
VOLUME 5.0 | 2023-24
AINA
AI & ANALYTICS MAGAZINE
AI Governance
Behind The Scenes
Machine Unlearning
AutoML & Path Ahead
AI in Pharma Marketing
The Invisible Shopkeeper

From The Team

Dear Readers,

Welcome to the fifth edition of AINA, India's only student-driven AI and analytics magazine. As we present this edition, we are filled with excitement and gratitude. Our journey has been driven by a passion for exploring the cutting-edge developments in AI and analytics, and we are thrilled to share these insights with you.

This edition is a testament to the hard work, creativity, and dedication of our entire team. We have meticulously curated content that highlights the latest trends, groundbreaking technologies, and innovative applications in the fields of AI and analytics. Our goal is to provide you with valuable knowledge that not only informs but also inspires you to think about the future possibilities in these dynamic fields.

We are deeply thankful for the unwavering support from our esteemed faculty at IIM Calcutta, ISI Kolkata, and IIT Kharagpur, whose guidance has been instrumental in shaping this magazine. A special thanks to Dr. Sujoy Kar and Alok Mani Singh for their insightful interviews, and to Pratyusha Addula for her thought-provoking article. Their contributions add depth and perspective to this edition.

We also want to extend our heartfelt thanks to the AINA 4.0 team for their continuous support and guidance. A special mention goes to Aditya Gadepalli, the Editor-in-Chief of the first edition, whose mentorship was invaluable during our journey.

As you delve into the pages of this magazine, we hope you find inspiration and insights that spark your curiosity and passion for AI and analytics. We have poured our hearts into creating this edition, blending our learnings, creativity, and honest effort to bring you something truly special.

Thank you for being a part of our journey. Happy reading!

Warm regards,
The AINA Team

HIMANSHU SINGH
Editor-In-Chief

DEEPT PANCHAL
Lead Designer

AKHIL RAYANKULA
Creative Head

ANIMESH PATEL
Content Architect

SOUMENDU MANDAL
Features Editor

From the Desk of PGDBA Chairperson, IIM Calcutta

Post Graduate Diploma in Business Analytics (PGDBA) is a full-time residential program in business analytics offered jointly by the Indian Institute of Management Calcutta (IIMC), the Indian Statistical Institute Kolkata (ISI-K), and the Indian Institute of Technology Kharagpur (IIT-KGP). Through a highly competitive selection process and a good mix of experienced professionals and fresh graduates, the PGDBA program has an impressive batch and alum base catering to the increasing global demand for business data scientists.

The PGDBA program's uniqueness stems from its integration of business, statistics, and technology and an extensive internship. Our students study various statistical and machine learning theories for analytics at ISI-K, technological aspects of analytics at IIT-KGP, and applications of analytics in the functional areas of management at IIMC. They are primed for success due to the rigorous and demanding learning at the three institutes through theory, case studies, labs, and business simulations.

What truly sets this program apart is its unique six-month industry internship that equips participants with hands-on experience and makes them wholly industry-ready for an upward and onward career in business analytics. Indeed, the fusion of business, technology, and statistics, along with the long industry internship, gives our PGDBA program students a head start in the professional world. I am confident that the PGDBA program will continue to evolve and expand, meeting the increasing demand for business data professionals worldwide and significantly contributing to the industry.

Sudhir S. Jaiswall
PGDBA Chairperson
IIM Calcutta

INDIAN INSTITUTE OF MANAGEMENT
CALCUTTA

From the Desk of PGDBA Coordinator, IIT Kharagpur

The Post Graduate Diploma in Business Analytics (PGDBA) program, a unique collaboration between three of India's premier institutions - Indian Institute of Management Calcutta (IIMC), Indian Statistical Institute (ISI), and Indian Institute of Technology (IIT) Kharagpur - is designed to provide rigorous training to equip students with the skills and knowledge required to excel in the rapidly evolving field of business analytics.

This program is not just about acquiring theoretical knowledge, but also about gaining practical experience through hands-on training and industry internships. One of the key highlights of the PGDBA program is the six-month industry internship which helps students bridge the gap between theory and practice, preparing them to tackle the challenges that they face in their professional careers.

The program also emphasizes the importance of continuous learning and staying updated with the latest advancements in the field of analytics. Students have access to a wealth of resources, including research papers, industry reports, and online courses, to help them stay ahead of the curve. I am confident that the dedication of the students and support of the faculty members and administration of the three Institutes will take the program to greater heights and enable it to cater to the increasing demand of data scientists globally.

Sangeeta Sahney
PGDBA Coordinator
IIT Kharagpur

From the Desk of PGDBA Coordinator, ISI Kolkata

The Post Graduate Diploma in Business Analytics (PGDBA) program offers a unique opportunity for students to receive high-quality education and training in business analytics. A key aspect of this program is that the students get trained in analytics, technology, and management from the top three institutes of the country: ISI, IITKGP, and IIMC. In addition, the six-month industry internship on analytics projects is a crucial component of the program, which fully prepares students for the professional world.

In essence, the curriculum of the PGDBA program is designed not only to meet the demands of the industry but also to prepare students comprehensively for the field of business analytics.

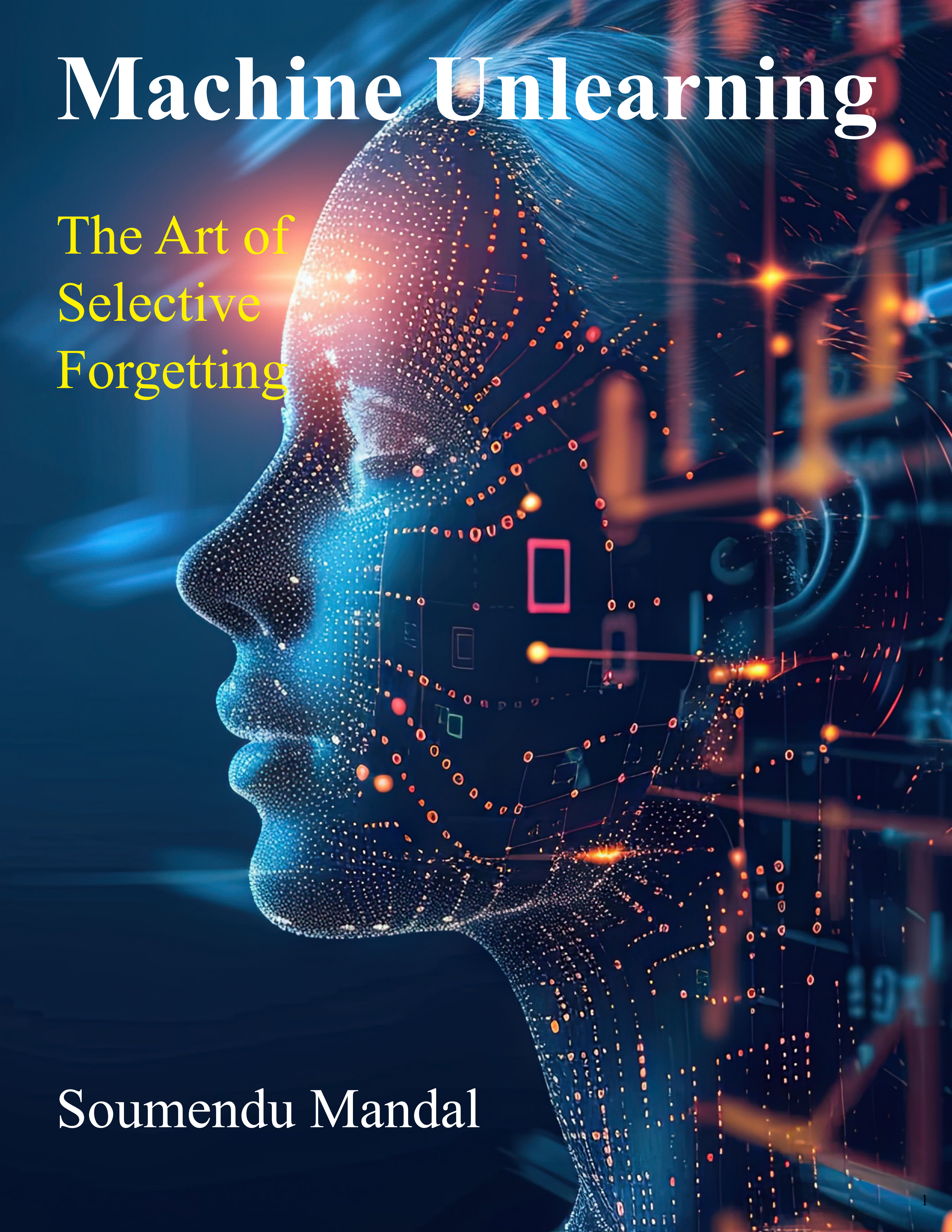
At ISI, we have a legacy of excellence in education and research. Our institution is renowned globally for its contributions to the fields of statistics, data science, computer science and mathematics. The faculty in the PGDBA program are experts in statistical and machine learning theories for analytics. Their expertise and guidance significantly enhance the quality of education provided. Besides maintaining meticulous academic standards, we encourage students to engage in sports activities, social engagement, and interaction with faculty and researchers. I am confident that the PGDBA program will continue to grow and make significant contributions to the analytics world.

Anisur Rahaman
PGDBA Coordinator
ISI Kolkata

Contents

Machine Unlearning

The Art of Selective Forgetting



Soumendu Mandal

Imagine this: You excitedly post a picture of yourself winning a baking competition on social media. Weeks later, you decide the photo isn't the best representation of you and request its removal. The platform removes the picture, but what about the countless copies that might exist across the web?

The idea of machines erasing memories might resonate with anyone who has ever wished to take back something they've posted online. Yet, for artificial intelligence, this concept of "artificial amnesia" demands pioneering techniques in computer science. Firms dedicate substantial resources to developing machine learning models for tasks such as facial recognition or social media content ranking, given their ability to process information faster than human programmers. However, these models, once trained, become infamously rigid and opaque, making them challenging to alter or fully comprehend. Traditionally, to diminish the impact of a particular data point, the entire model must be retrained from the ground up—a process that is not only costly but also labour-intensive.

This is where machine unlearning comes in. Machine unlearning allows artificial intelligence (AI) models to forget information they were trained on. Just like you want to forget that not-so-flattering baking picture, machine unlearning empowers AI to "forget" specific data points.

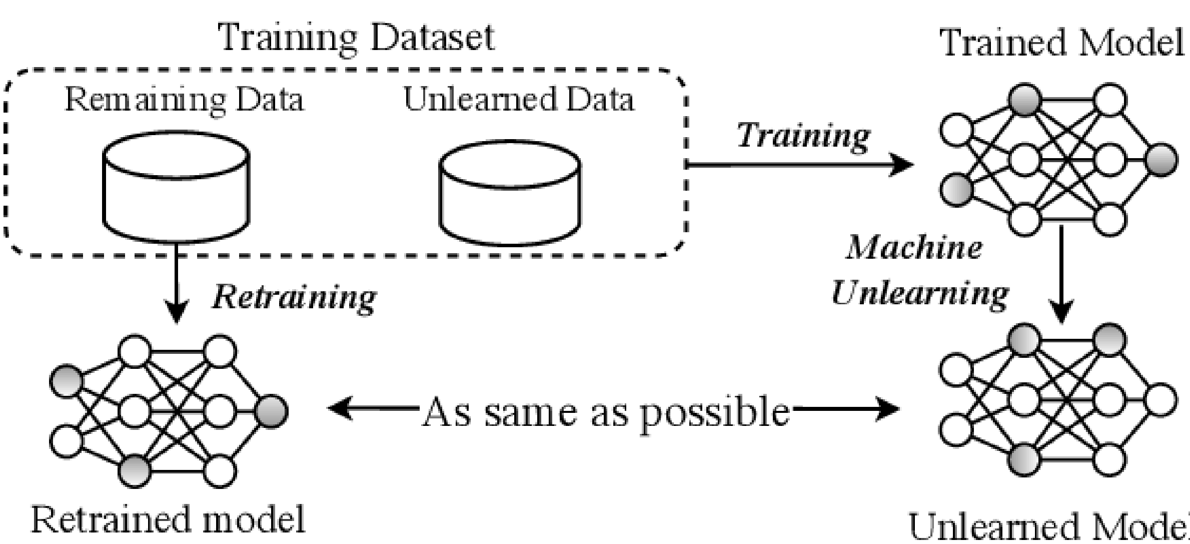
This concept is particularly relevant in today's data-driven world. As AI becomes more integrated into our lives, from influencing search results to personalizing recommendations, concerns about privacy and data security rise. Machine unlearning offers a potential solution, ensuring individuals have control over their data and the ability to be "forgotten" by AI systems.

However, machine unlearning is a complex challenge. Unlike deleting a picture from your phone, simply removing data from a training dataset isn't enough. AI models are complex, and the influence of a single data point can be interwoven throughout the entire system.

So, if you're curious about how AI can learn to forget, how it can be applied to real-world scenarios, and the potential impact it can have on data privacy, keep reading!

A Historical Perspective on Machine Unlearning

In the early days of AI, the prevailing mentality was that more data equals better results. Researchers strived to create algorithms that could consume and learn from ever-increasing datasets. This approach, inspired by the human brain's seemingly limitless capacity for learning, yielded significant advancements in areas like image recognition and natural language processing.



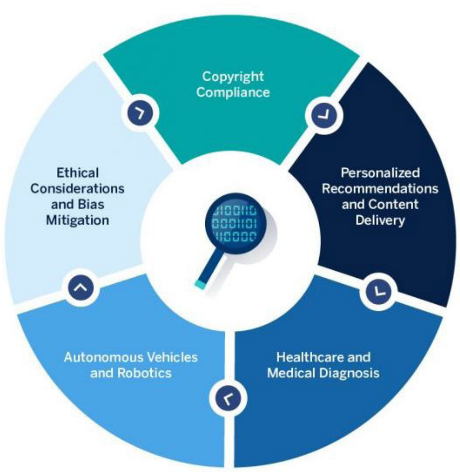
Training an unlearned model that closely resembles a retrained model using only the remaining data.

However, as AI became more integrated into our lives, concerns about privacy and data security began to surface. The realization that AI models, once trained on a specific data point, might retain that information indefinitely, even if the user wished it to be erased, sparked a critical shift in perspective.

This shift gave rise to the concept of machine unlearning. The idea was to equip AI with the ability to "forget" specific data points, like how an individual might choose to forget a negative experience. This would empower users with greater control over their personal information and ensure compliance with regulations like the General Data Protection Regulation (GDPR) in Europe, which grants individuals the "right to be forgotten."

The idea was to equip AI with the ability to "forget" specific data points, like how an individual might choose to forget a negative experience.

Developing effective unlearning techniques proved challenging. Unlike deleting a file from your computer, unlearning requires a nuanced understanding of how a particular data point has influenced an AI model's internal parameters. These parameters, often complex and interconnected, represent the AI's accumulated knowledge. Removing a single data point without disrupting the entire model's functionality is no easy feat.



Application of Machine Unlearning

Despite the hurdles, researchers have achieved substantial advancements in crafting unlearning algorithms. Techniques like SISA training. The story of machine learning and unlearning is a testament to the ongoing evolution of AI. As technology advances, we strive to create not just intelligent systems, but also responsible ones that respect user privacy and data security.

Why Machines Must Unlearn

Machines forgetting information? It might sound strange, Yet the ability for AI to selectively unlearn is no longer science fiction, but a necessity for several compelling reasons. This "right to be forgotten" empowers individuals with control over their digital footprint.

Training LLMs on massive datasets increases the risk of incorporating biased, copyrighted, or privacy-sensitive information. Unlearning algorithms act as a digital eraser, selectively removing these data points from AI models. This ensures your privacy and prevents these models from basing decisions on outdated or irrelevant information.

"The illiterates of the 21st century will not be those who cannot read and write but those who cannot learn, unlearn, and relearn."
— Alvin Toffler

Furthermore, unlearning tackles the pervasive issue of bias in AI. Trained on vast datasets, AI models can inadvertently inherit societal prejudices. Imagine an AI used for loan approvals that favours certain demographics based on historical data. Unlearning techniques offer a solution – acting as a corrective measure by identifying and removing biased data points. This allows AI to make fairer and more ethical decisions across various domains like finance, hiring, and even criminal justice. By forgetting biased data, AI can become a more inclusive and equitable force.

and data has a limited shelf life. Unlearning equips AI models with the ability to adapt to these changes. Imagine a self-driving car programmed on a specific traffic pattern. When a new road is built, the car needs to unlearn its previous behaviour and adjust to the altered environment. Unlearning empowers AI models to continuously update their knowledge base, ensuring they remain relevant and reliable in a constantly shifting world.

Finally, unlearning becomes crucial when a model has learned incorrect information during training. Imagine an AI system trained on an inaccurate historical timeline. Unlearning allows the model to "forget" these errors, essentially updating its knowledge base and improving the accuracy of its predictions and classifications. This ensures AI systems maintain a high level of precision and effectiveness in various decision-making processes.

In essence, machine unlearning is not just about forgetting; it's about fostering responsible AI development.

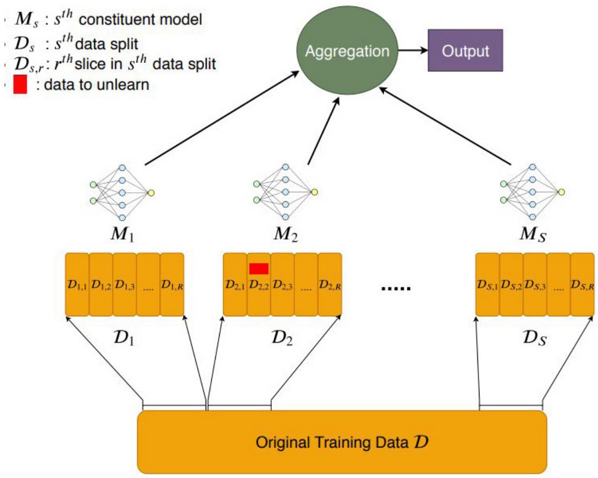
In essence, machine unlearning is not just about forgetting; it's about fostering responsible AI development. It empowers users, tackles bias, ensures adaptability, and allows AI to evolve by letting go of outdated information – like how humans learn and grow by discarding irrelevant experiences. This ability to unlearn paves the way for in a more balanced and responsible way, fostering privacy, fairness, and continuous learning for both.

Methodologies of Machine Unlearning

Exact unlearning stands out as the most definitive form of forgetting, where targeted data points are completely erased from the model's memory, ensuring the model behaves as if it had never encountered the erased data. This method is unequivocal and leaves no trace of the data points, aligning closely with compliance and ethical standards.

One of the most structured approaches to exact unlearning is the Sharding, Isolation, Slicing, and Aggregation (SISA) framework. This technique partitions the training data into multiple, distinct shards, isolates the data intended for removal, and then processes these shards independently to avoid data leakage. After processing, the results from the remaining shards are aggregated to form a new, cleaned model. This method, while thorough, demands significant computational resources and is typically reserved for simpler models where such granular control is feasible.

Alternatively, retraining from scratch is the traditional brute-force approach. By retraining the model on a pruned dataset that excludes the unwanted data points, this method ensures the purged data has no lingering influence. While effective, it's often prohibitively expensive and time-consuming, especially for larger, more complex models.



SISA training: dividing the original training data into multiple splits

Approximate unlearning offers a more feasible route for large-scale models, focusing on reducing—not necessarily eliminating—the influence of certain data. By implementing approximate unlearning, data scientists can significantly reduce computational, storage, and time costs associated with the unlearning process, making it more practical for real-world applications.

Weight decay employs a technique that adjusts a model's connection weights during its training phase. This method introduces a penalty term into the learning algorithm that discourages dependency on connections, thereby subtly reducing the influence of certain data points. By penalizing heavier weights, weight decay ensures that the model prioritizes the most pertinent information, making it an effective tool for refining models.

Fine-tuning involves retraining a model on a dataset that omits data points designated for unlearning. This method is quicker than retraining from the ground up but does not necessarily eradicate all traces of the unwanted data. Fine-tuning is commonly used by data scientists to quickly adapt models to new tasks or datasets, offering a balance between speed and thoroughness in the unlearning process.

Selective retraining updates a model by focusing solely on the data points that continue to be relevant, excluding those marked for unlearning. This approach enhances efficiency and precision over comprehensive model retraining. By concentrating on the essential data, selective retraining ensures the model remains current with minimal retention of undesired information.

Neural architecture modifications take a forward-looking approach to machine unlearning. This method designs models with inherent capabilities for selective memory, including modular structures that facilitate the exclusion of specific model segments or dynamic adjustments to reduce certain data influences. These architectural innovations enhance a model's adaptability and responsiveness, paving the way for more advanced and flexible AI systems capable of targeted data point unlearning.

> The choice of methodology depends on the specific scenario – the type of data, the AI architecture, and the desired level of forgetting.

The evaluation of machine unlearning algorithms involves metrics to assess effectiveness, efficiency, and utility. Challenges in the evaluation metrics include the need for standardized metrics, balancing forget quality with model utility and efficiency, and ensuring fairness in scoring. The evaluation aims to measure how different distributions are after unlearning, comparing the distribution obtained by retraining from scratch without the forget set to that obtained by a specific unlearning algorithm. This process involves defining an evaluation metric that considers forget quality, model utility, and unlearning algorithm efficiency, ultimately providing a comprehensive assessment of the unlearning process.

Additionally, developing metrics to accurately gauge how well data is forgotten and how well the model performs after unlearning is an ongoing challenge. Furthermore, unlearning algorithms must be scalable to handle massive datasets and complex models efficiently. Finally, ensuring fairness in the unlearning process itself is paramount, to avoid introducing new biases. Addressing these challenges will be key to unlocking the full potential of machine unlearning

Challenges of Machine Unlearning

Machine unlearning, while revolutionary, faces hurdles on its path to widespread adoption. Unlike deleting a computer file, unlearning requires a nuanced understanding of how a single data point has influenced a complex AI model. These models function like intricate ecosystems, where a single data point's influence can be woven throughout the entire network. Removing this data point without disrupting the entire model's functionality is a major challenge.

Another hurdle lies in efficiency and scalability. Techniques like fine-tuning require retraining the entire model on a new dataset, which can be computationally expensive and time-consuming. As AI continues to evolve and datasets grow exponentially, finding methods for scalable and efficient unlearning is critical.

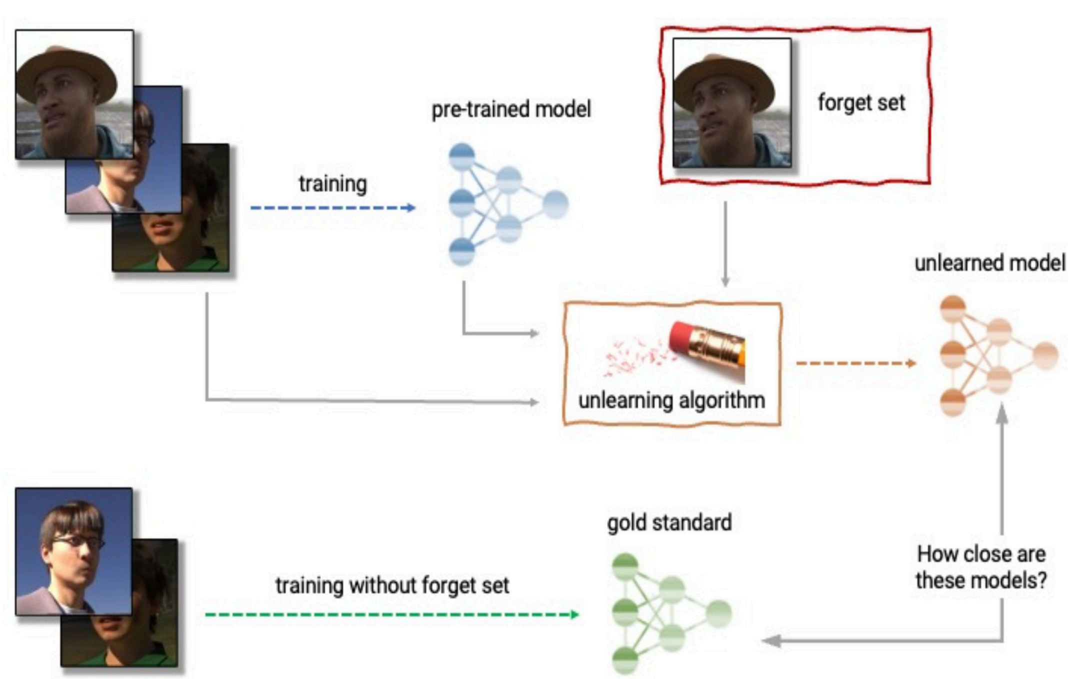
Finally, there's the challenge of balancing forgetting and remembering. Ideally, machine unlearning should be a targeted process, ensuring the model forgets specific information while retaining valuable knowledge. Striking this balance is tricky. Over-aggressive unlearning can lead to a loss of accuracy, while under-unlearning might not adequately address privacy concerns or adapt to evolving data.

Future Perspectives for Machine Unlearning

The field of Explainable AI (XAI) is developing methods to understand how AI models reach their decisions. By leveraging XAI techniques, researchers can pinpoint the influence of specific data points within the model, leading to more targeted and efficient unlearning processes.

Another exciting future perspective is the concept of dynamic parameter adjustment. Researchers are exploring techniques that allow AI models to dynamically adjust their internal parameters in response to unlearning requests. Imagine an AI model with built-in "forgetting mechanisms" that can selectively weaken the influence of specific data points. This approach holds promise for more efficient and scalable unlearning, particularly for large and complex models.

Researchers at The University of Texas developed a method for AI to "forget" unwanted information, such as copyrighted material, without retraining the entire model. This is particularly useful for image-based generative AI models that process massive amounts of internet data.



Google Machine Unlearning Challenge 2023: For face image based age predictor

Finally, machine unlearning aligns perfectly with the growing emphasis on user privacy. As regulations like GDPR become more commonplace, unlearning techniques will play a vital role in ensuring individuals have control over their data and the ability to be "forgotten" by AI systems.

Conclusion: A Paradigm Shift in Human-Machine Interaction

Machine unlearning represents a paradigm shift in human-machine interaction. No longer are AI models passive sponges, absorbing data indefinitely. Instead, they are evolving into dynamic entities capable of selective forgetting, much like humans. This newfound ability unlocks a future filled with possibilities.

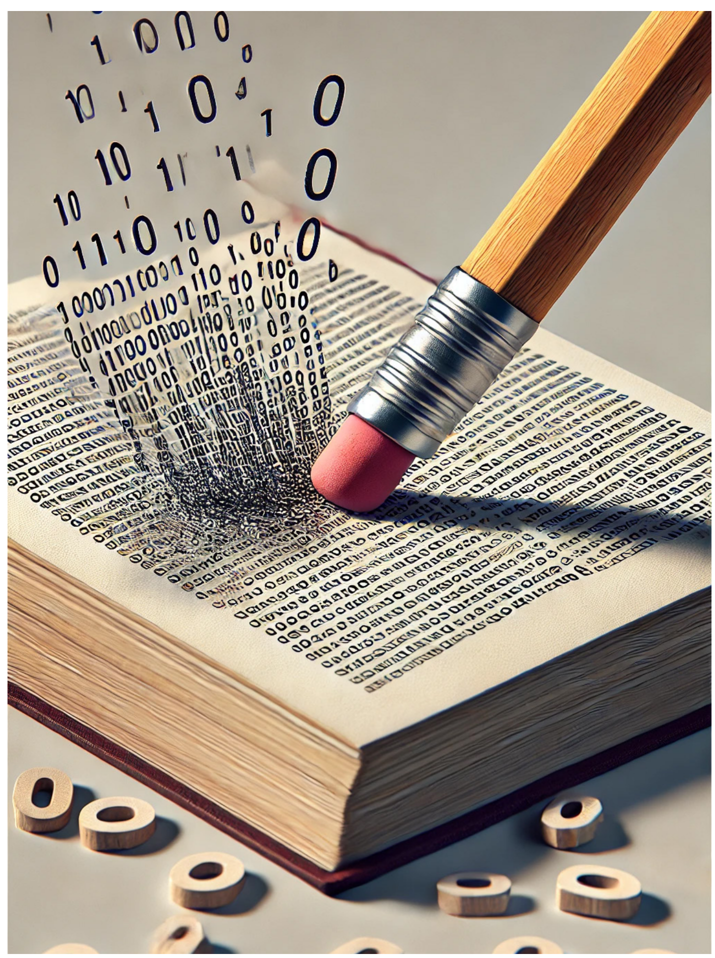
Firstly, machine unlearning empowers individuals with greater control over their digital footprint. The "right to be forgotten" becomes a tangible concept, allowing users to request the removal of outdated or irrelevant data from AI models that might influence decisions about them. This fosters trust and strengthens the ethical foundation of AI integration into our lives.

Secondly, unlearning tackles the issue of bias. AI models trained on vast datasets can inherit societal prejudices. Unlearning techniques can mitigate these biases by identifying and removing data points that perpetuate them. Imagine a loan approval system that learns to disregard creditworthiness based on a borrower's zip code – unlearning can eliminate this bias, ensuring fair and ethical decision-making.

Finally, unlearning ensures AI models remain adaptable in a constantly evolving world. As data becomes outdated models need to adjust. Imagine a self-driving car trained on a specific traffic pattern. When a new road is built, the car must "unlearn" the old route and adapt to the new one. Unlearning allows AI to continuously learn and improve, remaining relevant and reliable.

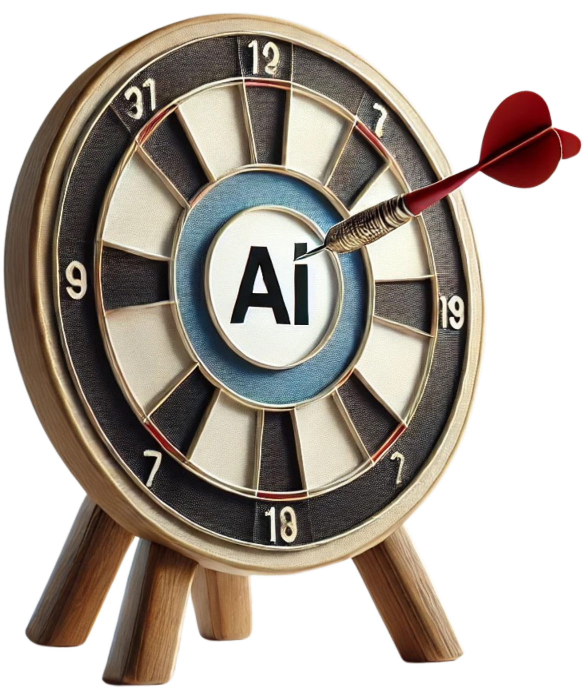
Out of Neural architecture modifications hold the most promise for future machine unlearning due to their efficiency and scalability. Unlike retraining the entire model, these modifications allow for targeted forgetting of specific data points, significantly reducing processing time and resources. Modular designs can be easily adapted to handle evolving datasets and unlearning needs, while dynamic parameter adjustments offer fine-grained control over the forgetting process. This approach paves the way for lifelong learning AI systems that can continuously adapt and forget as needed.

The journey towards widespread adoption of machine unlearning is ongoing, with challenges to overcome and exciting advancements on the horizon. However, one thing is clear: by equipping AI with the ability to forget, we are paving the way for a future where humans and machines co-exist in a more balanced and responsible way. Future respects privacy, combats bias, and fosters continuous learning – a future where both humans and AI can thrive.



Making your **Mark** with **AI**

Pratyusha Addulaz
Senior Data Scientist
Walmart Global Tech India



The role of Data Scientist first appeared in 2008 in companies like Facebook and LinkedIn and has since become very popular. Today, it is one of the most desirable tech jobs for graduates in India. Although HBR's prediction that data science would be the sexiest job of the 21st century seems to be coming true, it comes with its own challenges. The role itself is always changing with new advancements. Skills and tools that were important a few years ago are becoming outdated.

Don't get me wrong, it's exciting to see the latest LLMs becoming popular on social media almost every week. However, this also means that data scientists need to constantly learn and update their skills to keep up with the ever-advancing industry standards. Organizations must regularly check their needs and stay updated by adopting the latest tools and technologies. Along with this, they need to recruit and train the talent needed to manage these changes. These tasks take a lot of time and resources. Hence, companies might choose to buy technology as a service rather than spend time building it.

The build vs. buy dilemma is more common in traditional companies trying to go digital. Since technology isn't their main focus, they are often hesitant to invest time and effort in creating a digital process that might become outdated. At the same time, they don't want to be left behind by market changes. So, the natural choice is to buy these technologies from companies that offer them as a service. But there is a downside to this approach. They won't have control over how the data science products they buy will evolve or be affected by price changes. This can hinder companies from developing a long-term data strategy and executing it efficiently.

As a data scientist looking to build a successful career, navigating this fast-changing market can be challenging. Sometimes the projects you work on for months might become irrelevant due to a new product that the company can buy. Unless you are working at the cutting edge of technology, you might feel unsure about which path will create a lasting impact. Hence, young data science professionals often struggle with which roles to apply for and which companies to join. If the internet hype is to trust, the answer is MANGA (Meta, Amazon, Netflix, Google and Apple) are the leading companies in their respective markets to build your data science career. However, there might not be enough roles in these firms to house all the capable data science graduates. So, in this article, let's take a deeper look at what these companies do that sets them apart and how we can look for other promising companies like them which might not be as popular.

Data

The big five are obsessed with collecting data. Firms which collect proprietary data with customer consent have a competitive edge over others. They control how they use this information and can build sophisticated data science models that can drive efficiencies. Data is not a limitation for them to find new ways to lead and grow in their respective industries. Also, they have a lot of scope for experimentation and incremental improvements with the proprietary data compared to data collected through vendors. If you want to build cutting-edge data science solutions look for companies which collect proprietary information. An example of these are marketplace companies like Swiggy and Myntra. They collect data at a massive scale and use that to build new features in their apps. Eg. Myntra's recent launch of its ' Virtual try on' feature[1]. With the emergence of IoT devices, traditional companies have become a treasure trove of data. High quality and large quantities of data pouring in from various connected devices means there is a need for sophisticated and robust data storage systems to retain it and intelligent AI systems to analyze it.

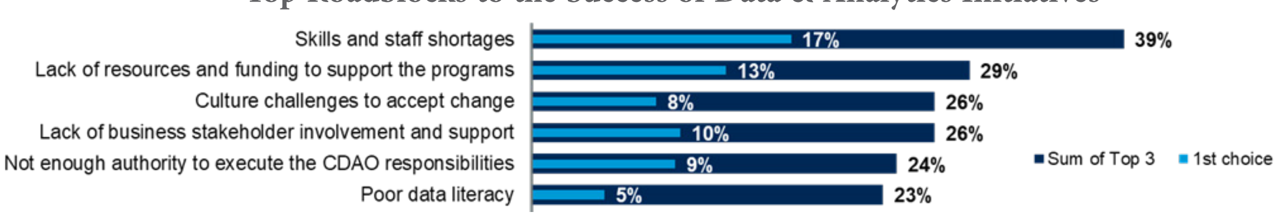
Data-oriented mission statement

A company's mission statement tells you how it envisions the future. Google's mission statement talks about organizing the world's data. Meta's mission is to help people build better communities. Amazon's mission statement is to be Earth's most consumer-centric company. One thing all of them have in common is the success of the mission statement depends on being at the forefront of technology. Hence, they are building cutting-edge tools to analyze the data they collect from their consumers. Data science is the protagonist in achieving their mission. In contrast to this, a lot of data science teams in traditional organizations play more of a supportive role to other departments. Although they make valuable contributions to shaping even long-term strategies for the business, they are not central to achieving the company's mission.

Chief Data Officers

Having a Chief data officer(CDO) or someone equivalent in the C Suite to oversee the data initiatives drives adoption and advances the transformation efforts for the business. A 2018 study published in Emerald Insight (2), looks at 128 firms to evaluate the impact of CDO appointments. After evaluating profit and cost ratios three years prior vs post these appointments, it concluded that CDO appointments significantly increased profit ratios while reducing some cost ratios. Furthermore, the effect on ROS is much more pronounced for large firms. CDO appointments show that firms are paying more attention to data initiatives and realize the importance of data-driven decision-making at the very top. CDO also helps the company produce

Top Roadblocks to the Success of Data & Analytics Initiatives



better products by designing long-term data strategies with the other members of the C suite. It also leads to measurement of tangible results and attribution of results to data initiatives. The business continuity provided by these long-term strategies is crucial when technologies are evolving and changing rapidly.

Data Maturity

With the AI race catching steam and the prophecy of AI singularity looming just years ahead of us, organizations which are further ahead in their digitization journey will benefit more. The origins of the data maturity model date back to 2008, that was when Gartner came up with a categorization for companies to track their digital transformation. There are five levels namely basic, opportunistic, systematic, differentiating and transformational.

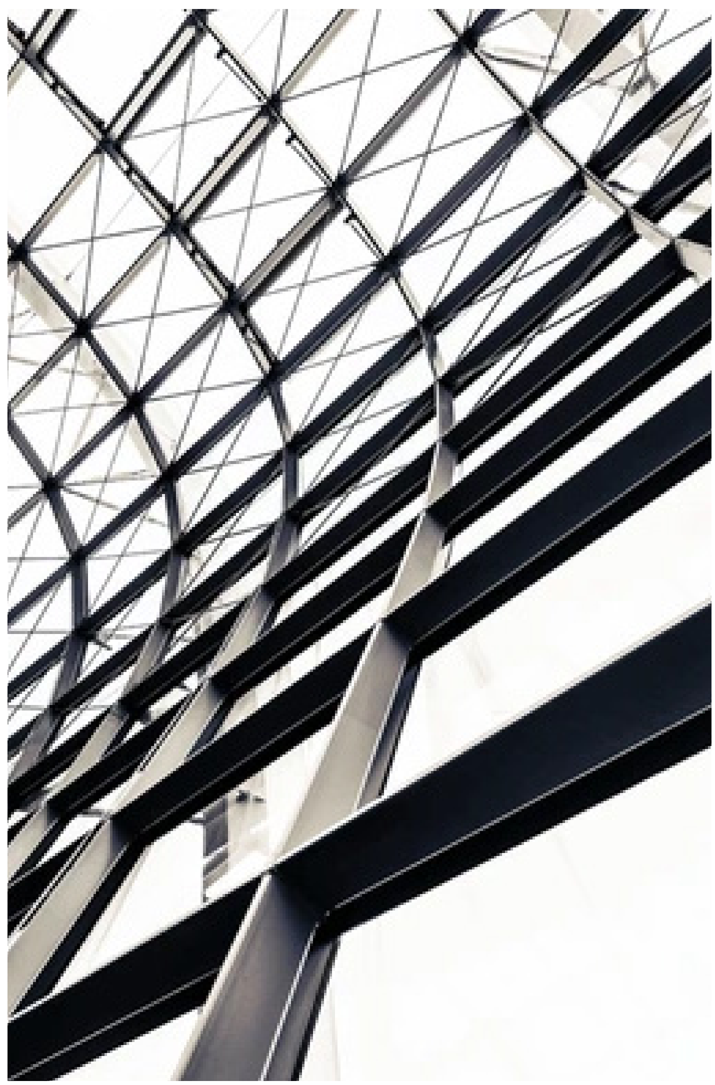
In 2018 a survey was conducted by Gartner [3] globally for 196 organizations, to evaluate their progress on these levels. The advancement proved to be slow. 91% of companies are yet to reach the transformational level. As high as 60% still rank themselves in the bottom three levels. Companies that are further along their maturity journey will have the necessary resources and processes in place and implement AI solutions.

Gartner's D&A effectiveness survey of 2022 [4] reiterated this point by finding that four out of six top hurdles in the D&A team's success are connected to poor data literacy, stakeholder support and lack of acceptance of process changes. Hence, organizations that have surpassed the «systemic level» in the maturity model will not only adopt new technologies with ease but also identify strategies for growth quickly, leading to better data initiatives overall.

It is no surprise when industry professionals' comment that more than 90 percent of data science initiatives fail. Often this failure rate is attributed to factors that have nothing to do with the project's merits. Even when the project meets all the organization's requirements, they are left with the task of getting multiple stakeholders on board to show tangible value and drive project adoption. Data scientists develop the hard skills needed to do this by having a learning mindset and constantly upskilling.

However, it is the soft skills that often prove to be more challenging. They can only be developed by networking and spending time understanding the organization. They help gain the trust of stakeholders and investors that the project will add value in the long term. Data-forward organizations make this journey less turbulent by setting up processes to help pick, advance and measure the impact of data science projects. Like the dotcom boom for web developers, we are experiencing an unprecedented era for data scientists.

The coming few years can turn out to be crucial in shaping your career in the field if you choose your organization wisely.



The Invisible Shopkeeper

Big Data's Silent Revolution in Retail

Akhil Rayankula

In the dynamic landscape of the retail industry, one force has reshaped the way businesses operate, compete, and engage with customers: Big Data. Imagine walking into a store where every aspect of your shopping experience feels personalized and intuitive, from the layout of the shelves to the promotions tailored just for you. This seamless and highly tailored experience is no longer a distant dream but a reality brought to life by the power of big data.

Big data, characterized by the vast volume, velocity, and variety of information generated every second, has become the cornerstone of modern retail strategy. Retailers, ranging from global giants to local boutiques, are leveraging this data to gain unprecedented insights into consumer behavior, preferences, and trends. By analyzing data collected from various sources such as online transactions, social media interactions, and in-store IoT devices like sensors, retailers can now predict what products will be popular, maintain the right level of inventory, and create marketing campaigns that resonate deeply with their target audience.

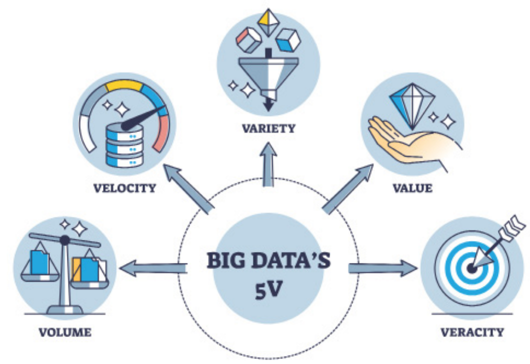
The transformation driven by big data extends beyond customer engagement. Retailers are making their supply chains better, cutting costs, and running more efficiently. Predictive analytics helps to predict demand more accurately, making sure the right products are available at the right time, which reduces the chances of running out of stock or having too much inventory. Additionally, big data analytics provides the capability to detect fraud and ensure secure transactions, safeguarding both the retailer and the customer.

> "Big data refers to the vast, complex datasets that traditional data processing software cannot manage effectively... require advanced tools and techniques..."

In this article, we will delve into the journey of big data in retail, exploring how cutting-edge advancements in data analytics have not only revolutionized customer experiences but also driven significant business growth. Welcome to the transformative world of big data in retail.

Big data

Big data refers to the vast, complex datasets that traditional data processing software cannot manage effectively. The sheer scale and intricacy of this data require advanced tools and techniques for storage, analysis, and visualization. Big data is not just about the volume of data but also encompasses various attributes that make it unique and powerful. These attributes are encapsulated in the concept of the 5Vs: Volume, Velocity, Variety, Veracity, and Value. Understanding these principles provides a foundation for exploring how leading retailers harness big data to transform their businesses.



Volume refers to the enormous amount of data generated every second. Retailers collect data from various sources, including sales transactions, customer feedback, social media, and IoT devices.

Velocity is the speed at which data is generated and processed. In retail, real-time or near-real-time data processing is crucial.

Variety refers to the different forms and sources of data. In retail, this includes structured data from databases, unstructured data from social media, and semi-structured data like emails.

Veracity relates to the accuracy and reliability of data. Ensuring data veracity is essential to avoid errors in decision-making.

Value is the ultimate goal of big data. To derive actionable insights that drive business growth and profitability.

Understanding these five dimensions of Big Data - Volume, Velocity, Variety, Veracity, and Value - is crucial for any business looking to leverage data analytics effectively. However, grasping these concepts is just the beginning. The real challenge lies in implementing them successfully to drive business growth and innovation.

To truly appreciate the transformative power of Big Data in retail, we need look no further than one of the industry's most prominent success stories. Amazon, the e-commerce giant, has not only embraced the 5Vs of Big Data but has built its entire business model around them. Let's delve into how this company has harnessed Big Data to become a dominant force in global retail.

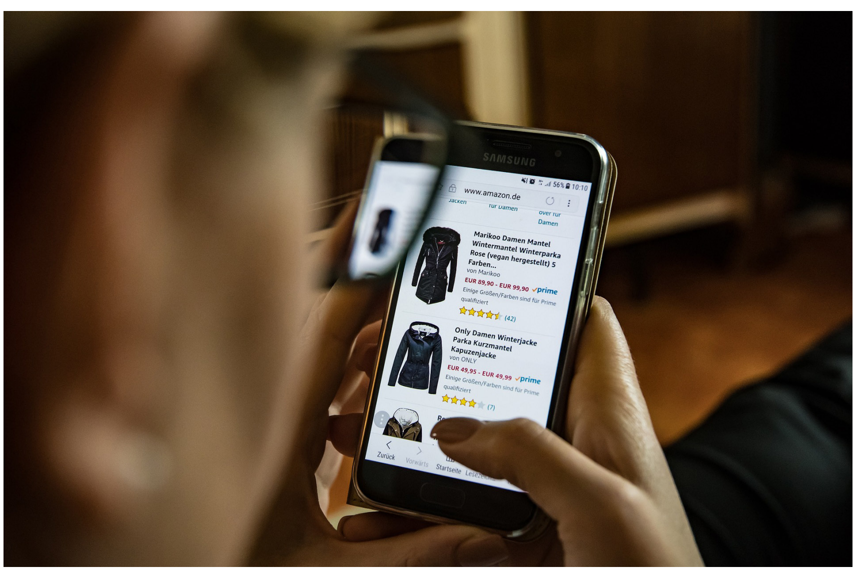
From Books to Behemoth: Amazon's Big Data Revolution

In the annals of retail history, few stories are as captivating as Amazon's meteoric rise from an online bookstore to a global e-commerce titan. At the heart of this transformation lies a powerful force: Big Data. By masterfully leveraging the 5Vs of Big Data - Volume, Velocity, Variety, Veracity, and Value - Amazon has not only reshaped its own destiny but has also redefined the entire retail landscape.

Picture a typical day at Amazon's headquarters. Terabytes of data flood in from millions of customers, transactions, and products. This sheer volume of information would overwhelm traditional systems, but Amazon's robust infrastructure, powered by its own Amazon Web Services (AWS), processes this digital deluge with ease.

> "This velocity of data analysis enables the company to adjust prices dynamically, offer personalized recommendations on the fly, and even predict what customers will buy before they know it themselves."

But it's not just about handling vast amounts of data - it's about speed. Amazon's real-time data processing capabilities are akin to a digital nervous system, instantly responding to changes in user behavior, market conditions, and inventory levels. This velocity of data analysis enables the company to adjust prices dynamically, offer personalized recommendations on the fly, and even predict what customers will buy before they know it themselves.



The true magic happens when Amazon weaves together diverse data strands. From structured transaction logs to the unstructured text of product reviews, the company's algorithms find patterns and insights in this variety of data. It's this ability to see the big picture that has allowed Amazon to expand confidently into new territories, from streaming services to cloud computing.

Of course, all this data would be worthless if it couldn't be trusted. Amazon's commitment to data veracity is evident in its sophisticated fraud detection systems and its continuous efforts to maintain the integrity of customer reviews. By ensuring the accuracy of its data, Amazon has built a foundation of trust with its customers and partners.

The ultimate proof of Amazon's Big Data mastery lies in the value it extracts from this digital gold mine. Every personalized product recommendation, every efficiently routed package, and every new service launched is a testament to Amazon's ability to turn data into dollars. The company's recommendation engine alone is said to drive 35% of its revenue - a staggering figure that underscores the power of data-driven decision making.

This seamless integration of the 5Vs of Big Data into Amazon's core operations exemplifies how theoretical concepts can be transformed into practical, revenue-generating strategies. As we've seen, Amazon's journey offers valuable lessons for other retailers looking to harness the power of Big Data.

> "This seamless integration of the 5Vs of Big Data into Amazon's core operations exemplifies how theoretical concepts can be transformed into practical, revenue-generating strategies."

However, the retail landscape continues to evolve rapidly, driven by technological advancements and changing consumer behaviors. In the next section, we'll explore some of the latest developments in Big Data analytics and how they're shaping the future of retail.

From Cloud to Edge: Innovations Amplifying Retail Big Data Capabilities

The retail landscape is undergoing a seismic shift propelled by innovative big data technologies. This evolution is fundamentally changing how retailers operate, engage with customers, and make strategic decisions. Let's explore the key innovations that are amplifying retail big data capabilities and reshaping the industry.

Cloud computing has emerged as the cornerstone of this retail revolution. Platforms like Amazon Web Services, Microsoft Azure, and Google Cloud offer scalable storage and processing power, enabling retailers to handle vast datasets with unprecedented ease. This shift to the cloud has democratized access to advanced analytics tools, allowing businesses of all sizes to leverage real-time insights and make data-driven decisions swiftly. The cloud's elasticity ensures retailers can manage fluctuating data volumes efficiently, particularly during peak seasons.

Artificial intelligence & machine learning are pushing the boundaries of retail analytics. These technologies excel at predicting customer behavior, optimizing pricing strategies, and streamlining supply chains with remarkable accuracy. AI-powered systems can forecast demand by analyzing historical sales data alongside external factors like weather patterns, social media trends, and upcoming events. This predictive prowess helps retailers maintain optimal inventory levels, reduce costs, and enhance customer satisfaction through improved product availability.

The Internet of Things (IoT) is bringing unprecedented granularity to retail insights. Smart shelves, sensors, and beacons generate real-time data on product movement, inventory levels, and customer interactions within stores. This wealth of information enables retailers to create personalized in-store experiences, such as delivering targeted promotions based on a customer's location. IoT devices also ensure timely stock replenishment, minimizing out-of-stock situations and optimizing store layouts based on customer flow patterns.

Augmented and Virtual Reality are creating immersive shopping experiences that blur the lines between physical and digital retail. AR allows customers to visualize products in their own spaces before purchasing, while VR can transport shoppers to virtual stores for a fully immersive experience. These technologies not only enhance customer engagement but also provide valuable data on preferences and behaviors, helping retailers refine their offerings and marketing strategies.


A Boy shopping using AR/VR headset

Intelligent assistants and visual search technologies are revolutionizing customer interactions. AI-powered assistants offer personalized recommendations and streamline the shopping process, while visual search allows customers to find products using images instead of text. These innovations improve customer satisfaction while offering retailers deep insights into preferences and emerging trends.

Blockchain technology is enhancing supply chain transparency and data security. By providing an immutable ledger, blockchain allows retailers to track products from supplier to shelf with unprecedented accuracy, ensuring authenticity and building customer trust. This transparency also helps retailers comply with increasingly stringent data privacy regulations.

Edge computing is emerging as a game-changer for real-time data processing in retail. By prcessing data closer to its source, such as within a store or on a local server, edge computing reduces latency and enables instant decision making. This is crucial for applications like fraud detection at checkout, real-time inventory management, and personalized customer interactions.

As these technologies converge, they're creating a retail ecosystem that's more responsive, personalized, and efficient than ever before. Retailers can now offer seamless omnichannel experiences, optimize their operations in real-time, and deliver highly targeted marketing campaigns. The future of retail is data-driven, and these innovations are paving the way for experiences that were once the stuff of science fiction.

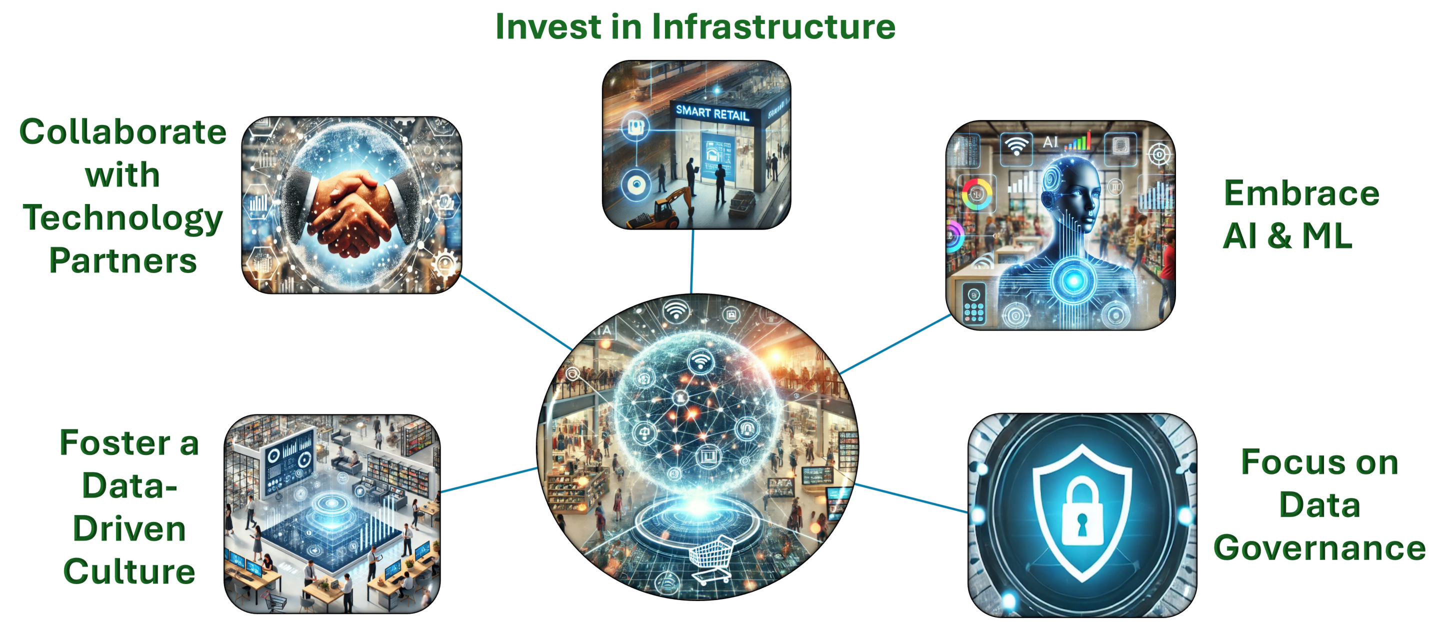
While the potential of these technologies is immense, their true value lies in how retailers integrate and leverage them to create meaningful customer experiences and drive business growth. As we look towards the future, it's clear that the retailers who successfully harness these big data innovations will be best positioned to thrive in an increasingly competitive landscape. In the conclusion, we'll explore the critical success factors for implementing these technologies and the potential challenges retailers may face along the way...

Challenges in Implementing Big Data Solutions

While the potential of big data in retail is huge, there are several challenges that retailers must navigate to fully harness its power.

Data Privacy and Security

One of the foremost challenges is ensuring data privacy and security. With the increasing volume of customer data being collected, retailers must comply with stringent data protection regulations such as GDPR and CCPA. Data breaches can result in significant financial penalties and damage to brand reputation. Retailers need to implement robust cybersecurity measures and maintain transparent data handling practices to build and retain customer trust.



Integration of Data Sources

Retailers often struggle with integrating data from diverse sources, including online transactions, in-store purchases, social media, and IoT devices. Ensuring that these disparate data sources communicate effectively and provide a unified view of customer behavior and operational metrics is a complex task. Implementing a cohesive data integration strategy and utilizing advanced ETL (extract, transform, load) tools can help mitigate this challenge.

Data Quality and Accuracy

Maintaining high data quality is critical for deriving meaningful insights. Inaccurate, incomplete, or outdated data can lead to poor decision-making. Retailers must establish stringent data governance frameworks that include regular data cleaning, validation, and monitoring processes to ensure data accuracy and reliability.

Skills and Expertise

The effective use of big data requires specialized skills and expertise in data science, analytics, and machine learning. Retailers often face a talent gap, struggling to recruit and retain skilled professionals who can interpret complex data and derive actionable insights. Investing in employee training programs and fostering partnerships with academic institutions can help bridge this skills gap.

Scalability and Infrastructure

As the volume of data grows, retailers must ensure that their infrastructure can scale accordingly. This includes not only storage capacity but also the computational power required to process and analyze large datasets in real time. Cloud computing solutions offer scalability, but retailers must carefully manage costs and ensure seamless integration with existing systems.

Change Management

Adopting big data technologies often requires significant changes in organizational culture and processes. Retailers must cultivate a data-driven mindset among employees and encourage collaboration between IT and business units. Effective change management strategies, including clear communication, training, and stakeholder engagement, are essential for successful implementation.

Conclusion

As retailers embark on their big data journey, the horizon is filled with boundless possibilities. The relentless flow of data, coupled with innovative technologies, is transforming the retail landscape. By addressing challenges head-on and fostering a culture of continuous learning and adaptation, retailers can navigate the complexities of the big data landscape. The future of retail is bright, and big data is the key to unlocking its full potential. Embrace the journey, harness the data, and watch as your business transforms in ways you never thought possible.

Alok Mani Singh

Alok is the Director of Data Science at Mastercard's AI Garage, bringing more than 7 years of experience building AI-enabled products and services in the healthcare and payment industries. He has honed his expertise in developing scalable machine learning solutions for fraud and risk detection.

With deep expertise in machine learning at scale, graph neural networks, knowledge graphs, and generative AI, Alok has made significant contributions to the research field, with over five research publications and multiple granted patents. Alok's innovative approach and technical prowess have propelled his career in data science, establishing him as a leading figure in the payment industry.

As an alumnus of the pioneering batch of the Post Graduate Diploma in Business Analytics (PGDBA) program, Alok continues to drive advancements and set new standards in the ever-evolving landscape of AI and data analytics.

AINA: Could you share with us your educational background and what led you to pursue PGDBA?

After graduating from IIT Guwahati in 2011 as a civil engineer, I spent four years working at BHEL as a structural engineer. Then, I decided to switch gears and become a data scientist, which seemed more in line with the current trend at the time.

Initially, I was considering an MBA after my four years in civil engineering. However, as you know, these days everything relies heavily on data, and AI is becoming increasingly important in various industries. That's why I chose to pursue PGDBA than an MBA.

During my undergrad, I had always leaned towards math courses because I was good at it and enjoyed problem-solving. Some friends who worked in analytics suggested PGDBA because of its comprehensive curriculum covering stats, machine learning, and practical business analytics.

Acting on their advice and choosing PGDBA proved to be the right move for me. It gave me the skills I needed to transition smoothly into the field of data science.

AINA: Could you walk us through your professional journey after completing PGDBA, and how did PGDBA helped you in that journey?

Right after graduating from the course I had the chance to work for a Dubai-based NBFC, a small financial institution. It was my first encounter with real-world datasets and the workings of the industry. There, I focused on credit risk and recovery models, gaining insight into the banking sector and the data it deals with.

PGDBA had helped me form a strong foundation in all the necessary skills. Compared to those without this background, PGDBA graduates require less training time when entering the professional realm. As a director at Mastercard now, I've noticed a significant difference between undergraduates and PGDBA graduates, with the latter being well-prepared and polished for industry roles.

PGDBA covers a wide range of skills, from creating presentations to understanding and implementing machine learning algorithms. This knowledge has proved to be valuable in contributing effectively from the outset. Being able to communicate complex ideas to stakeholders is crucial in the industry and PGDBA graduates are adept at conveying insights to stakeholders in a comprehensible manner.

"Being able to communicate complex ideas to stakeholders is crucial in the industry and PGDBA graduates are adept at conveying insights to stakeholders in a comprehensible manner."

I would reiterate that PGDBA's comprehensive curriculum laid the groundwork for my career growth and equipped me to excel in this analytics domain.

AINA: You have worked in United Health Group. Before that you interned with dunia finance also and you are currently working at Mastercard. How data analytics or data science in insurance industry is different from analytics in Finance industry?

Initially, I worked with a Non-Banking Financial Company (NBFC) and later moved to healthcare insurance, focusing on clinical analytics. Currently, I'm with Mastercard, a global technology company in the payments industry. Mastercard's mission is to connect and power an inclusive, digital economy that benefits everyone, everywhere by making transactions safe, simple and accessible.

In NBFCs, just like banks, data typically includes things similar to savings accounts and credit lines. The traditional scope was limited to building credit risk models and some applications for marketing, operations, and risk management. Before the fintech revolution, banking was mostly about reporting portfolio performance. Today, analytics and machine learning have enabled real-time risk assessment, significantly speeding up processes that once took days or weeks.

In healthcare insurance, the risk perspective changes. It's about predicting if a patient poses a financial risk due to potential health issues. We create models to identify and mitigate risks through early intervention, lifestyle changes, or proper medication. For example, I worked on a sepsis prevention model, predicting infection risks in hospitalized patients.

Another project addressed the opioid crisis in the U.S. by predicting potential opioid abuse through key indicators like pharmacy hopping, early refills, opioid dosages, and other co-morbidities of patients. This machine learning model and intervention framework helped in identifying and preventing substance abuse.

In the payments sector, the risk revolves around transaction fraud. We build machine learning models to assess transaction risks, detect fraudulent activities, and ensure safe payments. The definition of risk varies across sectors but fundamentally involves assessing and mitigating potential threats.

AINA: Can you share some of the big impact projects you have worked on or are currently working on?

In addition to the opioid crisis management project, another significant project was a fraud detection model in healthcare, using graph algorithms to detect collusive behaviour among different entities in healthcare such as Provider, Pharmacies and Medical equipment suppliers. Our three-pronged solution targeted provider-level fraud, claim-level fraud, and service line fraud. Traditionally, fraud detection occurred post-disbursement, but our machine learning approach allowed for pre-disbursement risk scoring. This proactive method prioritizes claims, enabling efficient fraud detection before payment.

Currently, we are building generative solutions for better model explainability. This system allows customers to ask questions about model outcomes and receive detailed explanations, enhancing trust and adoption of our products.

Additionally, we are working on various state-of-the-art technologies and payment optimization solutions to continuously improve our offerings.

AINA: What are some of the major challenges you have encountered in the healthcare and financial services industry, particularly related to data science and analytics?

There are primarily three major challenges in financial services concerning the use of data and machine learning models.

First, the distinction between payment networks and financial institutions is crucial. Unlike payment networks, financial institutions manage direct monetary accounts such as savings and checking accounts. They operate under stringent regulatory oversight to protect sensitive personal information (PI) and ensure financial security. Consequently, all models and decisions within these institutions are subject to rigorous audits by independent bodies, such as the Reserve Bank of India (RBI) in India or corresponding authorities in the U.S., to safeguard customer funds and maintain trust.

Second, addressing inherent biases in the data is imperative. Historical data often reflect societal biases related to gender, race, ethnicity, or nationality, which can inadvertently influence the training of machine learning models. As a financial institution, it's crucial to actively remove these biases to prevent discrimination against certain communities. This involves deploying strategies to remove bias from the data to ensure that decisions—such as loan approvals—are made solely based on an individual's ability to repay, without being influenced by sensitive attributes like age, gender, or ethnicity.

> **"addressing inherent biases in the data is imperative. Historical data often reflect societal biases related to gender, race, ethnicity, or nationality"**

Third, scaling, deploying, and monitoring ML models in the payment industry is challenging due to the need for high accuracy, data privacy, and regulatory compliance. These sectors handle vast amounts of sensitive data, requiring robust security measures and real-time performance monitoring to ensure reliability and protect against breaches. Additionally, integrating ML models into existing systems without disrupting operations adds another layer of complexity.

These challenges underscore the necessity for transparent and explainable machine learning models within financial institutions. Previously, simpler and more interpretable models like decision trees were preferred because they facilitated easier explanation during audits. How -ever, the rise of fintech has seen more sophisticated algorithms being adopted, demanding that these new models also adhere to high standards of explainability and fairness to prevent marginalizing certain groups.

AINA: You recently published a paper realted to Graph neural networks and knowledge graphs. Would you like to share something more about it?

The landscape of Graph learning algorithm has significantly evolved over time. Initially, we relied on basic network analysis techniques like spectral clustering and subgraph analysis, employing foundational statistics such as betweenness centrality and PageRank. These early methods laid the groundwork for more advanced graph convolutional algorithms, such as GraphSAGE, and later, even more sophisticated systems like Graph Attention Networks.

Despite these advancements, the industry primarily applied homogeneous graphs, or at most, bipartite graphs, like those representing transactions between cards and merchants. However, these didn't efficiently handle multiple entity interactions that include various attributes, such as card types (Platinum, Gold) or transaction modes (online, offline).

The introduction of knowledge graphs marked a significant shift by enabling a more nuanced representation of multiple relationships and entity types. For instance, in a transaction, a card isn't just a card; it might be categorized by its issuer or its transaction types, and similarly, a merchant might be classified by size or type. This depth of detail enhances the models' capability to manage entity relationships more effectively.

The paper focuses on how these evolved knowledge graph techniques significantly aid in tasks like entity matching. Previously, matching entities across different datasets without a common key was challenging. Knowledge graphs facilitate these connections by creating a comprehensive graph that links related information, making it much easier to match entities.

"In Mastercard, this evolution from using simple homogeneous models to sophisticated knowledge graphs has allowed for richer and more actionable insights."

In Mastercard, this evolution from using simple homogeneous models to sophisticated knowledge graphs has allowed for richer and more actionable insights. We're now able to generate more detailed embeddings that capture a wide array of interactions and attributes, which enhances our analytical capabilities. The paper discusses how these advancements have streamlined processes like entity matching, significantly impacting how we handle data and make decisions in the financial-sector.

AINA: What trends do you foresee in the fintech industry in data science and analytics in the next five years?

Over the next five years, one of the significant trend will be the widespread adoption of generative AI across various use cases. Generative AI is being utilized to detect fraud, among many other diverse applications. Foundational models, like OpenAI's GPT-4, are trained on extensive datasets available on the internet. Industry-specific foundational models can be built to perform multiple tasks. Traditionally, different machine learning models were needed for each specific use case. However, with foundational models trained on domain-specific data, a single master model can now handle multiple tasks due to the enormous amount of training data.

Another emerging trend is real-time prediction, particularly in customer onboarding. Automated machine learning algorithms can provide insights into Know Your Customer (KYC) processes. In the banking sector, image recognition algorithms can authenticate users by analyzing facial profiles and biometrics. This technology can quickly analyze multiple data points to authenticate and onboard customers efficiently. The adoption of image and natural language processing (NLP) algorithms will significantly impact the fintech industry.

Additionally, the fintech sector must incorporate robust fraud detection layers. Peer-to-peer (P2P) fraud detection is becoming more sophisticated, now even targeting social engineering and phishing schemes. Advanced products are being developed to detect and warn customers about suspicious activities, such as screen mirroring scams aimed at stealing one-time passwords (OTPs). Future products will further enhance fraud detection, preventing fraudulent transactions by leveraging extensive data networks and continuously improving models. These advancements will reduce P2P fraud, social engineering scams, and other fraudulent activities.

These are the trends that are shaping the future of generative AI and its impact on fintech industry.

AINA: Which machine learning techniques do you find yourself mostly using in your work ? Simple ones work better or the complex ones?

When working on a project, it's essential to approach it in phases. Initially, the fundamentals remain the same: defining hypotheses, identifying features, and assessing the correlation between these features and the target variable. It's crucial to avoid excessive multicollinearity among variables, as having multiple linearly dependent variables is counterproductive. As a machine learning data scientist, adhering to these foundational principles is essential.

Once the basics are established, you can focus on incremental improvements. For instance, you might start with basic algorithms like XGBoost. From there, you can explore enhancements such as incorporating graph-based methods or adding deep learning layers. These steps build on the foundation, aiming to boost accuracy and performance.

The core objective is to model a function, essentially represented as $y = f(x)$. Regardless of the algorithm — be it deep learning or otherwise — the goal is to learn the parameters that define this function. Understanding the problem, identifying the top features, and applying the right algorithms incrementally are key to success. Whether using parametric or non-parametric methods, the focus is on learning the function that best predicts the target variable.

In specific domains like image processing or NLP, specialized expertise and tools are required. For NLP, for example, you might need to delve into libraries for large language models and specific frameworks. However, in general predictive analytics, the foundational principles remain consistent. By systematically experimenting with different algorithms and assessing their incremental value, you can achieve meaningful improvements in your models.

AINA: Can you please share some insights which one gains after years of professional practice that are less known or seem to be counter intuitive to the students and new recruits ?

I believe the most valuable skill you develop over time is how to structure a problem when the statement is ambiguous. In the industry, you rarely receive clear instructions; instead, you must decipher the problem and design a system accordingly. This involves not only data science but also engineering aspects, like utilizing available resources efficiently, such as GPUs.

"Storytelling is another crucial skill. It's not enough to produce great work; you must also effectively communicate your findings and insights to stakeholders."

Ninety-five percent of the job is understanding where the data comes from, engineering features, deploying models, and interpreting results for end-users. It's not just about building models; it's about ensuring they're practical and impactful. This includes monitoring performance, understanding key metrics, and even aspects of MLops.

Storytelling is another crucial skill. It's not enough to produce great work; you must also effectively communicate your findings and insights to stakeholders. A good data scientist can translate complex analyses into actionable insights that drive value for clients.

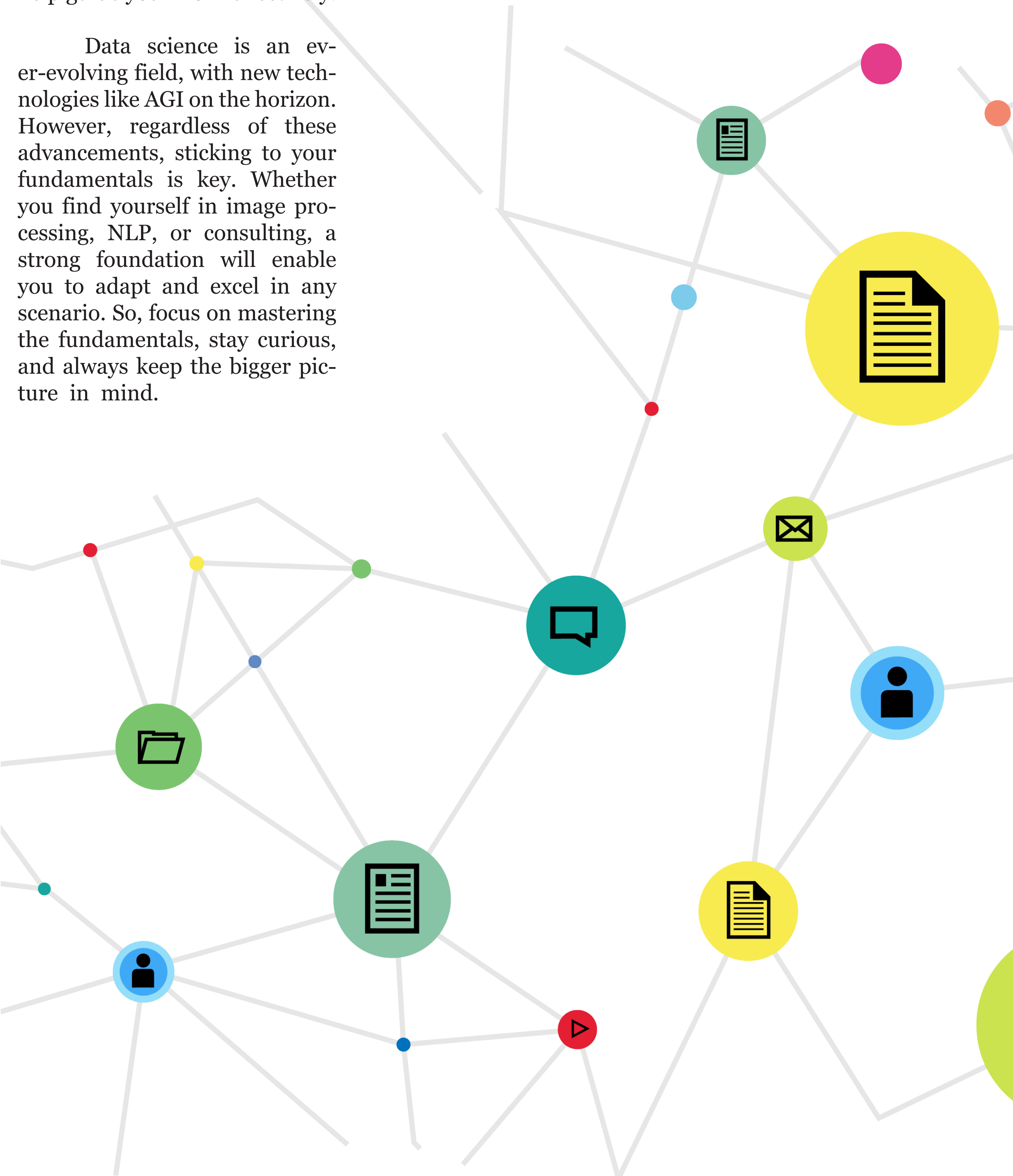
Ultimately, our job as data scientists is to create impact. This requires a holistic approach, from problem-solving to communication, ensuring that our work is not just confined to our laptops but is used to generate meaningful results for our clients.

AINA: What advice would you give to current students including your juniors who aspire to have a career in data science and business analytics ?

My advice is to prioritize brushing up your foundational knowledge. Regardless of the industry's rapid pace of innovation, understanding the core principles of algorithms is crucial. Take decision trees, for example—they serve as the foundation for various algorithms like bagging and boosting. Once you grasp decision trees' workings, comprehending subsequent algorithms becomes much easier, whether in image processing, NLP, or basic machine learning.

Similarly, delve into the basics of linear algebra and mathematics. These subjects form the backbone of data science, and having a solid grasp of them is indispensable. Additionally, maintain a curiosity to learn and understand the bigger picture. Whether you're interning or freelancing, always ask seek to comprehend why certain tasks or projects are initiated. This broader perspective will help guide your work effectively.

Data science is an ever-evolving field, with new technologies like AGI on the horizon. However, regardless of these advancements, sticking to your fundamentals is key. Whether you find yourself in image processing, NLP, or consulting, a strong foundation will enable you to adapt and excel in any scenario. So, focus on mastering the fundamentals, stay curious, and always keep the bigger picture in mind.



Silicon Superstars

How GPUs became the driving force of AI

Akhil Rayankula

Remember when computers were just for spreadsheets and solitaire? Those days are long gone. Now, machines can chat, create art, and even drive cars. But what's making all this artificial intelligence (AI) magic happen? Let's peek under the hood at the chips making it all possible.

The All-Rounder, the Math Genius, and the Multitasker

Imagine a team of workers: you've got the manager (CPU), the calculator whiz (TPU), and the multitasking marvel (GPU). Each has its strengths, but when it comes to AI, the GPU is stealing the spotlight.

CPUs (Central Processing Units) are the brain of your computer, great at handling a variety of tasks. TPUs (Tensor Processing Units), created by Google, are like math geniuses designed specifically for AI calculations. But GPUs (Graphics Processing Units) are the real stars of the AI show. Originally built to make video games look amazing, these chips excel at doing lots of math problems simultaneously – exactly what AI needs.

From Smart to Brilliant: How AI Got a Brain Boost

AI has come a long way from just figuring out cat pictures. A few years ago, AI was using LSTMs (Long Short-Term Memory) to understand language. It was smart, but had limits.

Then came the transformers – a new kind of AI model. They're like super-smart language experts that can understand context and generate human-like text. Purportedly, the biggest of these, GPT-4, has a "brain" which is rumoured to have around 1.76 trillion parameters!

These transformer models can write stories, answer questions, and even code computer programs. It's like having a super-smart assistant that can understand and create human-like text on almost any topic.

The Power Couple: GPUs and AI

As AI got smarter, it needed more brain power. Enter the GPU, ready to flex its muscles.

NVIDIA's A100 GPU can crunch numbers at mind-boggling speeds, enabling researchers to train massive language models. But it's not just about raw power. NVIDIA has built a whole ecosystem of tools and software around its GPUs, making it easier for developers to harness all that computational muscle.

This GPU revolution is having real-world impacts. Doctors use GPU-powered AI to spot diseases in x-rays faster. Climate scientists create more accurate weather models. Self-driving cars rely on GPUs to process sensor information in real-time.

The GPU Advantage in Scientific Discovery

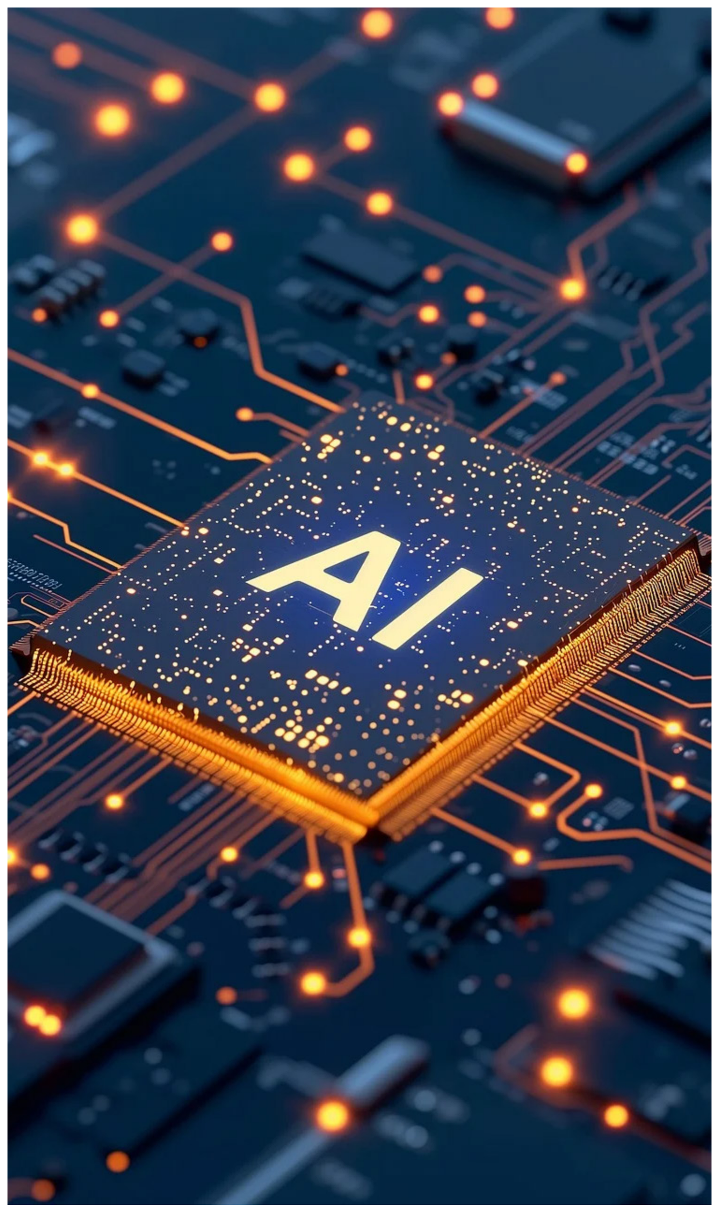
GPUs are also changing the game in scientific research. In drug discovery, researchers use GPU-powered simulations to test thousands of potential compounds quickly. Physicists can now model complex phenomena like black holes or plasma fusion with unprecedented detail.

Looking to the Future

As AI keeps growing, so does the need for more powerful chips. The next generation of GPUs promises to be even faster and more efficient. NVIDIA's upcoming Blackwell architecture is said to be able to run AI models with trillions of parameters, potentially accelerating breakthroughs in fields like climate change mitigation and medical research.

However, challenges remain. As GPUs get more powerful, they also use more energy. The AI industry is grappling with how to make these systems more environmentally friendly. From chatbots to self-driving cars, GPUs are the unsung heroes making our AI dreams come true. They've evolved from gaming sidekicks into the backbone of the AI revolution. As we look to a future where AI is part of everyday life, one thing's for sure – GPUs will be right there, powering the next big breakthrough.

So the next time you interact with a surprisingly human-like AI or see a computer-generated image that looks just like a photograph, remember: there's probably a hardworking GPU (or a whole bunch of them) behind the scenes, crunching the numbers to make it all possible. The AI future is here, and it's being built on a foundation of GPU power.



Leveraging AI to Revolutionize Sales and Marketing in Pharma

An AI Guide for Medical Representatives

Animesh Patel

In today's fast-paced and competitive pharmaceutical industry, sales and marketing teams face numerous challenges in effectively engaging with healthcare professionals (HCPs) and maximizing the impact of their efforts. However, with the advent of artificial intelligence (AI), these challenges are being met with innovative solutions that are transforming the way sales and marketing strategies are developed and executed.

Developing and Assessing marketing activation Plans for Top Customers

Developing and Assessing marketing activation Plans for Top Customers: Traditionally, developing business plans such as conducting CMEs (continuous medical education programs), RTMs (round table meetings) & medical conferences for top HCPs has been a time-consuming and resource-intensive task for pharmaceutical companies. However, with AI-powered analytics, medical representatives can now quickly assess HCPs data, identify trends, and develop customized CME/RTM & conference plans tailored to each customer's specific needs, preferences and knowledge gaps. By leveraging AI algorithms, sales teams can analyse vast amounts of data to gain valuable insights into HCPs behaviour, prescribing patterns, and market dynamics, enabling them to make informed decisions and develop strategies that drive sales growth and customer satisfaction. Sanofi Ukraine implementing similar solutions in the training domain, essentially empowering the HCPs and MRs.

Reviewing and Refining Call Plans & routes planning

Call plans generally made by medical representative himself. It is primarily based on his own convenience and not by any objective methods. This call-plan many times fail to complete the target of meeting certain number of doctors (TCFA) leading to missed calls & corridor calls because of unavailability of HCP at that time and also due to inefficient route planning. Planning efficient call routes for MRs is essential for maximizing their productivity and optimizing their time in the field.

AI-powered tools can analyse historical call data, customer interactions, market trends, customer locations, traffic patterns, and scheduling constraints to generate personalized call plans and optimal call routes for medical representatives that minimize travel time and maximize the number of customer visits reduce missed calls improve TCFA & HCP coverage in less field working days. These plans can include recommendations on the most effective communication channels and timing of calls. By leveraging AI, sales teams can ensure that their interactions with healthcare professionals are highly targeted and impactful, leading to increased engagement and sales success.

By automating the route planning process, AI enables medical representatives to focus more on engaging with customers and less on logistical challenges, ultimately leading to greater efficiency and effectiveness in their sales efforts. Salesforce already have a software where it does the route planning based on AI.

Profiling HCP - Customize Communication

Field force interaction with HCPs can often be generic and not fully aligned with the needs of individual HCP. Developing a personalized HCP engagement approach is critical for building strong relationships and driving sales success in the pharmaceutical industry.

AI-driven recommendation engines, customer segmentation and targeting tools can analyse HCP data such as specialty, prescribing behaviour, and engagement preferences and historical interaction patterns to generate & identify key segments such as key opinion leaders (KOLs) & key business leaders (KBLs), personas and personalized suggestions for MRs.

These suggestions can include tailored engagement approach, key talking points and specific messaging, product information, and promotional offers based on each HCP's preferences and interests leading to more meaningful interactions and increased sales opportunities. By providing MRs with targeted recommendations, AI empowers them to deliver more relevant and impactful messages to HCPs, leading to increased engagement and better return on time invested.

For example, a comprehensive profile of each HCP can be made based on customer feedback on websites like Practo etc. by which MR will know what are pain points of HCP and try to engage him/her on these points. Many consulting companies already provides such solutions. IQVIA's AI-powered HCP & KOL identification is one such tool amongst many.

Training & HR development

Effective product detailing is essential for conveying the value proposition of pharmaceutical products to HCPs. Hence Essential product knowledge (EPK) and Essential selling skills (ESS) becomes very important.

AI-powered sales enablement tools can provide medical representatives with access to up-to-date product information, clinical data, and educational resources to support their discussions with HCPs also such tools can tailor training programs to the specific needs of each representative, using data analytics to identify knowledge gaps and learning styles.

"Medical representatives bridge the gap between drug manufacturers & healthcare professionals, promoting medication effectively, ensuring their accessibility, and also helping in market research."

By equipping medical representatives with the knowledge and resources they need, AI enables them to deliver compelling and persuasive product presentations that resonate with healthcare professionals, leading to increased product adoption and sales. Furthermore, AI-powered simulations and virtual reality can provide immersive, hands-on experiences that are critical for mastering product demonstrations and patient/HCPs interactions. There are companies which are providing certification programs with AI powered training.

Improving Sales force Efficiency, Productivity & Automation

In a sales rep's daily routine, administrative tasks such as tracking and logging call activities, essential for monitoring sales performance and compliance in the pharmaceutical industry, often consume a significant amount of time. AI-driven sales reporting and analytics platforms can automate the process of logging call activities and generating comprehensive sales reports for analysis. Other such pharmaceutical solutions streamline tasks such as appointment scheduling, expense reporting etc.

This automation liberates sales reps to allocate more time to customer-facing activities. A PwC study found that automating these tasks can free up as much as 40% of a sales rep's time. AI's capacity to interpret and input data directly into CRM systems significantly reduces the time spent on manual data entry and reporting. This enhancement enables sales representatives to focus more on engaging with healthcare providers and less on paperwork. Companies leveraging AI for data entry report a 30% increase in sales productivity, according to data from McKinsey & Company.

Compliance

After every CME/RTMs, MR generally share the details of participation of HCPs with the compliance team of the company and compliance manager verify the signature and names of each HCPs so as to stop the leakage of funds and misutilisation of funds assigned for CME/RTMs. This verification process is very tardy and time taking process and one dedicated individual is needed. With the advent of AI these verification process can be automated to a large extent as signature recognition can be perfected by AI. In fact Microsoft already provides one such software namely Cogniware which identifies the authenticity of signature.

"Great Salespeople are relationship builder who provide value and help their customers win."
- Jeffrey Gitomer

Coordinating and meetings with Peers, office & hospital staff

Collaboration with peers, such as mirrored representatives or territory sales executive who takes care of supply side of the product, Medical Science Liaisons (MSLs), office staff responsible for administrative tasks and patient support services, is essential for maximizing the impact of sales and marketing efforts in the pharmaceutical industry.

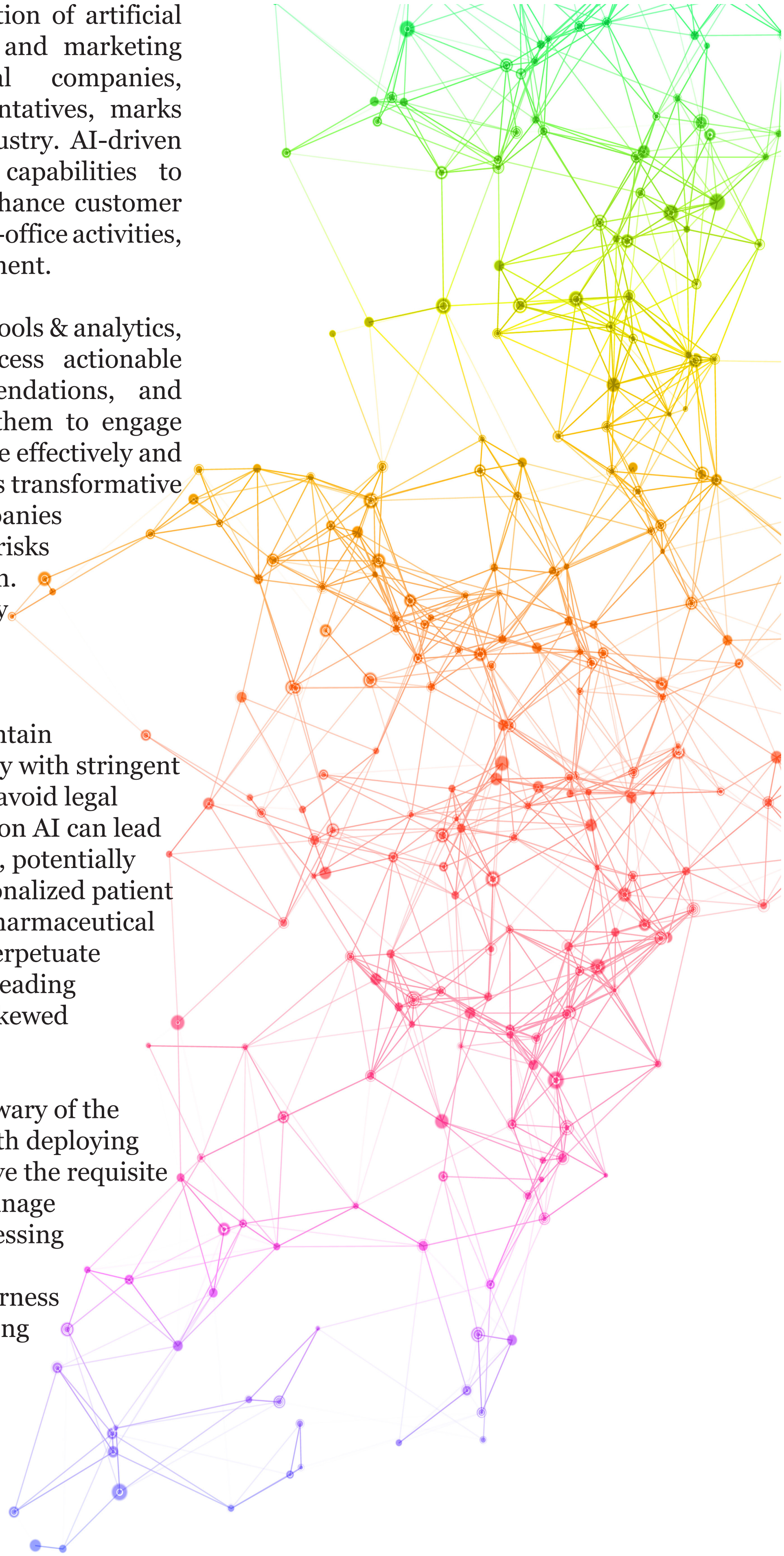
AI-powered collaboration platforms can facilitate communication and information sharing among team members, provide MRs with insights into office staff roles and responsibilities, enable them to tailor their interactions, coordinate their activities more effectively and leverage each other's expertise and insights to address specific needs and concerns. By fostering collaboration and knowledge sharing, AI empowers sales teams to work together towards common goals and drive collective success.

In conclusion, the integration of artificial intelligence (AI) into the sales and marketing operations of pharmaceutical companies, particularly for medical representatives, marks a transformative shift in the industry. AI-driven solutions offer unprecedented capabilities to streamline business planning, enhance customer visit planning, optimize call and in-office activities, and improve post-call engagement.

By leveraging AI-powered tools & analytics, medical representatives can access actionable insights, personalized recommendations, and efficient workflows that enable them to engage with healthcare professionals more effectively and drive sales success. While AI offers transformative potential, pharmaceutical companies must exercise caution to mitigate risks associated with its implementation. Ensuring data privacy and security is paramount, as AI systems often handle sensitive information.

There is also a need to maintain regulatory compliance, particularly with stringent healthcare industry standards, to avoid legal and ethical pitfalls. Over-reliance on AI can lead to a reduction in human oversight, potentially compromising the quality of personalized patient interactions that are vital in the pharmaceutical field. Moreover, AI systems can perpetuate biases if not properly monitored, leading to unequal training outcomes or skewed performance assessments.

Companies should also be wary of the cost and complexity associated with deploying AI technologies, ensuring they have the requisite infrastructure and expertise to manage these systems effectively. By addressing these challenges proactively, pharmaceutical companies can harness the benefits of AI while safeguarding against its potential drawbacks.



AI Governance

Himanshu Singh

In 2019, Apple found itself at the center of a public relations storm when its newly launched Apple Card was accused of gender discrimination. The AI algorithm determining credit limits allegedly offered significantly lower credit to women compared to men, even when they shared similar financial profiles. This incident ignited a heated debate about algorithmic bias and the lack of transparency in AI-driven financial decisions.

As artificial intelligence increasingly influences critical aspects of our lives, from credit worthiness to loan approvals, the Apple Card controversy underscores the urgent need for responsible AI development and deployment. Without comprehensive AI governance, we risk embedding and amplifying societal biases into automated systems, potentially creating a new frontier of digital discrimination, further fuelling the ever-increasing inequality.

What is AI Governance?

AI governance refers to the frameworks, policies, and practices designed to oversee the development and deployment of AI technologies. It involves setting standards and guidelines to ensure that AI systems are transparent, accountable, and ethical. Effective AI governance encompasses a range of activities, including regulatory compliance, risk management, ethical oversight, and promoting transparency and accountability.

These governance structures are essential for ensuring that AI systems adhere to relevant laws and regulations, identify and mitigate potential risks, align with ethical principles and societal values, and remain understandable to users and stakeholders. By implementing robust AI governance, we can hold developers accountable for the impacts of their creations and foster trust in these powerful technologies.

Principles of AI Governance

To address the complex challenges posed by AI, several core principles guide effective AI governance. These principles include fairness, transparency, accountability, privacy, safety and security, and inclusivity.

Fairness in AI systems means designing them to avoid bias and discrimination, ensuring that training data is representative and algorithms do not perpetuate existing inequalities. Transparency requires clear explanations of how AI decisions are made and what data informs those decisions. Accountability involves establishing mechanisms for auditing and reviewing AI systems, holding developers and operators responsible for their actions and outcomes.

Privacy protection is crucial, with robust data protection measures and respect for user consent. Safety and security principles focus on making AI systems resilient against malicious attacks and failures. Lastly, inclusivity ensures that AI development and use benefit all segments of society, addressing diverse community needs and preventing the widening of the digital divide.

Challenges of AI Governance

Despite the clear need for AI governance, implementing effective frameworks is fraught with challenges. The rapid pace of technological advancement often outstrips regulatory development, making it difficult to keep policies updated and relevant. Global coordination is essential, yet complicated by differing cultural values and regulatory approaches across borders.

The inherent complexity of AI systems poses another significant hurdle, as policymakers and the public struggle to fully grasp their workings and implications. This complexity impedes the development of effective and comprehensible governance frameworks. Additionally, striking the right balance between promoting innovation and ensuring safety and ethical compliance is a delicate task. Overly restrictive regulations could stifle AI development, while insufficient oversight could lead to harmful outcomes.

Navigating complex ethical dilemmas, such

UK's AI Summit developers and govts agree on testing to help manage risks as trade-offs between privacy and utility or the allocation of decision-making power between humans and machines, further complicates the governance landscape.

Global and National Initiatives in AI Governance

In response to these challenges, various global and national initiatives have emerged to develop comprehensive AI governance frameworks. The European Union's AI Act stands out as a pioneering regulatory framework, categorizing AI applications into different risk levels and imposing stricter requirements on higher-risk systems.

The **World Economic Forum's** AI Governance Alliance brings together stakeholders from various sectors to collaboratively address AI governance challenges, promoting responsible development and leveraging diverse expertise.

India's National Strategy for Artificial Intelligence, developed by **NITI Aayog**, outlines a strategic approach emphasizing inclusive AI technologies that benefit society as a whole. The strategy advocates for ethical AI practices while addressing the unique challenges of AI deployment in India.

The **United Nations** has also contributed to the global dialogue with its White Paper on AI Governance, calling for a unified global approach to address the ethical, social, and economic impacts of AI technologies.

In the private sector, initiatives like **IBM's Watsonx** Governance platform provide tools to help organizations comply with AI regulations and manage AI systems transparently, including features for tracking regulatory changes and mapping them to internal risk data.

Conclusion

As we reflect on incidents like the Apple Card controversy and consider the pervasive influence of AI in our daily lives, the importance of robust AI governance becomes increasingly clear. It is not merely a regulatory necessity but a crucial framework for ensuring that AI technologies develop in ways that are safe, ethical, and beneficial to all.

By learning from past failures, adhering to core governance principles, and fostering global cooperation, we can navigate the complexities of AI and harness its transformative potential responsibly. As we move forward, it is imperative that we continue to refine and strengthen our approach to AI governance, ensuring that these powerful technologies serve to enhance, rather than undermine, our collective well-being and societal values.

Behind the Scenes . . .

How AI is Transforming the Entertainment Industry

Deept Panchal

Picture yourself stepping onto a bustling film set, the air thick with the excitement of an impending blockbuster. Cameras are rolling, directors are barking out last-minute instructions, and actors are slipping into their roles. Amidst this whirlwind of creativity and chaos, a new kind of magic is at play—Artificial Intelligence (AI). This isn't a glimpse into a distant future or a scene from a sci-fi movie; it's happening right now, transforming the entertainment industry in ways we are just beginning to uncover.

As we navigate this space, it's impossible not to feel the rush of technological progress. Remember how it took humanity thousands of years to transition from writing to the printing press, and only a few centuries more to get to email? Now, AI is advancing at breakneck speed, promising to revolutionize industries faster than ever before. But what does AI really mean for entertainment? Is it Pandora's box, filled with both hope and potential peril?

AI is already making waves in film and television production, transforming everything from script analysis to casting, and even predicting box office success. In animation, AI breathes life into characters and special effects, creating stunning visuals that were once the domain of human animators laboring for months on end. Gaming, too, is experiencing an AI-led metamorphosis, with smarter NPCs and more immersive worlds than ever before.

Yet, with all this innovation come profound ethical questions. Can AI truly replicate human creativity and emotion, or will it always fall short? What happens when AI-generated content becomes indistinguishable from human-created works? And as Hollywood grapples with these changes, how will the industry balance technological advancement with the protection of creative rights?

Join us as we delve into these questions and more, exploring the myriad ways AI is reshaping the entertainment industry. From behind-the-scenes innovations to on-screen magic, we'll uncover how AI is not just a tool but a transformative force, poised to redefine the boundaries between technology and creativity. Welcome to the dawn of a new era in entertainment, where AI stands at the forefront of change.

AI in Film and Television Production

AI-Driven Script Analysis and Selection

"A tale of co-creation between man and machine"

In the age of AI, the process of script analysis and selection has been revolutionized. Traditionally, script analysis involved countless hours of reading and subjective judgment.

However, AI algorithms can now analyze hundreds of scripts in a fraction of the time, identifying themes, character development, and potential plot holes. For instance, the AI system developed by ScriptBook has been used to analyze scripts and predict their box office success. ScriptBook's AI examined the script for the movie "Passengers" and provided insights into character depth and plot dynamics, helping the filmmakers fine-tune the narrative to enhance its appeal.

" The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions. "

– Marvin Minsky

AI's Role in Post-Production

Post-production is a crucial phase where AI is making significant strides. AI-powered tools are now used for editing, color correction, and sound design, streamlining processes that once took weeks or months. Adobe's Sensei AI, for instance, automates repetitive editing tasks, allowing editors to focus on the creative aspects. The movie "Avengers: Endgame" benefited from AI in post-production, with AI algorithms used for advanced color grading and enhancing visual effects, ensuring a polished final product.

#WhatIsMorgan?

Case Studies of Successful AI Applications

Several recent productions have successfully integrated AI, showcasing its transformative potential. The film "Morgan," for instance, utilized AI to create its trailer. IBM's Watson analyzed hundreds of horror movie trailers to understand what makes them effective, then used this data to create a compelling trailer for the film. Another example is the use of AI in Disney's "The Lion King" (2019), where AI-driven photo-real rendering brought the animated characters to life, contributing to the film's visual success.

AI in Animation and Special Effects

Streamlining the Animation Process

AI is streamlining the animation process, making it more efficient and cost-effective. Traditional animation requires meticulous frame-by-frame work, but AI can automate much of this. For example, Autodesk's AI tools assist animators by predicting and generating in-between frames, speeding up the animation process. The movie "Spider-Man: Into the Spider-Verse" used AI-assisted animation techniques to create its unique visual style, significantly reducing production time while enhancing the film's aesthetic appeal.

Creating More Realistic and Complex Special Effects

AI is also transforming the realm of special effects (VFX), enabling the creation of more realistic and complex visuals. AI algorithms can analyze real-world physics and apply them to digital objects, making movements and interactions appear more natural. Deep learning techniques allow AI to enhance visual effects, such as generating realistic weather conditions or simulating intricate particle effects. The film "Avatar" utilized AI to create highly realistic environments and characters, setting a new standard for special effects in cinema.

AI-Powered Video Making Tools

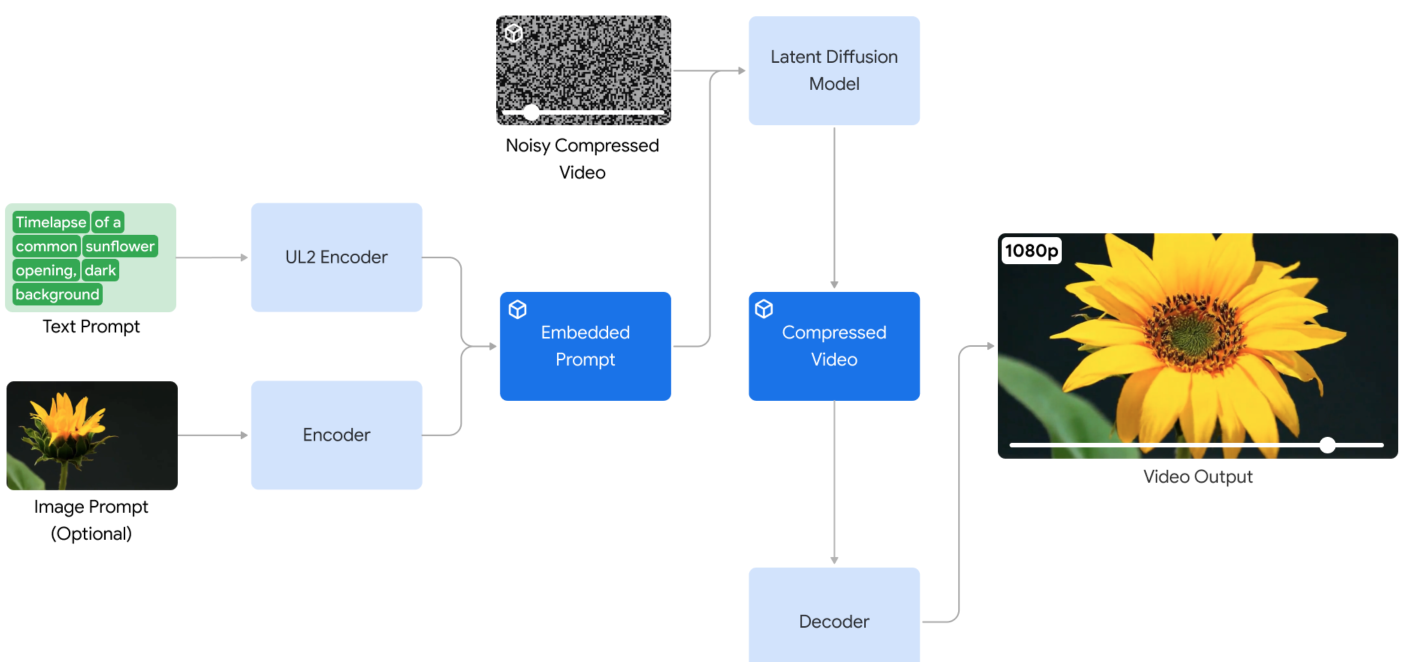
OpenAI Sora: OpenAI Sora is another groundbreaking tool in the realm of AI-driven video production. Sora can generate detailed storyboards from a script, complete with scene descriptions and camera angles. During the production of the animated film "Next Gen," Sora was utilized to create initial storyboards, allowing the directors to visualize scenes early in the production process. This not only streamlined the workflow but also provided a clear blueprint for the animation team to follow.

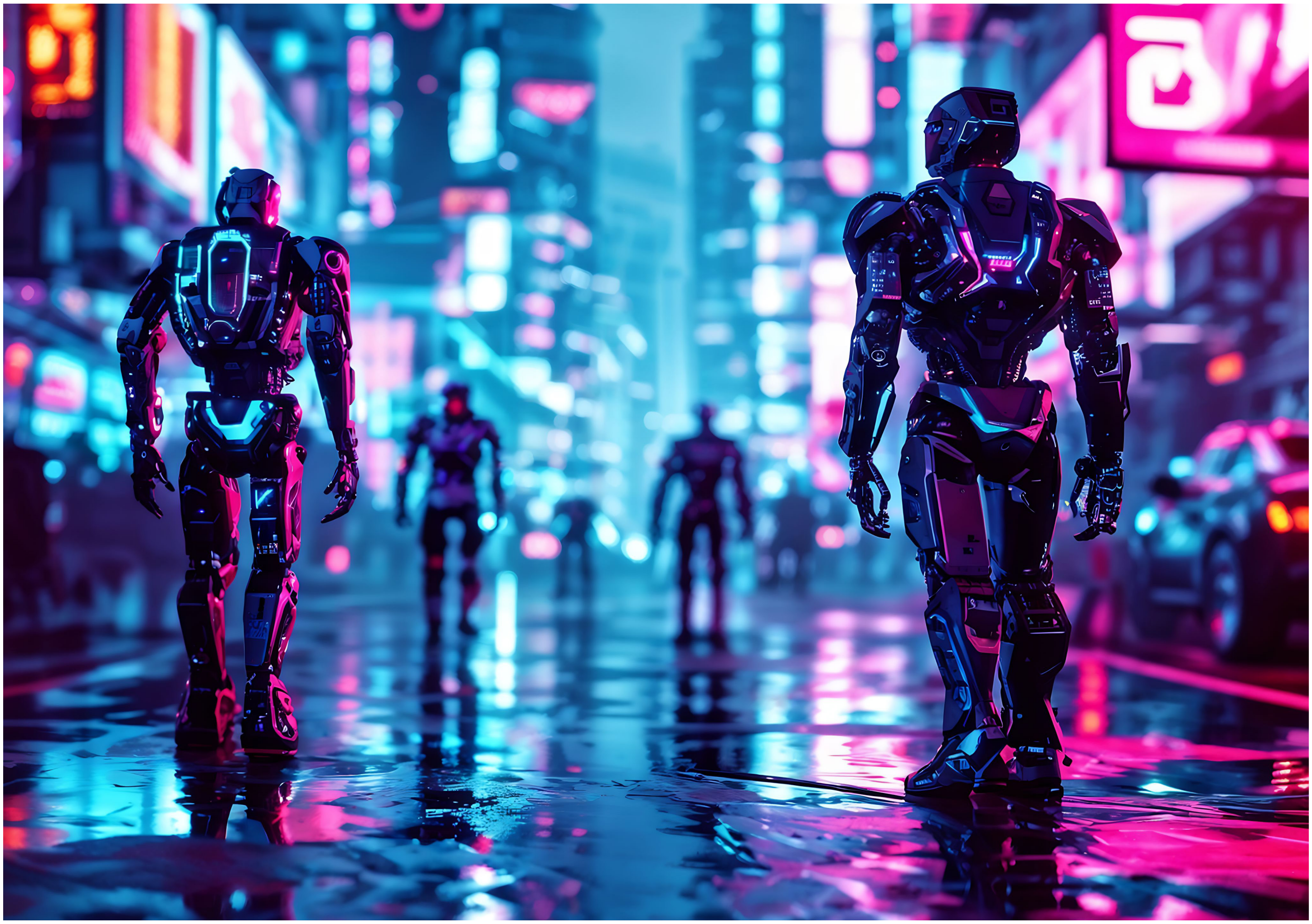
OpenAI Sora Prompt: "A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, ..."

Google Veo: Google Veo is an AI-powered video editing tool that enhances video production through advanced machine learning algorithms. Veo can analyze hours of raw footage, identify the most compelling moments, and automatically create highlight reels.

Google Veo's architecture is designed to generate high-quality videos from text prompts and optional image inputs. The process begins with users providing a detailed text description, such as "Timelapse of a common sunflower opening, dark background," and optionally an image to guide the video's style. The text is processed by the UL2 Encoder, while the image, if provided, is processed by the Image Encoder. These encodings are combined into an embedded prompt that serves as a compact and detailed representation of the input. This embedded prompt is then fed into a Latent Diffusion Model, which generates a noisy, compressed version of the video.

This model ensures visual consistency across frames, maintaining smooth and realistic video output. The compressed video is then refined through further processing and decoded into a high-definition (1080p) final video. Veo builds upon years of generative video model work, including Generative Query Network (GQN), DVD-GAN, Imagen-Video, Phenaki, WALT, VideoPoet, and Lumiere, as well as Google's Transformer architecture and Gemini.





Generative AI in Gaming: A New Era

Generative AI is making waves in the gaming industry, poised to transform everything from game development processes to player experiences. This was especially evident at the 2024 Game Developers Conference (GDC), where companies like Nvidia, Ubisoft, and Microsoft showcased their latest advancements in AI technology.

Hidden Door's Narrative Games

Hidden Door, a studio developing narrative-driven games, also demonstrated how AI could enable new types of gameplay experiences. Their game allows players to explore genre-based adventures that evolve in real-time, creating dynamic storylines and characters. This offers a glimpse into the future of interactive storytelling powered by AI.

Project AVA

One notable experiment at GDC was Project AVA by Keywords Studios, a sci-fi civilization-building game that showcased both the potential and limitations of generative AI. While AI tools were useful for generating static images and assisting with coding, they struggled with more complex tasks like creating user interfaces and fixing bugs. The project concluded that AI could not yet replace human creativity and expertise but could augment and streamline certain aspects of game development.

> "Generative AI is the most powerful tool for creativity that has ever been created. It has the potential to unleash a new era of human innovation. "
>
> \- Elon Musk

Ethical Considerations & Challenges in AI-Driven Entertainment

As generative AI continues to permeate various sectors, the entertainment industry is grappling with a host of ethical considerations and challenges. From deepfakes to the potential replacement of human talent, the implications of AI in entertainment are profound and complex.

Hollywood Writers' Battle Against AI

In Hollywood, the rise of AI has sparked significant concerns among writers. The Writers Guild of America (WGA) has been vocal about the potential threat AI poses to their profession. Writers fear that studios might use AI to generate scripts or storylines, potentially diminishing the demand for human creativity and reducing job opportunities. To combat this, the WGA has been advocating for contracts that explicitly limit the use of AI in writing. it should not replace the nuanced and emotive work that human writers bring to the table. Ensuring that AI serves as a tool rather than a replacement is a key aspect of their fight.

Hollywood Actors Secure Safeguards Around AI Use On Screen

Actors, too, are seeking protections against the encroachment of AI in their field. The use of AI to create digital replicas or to enhance performances raises concerns about consent and compensation. Actors want to ensure that their likeness and performances cannot be used without their explicit permission and that they are fairly compensated for any AI-generated content that uses their image or voice.

In recent negotiations, actors have pushed for clauses in their contracts that safeguard their rights in the digital age. These clauses often include provisions that prohibit the use of AI to replicate an actor's performance without consent and guarantee compensation if such technology is employed. By securing these safeguards, actors aim to protect their livelihoods and maintain control over their professional image.

Source: The Guardian (How Hollywood writers triumphed over AI – and why it matters)

Conclusion: Embracing the AI-Driven Future of Entertainment

As the sun sets on a bustling film set, the director calls for a wrap, and the actors retire to their trailers. The day's work, infused with the magic of AI, hints at a future where technology and creativity intertwine more deeply than ever before. The excitement and anxiety swirling through Hollywood, from film studios to animation houses, and even within the gaming industry, underscore the transformative potential and challenges of this new era.

Generative AI has already started to reshape how stories are told and how worlds are built. Script analysis tools now predict audience reactions, casting decisions are bolstered by data-driven insights, and post-production work is more efficient than ever. AI breathes life into digital characters, making animations more vibrant and visual effects more stunning. In gaming, NPCs are becoming more lifelike, enhancing player immersion, while development processes are becoming streamlined and efficient.

However, the ethical implications of these advancements cannot be ignored. The rise of deepfakes, the potential for job displacement, and the questions around authorship and consent highlight the need for careful consideration and regulation. Hollywood writers and actors are on the frontlines of this battle, fighting to ensure that AI serves as a complement rather than a replacement to human creativity.

Writers, through the efforts of the Writers Guild of America, are pushing for contracts that limit AI's role in scriptwriting, ensuring that the human touch remains central to storytelling. Actors are securing safeguards to protect their likenesses and performances, ensuring they are fairly compensated and their rights are respected.

As we stand at the dawn of this new era, it's clear that AI holds immense promise for the entertainment industry. It can enhance creativity, improve efficiency, and create more immersive experiences for audiences.

Yet, it also requires a balanced approach, where the benefits of technology are harnessed without compromising the rights and livelihoods of those who bring stories to life. The future of entertainment, illuminated by the glow of AI, is one where technology and human creativity coexist. It's a future where the boundaries between what's real and what's generated blur, creating new possibilities for storytelling. As we navigate this exciting landscape, it's crucial to remember that while AI can enhance our creative endeavors, the heart of entertainment will always be the human spirit that drives it.

So, as the cameras stop rolling and the set grows quiet, we look forward to a future where AI continues to inspire, innovate, and revolutionize the way we tell stories, always with the human touch at its core. Welcome to the new era of entertainment—where the magic of AI meets the timeless art of storytelling.

Dr. Sujoy Kar

Dr. Sujoy Kar, Chief Medical Information Officer & Vice President Apollo Hospitals, have been associated with the hospitas for 19 years now. For the past seven years he has been working in the Clinical AI Lab at the Apollo Hospitals to design, develop and deploy AI based solutions in clinical workstreams. His stint as Clinical Microbiology and Infectious Diseases MD, expertise gained from Six-Sigma Master Black Belt and an Executive PGM from MIT Sloan School have helped him to reach the heights he is currently at.
LinkedIn: linkedin.com/in/drsujoykar/

AINA: With your extensive leadership background in healthcare, how do you define the importance of analytics in the healthcare sector?

Analytics in healthcare is crucial, perhaps now more than ever. Traditional healthcare metrics like KPIs focused on hospital operations—admissions, discharges, transfers, length of stay, and complications from procedures are undoubtedly important. However, analytics goes beyond merely collecting data for these reporting and analysis. It is providing more and more deep understanding of the processes underlying these indicators.

For example, instead of just looking at overall hospital length of stay, analytics allows us to examine the throughput for specific specialties within the hospital, assessing specialty-based admissions, discharges, and other performance indicators.

Today's analytics digs deeper, moving from a simple 'What is happening?' to 'Why is it happening?' This transition is vital for developing thorough and credible insights, which are indispensable for making informed decisions, especially in healthcare. Understanding the 'why' behind data trends not only informs better decision-making but also enhances the overall quality of care provided.

AINA: With the established importance of analytics in healthcare, how do you ensure that your leadership understands and supports the integration of analytics in decision-making processes?

At Apollo Hospitals, we've fostered a culture deeply rooted in data analytics and understanding of the key performance indicators (KPIs). These KPIs have been rigorously monitored monthly across all Apollo units for at least the last 20 years, guiding clinical and operational decisions as well as strategic growth.

The practice at Apollo Hospitals exemplifies a broader industry trend where healthcare is increasingly driven by these

these critical indicators. Different hospitals may prioritize different KPIs based on their specific services, specialties, and strategic goals. Choosing the right KPIs involves a comprehensive understanding of the organization's priorities, which could be costs, patient safety, public health, and other specialty-specific needs.

Choosing the right KPIs involves a comprehensive understanding of the organization's priorities, which could be costs, patient safety, public health, and other specialty-specific needs.

And then there is a mix of 'bottom-up ' and 'top-down' methodologies: data is gathered and analyzed at the operational level, insights are then presented to the board, which discusses necessary actions and strategic directions. This dual approach ensures that decisions are both data-driven and aligned with our overarching goals, making analytics a cornerstone of our decision-making processes at Apollo.

AINA: Can you provide an example of the significant impact that analytics has had on the functioning of Apollo Hospitals?

Absolutely, I'll reference my first project during my Black Belt certification in 2010, which focused on reducing the hospital's average length of stay, which at the time exceeded five days in our Kolkata Hospital. We faced challenges in accommodating patients, prompting us to utilize tools from Lean Six Sigma, such as value stream mapping. This process involved analyzing the entire patient journey, including each activity's duration, manpower, and supply chain resources.

Our objective was to identify and eliminate non-value-added activities, both clinical and operational. By streamlining these processes, particularly around discharge timings and mobilizing support from consultants and nurses, we successfully reduced the average length of stay by about one day. This reduction significantly increased hospital throughput, allowing for more patients to receive care across various specialties.

Initially, there were concerns that reducing the length of stay might negatively impact financial performance, as an empty bed doesn't generate revenue. However, the outcome was quite the opposite. By increasing the number of treated patients, we not only improved accessibility but also passed on cost savings to our patients. This shift in approach underlines the importance of process and outcome efficiencies, ultimately enhancing clinical outcomes and generating incremental revenue for the hospital.

AINA: Could you discuss how you have adapted and evolved your analytics strategies over time at Apollo Hospitals?

Analytics and AI are often viewed as distinct, yet I see them as a continuum of evolving tools. Innovation lies at the heart of our analytics journey, underpinned by sustainable, data-driven experimentation. To me, real innovation isn't fleeting—it sticks around as a part of our analytics maturity, manifesting as incremental and breakthrough innovations.

Initially, we focused on incremental innovation, such as reducing hospital length of stay to decrease healthcare-associated infections while monitoring for potentially premature discharges. We then expanded our metrics to include readmission rates to ensure that reductions in stay did not compromise patient care post-discharge. My work on reducing readmissions formed the basis of my thesis for my Master Black Belt, where we managed to lower readmission rates by 3-4%, which is notable compared to global standards.

As we mastered these incremental changes, we began to explore breakthrough innovations using the same datasets. By applying machine learning and optimization theories, we examined capacity constraints and patient outcomes more deeply, determining optimal discharge times for specific specialties or clinical conditions. This approach not only supports operational efficiency but also ensures that each innovation contributes to sig-

-nificant cost savings, revenue generation, and enhanced clinical outcomes, proving the enduring value of our evolving analytics strategies."

AINA: Can you provide examples of instances where you had to pivot your strategy due to challenges faced?

Absolutely, strategic flexibility is crucial in healthcare. A prime example is our introduction of robotic surgery in India over a decade ago. Initially, there was significant hesitation around its adoption due to unfamiliarity and concerns about its accuracy and the required skills. Patients were apprehensive, imagining a scene from a science fiction film rather than a medical procedure.

To address these challenges, we had to reassess our strategy. We focused on training staff and educating patients about the benefits of robotic surgery, such as reduced hospital stays—for example, from four days to just 48 hours—and quicker recovery times, enabling patients to return to work in about seven days instead of fifteen. We also highlighted the substantial reductions in blood loss and infection rates.

It's about making technology relatable and beneficial while considering human behaviour and experience to ensure that analytics truly enhances decision-making."

AINA: In healthcare, it is tough to make different systems and data sources to work together, smoothly. So, how do you handle combining data from various places to get a complete picture of patient health and hospital operations?

That's a great question. In addressing healthcare data collection from a global perspective, we must recognize the significant variability across electronic medical records (EMRs). Healthcare data falls into four main categories: transactional and operational data, which show minimal global variation; clinical interactions and records, which often capture crucial data in unstructured textual formats; imaging and signal data, including X-rays, ECGs, and EEGs, which document physiological signals and images; and high-velocity, high-variability, and high-volume data such as genomics and wearable device data. These are typically collected electronically and form essential parts of the healthcare data ecosystem.

Healthcare data falls into four main categories: transactional and operational data; clinical interactions and records; imaging and signal data; and high-velocity, high-variability, and high-volume data.

Key technologies like Named Entity Recognition (NER) and Clinical Entity Disambiguation techniques are vital for linking unstructured text to specific medical conditions, such as connecting "fever and cough" to pneumonia. This linking extends to associating clinical findings with laboratory results, thus structuring the data effectively.

Additionally, sentiment analysis and bidirectional text analysis are used to interpret and structure symptom severity from clinical texts. Moreover, relation extraction techniques help build correlations coefficients between clinical terms, facilitating the construction of comprehensive Knowledge Graphs that can answer specific clinical queries, such as the effects of antibiotic treatments. Ensuring data privacy by anonymizing personal identifiers (PII) is crucial, allowing the data to be used safely for applications ranging from machine learning to Generative AI.

AINA: What are the common challenges you have faced in implementing analytics solutions in a healthcare environment, and how have you overcome them?

The first and the foremost challenge in healthcare data analytics is bias, originating from various sources such as methods of data collection, measurement devices, and inherent societal biases (gender, ethnic, geographic). These biases can distort data interpretation and outcomes. For instance, discrepancies in measurements from different devices can lead to signal bias, while selection bias in research might exclude important cohorts, affecting study results. Implementing fair learning practices and demographic parity is crucial to counteract these biases.

Ensuring the safety of analytical solutions is critical. The principle of non-maleficence requires that analytics are effective, accurate, fair, and harmless, underpinned by robust frameworks for Responsible AI. Moreover, the interoperability and universal applicability of analytics present challenges. Insights and AI technologies must be adaptable across diverse settings to bring in real world evidence. Human-machine interaction also adds complexity, with potential conflicts arising when AI insights conflict with human judgment, questioning decision-making authority. Lastly, the explainability and trustworthiness of technologies are essential. Systems must be understandable and trustworthy for practitioners to confidently relay information to patients, thereby creating a pipeline of trust. {https://link.springer.com/article/10.1007/s40012-023-00381-2}. These challenges underscore the need for ethically aligned and practically viable technical solutions in data analytics.

Systems must be understandable and trustworthy for practitioners to confidently relay information to patients, thereby creating a pipeline of trust.

AINA: In situations where AI predicts a disease with a certain probability, but a consulting physician disagrees based on their clinical experience, how should the discrepancy between the AI's recommendation and the doctor's judgment be handled?

The integration of AI in healthcare brings about significant ethical and regulatory challenges. Worldover, there have been instances where doctors faced consequences for not utilizing AI in detecting tumours, highlighting the push towards integrating technology in medical diagnostics. Conversely, excessive reliance on AI has led to overdiagnoses, underscoring the technology's limitations and the importance of cautious application. The critical issue lies in managing the balance between true negatives and false negatives and establishing appropriate thresholds for AI algorithms.

These approaches advocate for a firm shift from a binary deterministic view of disease states—where conditions are either present or absent—to a probabilistic understanding, which accommodates the complexities of real-world conditions. Such a nuanced perspective is essential, particularly in managing conditions like dense breast tissue, which could potentially develop into breast cancer. Regular monitoring, rather than immediate intervention, becomes the recommended course of action, leveraging AI to optimize screening processes and enhance patient follow-ups. This strategy not only improves patient care but also incorporates a more sophisticated understanding of risk in healthcare settings.

AINA: How do you ensure data security and compliance with healthcare regulations in your analytics work?

Data security and confidentiality in healthcare require robust systems due to stringent regulations like GDPR, HIPAA, HITRUST, DPDP (Digital Personal Data Protection Act 2024, India) and other compliance laws. These regulations mandate that personal health data must be handled with strict permissions, anonymization, and appropriate safeguards.

The evolving landscape of healthcare data now emphasizes encrypted environments and confidential computing, allowing for data processing without data leaving the premises of the healthcare institution. This approach is part of a model known as federated or distributed learning, which incorporates privacy-preserving technologies.

Additionally, the daily operation and maintenance of these data systems necessitate advanced security measures across the organization, including firewalls and encryption systems, to ensure data integrity and protection. On the top of these, there are multiple certification systems for information security, like ISO 27001.

AINA: How does the Indian healthcare system compare to those in developed economies, and what are some unique challenges it faces that might not be present in other countries?

India boasts excellent examples of healthcare practices that meet international standards, particularly in tertiary care where outcomes for specific surgeries are as effective as anywhere globally. This includes not only top private hospitals but also esteemed public institutions.

However, secondary care often takes place in smaller, less-documented settings like nursing homes and district hospitals, where outcome data is sparse and difficult to assess. Primary care, especially in rural and semi-urban areas, needs significant attention due to gaps in accessibility, affordability, and the credentials of healthcare providers. While the central and state governments have made strides in addressing these issues, there is still much work to be done.

Digital health companies are increasingly focusing on improving primary healthcare by providing solutions that enhance service delivery. Governance and integration are crucial to ensuring that these efforts are effective, and that data is shared securely, aligning with initiatives like the Ayushman Bharat Digital Mission.

Overall, while India's healthcare system has strong points, particularly in tertiary care attracting medical tourism, secondary and primary care sectors present challenges and opportunities for improvement. The enthusiasm from public health professionals and institutions gives hope for substantial progress in these areas.

AINA: Comparing healthcare analytics in India with abroad, I have an assumption that hospitals and colleges abroad are more advanced and collaborative in data science analytics compared to India. Can you provide some insights on this?

You are 'the brain', mark my words you are 'the brain', and you should recognize the potential of the PGDBA program. Since its inception, the alumni of your course have been doing fantastic work, which I've followed on LinkedIn. This kind of program is needed across many institutions in the country.

You are 'the brain', mark my words you are 'the brain', and you should recognize the potential of the PGDBA program.

While some are replicating your systems, which is positive, you will still remain the pioneers. There are competitive healthcare analytics courses emerging now, such as those at John Hopkins, Stanford and MIT Sloan School of Management. Your course is comparable with these and in many ways, better.

AINA: How do you ensure collaboration between medical officers and data scientists, considering there must be communication challenges within cross-functional teams?

I refer to the concept of X-teams from Professor Deborah Ancona at MIT Sloan, which involves people from diverse backgrounds solving specific problems. This ensures diversity of thought and brings out hidden perspectives when building solutions. To address a clinical or operational problem, data is gathered and then converted into a mathematical or statistical problem by data scientists. The mathematical solution then needs to be converted back into a clinical solution that is implementable. Therefore, cross-functional, multidisciplinary teams are essential for any organisationisation, and therefore Data scientists and clinical people must learn to work together to create practical solutions.

AINA: What do you think are the most significant trends or developments in healthcare analytics in the next 3 to 5 years?

From a directional perspective, we are on the right track by recognizing the value of data in every industry. We have figured out how to collect data, but now the challenge lies in computation and collaboration. We need to exchange data while maintaining privacy and using federated learning and cloud computing to build better solutions. In doing so, we must also address the cost, as it can be a barrier for startups and established companies alike. Energy consumption and climate impact are also concerns with the increased use of AI and analytics. In the next 5-10 years, quantum computing will likely transform data analytics, making it faster and more insightful. This will allow us to break many myths around conventional machine learning and better understand causality in healthcare through simulations.

AINA: Are there any mechanisms in place in Apollo or other healthcare institutions to assess the potential impact of new trends and developments?

The governance framework for understanding the ROI of investments in new technologies is evolving. We assess both monetization and operational gains, such as throughput optimization and length of stay reduction. For example, we reduced the length of stay from 5 to 4 days, and we are looking to see if throughput optimization algorithms can bring it down further to 3.5 days.

New technologies are piloted in hospitals to evaluate their impact and cost-effectiveness. Rigorous testing, validation, and ethical review are necessary before any new tool is implemented in patient care. We must consider accuracy, sensitivity, specificity, and predictive values to ensure the effectiveness and safety of new technologies.

AINA: What advice would you give to new students entering the field of healthcare analytics to be more relevant and successful?

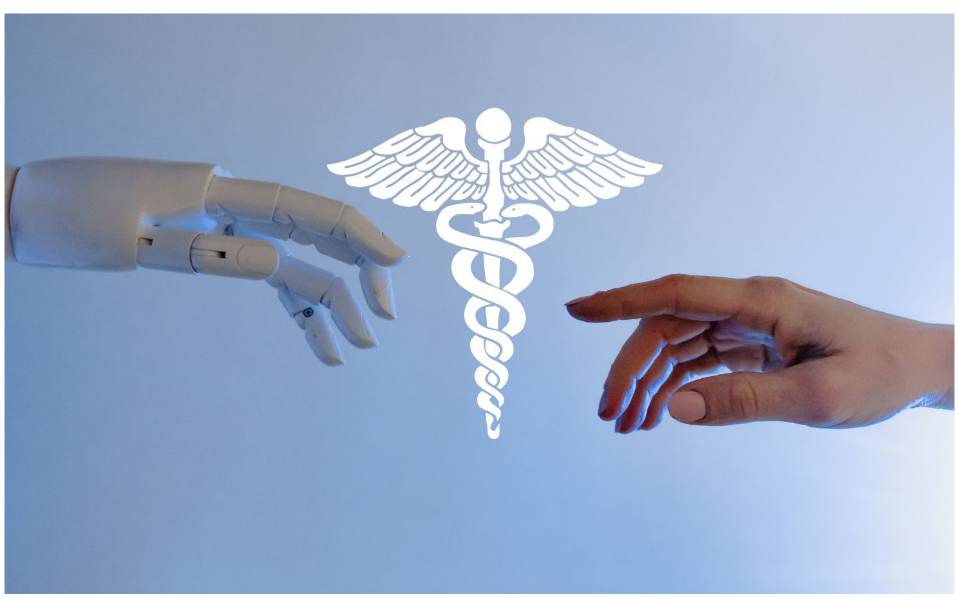
First, maintain intellectual honesty in data science. Don't manipulate data to fit preconceived notions. There's a saying that if you hammer the data too much, it will tell you whatever you want. Second, have a healthy understanding of the domain. Spend time understanding the data structures and insights relevant to your field. It is also important for domain experts to give data scientists enough time and resources to understand the domain.

Third, focus on delivering practical solutions to organizations using data science. The appeal of data science lies in its ability to deliver value, and you must constantly work on refining your models and methodologies to sustain its relevance. The field is no longer about the excitement of data science but about how you deliver value to your organization.

AINA: Healthcare is a complicated domain. How do you ensure that analytics results lead to the expected outcomes, and how do you identify where things might go wrong in implementation?

Solutions can fail at any stage: design, development, validation, deployment, or adoption. Rigorous testing, proper data management, and domain knowledge are crucial. Use frameworks like Failure Mode and Effects analysis (FMEA) to identify potential issues at each step. Intellectual honesty, domain expertise, and appropriate risk management are essential to ensure successful implementation.

Ensuring the right data at every source, using proper data engineering tools, curating content, choosing the right machine learning tools, building appropriate interfaces, and clearly stating limitations and disclaimers are all critical steps. Each step must be done correctly to avoid failures in implementation.



Unveiling Graph RAG

A New Era in NLP

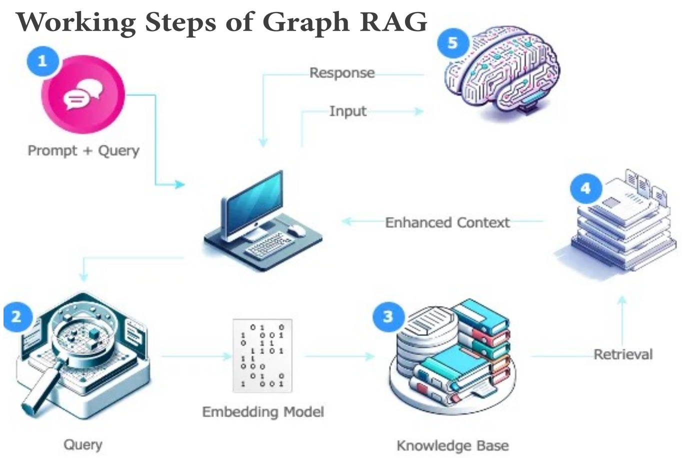
Soumendu Mandal

Graph RAG (Retrieval Augmented Generation) is an emerging and transformative technique in natural language processing (NLP). It promises to revolutionize the way large language models (LLMs) handle complex and contextually rich queries. By integrating external knowledge, particularly using knowledge graphs (KGs), Graph RAG aims to provide more accurate and contextually relevant answers. Traditional RAG approaches have limitations, mainly relying on vector similarity to retrieve information from plain text snippets. Graph RAG leverages the structured, interconnected nature of KGs to enrich the context available to LLMs, significantly enhancing their performance in a variety of applications.

Traditional RAG vs. Graph RAG

RAG (Retrieval Augmented Generation) is a technique designed to augment LLM outputs with external information, often referred to as "grounding context." This approach involves a retrieval component that fetches additional data from external sources (which are generally private files of the user), which is then incorporated into the LLM's prompt. This enhances the relevance and accuracy of the LLM's responses. Traditional RAG methods, while effective in many scenarios, often struggle with certain types of queries.

These include queries that require the synthesis of information from disparate sources or the holistic understanding of large datasets. The reliance on vector similarity to retrieve information can also limit the depth and accuracy of the responses.



Advantages of GraphRAGs

Graph RAG addresses the limitations mentioned above by utilizing knowledge graphs. Knowledge graphs provide a structured representation of information, connecting entities (such as people, places, or organizations) through various relationships. This structured format allows for a more nuanced understanding of context, capturing semantic relationships and properties of entities in a way that plain text cannot. By leveraging the structured data in KGs, LLMs can perform more sophisticated reasoning about the information, leading to more accurate and contextually relevant answers, especially for complex queries. This makes it particularly valuable for enterprise applications where domain-specific proprietary knowledge is crucial.

Banishing Hallucinations

One of the key advantages of Graph RAG is its ability to reduce the tendency of LLMs to generate "hallucinations"—responses that are factually incorrect or not grounded in the provided context. Traditional LLMs, when faced with a lack of specific context, may generate plausible-sounding but incorrect information. By anchoring the LLM's output in the structured, factual data from the knowledge graph, Graph RAG ensures that the generated responses are more reliable and accurate. This makes it an ideal solution for applications where precision is paramount, such as in legal, medical, or financial contexts. Ensuring the accuracy of responses in these fields is critical, as incorrect information can have serious consequences.

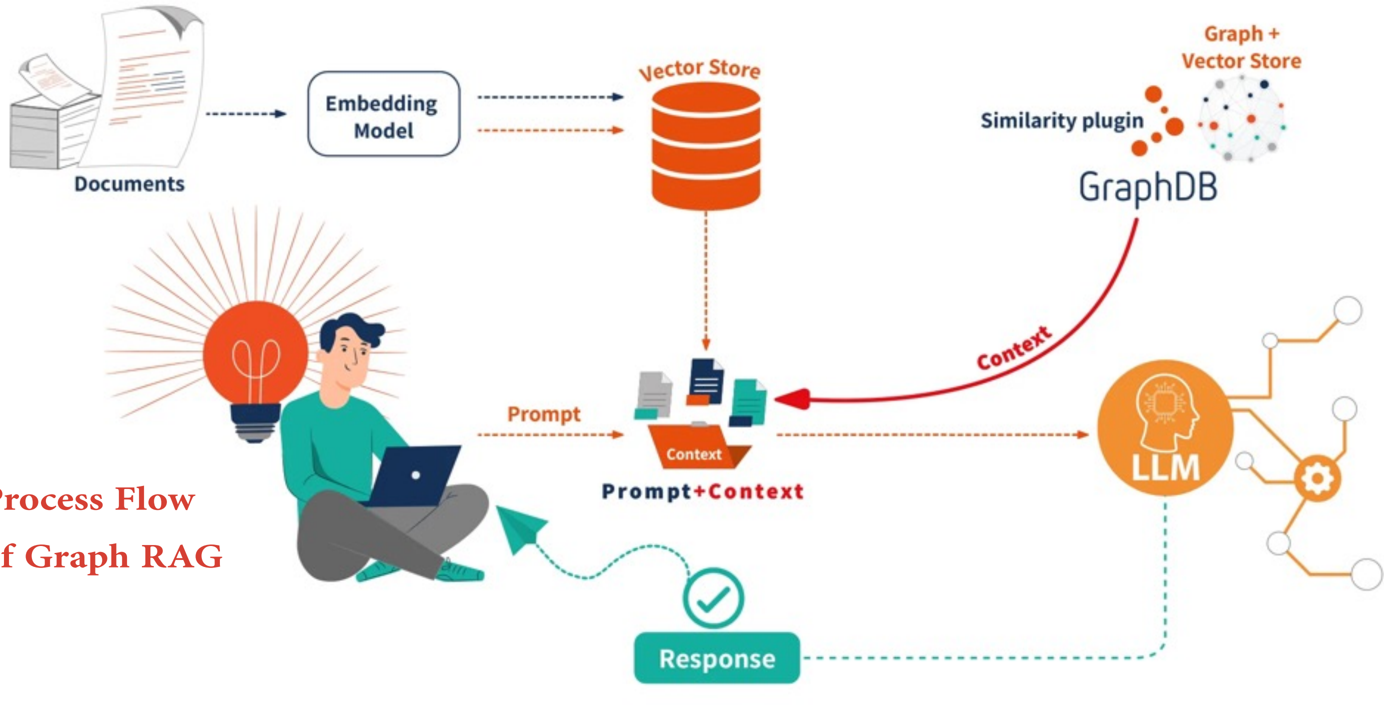
Blueprint for Implementation

Implementing Graph RAG involves several steps to ensure that the LLM can effectively utilize the structured data from knowledge graphs. First, the input corpus is divided into TextUnits, which are analyzable units of text. These units can be sentences, paragraphs, or other segments of text that contain meaningful information. An LLM is used to extract entities, relationships, and key claims from these TextUnits, creating a rich graph of interconnected information, where entities are nodes and relationships are edges.

The knowledge graph is subjected to hierarchical clustering using techniques like the Leiden algorithm. This algorithm groups entities into communities based on their relationships, creating clusters of related information. Summaries of these communities are generated from the bottom-up, providing a holistic understanding of the dataset. These summaries capture the key points and relationships within each community, making it easier for the LLM to access relevant information during query time.

Intelligent Query Modes

At query time, the structured information in the knowledge graph is leveraged to provide materials for the LLM context window when answering questions. There are two primary query modes: Global Search and Local Search. Global Search uses community summaries to reason about holistic questions concerning the entire corpus. This mode is useful for broad, overarching questions that require an understanding of the dataset. Local Search focuses on specific entities and their



Process Flow of Graph RAG

associated concepts, providing detailed insights. This mode is ideal for queries that require in-depth information about entities or relationships within the knowledge graph. This dual approach allows Graph RAG to handle a wide range of query types, from broad overviews to specific, detailed inquiries.

Diverse Flavors of Graph RAG

There are different varieties of Graph RAG, each with its own approach to integrating knowledge graphs with LLMs. In one approach, the knowledge graph serves as a content store, extracting relevant chunks of documents and using them to answer questions. This method involves retrieving specific pieces of information from the knowledge graph to enhance the LLM's responses. Another approach treats the knowledge graph as a subject matter expert, providing descriptions of concepts and entities relevant to the query, along with their relationships. This method leverages the structured data in the knowledge graph to offer detailed insights into specific topics.

A third approach maps natural language questions to graph queries, executing these queries, and summarizing the results. This method involves translating the user's query into a format that can be used to query the knowledge graph directly, providing precise and relevant answers. Each of these methods requires different config urations of the knowledge graph and varying degrees of integration with vector databases.

A New Horizon for NLP

Graph RAG represents a significant advancement in the field of NLP. By combining the strengths of retrieval-based and generative approaches, it enhances the capabilities of LLMs. Integrating knowledge graphs into the RAG process provides a richer, more reliable context for generating responses, improving accuracy, and reducing hallucinations. This technique is particularly valuable for enterprise applications where domain-specific knowledge is crucial, opening up new possibilities for advanced chatbots, natural language querying, and information extraction.

As the field continues to evolve, Graph RAG stands out as a pivotal development, paving the way for more sophisticated and contextually aware LLMs.

AutoML

Powers, Players, Potential, & the Path Ahead.

Himanshu Singh

Brief Overview of Machine Learning

As you would all know, Machine learning (ML) is a subset of artificial intelligence and it involves training algorithms to make predictions or decisions based on data. Machine Learning, or in the context of this article, Traditional machine learning, requires a significant amount of human intervention, from selecting the appropriate algorithms to feature engineering and parameter tuning. This process is not only time-consuming but also requires a high level of expertise, which can be a little too overwhelming for organizations without extensive data science resources.

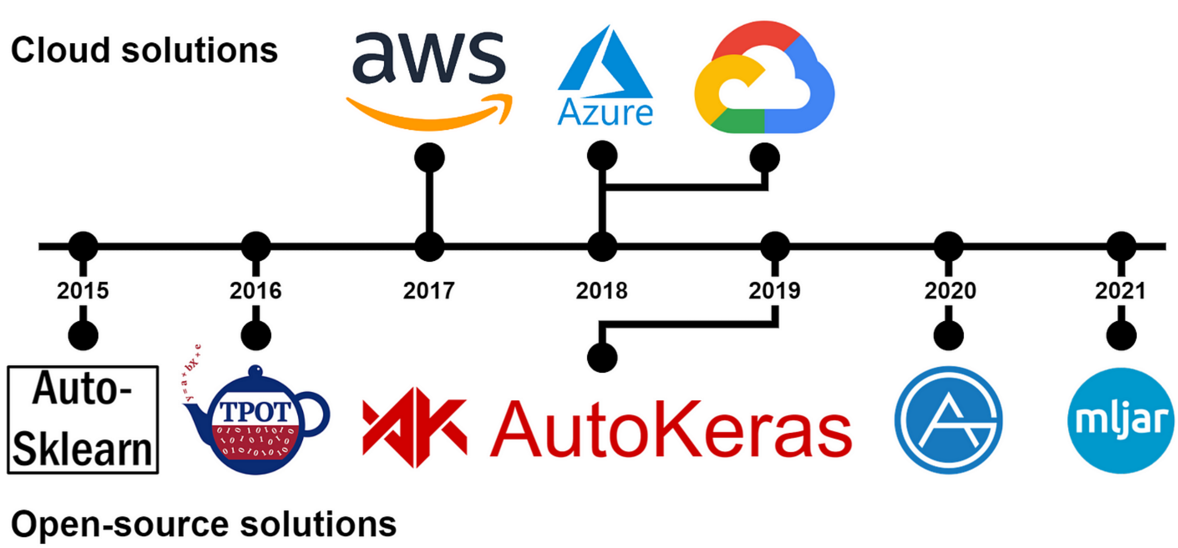
"AutoML tools empower users to focus on their data and objectives rather than the intricacies of algorithm implementation."

In order to address the above limitations few people started working on a new technology in the early 2000s, which ultimately was called Automated Machine Learning or AutoML for short. This was envisioned to automate many of the steps involved in a Machine Learning project like algorithm selection, feature engineering, and hyperparameter optimization. This automation could make machine learning more accessible to wider audiences and increase the pace of model development too. Thus, AutoML was born with an intention to reduce the barrier to entry and enable more organizations to leverage this potent technology.

Evolution of AutoML

Looking back we can say that the concept of Automated Machine Learning (AutoML) has come a long way since its inception. The journey began in the early 2000s with the development of meta-learning frameworks, which aimed to understand the best machine learning approaches for different types of data automatically. In 2013, platforms like Auto-WEKA leaped forward by integrating algorithm selection and hyperparameter tuning into a single framework. This development demonstrated the feasibility and benefits of AutoML and led to an increased interest and research in this area.

The mid-2010s saw rapid progress with the introduction of tools like Google's AutoML and Microsoft's Azure Machine Learning, which began to integrate cloud computing resources to AutoML solutions. These platforms opened the doors of machine learning to sizeable audience that didn't have deep technical knowledge. Since then a lot of players have entered the space viz. AutoKeras, mljar and H2O.ai, and others like Google and Amazon introduced updated versions of their AutoML platforms—Google's AutoML Tables and Amazon's SageMaker Autopilot.



"..global AutoML market is valued at $1 billion in 2023 and is projected to reach 6.4 billion by 2028"

AutoML Capabilities

AutoML relieves the pressure from its users for creating an ML model for their dataset by generating a suitable model autonomously. As depicted in the figure at the top, AutoML typically takes three inputs from the user: (i) a dataset, (ii) the ML task to be executed on this dataset, and (iii) the metric to be used to determine which model has performed best. For instance, the dataset can be ImageNet, the task can be image classification, and the metric can be accuracy.

The basic objective behind AutoML is automating the various steps involved in developing a machine learning model. It begins with Data Cleansing, addressing missing values, standardizing data types, identifying and rectifying anomalies, encoding text data, and partitioning the dataset.

This step is also called as **data preprocessing** at many places. The next phase is Feature Extraction, which involves transforming raw data into meaningful features, handling various data types like numeric, discrete, textual, image-based, time-series, and cross-features.

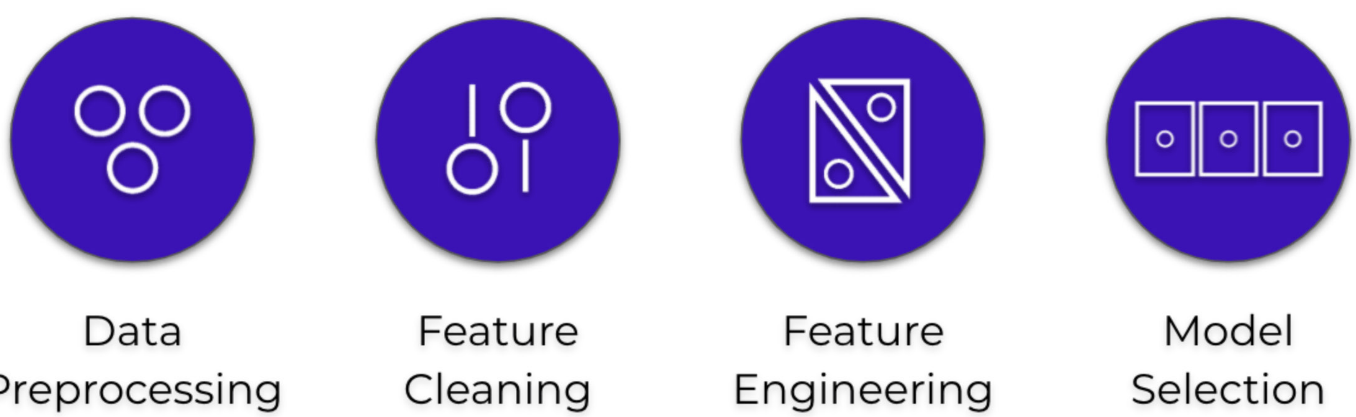
Following this is **Feature Selection**, where important features are identified using linear and non-linear projections, and dimensionality reduction techniques are applied to simplify the model without losing significant information and selecting or engineering the most relevant features to improve model performance.

Thereafter, **Model Selection** phase, here the system evaluates a variety of machine learning algorithms—from simple linear regressions to more complex ensembles and neural networks and uses techniques like meta-learning, which draws on knowledge from previously solved problems, to predict which models might perform best given the type of data at hand.

Finally, AutoML **optimizes hyperparameters** of selected models automatically. Traditionally, tuning these parameters is a trial-and-error process that can be very time-consuming. AutoML leverages algorithms like Bayesian optimization, evolutionary algorithms, or random search to efficiently explore the hyperparameter space and find the optimal settings for each model.

For example, if the model needs to be deployed on a device with limited computational power, it might prioritize simpler algorithms that meet practical requirement and deliver reasonable results quickly.

Steps performed by AutoML



Enterprise AutoML platforms

Google Cloud

Amazon SageMaker

Amazon Lambda

Amazon AutoGluon

IBM

AutoML Software Solutions and Frameworks

H2O AutoML from H2O.ai supports a broad range of models and tasks, while Google Cloud AutoML and Azure Machine Learning AutoML provide cloud-based solutions for various tasks, integrating well with their respective ecosystems. Amazon SageMaker Autopilot offers comprehensive tools for model training and optimization, integrating-with AWS and BigML's suite.

Auto-machine learning (AutoML) simplifies applying machine learning to real-world problems through automation. Key libraries include Auto-sklearn and TPOT, which build on scikit-learn, offering Bayesian optimization and genetic programming, respectively. AutoKeras focuses on deep learning with neural architecture search, and MLBox covers end-to-end machine learning processes.

> *"...decision points are guided by criteria such as data characteristics, computational resources, desired model performance, and constraints like the need for model interpretability."*

TransmogrifAI by Salesforce and Databricks are designed for large-scale tasks with Apache Spark, and Ludwig by Uber AI Labs allows model training without code. PyCaret provides a low-code solution for diverse machine learning tasks, making model selection and optimization more accessible.

Startup AutoML Platforms

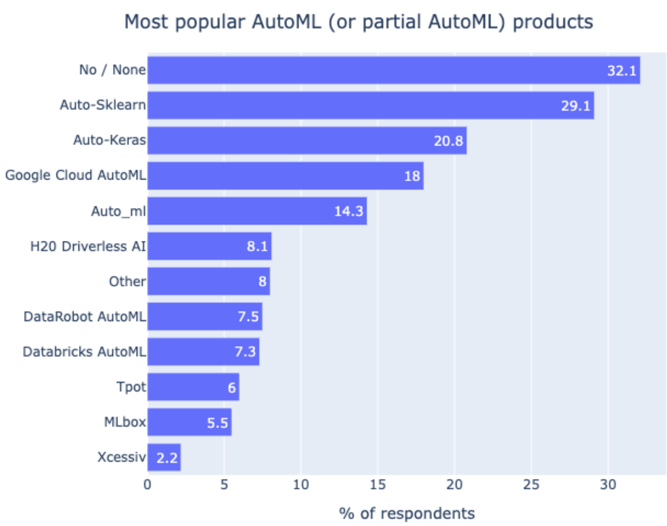
databricks

dataiku

H2O.ai

Big players in AutoML industry

According to the 2020 Kaggle State of ML and Data Science Survey, among tech giants only Google ranked in the top four most used AutoML frameworks. The other leaders are enterprise platforms that automate many steps of the machine learning process. These systems support big data technologies like Hadoop and Spark and make model deployment easy, whether on-premises or in the cloud (AWS, Google Cloud, or Microsoft Azure).



Real-life applications of AutoML

AutoML has been widely adopted across industries, demonstrating its versatility and impact. In the banking sector, PayPal employs AutoML to analyze transactions in real-time and help to identify and prevent fraudulent activities more effectively. Using Automl, PayPal can quickly adapt to new fraudulent strategies as they arise.

In the retail industry, Walmart uses AutoML to forecast demand and optimize stock levels across its numerous locations improving operational efficiency and customer satisfaction by ensuring product availability.

Mobile Broadband Network LTD (MBNL) a leading provider of telecommunication services in UK, implemented AI-driven predictive maintenance to avoid failures between regular checks and benefitted from AutoML. The automated approach helped to complete proof-of-concept within six weeks while typically it took one to two years.

The Adecco Group, the world's second largest HR provider, uses AutoML to speed up job filling. This approach enabled the launch of 60 machine learning projects with 3,000 models in just three weeks. The best models filter out up to 37 percent of unsuitable CVs, saving recruiters time and increasing productivity by 10 percent.

Last but not the least, in healthcare industry, University of Pittsburgh Medical Center (UPMC) had to previously contact 10,000 candidates against 1,500 parameters to identify one potential patient for liver donation. With the Squark platform's low-code AutoML solutions this number reduced to just 75, and model development time from months to hours.

Above examples underscore the transformative potential of AutoML across various sectors, making complex machine learning models more accessible and effective in addressing specific industry challenges.

Some Expert Opinions

AutoML is widely regarded as a game-changer in the field of machine learning. Dr. Fei-Fei Li, professor in Stanford Institute for Human-Centered Artificial Intelligence, praised AutoML for its ability to democratize AI, making powerful machine learning tools accessible to a broader range of users.

Citizen Data Scientists are individuals who use advanced analytics tools without being formally trained data scientists, either in professional or personal capacity. AutoML has made this phenomenon possible.

Similarly, Dr. Jeff Dean, Google Senior Fellow and leader of Google AI, has highlighted that it not only reduces the time required to develop effective models but also significantly improves the ability of businesses to customize solutions to their specific needs. At the same time, he cautions "if you learn from data and that data has biased decisions in it already, then the machine learning models who learn can themselves perpetuate those biases."

Furthermore, Dr. Joaquin Vanschoren, Assistant Professor of Machine Learning at the Eindhoven University of Technology, has pointed out that by automating the design of machine learning models, researchers can explore more complex model architectures and hypothesis spaces more efficiently than ever before, potentially leading to new breakthroughs in AI capabilities.

Challenges and Considerations of AutoML

While AutoML offers significant advantages, it also presents a range of challenges and ethical considerations. One of the primary limitations is the quality of data input; AutoML systems are highly dependent on the data they are trained with. Poor quality or biased data can lead to inaccurate models, perpetuating or even exacerbating existing biases, for example in recruitment for firms.

The impact of AutoML on the job market and skill requirements is another area of concern. While it enhances access to machine learning, there is an apprehension that it could lead to job displacement, particularly for roles that involve routine data analysis and model building. However, the counter argument that it also creates opportunities for new job roles focused on managing and interpreting AI systems is also strong enough. The skillset required is shifting towards a hybrid of domain expertise and AI literacy, meaning continuous learning and adaptation are becoming essential for professionals.

Ethical considerations are also paramount. The deployment of AutoML raises questions about accountability, especially in critical applications like healthcare or autonomous driving. Decisions made by models generated by AutoML could have serious implications, and tracing the decision-making process can be challenging. This opacity necessitates strict governance and ethical guidelines to ensure that AutoML systems are used responsibly and that there is clarity regarding liability in cases of failure.

These challenges highlight the need for a balanced approach to AutoML adoption, with an emphasis on ethical standards, quality data management, and continuous skill development. By addressing these issues, the potential of

AutoML can be harnessed effectively, ensuring it contributes positively to technological advancement and societal needs.

AutoML vs. ML Engineers

A 2015 Google paper on "technical debt" in machine learning highlighted that most ML systems consist predominantly of "glue code"—necessary code for data management within ML algorithms—making up about 95% of the total code. Despite AutoML's advancements in handling glue code, it still forms only a small part of an entire ML solution. AutoML is not yet a replacement for data scientists for several reasons.

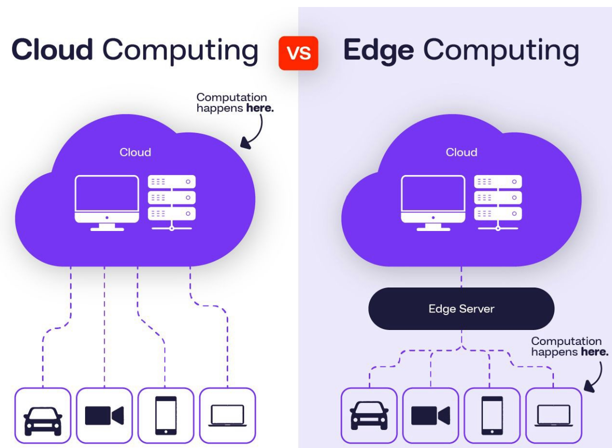
Firstly, autoML focuses mainly on model performance, which is just one aspect of real-world ML projects that may also require simpler models or models with high explainability due to practical constraints.

Secondly, in competitive environments like Kaggle, human data scientists still outperform AutoML solutions, demonstrating the irreplaceable value of human creativity and strategic insight. Thirdly, AutoML currently addresses only a limited range of problem types that are typically handled by data scientists.

Lastly, even the most advanced AutoML platforms can't incorporate domain expertise and replace human creativity and imagination that play a key role in crafting meaningful attributes. Therefore, while AutoML is a valuable tool, it cannot fully replace data scientists, who bring essential depth and flexibility to tackling diverse ML challenges.

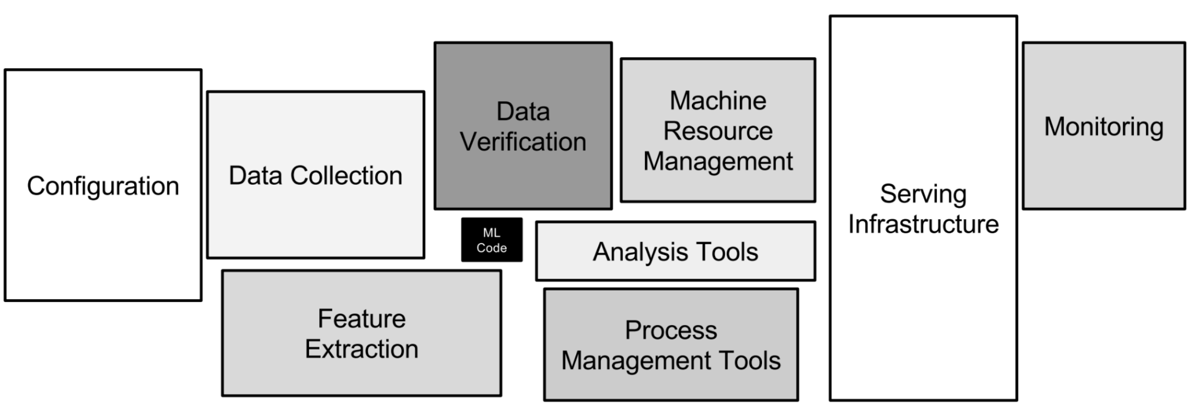
Future of AutoML

The development of AutoML is rapidly evolving, with current trends indicating a future where it becomes even more integrated into various industries and more accessible to a larger audience. One significant direction is the enhancement of interpretability and transparency. As businesses and regulators demand greater understanding of how AI models make decisions, particularly in sensitive areas like finance and healthcare, developers are focusing on creating more interpretable AutoML tools. This ensures that outputs can be explained and justified, which is crucial for building trust and compliance with regulatory standards.



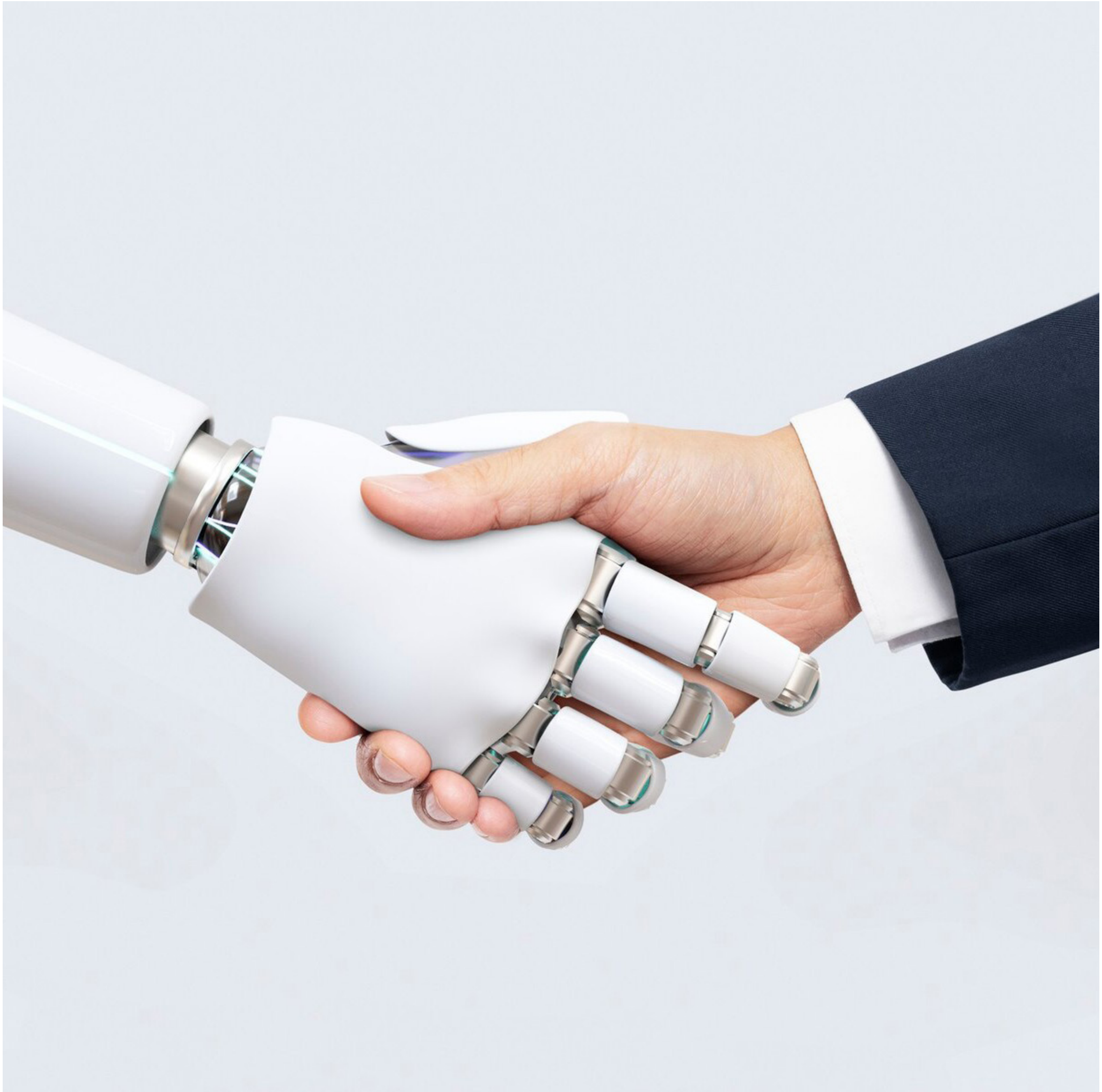
Another promising development is the integration of AutoML with edge computing. This involves deploying AI models directly on devices like smartphones and IoT devices, rather than running them on centralized servers. This approach reduces latency, increases privacy, and improves the responsiveness of applications. For example, Apple and Google are enhancing their mobile operating systems with capabilities that allow developers to incorporate AutoML models directly into apps that can run efficiently on users' devices.

Glue code makes up 95% of the total code in ML Systems - 2015 Google paper on "Technical Debt"



Furthermore, there is a growing focus on developing AutoML solutions tailored for specific industries. This specialization involves incorporating domain-specific knowledge into the AutoML process, which can drastically improve the relevance and effectiveness of the models produced. For instance, in the agricultural sector, companies are using AutoML to predict crop yields and optimize farming techniques, incorporating environmental variables and historical data to provide tailored recommendations.

These advancements are setting the stage for a future where AutoML not only simplifies the application of machine learning but also ensures that these technologies are sustainable, ethical, and beneficial across all sectors of society. As these trends continue, we can expect AutoML to become a cornerstone of innovation, driving forward a new era of technology that is more inclusive and impactful.



References:

Machine Unlearning: The Art of Selective Forgetting

[1] https://arxiv.org/pdf/1912.03817
[2] https://medium.com/@aliborji/a-fresh-perspective-on-machine-unlearning-with-a-real-world-solution-203821dd01c0
[3] https://pub.towardsai.net/machine-unlearning-in-2023-where-we-are-and-where- its-headed-2ef19ad0394
[4] https://research.google/blog/announcing-the-first-machine-unlearning-challenge/
[5] https://www.wired.com/story/machines-can-learn-can-they-unlearn/
https://arxiv.org/abs/2402.00351

The Invisible Shopkeeper: Big Data's Silent Revolution in Retail

[1] https://robots.net/fintech/how-walmart-uses-big-data/
[2] https://www.fortunebusinessinsights.com/industry-reports/infographics/big-data-technology-market-100144
[3] https://www.tokinomo.com/blog/big-data-in-retail
[4] https://dataforest.ai/blog/how-big-data-analytics-is-transforming-the-retail-industry
[5] https://www.turing.com/resources/big-data-in-retail
[6] https://hbr.org/2012/10/big-data-the-management-revolution
[7] https://bernardmarr.com/amazon-using-big-data-to-understand-customers/
[8] https://ideas.repec.org/a/eee/jbrese/v70y2017icp356-365.html
[9] https://doi.org/10.1080/00207543.2015.1061222

Silicon Superstars: How GPUs became the driving force of AI

[1] https://www.sesterce.com/blog/why-your-ai-dreams-depend-on-nvidia-gpus
[2] https://www.sesterce.com/blog/best-gpus-for-deep-learning
[3] https://www.sesterce.com/blog/gpu-trends-redefining-ai
[4] https://blogs.nvidia.com/blog/blackwell-scientific-computing/
[5] https://www.rtinsights.com/nvidia-intros-new-ai-chips-and-software-to-boost-artificial-intelligence-capabilities/

Leveraging AI to Revolutionize Sales & Marketing in Pharma: An AI Guide for Medical Representatives

[1] https://www.salesforce.com/blog/sales-route-planning-guide/
[2] https://www.handy.ai/post/enhancing-doctor-success-with-handy-ai-at-sanofi
[3] https://www.iqvia.com/solutions/real-world-evidence/health-data-apps-and-ai/commercial-and-brand-hcp-profiling
[4] https://www.mckinsey.com/industries/healthcare/our-insights/the-ai-opportunity-how-payers-can-capture-it-now?cid=other-soc----oth----ip&sid=soc-POST_ID&linkId=459875272
[5] https://www.quantified.ai/blog/how-ai-can-improve-your-pharmaceutical-sales-rep-certification-program/
[6] https://www.linkedin.com/pulse/enhancing-pharma-sales-rep-effectiveness-ai-dr-andr%C3%A9e-bates-vtooe/
[7] https://azuremarketplace.microsoft.com/en-US/marketplace/apps/cogniwaresro1624625552086.argos-004
[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931338/
[9] https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-021-00687-3

AI Governance

[1] https://www.nytimes.com/2019/11/09/business/Apple-credit-card-investigation.html
[2] https://hbr.org/2021/10/how-to-keep-your-ai-from-turning-biases-into-discrimination
[3] https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/an-executives-guide-to-ai-governance
[4] https://www.weforum.org/agenda/2020/01/why-we-need-a-global-framework-for-ai-governance/
[5] https://www.brookings.edu/research/the-ai-governance-challenge/
[6] https://digital-strategy.ec.europa.eu/en/policies/europe-fit-digital-age
[7] https://www.weforum.org/centre-for-the-fourth-industrial-revolution/partners
[8] https://www.niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf
[9] https://www.un.org/en/white-paper-ai-governance
[10] https://www.ibm.com/products/watsonx-governance

References:

Behind the Scenes: How AI is Transforming the Entertainment Industry

[1] https://www.scriptbook.io/#!/scriptbook
[2] https://www.awn.com/news/avengers-endgame-and-irishman-ai-powered-visual-effects-come-age
[3] https://www.forbes.com/sites/neilsahota/2024/03/08/the-ai-takeover-in-cinema-how-movie-studios-use-artificial-intelligence/
[4] https://news.disney.com/making-of-lion-king
[5] https://www.youtube.com/watch?v=PITNvuEV0C0
[6] https://skill-lync.com/blogs/what-are-the-technologies-used-in-avatar-way-of-the-water
[7] https://deepmind.google/technologies/veo/
[8] https://openai.com/index/sora/
[9] https://www.keywordsstudios.com/en/ava-at-gdc/
[10] https://www.hiddendoor.co/
[11] https://www.theguardian.com/culture/2023/oct/01/hollywood-writers-strike-artificial-intelligence
[12] https://www.bloomberg.com/news/articles/2023-10-26/hollywood-actors-still-hashing-out-basic-principles-of-ai-use-with-studios

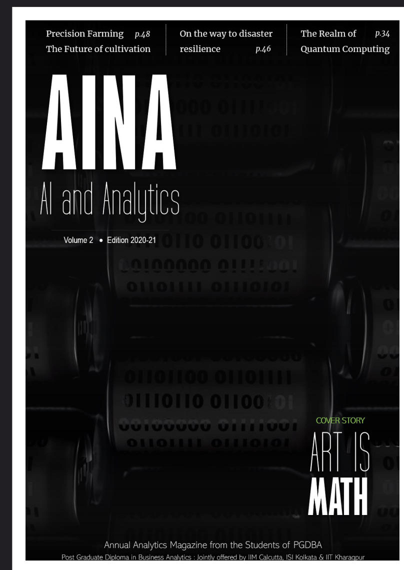
AutoML: Powers, Players, Potential, & the Path Ahead

[1] Machine Learning: The High-Interest Credit Card of Technical Debt; D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
[2] https://www.paypal.com/us/brc/article/paypal-machine-learning-stop-fraud
[3] https://cloud.google.com/blog/topics/inside-google-cloud/cloud-automl-making-ai-accessible-every-business/
[4] https://venturebeat.com/ai/googles-ai-chief-on-automl-autonomous-weapons-and-the-future/
[5] https://blog.mirkopeters.com/automl-the-revolutionary-tool-thats-changing-machine-learning-forever-c7a74ca8f338
[6] https://medium.com/walmartglobaltech/walts-walmart-automl-libraries-tools-and-services-936c73e5ec67
[7] https://www.altexsoft.com/blog/automl/
[8] https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-automl/
[9] Whitepaper on Importance of AutoML for Augmented Analytics - Dataiku
[10] https://media.geeksforgeeks.org/wp-content/uploads/20240206154951/1_0U3LAye3LxFcHg0UePmbSA.webp
[11] https://americancloud.com/blog/cloud-computing-vs-edge-computing-which-is-best-for-your-business

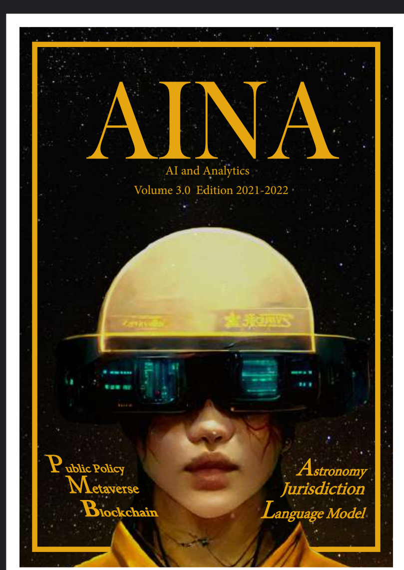
Batch 9 Magazine team signing off !!!

PREVIOUS EDITIONS

To read AINA Vol 1 scan the QR code or go to https://online.anyflip.com/lgfta/gxxz/mobile/index.html



To read AINA Vol 2 scan the QR code or go to https://online.anyflip.com/exokr/ehwf/mobile/index.html



To read AINA Vol 3 scan the QR code or go to https://online.anyflip.com/lgfta/wqew/mobile/index.html

To read AINA Vol 4 scan the QR code or go to https://online.anyflip.com/lgfta/vvas/mobile/index.html