organizations and governments. People are increasingly aware that their personal information is valuable; they're frustrated by lack of transparency and continuing misuses and breaches. Organizations act to mitigate the risks involved in securing and managing personal data, and governments are implementing strict legislation in this area.

Business Impact

Digital ethics strengthens the organization's positive influence and reputation among customers, employees, partners and society. Areas of business impact include influencing innovation, product development, customer engagement, corporate strategy and go-to-market. Intention is key. If ethics is simply a way to achieve business performance, it leads to window dressing. The goal to be an ethical company serves all parties and society more broadly and leads to better business trust and performance.

Drivers

- Despite the hype around digital ethics, many organizations are still ignoring it. They think it doesn't apply to their industry or domain without giving it a deliberate consideration.

- Board members and other executives are sharing concerns about the unintended consequences that the innovative use of technology can have.

- There is frequent, high-profile press coverage of stories that concern the impact of data and technology on business and society more broadly.

- With the emergence of artificial intelligence, for the first time the ethical discussion is taking place before — and during — a technology's widespread implementation. AI ethics and other responsible AI steps are a foundation to reverse the negative popular sentiment around AI and lead to a more responsible use of its powers.

- Government commissions and industry consortia are actively developing guidelines for ethical use of AI. Examples include Ethical Framework for Artificial Intelligence In Colombia, New Artificial Intelligence Regulation in the EU and Using Artificial Intelligence and Algorithms in the U.S.

- Over the past year, a quickly growing number of organizations declared their AI ethics principles, frameworks and guidelines. They have a long way to go from declaration to execution, although some organizations already have digital ethics practices.

- Gartner predicts that by 2024, 30% of major organizations will use a new "voice of society" metric to act on societal issues and assess the impact on their business performance. The voice of society will put more pressure on governments and public and private organizations alike to ethically use technology. "Big tech" is already a negative stereotype in societal jargon.

- More universities across the globe are adding digital ethics courses and launching programs and centers to address ethical, policy and legal challenges posed by new technologies.

Obstacles

- Digital ethics is seen as a moving target because of confusion on what society expects. It might even lead to opposing the majority's opinion, based on an organization's position and beliefs.

- Digital ethics is too often reactive and narrowly interpreted as compliance, or confined to the technical support of privacy protection or viewed as explainable AI only.

- AI ethics is an emerging area in overall digital ethics. Early high-level guidelines are inconsistent and will evolve over time.

- The voice of society is a new metric where digital ethics should be present, but its weight is still to be understood. Insufficient attention leaves organizations exposed to lost business, higher costs and increased risk.

- Opinions differ across people, regions and cultures on what constitutes "good" and "bad." Even in organizations where ethics have been recognized as an important issue, consensus between internal and external stakeholders (such as customers) remains sometimes difficult to achieve.

User Recommendations

Business and IT leaders responsible for digital transformation in their organizations:

- Identify specific digital ethics issues and opportunities to turn awareness into action.

- Discuss ethical dilemmas from diverse points of moral reasoning. Ensure that the ethical consequences have been accounted for and that you are comfortable defending the use of that technology, including unintended negative outcomes.

- Elevate the conversation by focusing on digital ethics as a source of societal and business value, rather than simply focusing on compliance and risk. Link digital ethics to concrete business performance metrics.

- Ensure that digital ethics is leading and not following digital transformation. Address digital ethics early "by design" to move faster by knowing methods to resolve ethical dilemmas.

- Organize training in ethics and run workshops to create awareness within all AI initiatives about the importance that AI design and implementation require an ethical mindset and clear accountability.

Gartner Recommended Reading

Digital Ethics: From Compliance Duty to Competitive Differentiator

AI Ethics: Use 5 Common Guidelines as Your Starting Point

Every Executive Leader Should Challenge Their Teams on Digital Ethics

Digital Ethics by Design: A Framework for Better Digital Business

Data Ethics and COVID-19: Making the Right Decisions for Data Collection, Use and Sharing

Use Privacy to Build Trust and Personalize Customer Experiences

AI Cloud Services

Analysis By: Van Baker, Bern Elliot

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition

Artificial intelligence cloud services provide AI model building tools, APIs and associated middleware that enable the building/training, deployment and consumption of machine learning models running on prebuilt infrastructure as cloud services. These services include automated machine learning, vision and language services.

Why This Is Important

The use and sophistication of AI cloud services continues to increase, with vendors competing to become the platform of choice for developers and citizen data scientists. Applications will increasingly use AI cloud services in language, vision and machine learning to help automate and accelerate achievement of business objectives. However, developers struggle with these offerings as current offerings are limited, but improvements in the service portfolios will drive continued adoption.

Business Impact

The impact of AI will extend to the applications that enable business, allowing developers and data scientists to enhance the functionality of these applications. With the incorporation of forecasts, next best actions and other capabilities, including automation of many workflows that are currently handled manually, these AI cloud services will enable advanced applications that improve business performance.

Drivers

- The explosion of data from both internal and third-party sources enable insight that has previously been unavailable to the business.

- Increasing need for human to machine interactions that are based on conversational capabilities.

- A mandate for businesses to automate processes to improve accuracy, improve responsiveness and reduce costs.

- Improving performance of both AI and machine learning models.

- Reduced need for large quantities of data to train models.

- Ease of accessibility for developers and citizen data scientists to AI and machine learning services due to the availability of API callable cloud hosted services will expand the use of AI.

- Multiple use cases spanning language, vision and automated machine learning services.

- Use of automated machine learning to tailor the off-the-shelf services to more precisely address the specific needs of the business.

- A wide range of AI cloud services from both hyperscaler cloud providers as well as specialized providers in the market.

- Business transformation efforts driving the development of applications with enhanced capabilities to improve business operations and workforce productivity.

- Increasing deployments of sensor networks in IoT-based solutions that facilitate data use to drive model development and facilitate proactive response to changes in the data rather than reactive response.

- The emerging AI model marketplaces should help developers adopt those techniques through AI cloud services.

Obstacles

- Lack of understanding by developers and citizen data sciences about these services and how they can be applied to specific business use cases.

- Pricing models for AI cloud services that make it challenging for businesses to determine the costs associated with use of these services.

- Lack of guidance for solutions that utilize multiple services to address specific use cases for developers and citizen data scientists.

- Lack of understanding about how to use automated machine learning to supplement and enhance the standard language and vision services.

- Minimal marketplaces for pre-built machine learning models that could be used by developers and citizen data scientists.

- Serious lack of ModelOps capabilities that contribute to challenges in integration of AI into applications.

User Recommendations

- Choose AI cloud services over building custom models to address a broader range of use cases and for quicker deployment and built in scalability.

- Improve the chances of success of your AI strategy by experimenting with different AI techniques and AI cloud services providers, using the exact same dataset and then selecting one that best addresses your requirements. Consider using an A/B testing approach.

- Use AI cloud services to build less complex models, giving the enterprise the benefit of more productive AI while freeing up your data science assets for higher priority projects.

- Use features like automated algorithm selection, data set cleansing and preparation, and feature engineering for project elements and leverage existing expertise on operating cloud services. This will assist technical professional teams with little to no data science expertise.

Sample Vendors

Amazon Web Services (AWS); Google; IBM; Microsoft

Gartner Recommended Reading

Critical Capabilities for Cloud AI Developer Services

Magic Quadrant for Cloud AI Developer Services

Deep Learning

Analysis By: Farhan Choudhary, Alexander Linden, Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition

Deep learning (DL) is a variant of machine learning algorithms that uses multiple layers to solve problems through extraction of knowledge from raw data, and transforming it at every level. These layers incrementally obtain higher-level features from the raw data, allowing the solution of complex problems with higher accuracy, less features and less manual tuning.

Why This Is Important

Deep learning can certainly outperform traditional machine learning or shallow learning techniques while working with complex and often high-dimensional data, such as images, speech and text. With sufficient training and inference, DL reduces the need for tedious feature engineering, and can generate superior results with complex quality data (especially in cases of fraud detection, quality analysis and demand prediction).

Business Impact

Deep learning allows organizations to generate insights from disparate, especially unstructured, and, at times, limited data, from disparate data sources. All this success is rooted in the ability of DL algorithms to exploit weak signals in the dataset, which in isolation may not carry any meaning, but in a group may highlight results that would have been neglected or not even surfaced.

Drivers

- DL applicability has been most successful in the domains of vision, speech and text, across industries such as healthcare, transportation, national security, military, criminal justice, cities, finance and social media.

- Methods such as reinforcement learning, transfer learning, deep belief networks, evolutionary learning algorithms are propelling the use of deep learning in certain domains.

- Organizations looking to enrich their decision-making process by leveraging wide unstructured data such as image, audio, video or text, can leverage DL techniques.

- Availability of off-the-shelf solutions and dedicated hardware is also driving the adoption of deep learning.

- Recent advancements in NLP techniques leveraging DL methods have propelled the advancement and use of transformers, which promise state-of-the-art results in conversational platforms.

Obstacles

- The infrastructure investments required to create and maintain deep learning solutions are high.

- DL methods are construed as black box in nature, so governing and ensuring explainability of these solutions is challenging.

- DL solutions rely on the availability of high volumes of quality and correctly labelled data, which is seldom available with an average client.

- DL in visual recognition tasks rely on extracting information from pixel-based features, which could lead to undesirable or suboptimal results.

- The skills required to create and manage DL solutions from scratch is hard to get by.

- There is limited support and capabilities around security, privacy and governance for vendors providing DL capabilities as a service, which adds a layer of complexity over already black-box implementations.

User Recommendations

- Explore deep learning techniques when shallow learning techniques have failed to generalize a learning model.

- Examine and select business areas, where deep learning can provide best value, especially where there is wide and heterogeneous data.

- Create a diverse talent pool from industry and academia that can ensure interpretability as well as privacy, compliance, ethics and governance in DL solutions.

- Examine prepackaged solutions first and then move on to custom-made solutions for the business using deep learning.

- Explore adversarial learning techniques to enhance the applicability of deep learning approaches.

Sample Vendors

Facebook; Google; Landing AI; Microsoft; NNAISENSE; NVIDIA

Gartner Recommended Reading

Innovation Tech Insight for Deep Learning

Introducing Deep Learning Abstraction Methods

3 Types of Machine Learning for the Enterprise

Sliding into the Trough

Data Labeling and Annotation Services

Analysis By: Anthony Mullen

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition

Data labeling and annotation services (DLA) support classification, segmentation, transformation or augmentation procedures to enrich data for artificial intelligence (AI) projects. These services and associated platforms route and allocate tasks to both internal staff and external third-party knowledge workers.

Why This Is Important

The need for labeled data has dramatically increased in order to remove the bottleneck in developing AI solutions — especially those particular to industry use cases. Given the typical lack of internal skills and systems, DLA services are often the best option (by cost, quality and availability) to provide necessary data for best AI results.

Today, at least, some AI solutions would not be possible without human-based labeling (e.g., car driving, image recognition, search engine tuning).

Business Impact

Major impacts are:

- Scenarios that do not require deep domain knowledge can accelerate annotation by using external knowledge workers.

- While mostly used in preproduction of models the real-time human-in-the-loop solutions where models are continually trained and calibrated, such as chatbots or recommendation engines, will provide ongoing benefit.

- Business users need to join the human-in-the-loop workflows to route and train handover and moderation tasks to subject matter experts.

Drivers

We see the following drivers accelerating use of these services:

- Growth of investment in AI requires scaling data pipelines for AI. The broader investment in AI raises demand for these services.

- Growth in language automation offerings. Natural language technology workload outsourcing for speech, conversational AI and document labeling is a major area of growth in this market.

- Semantic support. Not all data that needs labeling is in row, picture or video form. The last 12 months have seen outsourcing of graph labeling and stronger use of semantic assets (e.g., ontologies) to support quicker workflow in labeling.

- Increased diversity of use cases. These services can accelerate and unlock a wealth of use cases across all industries, and core competencies in natural language automation and computer vision. Vendors in the marketplace today have dedicated offerings for commerce, robotics and autonomous vehicles, retail, GiS/maps, AR/VR, agriculture, finance, manufacturing and transportation, and communications.

Obstacles

While the supervised learning approach is predominant, DLA services' usage will grow. Obstacles include:

- **Challenger methods**. Data labeling is one of many approaches to get data for models. Few shot learning, transfer learning, synthetic data, semantic platforms and data marketplaces compete for use.

- **Challenges remain around third-party knowledge workers' quality and security** to annotate the data, somewhat ameliorated by the development of reputation systems and prequalification tests.

- **No consolidation of AI-task-outsourcing marketplaces**. The translation ecosystem, the gig economy and data labeling and annotation are as yet not a simplified coherent "language operations" for organisations.

- **Supply outstrips demand and price points are often uneconomical for large-scale data**. Many vendors have entered this space in the last year and demand from buyers does not yet match supply. Pricing and business models vary considerably among providers, and buyers find it difficult to estimate costs.

User Recommendations

- Design development and production workflows to leverage a mixture of internal and external knowledge workers.

- Ensure the provider you choose has methods to test its pool of knowledge workers for domain expertise and measures around accuracy and quality.

- Model costs to avoid surprises by exploring and estimating the spend across the variety of business models, which range from label volumes and project-based to per annotator/seat costs.

- Allow data scientists to focus their time on more valuable tasks and lighten their load in classifying and annotating data by using these services.

- Use providers with real-time human-in-the-loop solutions for production systems like chatbots and recommenders to handle low-confidence thresholds, spikes in demand or access to real-time knowledge not present in the enterprise.

Sample Vendors

Alegion; Amazon; Appen; Cloudfactory; Datasaur; Infolks; Lionbridge; Playment; Scale; Yandex

Gartner Recommended Reading

Strategic Roadmap for AI: Natural Language Architecture

Individuals, Groups and Society in the Loop of Artificial Intelligence Design and Development

Market Guide for AI Translation Services (Bern Elliot)

Natural Language Processing

Analysis By: Bern Elliot, Erick Brethenoux

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition

Natural language processing (NLP) enables an intuitive form of communication between humans and systems; NLP includes computational linguistic techniques aimed at parsing and interpreting (and sometimes generating) human languages. NLP techniques deal with the pragmatics (contextual), semantics (meanings), grammatical (syntax) and lexical (words) aspects of natural languages. The phonetic part is often left to speech-processing technologies that are essentially signal-processing systems.

Why This Is Important

NLP enables the automated processing and leveraging of vast quantities and types of text-based information. These can include documents, literature, email, text messages, invoices, receipts and so forth. With speech-to-text, NLP can also process speech, including livestreams of text and speech. As a result, NLP enables a vast array of applications and automation that previously were not achievable by machine, offering businesses significant levels of process improvement.

Business Impact

Many applications outlined in the Hype Cycle for natural language technology use NLP in some way.

- NLP is an enabler typically useful when built into applications that support business workflows.

- Because so many tasks involving text have had to rely on human labor, the potential for savings and for new business processes is vast.

- Business value reported from some applications using NLP, for instance machine translation, are thousandfold efficiency improvements and operational cost savings.

Drivers

- Basic transcription and translation services.

- Language-generation applications that produce natural language descriptions of tabular data, making it easier for many to understand.

- Keyword tagging in documents, making it easier to determine relevant sections or to extract other information such as intent and entities.

- Content moderation services that analyze user-generated content (text or images), to flag potentially offensive content or to identify fake news in social media.

- Sentiment analysis to identify the affective states and subjective information in statements — for instance from negative to neutral, to positive.

- Search improvements through better understanding of the intent of a search query as well as through summaries of the retrieved content.

- Text analytics to quickly process large numbers of organizations' documents and determine their compliance or legal validity.

- Advancement in insight engine text capabilities combined with more advanced NLP functionality.

- The introduction of new machine learning techniques, including transformer-based approaches such as BERT and GPT-2 and GPT-3. This has enabled new use cases and improvements to existing use cases.

Obstacles

- Human language is complex, dynamic and deeply influenced by cultural and other idiosyncratic conditions.

- Despite the progress made in NLP methods, there are many subtleties and nuances to properly processing the complex and enormous variety found in human languages.

- Recent NLP methods leverage deep neural networks. While the progress of neural and neural-symbolic processing is significant, many of these are experimental and fragile.

- Despite advances in new techniques, the hyped expectations surrounding NLP may result in unrealistic expectations, leading to disappointment in the actual results.

- New use cases of emerging NLP opportunities are poorly understood and face issues with meeting expectations or defining a clear business value to companies.

User Recommendations

- Select the strongest and most immediate use cases for NLP. Examples include customer service (impacting cost, service levels, customer satisfaction and upselling) and employee support (including augmenting them as they perform their tasks). Another example is automation of paper- and document-based tasks (such as contract analysis, compliance enforcement, document generation, translation and transcription).

- Demonstrate success in initial projects by starting with modest goals. As experience is obtained, projects should iterate, and scope can increase. As enterprises enhance their NLP implementations, new skills should be explored that are better able to leverage the new NLP methods.

- Verify the effectiveness of their solutions before making significant commitments, because the quality of NLP solutions will vary.

Sample Vendors

Clarabridge; IBM; Microsoft; NLTK; Rasa

Gartner Recommended Reading

Infographic: Artificial Intelligence Use-Case Prism for Customer Service

Machine Learning

Analysis By: Farhan Choudhary, Carlie Idoine, Shubhangi Vashisth

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition

Machine learning is an AI discipline that solves business problems by utilizing statistical models to extract knowledge and patterns from data. There are three major approaches that relate to the types of observation provided. These are supervised learning, where observations contain input/output pairs (also known as "labeled data"); unsupervised learning (where labels are omitted); and reinforcement learning (where evaluations are given of how good or bad a situation is).

Why This Is Important

According to Gartner's 2019 AI in Organizations survey, machine learning (ML) is the AI initiative for which more POCs and production systems are conducted. Over the past few years, ML has gained a lot of traction because it helps organizations to make better decisions at scale with the data they have. ML aims to eliminate traditional trial-and-error approaches based on static analysis of data, which is often inaccurate and unreliable, by generalizing knowledge from data.

Business Impact

Machine learning drives improvements and new solutions to business problems across a vast array of business, consumer and social scenarios like:

- Automation

- Price optimization

- Customer engagement

- Supply chain optimization

- Predictive maintenance

- Fraud detection

Machine learning impacts can be explicit or implicit. Explicit impacts result from machine learning initiatives. Implicit impacts result from products and solutions that you use without realizing they contain machine learning.

Drivers

- As organizations continue to adopt these technologies, we recently see focus on aspects that relate to ML explainability and operationalization. Augmentation and automation (of parts) of the ML development process improve productivity of data scientists and enable citizen data scientists in making ML pervasive across the enterprise.

- In addition, pretrained ML models are increasingly available through cloud service APIs, often focused on specific domains or industries.

- Data science and machine learning education is becoming a standard at many academic institutions, therefore fueling the supply of newer talent eager to venture into this space.

- There's always active research in the area of machine learning in different industries — manufacturing, healthcare, corporate legal, defense and intelligence. Thus, its applicability is far and wide.

- Newer learning techniques such as zero, one, few or end shot learning are emerging that take away the burden of having high volumes of quality training data for ML initiatives. This lowers the barrier to entry and experimentation for organizations.

- New frontiers are being explored in synthetic data, new algorithms (e.g., deep learning variations) and new types of learning. These include federated/collaborative, generative adversarial, transfer, adaptive and self-supervised learning, all aiming to broaden ML adoption.

Obstacles

■ The triggers of its massive growth and adoption have been growing volumes of data, advancements in compute infrastructure and the complexities that conventional engineering approaches are unable to handle.

■ Even though ML is one of the particularly popular AI initiatives in the last few years, it is not the only one. Organizations also tend to rely on other AI techniques such as rule-based engines, optimization techniques, physical models to achieve decision augmentation or automation.

■ A significant portion of ML models at an organization doesn't make it into production, therefore adding to technical debt and risks mistrust in the initiative, often delaying value realization from ML at organizations.

■ The application of ML is often oversimplified as just model development but it's not so. Several dependencies which are overlooked, such as data quality, security, legal compliance, ethical and fair use of data, serving infrastructure, and so forth, have to be considered in ML initiatives.

User Recommendations

■ Build up and extend descriptive analysis toward predictive and prescriptive insights, which can be excellent candidates for machine learning.

■ Assemble a (virtual) team that prioritizes machine learning use cases, and establish a governance process to progress the most valuable use cases through to production.

■ Utilize packaged applications if you find one that suits your use case requirements. These often can provide superb cost-time-risk trade-offs and significantly lower the skills barrier.

■ Explicitly manage MLOps and ModelOps for deploying, integrating and monitoring analytical, ML and AI models.

■ Adjust your data management and information governance strategies to enable your ML team. Data is your unique competitive differentiator, and adequate data quality, such as the representativeness of historical data for current market conditions, is critical for the success of ML.

Sample Vendors

Amazon Web Services (AWS); Databricks; Dataiku; DataRobot; Domino; Google Cloud Vertex AI; H2O.ai; Microsoft Azure; SAS; TIBCO Software

Gartner Recommended Reading

Magic Quadrant for Data Science and Machine Learning Platforms

Critical Capabilities for Data Science and Machine Learning Platforms

Toolkit: RFP for Data Science and Machine Learning Platforms

3 Types of Machine Learning for the Enterprise

Understanding MLOps to Operationalize Machine Learning Projects

Intelligent Applications

Analysis By: Alys Woodward

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition

Intelligent applications (IAs) are enterprise business applications with embedded or integrated AI technologies, such as intelligent automation, data-driven insights and guided recommendations, that can deliver a more personalized interface, improve productivity and support decision making.

Why This Is Important

AI is the next major battleground for enterprise applications with many technology providers now incorporating AI and ML capabilities in their products. Enterprise application vendors are actively embedding AI technologies within their offerings — from ERP to CRM and HCM to productivity applications. Adding intelligence into applications, instead of more procedural features, allows applications to support decision-making processes alongside transactional processes.

Business Impact

Intelligent applications transform many core business processes by:

- Improving user experience through significant degrees of personalization and new interaction channels

- Creating efficiencies through automation

- Improving decision making through intelligent decision support within the application

- Scaling the ability to capture business value in core business processes

- Increasing the ability to work with many ecosystem partners to deliver value

- Decreasing risk through fraud or other risk detection

Drivers

- Organizations are demanding more and more functionalities from applications, whether built or bought, expecting them to enhance current processes for both transactions and decision-making with recommendations, insights and additional information.

- The trend toward composable business architectures has highlighted the possibilities around delivering advanced and flexible capabilities to support, augment and automate decisions. This has increased customers' appetite for intelligent augmentation within applications. However, to build this requires an underlying data fabric and packaged business capabilities.

- AI capabilities and features are increasingly being integrated into ERP, CRM, supply chain and knowledge management software within enterprise application suites.

- Embedded intelligence is typically a part of a core enterprise application, not a separate technology product. From the provider's perspective, including AI helps improve customer loyalty and customer lifetime value by extending and enhancing their applications with AI, rather than a separate revenue stream.

- The most important enabler for intelligent applications is ML, which allows features like recommendations, insights and personalization. Conversational UIs, while a part of AI, are more about improving the interface and less about adding intelligence to the application. As intelligent virtual assistants become more widely incorporated by application platforms, the line between interface and intelligence will blur.

- Emerging vendors in a number of enterprise domains are designed with AI first (known as AI-native applications), competing directly with suite-based incumbents in those same domains.

- Intelligent applications were the most popular technology in a survey of organizations implementing AI, and almost one-fifth of organizations surveyed intended to adopt intelligent applications in the next 12 months.

Obstacles

- **Lack of necessary data** — Intelligent applications require access to data from a range of systems, meaning application vendors need to think about data management technology and processes outside their own solutions.

- **Adding AI to the existing applications is often complex** — The models have to be trained and maintained, and the user has to understand the latency of the data. Insights generated with AI have to be contextualized for the process that the user is executing, which requires additional business metadata.

- **Overuse of "AI" in marketing** — Vendors have a tendency to overuse the term "AI" in marketing and neglect the focus on business impact, which can generate a cynical response in business buyers.

- **Trust in system-generated insights** — It takes time for business users to see the benefit and trust such insights, and they need some understanding of how the decision was taken in terms of inputs and logic.

User Recommendations

- Challenge your packaged software providers to outline in their product roadmaps how they are incorporating AI to add business value in the form of a range of AI technologies.

- Evaluate the architecture of your providers by considering that the best-in-class intelligent applications are built from the ground up to be constantly collecting data from other systems, with a solid data layer in the form of a data fabric.

- Prioritize investments in highly specialized and domain-specific intelligent applications delivered as individual point solutions, which help solve problem areas such as customer engagement and service, talent acquisition, collaboration, engagement and more.

- Bring AI components into your composable enterprise thinking to innovate faster and safer, to reduce costs, and to lay the foundation for business-IT partnerships. Remain aware of what makes AI different, particularly how to refresh and rebuild machine learning models, as this can cause implementation and usage challenges.

Sample Vendors

BizMerlinHR; Eightfold AI; JAGGAER; Salesforce; Sievo

Gartner Recommended Reading

Emerging Technologies: AI Survey Indicates That Vendors Must Take a Long-Term View

Predicts 2021: Artificial Intelligence in Enterprise Applications

Top Trends in Data and Analytics for 2021: Composable Data and Analytics

Chatbots

Analysis By: Magnus Revang

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Early mainstream

Definition

Chatbots are domain-specific or task-specific conversational interfaces that use an app, messaging platform, social network or chat solution for conversations. Chatbots range in use-case sophistication from simple, decision-tree-based, to implementations built on feature-rich platforms. They are always narrow in scope. A chatbot can be text-based or voice-based, or a combination of both.

Why This Is Important

Chatbots represent the No. 1 use of artificial intelligence (AI) in enterprises. Primary use cases are in customer service, human resources, IT help desk, self-service, scheduling, enterprise software front ends, employee productivity and advisory. Offerings in the market include developer self-service platforms, managed products, middleware offerings, integrated offerings and best-of-breed approaches.

Business Impact

Chatbots are the face of AI and will impact all areas with communication between machines and humans. Customer service is an area where chatbots are already very influential and will have a great impact on the number of service agents employed by an enterprise and how customer service is conducted. The change from "the user learns the interface" to "the chatbot learns what the user wants" has implications for onboarding, training, productivity and efficiency inside the workplace.

Drivers

- Chatbots in social media, service desk, HR or commerce, as enterprise software front ends and for self-service, are all growing rapidly.

- For enterprises, the main challenge with chatbots has been scaling and operationalizing them out of the proof-of-concept phase. As COVID-19 has accelerated adoption of chatbots, vendors seem to have "cracked the code" on operationalization. Vendors are now able to deliver multiple bots for multiple use cases, with no-code environments allowing multiple roles to participate in operationalization. This is creating a market for enterprise conversational AI platforms fueling the next generation of chatbots.

Obstacles

- Scaling and operationalizing still remain a challenge in some cases, due to lack of dedicated internal teams to work on continuous improvements.

- Figuring out the composition of teams, and the methodologies to iterate effectively, are still emerging practices with strong vendor dependency.

- Technology is improving at an astounding pace, but best practices on adoption and use of these technological advancements are still trailing, resulting in a lot of trial and error for enterprises.

- Selected vendors are sometimes unable to keep pace with the technology and the market dynamics.

- The vendor landscape comprises over 2,000 vendors, despite some consolidation during 2020. However, this is composed of many subcategories, majority of which are tactical. With this many vendors, the majority of chatbots will have to switch their underlying technology in the near to midterm future. Still a category of enterprise-grade platforms has emerged, with an estimated 120 vendors. These enterprise-grade platforms are becoming suitable as a more tactical choice.

User Recommendations

- Select an enterprise-grade platform to develop multiple use cases with orchestration of the assets needed.

- Focus on operationalization of chatbots as a product, with the necessary organization and roles in place, to evolve and maintain chatbots over time.

Sample Vendors

Amazon; Amelia; Cognigy; Google; IBM; Kore.ai; Microsoft; Pypestream; ServisBOT; Uniphore

Gartner Recommended Reading

[The 3 Decisions You Must Make Before You Begin a Chatbot Project](#)

[Consolidate Your Chatbot Initiatives Into a Single Enterprise Strategy](#)

[When Should I Use Embedded Conversational Assistants?](#)

Autonomous Vehicles

Analysis By: Jonathan Davenport

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition

Autonomous vehicles use various onboard sensing and localization technologies, such as lidar, radar, cameras, GPS and map data, in combination with AI-based decision making, to drive without human intervention. While self-driving passenger cars are getting most of the attention these days, the technology can also be applied to vehicles that transport goods.

Why This Is Important

Autonomous vehicles have the potential to change transportation economics, cutting operational costs and increasing vehicle utilization. In urban areas, cheap fares and high quality of service may cannibalize private car ownership. Road safety will also be increased as the AI systems will never be distracted, drive drunk or speed. Autonomous features on privately owned vehicles enable productivity and recreational activities to be undertaken, while the vehicle handles the driving operations.

Business Impact

■ Autonomous vehicles have the potential to disrupt established automotive business models.

■ Technology companies are building high-performance computers on which to run their self-driving software platforms.

■ After the office and home, vehicles will become a living space, like airplanes, where digital content is both created and consumed.

■ Over time, staff members currently undertaking driving roles must be retrained and redeployed to other, higher-value-adding roles within the company.

Drivers

■ Some progress is being made toward autonomous vehicle regulations and standards. Automated lane-keeping system (ALKS) technology has been approved by the United Nations Economic Commission for Europe (UNECE). This forms the first binding international regulation for SAE Level 3 vehicle automation, with a maximum operational speed of 37 mph. Likewise, the German government aims to enact laws that enable autonomous vehicles to operate without special permits by 2022. Companies like Intel, Waymo and Aurora are working on the IEEE 2846, which will create a standard that describes the scenarios to be considered when developing autonomous road-safety-related models.

■ To take advantage of the new regulatory landscape, automakers are beginning to announce Level 3 solutions. These autonomous vehicles provide drivers with safety and convenience features, reduce vehicle fuel consumption and improve traffic management. Honda is the first company to announce a commercially available Level 3 vehicle, though only 100 will be produced.

■ Improvements are also being made to the perception algorithms and broader self-driving systems for Level 4 vehicles that will operate as robotaxis. Fully driverless operations have started, with Waymo operating in Arizona and WeRide operating in California without safety drivers. The flexibility of vehicle operational design domains (ODDs) has been showcased — e.g., Mobileye's perception algorithm required minimal additional training when it tested vehicles in new locations. Mobileye has developed its self-driving software on the roads in Israel, but showcased its autonomous technology in both Munich and Detroit. Likewise, Yandex has made great strides, showcasing how its autonomous vehicles are capable of handling the harsh weather conditions of winters in Moscow.

Obstacles

- Designing an AI system that is capable of driving a vehicle is hugely complex. As a result, the cost of bringing a commercial autonomous vehicle to market has been greater than companies could have previously envisioned. This has required significant investments to be made in companies. Acquisitions have occurred, and further market consolidation is expected — e.g., Walmart has invested $2.75 billion in Cruise; Cruise acquired Voyage in March 2021; Aurora acquired Uber's ATG in December 2020; Amazon acquired Zoox for $1.2 billion in June 2020; Apple bought self-driving startup Drive.ai in June 2019.

- When autonomous vehicles are commercially deployed, autonomous vehicle developers, not the human occupants, will be liable for the autonomous operations of the vehicle. This raises important issues, should a vehicle be involved in an accident.

- Challenges increasingly include regulatory, legal and societal considerations, such as permits for operation and the effects of human interactions.

User Recommendations

Governments must:

- Craft national legislation that ensures that autonomous vehicles can safely coexist with an older fleet of nonautonomous vehicles.

Autonomous mobility operators should:

- Support consumer confidence in autonomous vehicle technology by remaining focused on safety to deliver on the vision of an accident-free road environment.

Self-driving system developers should:

- Seek out use cases, such as mining, agriculture or airports, where autonomous vehicles can operate in restricted areas safely without regulatory restrictions. Use these implementations to drive early revenue and gather data and insights to improve the performance of self-driving systems.

Traditional fleet operators looking to adopt autonomous technology into their fleets should:

- Minimize the disruptive impact on driving jobs (bus, taxi and truck drivers) by developing policies and programs to train and migrate these employees to other roles.

Sample Vendors

Baidu; Cruise; Mobileye; Waymo; Yandex; Zoox

Gartner Recommended Reading

Market Trends: Monetizing Connected and Autonomous Vehicle Data

Forecast Analysis: Autonomous Vehicle Net Additions, Internet of Things, Worldwide

Utilize Partnerships to Secure a Winning Position in the Autonomous Driving Ecosystem

Market Insight: Use Situationally Aware Platforms to Enable Safe Autonomous Vehicle Handovers

Tech Providers 2025: Product Leaders Must Strategize to Win in the Evolving Robotaxi Ecosystem

Computer Vision

Analysis By: Nick Ingelbrecht

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition

Computer vision is a process and set of technologies that involve capturing, processing and analyzing real-world images and videos to allow machines to extract meaningful, contextual information from the physical world.

Why This Is Important

Computer vision comprises a transformational collection of technologies that are becoming essential to understanding the physical environment. Computer vision technology is driving innovation across many industries and use cases and is pushing the business application of artificial intelligence to new frontiers.

Business Impact

Computer vision technologies are used across all industries and address a broad and growing range of business applications. Applications include physical security, retail and commercial property, automotive, robotics, healthcare, manufacturing, supply chain/logistics, banking and finance, agriculture, government, media and entertainment, and IoT. They are used in visible and nonvisible spectrum, including thermographic systems for remote fever and vital signs detection and facial recognition.

Drivers

Computer vision adoption is being driven by improvements in the application of machine learning methods, tooling and services, hardware processing efficiencies, and data generation and augmentation techniques:

- New neural network architectures, model and algorithm enhancements are steadily improving the price/performance of computer vision applications.

- The economics of computer vision are being enhanced by the rapid expansion of the market for computer vision tools and services. These include annotation and data preparation services and AutoML capabilities, reaching across computer vision data pipelines, from model development and training through to deployment and model management, maintenance, and governance.

- The proliferation of cameras and other sensors is generating exponential increases in image data, creating a critical and growing demand for methods to automate analysis and manage and extract value from that data.

- Edge-enabled cloud frameworks, developer ecosystems, products and support are further expanding the opportunity space and enabling nonexperts to train and deploy their own computer vision models.

- There is a long tail of innovation and adoption across all industry sectors. Technology advances in computer vision are enabling new business models, ranging from smartphone cameras and fun filters, through to global video content production and distribution, life-saving medical image diagnostics, autonomous vehicles, video surveillance for security, robotics, and manufacturing automation.

Obstacles

Enterprises struggle with how best to exploit their visual information assets and automate the analysis of exponential volumes of image data:

- High-end systems are expensive to maintain and support, and building business cases with adequate ROI is challenging.

- The computer vision market lacks independent standardization and performance benchmarks, and advanced solutions are far from being commoditized.

- Integration with existing systems is problematic due to a lack of open interfaces, off-the-shelf solutions and plug-and-play capabilities.

- Enterprises struggle to activate computer vision models in business processes and face data security and privacy risks.

- Scaling solutions is challenging due to the high levels of customization and service support needed.

- Adequate training and testing data may be hard or expensive to acquire, especially in areas where available open-source computer vision datasets are declining.

- Proprietary algorithms and patent pools deter innovation.

User Recommendations

- Assess change management impacts of computer vision projects on the organization and its people.

- Focus initially on a few small projects, using fail-fast approaches and scale the most promising systems into production using cross-disciplinary teams.

- Test production systems early in the real-world environment because lighting, color, object disposition and movement can break computer vision solutions that worked well in the development cycle.

- Build internal computer vision competencies and processes for exploiting image and video assets. This will enable the organization to make better procurement choices and lay the groundwork for more advanced innovation and product development opportunities.

- Exploit third-party computer vision tooling and services to accelerate data preparation and reduce costs.

- Evaluate legal, regulatory, commercial and reputational risks associated with computer vision projects at the outset.

Sample Vendors

Amazon Web Services; AnotherBrain; Baidu; Clarifai; Deepomatic; Google; Matroid; Microsoft; nyris; Tencent

Gartner Recommended Reading

Emerging Technologies: Tech Innovators for Computer Vision

Emerging Technologies Tool: Video Analytics Functionality Matrix

Emerging Technologies: Video Analytics Functionality Spectrum, 2021

Emerging Technologies: Top Advanced Computer Vision Use Cases for Retail

Climbing the Slope

Semantic Search

Analysis By: Stephen Emmott

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition

Semantic search improves the relevance of search by processing the relationship between words — as a proxy for meaning — in addition to the words themselves, i.e., "things, not strings." This serves to mediate between intent and outcome, raising and leveling the value of search for end users. Semantic search is a key part of insight engines as it augments the search technology that underpins them.

Why This Is Important

Search performance is essential for effective knowledge management, records management and support of automation involving data sourced from content in documents and records. Semantic search amplifies performance by analyzing the underlying meaning of documents/records, as well as the queries posed to retrieve them. This amplifies productivity through employee time spent, opportunity through connections and new ways of working, and reduces risks such as missing essential documents/records.

Business Impact

Semantic search impacts value streams across business functions and industry verticals. Use-case scenarios include:

- Facilitating employee experience by connecting employees to one another, or to information, based on their expertise, activities or need for knowledge.

- Facilitating customer experience by improving self-help in support of presale decisions or postsale support.

- Extending automation to processes currently restricted to employees receiving and processing documents manually.

Drivers

Progress is driven by:

- The continued need for improved search to support employee experience in the context of the digital workplace.

- Refining customer experience to stay ahead of expectations as well as increasing value and reducing costs.

- The pursuit of both augmentation and automation to create new ways of working for both employees and machines.

- The capability to extract textual content from audio and visual sources, thereby extending the reach of search to all content.

- Advances in word embedding (an application of machine learning) as a natural language processing technique to enable meaning to be processed reliably in the workplace.

- Advances in knowledge graphs, enabling explicit representation of the entities words represent, their attributes and relationships.

For these reasons and more, semantic search has reached 5 to 20% of its target audience as it leaves the Trough of Disillusionment and begins its ascent up the Slope of Enlightenment.

Obstacles

The progress of this innovation is obstructed by:

- Use and performance tend to be stronger in specific domains rather than across all, which holds back use in domains with fewer customers.

- While word embedding is flexible, rule-based approaches require further development, meaning those products that utilize a hybrid approach tend to be limited to specific languages.

- Performance improves with application, but much of this relates to confidential sources, resulting in confidential learning and adaptation that cannot be shared beyond individual customers.

- Not all products utilize a hybrid approach to techniques, which must involve a combination of techniques.

- Professional services from vendors and/or partners is essential to get started and, for many, to continue.

- Products that utilize knowledge graphs require them to be developed and maintained.

- Products that use word embedding require initial and subsequent training against content sources to remain current.

User Recommendations

To exploit this innovation:

- Treat search as a capability augmenting ways of working across multiple applications.

- Identify search capabilities across all applications and determine the level of semantic search provided, either at present or in terms of the roadmap.

- Steer review and selection processes to ensure your search and insight services include semantic search as a capability.

- Coordinate the development and maintenance of search and insight services to ensure a common hand behind configuration and training.

- Link your search and insight activities with your other NLT initiatives and in the context of the organization's wider data fabric to seek rationalization and consolidation where appropriate.

- Test the performance of semantic search in context. The performance of semantic search is highly dependent on its "semantic fabric."

- Engage and employ subject matter experts to contribute, ideally independently and proactively, to the maintenance of semantic search capabilities.

Sample Vendors

Expert.ai; Google; IBM; IntraFind; Microsoft; Ontotext; Semantic Web Company; ServiceNow
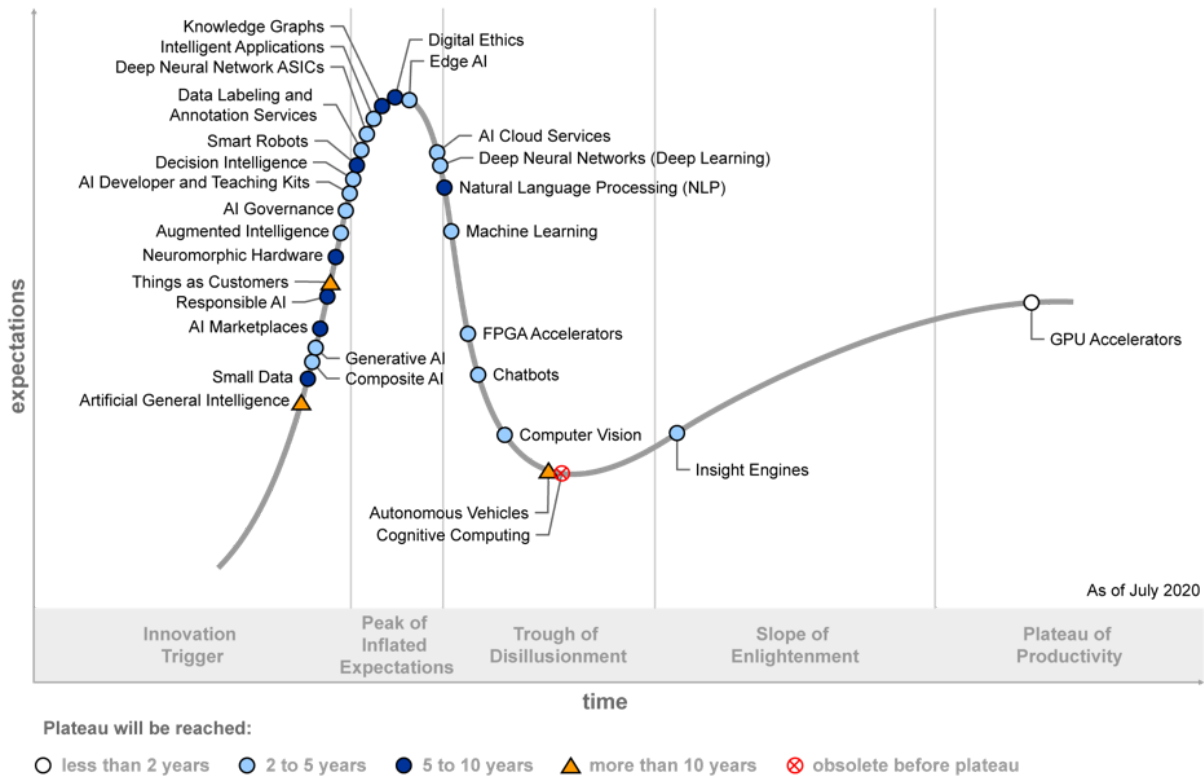
Gartner Recommended Reading

Magic Quadrant for Insight Engines

Critical Capabilities for Insight Engines

Appendixes

Figure 2: Hype Cycle for Artificial Intelligence, 2020

Source: Gartner (July 2020)

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 2: Hype Cycle Phases

(Enlarged table in Appendix)

Phase ↓	Definition ↓
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Source: Gartner (July 2021)

Table 3: Benefit Ratings

Benefit Rating ↓	Definition ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2021)

Table 4: Maturity Levels

(Enlarged table in Appendix)

Maturity Levels ↓	Status ↓	Products/Vendors ↓
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constraints replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source:Gartner (July 2021)

Document Revision History

Hype Cycle for Artificial Intelligence, 2020 - 27 July 2020

Hype Cycle for Artificial Intelligence, 2019 - 25 July 2019

Hype Cycle for Artificial Intelligence, 2018 - 24 July 2018

Hype Cycle for Artificial Intelligence, 2017 - 24 July 2017

Hype Cycle for Smart Machines, 2016 - 21 July 2016

Hype Cycle for Smart Machines, 2015 - 24 July 2015

Hype Cycle for Smart Machines, 2014 - 18 July 2014

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Understanding Gartner's Hype Cycles

Create Your Own Hype Cycle With Gartner's Hype Cycle Builder

Hype Cycle for Natural Language Technologies, 2021

Hype Cycle for Analytics and Business Intelligence, 2021

Artificial Intelligence Maturity Model

A Framework for Applying AI in the Enterprise

Magic Quadrant for Data Science and Machine Learning Platforms

Magic Quadrant for Cloud AI Developer Services

Table 1: Priority Matrix for Artificial Intelligence, 2021

Benefit	Years to Mainstream Adoption			
	Less Than 2 Years	2 - 5 Years	5 - 10 Years	More Than 10 Years
Transformational		Composite AI Computer Vision Decision Intelligence Deep Learning Edge AI Generative AI Human-Centered AI Intelligent Applications Machine Learning	Multiexperience Natural Language Processing Neuromorphic Hardware Physics-Informed AI Responsible AI Transformers	Artificial General Intelligence Autonomous Vehicles
High	Chatbots	AI Cloud Services AI Governance AI Maker and Teaching Kits AI Orchestration and Automation Platform Deep Neural Network ASICs Synthetic Data	AI TRiSM Digital Ethics Knowledge Graphs Model Compression ModelOps Small and Wide Data Smart Robots	Machine Customers
Moderate		Data Labeling and Annotation Services Semantic Search		
Low				

Source: Gartner (July 2021)

Table 2: Hype Cycle Phases

Phase ↓	Definition ↓
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Phase ↓	Definition ↓

Source: Gartner (July 2021)

Table 3: Benefit Ratings

Benefit Rating ↓	Definition ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2021)

Table 4: Maturity Levels

Maturity Levels ↓	Status ↓	Products/Vendors ↓
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constraints replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source:Gartner (July 2021)