
GOOGLE: VERTEX AI

ARTIFICIAL INTELLIGENCE | GENERATIVE AI

A White Paper

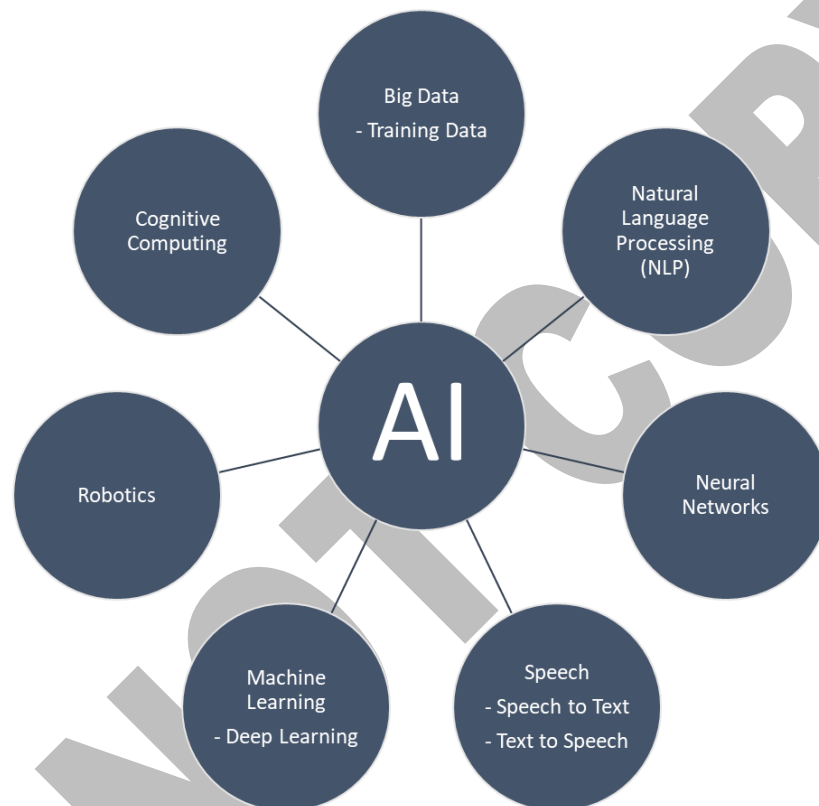
Prepared by Devang Maniar
Research Paper

Table of Contents

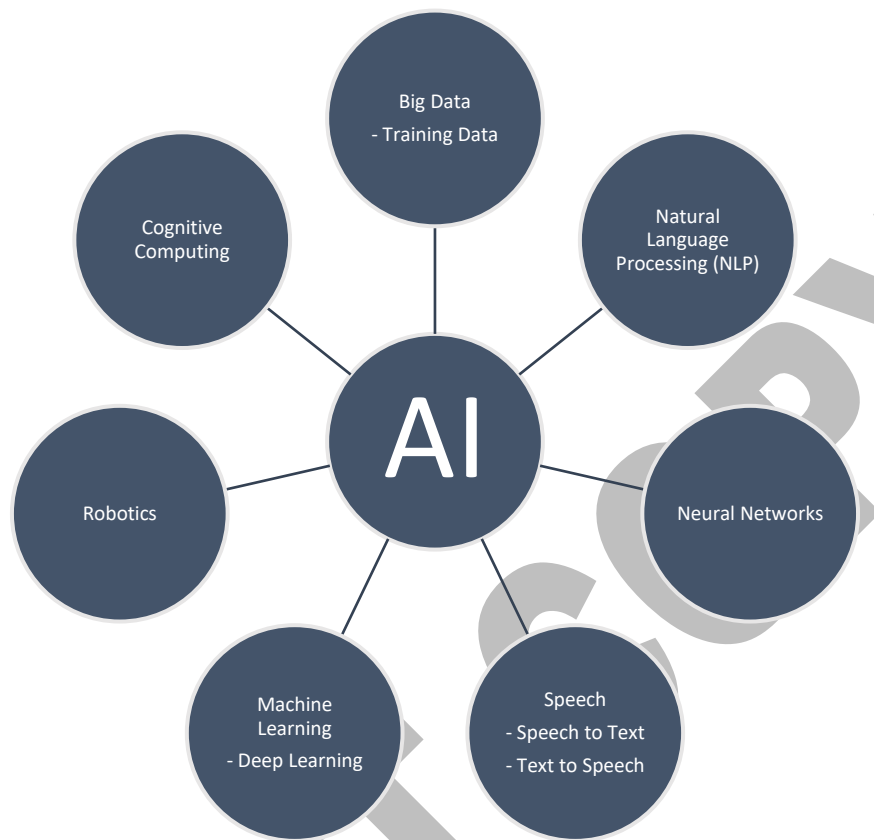
1.0	Introduction to Artificial Intelligence	3
1.1	AI Technology	3
1.2	Tools & Software implementing AI	4
2.0	Introduction to Generative AI	4
2.1	Generative Adversarial Network Architecture (GAN)	5
2.3	Various Types of Applications for Generative AI	6
2.4	Use Cases for Generative AI	7
2.5	Sectors of Generative AI	8
2.6	Features of Generative AI	9
2.7	Benefits of Generative AI	11
2.8	Industries exploring opportunities with Generative AI	11
3.0	Introduction to Machine Learning (ML)	12
3.1	Understanding MLOps	13
3.3	Capabilities of MLOps	14
3.3	Steps to Deploy MLOps into Production	15
4.0	Introduction GCP Vertex AI	16
4.1	Features of Vertex AI	17
4.2	Use Cases for Vertex AI	18
4.3	Benefits of Vertex AI	18
5.0	Conclusion	19
6.0	Reference Links	20

1.0 Introduction to Artificial Intelligence

Artificial Intelligence (AI) in simple words is training a machine to mimic human capabilities by performing tasks typically done by humans. In a nutshell, it is trying to make computers think and act like humans. Sounds scary, doesn't it? On the surface, it can sound scary but when you dig deeper, you will realize that it is not AI that we should be scared from but rather the person or the creator behind the AI.



See Figure 1a

**Figure 1a**

1.1 AI Technology

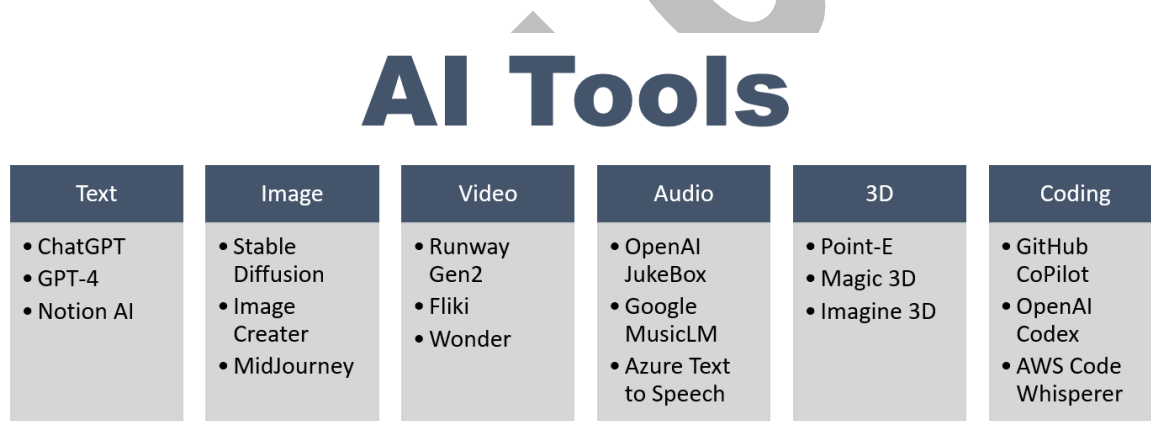
To dig deeper, one must first understand in general as to how the AI technology works. Let us say you're going on a family road trip, and you require directions to your destination, you're likely to nowadays use Google Maps to guide you with the directions. But as you are driving, the kids get hungry, and you are now suddenly finding yourself looking for a restaurant nearby that is both pet and kid friendly. Google Maps will obviously point out many restaurant choices, but your goal is to find something in the same direction as your route and within ten minutes of the pathway. This is where AI technology comes in handy because Google is not just producing the choices based on type of restaurant, but also considering the time, location and direction compared to your ongoing route.

AI basically takes large amounts of raw data, and after some cleansing and processing through algorithms, finds possible connections and patterns to deliver personalized recommendations back to humans. So, when Google Maps is giving your personalized restaurant recommendations, it is using its massive collection of data algorithms to provide an answer back to you. The process of these algorithms keeps improving over time as more data

gets collected and this is known as Machine Learning (ML). Just as how the human brain evolves over age so does Machine Learning because it is constantly learning from previous mistakes to get better. Below is a diagram showing the workflow of Machine Learning.

1.2 Tools & Software Implementing AI

Of course, Google is just one example of an AI Tool, but there are thousands of different AI Tools that are designed for different industries targeting different solutions and tasks. Microsoft has created an AI Tool known as Azure Cognitive which converts written text into a synthesized humanlike voice that not only sounds realistic but can also talk in 140+ languages and accents. Amazon has its own AI Tool known as Code Whisperer that generates actual code suggestions to help you quickly build software. There are AI Tools such as ChatGPT that can even generate valuable content based relevant to the topics you pick, and will generate an entire article within minutes. This makes a great segway to the next topic and that is Generative AI. See Figure 1b.



Text	Image	Video	Audio	3D	Coding
<ul style="list-style-type: none">• ChatGPT• GPT-4• Notion AI	<ul style="list-style-type: none">• Stable Diffusion• Image Creator• MidJourney	<ul style="list-style-type: none">• Runway Gen2• Fliki• Wonder	<ul style="list-style-type: none">• OpenAI JukeBox• Google MusicLM• Azure Text to Speech	<ul style="list-style-type: none">• Point-E• Magic 3D• Imagine 3D	<ul style="list-style-type: none">• GitHub CoPilot• OpenAI Codex• AWS Code Whisperer

Figure 1b

2.0 Introduction to Generative AI

Generative AI is when the technology is producing new content such as text, photo, video, code, data or even 3d renderings based on existing data. It refers to unsupervised and semi-supervised machine learning algorithms that enable computers to use existing content like text, audio and video files, images, and even code to create new possible content. The main idea is to generate completely original artifacts that would look like the real deal.

Generative AI is part of the Machine Learning framework because it is essentially making a prediction on an output of a content based on historical data. For example, there is an AI Tool by Google known as MusicLM. This tool essentially takes the user's input through text prompts, and uses the words to make a 10-second audio clip or even a brand-new full song. The tool is essentially compiling 280,000 hours of unlabeled music and creating its own new music for you. Below is a simplified workflow of how Generative AI works.

2.1 Generative Adversarial Network Architecture (GAN)

Generative AI commonly works by a deep learning model known as Generative Adversarial networks (GANs). GAN is a type of Machine Learning model that carries two sub models: generator model and discriminator model. While the generator is trained to produce content, the discriminator is trained to discriminate between real content from the dataset versus fake content generated by the generator. These two models work in tandem to create a more lifelike content.

Both the generator and the discriminator are neural networks. The generator output is connected directly to the discriminator input. When training begins, the generator produces obviously fake data, and the discriminator quickly learns to tell that it's fake. As training progresses, the generator gets closer to producing output that can fool the discriminator. Finally, if generator training goes well, the discriminator gets worse at telling the difference between real and fake. It starts to classify fake data as real, and its accuracy decreases.

- 1) Discriminative Modeling – It is used to classify existing data points (e.g., images of cats and guinea pigs into respective categories). It mostly belongs to supervised machine learning tasks. Most machine learning models are used to make predictions. Discriminative algorithms try to classify input data given some set of features and predict a label or a class to which a certain data example belongs.
- 2) Generative Modeling – Generative algorithms do the opposite of Discriminatory modeling. Instead of predicting a label given to some features, they try to predict features given a certain label. It tries to understand the dataset structure and generate similar examples (e.g., creating a realistic image of a guinea pig or a cat). It mostly belongs to unsupervised and semi-supervised machine learning tasks. See Figure 1c.

Generative Adversarial Network (GAN)

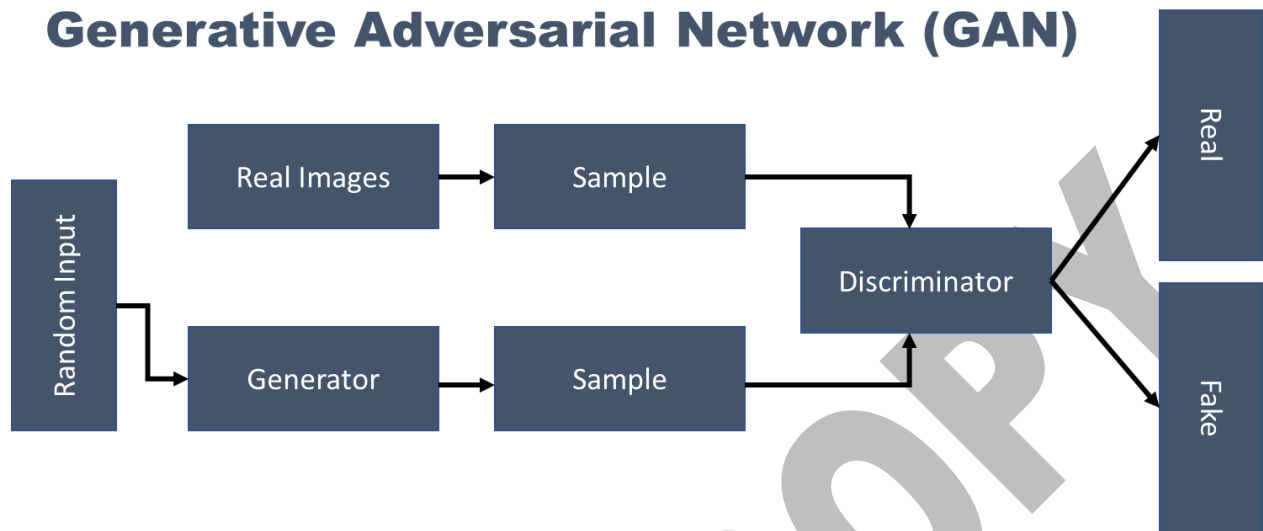


Figure 1c

2.2 Various Types of Applications for Generative AI

Since Generative AI covers a broad number of categories, let's take a look at some of the popular types of Generative AI that are currently making a buzz in the market. We start with ChatGPT where you're interacting with an AI through a chat box. You can ask any questions or request any information, and it is trained to respond back with an answer. ChatGPT is a type of Text based Generative AI. Similarly, there are other types of Generative AIs such as Images, Videos, Audios, and even Programming Codes. All these types are essentially using existing data or content and generating new content for the user based on the Machine Learning (ML) framework.

- 1) Image Generation
- 2) Text to Speech generation
- 3) Audio generation
- 4) Video generation
- 5) Data generation

2.3 Use Cases for Generative AI

Since there are various types of Generative AIs, we will explore various use cases pertaining to different industries separately:

- 1) Generate Image Use Case –
 - User enters text input describing what images they want, and the tool will process them to produce realistic images. Users can specify a subject, setting, style, object or location to the AI Tool, which will generate an output of an image based on your requirements. You can use these AI-generated images for Commercial purposes in media, design, advertising, marketing, education and more.
- 2) Create Music Use Case –
 - Generative AI can generate new music by learning the patterns and styles of input music and creating fresh compositions for advertisements. Just be aware of any copyrighted material in generated output that can lead to infringement issues.
- 3) Generate Text-to-Speech Use Case –
 - Create realistic speech audio with ability to even modify accents. You can even transform the voice into multiple different languages of your choice.
- 4) Generate Text Use Case –
 - Platforms like ChatGPT have become increasingly popular since their launch. Create content like articles, blogs, dialogues, translation, or even text for a website. Using the Natural Language Processing (NLP) and Natural Language Understanding (NLU), techniques of AI to read and understand text prompt and then respond back to users in an efficient way.
- 5) Creative Question Asking Use Case –
 - Let Generative AI create thought-provoking questions to stimulate your mind. It improves over time by incorporating previous answers into future generations of questioning.
- 6) Artificial Creativity Use Case –
 - Artificial creativity is different from image generation such that creativity seeks to create something totally new and original without human input. Instead of relying on a human artist to develop something unique and original, generative AI can do it all on its own!

7) Generate Videos Use Case –

- Generative AI can make videos by using image generation to create the visual content, text generation to create a script or storyboard, and music generation to create a soundtrack. It can take all sorts of input data (like images, blogs or articles, and music) and combine and manipulate it creatively to produce something new and unique.

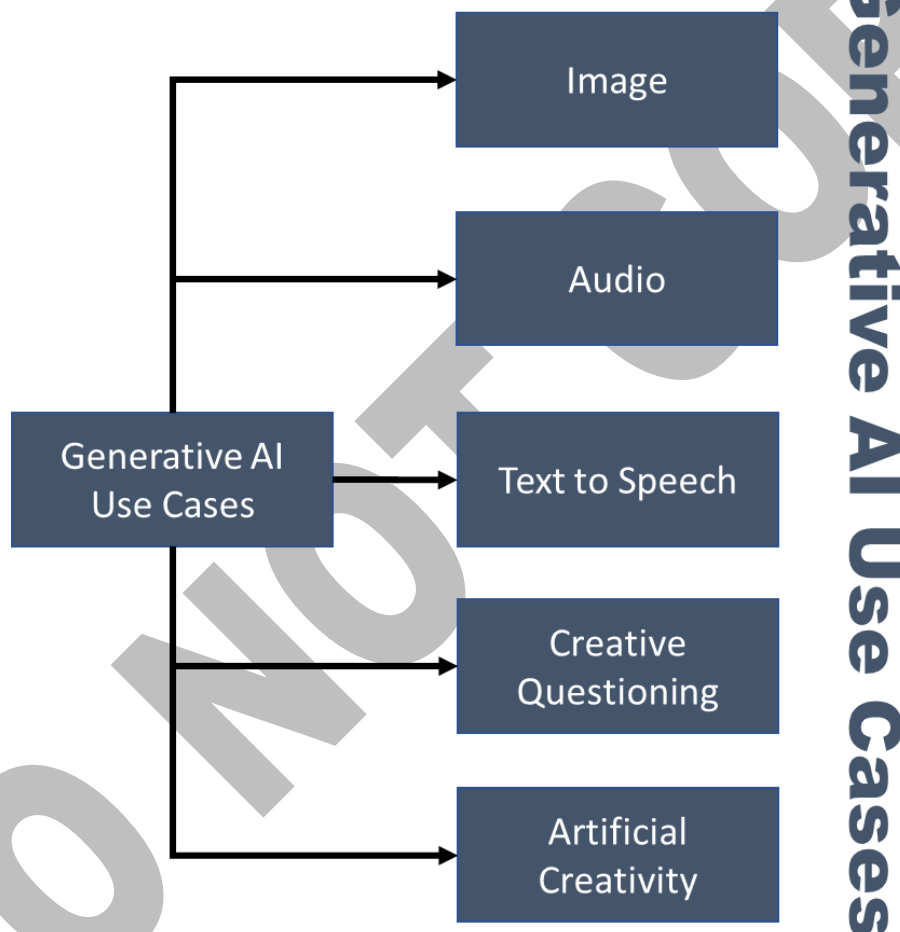


Figure 1d

2.4 Sectors of Generative AI

Generative AI has the potential to revolutionize several different sectors. These sectors have challenges that require human resources which they are either unable to afford or unable to obtain. Hence, Generative AI plays

a big lead in assisting these industries to solve certain challenges and improve product development process. Let us look at some of the popular sectors of Generative AI:

- 1) Logistics and Transportation – It goes without saying that companies like UPS, USPS, and FedEx rely on map locations and directions to navigate to their destinations.
- 2) Travel Industry – Face identification and photo verification systems at airports can make it easier for security to quickly differentiate travelers that can be of high risk.
- 3) Healthcare – CT scans and x-rays can be converted into helpful diagnosis by Generative AI. This can be especially useful for catching dangerous diseases like cancer in their early stages.
- 4) Marketing – Identifying and segmenting different groups to advertisements and marketing campaigns can be a valuable tool for companies targeting specific audiences and increasing sales

		Companies / Organizations using AI						
Industries/Sectors	Logistics	Alibaba Nike Amazon	GumGum EliseAI Affectiva	Hyatt Tripadvisor Kayak	Walmart Lowes Target	XPO DHL C.H. Robinson	UPS FedEx USPS	IBM Watson Health Google Health Butterfly
	Transportation					•	•	
	Travel			•				
	Healthcare							•
	Online Shopping	•						
	Marketing		•					
	Retail				•			

Figure 1e

2.5 Features of Generative AI

Generative AI has the potential to revolutionize several different industries. Some of the most successful AI Tools will combine and offer multiple capabilities to form a bigger ecosystem. At the end of the day, AI is simply an approach to problem solving but let us take a look as to what underlying features it has to offer.

- 1) Natural Language – There are over 6500 languages in the world, but less than 50% of these languages are currently represented online. AI bridges this gap for countries across the globe by taking human speech/text from one language, and translating that into a language of your choice combined with accent and tone. This is known as Natural Language Processing (NLP).
- 2) Predictions – AI uses large amounts of data and through processing and analyzing, it essentially predicts what will happen in future. This is drawn obviously from historical data but the more data AI can collect, the higher the probability of providing more accurate predictions. Many industries can use this capability to drive their own business decisions such as stock market, customer retentions, or even why an employee is likely to leave and why.
- 3) Personalization – When you're listing to music on Spotify, there are recommendations on new songs for you that you'll probably going to enjoy. These recommendations are personalization's based on how much the AI has learned about you as a person through choices and actions. This capability is widely used on many platforms such as Netflix, Amazon Shopping, social media and even other online shopping sites.
- 4) MLOps – MLOps stands for Machine Learning Operations and it is a core function of Machine Learning. It basically takes ML models to production and then maintains them. Google's MLOps promotes automation to deploy a production ready model with minimal risks quickly. This also guarantees quality and saves time.
- 5) Pre-trained APIs – Vertex AI offers pre-trained APIs for vision, video, natural language and more. You can integrate them easily into your existing applications or just spin a new application using that. So, you may not need to look for some of the other AI API platforms to get your work done.

Features of Generative AI

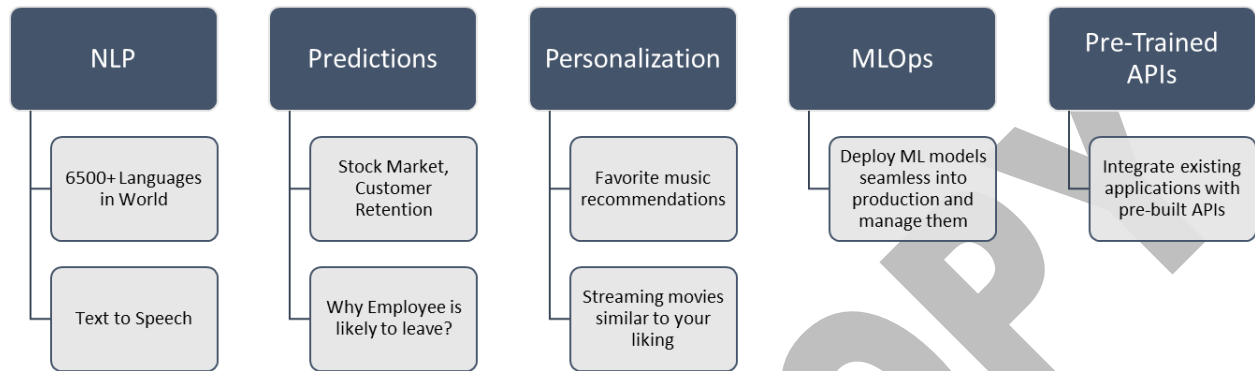


Figure 1f

2.6 Benefits of Generative AI

- 1) Boosted productivity — generative AI can help deal with repetitive tasks faster or, in some cases, completely automate them.
- 2) Reduced costs — by automating some processes, generative AI allows companies to function with fewer staff members, consequently reducing salary costs and expenses on office management and hardware and software maintenance.
- 3) Improved personalization — with generative AI solutions it takes just a few clicks to customize content such as visuals or offers to meet specific customers' needs.
- 4) Upgraded decision-making — generative AI can provide companies with valuable insights on different aspects of their processes, allowing them to find more effective and creative solutions.

2.6 Industries exploring opportunities with Generative AI

- 1) Finance can watch transactions in the context of an individual's history to build better fraud detection systems.
- 2) Legal firms can use generative AI to design and interpret contracts, analyze evidence and suggest arguments.
- 3) Manufacturers can use generative AI to combine data from cameras, X-ray and other metrics to identify defective parts and the root causes more accurately and economically.

- 4) Film and media companies can use generative AI to produce content more economically and translate it into other languages with the actors' own voices.
- 5) The medical industry can use generative AI to identify promising drug candidates more efficiently.
- 6) Architectural firms can use generative AI to design and adapt prototypes more quickly.
- 7) Gaming companies can use generative AI to design game content and levels.

3.0 Introduction to Machine Learning (ML)

Before you can create ML models with different platforms such as Vertex AI, one needs to have a basic understanding of what is a Machine Learning framework. Machine Learning relies on algorithms. Unless you are a data scientist or ML expert, these algorithms are very complicated to understand and work with. A machine learning framework, then, simplifies machine learning algorithms. An ML framework is any tool, interface, or library that lets you develop ML models easily, without understanding the underlying algorithms. These frameworks are oriented towards mathematics and statistical modeling (machine learning) as opposed to neural network training (deep learning). Here is a quick breakdown of some popular ML frameworks:

1. TensorFlow and PyTorch are direct competitors because of their similarity. They both provide a rich set of linear algebra tools, and they can run regression analysis.
2. Scikit-learn has been around a long time and would be most familiar to R programmers, but it comes with a big caveat: it is not built to run across a cluster.
3. Spark ML is built for running on a cluster, since that is what Apache Spark is all about.
4. Torch provides a little more complexity for functional programming but is also one of the easier ML frameworks

3.1 Understanding MLOps

Machine Learning Operations (MLOps) is a core function of Machine Learning and its job is to streamline the process of taking ML models to production and monitoring them. MLOps was designed based on the concept of DevOps, the existing practice of more efficiently writing, deploying, and managing enterprise applications. DevOps began as a way to unite software developers (the Devs) and IT operations teams (the Ops), destroying data silos and enabling better collaboration.

In a similar fashion, MLOps shares these aims but adds Data Scientists and ML engineers to the team. Data scientists curate datasets and analyze them by creating AI models for them. ML engineers are the people who use automated, disciplined processes to run the datasets through the models. Hence, MLOps is that deeply collaborative communication between Data Scientists and Ops (portions of the team focused on production or operations). ML operations intend to automate as much as possible, eliminate waste, and produce deeper, more consistent insights using machine learning.

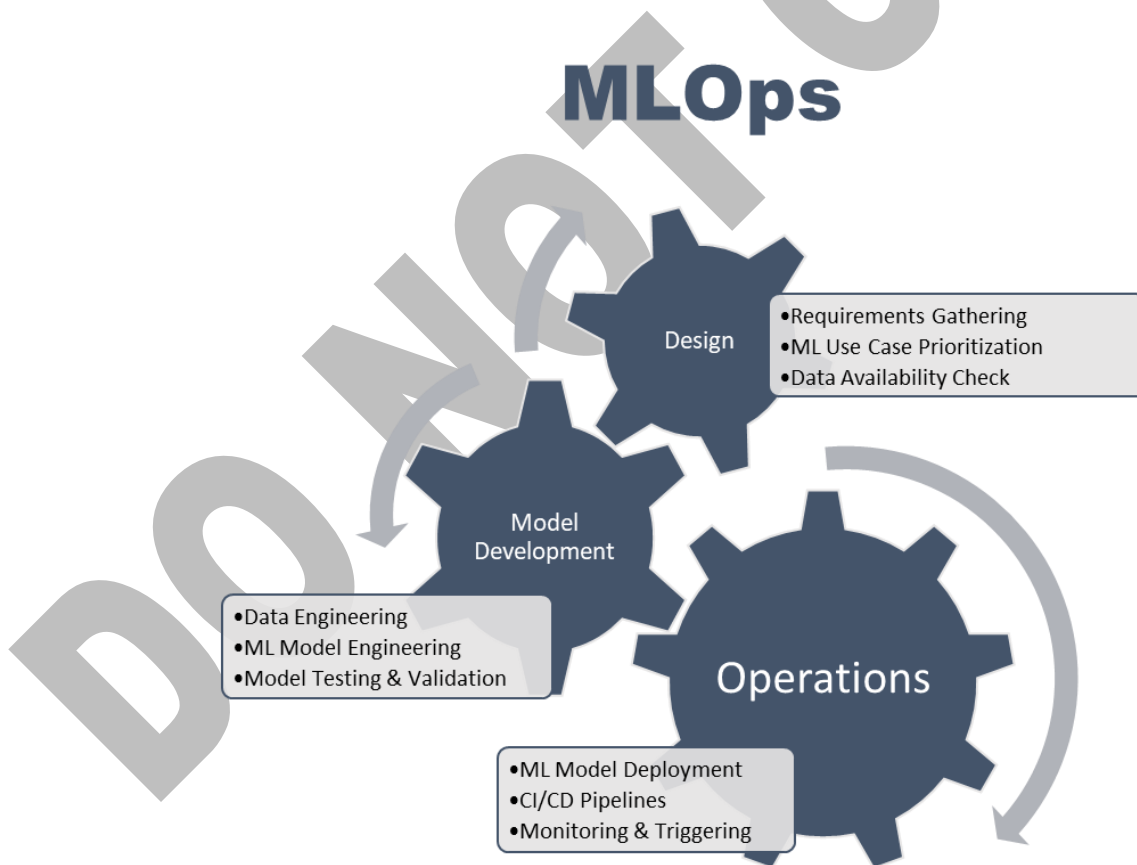


Figure 1g

3.2 Capabilities of MLOps

Some of the critical capabilities of MLOps that enable machine learning in production include:

- 1) **Simplified deployment** - Data scientists may use many different modeling frameworks, languages, and tools, which can complicate the deployment process. MLOps enables IT operations teams in production environments to deploy models more rapidly from various frameworks and languages.
- 2) **ML monitoring** - Software monitoring tools do not work for machine learning monitoring. In contrast, the monitoring that MLOps enables is designed for machine learning, providing model-specific metrics, detection of data drift for important features, and other core functionality.
- 3) **Life cycle management**. Deployment is merely the first step in a lengthy update lifecycle. To maintain a working ML model, the team must test the model and its updates without disrupting business applications; this is also the realm of MLOps.
- 4) **Compliance**. MLOps offers traceability, access control, and audit trails to minimize risk, prevent unwanted changes, and ensure regulatory compliance.

Capabilities of MLOps

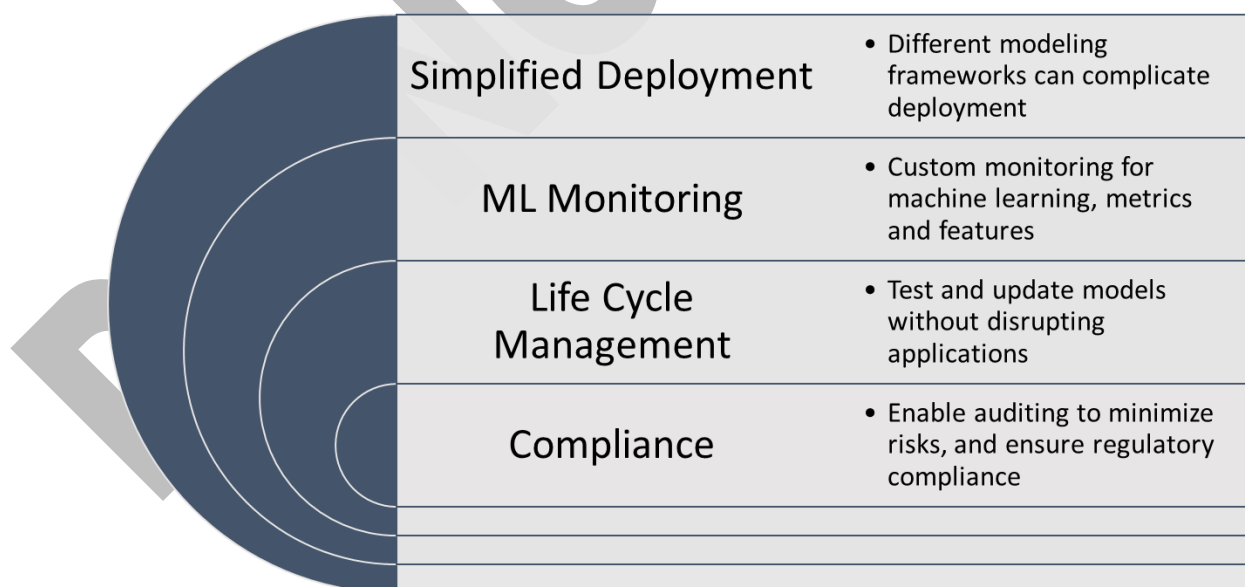


Figure 1h

3.3 Steps to Deploy ML Models into Production

In any ML project, the process of delivering an ML model to production involves the following steps.

- 1) Foundation. The team, including management and all stakeholders, defines the business use case for the data and establishes the success criteria for measuring model performance.
- 2) Data extraction. The data scientists on the team select relevant data from a range of sources and integrate it for the ML task.
- 3) Data analysis. Team data scientists perform exploratory data analysis (EDA) to analyze the data available for creating the ML model. The data analysis process allows the team to understand the characteristics and data schema the model will expect. It also enables the team to identify which feature engineering and data preparation the model needs.
- 4) Data preparation. The team prepares the data for the ML task, producing the data splits in their prepared format as the output. Data preparation involves applying feature engineering and data transformation to the model, and conducting data cleaning, in which the data scientist divides the data into sets for validation, training, and testing.
- 5) These data science steps allow the team to see what the data looks like, where it originates, and what it can predict. Next, the model operations life cycle, often managed by machine learning engineers, begins.
- 6) Model training. To train different ML models, the data scientist implements various algorithms with the prepared data. Next, to achieve the best performance from the ML model, the team conducts hyperparameter tuning on the implemented algorithms.
- 7) Model evaluation. The team evaluates the quality of the trained model on a holdout test set. This produces a set of metrics for assessing model quality as output.
- 8) Model validation. Next, the team confirms that the model's predictive power exceeds a specific baseline, making it adequate for deployment.

- 9) Model serving. The team deploys the validated model to a target environment to serve predictions as part of a batch prediction system; as an embedded model for a mobile device or edge device; or to serve online predictions as web services or microservices with a REST API.
- 10) Model monitoring. The team monitors the predictive performance of the model to determine when to invoke a new iteration.

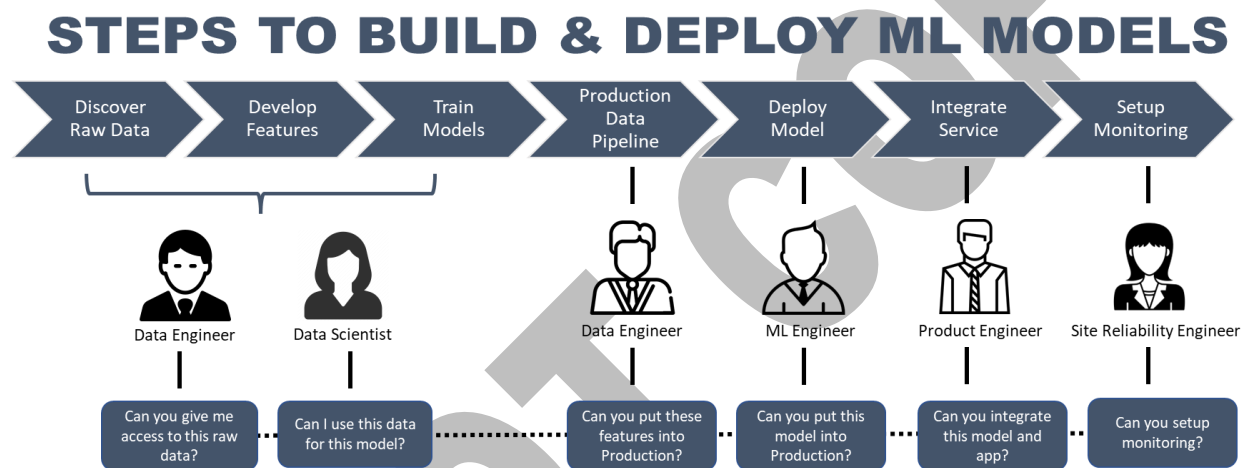


Figure 1i

4.0 Introduction to GCP Vertex AI

There are several choices of platforms that do Machine Learning (ML) and each platform offers advantages and disadvantages. However, to work with large amounts of data and creating ML models can be extremely time consuming and require experts and hardware resources. In addition, you are running millions of images or content to train your algorithms. So, Google provides a solution known as Vertex AI where even novice users can scale ML models faster with a unified approach.

With Vertex AI, users no longer need to have the knowledge of code to do custom ML modeling. Developers can deploy models faster with very little lines of code. Ultimately, Vertex AI is a complete ecosystem where it has its own pre-built ML models for Text, Image, Forecasting and even Video. Small or large enterprises can access Vertex AI by importing their own datasets and deploying their own models to carryout predictions for business decisions. Overall, Vertex AI is a unified platform that provides a

simplified approach to generate instant insights through fast deployment of ML models.

4.1 Features of Vertex AI

So now that we understand ML framework, we can now explore the features of Vertex AI which essentially is an ML framework. Below are some popular features that Vertex AI brings to Generative AI:

- 1) Supports all Open Source Frameworks – Users want choices and they don't want to be bounded to one platform. One of the key features of Vertex AI is that it is compatible with other open-source ML frameworks like TensorFlow, SparkML, Scikit, Huggingface or PyTorch.
- 2) Unified User Interface (UI) for entire ML workflow - It brings together the Google Cloud services for building ML under one unified UI and API. You can efficiently train and compare models using AutoML or custom code training. A central model repository stores all your models that can deploy to the same endpoints.
- 3) Pre-trained APIs - Vertex AI has pre-trained APIs for vision, natural language, video, among others. You can easily incorporate them into existing applications. You can also build new applications across use cases such as Translation and Speech to Text.
- 4) Integration with BigQuery Serverless Database – This is one of Google's cloud based serverless database that allows petabytes of data with a built-in query engine. Vertex AI allows you to leverage BigQuery

Features of Generative AI

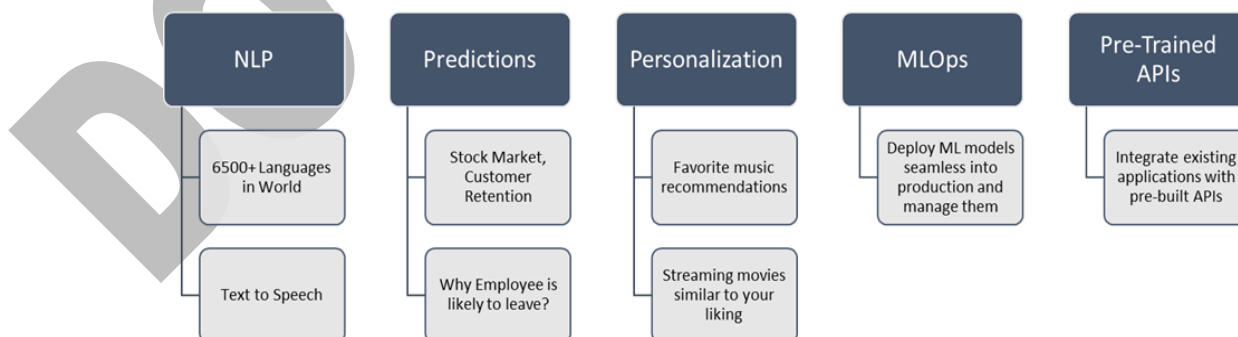


Figure 1j

4.2 Use Cases for Vertex AI

Several organizations and companies are using Vertex AI:

- 1) **Collecting Point of Sales Data Use Case** – Coca Cola is also ramping up its ML efforts, using Vertex AI and BigQuery to process billions of data records from 700,000 vending machines, helping the company to make strategic decisions about when and where to locate products. Create a prediction model of where to place vending machines, what products are lined up in the machines, at what price, and how much they will sell.
- 2) **Eliminate fatal accidents caused by Cars** – Vertex AI can be used to prevent fatal car accidents. The scientists found that a highway road, for example, has a higher risk than nearby residential roads, and ramps merging and exiting the highway have an even higher risk than other roads. Capturing high-risk locations can enable cities and states to build better road system.
- 3) **Forecasting Inventory in Retail Sector** – Sales can be region and city specific. Big chain stores like Lowes & Home Depot have to rely to on thousands of SKUs that are being entered into their system. It is crucial for these large chain stores to understand the performance of these SKUs, demographics on who are consumers, and inventory wastage. With this information, Vertex AI can create predictive models that can help pinpoint the shortcomings for these retail stores and improve their forecast.
- 4) **Fraud Detection in Banking Applications** - AI helps banks deal with fraud in many ways. It can improve their ability to detect fraud in real time, and it can reduce false positives which boosts accuracy and safeguards the customer experience.

4.3 Benefits of Vertex AI

- 1) **Spend less time and money on Infrastructure** - Because you only pay for what you use with Vertex AI, you can lower the cost of building your infrastructure. Moreover, it should make it much easier to set up or manage your infrastructure to train machine learning models. Hence even novice analysts or researchers can contribute to some progress.
- 2) **Ideal for Professionals and Novices** - Given that many tasks may be completed without writing (or configuring) them, it is an excellent option for novices. Machine learning models may be trained and used

with a short learning curve. Also, it saves time for professionals by giving them access to numerous managed tools and pre-trained APIs to do fundamental operations and activities.

- 3) **Help solve Complex Tasks** - It manages complex problems, such as running billions of iterations of a model to determine the best algorithm. Google believes that its unified approach to ML development allows users to speed up ML development and experimentation by 5%. In addition, it reduces the lines of code needing to be written by 80%.
- 4) **Reduce Risks for Production Deployment** – Vertex AI is equipped with user-managed notebooks instances that are protected by Google Cloud authorization and authentication, prevents unauthorized access to ML projects. These managed notebooks are prepackaged and setup within Jupyter notebook and will also sync with GitHub's data repositories.

5.0 Summary

There is every indication that Generative AI will establish itself as one of the most vital technologies a business should leverage to stay competitive and achieve success. Since the technology is highly versatile and can be customized to the needs of specific departments or companies, we will definitely see lots of real-world applications of Generative AI in the future.

Until now, when training algorithms, millions of test images, videos, or languages had to be run. Today, thanks to Vertex AI, this whole process can be simplified as it is the platform that does the heavy lifting. This technology has the means to perform the calculations and solve the most complex problems, being able to make billions of iterations and create the best algorithms. In a nutshell, Vertex AI is a one-stop shop for data scientists, providing all of the tools they'll need to manage, construct, deploy, interpret, and monitor their models. Newbies and experts alike can instantly start utilizing Vertex AI without any kind of formal machine learning training. Vertex AI is a powerful Google AI & AutoML platform that offers a lot of potential for any company that has been trying to get true benefit from their machine learning efforts. With its user-friendly interface, it is ideal choice everybody from novices to professional data scientists regardless of their level of experience. In addition to a central repository where you can manage the lifecycle of your ML models, it introduces new ways to work with models you've trained outside of Vertex AI. With all these benefits of the Vertex AI Model Registry, you can confidently move your best models to production faster.

5.1 References

“Get Started.” Google Cloud, <https://cloud.google.com/vertex-ai/docs/start>. Accessed May 12, 2023.

Tappen, Henry. “How Businesses Use Google Cloud VertexAI.” <https://cloud.google.com/blog/products/ai-machine-learning/how-businesses-use-google-cloud-vertex-ai>. Accessed May 12, 2023.

Mahendra, Sanksshep. “Democratizing Artificial Intelligence.”, <https://www.aiplusinfo.com/blog/democratizing-artificial-intelligence/>. Accessed May 12, 2023.

Tech, Google Cloud. “What Is Vertex AI?” YouTube, <https://www.youtube.com/watch?v=gT4qqHMiEpA>. Accessed May 12, 2023.

Campbell, Pryor. “Google Vertex AI: A Powerful Tool to Solve Your Machine Learning Woes.” <https://www.contino.io/insights/google-vertex-ai>. Accessed May 12, 2023.

“Background: What is a Generative Model?”. Google Cloud, <https://developers.google.com/machine-learning/gan/generative>. Accessed May 12, 2023.

Tech. “MLOps Explained: A Complete Introduction”, <https://www.arrikto.com/mlops-explained/>. Accessed May 12, 2023.

Tech. “MLOps: Step by Step Guidelines for Beginners”. <https://guide4info.com/mlops-for-dummies>. Accessed May 12, 2023.