Hype Cycle for Data Science and Machine Learning, 2023

Published 27 July 2023 - ID G00792722 - 124 min read

By Analyst(s): Peter Krensky

Initiatives: Analytics, BI and Data Science Solutions; Evolve Technology and Process Capabilities to Support D&A

The data science and machine learning landscape is reacting to the massive hype this year has brought to generative AI. DSML and its adjacent fields are enjoying the spotlight and data and analytics leaders are more receptive to novel techniques and cognizant of accompanying challenges.

More on This Topic
This is part of an in-depth collection of research. See the collection:

- 2023 Hype Cycles: Deglobalization, AI at the Cusp and Operational Sustainability

Analysis

What You Need to Know

This document was revised on 22 August 2023. The document you are viewing is the corrected version. For more information, see the Corrections page on gartner.com.

As generative AI and other concepts garner unprecedented spotlight and eye-popping early investment, the data science space basks in the glow as it continues on what was already an exciting trajectory. At the technology and technique level, innovation is abundant and adoption only bolstered by modern AI entering the mainstream.

Gartner sees major innovations in data science and machine learning (DSML) coming from corporate research labs, open-source communities and academia. Organizations are fully engaged in exploring machine learning techniques and solutions, often with the help of vendors, systems integrators and consultants.

In addition to this Hype Cycle, D&A leaders should consult the following Hype Cycles in adjacent areas:

- Hype Cycle for Analytics and Business Intelligence, 2023
- Hype Cycle for Artificial Intelligence, 2023
- Hype Cycle for Natural Language Technologies, 2023
- Hype Cycle for Data Management, 2023
- Hype Cycle for Data and Analytics Governance, 2023
- Hype Cycle for Data and Analytics Programs and Practices, 2023
- Hype Cycle for Privacy, 2023

The Hype Cycle

Three major storylines stand out in this year's Hype Cycle:

- **Generative AI dominates the conversation and the Peak of Inflated Expectations:** There has never been a year of hype in DSML to match this one. Generative AI has captured the attention and imagination of the global public and data science community. None of the hype represented on this Hype Cycle comes close to that currently surrounding "generative AI" and "foundation models." This hype has also introduced new innovation profiles such as prompt engineering and machine learning code generation. For more on hype specific to generative AI, please see Hype Cycle for Artificial Intelligence, 2023.

- **Generative AI influences and even resets the hype around other innovations and technologies:** This Hype Cycle is replete with innovations that are being disrupted and transformed by generative AI. This has frozen or even regressed the journey through the Hype Cycle for innovations such as augmented analytics, natural language query, citizen data science, MLOps, explainable AI and data science education. These will now become thoroughly infused with generative AI before they continue on toward the trough and eventually plateau.

- **Highly hyped data science concepts, technologies and techniques climb a steepening slope toward a rising peak:** There is so much more happening in data science than breakthroughs in generative AI. Innovations that were steadily rising toward the peak are still doing so, though the summit is now higher and the ascent slower as generative AI has its moment. This grouping of key technologies and techniques includes transfer learning, reinforcement learning, federated machine learning, multistructured analytics, composite AI and graph data science. These innovations are of great importance and interest to practitioners at the technical level as most current attention is on higher-level concepts in AI.

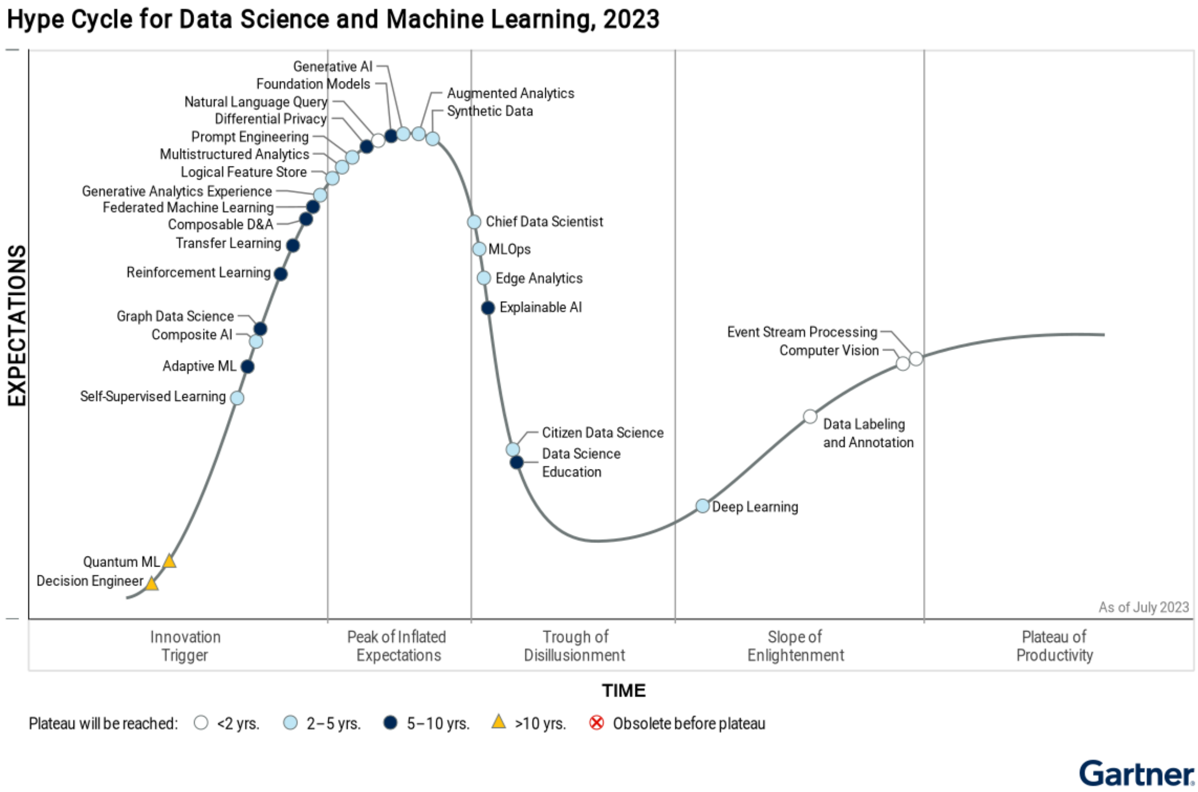
The Trough of Disillusionment is bare as deep learning begins to climb the Slope of Enlightenment and generative AI keeps expectations high for many postpeak profiles. The Peak of Inflated Expectations is owned by the story of generative AI. The Innovation Trigger remains crowded and active, with longstanding innovations continuing to gain attention and maturity and the emergence of intriguing new concepts with manifest short and long-term impact.

The following innovations are new to this Hype Cycle this year:

- Decision engineer
- Edge analytics
- Generative AI

- Generative analytics experience
- Prompt engineering

Figure 1: Hype Cycle for Data Science and Machine Learning, 2023



The Priority Matrix

The Priority Matrix arranges each innovation profile by its level of benefit attainable versus its predicted time to plateau. The darker the color, the more attention D&A leaders should pay (generally) to those innovations.

Innovations of transformational benefit will dramatically boost the productivity of expert data scientists and lower the barriers to starting data science projects. This will, however, increase the need for strong MLOps and the demand for validation and governance. Composable D&A will allow organizations to have flexibility and heterogeneity in their DSML systems, which are more vital than ever with the diverse explosion of generative AI technologies.

Innovations of high benefit help organizations refine and optimize their data science initiatives as they work to address a wider spectrum of use cases. This means better systems, stronger leaders and a broader set of techniques. Organizations will also demand innovations that address strategic challenges such as privacy, security and risk.

Table 1: Priority Matrix for Data Science and Machine Learning, 2023

(Enlarged table in Appendix)

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational	Computer Vision Event Stream Processing	Citizen Data Science Composite AI Deep Learning Generative AI Self-Supervised Learning	Composable D&A Foundation Models	Quantum ML
High	Data Labeling and Annotation Natural Language Query	Augmented Analytics Chief Data Scientist Edge Analytics Generative Analytics Experience MLOps Multistructured Analytics Prompt Engineering Synthetic Data	Adaptive ML Data Science Education Explainable AI Federated Machine Learning Graph Data Science Reinforcement Learning Transfer Learning	Decision Engineer
Moderate		Logical Feature Store	Differential Privacy	
Low				

Source: Gartner (July 2023)

Off the Hype Cycle

This Hype Cycle is intended to provide insight into only the most important and prevalent concepts in the field of DSML today. As a result, the following entries in Hype Cycle for Data Science and Machine Learning, 2022 were modified or removed in this year's edition, and a few remain relevant and important in the analytics field and appear within other innovation profiles:

- **Augmented DSML:** The introduction of generative AI to the already very healthy hype around augmented DSML and augmented analytics has helped to consolidate these technologies. The augmented analytics profile now captures the hype for augmented DSML as the vendors in this space rapidly innovate around generative AI.

- **Generative adversarial networks:** The hype around this specific technique has been subsumed by a combination of highly hyped innovations.
- **tinyML:** Hype related to tinyML in this context is best captured in the edge analytics innovation profile in this Hype Cycle.

On the Rise

Decision Engineer

Analysis By: David Pidsley

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Decision engineers apply analytics and engineering skills to decision intelligence platforms and practices. This practical discipline advances decision-making experiences with design thinking by engineering how decisions are modeled and made and how outcomes are evaluated and improved via feedback. Practitioners foster multidisciplinary collaboration in decision support, augmentation and automation by applying decision engineering to processes with embedded analytics, data science and AI.

Why This Is Important

Decision engineer is a decision-centric role that operationalizes embedded decision models, monitors feedback and optimizes outcomes with decision intelligence practices and platforms. The emerging role is not only focused on implementation. It is also essential in fostering multidisciplinary collaboration to bridge the gap between business domain and process experts on the one hand, and data scientists and AI experts on the other. They collaborate with the business, inventory decision models to manage and monitor, and report on the business value of decisions under management to promote reuse and deduplication of decision models.

Business Impact

Most decisions that currently use data are or soon will be at least partially automated. Decision engineers apply process, (computational) software engineering and mathematical techniques to help organizations make better decisions. Decision engineering leverages data, analytics, DSML, optimization and simulation to support decision augmentation and automation across a range of industries and contexts.

Drivers

- Though they recognize the need to collaborate, executives report that too many stakeholders and unclear decision ownership cause problems and delays in taking action. Instead of supporting multiple decision types for a single business unit, decision engineers can support a specific decision type, such as cost management or product improvements, across a number of business units.

- The shift from data-driven to decision-centric enterprise accelerates the demand for emerging roles that apply analytics and engineering skills to decision intelligence platforms and practices.

- Gartner identifies decision intelligence as a top strategic technology trend that is disrupting decision-making culture, and a decision engineer describes “who” plays a key role in this.

- Data and analytics leaders are upgrading their operating models, especially for organizations and people, to ensure they can enable dynamic business outcomes amid disruptive market conditions.

- Challenges in executing high-impact reengineering of decisions will accelerate common definitions of embryonic roles to have a high benefit and mature into productivity in the coming decade.

- Decision engineers bring a deeper understanding of how effective decision-making processes work, and they provide human and social perspectives. Some top data science teams will be rebranded as cognitive science or science consultancies, increasing diversity in staff skills.

- Skills in demand for decision engineering include data science, simulations, optimization, SQL, Python, R, DAX, VizQL, process methods, software engineering techniques, design thinking and communication skills.

Obstacles

- Decision engineers may become a role, but not a job title. Similarities to other roles like citizen data scientist, data steward or D&A translator lead to confusion among candidates and hiring managers.

- Embryonic roles have less than 1% of target market adoption and are undifferentiated from adjacent D&A and decision-making roles. Despite different focuses and responsibilities, these roles all involve working with data and using quantitative methods to augment and reengineer decision making.

- Skill and staff shortages are the top roadblocks for success in D&A initiatives. Late adopters will struggle to recruit decision engineers and must instead focus on upskilling, motivating and retaining decision intelligence talent.
- Failure to operationalize decision intelligence or embed decision models into workflows and business processes makes for ineffectively integrated decision making.
- Not employing adaptive governance of decisions to ensure ongoing optimization of business outcomes by establishing clear decision-making processes, proactively identifying and addressing issues, continuously refining and optimizing decision models based on data-driven insights, and aligning decision intelligence with the organization's goals and values.
- Existing organizational structures silo decision intelligence approaches which, in fact, go across domains. Tactical, functional decisions are often compartmentalized by technology vendor or product (e.g., CRM, ERP, HCM, FP&A).
- Localized implementations may create fragmentation in organizational units, where decisions are very similar but regulated differently.

User Recommendations

D&A leaders responsible for analytics, BI and data science solutions should:

- Evolve their D&A approaches to support data-driven decisions by empowering and supporting business units to embed D&A in business processes.
- Assess the impact of the transition from data-driven to decision-centric and update your operating model for decision intelligence practices.
- Assess the impacts that demand for decision engineering will have upon existing skills shortages and how to fill the role by considering how other companies do this (see sample vendors).
- Define decision engineers' roles, responsibilities, requirements and qualifications (a bachelor's or master's degree in computer science, mathematics, statistics, operations research or a related discipline).
- Foster and develop decision intelligence talent to address staff shortages by recruiting decision engineers and data scientists, forming fusion teams with business experts and fostering communities of practice.

- Define the role's key responsibilities as collaboration with business functions outside of D&A, decision modeling using frameworks, decision model management (especially deduplication, reuse and mitigation of model drift), valuation and data storytelling, and continuous learning and trendspotting.
- Involve relevant stakeholders in the business, D&A and adjacencies in a collaborative way by applying best practices to fill the role through upskilling, attracting recruits, motivating and retaining decision engineers.

Sample Vendors

Airbnb; Amazon; Google; LinkedIn; Meta; Microsoft; Netflix; Philips; Uber

Gartner Recommended Reading

What Are the Essential Roles for Data and Analytics?

Maverick Research: Data and Analytics Roles Will No Longer Be a Priority

The Future of Data and Analytics: Create Competitive Differentiation Through Better Decision Making

Predicts 2023: Analytics, BI and Data Science Composability and Consolidation

Redefining Analysts as Decision Experts (Philips)

Quantum ML

Analysis By: Chirag Dekate, Matthew Brisse

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Quantum machine learning (QML) combines the principles of quantum mechanics and machine learning (ML) to develop new algorithms, hybrid methods and new processing techniques to analyze data faster, with more accuracy and efficiency than classical approaches alone. As of this writing, there are no competitive advantages using QML today due to hardware limitations. There is a large amount of research dedicated to the QML space for quantum computing and quantum simulation.

Why This Is Important

ML techniques, including quantum neural networks, or quantum kernel methods, rely on two key parts: quantum embedding of data that maps feature space using quantum approaches and an evaluation function applied to data embedding. Theoretically, quantum embedding of feature space may result in a quantum advantage for a subset of problems over classical alternatives.

Quantum ML is nascent, and increasing awareness of quantum ML plays a key role in determining its potential value.

Business Impact

Quantum ML continues to be in an embryonic stage, with most R&D activities clustered around devising quantum algorithms for key ML kernels. However, the scale of the systems and algorithms and the challenges associated with “data loading” will limit adoption in the near term. Potential applications of quantum computing in AI and ML include quantum search, recommendation algorithms, quantum algorithms for game theory,and quantum algorithms for decisions and learning.

Drivers

- Early research in developing quantum ML initially indicated the potential for applicability across a growing set of ML algorithms, including k-means, k-medians, hierarchical clustering, principal component analysis, neural networks, support vector machines, nearest neighbors, regression and boosting.
- However, new research in this ever-evolving field seems to call into question the potential applicability of quantum computing in ML. Additionally, considerable hardware and software challenges remain.
- R&D today is focused on developing different quantum algorithms for ML kernels. Vendors such as Google, IBM, Rigetti Computing and Xanadu have prototype ML algorithms implemented for very selected use cases.
- Developing scalable ML systems will require many qubits and fundamental advances in applicable quantum algorithms.
- There are five well-known QML algorithms in use today: Quantum Principal Component Analysis (PCA), Quantum Support Vector Machines (SVM), Quantum k-Nearest Neighbors (k-NN), Quantum Boltzmann Machines (QBM) and Variational Quantum Classifier (VQC).
- Vendors and academia are investigating leveraging quantum simulation on classical machines for specific use cases with promising results that are limited by noisy intermediate-scale quantum (NISQ) computers.

Obstacles

While quantum ML is theorized to work effectively in NISQ computers, it is not ready for mainstream adoption today, with key obstacles including:

- **A nascent quantum computing ecosystem —** Quantum computing is still at a very early stage of development, with many systems offering scaling limited to tens of qubits. As a result, algorithms executed on these systems are primarily exploratory in nature.
- **Data encoding —** Although quantum computing can hypothetically deliver dramatic boosts for certain classes of data, one challenge is encoding input data. For quantum ML to work at scale, large amounts of data must be encoded and loaded into the quantum system.

- **Lack of mature algorithms —** New algorithms that can take advantage of capabilities offered by near-term noisy quantum systems will need to be discovered.

User Recommendations

Data and analytics leaders seeking to leverage risk-minimized quantum ML should:

- **Reinvest budget in your classical ML ecosystems**, where the value return will be demonstrably higher than in simulated quantum environments. Explore quantum ML environments at your own risk created by a lack of near-term meaningful results and uncertain timeframes to value.
- **Increase your awareness of quantum computing capabilities** and the potential for applicability in ML use cases by exploring early quantum ML algorithm prototypes on current systems. Leverage free quantum computing access and tutorials to develop quantum computing competency.
- **Prepare for quantum ML** by partnering with quantum computing solution providers and consulting experts to devise new ML algorithm kernels.
- **Leverage quantum-as-a-service** capabilities for validating hypotheses involving quantum ML to minimize risk and maximize the accessibility of quantum computing resources.

Sample Vendors

Amazon Web Services (AWS); Google; IBM; Microsoft; Multiverse Computing; QC Ware; Quantinuum; Quantum Metric; Rigetti Computing; Xanadu

Gartner Recommended Reading

How Use Cases Are Developed and Executed on a Quantum Computer

Innovation Insight for Quantum Computing for the Automotive Industry

Cool Vendors in Quantum Computing

Self-Supervised Learning

Analysis By: Pieter den Hamer, Erick Brethenoux

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Self-supervised learning is an approach to machine learning in which labels or supervisory signals are created from the data itself, without having to rely on historical outcome data or external (human) supervisors that provide labels or feedback. It is inspired by the way humans learn through observation, gradually building up general knowledge about concepts, events and their relations, or spatiotemporal associations in the real world.

Why This Is Important

Self-supervised learning aims to overcome one of the biggest drawbacks of supervised learning: the need to have access to typically large amounts of labeled training data. This is not only a practical problem in many organizations with limited relevant data, or where manual labeling is prohibitively expensive, but also a more fundamental problem with current AI, limiting its learning versatility and broader applicability.

Business Impact

Self-supervised learning enables the extended applicability of machine learning to use cases where labeled training datasets are not available. It may also shorten development time and improve the robustness and accuracy of models. Its relevance is most prominent in computer vision (CV), natural language processing (NLP) — including large language models (LLMs) such as GPT-4, Internet of Things (IoT) analytics/continuous intelligence, robotics, or other AI applications that rely on unstructured data or typically unlabeled sensor data.

Drivers

- **Making ML feasible in the absence of labeled training data:** In self-supervised learning, labels can be generated automatically from the data itself, without the need for human annotation. In essence, this is done by masking elements in the available data (e.g., a part of an image, a sensor reading in a time series, a frame in a video or a word in a sentence) and then training a model to "predict" the missing element. Thus, the model learns how one part relates to another, how one situation (captured through video and/or other sensors) typically precedes or follows another, and which words often go together, for example. In other words, the model increasingly represents the concepts and their spatial, temporal or other relations in a particular domain. This model can then be used as a foundation to further fine-tune the model (e.g., using "transfer learning") for one or more specific tasks with practical relevance.

- **Helping derive more value from the growing availability of IoT sensor data and other diverse, possibly external, sources of data:** Taken alone, these data sources (e.g., visual, sound, pressure, temperature or textual data) may be of limited value. More value can be derived from data by identifying associations between data sources, in essence, using the elements or events in one source to label elements or events in another source.

- **Stepping toward broader AI with more efficient learning:** Self-supervised learning has the potential to bring AI closer to the way humans learn. This occurs mainly via observation and association, building up general knowledge about the world through abstractions and then using this knowledge as a foundation for new learning tasks, thus incrementally building up ever-more knowledge that in future AI scenarios may serve as common sense. For example, ChatGPT and other generative AI rely heavily on this use of self-supervised learning.

Obstacles

- **Skills and experience are still very scarce:** Self-supervised learning is currently only practiced by a limited number of innovative AI companies. This includes its use by large tech firms in the context of foundation models for natural language processing and computer vision.

- **Tool support is still limited:** Although open-source ML frameworks, such as TensorFlow and PyTorch, have started to support self-supervised learning, broader tool support is lacking, which makes implementation a knowledge-intensive and low-level coding exercise.

User Recommendations

- **Apply self-supervised learning only when the value of such application justifies the risks of a still experimental approach.** Scarce, highly experienced ML experts are needed to carefully design a self-supervised learning task, based on masking of available data, which allows a model to build up knowledge and representations that are meaningful to the business problem at hand.
- **Apply self-supervised learning when manual labeling or annotating of data is too expensive or infeasible** — but only after comparing alternative approaches, such as the use of (external) data labeling and annotations services, synthetic data, reinforcement learning, active learning or federated learning.
- **Track the developments in self-supervised learning,** once more mature, self-supervised learning has the potential of becoming a pervasively used foundation for a next generation of applications with AI and machine learning, not limited to foundation models.
- **Complement self-supervised learning with other machine learning approaches.** Self-supervised learning can be used to create a baseline model, which can then be further improved or fine-tuned by using a (smaller) labeled dataset for supervised learning, or by applying reinforcement learning.

Sample Vendors

Amazon; craftworks; Google; Helm.ai; Microsoft; OpenAI; Speechmatics; V7

Gartner Recommended Reading

Three Steps to Boost Data for AI

Innovation Insight for Artificial Intelligence Foundation Models

Innovation Insight for Generative AI

Quick Answer: What Is GPT-4?

Adaptive ML

Analysis By: Pieter den Hamer, Erick Brethenoux

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Adaptive machine learning (ML) is the capability of frequently retraining ML models online in their runtime environment, rather than only training ML models when offline in their development environment. This capability allows ML applications to adapt more quickly to changing or new real-world circumstances that were not foreseen or available during development.

Why This Is Important

Adaptive ML gets AI much closer to self-learning — or at least to more frequent learning or better contextualization — compared with most current AI applications, which only use static ML models that depend on infrequent redeployment of new model updates to improve themselves. Adaptive ML can respond more quickly and effectively to change, enabling more resilient and, ultimately, autonomous systems that are responsive to the dynamics of gradual change and massive disruptions.

Business Impact

Adaptive ML is most relevant in areas where the context, conditions, actor behavior or actor preferences (or the behavior or preferences) change persistently. It is also used to fine-tune, contextualize or personalize models once in production. Example application areas include customer churning in highly competitive markets, gaming, organized crime fighting and anti-terrorism, fraud detection, cybersecurity, quality monitoring in manufacturing, virtual personal assistants and chatbots, semiautonomous cars, and smart robotics.

Drivers

- The ever-increasing complexity, pace and dynamics in society and business require ML models that frequently adapt to changing circumstances and impactful events. More automation requires systems that process data in real time with very frequently updated models for continuous decision intelligence.
- With adaptive ML, models remain accurate longer and suffer less from model drift. Data science teams can improve their productivity by leveraging adaptive ML to reduce the need for conventional, time-consuming model monitoring, periodic retraining and redeployment — also known as MLOps.
- Adaptive ML can be used to compensate for limited availability of training data or "small data," hindering offline (for example, supervised or reinforcement) learning. Adaptive ML may start out with a minimal viable model pretrained offline and then incrementally improved or fine-tuned during the actual online usage. For this reason, adaptive ML is also known as continual or continuous learning. For example, reinforcement learning may be done in a simulated environment during development and continued in a real environment during production.
- Adaptive ML allows for the personalization or contextualization of ML models, using a more general ML model as a starting point and then adjusting this to their user's preferences or the specifics of their context. This, for example, happens through the user prompting of ChatGPT, where style, format and contents of the response are adjusted to the context set through one or multiple session prompts.
- More conventional, offline ML is becoming more adaptive too, although to a lesser extent. With automation in MLOps, models can be retrained with the latest data and redeployed more frequently. Also, some adaptability can be achieved through automated champion-challenger model rotation.
- Combined with federated or swarm ML, adaptive ML can benefit from model improvements in multiple locations and usage contexts. Together, these approaches can ultimately enable autonomous systems, such as self-driving vehicles or smart robots, which should be able to demonstrate resilience in their ever-changing contexts.
- Adaptive ML will evolve to adaptive AI. Self-configuring/self-learning composite AI or multiagent systems — not limited to ML techniques — will offer a more generic approach to adaptive AI.

Obstacles

- Adaptive ML depends on the availability of (real-time) feedback from users, from operating environment or from closed loop data about the quality of the ML output (for example, prediction errors) while online.
- There may be no time for repeated full retraining of the model but only for incremental retraining while online. This requires online or incremental learning algorithms that must be tuned in terms of weighting new data versus older data, mitigating the risk of so-called "catastrophic forgetting."
- Responsible AI and fairness must be addressed not only during offline development but also during online operations. Bias may creep in with continuous model changes, which can be mitigated with periodic offline checks for bias (and also drift and overfitting) and online fairness feedback.
- Implementation is technically challenging with tool and open-source framework support — varying per learning algorithm — only just emerging.
- Nontechnical challenges include ethical, reliability, liability, safety and security concerns that come with self-learning and autonomous systems.

User Recommendations

- Use adaptive ML not to replace but to complement current ML. Most adaptive ML applications will start out with a model that was first trained offline. Adaptive ML can further improve, maintain, contextualize, personalize or fine-tune the quality of ML models once online.
- Accompany adaptive ML with model monitoring for accuracy, bias and relevancy as well as with proper risk analysis and risk mitigation activities, if only to frequently monitor the quality and reliability of adaptive ML applications. Even with adaptive ML, a periodic offline full retraining of the model may be required, as incremental learning has its limitations.
- Manage the required talent, infrastructure and enabling technology, actively. For example, adaptive ML is likely to be more demanding in terms of compute power in runtime environments and will require the development of knowledge about new (incremental learning) algorithms and tools.

Sample Vendors

Accern; Guavus; IBM; Microsoft; Pandio; SAS; TAZI.AI; Toloka

Gartner Recommended Reading

Top Strategic Technology Trends for 2023: Adaptive AI

Use Gartner's MLOps Framework to Operationalize Machine Learning Projects

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 (or fusion) of different AI techniques to improve the efficiency of learning to broaden the level of knowledge representations. Composite AI broadens AI abstraction mechanisms and, ultimately, provides a platform to solve a wider range of business problems in a more effective manner.

Why This Is Important

Composite AI recognizes that no single AI technique is a silver bullet. Composite AI currently aims to combine "connectionist" AI approaches, like machine learning (ML), with "symbolic" and other AI approaches, like rule-based reasoning, graph analysis or optimization techniques. The goal is to enable AI solutions that require less data and energy to learn, embodying more abstraction mechanisms. Composite AI is at the center of the generative AI and decision intelligence market emergence.

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

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

- **Small datasets, or the limited availability of data, have 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. This can be used as a feature to train a neural network to assess and predict the asset's health, also integrating causal AI capabilities.

- **Proliferation of computer vision and NLP solutions.** 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, representing the image entities and their relationships.

- **Agent-based modeling is the next wave of composite AI.** A composite AI solution can be composed of multiple agents, each representing an actor in the ecosystem. Combining these agents into a "swarm" enables the creation of common situation awareness, more global planning optimization, responsive scheduling and process resilience.

- **The acceleration of generative AI.** The advent of generative AI is accelerating the research and adoption of composite AI models (through artifacts, process and collaboration generations), which are the foundation of decision intelligence (DI) platforms.

Obstacles

- **Lack of awareness and skills in leveraging multiple AI methods.** This could prevent organizations from considering the techniques particularly suited to solving specific problem types.
- **Deploying ModelOps.** The ModelOps domain (i.e., the operationalization of multiple AI models, such as optimization models, rule models and graph models) remains an art much more than a science. A robust ModelOps approach will be necessary to efficiently govern composite AI environments and harmonize it with other disciplines, such as DevOps and DataOps.
- **Trust and risk barriers.** 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, observability, model 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, in cases when enough data is not available or when the pattern cannot be represented through current ML models.
- **Capture domain knowledge and human expertise** to provide context for data-driven insights by applying decision management with business rules and knowledge graphs, in conjunction with ML and/or causal models.
- **Combine the power of ML, image recognition or natural language processing with graph analytics** to add higher-level, symbolic and relational intelligence.
- **Extend the skills of ML experts, or recruit/upskill additional AI experts,** to also cover graph analytics, optimization or other techniques for composite AI. For rules and heuristics, consider knowledge engineering skills, as well as emerging skills such as prompt engineering.
- **Accelerate the development of DI projects** by encouraging experimentation with generative AI, which will in turn accelerate the deployment of composite AI solutions.

Sample Vendors

ACTICO; Aera Technology; FICO; Frontline Systems; IBM; Indico Data; Peak; SAS

Gartner Recommended Reading

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

Top Strategic Technology Trends for 2022: AI Engineering

Innovation Insight for Decision Intelligence

Innovation Insight for Decision Intelligence Platforms

How to Choose Your Best-Fit Decision Management Suite Vendor

Graph Data Science

Analysis By: Afraz Jaffri, Svetlana Sicular

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Graph data science (GDS) is a discipline in which data science techniques are applied to graph data structures to identify behavioral characteristics that can be used to build predictive and prescriptive models. Graph data science and machine learning algorithms compute distances and paths, similarity, and communities; simulate the effects of changes in the graph; and allow predictions to be made for inferring new nodes or edges or classifying whole graph structures.

Why This Is Important

Graphs are being used in a wide variety of scenarios, such as financial crime prevention and recommendation systems, to address business problems that need data models to represent complex interactions. GDS allows insight to be derived from these structures beyond the expression of queries and visualizations to build predictive and prescriptive analytical models. GDS includes a wide variety of graph machine learning models that work directly on graph-structured data for node and link prediction.

Business Impact

Extending data science and machine learning techniques with GDS opens opportunities to tackle complex challenges where network effects and relationships cannot be easily modeled using tabular data, but are often better indicators of predicting an outcome. Such challenges are present in many industries. GDS used in conjunction with other techniques is part of the emerging discipline of decision intelligence and forms part of an organization's advanced analytics ecosystem.

Drivers

- Graph databases and knowledge graphs are increasingly used as a unifying layer for data access and usage for downstream applications.
- Academia and industry research centers have made significant advancement in the creation and implementation of graph data science and machine learning algorithms.
- Cloud computing has reduced the burden on organizations to set up and maintain the infrastructure required to run complex graph workloads and has also made specialist hardware accessible that is suited to graph algorithms.
- GDS libraries, both open-source and commercial, enable access to GDS techniques for data scientists using familiar languages and tooling.
- Low-code applications enable graph algorithms and machine learning to be done by domain experts and citizen data scientists.
- Technology and digital native companies such as Google, Uber, Pinterest and Amazon are pioneering the use of graph machine learning and publicizing the value they bring.
- GDS applied to knowledge graphs can provide a level of explainability to predictive models that existing methods cannot provide.

Obstacles

- Expertise is required to understand the various algorithms in the GDS domain and when to apply the appropriate technique to a business problem.
- Awareness of graphs as a data representation solution is not widespread among business leaders and data scientists, who tend to stick to traditional approaches or off-the-shelf solutions.
- Some machine learning tasks on very large graphs still require a significant amount of compute infrastructure and require the manipulation and preprocessing of the graph into the correct structure.
- Operationalization platforms for models built using GDS algorithms are immature compared to traditional DSML platforms.

User Recommendations

- Identify business problems that have potential to be solved using graphs by engaging with domain experts and assessing the amount of data integration, processing and analytical workloads that can be optimized.
- Dedicate time for data scientists to explore graph frameworks and libraries, and create sandbox environments where ideas can be tested.
- Apply graph features to existing predictive models, where relationships are a key characteristic of the data and measure results.
- Create a team to educate and inform different audiences and stakeholders on the uses of graph machine learning and applicable use cases.

Sample Vendors

Amazon Web Services; ArangoDB; Graphistry; Kumo.ai; Neo4j; Oracle; TigerGraph; Virtualitics

Gartner Recommended Reading

3 Ways to Enhance AI With Graph Analytics and Machine Learning

How Graph Techniques Deliver Business Value

Reinforcement Learning

Analysis By: Peter Krensky, Shubhangi Vashisth

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Reinforcement learning (RL) is a type of machine learning (ML) where the learning system receives training only in terms of positive feedback (rewards) and negative feedback (punishments). During problem solving, the system fosters actions or situations so that the overall reward is maximized while minimizing punishments.

Why This Is Important

Some problems can best be solved with RL, especially when other ML approaches are not feasible due to a lack of labeled training data.

Business Impact

The primary potential of RL is in industrial control and design, marketing and advertising, recommendation systems, and gaming industries. The technology can lead to significant improvements in self-driving cars, robotics, vehicle routing, warehouse optimization, logistics, predictive maintenance and other industrial control scenarios.

Drivers

- Recent successes across various industries (For example, text summarization and machine translation, real-time bidding for marketing and advertising, creation of dynamic treatment regimes in healthcare, optimized design of chip layouts in manufacturing, and optimization of robotic players in gaming.)
- Commercial vendors launching new RL products and products with embedded RL
- Sustained data scientist interest in the RL framework because it involves much less training data and supervision than currently dominant supervised learning schemes
- Faster compute capabilities are enabling more application scenarios for RL
- Better simulation capability is also an enabler of RL scenarios
- Reinforcement Learning from Human Feedback (RLHF) in which feedback from an AI community or user group is used to train better models
- Increased attention, interest and potential recognition due to generative AI hype

Obstacles

- Limited RL capabilities offered by current data science and machine learning (DSML) platforms
- Often exceedingly high computational requirements
- Lack of good-enough simulations in many business situations
- Difficulty in designing the reward structure of the RL model for most business scenarios
- Often brittle or difficult-to-implement solutions with applicability in limited use cases
- Lack of staff with reinforcement learning experience
- Lack of explainability

User Recommendations

- Apply RL in use cases requiring frequent model retraining with traditional techniques, because RL can adapt to new environment and circumstances
- Apply RL when the business outcomes and constraints are clear but you lack sufficient labeled data to build robust ML models.
- Acquire special expertise or engage a service provider with risk management support. The application of RL is currently riskier than most traditional techniques.
- Leverage off-the-shelf capabilities available from major vendors in the market, and seek out embedded reinforcement learning.

Sample Vendors

AgileSoDA; Amazon Web Services (AWS); Dataiku; MathWorks; Microsoft; Pathmind; RISELab; TensorFlow

Gartner Recommended Reading

Innovation Insight for Generative AI

Innovation Insight: AI Simulation

Go Beyond Machine Learning and Leverage Other AI Approaches

Transfer Learning

Analysis By: Ben Yan, Shubhangi Vashisth, Radu Miclaus, Wilco van Ginkel

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Transfer learning reuses previously trained machine learning models as an advanced starting point for new purposes, in order to reduce the learning time and data required to attain acceptable performance.

Why This Is Important

Transfer learning is attractive. It enables rapid training, reduces the amount of data needed and may provide better predictive performance than models trained from scratch. A starting point is a repository of models, and these models can be customized based on internal or external data. Transfer learning advances the broader field of AI, as it allows AI to generalize — i.e., use what is learned in one task to more quickly learn another, related task. Transfer learning can also be used to further refine existing models with smaller amounts of data.

Business Impact

The business impact of transfer learning can be summarized as follows:

- It will impact how organizations apply machine learning (ML), especially for natural language processing (NLP) scenarios.
- It promises to attain acceptable performance with less training data, significantly less computational overhead (green IT) and faster development speed.
- It utilizes the model from the source (data-rich) domain, opening ML use cases that were previously infeasible due to lack of data.

Drivers

- **Broader application in foundation models:** The popularity of ChatGPT and image generation models is driving attention to the fine-tuning (a form of transfer learning) of foundation models. Compared with prompt engineering, fine-tuning techniques can incorporate more data into the models, and build customized models for organizations.
- **Increase in model marketplaces and community:** The availability of repositories, such as Hugging Face, allows developers to find and reuse pretrained models with easy community collaboration. The open-source repositories also enable organizations to build models with fewer barriers to entry.
- **Applicability to multiple use cases and verticals:** We see transfer learning being used across a number of model types (language, computer vision, predictive and multimodal) in many domains, such as finance, healthcare, gaming, autonomous driving and e-commerce.
- **Proliferation of AI models within organizations:** Many AI models can be reused. As more models are created, opportunities for transfer learning increase. These models and their datasets can be reused between departments, or even in external organizations.

Obstacles

- The adoption of transfer learning highly depends on the availability of existing models and the relevance/similarity between domains. It is hard to determine upfront whether transfer learning works.
- Transfer learning today is a capability embedded into existing platforms or a method applied by systems integrators and analytics consultancies.
- Transfer learning remains a technical challenge. Fine-tuning foundation models is even harder, and requires proper data, computing resources and talents. The ROI of fine-tuning customized models needs to be measured case by case.
- As the AI field becomes more regulated, documentation of source data and model lineage may be required to support explainability and trust. Not all model providers could provide sufficient information.
- In the quest for more explainable, fair and transparent AI, transfer learning can be seen as a further complication to the AI development process.

User Recommendations

- **Maintain repositories of AI models and datasets:** Work with your data and analytics leaders to utilize metadata management initiatives to identify AI models and their datasets. Document the successful transfer learning examples. AI centers of excellence (COEs) or similar should facilitate.
- **Explore useful internal and external models:** Seek transfer learning opportunities to reuse AI models. Organizations with a more mature level of AI adoption should additionally assess how their current models might be reused in related domains and/or similar tasks.
- **Check the ML tools you use to create and train models, and determine their support for transfer learning:** ML tools should include capabilities that facilitate transfer learning, such as fine tuning.
- **Loop in CSOs, legal teams and business stakeholders:** Teach them about transfer learning to develop your initial position on AI risk. For example, educate them on the lineage of the original data for the base model and the security risks of open base models you may use.

Sample Vendors

4Paradigm; Alibaba Group; Amazon; Google; H2O.ai; Hugging Face; IBM; Microsoft; NVIDIA; OpenAI

Gartner Recommended Reading

Innovation Insight: Transfer Learning

Transfer Learning in China: Increase the Value of Your Data Every Time You Use It

Three Steps to Boost Data for AI

AI Design Patterns with ChatGPT

Composable D&A

Analysis By: Peter Krensky, Erick Brethenoux, Julian Sun, Carlie Idoine

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Composable data and analytics (D&A) utilizes container or microservices architectures and data fabric to assemble flexible, modular and consumer-friendly D&A capabilities from existing assets. This transforms monolithic data management and analytics applications into assemblies of D&A building blocks. This is achieved via composition technologies enabled by low- and no-code capabilities, supporting adaptive and intelligent decision making.

Why This Is Important

Organizations are looking for flexibility in assembly/reassembly of D&A capabilities, enabling them to blend more insights into actions. Time to insight, reuse and agility are top requirements. Modular D&A capabilities enable faster and more proactive insight delivery.

Business Impact

The transition from monolithic D&A applications to composable D&A capabilities can be used with application development to assemble AI-augmented decision-making solutions. The focus of collaboration will transition from technology integrations to business problem solving. Organizations can create advanced analytics capabilities by composing the best capabilities from different vendors, rather than using them separately. Composability also relates to data fabric and data mesh in terms of being able to correctly identify data objects that exhibit wide reuse and separating them from those that are business-process-unique.

Drivers

- Container- or microservices-based analytics and business intelligence (ABI) and data science and machine learning (DSML) platforms with improved APIs enable the assembly of analytics applications in a more flexible way than custom code-based solutions.
- For most organizations, AI is still at the piloting stage, but ABI has been in production for years. Organizations can use composition to connect ABI to AI, extending ABI capabilities and empowering users with a comprehensive, tailored and even personalized solution without having to use different applications.
- Organizations need to assemble descriptive, diagnostic, predictive and prescriptive analytics capabilities dynamically to generate insights along with the decision-making process. Use analytics to inform decision making and drive effective actions in a more connected, continuous and contextual way.
- Both D&A and software development teams will need composable data and analytics to enable emerging business technologies.
- As more data and analytics are integrated into digital platforms, traditional embedded analytics will need more modular capabilities to be assembled and reassembled for faster delivery.
- Embedded analytics are usually implemented by IT, but business users can use low- or no-code capabilities to source more data and compose more capabilities, such as interactive data visualization and predictive modeling, independently enriching more comprehensive embedded analytics.
- Cloud-based marketplaces are becoming an effective channel for organizations to distribute and share analytics applications, and composable D&A enables them to easily find the required components and add value to their applications by infusing analytics.

Obstacles

- New technologies and data have been the key drivers to evolve an analytics platform, resulting in less of a connection with business outcomes. Making data more accessible and composable often raises quality, governance and security concerns, among others.
- Software application development teams and data and analytics teams have not collaborated closely before. Composable D&A requires more involvement from the application development side, including applying XOps practices to maximize its value.
- Today's ABI and DSML markets are not zero-sum games. Many vendors of all sizes and specialties can thrive. No single vendor or tool offers all functions at the same level. It is unrealistic to implement a full D&A stack all at once, so many companies do so in stages. The composability of the existing products is not mature enough without technology partnership.

User Recommendations

- Improve decision making and business impact of data and analytics by incorporating and assembling modular, reusable D&A capabilities.
- Leverage composable analytics to drive innovation by incorporating advanced DSML capabilities into analytics applications.
- Exploit opportunities to add analytics capabilities to applications by building a joint team of application developers and business analysts with ongoing collaboration. Rethink organization, processes and skills to support agile assembly and reassembly of analytics services.
- Pilot composable analytics in the cloud, establishing an analytics marketplace to drive and support collaboration and sharing.

Gartner Recommended Reading

Use Gartner's Reference Model to Deliver Intelligent Composable Business Applications

Adopt Cloud Analytics to Drive Innovation

3 Steps to Build and Optimize a Portfolio of Analytics, Data Science and Machine Learning Tools

Federated Machine Learning

Analysis By: Ben Yan, Svetlana Sicular, Pieter den Hamer, Mike Fang

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Federated machine learning aims at training a machine learning (ML) algorithm on multiple local datasets contained in local nodes without the explicit sharing of data samples. Federated ML helps to protect privacy, enables ML and specifically deep neural networks (DNNs) to use more data, resolves data transfer bottlenecks, and empowers collaborative learning for better accuracy.

Why This Is Important

Federated ML (FedML) highlights an important innovation in (re)training ML algorithms in a decentralized environment without disclosing sensitive business information. FedML enables more personalized experiences with local learning in smartphones, softbots, autonomous vehicles or IoT edge devices, and also facilitates organizations to build collaborative learning models across data silos.

Business Impact

FedML enables collaborative ML by sharing local model improvements at a central level, while keeping the data locally. It especially benefits the Internet of Things (IoT), cybersecurity, privacy, data monetization and data sharing in regulated industries. For example, the U.S. Department of Health and Human Services recently reported an average improvement of 16% and a 38% increase in generalization over local models, as a collaboration result of 20 institutes.

Drivers

- The proliferation of privacy regulations requiring protection of local data.
- With the increasing hype around edge AI, the data becomes distributed across multiple, heterogeneous edge devices and clouds. FedML allows organizations to keep the data in place.
- Data volumes are still growing rapidly, making it more challenging to collect and store big data centrally. This is especially pronounced in the IoT scenarios, where sensor data is collected on the devices and often there is no time or reason to pass it centrally.
- Due to scalability issues, excessive power consumption, connectivity and latency, we see a move toward edge infrastructure in the form of FedML.
- Organizations need collaboration with upstream and downstream partners to improve the overall operation efficiency.
- As large language model (LLM) evolves, research on federated LLM emerges so that a group of organizations could collaborate to train LLM together.
- Swarm (federated) learning is emerging as a promising approach in decentralized ML, uniting edge computing, peer-to-peer networking and coordination, enabled by blockchain.
- FedML is often combined with other privacy enhancing computation techniques as complete secured computing solutions.

Obstacles

- Building trust between organizations for collaborative learning models takes time.
- The incentive mechanism needs to be defined and agreed with all parties engaged to keep participants motivated and keep the FedML group in the long run.
- System and data heterogeneity requires a lot of coordination and standardization among systems to be fully functional.
- Enabling FedML requires a complete end-to-end infrastructure stack that integrates capabilities across DataOps, ModelOps, deployment and continuous tracking/retraining, necessitating a high degree of implementation maturity.
- Creating a new, more accurate and unbiased central model from local model improvements can be nontrivial, as the diversity or overlap between local learners and their data may be hard to assess and may vary greatly.
- FedML is still not widely known in the enterprise, as it lacks marketing on the vendor and researcher sides.
- Security and privacy validation concerns require additional steps.

User Recommendations

- Apply FedML to create and maintain decentralized smart services or products, while protecting the privacy of users and preventing the need to centrally collect massive amounts of data.
- Explore FedML use cases with upstream and downstream partners and look for opportunities to improve overall operation efficiency.
- Give a head start to decentral ML applications by deploying a common, centrally pretrained model while still providing personalization and contextualization by locally retraining the model based on local data and feedback.
- Enable continuous improvement of decentralized ML applications with collaborative learning by repeatedly collecting local model improvements to create a new, improved central model and then redeploying it for decentral usage and fine-tuning.
- Keep a central reference model to ensure “cognitive cohesion” across distributed models — that is, by avoiding decentralized models that veer off too far from its original purpose.

Gartner

Sample Vendors

Alibaba Group; Devron; Ederlabs; F-Secure; Google; Intel; NVIDIA; Owkin; WeBank

Gartner Recommended Reading

Innovation Insight for Federated Machine Learning

Quick Answer: Why Is Federated Learning Prominent in China?

Explore Secured, Accurate and Green AI With Federated Machine Learning

Generative Analytics Experience

Analysis By: Julian Sun, Edgar Macari, Peter Krensky

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Generative analytics experience is an analytics evolution that uses business terminology expressed as natural language via prompts to augment the entire data and analytics process, connecting qualitative and quantitative analytics. It leverages generative AI to empower business users by linking the interpretation of business problems and D&A problems in both directions, autogenerating code (Python, R and SQL) that interacts with the data, and narrating the insights with the business context.

Why This Is Important

Enterprises are using generative AI to improve the experience of analytics for business by better understanding the business questions, translating them into analytical questions and curating into a business-friendly data storytelling. Generative AI enhances augmented analytics with autogenerated content including data, text and code. Technology providers are developing new product lines and innovative customer engagement by integrating generative AI into the main augmented analytics offering.

Business Impact

Enterprises can implement generative analytics experience to:

- Enable business executives without an in-depth understanding of D&A context or who have low data literacy to answer critical business questions and improve decision support.
- Evolve the enterprise's analytics capabilities with more advanced analytical functions triggered by autogenerated R and Python code.
- Improve both quality and quantity of analytics content with more business users acting as analytics creators, and with more narrative business context.
- Improve and expand composable analytics by using generative analytics as the new interface to connect to insight engines.
- Activate the metadata usage to autogenerate the business context of data by incorporating generative AI with semantic layer.

Drivers

- Achieving the business outcome of data and analytics requires connection from insights to actions — a closed-loop activity. Generative analytics experience expands the connections of D&A solutions to a broader generative AI business application with natural language.

- Enterprises are accelerating the adoption of analytics to support more complex decision makings that used to have to use code-oriented data science and machine learning (DSML) solutions. Generative analytics experience can work as a unified experience layer to compose both analytics and business intelligence (ABI) and DSML solutions with better explainability compared to many augmented analytics solutions today, as the code (Python, SQL and R) is clearly generated during the process.

- The market lacks talent. Few people have both the business domain knowledge and advanced analytics skills. Generative analytics experience can fill the gap by enabling more business users with domain knowledge to ask complex business questions.

- ABI and DSML solution providers, especially the search-first vendors, have good technological foundations to integrate with generative AI technologies, which will bring immediate value to the clients if integrated properly. These include semantic layer, MLOps, knowledge graph, natural language query and catalog technologies.

- Digital workplace applications such as Slack and Teams have already integrated with ABI and DSML solutions. The use of natural language from digital workplace applications to perform analytics will form the generative analytics experience as both sides incorporate large language models.

- Enterprises that adopt a data-centric AI approach will proceed with generative analytics experience as one use case among many to achieve the outcome of new technology innovation.

Obstacles

- Natural language query is not a capability in high demand, according to Gartner client interactions over the past year. Enterprises still consider it a nice-to-have feature, and incorporating generative AI would not improve the adoption in the short term.
- Generative AI, and all the ChatGPT hype, is still new to the market. Vendors are still innovating new product lines with immature capabilities. It remains a challenge to incorporate large language models with the right governance to seamlessly integrate with the existing analytics capabilities, especially considering that the accuracy of generative AI is based on activating D&A metadata across multiple vendors.
- Use of generative analytics experience will bring extra cost as most vendors are incorporating GPT-3 or GPT-4 APIs. The usage from broader business users will drive more cost concerns.

User Recommendations

- Target the automation of certain closed-loop business outcomes of D&A by piloting use cases that leverage generative analytics experience to connect insights to action by natural language.
- Start with generative analytics experience in digital workplace applications, mobile BI and natural language query capabilities by evaluating and monitoring existing vendors’ roadmap items.
- Establish the governance of generative analytics experience to minimize errors and “hallucinations” by assessing vendor’s accuracy and veracity of their outputs and its feedback loop to correct and monitor the errors.

Sample Vendors

Aible; AnswerRocket; Hex Technologies; Microsoft; Pyramid Analytics; Tellius

Gartner Recommended Reading

Emerging Tech: Generative AI Needs Focus on Accuracy and Veracity to Ensure Widespread B2B Adoption

Quick Answer: What Are the Short-Term and Midterm Implications of ChatGPT for Data and Analytics?

At the Peak

Logical Feature Store

Analysis By: Georgia O'Callaghan

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Logical feature stores are solutions to address the need for feature reusability, reproducibility and reliability in machine learning (ML) portfolios. Code used to create features for ML can be stored, versioned and accompanied by additional metadata. Most feature stores include a repository of these features, from which users can pull training and testing datasets, accelerating ML model development. Some can also orchestrate feature transformations and monitor data serving models in production.

Why This Is Important

Logical feature stores aim to break down analytical silos, promote collaboration, reduce time spent on feature engineering and enable the reusability of features. A common understanding of how data was transformed into features creates consistency, trust and explainability. Feature stores increase the reliability of data serving models in production by providing a consistent view of features across training and inference, reducing training-serving skew.

Business Impact

Logical feature stores scale to support organizations as they expand ML portfolios and develop a more complex array of ML use cases. They accelerate ML model development by providing datasets for training and testing. They accelerate time to production and ensure data reusability and reliability by providing a consistent view of features across development and production environments. They support model audit and retraining because datasets can be recreated with point-in-time correctness.

Drivers

- Organizations want to expand their use of ML but find that AI teams spend more time sourcing and preparing data to create training and test datasets than developing the ML models themselves.

- Despite the time and effort involved in selecting and engineering ML features, many organizations lack an effective feature management system. This leads to inefficiency, duplication of work and inconsistency between feature definitions.
- Considerable overlap can occur in features used by ML models and, therefore, the ability to reuse features across ML models would lead to faster development times. However, feature engineering within data science teams is typically a siloed practice. Organizations need a mechanism to break down these analytical silos to enable the reusability of features across ML workloads.
- Organizations are increasingly concerned with reproducibility, explainability, audit and governance of ML. Siloed feature engineering efforts make it difficult to capture the lineage of features and other metadata that would enable these efforts.
- Organizations want to leverage ML models in production, yet the process of recreating and implementing data pipelines in a production environment is prone to issues like training-serving skew.
- Issues like data drift, outliers, poor data quality and missing data impact model performance. Organizations understand the importance of monitoring ML model performance, however many lack the means to monitor data serving models in production.

Obstacles

- Confusion still exists about the definition of feature stores and what capabilities they should have. This is further complicated by overlap with the capabilities of data management, DSML and MLOps tooling.
- The logical feature store is the practice of feature management and the capabilities that support this objective. The scale of ML operations within an organization determines how comprehensive its feature management solution must be. Organizations need help deciding whether to build their own logical feature store or use a vendor solution.
- Within the feature store product landscape, some vendors focus on solution completeness and others on optimizing specific capabilities. Many MLOps and DSML platforms have added feature stores or other feature management capabilities into their solutions.

- Feature store solutions only support structured data. Organizations with unstructured data face additional data management challenges that cannot be solved by feature store products.

User Recommendations

- Increase the reproducibility of ML features by capturing version-controlled transformation logic.
- Deliver a consistent view of data across ML development and production environments by storing curated features in a repository that meets requirements for storage scalability and low-latency retrieval of features.
- Increase the efficiency of ML model development by providing a searchable catalog of features within the repository, enabling their reuse across ML workloads. Include a means to pull datasets for training and testing models.
- Promote the understandability of features by establishing their upstream and downstream lineage, including what features are serving what models in production.
- Facilitate model audit and retraining by providing a means to recreate datasets used during model development.
- Prevent ML model degradation by monitoring data and automating the detection of problems known to impact performance, such as data drift, outliers, poor data quality and missing data.

Sample Vendors

Amazon Web Services (AWS); Databricks; Dataiku; Feast; Featureform; Google; H2O.ai; Logical Clocks; Scribble Data; Tecton

Gartner Recommended Reading

Feature Stores for Machine Learning (Part 1): The Promise of Feature Stores

Feature Stores for Machine Learning (Part 2): Current State and Future Directions

Multistructured Analytics

Analysis By: David Pidsley, Stephen Emmott, Tim Nelms, Anthony Mullen

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Multistructured analytics describes techniques applied to unlocking the value of wide data — the spectrum of multistructured information (structured, semi-structured, and unstructured) of any formats, including language (text/audio) and vision (image/video), sourced internally or externally. Multistructured analytics decomposes meaningful features of human-generated data for DSML modeling. It composes machine-generated data into context-enriched analysis for expert decision making.

Why This Is Important

D&A leaders must bridge the gap between the analytics of today and the context-enriched analysis for decision makers to uncover unique insights. Exponential growth in the spectrum of multistructured information sources/formats requires techniques beyond traditional, structured, transactional or relational data. To get a grip on business complexity, harness multistructured analytics to support expert decision making with richer situational awareness, augment business workflows and automate decisions.

Business Impact

Across industries/business functions, organizations accelerate their application of multistructured analytics to wide data sources/formats to reduce costs, address new uncertainties, drive growth and enable innovation in

the use cases for analytics. Multistructured analytics of audio/video streams in sales and marketing can identify behaviors and sentiments of customers and influencers across channels for new insight, experience optimization, real-time dynamic pricing and competitive intelligence.

Drivers

- Decisions are more complex, with more stakeholders and choices than two years ago. Scenarios need context-sensitive evaluation, beyond individual events, using multidimensional models of real-world uncertainties. Lacking the right variety of data stifles this.

- By 2025, 70% of organizations will shift their focus from big to wide data, providing more context for analytics. Internal (digital workplace) and external (business ecosystem) data sources continue to increase silos, leaving hidden intelligence for competitive advantage. Data marketplaces and exchanges make larger, pretrained and more diverse data assets widely available.

- Organizations are adopting multistructured analytics to move beyond storing content, to extract meaningful features and insights. By 2023, over 80% of organizations will use some form of computer vision to analyze images and videos.

- All forms of wide data can now be processed. Document topics can be tagged, speech transcribed, imaged environments annotated, emotions predicted from video, gauges digitized, opening new doors for analytics, data science and machine learning (DSML) and AI.

- Transformer models (via BERT and GPT techniques), advanced text analytics and deep learning have been a catalyst for linguistic and visual analysis. By 2025, AI for video, audio, vibration, text, emotion and other content analytics will trigger major innovations and transformations in most global enterprises.

- Improved price/performance ratio of cloud AI developer services has made experimenting accessible and scalable.

- Multistructured analytics enriches structured data with categorization and tagging. Analytics and BI and DSML platforms are adding multistructured analytics and graph capabilities so that by 2025, context-driven analytics and AI models will replace 60% of existing models built from traditional data sources, consolidating a mix of analytics solutions.

Obstacles

- Although different forms of (text) content analytics have been deployed for years, many organizations avoid leveraging multistructured information due to limited competencies, specialized tools and their perceived difficulty fueled by confusion around terminology.
- While the tools now exist to deploy multistructured analytics uniting the spectrum of multistructured information to model complex context, it's a shift in the way D&A teams undertake data modeling especially in natural language processing.
- The difficulty of combining techniques (composite AI) to handle specific formats/sources — like deep learning for videos, symbolic algorithms for text analytics, and knowledge graphs — is a challenge.
- Data sourcing, quality and privacy are common challenges that can be cost prohibitive for large datasets. Finding suitable data for a specific use case can be difficult and require governance.
- The market for multistructured analytics tools is fragmented and will likely require leveraging multiple vendors, increasing costs.

User Recommendations

- Leverage multistructured analytics for richer situation awareness and expert decision support.
- Conduct proof of value/pilots and understand the data, technical and organizational gaps.
- Apply text analytics for supply chain optimization, image analytics for diagnostic maintenance, video analytics for conferences and audio analytics for fraud prevention.
- Provide context-enriched analysis for decision makers by applying multistructured analytics to multistructured information.
- Explore multistructured analytics capabilities and roadmaps of vendors, including insight engines for text content and cloud AI developer services for image, video and audio analytics.
- Engage startups and hyperscale cloud providers for innovation.
- Estimate your compute and storage needs to train/run effective ML models that leverage multistructured information.
- Invest in taxonomy/ontology skills to accelerate the refinement and automation of information tagging/classification.
- Revise data collection, management and integration practices to take advantage of multistructured analytics.

Sample Vendors

Amazon Web Services; Databricks; Elastic; Google; IBM; Microsoft; OpenAI

Gartner Recommended Reading

Use Multistructured Analytics for Complex Business Decisions

Quick Answer: What Are the Short-Term and Midterm Implications of ChatGPT for Data and Analytics?

Buyer’s Selection Spotlight: Insight Engines

Magic Quadrant for Cloud AI Developer Services

Prompt Engineering

Analysis By: Frances Karamouzis, Afraz Jaffri, Jim Hare, Arun Chandrasekaran, Van Baker

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Prompt engineering is the discipline of providing inputs, in the form of text or images, to generative AI models to specify and confine the set of responses the model can produce. The inputs prompt a set that produces a desired outcome without updating the actual weights of the model (as done with fine-tuning). Prompt engineering is also referred to as "in-context learning," where examples are provided to further guide the model.

Why This Is Important

Prompt engineering is the linchpin to business alignment for desired outcomes. Prompt engineering is important because large language models (LLMs) and generative AI models in general are extremely sensitive to nuances and small variations in input. A slight tweak can change an incorrect answer to one that is usable as an output. Each model has its own sensitivity level, and the discipline of prompt engineering is to uncover the sensitivity through iterative testing and evaluation.

Business Impact

Prompt engineering has the following business impacts:

- **Performance:** It helps improve model performance and reduce hallucinations.
- **Business alignment**: It allows subject data scientists, subject matter experts and software engineers to steer foundation models, which are general-purpose in nature, to align to the business, domain and industry.
- **Efficiency and effectiveness:** Alternative options, such as building a model from scratch or fine-tuning, can be much more complex, drive longer time to market and be more expensive.

Drivers

- **Balance and efficiency:** The fundamental driver for prompt engineering is it allows organizations to strike a balance between consuming an “as is” offering versus pursuing a more expensive and time-consuming approach of fine-tuning. Generative AI models, and in particular LLMs, are pretrained, so the data that enterprises want to use with these models cannot be added to the training set. Instead, prompts can be used to feed content to the model with an instruction to carry out a function.

- **Process or task-specific customizations or new use cases:** The insertion of context and patterns that a model uses to influence the output generated allows for customizations for a particular enterprise or domain, or regulatory items. Prompts are created to help improve the quality for different use cases — such as domain-specific question answering, summarization, categorization, and so on — with or without the need for fine-tuning a model, which can be expensive or impractical. This would also apply to creating and designing new use cases that utilize the model’s capability for image and text generation.

- **Validation and verification:** It is important to test, understand and document the limits and weaknesses of the models to ensure a reduced risk of hallucination and unwanted outputs.

Obstacles

- **Embryonic nature of the discipline:** Prompt engineering processes and roles are either unknown or enterprises have a low level of understanding and experience. Gartner webinar polling data (over 2,500 responses; see Executive Pulse: AI Investment Gets a Boost From ChatGPT Hype) revealed that approximately 60% of respondents self-reported that they had not heard of prompt engineering. And 90% of those same respondents revealed that their organization did not currently have prompt engineers.

- **Role alignment:** Data scientists are critical to understanding the capabilities and limits of models, and to determine whether to pursue a purely prompt-based or fine-tuning-based approach (or combination of approaches) for customization. The ultimate goal is to use machine learning itself to generate the best prompts and achieve automated prompt optimization. This is in contrast to an end user of an LLM who concentrates on prompt design to manually alter prompts to give better responses.

- **Lack of business alignment:** There is often a lack of consensus on prompt engineering's business approach, as well as agreed-upon standards, methodology and approaches. This has led to fierce debates on the value of prompt engineering and how to establish governance.

- **Risk:** Beyond the early stages of awareness and understanding, the biggest obstacle may be that prompt engineering is focused on verification, validation, improvement and refinement; however, it's not without risk. Prompt engineering is not the panacea to all of the challenges. It helps to manage risk, not remove it completely. Errors may still occur, and potential liability is at stake.

User Recommendations

- Rapidly build awareness and understanding of prompt engineering in order to quickly start the journey of shape-shifting the appropriate prompt engineering discipline and teams.
- Build critical skills across a number of different team members that will synergistically contribute critical elements. For example, there are important roles for data scientists, business users, domain experts, software engineers and citizen developers.
- Communicate and cascade the message that prompt engineering is not foolproof. Rigor and diligence need to permeate and work across all the enterprise teams to ensure successful solutions.

Sample Vendors

FlowGPT; HoneyHive; LangChain; PromptBase; Prompt Flow; PromptLayer

Gartner Recommended Reading

Quick Answer: How Will Prompt Engineering Impact the Work of Data Scientists?

Quick Answer: What Impact Will Generative AI Have on Search?

Accelerate Adoption of Generative AI by Offering an FMOps- or a Domain-Specific Partner Ecosystem

Glossary of Terms for Generative AI and Large Language Models

Differential Privacy

Analysis By: Bart Willemsen

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Differential privacy is an approach to using or sharing data while withholding or distorting certain elements about individual records in the dataset. It uses exact mathematical algorithms that randomly insert noise into the data, and add parameters for distinguishability, closeness and diversity of outcomes at each query. When applied correctly, this prevents the disclosure of identifiable information while ensuring that the resulting analysis does not significantly change informationwise.

Why This Is Important

Concerns continue to exist about privacy and the use of personal data in algorithms to serve content or personalize recommendations. As regulatory measures are employed to prevent unauthorized use of personal data, businesses are looking for ways to protect personally identifiable information while still using the data. One technology that can be deployed to accomplish this is differential privacy.

Business Impact

Business data holds value and much of it is personal data. Regulations that constrain the use of personal data are increasing, and the liability for misusing personal data can be substantial. Businesses need to ensure their reputation reflects a company that protects customer data. There are many techniques to address problems in preserving privacy when training AI models. Differential privacy ensures the privacy of individual rows of data while supporting meaningful analysis of aggregate data.

Drivers

- Differential privacy helps to not only reduce risk but also unlock data for AI that was previously too difficult to access.
- Businesses need to uncover value from data without crossing the boundaries of ethical or regulatory restrictions on the use of personal data.
- It is increasingly likely that more restrictive regulations will be enacted, including on the use of personal data in training of algorithms and on how algorithms handle personal data in turn.
- The risk from sophisticated, state-sponsored bad actors that target theft of personal information to facilitate fraudulent actions, remains on the rise.
- Business reputations and trust can be significantly damaged by information breach or misuse.
- Exposure is not limited to datasets in control of the business, as malicious actors can increasingly combine data sources to reidentify individuals even if the data used by the business is anonymized.
- With differential privacy, source data is not altered because the answer to each query is treated "on the fly," protecting the data in use while retaining source data integrity.
- With differential privacy, information value is maintained in a controllable manner via a privacy budget, delivering the desired level of anonymity.
- Some providers have started to add collaborative differential privacy capabilities in their offerings for further privacy protections.

Obstacles

- Solutions that reference the use of differential privacy are not always comparable or equally easily implemented.
- Privacy protection solutions use a variety of techniques and they vary in effectiveness. Organizations often lack a framework to consistently determine the appropriate approach based on use-case requirements, technology maturity and fit.
- Most tools cover anonymity in different degrees and focus on the extent to which reidentification can occur. Other deployments add measures to diversity and closeness of outcome, apart from reidentification protection. This can cause confusion in comparison.
- Lack of familiarity with differential privacy — and the skilled staff to effectively deploy and manage it — hinders adoption. This is exacerbated by how jurisdictions define and determine “anonymous” versus “pseudonymous” data differently.
- Lack of transparency around setting of the privacy budget (the extent to which controls are implemented) undermines trust, whereas increased transparency could elevate trust.

User Recommendations

- Explore the use of differential privacy techniques to decrease the likelihood of sensitive data exposure.
- Use a privacy impact or data protection impact assessment to establish whether additional means are necessary and relevant to the use case.
- Compare differential privacy with other privacy-enhancing computation techniques when operating in high-performance environments that require a high level of precision in analytics models.
- Prioritize differential privacy techniques if you’re operating in a highly regulated industry, such as financial services or healthcare.
- Explore differential privacy techniques when using data across regions where privacy regulations may vary, and always be transparent about where you have set the privacy budget.

Sample Vendors

Immuta; LeapYear; LiveRamp; PHEMI Systems; Privitar; Tumult Labs

Gartner Recommended Reading

Three Critical Use Cases for Privacy-Enhancing Computation Techniques

Natural Language Query

Analysis By: David Pidsley, Rita Sallam

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Natural language query (NLQ) allows business users to query information using business terms typed into a search box or chatbot, or via voice. Vendors' techniques differ in analytical complexity of queries, data volumes and types supported. These keyword searches translate terms into natural language questions using natural language processing technologies and LLM like ChatGPT. Some support querying structured data, and others enable semantic search of multistructured information.

Why This Is Important

- Business users need to make faster data-driven decisions and get context-enriched analysis that includes reasoning about location and time-sensitive situations.
- Despite significant advances in the usability of the point-and-click visual-based analytics, business intelligence (ABI) platforms and other knowledge bases, traditional access paradigms are still too hard for most business users.
- Flattening the learning curve for ABI platform users enables adoption by the remaining two-thirds of employees in organizations that do not use them.

Business Impact

NLQ drives adoption by nontechnical users, offering the ability to ask questions to gain insights, overcoming resistance to visual-based self-service analytics interfaces. NLQ is an increasingly important interface for analytic content development and consumption in data-driven decision making accessible to those unfamiliar with SQL. For data pipelines to enable multistructured analytics across a spectrum of structured data and unstructured content, NLQ can unify a multiexperience user interface.

Drivers

- Foundation models like BERT, large language models (LLMs) and ChatGPT see NLQ repositioned at the Peak of Inflated Expectations and a high benefit rating with less than 2 year time to plateau.
- Generative AI hype is accelerating NLQ capabilities with advanced text analytics and deep learning as catalysts of natural language technologies, including natural language generation (NLG) and NLQ. They enable two-way communication between the human questioner and the machine-generated answer based on the data.
- Demand for generative D&A is substantial with the substantial increase in entrants in 2023. Established ABI platform vendors responded to ChatGPT by improving support for and innovations in NLQ, which is a well-established critical capability of the platforms. Adoption continues to grow as NLQ awareness, availability and solution capabilities improve.
- Orchestration of the entire analytics workflow will increasingly become NLQ-driven and used to manage the analytics and application development activities.
- Augmented analytics capabilities make the analytics consumer of tomorrow a power user by today's standards. Most analytics consumers enter the data story workflow when viewing content that has been created from prepared components and existing data visualizations. Their interaction is typically followed by NLQ or conversational analytics.
- NLQ is becoming central to personalized, consumer-oriented user experiences that combine augmented analytics or automated insights into automated data stories, scenario analysis and conversational analytics. Analytics collaboration enables NLQ engines to learn from team-usage preferences.
- Increasingly mobile workforces using handheld devices and voice interfaces need NLQ to interpret geospatial questions and immediately deliver location-based answers and business insights as a best-fit map visualization. Geospatial analytics and algorithm advances enable NLQ to deliver geospatial reasoning of distance, route calculations and analytics about entities near, farther than or within a certain proximity or boundary, based on business-defined regions or geocoded reference data.

Obstacles

- Limitations in real-time type ahead search-bar suggestions can frustrate users, reduce usefulness and hinder adoption. Some users may not understand the implicit structure of underlying data, rendering queries uninterpretable by the NLQ parser.
- Unindexed datasets often hinder bringing search into an ABI platform. The effort/costs to map/model wide data are high, although generative AI is enabling NLQ of unstructured data to expand the scope and enable multistructured analytics.
- A substantial variety exists in the analytical complexity of queries, NLQ reasoning, support for suggestions for the next questions to ask, NLG to explain findings and support for large data volumes, structured and formats.
- Poor support of spoken languages beyond English, limited domain and industry ontologies, difficulty in configuration, and the need to be predefined in advance means optimizing NLQ implementations often requires customizing the platform and curating synonyms.
- Consistency is lacking for where users can ask questions across platforms and where implementations embed NLQ into the decision making or business process.

User Recommendations

- Help users adopt NLQ for decision making and orchestrating workflows.
- Promote NLQ-specific data literacy training for augmented consumers, business analysts and analytics developers.
- Assess the NLQ roadmaps of vendors and augmented analytics startups.
- Prioritize vendors based on how and what a platform learns (from activate metadata for personalization) via a proof of concept with real data and users.
- Evaluate how NLQ fits into analytics solution architectures. Involve IT in evaluation, data preparation and deployment of ABI platforms.
- Support multiple use cases with multiexperience UIs including evaluating enterprise conversational AI platforms.
- Invest in design thinking on dialogue flows and in competencies to connect conversational analytics to the ecosystem of APIs; for example, ABI platforms and insight engines that enable semantic search and analyzing results sets of wide data with multistructured analytics.

Sample Vendors

ConverSight; iGenius; Pyramid Analytics; Qlik; Tellius; ThoughtSpot

Gartner Recommended Reading

Magic Quadrant for Analytics and Business Intelligence Platforms

Is Your Business Intelligence Enabling Intelligent Business?

Quick Answer: What Are the Short-Term and Midterm Implications of ChatGPT for Data and Analytics?

Magic Quadrant for Insight Engines

Magic Quadrant for Enterprise Conversational AI Platforms

Foundation Models

Analysis By: Arun Chandrasekaran

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Foundation models are large-parameter models that are trained on a broad gamut of datasets in a self-supervised manner. They are mostly based on transformer or diffusion deep neural network architectures and will potentially be multimodal in the near future. They are called foundation models because of their critical importance and applicability to a wide variety of downstream use cases. This broad applicability is due to the pretraining and versatility of the models.

Why This Is Important

Foundation models are an important step forward for AI due to their massive pretraining and wide use-case applicability. They can deliver state-of-the-art capabilities with higher efficacy than their predecessors. They've become the go-to architecture for NLP, and have also been applied to computer vision, audio and video processing, software engineering, chemistry, finance, and legal use cases. Primarily text-based, large language models (LLMs) are a popular subset of foundation models. ChatGPT is based on one (GPT-4).

Business Impact

With their potential to enhance applications across a broad range of natural language use cases, foundation models will have a wide impact across vertical industries and business functions. Their impact has accelerated, with a growing ecosystem of startups building enterprise applications on top of them. Foundation models will advance digital transformation within the enterprise by improving workforce productivity, automating and enhancing CX, and enabling rapid, cost-effective creation of new products and services.

Drivers

Foundation models:

- **Require only limited model customization to deliver effective results.** Foundation models can effectively deliver value through prebuilt APIs, prompt engineering or further fine-tuning. While fine-tuning may deliver the best value because of customization possibilities, the other two options are less complex.
- **Deliver superior natural language processing.** The difference between these models and prior neural network solutions is stark. The large pretrained models can produce coherent text, code, images, speech and video at a scale and accuracy not possible before.
- **Enable low-friction experimentation.** The past year has seen an influx of foundation models, along with smaller, pretrained domain-specific models built from them. Most of these are available as cloud APIs or open-source projects, further reducing the time and cost to experiment.
- **Have accelerated AI innovation with massive model sizes.** Examples include OpenAI's GPT-4; Google's AI's PaLM; Google DeepMind's Gopher and Chinchilla; Meta AI's LLaMA; and Alibaba's M6. In addition, companies such as Hugging Face, Stability AI and EleutherAI have open-sourced their models.

Obstacles

Foundation models:

- **Do not deliver perfect results.** Although a significant advance, foundation models still require careful training and guardrails. Because of their training methods and black-box nature, they can deliver unacceptable results or hallucinations. They also can propagate downstream any bias or copyright issues in the datasets.
- **Require appropriate skills and talent.** As with all AI solutions, the end result depends on the skills, knowledge and talent of the trainers, particularly for prompt engineering and fine-tuning.
- **Expand to impractical sizes.** Large models are up to billions or trillions of parameters. They are impractically large to train for most organizations because of the necessary compute resources, which can make them expensive and ecologically unfriendly.
- **Concentrate power.** These models have been mostly built by the largest technology companies with huge R&D investments and significant AI talent, resulting in a concentration of power among a few large, deep-pocketed entities. This situation may create a significant imbalance in the future.

User Recommendations

- **Create a strategy document** that outlines the benefits, risks, opportunities and execution plans for these models in a collaborative effort.
- **Plan to introduce foundation models into existing speech, text or coding programs.** If you have any older language processing systems, moving to a transformer-based model could significantly improve performance. One example might be a text interpretation, where transformers can interpret multiple ideas in a single utterance. This shift in approach can significantly advance language interfaces by reducing the number of interactions.
- **Start with models that have superior ecosystem support,** have adequate enterprise guardrails around security and privacy, and are more widely deployed.
- **Explore new use cases,** such as natural language inference, sentiment analysis or natural-language-based enterprise search, where the models can significantly improve both accuracy and time to market.
- **Designate an incubation team** to monitor industry developments, communicate the art of the possible, experiment with BUs and share valuable lessons learned companywide.

Sample Vendors

Alibaba Group; Amazon; Baidu; Cohere; Google; Hugging Face; IBM; Microsoft; OpenAI; Stability AI

Generative AI

Analysis By: Svetlana Sicular, Brian Burke

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Generative AI technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. Generative AI has profound business impacts, including on content discovery, creation, authenticity and regulations; automation of human work; and customer and employee experiences.

Why This Is Important

Generative AI exploration is accelerating, thanks to the popularity of Stable Diffusion, Midjourney, ChatGPT and large language models. End-user organizations in most industries aggressively experiment with generative AI. Technology vendors form generative AI groups to prioritize delivery of generative-AI-enabled applications and tools. Numerous startups have emerged in 2023 to innovate with generative AI, and we expect this to grow. Some governments are evaluating the impacts of generative AI and preparing to introduce regulations.

Business Impact

Most technology products and services will incorporate generative AI capabilities in the next 12 months, introducing conversational ways of creating and communicating with technologies, leading to their democratization. Generative AI will progress rapidly in industry verticals, scientific discovery and technology commercialization. Sadly, it will also become a security and societal threat when used for nefarious purposes. Responsible AI, trust and security will be necessary for safe exploitation of generative AI.

Drivers

- The hype around generative AI is accelerating. Currently, ChatGPT is the most hyped technology. It relies on generative foundation models, also called “transformers.”
- New foundation models and their new versions, sizes and capabilities are rapidly coming to market. Transformers keep making an impact on language, images, molecular design and computer code generation. They can combine concepts, attributes and styles, creating original images, video and art from a text description or translating audio to different voices and languages.
- Generative adversarial networks, 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 (ML) and natural language processing platforms are adding generative AI capabilities for reusability of generative models, making them accessible to AI teams.
- Industry applications of generative AI are growing. In healthcare, generative AI creates medical images that depict disease development. In consumer goods, it generates catalogs. In e-commerce, it helps customers “try on” makeup and outfits. In manufacturing, quality inspection uses synthetic data. In semiconductors, generative AI accelerates chip design. Life sciences companies apply generative AI to speed up drug development. Generative AI helps innovate product development through digital twins. It helps create new materials targeting specific properties to optimize catalysts, agrochemicals, fragrances and flavors.
- Generative AI reaches creative work in marketing, design, music, architecture and content. Content creation and improvement in text, images, video and sound enable personalized copywriting, noise cancellation and visual effects in videoconferencing.
- Synthetic data draws enterprises’ attention by helping to augment scarce data, mitigate bias or preserve data privacy. It boosts the accuracy of brain tumor surgery.
- Generative AI will disrupt software coding. Combined with development automation techniques, it can automate up to 30% of the programmers’ work.

Obstacles

- Democratization of generative AI uncovers new ethical and societal concerns. Government regulations may hinder generative AI research. Governments are currently soliciting input on AI safety measures.
- Hallucinations, factual errors, bias, a black-box nature and inexperience with a full AI life cycle preclude the use of generative AI for critical use cases.
- Reproducing generative AI results and finding references for information produced by general-purpose LLMs will be challenging in the near term.
- Low awareness of generative AI among security professionals causes incidents that could undermine generative AI adoption.
- Some vendors will use generative AI terminology to sell subpar “generative AI” solutions.
- Generative AI can be used for many nefarious purposes. Full and accurate detection of generated content, such as deepfakes, will remain challenging or impossible.
- The compute resources for training large, general-purpose foundation models are heavy and not affordable to most enterprises.
- Sustainability concerns about high energy consumption for training generative models are rising.

User Recommendations

- Identify initial use cases where you can improve your solutions with generative AI by relying on purchased capabilities or partnering with specialists. Consult vendor roadmaps to avoid developing similar solutions in-house.
- Pilot ML-powered coding assistants, with an eye toward fast rollouts, to maximize developer productivity.
- Use synthetic data to accelerate the development cycle and lessen regulatory concerns.
- Quantify the advantages and limitations of generative AI. Supply generative AI guidelines, as it requires skills, funds and caution. Weigh technical capabilities with ethical factors. Beware of subpar offerings that exploit the current hype.
- Mitigate generative AI risks by working with legal, security and fraud experts. Technical, institutional and political interventions will be necessary to fight AI's adversarial impacts. Start with data security guidelines.
- Optimize the cost and efficiency of AI solutions by employing composite AI approaches to combine generative AI with other AI techniques.

Sample Vendors

Adobe; Amazon; Anthropic; Google; Grammarly; Hugging Face; Huma.AI; Microsoft; OpenAI; Schrödinger

Gartner Recommended Reading

Innovation Insight for Generative AI

Emerging Tech Roundup: ChatGPT Hype Fuels Urgency for Advancing Conversational AI and Generative AI

Emerging Tech: Venture Capital Growth Insights for Generative AI

Emerging Tech: Generative AI Needs Focus on Accuracy and Veracity to Ensure Widespread B2B Adoption

ChatGPT Research Highlights

Augmented Analytics

Analysis By: David Pidsley, Anirudh Ganeshan

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Augmented analytics uses AI to automate analytics workflows in platforms, contextualizing user interfaces with automated insights, generative storytelling explanations and collaborative exploration. Driven by ML and generative AI, it enables natural language queries and personalized analytics catalogs. It democratizes advanced analytics with augmented data ingestion, preparation, analytics content and DSML model development. It also curbs human biases and accelerates insights for diverse users.

Why This Is Important

Many activities associated with data, including preparation, pattern identification, transformation, model development and insight sharing, remain highly manual. This friction limits the user adoption and business impact of analytics. Enhancing these capabilities with generative AI democratizes analytics and reduces barriers to entry by allowing users to perform complex analytics tasks with low/no code.

Business Impact

Augmented analytics is transforming how users interact with analytics content. Features like conversational interfaces are making analytics more accessible, explainable and expedient. Generative AI is changing how people interact with augmented analytics, enabling access to deeper insights from data. Once confined to experts only, insights from advanced analytics are now in the hands of business analysts, decision makers and operational workers across the enterprise. These augmented consumers are driving new sources of business value.

Drivers

- Organizations increasingly want to analyze more complex datasets combining diverse data from both internal and external sources. With an increasing number of variables to explore in such harmonized data, it is practically impossible for users to explore every pattern combination. It is even more difficult for users to determine whether their findings are the most relevant, significant and actionable. Expanding the use of augmented analytics will reduce the time users spend on exploring data, while giving them more time to act on the most relevant insights.
- Generative AI has accelerated market interest in dynamic data stories and other combinations of augmented analytics features that automate insights. Generative AI combines augmented analytics with natural language query, natural language generation, and anomaly detection to dynamically generate data stories for users in their contexts. This type of multiexperience UI will reduce the use of predefined dashboards for monitoring and analysis, and increase the use of augmented analytics.
- Vendor technology innovation is pushing augmented analytics forward. With the explosion of generative AI, augmented analytics is receiving heightened attention. ABI platforms are now integrating large language models like GPT-4, allowing users to generate, debug and convert code, create data stories, and aid in data preparation. This integration has also enabled newer users to emerge, fueling analytics adoption. In a next wave of generative analytics experiences, users may see the entire workflow become AI-driven.
- Most organizations leverage multiple ABI platforms, causing exponential proliferation of analytics content. Coupled with a lack of governance, this proliferation often leads to inconsistencies in metrics and insights, duplication of reports and dashboards, and an overall decline of trust in data. Hence, analytics catalogs, powered by augmented analytics capabilities with generative AI, are becoming key in allowing users to find and recommend analytics content.
- By integrating with digital workplace applications (e.g., Microsoft Teams and Slack), augmented analytics features allow users to share and collaborate on insights.

Obstacles

- **Lack of trust in autogenerated models and insights:** Organizations must ensure that the augmented approach is transparent and auditable for accuracy and bias. They must establish a process to review and certify analyses created. These guardrails are especially important with generative AI being included within ABI platforms.

- **Training and rapidly evolving skills needs:** Obtaining desired skill sets and data literacy standards is a never-ending challenge, and leaders need broad and diverse training for multiple personas.

- **Ecosystem requirements:** It will be critical to build an ecosystem that includes not only tools, but also data assets, people and processes to support the use of augmented analytics.

- **Cultural barriers:** Analytics developers writing analytics-as-code and business analysts accustomed to visual self-service analytics may regard augmented analytics as a “nice to have” feature. However, they neither utilize nor rely on it in their analytics content production workflows.

User Recommendations

- Identify the personas and use cases that will benefit most from augmented analytics capabilities.

- Ensure that users can get value from new augmented analytics features by providing targeted and context-specific training. Invest in data literacy to ensure responsible adoption.

- Focus on explainability as a key feature to build trust in autogenerated models. Create learning opportunities for those who wish to know more about the theory and inner workings of augmented analytics solutions.

- Assess the augmented analytics capabilities and roadmaps of ABI platforms, data science platforms, data preparation platforms and startups as they mature. Look into the upfront setup and data preparation required, the range of data types and algorithms supported, the integration with existing tools, the explainability of the models, and the accuracy of the findings.

- Provide incentives for citizen data scientists to collaborate with, and be coached by, specialist data scientists who still need to validate models, findings and applications.

Sample Vendors

AnswerRocket; iGenius; Microsoft; Oracle; Pyramid Analytics; Qlik; Sisense; Tableau; Tellius; ThoughtSpot

Gartner Recommended Reading

Market Guide for Augmented Analytics

Magic Quadrant for Analytics and Business Intelligence Platforms

Critical Capabilities for Analytics and Business Intelligence Platforms

Is Your Business Intelligence Enabling Intelligent Business?

Top Trends in Data and Analytics, 2023

Synthetic Data

Analysis By: Arun Chandrasekaran, Anthony Mullen, Alys Woodward

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 rather than obtained from direct observations of the real world. Synthetic data is used as a proxy for real data in a wide variety of use cases including data anonymization, AI and machine learning development, data sharing and data monetization.

Why This Is Important

A major problem with AI development today is the burden involved in obtaining real-world data and labeling it. This time-consuming and expensive task can be remedied with synthetic data. Additionally, for specific use-cases like training models for autonomous vehicles, collecting real data for 100% coverage of edge cases is practically impossible. Furthermore, synthetic data can be generated without personally identifiable information (PII) or protected health information (PHI), making it a valuable technology for privacy preservation.

Business Impact

Adoption is increasing across various industries. Gartner predicts a 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.
- Improves ML performance as more training data leads to better outcomes.
- Enables organizations to pursue new use cases for which very little real data is available.
- Is capable of addressing fairness issues more efficiently.

Drivers

- In healthcare and finance, buyer interest is growing as synthetic tabular data can be used to preserve privacy in AI training data.
- To meet increasing demand for synthetic data for natural language automation training, especially for 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 AI simulation techniques is accelerating synthetic data.
- There is an expansion to other data types. While tabular, 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.
- The explosion of innovation in AI foundation models is boosting synthetic data creation. These models are becoming more accessible and more accurate.

Obstacles

- Synthetic data can have bias problems, miss natural anomalies, be complicated to develop, or not contribute any new information to existing, real-world data.
- Data quality is tied to the model that develops the data.
- Synthetic data generation methodologies lack standardization.
- Completeness and realism are highly subjective with synthetic data.
- Buyers are still confused over when and how to use the technology due to lack of skills.
- 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 or teacher for labels.
- If fringe or edge cases are not part of the seed dataset, they will not be synthetized. This means the handling of such borderline cases must be carefully accommodated.
- There may be a level of user skepticism as data may be perceived to be “inferior” or “fake.”

User Recommendations

- Identify areas in your organization where data is missing, incomplete or expensive to obtain, and is thus currently blocking AI initiatives. In regulated industries, such as healthcare 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.
- Educate internal stakeholders through training programs on the benefits and limitations of synthetic data and institute guardrails to mitigate challenges such as user skepticism and inadequate data validation.
- Measure and communicate the business value, success and failure stories of synthetic data initiatives.

Sample Vendors

Anonos (Statice); Datagen; Diveplane; Gretel; Hazy; MOSTLY AI; Neuromation; Rendered.ai; Tonic.ai; YData

Gartner Recommended Reading

Innovation Insight for Synthetic Data

Innovation Insight for Generative AI

Data Science and Machine Learning Trends You Can't Ignore

Cool Vendors in Data-Centric AI

Case Study: Enable Business-Led Innovation with Synthetic Data (Fidelity International)

Sliding into the Trough

Chief Data Scientist

Analysis By: Peter Krensky, Erick Brethenoux, Carlie Idoine, Aura Popa

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Chief data scientist is an evolving leadership role responsible for translating analytics and AI strategy into efficient and effective implementations of advanced D&A products and services. The role is typically the most senior data science position within an organization and has a specific focus on applied data science approaches.

Why This Is Important

The complexity, pervasiveness and criticality of advanced analytics and AI requires dedicated attention. The chief data scientist ensures alignment with critical business priorities. Chief data scientists are also responsible for enterprise data science teams and for execution of the data science vision. They must think tactically, assess the current situation and deliver value today while planning strategically, in parallel, to coordinate and maximize the future use of advanced analytics and AI.

Business Impact

The business value derived from advanced analytics and AI done right is more than what can be accomplished by other means. Chief data scientists propel the use of advanced analytics and AI forward by:

- Transforming the business vision to advanced analytics and AI initiatives.
- Coordinating all advanced analytics and AI initiatives across the organization, reliably delivering advanced analytics and AI solutions and accelerating time to value.
- Delivering measurable results and value to the organization.

Drivers

- Communicating the value of advanced analytics and AI, not just delivering projects, is a key business priority. Efforts are required to manage and coordinate both centralized and decentralized teams to deliver measurable business outcomes.

- Growing data science teams are fueled by generative AI hype, excitement and strategic funding. These teams especially need help with communication, coordination and stakeholder management.

- Organizations are digitizing and automating more of their processes with AI and analytics at the core. As AI becomes a critical function in processes underlying digital businesses, it requires leadership skills and oversight.

- Siloed, unstructured approaches to advanced analytics and AI not only consume significant time and resources but also increase risk and minimize return on investment and overall trust in these techniques. The role of chief data scientist aids breaking down and eliminating silos.

- Value from advanced analytics and AI requires a consistent, managed approach. Chief data scientists are responsible for establishing processes resulting in consistent delivery of high-value advanced analytics solutions.

- The democratization of data science has generated an increasing number of ML models that are often not operationalized. The need for better coordination with the lines of business and IT and the harmonization of DSML practices requires a chief data scientist role.

- A shortage of staff with data science skills requires a concerted effort to recruit, retain, organize and develop data science talent across the organization.

Obstacles

- Recruiting and retaining an experienced chief data scientist, with the right blend of management, technical, business and communication skills, is challenging.

- D&A leaders and their IT and business partners often lack influence, organization, process and practice to deliver, operationalize and scale advanced analytics and AI solutions and approaches.

- The chief data scientist role may not have enough organizational clout and defined authority to drive the enterprisewide changes required to reap the benefits of AI.

- The lack of business recognition often leaves the organization open to data science poaching, weakening the role of the chief data scientist.

- Keeping roles and responsibilities clearly defined is challenging when organizations also have (sometimes multiple) CDOs and CAOs.
- Organizations may think they need a CDS role when it is still too early in their maturity to warrant the position.

User Recommendations

- Define the chief data scientist role as a complement to other CxO roles, recognizing that alignment between these roles is critical.
- Work both within (the internal IT team and the broader organization) as well as outside the organization to orient the chief data scientist within the broader community and identify opportunities for learning and partnership.
- Empower chief data scientists to build a diverse team, develop processes and procure tools to deliver models in a way that builds trust while tracking the impact on key business priorities and value generated.
- Leverage the chief data scientist role to drive and coordinate application, exploration and delivery of advanced analytics and AI methods and techniques to align those methods with real, prioritized business problems.

Gartner Recommended Reading

Lessons From Data Scientists on Their Education and Career Development

How CDAOs Can Lead Upskilling Initiatives in Data Science and Machine Learning

MLOps

Analysis By: Peter Krensky, Pieter den Hamer, Jim Hare, Erick Brethenoux

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Machine learning operationalization (MLOps) aims to streamline the end-to-end development, testing, validation, deployment and instantiation of ML models. It supports the release, activation, monitoring, experimentation and performance tracking, management, reuse, updating, maintenance, version control, risk and compliance management, and governance of ML models.

Why This Is Important

Operationalization of machine learning projects is often an afterthought, which keeps organizations from realizing the true value of their investments. MLOps aims to standardize the development, deployment and management of ML models by supporting the release, activation, monitoring, performance tracking, management, reuse, maintenance, risk and compliance management, and governance of ML artifacts.

Business Impact

MLOps supports the following:

- **Integration:** Integrate advanced analytics and ML platforms to provide a unified ML operationalization pipeline that rapidly reduces time to value.
- **Catalogs:** Store and secure data, analytical artifacts and ML artifacts for ease of collaboration and reusability.
- **Governance:** Ensure auditability, enforce adherence to internal and external security policies and procedures, and address potential privacy issues.
- **Coherence:** Provide functional bridges between the development and operationalization cycles.

Drivers

- Organizations often face machine learning model debt, which keeps them from realizing the true value of their investments. MLOps helps organizations pave a clear path from experimentation to production.
- Organizations want to ensure the integrity (technical and business), transparency and sustainability of deployed ML models by establishing a systematic operationalization process for their machine learning projects — one that differs from the process for traditional software engineering projects.
- Organizations want to maximize their operationalization success by securing the help of domain experts, ML engineers, IT professionals and business practitioners, in addition to existing data science talent. MLOps brings all these personas together by providing a common management and governance framework.
- Organizations seek to simplify the maintenance of deployed machine learning models by monitoring and revalidating their business value on an ongoing basis.
- The number of deployed ML models is increasing rapidly. In the 2021 Gartner AI in Organizations Survey, some respondents said they have hundreds of thousands of models, making MLOps a must-have capability.

Obstacles

- Organizations tend to think of MLOps as a technology or procedure, rather than a collaborative way of working. MLOps brings different personas together to productionize ML workflows.
- MLOps tools and platforms primarily focus on the management, monitoring and governance of machine learning models. In most cases, they do not assist in the end-to-end development, deployment, management and governance of machine learning pipelines.
- Most organizations overlook the critical aspect of operationalization.
- The vendor landscape for MLOps is rapidly evolving, creating confusion for organizations seeking the most efficient way to operationalize their machine learning workflows.
- MLOps focuses on the end-to-end management and governance of ML models, whereas ModelOps focuses on the end-to-end governance of all logic-based models and decision models, including knowledge graphs, heuristics, ML-based models and other AI models. This subtle distinction adds a layer of confusion for organizations making platform decisions.

User Recommendations

- Establish a systematic MLOps process through Gartner's MLOps framework.
- Consider implementing a wider ModelOps strategy if you are a mature organization with a variety of AI models, such as heuristics, agent-based models, knowledge graphics and ML models.
- Make active investments to upskill your workforce for operationalization.
- Ensure the business value of ML deployments while prioritizing use cases by establishing close, ongoing dialogue with, and explicit buy-in from, business counterparts. The earlier the dialogue happens, the more successful the model operationalization will be.
- Distribute delivery, but centralize oversight, by organizing for MLOps. Connect delivery and oversight within the data science lab, IT or the LOBs, depending on the expected business outcomes, size of the ML project team and complexity of the initiatives.
- Define roles and responsibilities of data scientists, IT and MLOps by aligning team members with stages in the ML life cycle.

Sample Vendors

Amazon Web Services; Databricks; Dataiku; DataRobot; Datatron; Google; Microsoft; Valohai

Gartner Recommended Reading

A Mandate for MLOps, ModelOps and DevOps Coordination

Use Gartner's MLOps Framework to Operationalize Machine Learning Projects

Edge Analytics

Analysis By: Peter Krensky

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Analytics is the discipline that applies logic (e.g., “rules”) and mathematics (“algorithms”) to data to provide insights that drive organization strategy and decision making. “Edge” analytics means that the analytics are executed in distributed devices, servers or gateways located outside of data centers and public cloud infrastructure closer to where the data and decisions of interest are created and executed.

Why This Is Important

Gartner client inquiries about the impact of edge on data and analytics continue to increase. With a growing relevance, by 2025, more than 50% of enterprise-managed data will be created and processed outside the data center or cloud. Demand for real-time decision making closer to where the data of interest is created and stored is one of many drivers for edge analytics.

Business Impact

The origins of edge analytics offerings were primarily in the support of decentralized deployments for device-isolated insights. However, connectivity advances, demands for cross-device analytics and innovations surrounding IoT have dramatically increased the scale and complexity of edge analytics use cases. Real-time event analytics and decision making, autonomous behavior of assets, and fault-tolerant applications hold tremendous potential value for enterprises in many industries.

Drivers

- Advantages of edge analytics include faster response times, reduced network bottlenecks, data filtering, reliability, increased access to data and reduced communications costs.
- Data sovereignty and governance issues related to sensitive/regulated data can constrain D&A teams from adopting centralized/cloud-based environments — moving data outside its originating geography can violate sovereignty regulations. By locating analytics in edge environments, the data remains in the originating locations, increasing the likelihood of compliance.
- The increase of distributed cloud and hyperconverged solutions from public cloud providers, including Amazon Web Services (AWS Outposts), Microsoft (Azure Stack Hub) and Google Cloud (Anthos), are further decentralizing previously cloud-restricted workloads. This perimeter expansion of the cloud brings compute and storage closer to the edge — creating new possibilities for edge-centric analytic workloads.
- 5G networks continue to grow in relevancy and, combined with mobile edge computing, will increase edge analytics use cases — particularly for latency-sensitive deployments.
- More analytics solutions, such as those supporting IoT use cases, need to operate in disconnected (or intermittently connected) scenarios. By bringing more powerful analytics capabilities to edge environments, these solutions need not rely on centralized data centers or cloud resources. As demand grows for “smarter” physical assets in many industries, supporting autonomous behavior will be a common requirement.

Obstacles

- Some of the disadvantages of edge analytics include increased complexity, lack of cross-device analytics, overhead of device maintenance and technical currency demands.
- Architectural design and development best practices for traditional or cloud-resident analytics typically assume or prioritize data/analytics centrality and do not carry over directly for edge analytics use cases.
- Vendor choices include two extremes in terms of provider scale — with early and unknown startups competing head-to-head with global megavendors. This drives a mix of platform/protocol standards and complicates going concern considerations for prospective buyers.
- Edge analytics can increase the complexity of enterprise standards and governance (data privacy, security, etc.), which has the potential to delay overall value realization objectives.

User Recommendations

Analytics leaders should consider edge analytics across the following five imperatives:

- Provide analytic insights for individual devices, assets or a larger distributed site even in the midst of disconnection from cloud or data center infrastructure and resources (e.g., driverless cars).
- Provide data sovereignty. Many regulations or data privacy laws require data be kept in the location of origin or the organization deems the transfer of data to introduce too many security vulnerabilities.
- Adapt to scenarios where network connectivity does not have the ability to support desired latency or stability requirements.
- Address scenarios where cross-device interdependencies serving as part of a larger system require edge-resident analytics.
- Redesign analytic strategies where it costs too much to upload the full volume of generated data and where there is no benefit to moving device-level data to a central location for aggregated analysis.

Sample Vendors

Amazon Web Services; Arundo; CloudPlugs; FogHorn; Microsoft; PTC; Samsara; TIBCO Software

Gartner Recommended Reading

Market Guide for Edge Computing

Innovation Insight for Edge AI

The Edge of the Edge Overview

Emerging Technologies Impact Radar: Edge AI

Explainable AI

Analysis By: Peter Krensky, Sumit Agarwal

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Explainable AI (XAI) is a set of capabilities that describes a model, highlights its strengths and weaknesses, predicts its likely behavior, and identifies any potential biases. It can clarify a model's functioning to a specific audience to enable accuracy, fairness, accountability, stability and transparency in algorithmic decision making.

Why This Is Important

XAI gives visibility into how a model arrived at a particular decision. This helps in building trust, confidence and understanding in AI systems. In highly regulated sectors such as insurance or banking, regulations directly or indirectly mandate the need for model explainability to properly manage model risk.

Business Impact

XAI is the responsibility of both vendors (data scientists and solution developers) and also of end-user organizations that consume them. Not supporting this capability puts businesses and decision making at risk. However, different levels of explainability are required for customers, the organization's IT and management, society, and regulators to direct AI governance.

Drivers

- The lack of model transparency or interpretability among model users, managers and consumers impacted by models' decisions severely limits an organization's ability to manage AI risk. Fairly or unfairly, consumers hold the originating organization responsible for the performance and behavior of AI.
- Not ensuring explainability invites model risk that can lead to financial loss, poor business and strategic decision making, or damage to organizational reputation.
- A lot of organizations are shifting to augmented decision-making capabilities with the use of AI models. As a result, they should be able to explain how an AI model arrived at a particular prediction or decision.
- XAI capabilities are prebuilt into both platforms and innovations in the open-source community to explain and interpret models are on the rise.
- Ethical and moral considerations need to be accounted for while relying on augmented decision making, often supported by thorough governance and auditing capabilities for these models.
- New regulations and legal interventions are taking place that mandate the use of explainable AI methodologies.
- Explainable models also help with attrition, so data scientists who quit the job do not leave black boxes behind them. Models that are interpretable help business audiences gain trust in AI.

Obstacles

- Explainability is often confused with ML interpretability. Although the latter serves data scientists, the former applies to different personas interacting with the AI life cycle.
- XAIs are often looked at as a task or a step required while creating AI projects toward the end of the AI life cycle, but they have to be continuous and tested throughout training, development and production phases.
- An inherent lack of trust exists in AI systems that keeps organizations from adoption, since they're simply not aware of XAI techniques or frameworks.
- Explainability tools are fragmented, and XAI is often consumed in an oversimplification such as showing feature importance to end users. Although that approach works in the beginning, XAI is much wider than that, and requires a deep understanding of the subject.
- Organizations that focus on the accuracy of the models rather than on the interpretability stall their decisions on creating a more explainable AI.

User Recommendations

- Define a range of actions that can be taken independently that identify unacceptable results and that flag those results for human intervention. Minimizing the number of incorrect results derived from AI is critical, because users will lose trust in a poorly performing system.
- Educate, train and foster ongoing conversations with key stakeholders, including line-of-business managers, legal and compliance, to understand the AI model's explainability requirements, challenges and opportunities.
- Strive for XAI for each model along the dimensions of business, data, algorithms, models and production.
- Accept deficiencies in explainability as a natural consequence of systems becoming increasingly complex. Document notable deficiencies or potential biases so that they can be used to make corrections in the future.
- Establish the role of AI model validator, a data scientist whose job is to ensure that models are explainable and robust, and meet all possible constraints.

Sample Vendors

Dataiku; EazyML; Fiddler AI; Google; H2O.ai; IBM; Microsoft; Modzy; Superwise; TruEra

Gartner Recommended Reading

Innovation Insight for Bias Detection/Mitigation, Explainable AI and Interpretable AI

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

Market Guide for AI Trust, Risk and Security Management

Incorporate Explainability and Fairness Within the AI Platform

Citizen Data Science

Analysis By: Peter Krensky, Rita Sallam, Carlie Idoine, Shubhangi Vashisth, Frances Karamouzis

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Citizen data science is the collective set of capabilities applied to deliver analytic insights where the personnel are not the experts and their role or job function may not be within the data and analytics (D&A) discipline. Citizen data scientist is a persona rather than a title or role within an organization.

Why This Is Important

- The collective personnel (citizen data scientists) delivering these insights add to the impact of the D&A discipline on the organization through the creation and delivery of insights.
- The functional knowledge of citizen data scientists adds a dimension of efficiency, efficacy and depth to the solutions and experience. Citizen data scientists often unlock new insights beyond the use of basic descriptive and diagnostic capabilities.
- Citizen data scientists serve to reduce the talent gap caused by the shortage and high cost of data scientists.

Business Impact

The business impact can range from a synergistic force multiplier effect to governance challenges. The most powerful and impactful business benefits come when citizen data scientists are actively recruited to fusion teams, are provided tools, and perform specific phases of the analytics life cycle (such as feature generation and selection, and algorithm selection) to best leverage their expertise. Ultimately, this puts the power of the tooling in the hands of those who know best how to apply it and align to making business decisions. The challenges arise when the citizen data scientists are reaching beyond their expertise and the appropriate guardrails are not in place.

Drivers

The most significant drivers of citizen data science include:

- **Talent gap** — The sheer volume of personnel needed continues to outstrip demand. Citizen data scientists help fill a portion of that gap. Historically, building data science and machine learning (DSML) models required expert data scientists, who are difficult and expensive to hire and retain. Citizen data science helps overcome such limitations.

- **Generative AI excitement and possibilities** — The popularity of ChatGPT and the dawning of the generative AI era has had a profound effect on citizen data science. The full user spectrum from experts to beginners is experimenting with novel approaches and techniques for low-code/no-code data science. Data preparation exploration and model development will be dramatically accelerated and democratized, contributing to a rewritten art of the possible for citizen data science.

- **Functional knowledge** — Citizen data scientists' primary knowledge base is an in-depth understanding of the business domain. It is the combination of functional knowledge, data science skills and technology that drive results.

- **Vendor offerings** — Vendors have recognized this additional population as a target-rich environment for their offerings. As such, many vendor offerings now commonly include tools and features designed specifically for usage by citizen data scientists.

- **Augmented analytics capabilities —** These include automated, streamlined data access and data engineering; augmented user insight through automated data visualization and exploration; modeling and pattern detection including feature engineering, model selection and validation; automated deployment and operationalization; and capabilities to support collaboration and sharing.

Obstacles

- Upskilling in advanced DSML techniques and approaches is important to derive value from citizen data science. Classroom learning provides a foundation but must be supported by on-the-job learning and experimentation.
- Tools with augmented analytics capabilities and additional processes to manage creation and sharing of models will be required to support citizen data science.
- There is still a need to (statistically) validate results of citizen data science by expert data scientists.
- Expert data scientists often resist or underestimate the effectiveness of citizen data science approaches.
- Citizen data science is often deemed to be just a preliminary, elementary step and not a fully functional DSML approach.
- Citizen data science leveraged in silos with no oversight or collaboration among experts and others with a vested interest in DSML success could lead to duplication of data engineering and analytic effort, lack of operationalization, and limited visibility and standards.

User Recommendations

- **Success starts with leadership —** Educate business leaders and decision makers about the potential impact of a broader range and larger pool of delivery capability. Work with leadership to scan opportunities for citizen data science to complement existing analytics and expert data science initiatives across the data science life cycle.
- **Inviting and inclusive environment —** Create communities of practice, and provide training and tools to make an inviting and supportive environment for all to explore the value of the citizen data scientist persona. This involves defining the citizen data scientist as a formal persona. Define its "fit" relative to other roles, and identify those who fit the citizen data scientist profile.

- **Expert data scientist value —** Acknowledge that you still need specialist data scientists to validate and operationalize models, findings and applications.

- **Tools and technologies —** Provision augmented analytics tools (including but not limited to augmented data science and machine learning tools), platforms and processes to support and encourage collaboration between business users, application developers and data science teams. Track the capabilities (technology) and roadmaps of existing business intelligence (BI) and data science platforms and emerging startups for support of augmented features.

Sample Vendors

Aible; Alteryx; Dataiku; DataRobot; H2O.ai; Microsoft; Qlik; SAS; Tellius

Gartner Recommended Reading

Build a Comprehensive Ecosystem for Citizen Data Scientists to Drive Impactful Analytics

Maximize the Value of Your Data Science Efforts by Empowering Citizen Data Scientists

Best Practices to Enable Effective Citizen Data Science

Data Science Education

Analysis By: Peter Krensky, Aura Popa

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Higher education institutions and solutions vendors offer learning experiences for credit-bearing courses and independent study of data science capabilities. These include massive open online courses (MOOCs), certifications, diplomas, and undergraduate and postgraduate degrees.

Why This Is Important

As demand for data science talent stays high and data science careers remain attractive, many organizations and forward-thinking professionals are exploring avenues for upskilling in data science and machine learning (DSML). There are a variety of tools designed for users with minimal DSML experience and newly acquired skills. Education is the best answer to further narrow the data science talent gap, and prepare organizations for optimal adoption of next-generation AI and ML technologies.

Business Impact

Data science education helps leaders take individuals with business or technical experience and domain expertise and equip them to become citizen data scientists. Education is becoming vital to preparing the large populations of augmented consumers to better engage with and understand embedded machine learning. Continuing education is also a vital practice to develop and retain expert data scientists.

Drivers

- Data science class sizes, faculties and degree programs have been expanding in traditional universities for the past decade.
- General upskilling and online education in data science are thriving, and online education activity has only intensified over the last year.

Other key drivers of data science education at all levels include:

- Sudden and widespread interest in generative AI
- Abundant and lucrative job opportunities for those with data science skills and ML literacy
- New and expanded university degree programs and curriculums in analytics and data science
- Large populations of knowledge workers interested in upskilling in data science
- Training and certification programs of all kinds from software vendors, some tied to specific technologies
- Low technical barrier to entry in DSML with the advent of low-code/no-code platforms and augmented DSML, making the area well within the reach of interested individuals

- DSML talent development as a retention strategy
- Rapid progress of research and new technologies coming out of academia and corporate research labs
- The low cost of a foundational data science education (under $1,000 for strong independent learners)

Obstacles

- An overwhelming number of online course options designed for different personas and experience levels
- Broad interpretations of the terms “data science” and “data scientists” in terms of training approach, curriculum and desired skills
- Plenty of content failing to justify cost despite there being many popular and excellent options
- Students and organizations recognizing that data science is a moving target, and certain areas of study can quickly fall out of favor or become obsolete
- A sporadic and inconsistent approach to training and certification
- Training people out the door (that is, sponsoring the development of skills that attract recruiters and lead to the departure of valued employees)
- No expert mentorship available for citizen data scientists after completing foundational education
- Experts lacking education and experience in business considerations around data science use cases and applications
- Underestimating DSML complexity and oversimplifying it to foundational elements

User Recommendations

- Build and expand relationships with local universities to establish internship programs and a data science talent pipeline into your organization.
- Sponsor all MOOCs under $1,000 per employee in total cost with the expectation that the majority of classroom work will be done outside of working hours.
- Expect leadership from expert data scientists on key new topics, leading tools and technologies, and how to manage common pitfalls that less experienced practitioners will encounter.
- Embrace the fact that the majority of learning and skill development is done after classroom learning in the course of experimentation and project delivery. Quick wins build momentum, but lasting and meaningful change takes time because it requires people to learn new skills and behave in new ways.

Sample Vendors

The Center of Applied Data Science (CADS); Coursera; DataCamp; DeepLearning.AI; edX; Kaggle; Pluralsight; Skillsoft; Udacity; Udemy

Gartner Recommended Reading

Lessons From Data Scientists on Their Education and Career Development

Tool: Data Literacy Playbook

Climbing the Slope

Deep Learning

Analysis By: Mike Fang, Svetlana Sicular, Alexander Linden

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Deep learning (DL) is a variant of machine learning algorithms. It uses multiple layers to solve problems by extracting 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 more complex problems with higher accuracy and less manual tuning.

Why This Is Important

Deep learning often outperforms traditional machine learning or shallow learning techniques in the presence of complex and very high dimensional data, such as images, speech and text. It can reduce the need for tedious feature engineering.

Business Impact

Deep learning is key for a lot of breakthroughs in the field of AI in many domains. It allows organizations to generate insights from disparate, especially unstructured, 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 much meaning, but in a group may highlight results that would have been neglected or not even surfaced. DL applicability has been most successful in vision, speech and text domains.

Drivers

- Algorithmic breakthroughs, such as the recent advancements in natural language processing (NLP) techniques using DL methods, have propelled the use of foundational models that promise state-of-the-art results in conversational platforms.
- The performance improvement and affordability from vast computing architectures, such as graphics processing unit (GPU), clustered computing and cloud, are driving adoption.

- Availability of off-the-shelf solutions is also a driver for DL.
- The vast availability of huge training datasets, such as image, audio, video or text, is enabling the use of DL techniques to help organizations enrich their decision-making process.
- Methods such as reinforcement learning, transfer learning, deep belief networks and evolutionary learning are driving the use of DL.
- Graph neural networks, used to embed dependencies between nodes from the graph, together with the deep neural network, are increasingly being used in computer vision, recommender systems and industries, such as transportation and chemical.

Obstacles

- The infrastructure investments required to create and maintain DL solutions are still high.
- DL methods are construed as a black box by nature, so governing and ensuring the explainability of these solutions is a constant challenge.
- AI energy consumption is particularly high for advanced DL models and is harmful to the environment.
- The skills required to create and manage DL solutions from scratch are hard to come by.
- Support and capabilities around security, privacy and governance for vendors providing DL capabilities as a service are limited, which adds a layer of complexity over already black-box implementations.

User Recommendations

- Use DL techniques only when shallow learning techniques have failed to deliver a suitable AI solution.
- Examine and select business areas where deep learning can provide the best value, especially where there is wide and heterogeneous data and the back-box nature of DL isn’t a concern.

- Explore prepackaged solutions first and then move on to custom-made solutions for the business using DL. Create a diverse talent pool from industry and academia to ensure interpretability, as well as privacy, compliance, ethics and governance in DL solutions.
- Connect symbolic systems with deep learning to pilot some state-of-the-art composite AI techniques, such as neural symbolic AI, graph neural networks, and deep reinforcement learning to reduce technical debt and promote reusability, consistency and explainability.

Gartner Recommended Reading

Innovation Tech Insight for Deep Learning

Market Guide for DSML Engineering Platforms

Market Guide for Multipersona Data Science and Machine Learning Platforms

Introducing Deep Learning Abstraction Methods

3 Types of Machine Learning for the Enterprise

Data Labeling and Annotation

Analysis By: Svetlana Sicular, Alexander Linden, Anthony Mullen

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Data labeling and annotation (DL&A) is a process where data assets are further classified, segmented, annotated and augmented to enrich data for better analytics and artificial intelligence (AI) projects. Associated services and platforms route and allocate these tasks to both internal staff and external third-party knowledge workers to optimally manage the required workflows and thus improve the quality of training data.

Why This Is Important

The need for better training data has increased to remove the bottleneck in developing AI solutions — especially those particular to generative AI and industry use cases. Given the typical lack of internal skills and systems, DL&A services and tools 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 at their current levels without human-based labeling and its further automation.

Business Impact

Major impacts of DL&A are:

- Enables AI solutions where they may not have been possible before due to lack of training data
- Improves accuracy of foundation models, mostly via reinforcement learning with human feedback (RLHF)
- Adjusts generative AI outcomes to organization-specific needs
- Drives higher performance of AI solutions because of richer, annotated datasets
- Speeds times around model development — can flex to accommodate varying workloads
- Can support a reboot of metadata management practices

Drivers

Growth of investments in AI:

- **Increased exploitation of pretrained, off-the-shelf models**: When models are pretrained, data is the main means for customization and fine-tuning of the models.
- **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.

- **A shift from tactical to strategic offerings:** Some language-centric platforms such as insight engines, semantic AI platforms and intelligent document processing solutions connect labeling tasks to master metadata management, yielding reusable strategic assets like taxonomies, ontologies and knowledge graphs.
- **Major AI vendors offer DL&A crowdsourced and automated services** to streamline the AI process on their AI and ML platforms.

Rise of generative AI:

- Growth of large language model (LLM) offerings: Natural language technology workload outsourcing for speech, conversational AI and document labeling is a major area of growth in this market. The DL&A market is expanding beyond row data to look at corpora quality to make LLMs more performant.
- RLHF, a type of LLM fine-tuning, involves a reward model that rates how good a model's response is from the perspective of a human labeler — this guides the model's learning process. The process involves human-labeled prompts and labeling preferred responses.
- Generative AI methods allow lowering the cost of DL&A through automation. LLMs are increasingly used to extract labels from text data through zero-shot learning.

Obstacles

- **Challenger methods:** Few-shot learning, transfer learning, synthetic data, semantic platforms and data marketplaces compete for use.
- **Third-party worker quality:** 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 operation" for organizations.
- **Supply outstrips demand and price points are often uneconomical for large-scale data:** Many vendors have entered this space in the last few years, 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.
- **Security concerns:** Especially for those DL&A services that bring in public crowds, many clients feel uneasy distributing certain data to virtually unknown parties.

User Recommendations

- Design development and production workflows to leverage a mixture of internal and external knowledge workers to support data labeling.
- Prioritize DL&A automation options when possible.
- Ensure the provider you choose has methods to test its pool of knowledge workers for domain expertise and measures of 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 on more valuable tasks and lighten their load in classifying and annotating data by using DL&A services.
- Use vendors 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.
- Mature ad hoc tactical labeling activities to a systemwide in-the-line-of-work approach to curate ongoing master metadata schemas.

Sample Vendors

CrowdWorks; Defined.ai; Diffgram; Heartex; Isahit; Labelbox; Mindy Support; Scale AI; Snorkel AI

Gartner Recommended Reading

Best Practices for the Responsible Use of Natural Language Technologies

Market Guide for AI-Enabled Translation Services

Emerging Tech: Tech Innovators in Synthetic Data for Image and Video Data — Domain-Focused

Computer Vision

Analysis By: Nick Ingelbrecht, Shubhangi Vashisth

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

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

Why This Is Important

Computer vision comprises a transformational collection of technologies that are essential to sensing and understanding the physical environment. Computer vision technology is driving innovation across many industries and use cases and is creating unprecedented business applications and opportunities.

Business Impact

Computer vision technologies are used across all industries and address a broad and growing range of business applications. These include physical security, retail and commercial property, automotive, robotics, healthcare, manufacturing, supply chain/logistics, banking and finance, agriculture, government, media and entertainment, and Internet of Things (IoT). Computer vision exploits the visible and nonvisible spectrum, including infrared, hyperspectral imaging, lidar, radar and ultraviolet.

Drivers

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

- **New neural network architectures, models and algorithm enhancements** are steadily improving the price/performance of computer vision applications; combinations of CNNs and vision transformers are delivering leading levels of performance; model compression and chip advancement enable larger workloads to be run on edge devices.
- **The economics of computer vision are being enhanced by the growth of the market for computer vision tools and services.** These include annotation and data preparation services and automated machine learning (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. Dynamic vision systems and lower cost lidar products are opening new areas for innovation.
- **Edge-enabled cloud frameworks,** developer ecosystems, products and support are further expanding the opportunity and enabling non-experts to train and deploy their own computer vision models.
- **New business models and applications** are emerging, 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.

- **Sensor fusion,** multimodal analysis, generative AI, multispectral and hyperspectral imaging are expanding the opportunities.
- **Improved reliability,** price, performance and functionality are generating compelling business value and driving adoption.

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.
- 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.
- Reduce the barrier to computer vision adoption by addressing two of the main challenges, lack of training data and costly and constrained hardware, by investing in synthetic and augmented data solutions and model compression to improve model performance and expand the range of more valuable use cases.

Sample Vendors

Amazon Web Services; Baidu; Clarifai; Deepomatic; Google; Matroid; Microsoft Azure; Tencent

Gartner Recommended Reading

Emerging Technologies: Emergence Cycle for Computer Vision

Emerging Tech: Revenue Opportunity Projection of Computer Vision

Emerging Technologies: Computer Vision Is Advancing to Be Smarter, More Actionable and on the Edge

Emerging Technologies Tool: Video Analytics Functionality Matrix

Emerging Technologies: Tech Innovators for Computer Vision

Event Stream Processing

Analysis By: W. Roy Schulte, Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Event stream processing (ESP) is computing that is performed on streaming data (sequences of event objects) for the purpose of stream analytics or stream data integration. ESP is typically applied to data as it arrives (data “in motion”). It enables situation awareness and near-real-time responses to threats and opportunities as they emerge, or it stores data streams for use in subsequent applications.

Why This Is Important

ESP enables continuous intelligence and real-time aspects of digital business. ESP’s data-in-motion architecture is a radical alternative to the conventional data-at-rest approaches that have historically dominated computing. ESP platforms have progressed from niche innovation to proven technology, and now reach into the early majority of users. ESP will reach the Plateau of Productivity in less than two years and eventually be adopted by multiple departments within every large company.

Business Impact

ESP transformed financial markets and became essential to telecommunications networks, smart electrical grids, and some IoT, supply chain, fleet management and other transportation operations. However, most of the growth in ESP during the next 10 years will come from areas where it is already established, especially IoT and customer engagement. Stream analytics from ESP platforms provide situation awareness through dashboards and alerts, and detect anomalies and other significant patterns.

Drivers

Six factors are driving ESP growth:

- Organizations have access to ever-increasing amounts of low-cost streaming data from sensors, machines, smartphones, corporate websites, transactional applications, social computing platforms, news and weather feeds, and other data brokers. Many new AI and other analytical applications need this streaming data to satisfy business requirements for situation awareness and faster, more-accurate decisions.

- The wide use of Apache Kafka and similar streaming messaging systems is reducing the cost and complexity of ingesting, storing and using streaming data.

- Conventional data engineering pipelines take hours or days to prepare data for use in BI and analytics, causing delays that are unacceptable for some purposes. Therefore, an increasing number of data engineering pipelines are being reimplemented as real-time data flows (continuous ETL) in ESP platform products or stream data integration tools with embedded ESP. These real-time data flows filter, aggregate, enrich, and perform pattern detection and other transformations on streaming data as it arrives.

- ESP products have become widely available, in part because open-source ESP technology has made it less expensive for more vendors to offer ESP. More than 30 ESP platforms or cloud ESP services are available. All software megavendors offer at least one ESP product, and numerous small-to-midsize specialists also compete in this market. Cloud ESP platforms have lowered the cost of entry.

- Vendors are embedding ESP platforms into a wide variety of other software products, including industrial IoT platforms, stream data integration tools, unified real-time platforms (aka continuous intelligence platforms), insider threat detection tools and AI operations platforms.

- Vendors are adding highly productive development tools that enable faster ESP application development. Power users can build some kinds of ESP applications via low-code techniques and off-the-shelf templates.

Obstacles

- ESP platforms are overkill for many applications that process low volumes of streaming data (i.e., under 1,000 events per second), or that do not require fast response times (i.e., less than a minute). Conventional BI and analytics tools with data-at-rest architectures are appropriate for most stream analytics with these less-demanding requirements.
- Many architects and software engineers are still unfamiliar with the design techniques that enable ESP on data in motion. They are more familiar with processing data at rest in databases and other data stores, so they use those techniques by default unless business requirements force them to use ESP.
- Some streaming applications are better-implemented on unified real-time platforms that process both data in motion and data at rest. Some unified platforms use embedded open-source ESP platform products, while others get their ESP capabilities from custom internal code.

User Recommendations

- Use ESP platforms when conventional data-at-rest architectures cannot process high-volume streams fast enough to meet business requirements.
- Acquire ESP functionality through a SaaS offering, an IoT platform or an off-the-shelf application that has embedded ESP logic if a product that targets specific business requirements is available.
- Use vendor-supported closed-source platforms or open-core ESP products that mix open-source with closed-source extensions for applications that need enterprise-level support. Use free, community-supported, open-source ESP products if developers are familiar with open-source software, and license fees are more important than staff costs.
- Use ESP platforms or stream data integration tools to ingest, filter, enrich, transform and store event streams in a file or database for later use.
- Choose a unified real-time platform with embedded ESP capabilities over a plain ESP platform if the application uses both data at rest and data in motion.

Sample Vendors

Confluent; EsperTech; Google; Hazelcast; IBM; Microsoft; Oracle; SAS; Software AG; TIBCO Software

Gartner Recommended Reading

Market Guide for Event Stream Processing

5 Essential Practices for Real-Time Analytics

Create an Optimal IoT Architecture Using 5 Common Design Patterns

Adopt Stream Data Integration to Meet Your Real-Time Data Integration and Analytics Requirements

Appendixes

See the previous Hype Cycle: Hype Cycle for Data Science and Machine Learning, 2022

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 2023)

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 2023)

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 constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (July 2023)

Document Revision History

Hype Cycle for Data Science and Machine Learning, 2022 - 29 June 2022

Hype Cycle for Data Science and Machine Learning, 2021 - 2 August 2021

Hype Cycle for Data Science and Machine Learning, 2020 - 28 July 2020

Hype Cycle for Data Science and Machine Learning, 2019 - 6 August 2019

Hype Cycle for Data Science and Machine Learning, 2018 - 23 July 2018

Hype Cycle for Data Science and Machine Learning, 2017 - 28 July 2017

Hype Cycle for Data Science, 2016 - 25 July 2016

Hype Cycle for Advanced Analytics and Data Science, 2015 - 6 July 2015

Recommended by the Author

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 2021

Top Trends in Data and Analytics, 2022

The State of Data Science and Machine Learning

Market Guide for Multipersona Data Science and Machine Learning Platforms

Market Guide for DSML Engineering Platforms

Table 1: Priority Matrix for Data Science and Machine Learning, 2023

Benefit	Years to Mainstream Adoption			
↓	Less Than 2 Years ↓	2 - 5 Years ↓	5 - 10 Years ↓	More Than 10 Years ↓
Transformational	Computer Vision Event Stream Processing	Citizen Data Science Composite AI Deep Learning Generative AI Self-Supervised Learning	Composable D&A Foundation Models	Quantum ML
High	Data Labeling and Annotation Natural Language Query	Augmented Analytics Chief Data Scientist Edge Analytics Generative Analytics Experience MLOps Multistructured Analytics Prompt Engineering Synthetic Data	Adaptive ML Data Science Education Explainable AI Federated Machine Learning Graph Data Science Reinforcement Learning Transfer Learning	Decision Engineer
Moderate		Logical Feature Store	Differential Privacy	
Low				

Source: Gartner (July 2023)

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 2023)

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 2023)

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 constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (July 2023)