JON BONSO, KAYNE RODRIGO,
& SAMANTHA SERVO

AWS CERTIFIED
AI PRACTITIONER
AIF-C01
EXAM

Tutorials Dojo Study Guide

TABLE OF CONTENTS

INTRODUCTION

We live in an era where artificial intelligence (AI) and machine learning (ML) are revolutionizing industries and enhancing human capabilities. AWS, as a pioneer in cloud computing, offers an extensive array of AI/ML services that empower businesses to innovate and thrive. The AWS Certified AI Practitioner (AIF-C01) certification is designed to validate your understanding of AI, ML, and generative AI concepts and their applications on AWS, as well as your ability to identify appropriate AWS services to implement AI solutions. This certification is ideal for professionals who are familiar with AI/ML technologies and use them in their roles but do not necessarily build AI/ML solutions themselves.

AWS has consistently been recognized as the leading cloud provider in the market[1]. They are committed to continuously enhancing their services to ensure customer success and satisfaction. By earning the AWS Certified AI Practitioner certification, you demonstrate your expertise in identifying appropriate AI solutions based on specific business needs, which can significantly elevate your career prospects. This certification signals to employers your proficiency in leveraging AWS AI services to drive innovation and efficiency. It also positions you for career growth and higher earnings, as employers are willing to pay a premium for AI-skilled workers.

The journey to mastering AI and ML on AWS is both challenging and rewarding. The AWS Certified AI Practitioner credential serves as a robust foundation for further specializations in AI and ML. Whether you aim to become a data scientist, ML engineer, or other AI/ML-focused roles, this certification will equip you with the knowledge and skills necessary to excel in your chosen field. The credential also opens doors to new career opportunities and professional growth in a rapidly evolving technological landscape. AWS provides comprehensive resources and training to help you prepare for the exam and achieve your certification goals.

This Study Guide eBook is designed to provide you with the knowledge and practical skills needed to pass the AWS Certified AI Practitioner (AIF-C01) exam. It includes essential concepts, exam domains, tips for success, sample questions, cheat sheets, and other relevant information about the AIF-C01 exam. The guide begins with an overview of the exam structure, offering insights into the types of questions, the exam domains, the scoring scheme, and the benefits of passing the exam.

The contents are organized according to the official AIF-C01 exam guide, covering all pertinent AWS topics for each exam domain. The guide discusses various AWS concepts, related AI/ML services, and technical implementations to provide a clear understanding of what to expect on the actual exam.

Note: We've created these study guides and cheat sheets as supplementary resources. To enhance your exam readiness, we recommend using them alongside our high-quality practice exams, which will help you assess your knowledge and identify areas for improvement.

1 https://aws.amazon.com/blogs/aws/aws-named-as-a-leader-in-gartners-infrastructure-as-a-service-iaas-magic-quadrant-for-the-9th-consecutiveyear/

AWS CERTIFIED AI PRACTITIONER AIF-C01 EXAM OVERVIEW

Amazon Web Services began its Global Certification Program in 2013 with the primary purpose of validating the technical skills and knowledge of IT Professionals in building secure and reliable cloud-based applications using the AWS Cloud. In April 2013, AWS launched its first-ever AWS Certification test called the AWS Certified Solutions Architect Associate exam. This was followed by the AWS Certified SysOps Administrator and AWS Certified Developer Associate exams.

Amazon has been continuously expanding and updating its certification program year after year. They launched a series of Professional and Specialty-level certifications that cover various topics like DevOps, machine learning, data analytics, advanced networking, and many others. As the number of AWS services increases, a new and updated version of the AWS certification exam is released regularly to reflect the recent service changes and include the new knowledge areas.

In August 2024, AWS launched the AWS Certified AI Practitioner (AIF-C01) exam, a foundational certification for individuals who are familiar with AI/ML technologies on AWS. The AI Practitioner exam evaluates your understanding of concepts and use cases related to artificial intelligence (AI), machine learning (ML), and generative AI. This certification is ideal for roles such as business analysts, IT support, marketing professionals, product or project managers, and sales professionals. The exam has a duration of 90 minutes and consists of 65 questions. There are no prerequisites, allowing you to take the exam without the need for any prior certification, degree, or training.

The exam contains a mixture of scenario-based and multiple-choice questions, including multiple-response formats. The scenario-based questions have one correct answer and three incorrect responses, while the multiple-response format requires you to select two or more correct answers out of five or more options. Additionally, AWS has introduced three new question types: Ordering, where you get a list of 3 to 5 responses to complete a specific task, and you need to select the right ones and place them in the correct order; Matching, where you get a list of responses to match to 3 to 7 prompts; and Case Study, which involves detailed real-world scenarios requiring in-depth analysis and application of AWS AI principles. The exam costs **100 USD** and can be taken either at a local testing center or online from the comfort of your home.

The AI Practitioner certification exam has a total of 65 questions that you should complete within 90 minutes or one hour and a half. The score range for this test is from 100 to 1,000, with a minimum passing score of 700. AWS is using a scaled scoring model to equate scores across multiple exam types that may have different difficulty levels. An email of your result will be sent to you after a few days, and the complete score report will be available to your AWS Certification account afterward.

Individuals who unfortunately did not pass the AWS exam must wait for 14 days before they are allowed to retake the exam. There is no hard limit on the number of exam attempts, so you can try again and again until you pass the exam. Take note that on each attempt, the full registration price of the exam must be paid.

Your AWS Certification Account will have a record of your complete exam results within 5 business days of completing your exam. The score report contains a table of your performance for each exam domain, which indicates whether you met the competency level required for these domains or not. AWS uses a compensatory scoring model, which means that you do not necessarily need to pass each and every individual section.

You will pass this exam as long as you get an overall score of 700 across 4 domains. Each section has a specific score weighting that translates to the number of questions; hence, some sections have more questions than others. Your Score Performance table highlights your strengths and weaknesses that you need to improve on.

Exam Details

The AWS Certified AI Practitioner (AIF-C01) exam is designed for individuals who can effectively demonstrate a comprehensive understanding of AI/ML, generative AI technologies, and related AWS services and tools, regardless of their specific job role.

The exam includes various question types such as multiple-choice questions, where you select one correct response out of four options, and multiple-response questions, where you choose two or more correct responses out of five or more options. Additionally, it features ordering questions that require arranging responses in the correct order to complete a specified task, matching questions that involve pairing responses to a set of prompts correctly, and case study questions based on a scenario with two or more related questions. Each question is evaluated separately, and you receive credit for each correct response. You have the flexibility to take the exam either via online proctoring or at a testing center near you.

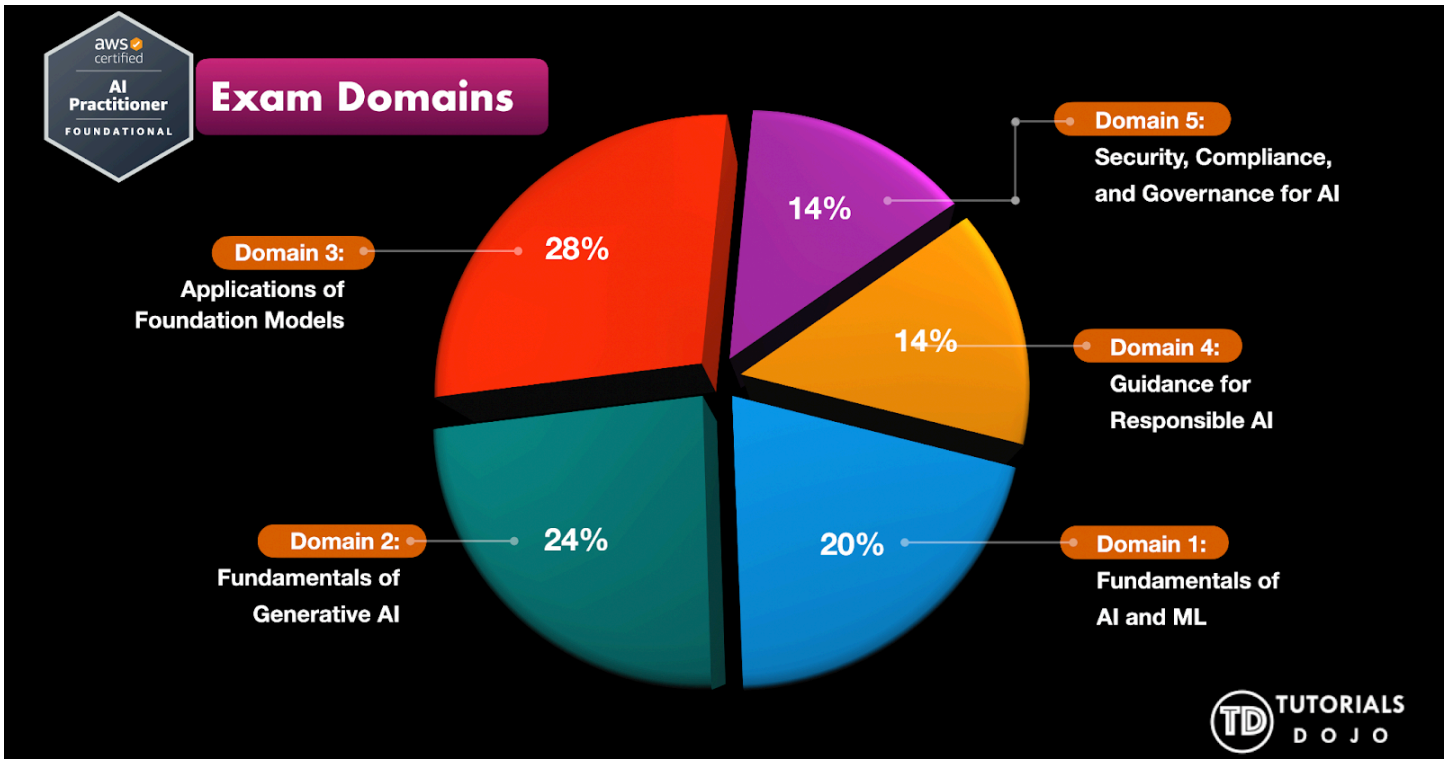
Exam Code:	AIF-C01
Prerequisites:	None
No. of Questions:	65
Score Range:	100-1000
Cost:	100 USD
Passing Score:	700
Time Limit:	90 minutes

Exam Domains

The **AWS Certified AI Practitioner (AIF-C01)** exam has five different domains, each with a corresponding weight and topic coverage. The domains are as follows:

- Domain 1: Fundamentals of AI and ML (20%)
- Domain 2: Fundamentals of Generative AI (24%)
- Domain 3: Applications of Foundation Models (28%)
- Domain 4: Guidelines for Responsible AI (14%)
- Domain 5: Security, Compliance, and Governance for AI Solutions (14%)

One exam domain is comprised of several task statements. A task statement is a sub-category of the exam domain that contains the required cloud concepts, knowledge, and skills for you to accomplish a particular task or activity in AWS. In the AWS Certified AI Practitioner (AIF-C01) test, the **Domain 3: Applications of Foundation Models** has the biggest weighting in the exam at 28%, so expect to see a lot of foundation models-related scenarios in the exam. Conversely, the exam domain with the least exam weighting is **Domain 4: Guidelines for Responsible AI** and **Domain 5: Security, Compliance, and Governance for AI Solutions** so you have to limit the time you spend studying under this knowledge area.



Domain 1: Fundamentals of AI and ML

1.1. Explain basic AI concepts and terminologies.
1.2. Identify practical use cases for AI.
1.3. Describe the ML development lifecycle.

Domain 2: Fundamentals of Generative AI

2.1. Explain the basic concepts of generative AI.
2.2. Understand the capabilities and limitations of generative AI for solving business problems.
2.3. Describe AWS infrastructure and technologies for building generative AI applications.

Domain 3: Applications of Foundation Models

3.1. Describe design considerations for applications that use foundation models.
3.2. Choose effective prompt engineering techniques.
3.3. Describe the training and fine-tuning process for foundation models.
3.4. Describe methods to evaluate foundation model performance.

Domain 4: Guidelines for Responsible AI

4.1. Explain the development of AI systems that are responsible.
4.2. Recognize the importance of transparent and explainable models.

Domain 5: Security, Compliance, and Governance for AI Solutions

5.1. Explain methods to secure AI systems.
5.2. Recognize governance and compliance regulations for AI systems.

Exam-Related AWS Topics and Services

The official exam guide contains a list of key tools, technologies, and concepts that may show up on the AI Practitioner AIF-C01 test. Keep in mind that this is just a non-exhaustive list of the tools and technologies that may or may not appear on the exam. This list can change at any time and is primarily given to test-takers to help them understand the general scope of services, features, or technologies for this certification. In addition, the general tools and technologies in this list appear in no particular order.

Here are the topics, AWS services, and concepts that you should focus on for your upcoming exam. You have to review your knowledge on the following:

• AI/ML • AI/ML Use Cases and Applications • Generative AI technologies • Types of AI/ML technologies • AWS shared responsibility model • Management and governance	• Amazon Translate • Amazon Comprehend • Amazon Lex • Amazon Polly • Amazon Bedrock • Amazon SageMaker AI

• AWS global infrastructure • Prompt Engineering • Foundation model • Foundation model lifecycle • Advantages of generative AI (for example, adaptability, responsiveness, simplicity). • Disadvantages of generative AI solutions (for example, hallucinations, interpretability, inaccuracy, nondeterminism). • Retrieval Augmented Generation (RAG) • Recall-Oriented Understudy for Gisting Evaluation [ROUGE] • Bilingual Evaluation Understudy [BLEU] • BERTScore • Features of responsible AI • Amazon Transcribe	• Amazon SageMaker JumpStart • Amazon SageMaker Clarify • Amazon SageMaker Data Wrangler • Amazon SageMaker Model Monitor • Amazon SageMaker Feature Store • Amazon Augmented AI [Amazon A2I] • Amazon SageMaker Model Cards • Amazon Macie • Data cataloging • International Organization for Standardization [ISO] • System and Organization Controls [SOC] • Data lifecycles • AWS Support plans • AWS Well-Architected Framework

Remember that out of the 5 exam domains, the **Applications of Foundation Models** domain has the biggest coverage in the exam, at 28 percent. This means that more than a quarter of the questions in the entire AWS Certified AI Practitioner exam cover the applications of foundation models.

The Appendix section of the exam guide also includes a list of relevant AWS services that you should focus on, so in your exam, make sure that you review the following AWS services.

When it comes to Analytics, ensure you study AWS Data Exchange, Amazon EMR, AWS Glue, AWS Glue DataBrew, AWS Lake Formation, Amazon OpenSearch Service, Amazon QuickSight, and Amazon Redshift.

For Cloud Financial Management, Compute, and Containers, you should learn about AWS Budgets, AWS Cost Explorer, Amazon EC2, Amazon Elastic Container Service (Amazon ECS), and Amazon Elastic Kubernetes Service (Amazon EKS).

In the Databases category, focus on Amazon DocumentDB (with MongoDB compatibility), Amazon DynamoDB, Amazon ElastiCache, Amazon MemoryDB, Amazon Neptune, and Amazon RDS.

The Machine Learning section includes essential services such as Amazon Augmented AI (Amazon A2I), Amazon Bedrock, Amazon Comprehend, Amazon Fraud Detector, Amazon Kendra, Amazon Lex, Amazon Personalize, Amazon Polly, Amazon Q, Amazon Rekognition, Amazon SageMaker AI, Amazon Textract, Amazon Transcribe, and Amazon Translate.

The AI Practitioner exam covers a handful of services related to Management and Governance, including AWS CloudTrail, Amazon CloudWatch, AWS Config, AWS Trusted Advisor, and AWS Well-Architected Tool.

For the Security, Identity, and Compliance category, prepare to see a range of AWS services that you can use to secure your enterprise applications and AWS resources. These include AWS Artifact, AWS Audit Manager, AWS Identity and Access Management (IAM), Amazon Inspector, AWS Key Management Service (AWS KMS), Amazon Macie, and AWS Secrets Manager. Pay attention to how these services work together and know the appropriate AWS service to use for a particular use cases or situation.

Lastly, for Networking and Content Delivery, make sure to understand Amazon CloudFront and Amazon VPC. For Storage, review Amazon S3 and Amazon S3 Glacier.

Exam Scoring System

You can get a score from 100 to 1,000 with a minimum passing score of **700** when you take the AWS Certified AI Practitioner exam. AWS uses a scaled scoring model to associate scores across multiple exam types that may have different levels of difficulty. Your complete score report will be sent to you by email 1 - 5 business days after your exam.

For individuals who unfortunately do not pass their exams, you must wait 14 days before you are allowed to retake the exam. There is no hard limit on the number of attempts you can retake an exam. Once you pass, you'll receive various benefits such as a discount coupon which you can use for your next AWS exam.

Once you receive your score report via email, the result should also be saved in your AWS Certification account already. The score report contains a table of your performance on each domain and it will indicate whether you have met the level of competency required for these domains. Take note that you do not need to achieve competency in all domains for you to pass the exam. At the end of the report, there will be a score performance table that highlights your strengths and weaknesses which will help you determine the areas you need to improve on.

		Score Performance	
Section	**% of Scored Items**	**Needs Improvement**	**Meets Competencies**
Domain 1: Fundamentals of AI and ML	20%		
Domain 2: Fundamentals of Generative AI	24%		
Domain 3: Applications of Foundation Models	28%		
Domain 4: Guidelines for Responsible AI	14%		
Domain 5: Security, Compliance, and Governance for AI Solutions	14%		

Exam Benefits

If you successfully passed any AWS exam, you will be eligible for the following benefits:

- **Exam Discount** - You'll get a 50% discount voucher that you can apply for your recertification or any other exam you plan to pursue. To access your discount voucher code, go to the "Benefits" section of your AWS Certification Account, and apply the voucher when you register for your next exam.

- **Certification Digital Badges** - You can showcase your achievements to your colleagues and employers with digital badges on your email signatures, Linkedin profile, or on your social media accounts. You can also show your Digital Badge to gain exclusive access to Certification Lounges at AWS re:Invent, regional Appreciation Receptions, and select AWS Summit events. To view your badges, simply go to the "Digital Badges" section of your AWS Certification Account.

You can visit the official AWS Certification FAQ page to view the frequently asked questions about getting AWS Certified and other information about the AWS Certification: https://aws.amazon.com/certification/faqs/.

AWS CERTIFIED AI PRACTITIONER (AIF-C01) EXAM STUDY GUIDE

The AWS Certified AI Practitioner or AWS AIF-C01 exam is designed to assess the candidate's understanding of AI and machine learning concepts within the AWS environment. This certification covers most, if not all, fundamental knowledge that one should know when venturing into the Cloud, together with AI/ML applications. The AWS AIF-C01 course intends to provide practitioners with a fundamental understanding of the AWS Cloud without having to dive deep into the technicalities. This includes the AWS Global Infrastructure, best practices in using AWS Cloud, pricing models, technical support options, and many more. You can view the complete details and guidelines for the certification exam here.

In addition to core cloud concepts, the exam emphasizes AI/ML services and solutions available through AWS, such as Amazon SageMaker AI, Rekognition, and Polly. It also touches on the basics of machine learning workflows, from data collection to model deployment and monitoring. A candidate will be tested on their ability to apply AI/ML models in real-world scenarios, leveraging AWS tools to solve business challenges. While no deep programming expertise is required, a solid understanding of machine learning principles, data handling, and AWS services is essential to passing the exam. This certification is ideal for those aiming to advance their careers in AI, machine learning, or cloud computing while learning about AWS's vast array of AI solutions.

What to review

1. AWS AI and ML Services

Familiarize yourself with the wide range of AWS services designed for artificial intelligence (AI) and machine learning (ML). Understand their key features and use cases, and learn how these services can help you in building, deploying, and managing AI and ML solutions efficiently. The available services include tools for data preparation, model building, training, deployment, and inference. Additionally, understanding the AWS Cloud and its various services is crucial. This includes:

- Core AWS services like Amazon EC2, Amazon S3, AWS Lambda, and Amazon SageMaker AI, and their operational applications.
- The AWS shared responsibility model for security and compliance.
- AWS Identity and Access Management (AWS IAM) for managing security and access to AWS resources.
- The AWS global infrastructure, including the concepts of AWS Regions, Availability Zones, and edge locations.
- AWS services pricing models.

To quickly view over the different categories, you may visit this link. For a comprehensive introduction, this AWS whitepaper contains an overview of the different AWS services along with their definitions and use cases.

2. Fundamentals of Artificial Intelligence (AI) and Machine Learning (ML)

Grasp the foundational concepts of artificial intelligence and machine learning. Understand the differences between supervised, unsupervised, and reinforcement learning, as well as common algorithms. Learn how these technologies are applied to solve various problems across different domains.

3. Prompt Engineering

Delve into the art of prompt engineering to effectively customize AI models. Learn about different types of prompts and how to use prompt templates to enhance model responses without changing the underlying model architecture.

4. Data Wrangling and Preprocessing

Understand the importance of data preparation in the machine learning pipeline. Discover how AWS services can simplify and automate data wrangling tasks to improve the quality and reliability of your machine learning models.

5. Model Training and Evaluation

Gain insights into training machine learning models, selecting appropriate algorithms, tuning hyperparameters, and evaluating model performance. Learn key metrics like accuracy, precision, recall, and F1 score to evaluate your models' performance.

6. Deployment and Operationalization

Study the process of deploying machine learning models into production environments. Learn about best practices for monitoring, maintaining, and scaling deployed models to ensure they perform well over time.

7. Ethical and Responsible AI

Understand the ethical considerations in AI, such as fairness, transparency, and bias detection. Learn best practices for implementing AI solutions responsibly to mitigate ethical risks.

8. Real-World Applications

Examine real-world case studies and examples of how various industries are leveraging AWS AI and ML technologies. This will help you understand practical applications and the impact of these technologies in different sectors.

How to review

As with any exam, the very first step is always the same - **KNOWING WHAT TO STUDY**. Although we have already enumerated them in the previous section, I highly suggest you go over the AWS Certified AI Practioner AIF-C01 Exam Guide again and see the exam contents.

AWS already has a vast number of free resources available for you to prepare for the exam. I suggest you first read Overview of Amazon Web Services whitepaper, and gain a good understanding of the different AWS concepts and services. Check out the amazing Tutorials Dojo cheat sheets to supplement your review for this section. Also check out this article: **Top 5 FREE AWS Review Materials**.

Beyond understanding the core AWS services, it's essential to dive into AWS's AI/ML offerings. Familiarize yourself with services like Amazon SageMaker AI, Amazon Bedrock, Amazon Rekognition, Amazon Lex, Amazon Polly, and many more, as these are crucial for the exam.

AWS also offers a comprehensive digital course on Machine Learning, consisting of a collection of free courses designed to enhance your understanding and skills. This collection includes:

- AWS Machine Learning Services Overview
- AWS DeepRacer
- Data Science
- Math for Machine Learning
- MLS-C01 Exam Readiness

These courses not only cover the MLS-C01 exam topics but also provide valuable insights for the AWS Certified Machine Learning Engineer – Associate MLA-C01 and AWS Certified AI Practitioner AIF-C01 certifications.

Common Exam Scenarios

Scenario	Solution
Domain 1: Fundamentals of AI and ML	
A branch of computer science focused on addressing cognitive challenges typically linked to human intelligence.	Artificial Intelligence
It is a type of AI focused on developing methods that enable machines to learn and understand.	Machine Learning
It is an AI technique that enables computers to process data in ways that mimic the human brain.	Deep Learning

It is a technique of artificial intelligence that teaches computers to interpret data in a way inspired by the human brain.	Neural Networks
It allows machines to identify people, places, and objects in photos with accuracy similar to humans, all while operating much faster and more efficiently.	Computer Vision
.Analyzing data sets to summarize their main characteristics, frequently using visualizations to reveal patterns, trends, and relationships.	Exploratory Data Analysis (EDA)
It is a branch of artificial intelligence that focuses on how computers interact with human languages.	Natural Language Processing (NLP)
A performance metric for classification models that indicates the model's effectiveness in distinguishing between classes at different thresholds.	Area Under the ROC Curve (AUC)
Domain 2: Fundamentals of Generative AI	
The process of guiding generative artificial intelligence (generative AI) to produce specific outputs.	Prompt Engineering
These models are created to handle inputs from multiple sources, such as text, images, audio, and video.	Multi-modal Models
It involves training an already established model on a new dataset instead of starting from the beginning. This technique, known as transfer learning, can produce reliable models from smaller datasets and requires less training time.	Fine-tuning
A numerical representation of real-world objects used by machine learning (ML) and artificial intelligence (AI) systems to comprehend complex knowledge domains like humans do.	Embedding
Generative AI is capable of adapting to various activities and domains by learning from data and creating content tailored to specific situations or needs. Its flexibility allows it to be applied across a wide range of sectors.	Adaptability

It generates content in real-time, leading to quicker responses and more dynamic interactions. This is particularly advantageous for chatbots, virtual assistants, and other interactive applications that require immediate feedback.	Responsiveness
AI language models can simplify challenging tasks by automating content generation processes. For instance, they can produce text that resembles human writing, thus reducing the time and effort required for content development.	Simplicity
With JumpStart, you can quickly evaluate, compare, and choose FMs for tasks like summarizing articles and creating images based on established quality and accountability criteria.	Amazon SageMaker JumpStart
It is a fully managed solution that provides access to high-performing foundation models (FMs) from leading AI startups and Amazon through a common API.	Amazon Bedrock
It is an Amazon Bedrock Playground that enables users to easily and intuitively build generative AI applications. The platform offers a fun, hands-on environment where users can create a variety of AI-driven applications in just a few steps.	PartyRock
Domain 3: Application of Foundation Models	
It is a technique used to help a model generalize from a few examples. The model leverages these examples to make more accurate predictions without the need for re-training or fine-tuning.	Few-shot Prompt Engineering
It is a method for customizing a pre-trained FM by fine-tuning the model on a specific task or domain-specific information.	Domain adaptation fine-tuning
A technique that utilizes labeled examples to enhance the performance of a model for a specific task.	Instruction-based fine-tuning
One continuously improves their model by analyzing feedback from earlier versions. In reinforcement	Reinforcement learning

learning, an agent learns through trial and error while interacting with its environment.	
It is a method to optimize the output of a large language model (LLM) by referencing a knowledge base containing company-specific or industry-specific data.	Retrieval Augmented Generation
A managed service for search, monitoring, and data analysis that provides real-time search and analytics for various applications.	Amazon OpenSearch Service
A relational database compatible with MySQL and PostgreSQL, designed for high performance and availability while supporting complex applications.	Amazon Aurora
A fully managed graph database service designed for efficient storage and querying of highly connected datasets, making it perfect for graph-based applications.	Amazon Neptune
In generative models, these prompts indicate which content to exclude from the generated output.	Negative prompts
The conceptual space where machine learning models transform input data into feature representations that are used to generate outputs.	Model Latent Space
It is a technique that uses human feedback to help machine learning models make predictions more efficiently and accurately while maximizing rewards.	Reinforcement Learning from Human Feedback (RLHF)
A metric for assessing text summarization quality by comparing the overlap between produced and reference summaries.	Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
A metric used to evaluate the quality of machine-translated text by measuring the similarity between the machine output and human reference translations.	Bilingual Evaluation Understudy (BLEU)
A metric for assessing text generation models by comparing token-level similarities using BERT embeddings.	BERTScore

Domain 4: Guidelines for Responsible AI	
It outlines the procedures and principles that ensure AI systems are transparent and trustworthy, while also minimizing potential risks and negative effects.	Responsible AI
A tool for machine learning that offers insights to enhance model fairness and transparency by analyzing potential biases in datasets and model predictions.	Amazon SageMaker Clarify
A service that simplifies incorporating human review into machine learning predictions, ensuring high-quality outcomes.	Amazon Augmented AI (Amazon A2I)
Documents that describe important details about machine learning models, such as performance metrics, intended uses, and compliance information.	Amazon SageMaker Model Cards
Domain 5: Security, Compliance, and Governance for AI Solutions	
A fully managed data security and privacy solution utilizing machine learning to identify, categorize, and safeguard sensitive data in AWS.	Amazon Macie
A service that offers secure, private connectivity between VPCs and AWS services, ensuring scalable access to critical resources.	AWS PrivateLink
A security vulnerability occurs when malicious input is used in prompts to alter the output of language models.	Prompt Injection
A global organization that sets standards and provides guidelines for various industries to ensure quality, safety, and efficiency.	International Organization for Standardization (ISO)
A collection of reports that provide details about the controls at a service organization, related to security, availability, processing integrity, confidentiality, or privacy.	System and Organization Controls (SOC)
A framework for identifying and managing security risks of generative AI models.	Generative AI Security Scoping Matrix

Validate Your Knowledge

When you are feeling confident with your review, it is best to validate your knowledge through sample exams. **Tutorials Dojo** offers a very useful and well-reviewed set of practice tests for the AI Practitioner exam takers **here** to help you prepare well. Each test contains many unique questions which will surely help you verify if you have missed out on anything important that might appear on your exam. You can pair these practice exams with this study guide eBook for comprehensive preparation.

If you have scored well on the **Tutorials Dojo AWS Certified AI Practitioner Practice Tests** and you think you are ready, then go earn your certification with your head held high. However, if you find certain areas challenging, take the time to review them again and pay attention to any hints in the questions that can guide you to the correct answers. If you're not entirely confident about passing, consider rescheduling your exam to allow yourself more preparation time. In the end, the efforts you invest will undoubtedly pay off.

Sample Practice Test Questions:

Question 1

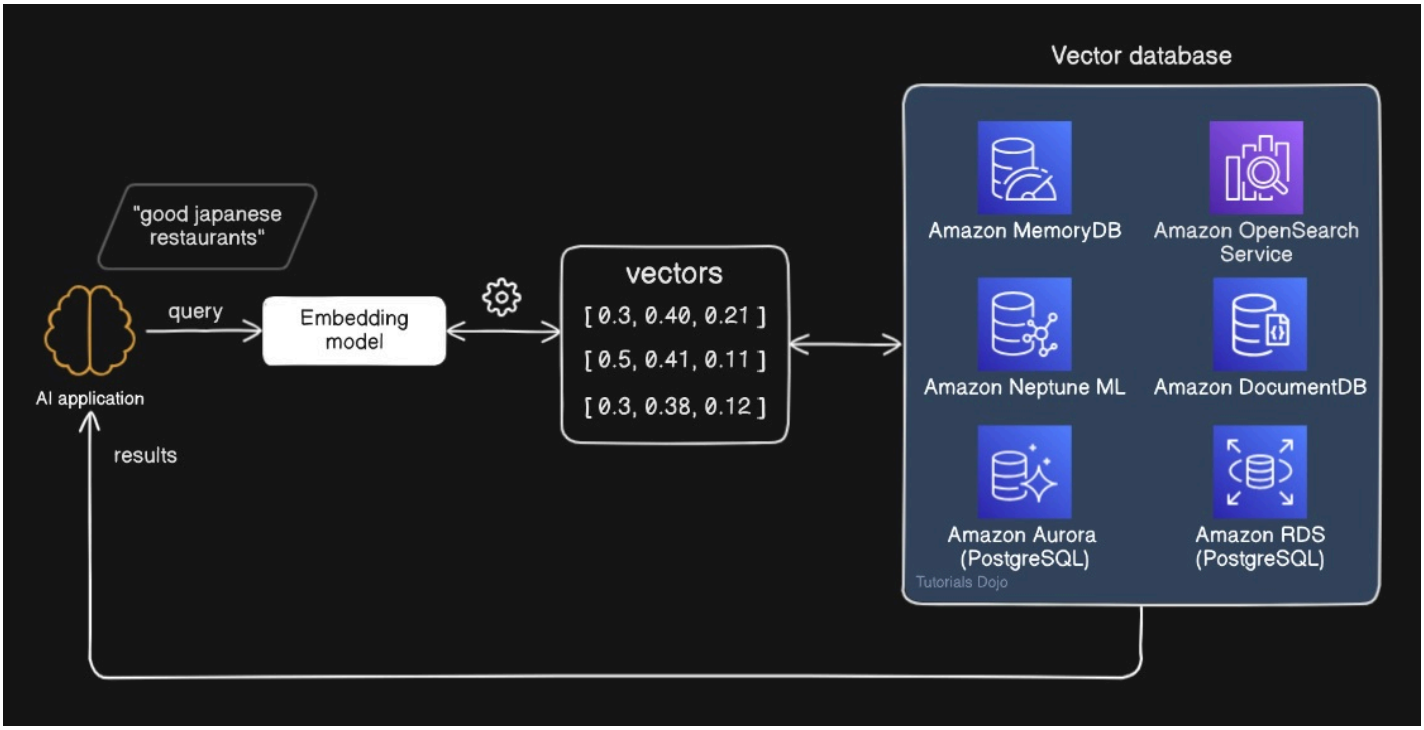
A company has a recommender system that generates embeddings from customer interaction data. They want to improve the speed and efficiency of retrieving similar product recommendations.

Which AWS services are best suited for implementing vector search to optimize the recommendation system? (Select THREE).

1. Amazon OpenSearch Service
2. Amazon Neptune ML
3. Amazon S3
4. Amazon DocumentDB (with MongoDB compatibility)
5. Amazon Redshift
6. Amazon Quicksight

Correct Answer: 1,2,4

Vector search is a method used in machine learning to find similar data points to a given data point by comparing their vector representations using distance or similarity metrics. The closer the two vectors are in the vector space, the more similar the underlying items are considered to be. This technique helps capture the semantic meaning of the data. This approach is useful in various applications, such as recommendation systems, natural language processing, and image recognition.



A vector database, specifically designed to handle vector representations of data efficiently, excels in storing, indexing, and retrieving vectors compared to traditional databases. Traditional databases typically handle scalar values and are optimized for transactional operations, which can be limiting for AI-driven queries that require rapid computation of vector similarities. In contrast, vector databases use specialized indexing algorithms, such as approximate nearest neighbor (ANN) search, which significantly speeds up query times and maintains high accuracy, even in large datasets. This makes them particularly advantageous for AI applications, where quick and precise retrieval of similar items based on complex, high-dimensional data is crucial. These capabilities allow for more dynamic and responsive AI systems, such as real-time personalized recommendation engines and instant image or voice recognition services.

Here are some services in AWS that you can use for your vector database requirements:

- **Amazon OpenSearch Service** makes it easy for you to perform interactive log analytics, real-time application monitoring, website search, and more. For vector databases, you can read about k-Nearest Neighbor (k-NN) search in OpenSearch Service.
- **Amazon Aurora PostgreSQL-Compatible Edition** and **Amazon Relational Database Service (Amazon RDS)** for PostgreSQL support the pgvector extension to store embeddings from machine learning (ML) models in your database and to perform efficient similarity searches.
- **Amazon Neptune ML** is a capability of Neptune that uses Graph Neural Networks (GNNs), an ML technique purpose-built for graphs, to make easy, fast, and more accurate predictions using graph data.
- **Amazon MemoryDB** supports storing millions of vectors, with single-digit millisecond query and update response times, and tens of thousands of queries per second (QPS) at greater than 99% recall.

- **Amazon DocumentDB (with MongoDB compatibility)** supports vector search, a new capability that enables you to store, index, and search millions of vectors with millisecond response times. With vector search for Amazon DocumentDB, you can simply set up, operate, and scale databases for your ML applications.

Hence, the correct answers are:

- Amazon OpenSearch Service
- Amazon Neptune ML
- Amazon DocumentDB (with MongoDB compatibility)

The option that says: **Amazon S3** is incorrect. This service is primarily an object storage service designed for storing and retrieving large amounts of data. While it can store vectors as files, it lacks built-in capabilities for indexing and searching vectors efficiently. For vector search, you need a service that supports real-time querying and similarity search, which S3 does not provide.

The option that says: **Amazon Redshift** is incorrect because this is mainly a data warehouse service designed for running complex queries on large datasets. Although it may be possible to implement vector search using custom User Defined Functions (UDFs), this approach would involve additional complexity and introduce potentially higher query latency.

The option that says: **Amazon Quicksight** is incorrect. This service is simply a business intelligence and visualization service for creating dashboards and reports. It is not designed for vector search or indexing.

References:
https://aws.amazon.com/what-is/vector-databases/
https://docs.aws.amazon.com/neptune-analytics/latest/userguide/vector-similarity.html
https://docs.aws.amazon.com/documentdb/latest/developerguide/vector-search.html
https://docs.aws.amazon.com/opensearch-service/latest/developerguide/serverless-vector-search.html

Check out these cheat sheets for Amazon OpenSearch Service, Amazon Neptune, and Amazon DocumenDB:
https://tutorialsdojo.com/amazon-neptune/
https://tutorialsdojo.com/amazon-documentdb/
https://tutorialsdojo.com/amazon-opensearch-service/

Question 2

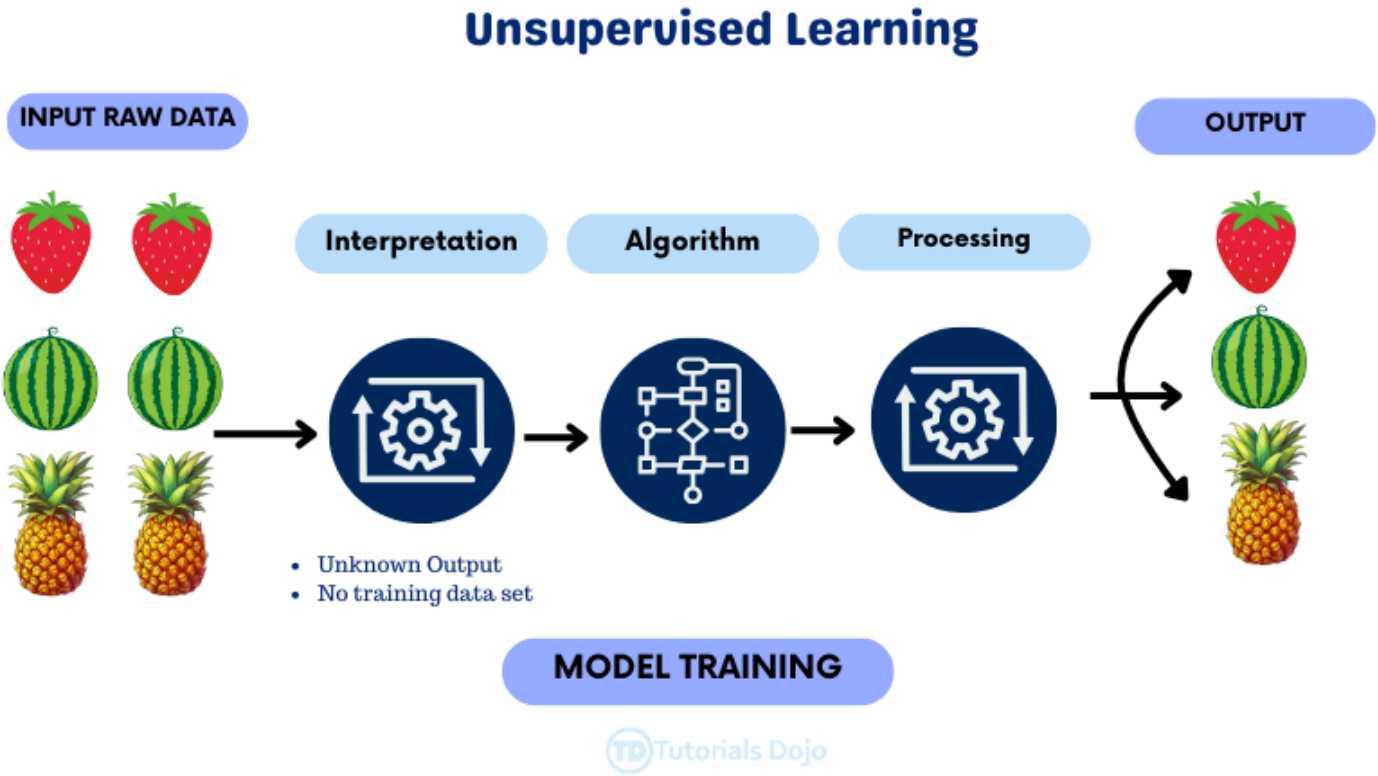
Which machine learning approach is used to classify and organize unlabeled data by identifying hidden patterns without requiring predefined categories or labels?

1. Transfer Learning
2. Reinforcement Learning

3. Few-shot Learning
4. Unsupervised Learning

Correct Answers: 4

Unsupervised learning is a machine learning technique used to analyze and cluster unlabeled data. Unlike supervised learning, where models are trained with labeled datasets, unsupervised learning algorithms operate without any predefined categories or labels. These algorithms discover hidden patterns, structures, and relationships in the data. One of the most common applications of unsupervised learning is clustering, where data points with similar characteristics are grouped. Another use case includes dimensionality reduction, which simplifies data by reducing the number of variables while maintaining its important features. This is highly effective when handling large datasets.



Amazon SageMaker AI supports unsupervised learning for various tasks, such as clustering, anomaly detection, and association rule learning. Amazon SageMaker AI provides pre-built algorithms such as K-Means and Principal Component Analysis (PCA) that can be used to analyze unlabeled datasets. By utilizing SageMaker's built-in unsupervised learning capabilities, users can quickly build and deploy models for tasks like customer segmentation, recommendation engines, or detecting anomalies in system logs. SageMaker streamlines the entire machine learning workflow, from data preparation to deployment, making it easier to leverage unsupervised learning for real-world business needs.

Hence, the correct answer is: **Unsupervised Learning.**

The option that says: **Transfer Learning** is incorrect because this type of learning uses a pre-trained model from one task or domain and applies it to a different but related task. Transfer learning is typically used when there is a shortage of labeled data in the target domain but an abundance of labeled data in a related domain. This approach is not designed for working with unlabeled data or discovering hidden patterns, making it unsuitable for the scenario described.

The option that says: **Reinforcement Learning** is incorrect because this primarily focuses on learning through an agent's interactions with an environment. In reinforcement learning, the agent receives feedback in the form of rewards or penalties based on its actions and learns to maximize cumulative rewards over time. This approach is not meant for classifying or grouping unlabeled data, but rather for optimizing decision-making processes in dynamic environments.

The option that says: **Few-shot Learning** is incorrect because it involves training models with a very limited amount of labeled data, often just a few examples. Few-shot learning is particularly useful when labeled data is scarce, but it still requires some level of labeled data to function. It does not apply to situations with no labeled data, as is the case with unsupervised learning.

References:
https://docs.aws.amazon.com/sagemaker/latest/dg/algorithms-unsupervised.html
https://aws.amazon.com/compare/the-difference-between-machine-learning-supervised-and-unsupervised/

Check out this Amazon SageMaker AI Cheat Sheet:
https://tutorialsdojo.com/amazon-sagemaker/

What to expect from the exam

There are five types of questions on the examination:

- **Multiple-choice:** Has one correct response and three incorrect responses (distractors).
- **Multiple-response:** Has two or more correct responses out of five or more options.
- **Ordering:** These questions provide a list of 3–5 responses that need to be arranged in the correct order to complete a specified task.
- **Matching:** You need to match a list of responses with a set of 3–7 prompts. All pairs must be matched correctly to receive credit.
- **Case study:** These questions are based on a single scenario with two or more related questions. Each question within the case study is evaluated separately. Credit is given for each correct response.

Distractors, or incorrect answers, are response options that an examinee with incomplete knowledge or skill would likely choose. However, they are generally plausible responses that fit in the content area defined by the test objective.

Unanswered questions are scored as incorrect; there is no penalty for guessing.

Majority of questions are usually scenario-based. Some will ask you to identify a specific service or concept. While others will ask you to select multiple responses that fit the given requirements. No matter the style of the question, as long as you understand what is being asked, then you will do fine.

Your examination may include unscored items that are placed on the test by AWS to gather statistical information. These items are not identified on the form and do not affect your score.

The AWS Certified AI Practitioner (AIF-C01) examination is a pass or fail exam. Your results for the examination are reported as a scaled score from 100 through 1000, with a minimum passing score of 700. Right after the exam, you will immediately know whether you passed or you failed. And in the succeeding business days, you should receive your complete results with the score breakdown (and hopefully the certificate too).

A few more tips:

1. Be sure to get proper sleep the night before, and don't be lazy in preparing for the exam. If you feel that you aren't ready enough, you can just reschedule your exam.
2. Come early to the exam venue so that you have time to handle mishaps if there are any.
3. Read the exam questions properly, but don't spend too much time on a question you don't know the answer to. You can always go back to it after you answer the rest.
4. Keep your reviewer if you plan on taking other AWS certifications in the future. It will be handy for sure.
5. And be sure to visit the Tutorials Dojo website to see our latest AWS reviewers, cheat sheets and other guides.

AI AND ML FUNDAMENTALS

Understanding Basic AI Concepts and Terminologies

Identifying Practical Use Cases for AI

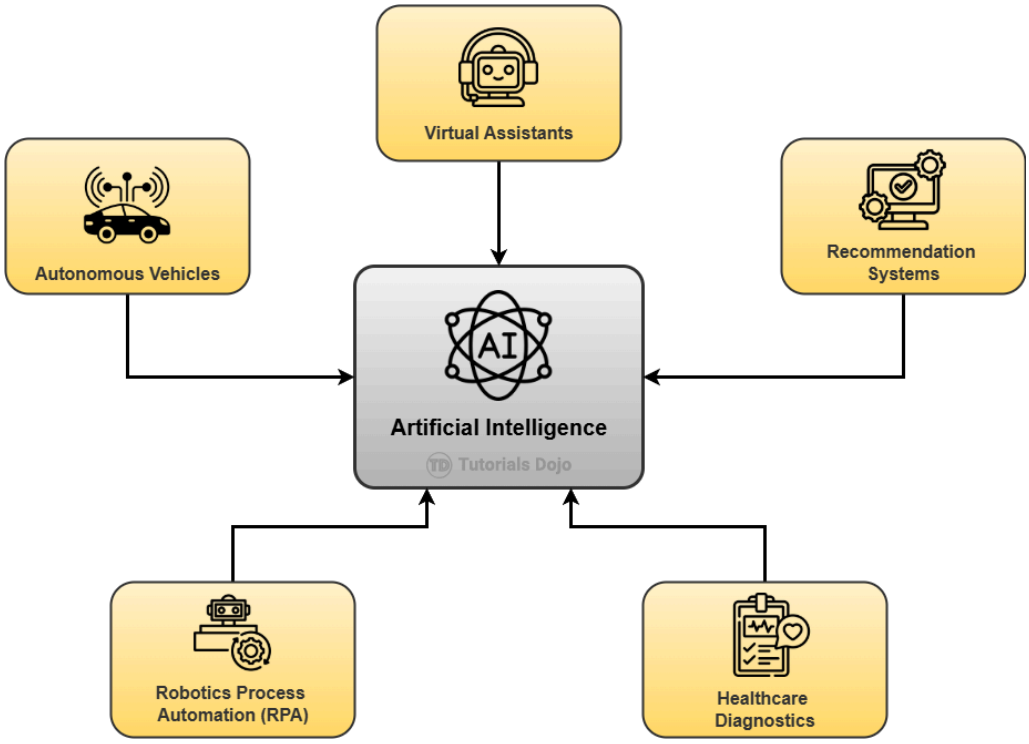
The Machine Learning Development Lifecycle

A. Explain Basic AI Concepts and Terminologies

Before we proceed in learning and understanding the AWS tools and services under AI, it is important for us to know first the fundamentals. Here, we crafted the essentials needed for you to understand how AI/MLalgorithm work theoretically, without writing a single code.

Chapter 1.1: Understanding Basic AI Concepts and Terminologies

Artificial Intelligence



Artificial intelligence (AI) is a field within computer science that focuses on creating systems capable of performing intelligent tasks that mimic human intelligence. These tasks include understanding the context behind natural human languages, recognizing information patterns and events, solving complex challenges and decisions, and having an awareness of their environment. For example, you can think of AI as a talented chef who can cook a wide range of Filipino dishes such as adobo, sinigang, or lechon without asking a co-worker for assistance with the recipes or even looking at a cookbook again.

Additionally, AI represents the overarching concept, but within AI, there are specialized techniques like machine learning, where algorithms learn patterns within data and deep learning. It may sound too overcomplicated, but we will discuss this each by each in a bite-sized approach in the later parts of this chapter.

Furthermore, to give a wider understanding of how AI is applied, below are the practical real-world examples of AI across different sectors, showcasing how AI technologies solve real-world problems and improve outcomes. *Take note that understanding these use-cases might be asked during the exam.*

1. **Virtual Personal Assistants**
    a. Examples: Apple Siri, Amazon Lexa, Google Assistant
    b. Description: These AI-powered assistants understand and respond to the user's voice commands, perform tasks like setting reminders, controlling our smart home devices, and providing information through speech-to-text mechanisms.

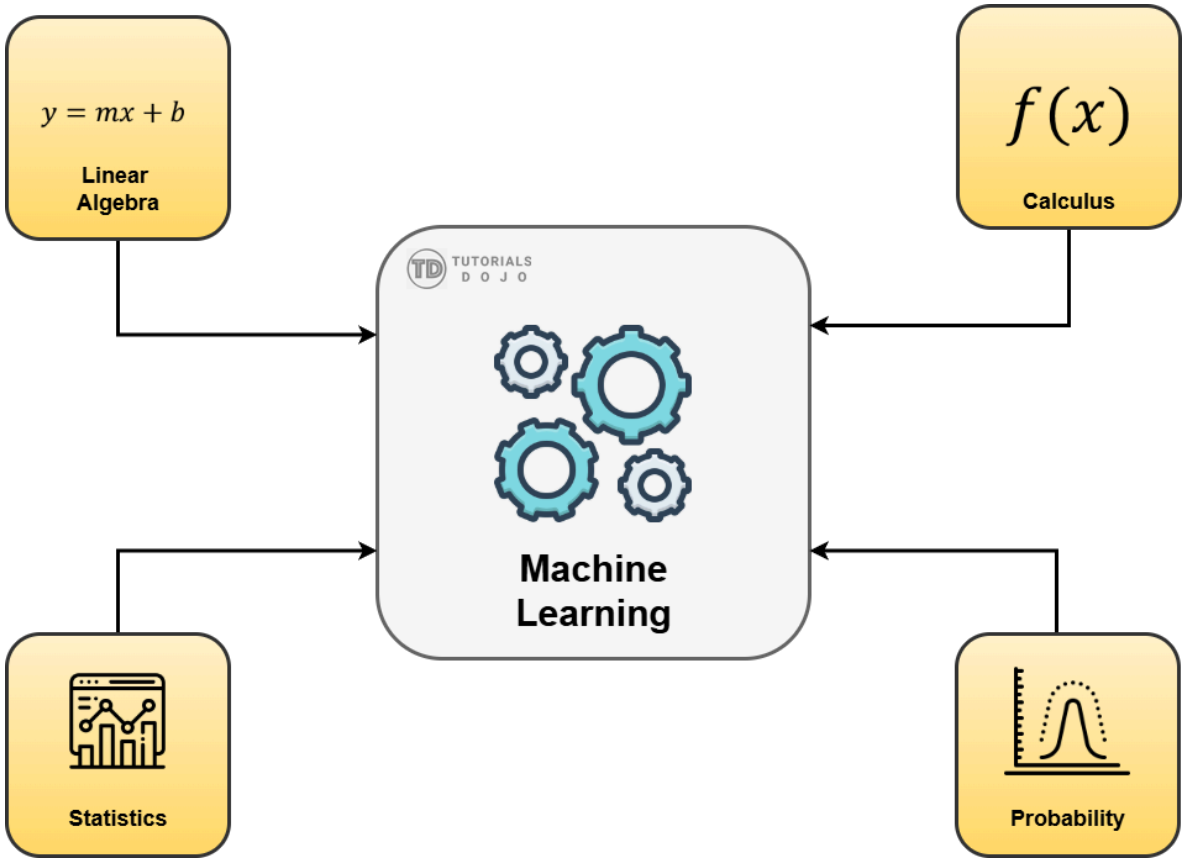
2. **Autonomous Vehicles**
    a. Examples: Tesla Autopilot and Full Self-driving, Wayno
    b. Description: These AI algorithms use smart vehicle sensors and cameras to navigate roads, make driving decisions, and ensure that passengers arrive at their destination safely without human intervention.

3. **Recommendation Systems**
    a. Examples: Netflix, Spotify, Amazon
    b. Description: These AI algorithms analyze user behavior and preferences to predict and suggest the next movies, music, or products which they may like, enhancing user experience and engagement. Through this approach, these algorithms became beneficial for a lot of e-commerce applications as well.

4. **Robotics Process Automation (RPA)**
    a. Examples: UiPath, Blue Prism
    b. Description: AI-driven software robots like these help create automations and constant repetitive tasks, such as data entry, invoice processing, and customer service which in return helps reduce operational costs with reduced human error.

5. **Healthcare Diagnostics**
    a. Examples: IBM Watson Health
    b. Description: AI systems like these help diagnose health-related diseases by analyzing medical data, imaging, and patient history, supporting healthcare professionals in making informed decisions for the patient.

Machine Learning



Machine Learning Is a subset of artificial intelligence that blends science and art to teach computers how to learn from data. This allows the technology to improve their computing and predicting performance over time without needing explicit programming for each task, similar to how humans learn from mistakes or experience.

For instance, imagine that the chef we mentioned starts learning their general customer's tastes by asking for feedback. After cooking their recipes several times on a daily basis, the chef begins to adjust the flavor levels based on the feedback received. This learning from experience and received feedback is the similar structure behind Machine Learning, it is under the work of an AI, where ML systems learn from data.

Furthermore, ML algorithms use mathematical concepts (see the picture above) and algorithmic logic to adapt to new information and make predictions. Hence, this enables computers to perform intricate tasks such as recognizing images, understanding speech, and making predictions.

Common Machine Learning Applications

To take it a step further, here are some real world Machine Learning (ML) applications that you may need to take note of:

1. **Fraud Detection in Finance**

   Fraud Detection is important to avoid suspicious activities, theft, and misuse of resources. To avoid these incidents. Machine Learning performs an analysis of transaction patterns from historical data to identify and prevent fraudulent activities in real-time.

   Example: Paypal, Visa, MasterCard

2. **Predictive Maintenance**

   Business continuity plans are critical to avoid company revenue loss and customer trust. Machine Learning models can predict potential equipment failures by analyzing material performance data. This enables organizations to perform maintenance ahead of time, which reduces downtime, maintenance cost, and increases the asset lifespan.

   Example: IBM Maximo, SAP Predictive Maintenance and Service

3. **Personalized Marketing**

   Personalized marketing offers a way for business owners to reach their target market effectively. There are machine learning algorithms that analyze user data and behavior to recommend products and services via advertisements to their target market, improving marketing effectiveness and return of investment (ROI) for businesses.

   Example: Specialized Facebook Ads, Lazada, Shoppee

4. **Email Filtering and Spam Detection**

   A lot of companies get their data breached every year being a victim of email phishing. To mitigate the risk of clicking phishing sites, Machine Learning-based algorithms can also classify an incoming email as either legitimate or potentially a spam, based on patterns and specific features from the email content and metadata.

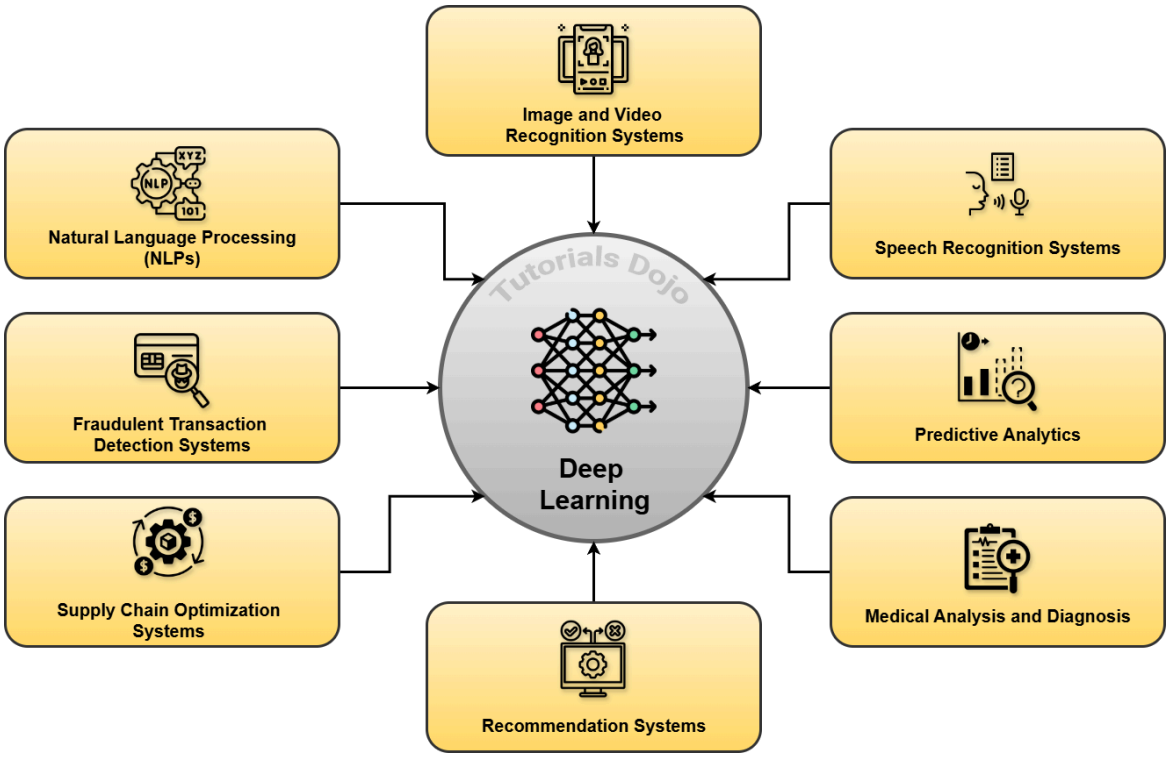
   Example: Gmail Spam Filter

5. **Credit Scoring**

   Financial institutions that offer loaning services need to track the credibility of their clients to ensure that they can repay on time. ML algorithms can assess an individual's credit worthiness by analyzing financial history, transaction data, and other relevant factors to determine loan eligibility.

   Example: FICO Score, Zest AI

Deep Learning



Deep Learning (DL) is a specialized subset of ML, which is also a subset of AI, where it employs **Artificial Neural Network (ANN)**, which is known for implementing layers of computational learning, giving the learning term "deep", which enables a model to understand complex patterns and representations within large datasets, such as Big Data. It was inspired by the functional structure of a human brain, as these consist of interconnected layers of nodes called "neurons" to enable systems to learn intricate relationships and create predicted decisions. We will further discuss its structure on the following chapter.

Hence, to give it a better representation, going back with our Chef example, imagine now that the chef has mastered cooking their signature dishes, then starts to analyze other factors such as when is a dish suitable depending on the season or events and as well as learning how to adjust the taste depending on the

preference of some specific customers they note of. As the chef becomes capable of knowing these specific patterns without being instructed explicitly in the first place is similar to how Deep Learning works.

Use Cases of Deep Learning

Deep learning (DL) has revolutionized various industries by enabling businesses to harness vast amounts of data, leading to increased productivity through enhanced decision-making, automation, and improved customer experiences. By leveraging sophisticated algorithms, deep learning helps organizations unlock valuable insights and streamline operations. Below are several prominent use cases of deep learning:

1. **Image and Speech Recognition Systems**
    a. Examples: Google Photos, Apple Face ID, and Amazon Alexa Speech Recognition.
    b. DL algorithms process and recognize unique objects, understand speech context within sounds, and enable functionalities such as automatically tagging individuals, securing facial authentication, and voice-controlled devices.

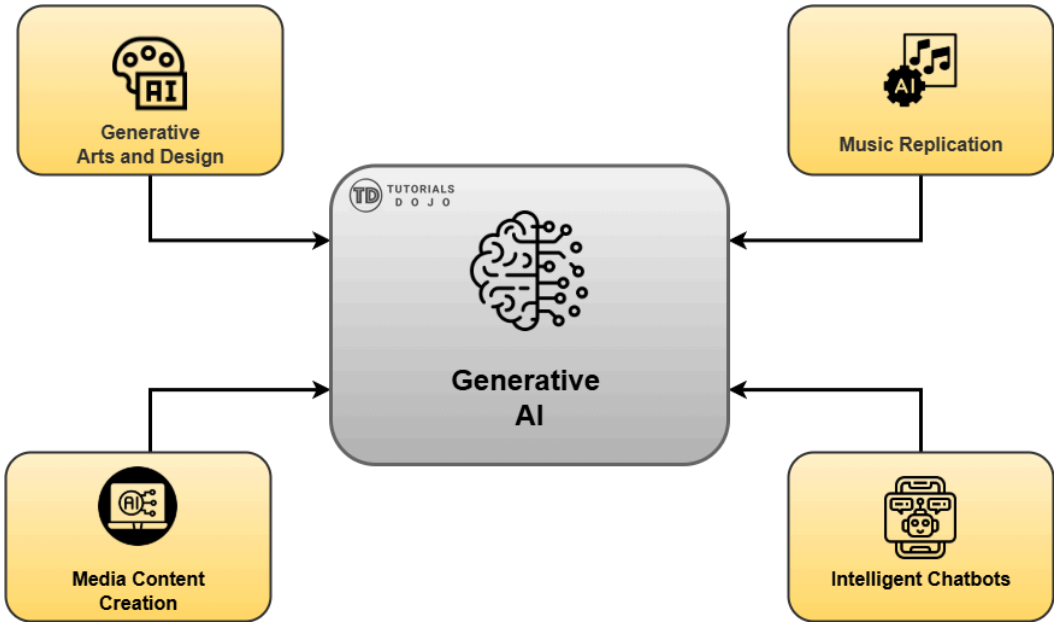
2. **Natural Language Processing**
    a. Examples: OpenAI's GPT-4o, Claude Anthropic, Google Gemini
    b. These specialized DL algorithms under natural language processing (NLP) enable computers to understand the context and generate human language, which facilitates language translation, chatbots, and content creation.

3. **Autonomous Vehicles**
    a. Example: Tesla Full Self-Driving (FSD)
    b. DL algorithms, sometimes combined with other ML algorithms, serve as the underlying mechanism of AI to enable self-driving cars. It analyzes sensor data, images, and videos at real-time to recognize traffic signs, pedestrians, and navigate complex environments ahead.

4. **Health Imaging Applications**
    a. Examples: PathAI, Aidoc
    b. Health Images with DL algorithms can analyze X-rays or MRIs to detect diseases such as cancer or tumors accurately, serving as a tool to assist radiologists in diagnostic procedures and treatment planning.

5. **Generative Adversarial Networks**
    a. Examples: DeepArt, OpenAI's DALL-E, "This Person Does Not Exist"
    b. DL models in general can generate new images based on the learned image datasets. It can create synthetic data by learning on large datasets, enabling creative applications, and data augmentation for training other models.

Generative AI



Generative AI, by the word itself "Generative", means it is a subset of deep learning domain that focuses on producing new data in various formats, used to inform people or create quality work similar to how humans do. Generative data can be in the form of text, files, audio, images, and synthetic data. It was proposed in the 1960s but became widespread in 2014 after the proposal of *Generative Adversarial Networks (GAN)*. These shall be discussed further in Chapter 2 of this book.

Use Cases for Generative AI.

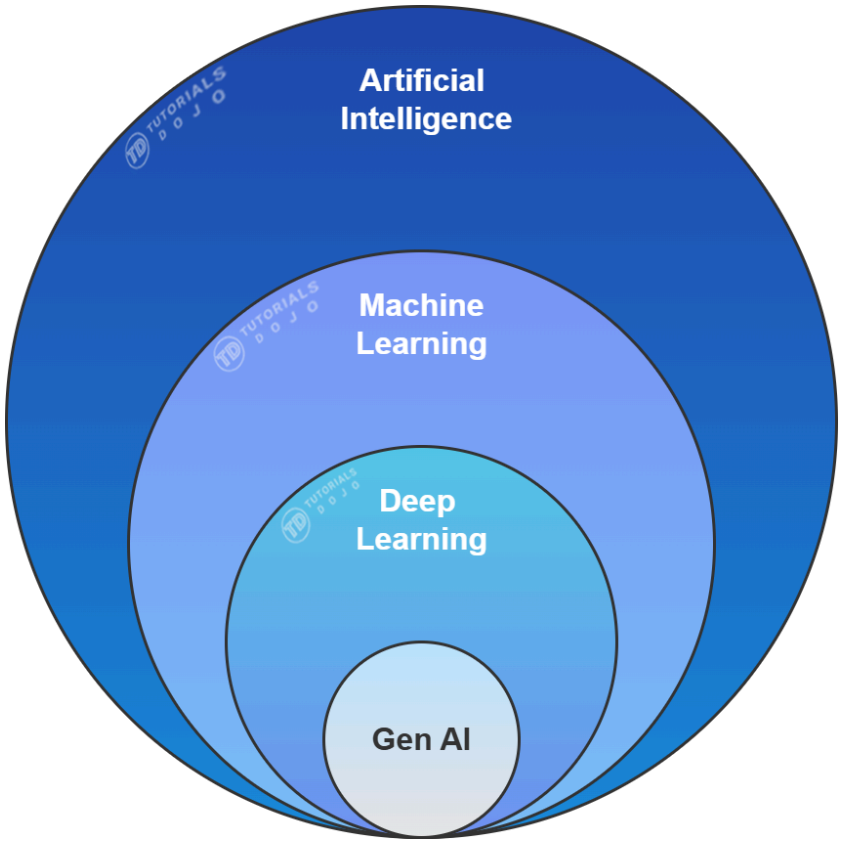
1. **Arts and Design** - Generative AI models can generate images patterned at various artworks and design images.

2. **Music Replication** - Generative AI models can create new sound wavelengths patterned at different audio files, such as human voice, environment sounds, or instrumental beats, which can compose music. One of the popular applications of this is called AI vocal transformation, where the singer from a song is artificially replaced by another singer based on his or her voice.

3. **Media Content Creation** - Generative AI models can now create blogs, documents, and other data content generation which can be used to inform people about a certain topic.

4. **Intelligent Chatbots** - One of the most common use-case of generative AI is the creation of chatbots, where a user can interact to gather information based on its knowledge bases.

Similarities and Differences Between AI, Machine Learning, Deep Learning, and Generative AI



Overall, despite the differences, there are profound similarities within AI, Machine Learning, Deep Learning, and Generative AI. Firstly, these fields all aimed to mimic human intelligence, such as understanding the context of the data, pattern recognition, and decision making. Secondly, all accept input data to make a basis for the aforementioned tasks. Lastly, based on the image itself, it shows a hierarchical relationship, sharing the core principles for the use-case of AI.

However, we must denote the following distinctions:

- **AI** refers to rule-based systems, a general terminology for these intelligent processes, and represents the symbolic reasoning behind intelligent systems.

- **Machine Learning** is a subset of AI which simply learns from the data and makes predictions solely depending on the learned data.

- **Deep learning** is a subset of machine learning that uses complex layers for learning a particular dataset, enhancing its ability to learn more, as it also mimics the human brain.

- **Generative AI** is a subset of deep learning that uses generative AI models to create new usable data derived from its knowledge bases. It can be used to generate humanistic documents, images, and media.

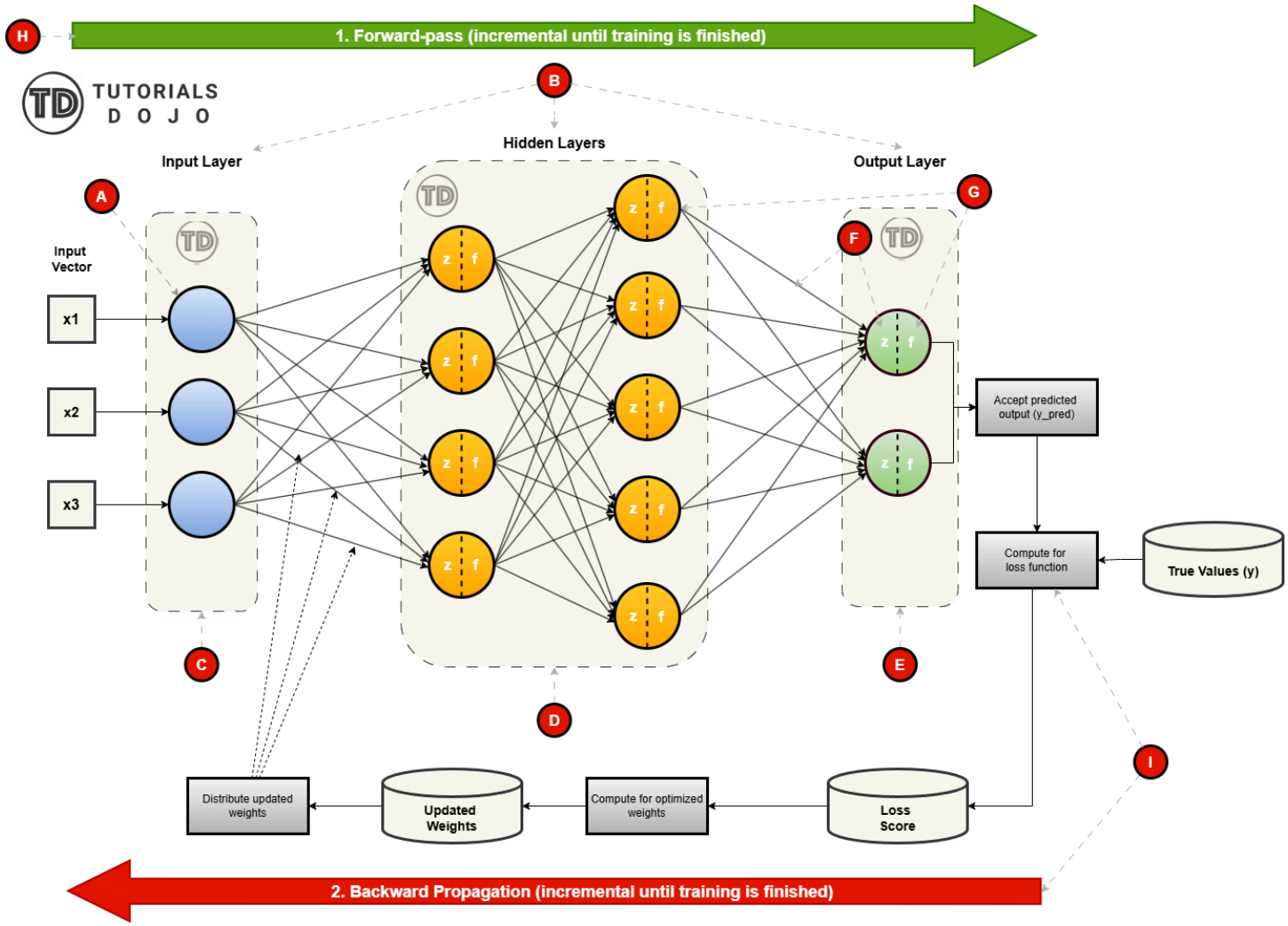
Neural Network

Neural Network (NN) is a computational AI model built within layers, inspired by the structure of the human brain. It is a type of machine learning process called **deep learning,** where it consists of interconnected units called **neurons** or **nodes** organized in structured layers which work along to process and learn data. By using its adaptive system, computers can learn from their errors and improve incrementally. Neural networks are fundamental to many deep learning algorithms as they are efficient in recognizing patterns, following complex instructions, and making decisions.

Furthermore, common use-cases of neural networks would involve:

- Assistive medical diagnosis using image classification algorithms.
- Recommendation systems to social networks using user-behavior analysis.
- Stock market and other finance-related forecasting using time-series and financial historical records.

Architecture of a Basic Neural Network



Neurons (A)

Neurons, represented as nodes in neural network diagrams, are the fundamental units of artificial neural network algorithms, analogous to biological neurons. They receive input vectors, compute weighted sums with bias in linear function (z), and apply activation functions (f) to produce outputs. Organized into layers, these neurons process data sequentially until reaching an output layer with a limited number of neurons corresponding to the expected labels.

Layers (B)

Layers serve as the core components of neural network algorithms that define the logical structure and function of a network. It consists of sequentially arranged neurons, where they process the input data, transforming it into a specific output. A conventional artificial neural network will have the following layers:

Input Layer (C)
This layer receives the vectorized data and passes it to its next layer. Each neuron in the input layer represents a feature in the data. A common misconception of the input layer is that its neurons already have their weight and biases disbursed. However, it only receives the vectorized data as an input and no calculations are performed yet.

Hidden Layers (D)
These are the layers that reside within the input and output layer. The neurons under these layers contain a set of weights and activation functions to transform the input data into smaller learnable segments. For this to happen, they use non-linear computations to the input vectors. Furthermore, these layers stack together in multiple segments which allows the network to have hierarchical feature representations.

Output Layer (E)
This layer receives the final processed results of the network, and it is now limited to prediction or classification tasks. For classification tasks, the output layer typically uses a softmax activation function to produce the probabilities for each class.

Weights and Biases (F)
Weights are numbers assigned to connections in a neural network. They control how strongly one neuron's output influences another. Adjusting weights during training helps the network predict accurately. The goal is to reduce the gap between predictions and true labels. On the other hand, Biases are constants added to the input sum of each neuron. They change the final result before using the activation function. Biases adjust how quickly a neuron activates. Even if the input is zero, the bias allows the neuron to still send a signal. This keeps the process moving forward.

Activation Function (G)
Activation functions help compute the result of the weighted sum plus bias to minimize the value in determining whether the neuron should activate or remain dormant on the next input value. It is a non-linear mathematical function applied to every neuron output, with the aim of learning and representing patterns in the data. In practice, there are various activation functions depending on the use cases.

Forward pass / Forward propagation algorithm (H)
An algorithm that happens during training and inference. The input data moves through the neural network's hidden layers. Each neuron in every layer uses weights and biases repeatedly. The activation function adds non-linearity and changes the input to create the neuron's output.

Backpropagation algorithm (I)
An algorithm that is crucial during the training of neural networks. It reduces the error gap between predicted and actual values. Everytime it performs, it updates its weights and biases after a feedforward training cycle. The process begins by calculating gradients, where it contains the adjusted weights by the calculated error and distributes it to the whole architecture.

Gradient Descent
An optimization algorithm used to minimize the functions occurring within the network by iteratively moving toward the lowest point of the function’s slope. In other words, this optimizes the model parameters to reduce the loss function or the distance of actual vs. predicted labels, enhancing the model’s overall accuracy.

Vanishing Gradients
An occurrence happens when the gradients of the loss function become really small as they move backward through the network layers. This event makes weight updates hard. It stops the network from learning from its

errors efficiently. This issue prevents the model from capturing long-range neural network dependencies. Hence, it hinders the successful training of deep architectures.

Exploding Gradients

This is another occurrence where the gradients grow exponentially during back propagation, which leads to fluctuating updates to weights, making the model parameters unstable. Hence, this makes the training process unstable and can prevent networks from converging to a meaningful solution.

Normalization

Potentially transforms and scales the input data or other intermediate activations within a neural network to ensure consistent distributions of activation across different layers. It helps the neural network stabilize during the forward pass by mitigating potential issues such as vanishing or exploding gradients.

Regularization

It is a process that encompasses a set of techniques that only aims to prevent overfitting, a terminology used to describe a poor prediction ability on unseen values despite performing well during the training phase based on the training data. Overall, regularization simply aims to ensure better generalization.

Types of Neural Network Architectures

Feedforward Neural Networks (FNN)

A Feedforward Neural Network (FNN) or Multilayer Perceptron (MLP), is a basic type of artificial neural network. The design is simple. Connections between nodes never create loops or work in cycles, unlike in Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs). FNNs help with tasks like classification, regression and recognizing patterns. People use them often for these jobs.

An **FNN** consists of the following:

- **Input Layer -** This layer gets the raw data. Each neuron in this layer represents a part of the input.
- **Hidden layers** - These layers do computations and transformations. These help the network find complex patterns.
- **Output layer** - This layer gives the final prediction. It labels or classifies the data.
- **Loss Function** - It measures how much the predicted results differ from real targets. The loss function guides adjustments.

Commonly used for **Pattern Recognition, Classification**, and **Regression tasks**.

Recurrent Neural Networks (RNN)

Recurrent Neural Networks are specialized neural networks designed for processing sequential data. They feature loops in their architecture, enabling information to persist across different time steps, making them

ideal for tasks where context and order are essential.

Sequential Data: Consists of consecutively ordered data such as words in sentences, numerical time-series data, or sequences in video frames. The position and order of data points contribute significantly to their context and meaning.

An **RNN** consists of the following:

- **Input Layer** - The input layer receives the sequence data. The network takes each element in the order, one by one. In language tasks, each word or letter from a sentence enters as input at a certain moment.
- **Recurrent Hidden Layers** - These layers perform the main calculations in the RNN. Every hidden layer holds a state. This state keeps details from earlier inputs. It helps the network show patterns over time. The layers really process temporal information.
- **Output Layer** - Gives the network's final result. This result is based on what the hidden layers compute.
- **Loss Function** - Checks the difference between the predicted results and the real targets. For classification tasks, Cross-Entropy Loss is common. Regression tasks often use Mean Squared Error. It measures errors.

Common applications of **RNN** would include but are not limited to:

- **Natural Language Processing (NLP)** - RNNs became a foundation to more complex algorithms such as Transformers, which power advanced algorithms such as Natural Language Processing (NLP) applications like ChatGPT, Claude Anthropic, Llama. Examples would include language translation, sentiment analysis, text generation.
- **Time Series Prediction** - Ideal for forecasting tasks where temporal dependencies are crucial. Examples would include Stock market prediction, weather forecasting, sales and inventory forecasting.
- **Video Analysis** - Employed in tasks requiring understanding of sequences within video data. Examples would include activity recognition, video captioning, and sequence generation in video events.
- **Speech Recognition** - This converts spoken language into text by captioning patterns in audio signals. Examples would include virtual agent assistants and voice transcription services.

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a unique type of neural networks designed to process batches of image data by dividing collection of images into smaller segments and applying grid-like scanning algorithms

through convolutions. This structure enables tasks such as image classification, segmentation, and video object detection.

A **CNN** consists of the following:

- **Input Layer** - Takes in raw data, commonly as images. These images are arrays with height, width and color channels such as RGB.
- **Convolutional Layers** - This is where convolution operations perform to extract local features of the data. Filters, also called kernels, slide over the grid of data, creating an output called *feature maps* to store the relevant information.
- **Pooling Layers** - They make feature maps smaller. This helps less work for the computer and prevents mistakes.
- **Fully Connected (Dense) Layers** - Mix different features from the previous layers for deep understanding. They really help in final decision-making.
- **Output Layers** - Gives the last prediction or result.
- **Loss Function** - Checks the difference between the guess and the real answer. It shows how much the guess is off. Cross-Entropy Loss and Mean Squared Error are often used here.

Common applications of **CNN** would include but are not limited to:

- **Image Classification:** CNN can be used for creating models that classify images based on its features. Examples would include utilizing CNN for tasks like identifying diseases in leaves, animals in wildlife, or detecting defects in manufacturing.
- **Object Detection:** CNN can be used for locating and classifying specific objects within images or videos. or images. Examples would include employing object-detection CNN models for real-time detection of vehicles in traffic monitoring systems.
- **Image Segmentation:** A way to manipulate the images to make the foreground distinct from its background. Examples would include training Image Segmentation CNN models for medical imaging to segment tumors or for autonomous vehicles to recognize road signs and pedestrians.

Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GANs) is a deep learning architecture that trains two competing neural networks, which in return, provides new authentic data derived from the given training dataset.

A **GAN** consists of the following:

- **Generator** - a neural network model used to train for generating new authentic data.

- **Discriminator** - a neural network model that tries to classify examples as either real (from the domain) or fake.
- **Zero-sum** - a game that checks if the discriminator is “fooled” with less time, meaning that GAN is generating almost realistic examples.

Common applications of **GAN** would include but are not limited to:

- **Training Data Generation** - when datasets are quantifiably imbalanced, GANs are used to generate augmented images that replicates the training data.
- **Completion of Missing Values** - GANs can accurately guess missing values based on its surrounding values

Transformer Models

- A neural network model that learns features from data by understanding context. It follows relationships in sequences like text or images. Its special trait is using *self-attention mechanisms* to find faraway features in the data. It notices details based on position.
- This helps language tools like translators. They translate words not just by their shape but by their meaning. For example, in "He recorded the match" and "I lit the match stick," the word “match” looks the same. However, these systems see it as a contest or as a tool to start fires.

A **Transformer Model Block** would include the following:

- **Encoder** - processes and changes input data into position-based representations. It employs self-attention to grasp links between all elements in the input sequence. It then moves through a feed-forward layer with a residual connection to keep details while normalizing layers. Information is important.
- **Decoder** - receives data from the encoder and creates an output sequence. It includes a masked multi-head attention from the output embedding. These values help understand words generated by the encoder. It finishes with a linear and softmax activation.

Common applications of a **Transformer** would include but are not limited to:

- **Generative Pre-trained Transformer (GPT)** - creates new contextual information based on the prompts of its users.
- **Vision Transformers (ViT)** - splits the image into grids called patches, and align its pixels as an input embedding, and by only using the encoder of a standard Transformer, it can identify images more accurately than CNNs.

Chapter 1.2 Key Domains in AI

Computer Vision

Enables machines to identify people, places, and things in photographs with accuracy equal to or greater than humans while also operating at considerably higher speeds and efficiency.

Natural Language Processing (NLP)

A branch of artificial intelligence dedicated to enabling computers to understand, interpret, and interact with human languages.

Large Language Models (LLMs)

Large Language Models (LLMs) are deep learning models. They are trained on large volumes of data. Neural networks comprising an encoder and a decoder make up the transformer structure used in these models. Transformers are special because they can learn on their own without continual guidance.

One application of LLMs would be the development of intelligent chatbots for customer support that can comprehend and reply to queries from clients in regional or Filipino languages. By being aware of regional dialects and idioms, these models can enhance services like online banking and shopping by facilitating more efficient and realistic interactions with Filipino clients.

Small Language Models (SLMs)

Small Language Models (SLMs) are efficient versions of Large Language Models (LLMs) that require less computational power and memory. By leveraging cloud processing, SLMs enable devices with limited capacity to perform advanced tasks like natural language processing and predictive automation. They provide faster processing and lower resource demands, making them ideal for specialized tasks and edge computing where larger models are impractical.

Models and Algorithms

In general, an **algorithm** is a finite step-by-step instructions that a computer follows in order to solve a problem. In the context of machine learning, it is the step-by-step process or procedure that data goes through in order to create a model. On the other hand, a **model** is the output of a machine learning algorithm.

An example to make this clearer, linear regression is used to model the relationship between monthly household income and electricity consumption. This will output in a model that estimates electricity usage based on these factors, which can help power companies forecast demand and optimize resource allocation.

Common Algorithms in AI/ML

The following is the list of common algorithms used in machine learning and artificial intelligence. However, it is not limited to only the ones in this list, rather this is a good list to be familiar with.

- **Linear Regression**

  By fitting a straight line through the data points, linear regression is a technique for determining the relationship between two variables. Based on the value of another variable (the independent variable), this line aids in the prediction of the dependent variable. In order to forecast actual values, such as home prices or sales, depending on relevant criteria, the objective is to choose the line that most accurately depicts the pattern in the data.

  The best fit line is also known as the **regression line**, represented by a linear equation: Y = a * X + b.

  - Y - Dependent variable
  - a - slope
  - X - Independent variable
  - b - Intercept

  An example of this algorithm is if someone wants to predict the house prices in Metro Manila based on their size (square meters). After collecting data on recently sold houses with their prices and sizes, a linear regression model can be used to estimate the relationship between these two variables. Once the model is trained, it can be used to predict the price of a house based on its size.

- **Logistic Regression**

  To predict a true/false or yes/no result, logistic regression employs certain factors. It is also known as logistic regression since it estimates the likelihood that an event will occur and produces a value between 0 and 1. In other words, it is a classification algorithm despite the “regression” in its name.

For instance, if a student wants to predict whether they will pass their incoming entrance exams based on their high school grades, it is a classification problem. Logistic Regression can be used to get an outcome of either yes or no.

- **Decision Tree**

  Decision Tree is mostly used for classification problems too. It accurately predicts both continuous and categorical outcomes. This approach divides the data into two or more related categories depending on essential characteristics. These factors, known as independent variables, contribute to the formation of separate groups.

  For example, if a Philippine bank wants to predict whether loan applicants will likely repay their loans or default, they can use this algorithm. Factors such as income, employment status, and credit score can be analyzed in order to classify the applicants into different risk groups.

- **Support Vector Machine (SVM)**

  Another classification method, wherein every data point in the SVM approach is positioned in a n-dimensional space (n is the number of features), with each data characteristic representing a distinct dimension. The point's position in that space is determined by the value of each characteristic.

  For example, if someone wants to categorize homes in Cebu according to size and cost, the SVM method will treat each house as a point in a space with two dimensions—one for size and one for price. The SVM model will then find the best way to separate houses into categories, like affordable vs. expensive.

- **Naive Bayes**

  Assuming that all features are independent of one another, the Naive Bayes classification approach is based on the Bayes theorem. This classifier makes predictions based on the assumption that the existence of one attribute has no bearing on the existence of another.

  To illustrate, if someone is classifying whether a person in the Philippines prefers basketball or volleyball based on factors like age, location, or favorite athlete, the Naive Bayes classifier would treat each factor (age, location, favorite athlete) as independent, even though they may be related in real life. It wouldn't consider how location might affect someone's favorite athlete.

- **k-Nearest Neighbors (kNN)**

  KNN can be applied to regression and classification. All available data points are stored, and new data is categorized using the majority class of its closest neighbors. The "K" stands for the number of nearest neighbors taken into account, and the most common class among them is given to the new

case. To determine the closest neighbors, distance functions such as Manhattan or Euclidean are utilized.

An example of this is a farmer in the Philippines who wishes to categorize various crop types according to rainfall and soil quality data. By comparing the closest samples from historical data, KNN can be used to classify new soil samples into crop types.

- **K-Means**

  K-Means uses an unsupervised approach to tackle the data grouping problem. The data is arranged into a certain number of groups or clusters in order for it to function (e.g., k clusters). Although the data points in each cluster differ from those in other clusters, they are comparable to one another.

  For instance, k-means clustering can be used to group cities in the Philippines based on similar economic indicators like income levels and employment rates. This helps identify regions with similar economic conditions, allowing for targeted development programs.

- **Random Forest**

  Multiple decision trees are used in Random Forest, an ensemble learning technique, to generate predictions. Each tree in the "forest" votes for a specific outcome depending on the input data, and the most common vote among the trees is the final prediction.

  An example could be to predict whether rice crops will have a high or low yield based on weather conditions and soil quality across different regions in the Philippines. Each tree in the forest analyzes the data and "votes" on whether the yield will be high or low, and the majority vote determines the final prediction.

Fit and Overfitting

Model Fit
The ability of a machine learning model to adjust to new data that is comparable to the data it was trained on is known as model fitting.
Overfitting and Underfitting
Overfitting is when a model can give accurate predictions for the training data but not for new or unseen data, while underfitting is when a model is too basic to adequately represent the patterns in the data. This is frequently the result of insufficient model training time.
Preventing Underfitting and Overfitting
For preventing underfitting, the common methods used are the following: increasing model complexity, performing feature engineering, removing noise from the data, and increasing the duration of training. On the other hand, preventing overfitting can be done with preprocessing the data (for better quality), increasing the training data, reducing model complexity, and early stopping during the training phase.

Inferencing in AI

Inference in AI is the process of trained models, drawing insights or conclusions from new, unseen data. An AI model will first go through the training phase before inference. For instance, a healthcare application can train on past medical records and local health data to predict the possible diagnosis. Then the application can provide the next recommended steps based on the inferred diagnosis. The decision-making process based on the data is the inference done by the AI.

Batch inference

When processing vast volumes of data and efficiency is more important than quick results, batch inference is the type of inference suited for the job. It is more economical and efficient than real-time inference because it processes all the data at once.

Real-time inferencing

Real-time inference processes data quickly and efficiently, making it vital for applications such as fraud detection, speech recognition, and recommendation systems. It is more complicated and expensive than batch processing.

Chapter 1.3 Data in AI Models

Types of Data

Labeled data is when the data has been given a specific tag or category, showing the correct classification for each piece of information. The labels here are usually provided by a human annotator. This kind of dataset is used for supervised machine learning algorithms. The data is unlabeled if there is no category or labels assigned to it.

To further illustrate the difference between the two, consider a Filipino-language sentiment analysis tool for customer feedback in social media. If the data is labeled, the customer comments will be tagged whether they are positive or negative. Otherwise, the data will only compose all the customer comments without any kind of labeling or indicator.

Datasets

A machine learning dataset is a set of data used to train a model and instruct a machine learning algorithm on how to generate predictions.

Features and Labels

Features and labels are both in a dataset. **Features** are the unique attributes of the data that the model uses to predict labels. The **label** is the value or output that is aimed to be predicted.

For example, an application for farmers in the Philippines will use a supervised learning model to predict the yield of rice crops. The features can include variables that will influence the rice yield (temperature, rainfall, etc.) and the label is the output being predicted, in this case, the rice yield measured in kilograms per hectare.

Data Format

Structured Data
This type of data has a standardized format, usually tabular. This is easier to process due to its quantitative nature. It can be stored and managed using SQL (Structured Query Language) which lets the user define a data model called schema. Schema defines the rules such as fields, formats, and values for the data. Some structured data examples include: excel files, SQL databases, point-of-sale data, web form results, etc.
Semi-structured Data
Semi-structured data sits between the two types. Since it lacks a specific relational or tabular data model, it cannot be considered as structured. However, with its metadata it can still be analyzed. The metadata can be tags or similar markers. Semi-structured data examples include: JSON, XML, web files, email, and zipped files.
Unstructured Data
Unlike the first two, unstructured data has no set data model. It is qualitative data and requires different technologies to analyze it properly and effectively. Storing it in NoSQL database or data lakes are some of the examples. Other examples of unstructured data include: text files, video files, report, email, and images.

Chapter 1.4 Machine Learning Paradigms

Types of Machine Learning

Supervised Learning
Supervised learning algorithms are trained using sample data that consists of the matching output as well as the input. This type of machine learning has a wide variety of applications and is called as such because there must be a supervisor.

For example, the data could be different reviews about a restaurant, each of which are labeled with either negative, positive, or neutral. The model can use this as its guide and eventually recognize the labels associated with a corresponding group of comments.

The types of supervised machine learning are as follows:

- **Classification**
  This type of supervised learning method assigns labels or categories to unseen data using a learned model. Similar to the above mentioned example.
- **Regression**

This type of supervised learning method mainly uses continuous or numerical values with one or more input variables to predict. An example is using historical data of stocks to predict its future price.

Unsupervised Learning

In contrast, unsupervised learning algorithms are trained on unlabeled data. It finds structures, groups, or patterns in the supplied data without any prior information or guidance. Going back to the previous example (restaurant reviews), even without the labels (negative, positive, neutral), this kind of learning will be able to group the reviews into different categories.

The types of unsupervised machine learning are as follows:

- **Clustering**
  This type of unsupervised learning method classifies the data into "clusters" according to shared characteristics or the separations between individual data points. This will improve the model's comprehension of a specific cluster's properties.

- **Dimensionality Reduction**
  This type of unsupervised learning method keeps the most important data or patterns in a dataset while reducing the amount of features or dimensions.

Reinforcement Learning

Through interactions with the environment and the provision of rewards or penalties for actions, an agent learns to make decisions by trial and error. It will continuously improve by analyzing feedback from previous versions.

An example is when an agent is tasked to make stock trading decisions. The following is how it will work:

- The agent will take an action of either selling or buying a stock.
- Then the environment (stock market) will respond and the agent will observe the new stock price and its portfolio's value.
- The agent will get a reward if its action results in profit, otherwise it will be a penalty.
- Over time the agent will be able to strategize its actions to achieve profit (for rewards) while avoiding losses (and penalty).

  Reinforcement Learning from Human Feedback (RLHF)
  A technique wherein the machine learning model learns through human feedback. Its actions will then improve based on what people consider correct or not.

  For instance, a local restaurant chatbot app has a knowledge base for local Filipino dishes and the common phrases that Filipino customers use in a restaurant. This can include phrases such as, *"Pwede bang magdagdag ng kanin?"* (Can I add rice?), or *"Ano ang best-seller niyo?"* (What's your best-seller?). By giving users the choice to rate a chatbot's response with a thumbs up or down, the model can eventually learn and pick-up the Filipino expressions or slang that customers use when ordering.

Semi-Supervised Learning

Next is the semi-supervised learning which combines labeled and unlabeled data for training, often using a small quantity of labeled data to guide the learning process while improving accuracy using a larger set of unlabeled data. This learning is used when there is a lot of unlabeled data and labeling all of them will take too much work.

For instance, in a culinary school, the instructor demonstrates how to prepare and cook *adobo*, while the students follow along. After the session, they will be tasked to cook similar but varied dishes such as *sinigang*, *kaldereta*, or *kare-kare*. They must apply all the fundamentals and techniques taught to other dishes. The instructor's demonstration of cooking is the labeled data, while the assigned dishes represent the unlabeled data.

Self-Supervised Learning

Lastly, self-supervised learning is a kind of learning in which the model learns without the need for labeled data by creating its own labels from the input data by predicting certain portions of the data from other parts.

For example, a student is learning to read and write *baybayin*, an ancient Filipino script, on their own. They use a book containing phrases in both *baybayin* and Filipino, translating back and forth. By studying these translations, the student starts to recognize patterns and understand the script. Later, they practice on new phrases, attempting to translate them to and from Filipino and *baybayin*. In this process, the student uses the Filipino translations from the book as self-created "labels," which they then apply to interpret new, unseen phrases.

B. Identifying Practical Use Cases for AI

Chapter 1.5 Real-World AI Applications

For this part, we shall do a deep dive on how AI has become a transformative tool across several industries, which enables businesses to refine their operations for better productivity, an enhanced customer experience, and bring innovative solutions. Identifying practical use cases for AI involves recognizing areas where AI technologies can solve real-world problems effectively. Below are the main applications of AI, along with context related to Amazon Web Services (AWS).

Applications in Computer Vision

Computer vision is an AI technology that enables machines to automatically classify and accurately recognize real-world images. In today's world computer systems can access vast amounts of image and video data generated by our smartphones, traffic, security, and other variants of cameras.

The following are the real-world use-cases of Computer Vision, but are not limited to:

- **Security Surveillance** - Governments and other security enterprises use computer vision to improve the security of public or private assets, sites, and facilities.

  Examples: Computer-vision embedded cameras can send automatic intrusion alerts if an unauthorized individual enters a highly restricted area, improving the discretionary access control (DAC) of the facility.

- **Operational Efficiency** - Computer vision can be utilized to enhance business intelligence by securing the quality of company assets and generated products and maintaining employee attendance, creating operational efficiency. It can also be used in analyzing the images of customer demands to discover trends and patterns for target market behavior.

  Example:
  - Automatically identify the factory defects before exporting based on the structure of the components for each product.
  - Detects factory machine issues based on the structure of its parts.
  - Authenticate employees with facial recognition.
  - Analyze the behavior of customers by using social media images to discover trends.

- **Healthcare and Medical Imaging** - medical industries serve as one of the most prominent users of computer vision technology, due to its capability to assist medical physicians to speed up the process of accurate diagnosis. Having this technology will result in securing the correct treatment and increase life expectancy of the patients.

  Example:
  - Given the thousands of open-source tumor datasets, a computer vision model can be created to classify tumors in human skin based features.
  - Identify and detect bone issues based on X-ray analysis.
  - Classify potential symptoms based on MRI scans.

- **Autonomous Vehicles** - uses computer vision to analyze real-time images, which enables them to have a road awareness, such as detecting pedestrians, road signs, or obstacles based on their auto-generated maps as they traverse the road while driving automatically.

  Example: Given an autonomous car, which consists of multiple embedded cameras with computer-vision algorithms, it can have a 360-view of the objects moving, traffic signs, and pedestrians.

- **Semi Autonomous Vehicles** - other vehicle companies utilize self-autonomous vehicles using Machine Learning, to assist the driver to have a secured trip by alerting the driver for potential hazards due their behavior.

Example: Using computer-vision cameras on the driver-seat to detect driver behaviors and symptoms, where the vehicle can automatically stop to prevent unwanted car-crash.

- **Agriculture** - computer vision can be used to boost animal and vegetable and other plant-based product productivity by implementing intelligent automations.

  Example:

  - Using computer-vision to UAV footages to detect geographical information, soil and nutrient quality to identify how to cultivate the farming site.
  - Using computer vision to detect diseases based on plant leaves to provide initial treatment.
  - Using computer vision to monitor lifestocks, implementing smart-farming.

AWS Services related to Computer Vision

- **Amazon Rekognition** - is a service from Amazon Web Service (AWS), which aims to provide an easy-to-use computer vision image classification service, to be able to detect image objects based on common patterns.

- **Amazon SageMaker AI** - is a service from Amazon Web Services (AWS) that provides tools to build machine learning models, train, and deploy computer vision models at scale.

Natural Language Processing Use Cases

Natural Language Processing (NLP) is a branch of AI that focuses on enabling computers to understand the context behind natural human language, and generate similar language after learning. It helps produce better outcomes based on human requirements as it understands the context behind it. It became important to many industries due to its ability to fully and efficiently analyze text and speech data from different dialects, slang, and uncommon grammars from day-to-day transactions.

The following are the real-world use-cases of NLP, but are not limited to:

- **Automated Sensitive Data Censoring** - For businesses that require critical adherence to data privacy, such as those in the insurance, legal, and healthcare sectors, editing documents with sensitive information for censoring can be a tedious activity to do, especially if done manually. Hence NLP offers services to replace Personally Identifiable Information (PII) with filtered words.

- **Business Analytics** - using NLP tools to gain a smart interpretation of customer reviews towards the company product, and analyze which features to improve. For instance, finding a particular phrase within the feedback can predict a potential mood and emotions. This can be useful when setting up the Key Performance Indicators of the company.

- **Customer Engagement** - another use-case of NLP as it aims to create intelligent chat and voice agents to replicate human interaction when conversing with actual clients. This will reduce the business operational cost to a minimum.

- **Human Language Translation** - a use-case of NLP would include understanding the context in a foreign language, then translating it with respect to meaning, context, and sentiments to the target language. This helps businesses interact with foreign clients and be able to serve them accurately based on their needs.

AWS Services related to Natural Language Processing (NLP)

- **Amazon Comprehend** - An NLP service from Amazon Web Services (AWS) that uses deep learning to find insights within the human context.

- **Amazon Translate** - An efficient NLP translator from Amazon Web Services (AWS) that enables fast and affordable language translation.

- **Amazon Lex** - An NLP service provided by Amazon Web Services (AWS) that powers conversational interfaces using voice and text using chatbots.

Speech Recognition Systems

Speech Recognition Systems, commonly known as “Speech-to-Text”, is a speech recognition algorithm that translates the audio of spoken human language. It allows machines to process and respond to vocal inputs, which enables hands-free control and accessibility features. Specific applications nowadays can transcribe audio streams in real-time to display text.

The following are the real-world use-cases of Speech Recognition, but are not limited to:

- **Voice Assistants** - using Speech Recognition such as Amazon Alexa, Siri, and Google Assistant that analyzes the response from voice commands then performing such action or recommendation to it.

- **Transcription Services** - Converting speech audio into a text document for transcription, which can be beneficial for interviews.

- **Accessibility Tools** - Using speech-to-text services as an assistive technology for persons with movement disability to interact with their devices using voice.

AWS Services related to Speech Recognition

- **Amazon Transcribe** - a service provided by AWS that makes it possible to integrate efficient speech-to-text capabilities to applications.

- **Amazon Polly** - a text-to-speech service provided by AWS that turns text into human-like speech. This enables the service to talk to its users.

Recommendation Engines
Recommendation Engines analyze user behavior patterns from product to suggest products, services, or content that users are likely to be interested in. In return, this algorithm helps user experience by providing personalized recommendations, reaching the product that they really want or need on their end.

The following are the real-world use-cases of Recommendation Engines, but are not limited to:

- **E-commerce Personalization** - Recommendation Engines helps in suggesting products based on browsing and purchase history of the past customers.
- **Content Streaming Services** - Recommending movies, music, or articles tailored to user preferences.
- **Social Media Feeds** - Recommendation Engines helps in curating the content that aligns with user interest and engagement patterns.

AWS Services related to Recommendation Engines:

- **Amazon Personalize** - A machine learning service from Amazon Web Service (AWS) that makes it easy for developers to create unique recommendations for the application users.
- **Amazon SageMaker AI** - Can also be used to build, train, and deploy recommendation system models.

Fraud Detection Mechanism

Fraud detection mechanisms use AI to identify unusual patterns, which may result in anomalies or incidents of potential fraudulent activity. Nowadays, due to the vastness of data quantities derived from big data, AI models can accurately identify fraud and other suspicious transactions. This helped several organizations add an extra layer of security and adhere to transaction security guidelines provided by their government.

The following are the real-world use-cases of Fraud Detection Mechanisms, but are not limited to:

- **Financial Services** - for detecting credit card fraudulent transactions.
- **Insurance Claims** - for checking transactional patterns in insurance agencies to avoid fraudulent claims.
- **Cybersecurity** - for monitoring unauthorized access or activities within networks, commonly under Security, Orchestration, Automation and Response (SOAR).

AWS Service related to Fraud Detection Mechanisms:

- **Amazon Fraud Detector** - A fully managed Amazon Web Services (AWS) service that makes it easy to identify potentially fraudulent online activities such as payment fraud and fake account creation.

Forecasting Models

Forecasting Models uses historical time-series data to anticipate future trends and events. These help data analysts and data scientists to make informed decisions and actionable insights which they can use for anticipating demand, resource requirements, trend seasonalities, or market movements.

The following are the real-world use-cases of Forecasting Models, but are not limited to

- **Demand Planning** - predicting product demand to optimize inventory levels to avoid resource scarcity especially during high demands of a product.
- **Financial Forecasting** - estimating future revenue, expenses, and cash flows for the companies.
- **Resource Allocation** - anticipating manpower needs or energy consumption to avoid unwanted usage.

AWS Services related to Forecasting Models

- **Amazon SageMaker Canvas**: Enables users to build custom forecasting models designed to meet unique business requirements without needing deep machine learning expertise. It provides a no-code interface to create, train, and deploy predictive models for various forecasting use cases.

Chapter 1.6 When AI Solutions May Not Be Appropriate

Despite the advancements of AI for business and technology, let it be known that AI is not a one-size-fits all solution. To some extent, there are scenarios where AI implementation may not be appropriate, due to cost, complexity, or regulatory constraints. In fact, AI systems aren't always perfect in terms of their functionality. Hence having a clear understanding and business analysis to identify when to opt for traditional solutions over AI is important for effective decision-making.

Cost-Benefit Analysis of AI Implementation

A Cost-Benefit Analysis (CBA) of AI implementation to a business involves estimating the financial and operational cost associated with adopting AI technologies against the expected benefits. It involves analyzing the one-time expenses and recurring expenses within the development, deployment, and software

maintenance. This will then help the organization determine whether the investment in AI will yield a positive return and will contribute to their company mission.

Key Considerations

- **High Initial Costs** - Developing AI models can require significant investment in data collection, computing and storage services, and specialized expertise by the developer.
- **Operational Expenses** - It is expected that renting cloud computing and storage services will incur a cost.
- **Return on Investment (ROI) Uncertainty** - AI projects do not always guarantee ROI, hence a further analysis is needed.

When AI systems may not be cost-effective

- **Return on Investment (ROI) Uncertainty** - AI projects do not always guarantee ROI, hence a further analysis is needed.
- **Limited Data Availability** - during the data gathering, if the dataset for model training is significantly low, it can impact the ability of the model to predict accurately. An alternative would be using publicly-available datasets or utilize GAN models to create synthetic data.
- **Small Scale Operations** - for startups or small-scale businesses with starting operations, creating a forecasting model with historical data may not be sufficient yet and the costs may outweigh the benefits.
- **Alternative Solutions** - utilizing traditional software or manpower may achieve similar benefits with lower costs.

AWS Services that may help in the Cost-Benefit Analysis

- **Alternative Solutions** - utilizing traditional software or manpower may achieve similar benefits with lower costs.
- **AWS Services that may help manage investments for AI:**
    - **Amazon SageMaker AI** - despite that amazon sagemaker offers an extensive set of services for building, training, and deploying Machine Learning models, its computing costs and the developer team may take a significant revenue in AWS .

    - **AWS Cost Management** - to monitor and optimize AWS expenses, services such as AWS Cost Explorer may help.

- **Economies of Scale** - As AWS implements a Pay-As-You-Go model pricing, depending on the scale of resources used, can help manage costs but requires careful monitoring. Fortunately, as many companies use AWS services on their operations, the price of renting services went cheaper.
- **Total Cost of Ownership (TCO)** - AWS infrastructure can reduce the operating and maintaining costs compared to investing in on-premise solutions and at the same time, it consists of several AI-services as well.

Situations Requiring Deterministic Solutions

Deterministic systems are a type of solution that expects a consistent outcome every time a model is run with the same input. This is in contrast to traditional Machine Learning algorithms where it uses probability as an output and may represent a non-deterministic behavior due to randomized initialization and data variability.

Key Considerations

- **Predictability** - Critical systems require consistent outputs to ensure reliability and safety.
- **Explainability** - When the use case needs to have clear logic paths, which are essential for auditing and compliance.
- **Latency** - Real-time systems that cannot tolerate the processing overhead or delays coming from AI models.

Examples where deterministic solutions are preferred

- **Embedded Systems** - systems that are embedded within medical devices or aerospace, where consistent outcome is mandatory.
- **Financial Transactions** - systems used for client transactions require a consistent result to prevent errors.
- **Legal Compliance** - for creating IT Auditing reports, where security compliance details do not change, transparent, and justifiable.

Alternative AWS Services that assists in deterministic solutions

- **Rule-Based or Event-driven Systems**

- **AWS Lambda** - automatically runs the code in response to events, suitable for deterministic, rule-based processing.
- **Amazon EventBridge** - Facilitates event-driven architectures for deterministic workflows.

Limitations of AI in Regulated or Sensitive Areas

Certain organizations or industries are subject to strict regulations especially when handling and using data for actionable insights. These would include healthcare, finance, or governmental sectors. Hence, the usage of AI systems is limited to adhere to legal, ethical, and security protocols

Key Challenges

- **Data Privacy and Security** - Regulations such as Data Privacy act from the Philippines, General Data Protection Regulation (GDPR) of Europe, and Health Insurance Portability and Accountability Act (HIPAA) of United States which imposes stringent requirements on data handling.
- **Explainability and Accountability** - AI models, such as deep learning networks, usually result in "black boxes", where decisions cannot be explicitly tracked due to computational complexity.
- **Bias and Fairness** - AI models may inadvertently incorporate biases or discrimination present in training data, leading to unfair outcomes. Remember that the prediction which came from a Machine Learning model will heavily rely on the training data.

Examples of Limitations

- **Healthcare**
  - **Patient Data Protection** - AI systems must ensure compliance with health information and privacy laws, depending on where the organization and their clients are located.
  - **Diagnostic Accountability** - Clinicians need to understand AI recommendations to make informed decisions. Due to the potential "black boxes" from deep learning algorithms, precisely understanding the computation will be difficult.
- **Finance**
  - **Regulatory Compliance** - Financial AI models must adhere and be transparent with auditing and anti-fraud regulations.

- **Risk Management** - Unexplainable AI decisions can introduce unacceptable risks, which in return, AI cannot be used.

- **Government**
  - **Security Clearance** - AI models must be justified that it meets the security requirements for handling citizen information.
  - **Policy Adherence** - Government agencies may have policies restricting the use of certain AI technologies, due to its potential for data leak.

AWS Services that helps address AI Regulations

- **Compliance and Security Services**
  - **AWS Artifact** - an AWS security tool that automatically checks for compliance of AWS cloud infrastructure to several security frameworks such as General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), and Payment Card Industry Data Security Standard (PCI DSS).
  - **AWS Key Management Services (KMS)** - Allows for encryption key management, essential for protecting data for unauthorized users.
  - **AWS Identity and Access Management (IAM)** - Enables zero-trust policy, as it enables fine-grained access control to AWS cloud infrastructure resources.
- **Specialized Services**
  - **Amazon Macie** - a service from AWS that helps discover and protect sensitive data stored in AWS.
  - **AWS GovCloud (US)** - available only for regional areas of the United States, where it offers a region to host sensitive data and regulated workloads.
- **AI Explainability Tools**
  - **Amazon SageMaker Clarify** - a SageMaker tool from AWS, where it helps detect biases within Machine Learning models and provides explainability reports which end-users such as medical practitioners can use.

Limitations in AWS Services:

- **Shared Responsibility Model** - Despite the services that AWS can offer for secured and compliant AI implementation, they can only manage the security within their own physical infrastructure, but customers are responsible for securing their data and ensuring compliance.

- **Regulatory Restrictions** - Some regulations may prohibit cloud storage or require data to remain on-premise. Hence, a hybrid-approach can be an alternative.
- **Explainability Gaps** - some AWS services can provide a way to understand models created in AWS services, but explanation may not be sufficient for transparency.

Chapter 1.7 Capabilities of AWS Managed AI Services

This part discusses the capabilities of some of the AWS-managed services used for implementing AI systems. These are some of the examples of AWS services with use-cases, and will be discussed further in the later chapters of this book.

Generative AI Apps

Generative Artificial Intelligence (Generative AI) describes AI models that create new content like text, pictures, sounds and even complex data. This will change the course of many industries, such as entertainment, healthcare and finance. This will be further discussed in Chapter 2 of this book. **Amazon Web Services (AWS)** gives developers and organizations strong tools, services and frameworks. They use these to build, deploy and grow Generative AI applications safely and without issues. The following sample services for Generative AI apps are as follows:

Amazon Bedrock

A fully managed AWS service that provides an API to access the cutting-edge foundational models (FM) from leaning AI companies, where a user can build generative AI applications with security and responsible AI practices. Bedrock eases the use of these models in your applications. Customize and put in place AI solutions without requiring deep machine learning knowledge or managing complex systems.

Key Characteristics:

- **Accessibility to Foundational Models** - Has access to Foundational Models (FMs) from AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI and Amazon for works like writing text, summarizing and replying to questions.
- **Personalize** - Bedrock FMs permit modification with personal labeled information or extra training with unlabeled data to adjust models to your specific needs.
- **Retrieval Augmented Generation (RAG) Support** - where FM answers come from company knowledge sources, using these for the Amazon Bedrock feature.

- **Create AI Agents** - which are systems proficient in finishing tasks for the user without assistance. They organize and accomplish step-by-step tasks across your business systems, knowledge from RAG and use API calls.
- **Offers Security and Privacy** - this is for data usage to implement responsible AI practices, such as **guardrails** for safe model outputs.
- **Invocation Logging**: Amazon Bedrock enables you to track model invocations by gathering logs, input data, and output data from all invocations within your AWS account. This functionality is switched off by default but can be activated to direct logs to destinations like Amazon CloudWatch Logs or Amazon S3. Configuration can be done through the console or API, supporting logging for operations such as Converse, ConverseStream, InvokeModel, and InvokeModelWithResponseStream.

Use Cases

- **Content Generation** - By using prompts, bedrock can create articles, blog posts, and other forms of media.
- **Customer Support Chatbots** - receive intelligent responses with natural language context.
- **Inquiries derived from documents** - Create web apps that provide intelligent responses derived from business documents.

Amazon Bedrock offers a flexible pricing model designed to cater to different usage needs. There are two main pricing plans available:

1. **On-Demand and Batch Pricing**:
    - **On-Demand**: You pay for what you use, with no time-based commitments. Charges are based on the number of input and output tokens processed for text-generation models, and the number of images generated for image-generation models.
    - **Batch**: This is ideal for large-scale predictions. It enables you to submit a collection of prompts in a single input file and receive the responses as a single output file. Batch inference is offered at a rate 50% lower than on-demand inference.
2. **Provisioned Throughput Pricing:** This plan allows you to buy model units for a specific base or custom model, ensuring a guaranteed throughput capacity in exchange for a time commitment.

Amazon Q

A generative AI-powered assistant designed for assisting business organizations and software developers as well as providing company and software development knowledge. This must not be confused with Amazon QuicksightQ, where it serves as a search bar and Q & A experience.

Key Characteristics

- **Empowered by Amazon Bedrock** - a fully managed service that makes foundation models (FMs) available through an API.
- **Embedded security and privacy protocols** - to make it easier for organization implementation.

- **Amazon Q Products:**
  - **Amazon Q Business** - A type of Amazon Q product, where it serves as a generative AI-powered assistant that you can tailor to your business needs.
    - A fully-managed, generative-AI powered assistant that you can configure using **Amazon Q Business API** to answer questions, provide summaries, generate content, and complete tasks based on your enterprise data.
    - It allows users to attain immediate permissions-aware insights from enterprise data with citations coming from IT, HR, and Help Desk documentations.
    - It can be integrated with **Amazon Kendra, S3, Microsoft SharePoint,** and **Salesforce.**
  - **Amazon Q Developer** - an AWS generative AI assistant that helps you understand, build, extend, and operate applications and workloads on AWS.
    - A user can inquire or ask assistance about AWS Architecture, AWS resources, best practices, documentation references, and many more.
    - It can also be integrated within IDE to provide software development assistance.

Use Cases

- **Resume Enhancer** - Use Amazon Business Q to create a chatbot interface that specifically enhances the resume of an applicant based on the business requirements of the company they are applying to.
- **Programming Support from Amazon Q Developer** - Use Amazon Q Developer to assist developers in integrating modules to AWS.

Language AI

Language Artificial Intelligence (Language AI) includes various technologies and tools that help computers to understand, interpret, generate and reply to human language. It allows for real-time translations, where machines pull useful information from text. **Amazon Web Services (AWS)** provides many services and tools. These help create, launch and grow Language AI applications. They work safely and very well. AWS becomes a key resource in this field. The following sample AWS AI services for Language AI are as follows:

Amazon Transcribe

An AWS fully managed Automatic Speech Recognition (ASR) that helps people use speech-to-text features in apps easily. It is a powerful model with many advanced parameters that helps it function effectively. It works with many languages and audio types, allowing developers to include speech-to-text features in their apps.

Key Characteristics

- **Create effective voice technologies** - this is in your applications with Amazon Transcribe,as it can instantly convert real-time or recorded speech into text.
- **Trained on several audio data** - It is trained on audio data comprising millions of hours across several human languages and dialects.
- **Resilient to background noise** - It is aware of different language accents, noisy background, and other sound conditions which makes it produce accurate outputs for transcription.
- **Can transcribe several languages -** Can accommodate 100+ languages that make it easy to use and customize. It can also perform redaction of sensitive information, automatic language detection, content moderation, and custom language models.

Use Cases

- Enhance customer call experience and track insights, where this can be done using Amazon Transcribe Call Analytics and Amazon Connect Contact Lens to enhance customer experiences and increase agent efficiency by providing real-time or post-call conversation insights.
    - **Amazon Transcribe Call Analytics** - is a generative AI-powered API for generating highly accurate call transcripts and extracting conversation insights which can be used to discover areas for improvement.
    - **Amazon Connect Contact Lens** - this provides contact center analytics and quality management capabilities that helps an agent monitor, measure, and continuously improve contact quality and agent performance for better overall customer experience
- **Subtitles for videos and meetings** - as Amazon Transcribe can perform real-time transcription of audio, subtitles on demand can be possible for accessibility and better experience.
- **Audio detection of inappropriate content** - Amazon Transcribe can analyze the toxic languages and replace it with other content using **Amazon Transcribe Toxicity Detection.**
    - **Amazon Transcribe Toxicity Detection** - an ML powered voice-based toxicity detection that uses speech-based toxicity detection from the emitted audio language.
- **Clinical documentation** - Amazon Transcribe developed a tool for doctors like **Amazon Transcribe Medical** and **AWS Healthscribe.** This tool listens and changes spoken words into text during doctor talks for paperwork. It sends this text to health record systems for study. It also follows HIPAA rules for safety.
    - **Amazon Transcribe Medical** - a service that turns spoken words into written text, especially for health talk. Helpful in changing spoken words into text during chats about patient problems, care plans and clinical paperwork.
    - **AWS HealthScribe** - a tool for developers building health programs with speech detection and smart tech. It fits HIPAA rules and helps users write first clinical notes during doctor visits and other health meetings.

Amazon Polly

An AWS fully managed text-to-speech (TTS) service that uses deep learning techniques to synthesize speech that closely mimics human intonation and rhythm. It can convert articles, web pages, PDF documents into speech audio. Polly supports different voices from different languages with natural sounding voices.

Key Characteristics:

- **Can analyze various languages for TTS** - It offers many voices and languages covers as it has many languages and dialects with several voice choices for each.
- **Offers real-time streaming** - Provides low latency for real-time applications.

Use-cases

- **Voice-powered applications**
  - For mobile apps, especially when powered by AWS services, polly can be integrated for audio-based navigation.
  - For smart devices, voice responses can be provided by polly to enable your Audio-based IoT, Smart Home, and Wearable technology talk.
- **Accessibility**
  - **Assistive Technology** - polly can provide a significant help by verbally reading the content from websites, ebooks, and other digital media.
  - **Educational tools** - polly can be implemented on learning management systems for audio-based approach when studying.
- **Media and Entertainment**
  - **Audio-books** - For an audiobook application, polly serves an important role especially when a user prefers to listen than to read ebooks.

Amazon Lex

A fully managed AWS service to build conversation tools like chatbots and voice helpers inside apps. It uses deep learning tech, just like Amazon Alexa, allowing developers to create advanced natural language understanding and automatic speech recognition. Lex makes it easier to design, put into use and launch conversation agents on different platforms.

Furthermore, It easily connects with AWS Lambda, Amazon CloudWatch and other AWS services for added features, and available for multi-platform deployment, where we place bots on platforms like Facebook Messenger, Slack, Twilio SMS and others.

Key Characteristics

- **Natural Language Understanding (NLU)** - which understands what users want and takes important information from talks.

- **Automatic Speech Recognition (ASR)** - where it changes spoken words into written form for apps using voice.
- **Integration with AWS Services** - where it works smoothly with AWS Lambda, Amazon CloudWatch and other AWS tools for more features.
- **Enables multi-platform deployment** - where it puts bots on different platforms like Facebook Messenger, Slack, Twilio SMS and others.
- **Built-in security** - where it keeps data safe and private with AWS's strong systems.

Use-cases

- **Customer Support Chatbots** - Has 24/7 Support Deliver support, nonstop help by resolving frequent questions and problems.
- **Virtual Assistants**
  - **Employee Helpers** - Assist employees with HR questions, IT support and other internal matters.
  - **Personal Aides** - Supply scheduling, reminders and information finding for people.
- **E-Commerce**
  - **Shopping Helpers** - Lead users through product searches, suggestions and buying steps.
  - **Order Tracking -** Permit customers to ask about their order status through normal conversations.
- **Healthcare**
  - **Patient Interaction** - Set appointments, give medication reminders and share basic health data.
  - **Telemedicine Help** - Simplify initial patient checks and information collection before visits.

Amazon Translate

Another AWS fully managed, Neural Machine Translation service provides efficient language translation. Many languages and dialects are available where developers localize content, applications and websites for people around the world easily.

Amazon Translate connects simply with other AWS services as well. It has options to fit different business needs better.

Key Characteristics

- **Can translate in various languages** - provides many languages, even some that are not spoken often.
- **Performs in Real-Time and Batch Translation**, where it supplies both quick and bulk translation abilities.
- **Enables personalized translation** - where it permits unique words and flexible translation to fit trade-specific language.
- **Scalability -** as it manages big amounts of translation demands well.
- **Security and compliance** - where it keeps information secure with secret codes and follows various industry standards.

Use-cases

- **Website and Application Translation**
  - **Multilingual Websites** - Automatically change website text for visitors from different countries.
  - **Mobile Apps** - Provide in-app translations to reach more users.
- **E-Commerce**
  - **Product Descriptions** - Translate product details to draw in customers who speak different languages.
  - **Customer Reviews** - Change user comments to understand and address feedback globally.
- **Customer Support with Multilingual Support** - Allow support teams to talk with customers in their languages.
- **Travel and Tourism**
  - **Travel Guides** - Present translated travel details and guides for visitors.
  - **Booking Platforms** - Deliver multilingual help for booking services and talking with customers.
- **Healthcare**
  - **Medical Records** - Convert patient details and medical records to help communication among global healthcare providers.
  - **Patient Communication** - Aid in translating medical directions and information for patients with limited language skills.
- **Legal and Compliance**
  - **Document Translation** - Translate legal documents, contracts, and compliance materials precisely.
  - **Regulatory Filings** - Help in translating required papers for international regulatory needs.

Augmented Analysis

Augmented Analysis uses machine learning and artificial intelligence to improve extraction of data and find insights from it. It makes hard tasks easier, finds hidden patterns and offers useful insights. The following sample AWS AI services for augmented analysis are as follows:

Amazon Textract

Is an AWS fully managed machine learning service that extracts written words and details from scanned physical documents without needing human intervention.

It uses Optical Character Recognition (OCR), combined with machine learning systems to do more than just pull out text. It recognizes document layouts like forms and tables, allowing people to handle and examine many papers quickly.

Key Characteristics

- **Automatic Text and Data Extraction** - Pulls out printed words, handwritten notes and information from different kinds of papers.
- **Structured Data Extraction** - Finds and keeps the layout of papers, including forms, charts and key-value combinations.
- **Support for Multiple Formats** - Works with documents in styles like JPEG, PNG, PDF and TIFF.
- **Integration with AWS Services** - Easily connects with programs like Amazon S3, Amazon Comprehend and AWS Lambda for smooth workflows.
- **Highly Accurate** - Uses deep learning systems to promise high exactness in word and information pulling.
- **Scalable and Secure** - Automatically adjusts to handle changing tasks while keeping information safe with encryption and access controls.

Use Cases

- **Automated Document Handling**
  - **Invoice Management** - Takes important details like invoice numbers, dates and totals, automating accounts workflows.
- **Expense Report Handling** - Automates the pulling and checking of expense data from receipts and reports.
- **Data Entry Automation**
  - **Form Handling** - Changes paper or scanned forms into digital information, lowering manual data work.
  - **Customer Joining** - Pulls and checks information from identity papers, smoothing customer joining steps.
- **Content Handling**
  - **Digital Storage** - Changes physical papers into digital forms that are easy to search, improving document finding.
  - **Rules and Checks** - Automates pulling and checking of data from rule-related papers for checking needs.
- **Healthcare**
  - **Medical Record Handling** - Takes patient details from handwritten or printed medical records.

Amazon Augmented AI (A2I)

Amazon Augmented AI (A2I) is a fully managed service. It helps add human review to machine workflows. A2I permits groups to insert human analysis into automatic systems. This helps increase accuracy and trust where machine models might need more control. This mixed method uses AI's growth but also keeps the careful thought humans bring.

Key Characteristics

- **Human Task Integration** - Easily joins human review tasks with current machine learning processes.

- **Flexible Worker Options** - Uses Amazon Mechanical Turk, other vendors or private groups for human reviews.
    - **Amazon Mechanical Turk** - is a crowdsourcing marketplace from Amazon that simplifies the task for people and companies to delegate their work and duties to a group of workers who accomplish these tasks online.
- **Customizable Workflows -** Lets users create specific workflows for unique task needs and business rules.
    - **Built-In Quality Control -** Includes systems like multi-level reviews and validation steps for very good results.
    - **Scalability and Security -** Adjusts to handle different work levels while keeping data safe and following rules.
    - **Integration with AWS Services -** Connects smoothly with Amazon SageMaker, Amazon Textract and Amazon Comprehend for complete workflow automation.

Use Cases

- **Content Moderation**
    - **Social Media Platforms** - Automatically check and route likely bad content for humans to look at, keeping community rules safe.
    - **User-Generated Content Sites -** Review and approve uploads like pictures, videos and comments to keep the site honest.
- **Document Processing**
    - **Invoice Verification -** Combine automated data pull with verification to keep numbers right in money documents.
    - **Legal Document Review** - Help check contracts, deals and legal papers for correctness and rules.
- **Healthcare**
    - **Medical Record Analysis** - Add human skill to automated data pulled from health records for correct details.
    - **Clinical Trial Data Review -** Confirm data entries and results in clinical studies to keep research honest.
- **E-Commerce**
    - **Product Categorization -** Improve automatic product grouping with human monitoring to reach very good accuracy.

Amazon Comprehend

Is a fully managed Natural Language Processing (NLP) service that uses machine learning to find insights and connections in text. This service performs tasks like sentiment analysis, entity identification, key phrase finding, language detection and topic modeling. Amazon Comprehend helps groups study big amounts of

unstructured text data to get useful business knowledge and create data-driven decisions.

Key Characteristics

- **Sentiment Analysis** - Finds out if the feeling in text documents is good, bad, neutral or mixed.
- **Entity Recognition** - Identifies and sorts things like people, groups, places and dates.
- **Key Phrase Extraction** - gets important parts and ideas from text to sum it up.
- **Language Detection** - Can comprehend the language of the input text.
- **Topic Modeling** - Finds hidden subjects in many documents, which helps with sorting and study.
- **Custom Classification and Entity Recognition** - Trains special models to fit specific business goals using your own marked data.
- **Integration with AWS Services** - Works easily with tools like Amazon S3, AWS Lambda and Amazon QuickSight for smooth tasks and better data study.
- **Compliance and Security -** Keeps data safe with encryption both stored and during transfer and follows different industry rules.

Use Cases

- **Customer Feedback Analysis**
  - **Sentiment Tracking** - Study customer reviews to understand satisfaction and find areas for better service.
  - **Feature Extraction** - Find common themes in customer comments to guide product creation.
- **Content Organization**
  - **Document Categorization** - Sort and arrange large piles of documents by their content automatically.
  - **Metadata Tagging** - Improve search capabilities by taking out important metadata from documents and media. Metadata is data about data.
- **Compliance and Risk Management**
  - **Regulatory Compliance** - Monitors communications and documents to follow industry rules.
  - **Risk Assessment** - Identify potential risks by examining unstructured data for relevant indicators.
- **Healthcare**
  - **Clinical Data Review** - Find important facts in medical records and research reports. Clinical trials are tests done to study new treatments.
  - **Patient Feedback** - Study patient surveys to improve healthcare services.
- **Finance**
  - **Market Sentiment Study** - Track news and social media to understand market feelings. This helps in planning trades.
  - **Fraud Detection** - Study transaction details for signs of fraud.
- **Human Resources**
  - **Resume Screening** - Find important details in resumes. This helps hiring be more efficient.
  - **Employee Feedback** - Study employee surveys to make workplaces better and more productive.

- **Legal Services**
  - **Contract Study** - Find and study important parts of legal papers to meet rules and spot risks.
  - **Case Law Review** - Summarize legal cases to help legal study and decisions. A legal case involves a judge's decision on a matter.

Computer Vision

Computer Vision helps machines understand images and videos. AWS offers tools for computer vision, where these tools allow businesses to set up automated checks from visual media. They can also help improve security systems. The following sample AWS AI services for Computer Vision are as follows:

Amazon Rekognition

Amazon Rekognition is a fully managed Computer Vision service from AWS. Developers use this service to add image and video analysis to their apps. Computer Vision uses computers to understand images. Rekognition identifies objects, people, text, scenes and activities in pictures and videos.

It also finds inappropriate content. This service provides facial analysis and recognition for strong security and user checks. Rekognition is very scalable and secure. It is also easy to connect to other applications. Many industries find it useful for different tasks.

Key Characteristics

- **Image and Video Analysis**
- **Object and Scene Detection** - The system finds objects like cars, animals and scenes like beaches in images and videos.
  - **Activity Detection** - The system sees actions in videos, such as running, jumping or dancing.
  - **Facial Analysis** - The system spots faces and details like age, gender, emotions and facial features.
  - **Facial Recognition** - The system matches faces to identify people in images and videos.
  - **Text Detection on Images** - The system can extract and read text in images and videos in different languages and fonts.
  - **Content Moderation** - The system spots bad or unsafe content, like violence, to keep material safe.
- **Real-Time and Batch Processing**
  - **Streaming Video Analysis** - The system checks live video for fast information and alerts.
  - **Batch Processing** - The system looks at many videos or images at once.
- **Customization and Integration**
  - **Custom Labels** - Users train Rekognition to find special objects and scenes important for their business needs.
  - **API and SDK Support** - APIs and SDKs are available to connect easily with apps on different platforms and languages.

    - **Integration with AWS Services** - Works well with Amazon S3 for storage, AWS Lambda for computing and Amazon CloudWatch for monitoring.
- **Security and Compliance**
    - **Data Privacy** - All data is encrypted both when stored and when shared.
    - **Compliance Certifications** - Meets industry rules and regulations like GDPR and HIPAA.
- **Scalability and Performance**
    - **Automatic Scaling** - The system adjusts to handle different workloads without problems.
    - **High Accuracy** - Advanced deep learning methods provide very accurate analysis results.

Use Cases

- **Security and Surveillance**
    - **Access Control** - Use facial recognition for secure building entrances in offices, schools and other places. Facial recognition software identifies people by their faces.
    - **Threat Detection** - Check live camera footage to spot unusual activities or people without permission.
- **Content Moderation**
    - **Social Media Platforms** - Automatically block offensive or bad content to follow community rules. Community rules are guidelines for proper behavior.
    - **User-Generated Content Sites** - Examine and control photos and videos added by users to keep the site safe and trustworthy.
- **Retail and E-Commerce**
    - **Customer Analytics** - Study customer actions in stores by observing their paths and product interests. This helps arrange store layouts better.
    - **Personalized Shopping Experiences** - Identify returning shoppers with facial recognition to customize their visit based on past choices.

Customer Experience

Amazon Personalize

It focuses more on having a personalized experience by helping developers create custom recommendations even without ML expertise. It uses technology derived from Amazon.com for having personalized product recommendations, personalized search results, and personalized marketing. Businesses typically utilize Amazon Personalize to improve user engagements and satisfaction for its highly relevant content.

Key Characteristics

- **Real-Time Personalization** - The system suggests items and changes content quickly based on user interactions.
- **Automated Machine Learning** - The software manages data cleanup and model teaching to improve results.

- **Scalable Infrastructure** - The setup supports big systems to serve millions of users and activities.
- **Customizable Algorithms** - It provides different algorithms and models to match unique company requirements.
- **Integration with AWS Ecosystem** - It works well with AWS tools like Amazon S3, AWS Lambda and Amazon DynamoDB. AWS stands for Amazon Web Services, which offers cloud computing.
- **User Segmentation** - The feature groups users by actions and likes to target specific groups.

Use-Cases

- **E-Commerce Recommendations**
  - **Product Suggestions** - Show items to users based on what they have browsed or bought before.
  - **Upselling and Cross-Selling** - Offer additional or premium items to raise the total order cost.
- **Media and Entertainment**
  - **Content Suggestions** - Give personalized movie, music or article ideas matching user likes and habits.
  - **Tailored Playlists** - Create playlists fitting each user's preferences and listening patterns.
- **Personalized Marketing**
  - **Targeted Campaigns** - Send market messages and deals to specific groups of users.
  - **Email Customization** - Adjust email text and item suggestions to improve interaction and buying rates.
- **News and Publishing** - Recommend news stories and blog posts to readers based on their likes and past reads.
- **Travel and Hospitality**
  - **Travel Tips** - Propose places, hotels and activities based on what the user likes and past trips. Travelers should have options that suit them.
  - **Custom Trip Plans** - Create special travel plans that really improve the user's trip experience. Better plans help with travel enjoyment.
- **Education**
  - **Course Suggestions** - Offer courses and study materials that match student habits and interests.
  - **Custom Learning Paths** - Design personalized study plans that improve learning results for students.
- **Gaming**
  - **Game Suggestions** - Recommend items, challenges or content based on how players act and what they like.
  - **Player Retention Plans** - Customize game experiences to keep players interested and coming back.

Amazon Kendra

Amazon Kendra is an intelligent search tool that uses machine learning. It offers very accurate and relevant search results. It searches through both organized and unorganized data sources. Organizations use Kendra to

create strong search abilities that understand natural language questions. Users find information quickly and easily.

Key Characteristics

- **Natural Language Understanding (NLU)**
  - **Conversational Queries** - Users search with simple conversations. This makes their interaction easy and more natural.
  - **Query Suggestions** - Users get smart suggestions for their searches. These suggestions depend on what they type and their current context.
- **Wide Data Source Integration**
  - **Pre-Built Connectors** - The system links easily to well-known data sources. Sources include Microsoft SharePoint, Amazon S3 and others.
  - **Custom Data Sources** - Specialized data storage can also integrate for custom use cases. This connection happens through APIs and SDKs.
- **Relevance Tuning**
  - **Custom Ranking** - Companies choose which documents or data sources appear first. This really helps users find better results. Results become more relevant.
  - **Contextual Understanding** - Machines learn to grasp what queries mean. This understanding helps give users the most accurate answers.
- **Faceted Search and Filtering**
  - **Advanced Filtering** - Offers different ways to narrow down search results. Users sort these by specific features or groups.
  - **Dynamic Facets** - Creates useful filters by looking at the search content and context.
- **Continuous Learning**
  - **Feedback Loop** - The system learns from user actions and feedback. It improves search accuracy over time. This helps keep searches relevant.
  - **Analytics and Insights** - The platform provides detailed reports. These reports show search queries, user actions and result success. This information helps with future improvements.

Use Cases

- **Enterprise Knowledge Management**
  - **Intranet Search -** Kendra can improve search tools inside the company. It helps employees locate documents, rules and other important information fast.
  - **Knowledge Bases** - Help support teams and help desks with precise information. This can lessen time to retrieve results.
- **Customer Support**
  - **Self-Service Portals** - Customers search through FAQs and manuals. They find answers themselves and feel happier. Support costs go down as a result.
  - **Agent Assistance** - Support staff get quick access to needed information. This speeds up problem resolution.
- **Healthcare**

    - **Medical Research** - Professionals search through huge collections of medical papers. They discover important details for patient care and research.
    - **Patient Information Retrieval** - Clinicians need fast access to patient records, Kendra can help provide quick access to patient records, medical histories, and medication protocols which are needed for decision making.
- **Financial Services**
    - **Regulatory Compliance** - Compliance officers explore regulations, policies and legal papers efficiently through efficient search.
    - **Investment Research** - Analysts can efficiently search for financial reports, market data and research articles. These sources inform their investment decisions, where analysts rely on accurate information.
- **Legal Services**
    - **Case Law Search** - Allows lawyers to search through regulations, policies and legal papers efficiently.
    - **Contract Management** - People search contract terms and clauses quickly. This speeds up contract reviews. The process becomes really efficient.
- **Education**
    - **Academic Research** - Students and researchers discover academic papers, journals and educational resources.
    - **Course Materials** - Course content, lectures and study materials become easier to find. This helps students and teachers. Learning experiences improve significantly.
- **Media and Publishing**
    - **Content Discovery** - Editors and writers look through archives and articles. They check multimedia content for creating and curating new content.
    - **Digital Libraries:** Users search for books and articles with ease.
- **Manufacturing**
    - **Product Documentation** - Employees find manuals and technical details fast. This helps them work better.
    - **Knowledge Sharing** - Teams share and find best practices. They can access processes and guidelines easily.
- **Retail and E-Commerce**
    - **Product Information Search** - Customers can easily find product details and specifications. They check availability on different platforms.
    - **Inventory Management** - Staff manage inventory data effectively. Stock tracking and replenishment improve.
- **Government and Public Sector**
    - **Public Information Services** - Efficient search to access open-source government documents, public records and information for both citizens and officials.
    - **Policy Analysis** - Policymakers need to read policy documents, reports and legislative texts. This helps in making better decisions.

Amazon Connect

Amazon Connect is a cloud service from Amazon Web Services (AWS). It helps businesses offer better customer service at a lower price. This service is easy to set up and scalable. Amazon Connect provides a smooth managed contact center solution that can handle customer calls and chats. This enables customers to have a personalized experience without the need of complex IT resources.

Key Characteristics

- **Cloud-Based and Scalable**
  - **No Infrastructure to Manage** - Being cloud-native, Amazon Connect eliminates the need for on-premises hardware, allowing businesses to deploy a contact center quickly without significant upfront costs.
  - **Automatic Scaling** - Seamlessly scales to handle varying call volumes, ensuring that customer service operations remain efficient during peak times.
- **Omnichannel Support**
  - **Voice and Chat Integration** - Customers can use voice calls and chat through Connect. This service combines both methods.
  - **Integration with Customer Relationship Management (CRM) and Other Systems** - The system works well with CRM tools. Agents easily see customer information and history. They can now work more effectively with these insights.
- **Interactive Voice Response (IVR) and Routing**
  - **Customizable IVR** - Create and set up IVR systems with a visual tool without coding needed. This allows for personalized and smooth call routing.
  - **Skill-Based Routing** - Routes customer calls to the right agents based on their skills. This may lead to quicker resolutions and satisfied customers.
- **AI and Machine Learning Integration**
  - **Amazon Lex Integration -** Uses Amazon Lex to build chatbots that manage simple questions. This lets agents focus on harder tasks.
  - **Amazon Polly Integration** - Supports the use of Amazon Polly for natural-sounding voice responses. It improves the IVR experience.
  - **Real-Time Analytics and Insights** - Uses AWS analytics to understand customer interactions and agent results. This helps to measure contact center efficiency.
- **Security and Compliance**
  - **Data Encryption** - Data stays encrypted when stored or moved. This really keeps customer information private and safe.
  - **Industry Compliance Certifications** - Connect follows many industry rules like GDPR, HIPAA and PCI DSS. This helps contact centers stay within legal boundaries.
- **Agent Experience**
  - **Agent Desktop** - Agents use an easy-to-use, flexible desktop interface to handle interactions smoothly.
  - **Real-Time and Historical Metrics** - The system shows live dashboards and past reports. Managers watch performance closely.

- **Cost-Effective Pricing**
  - **Pay-As-You-Go** - Businesses only pay for what they use. There are no long-term commitments or upfront fees.
  - **Flexible Licensing** - Offers different licensing choices. Companies of any size, from startups to large firms, find a suitable option.
- **Global Reach**
  - **Multi-Region Deployment** - Contact centers can now exist in many AWS regions worldwide. This ensures that service to customers is low-latency and highly reliable.
  - **Language Support** - Supports many languages and dialects. Businesses serve diverse customer groups successfully.

Use Cases

- **Customer Support**
  - **Automated Call Routing** - IVR systems send customers to the right department based on their questions. This reduces wait times which helps solve issues efficiently.
  - **24/7 Support Availability:** Deploy contact centers that can work all day and night. They provide consistent customer assistance without office hour constraints.
- **Sales and Marketing**
  - **Proactive Customer Outreach** - Utilize outbound calls to reach new customers. Calls promote products and ask for feedback.
  - **Lead Qualification** - Automated chatbots can interact with website visitors. They check leads and transfer them to sales teams for later contact.
- **Healthcare**
  - **Patient Appointment Scheduling:** Automated phone and chat systems help schedule appointments. They also send reminders. This approach reduces missed appointments and keeps patients involved.
  - **Telehealth Support:** An integrated contact center helps with telehealth services. It manages bookings, answers questions and follows up on appointments.
- **Financial Services**
  - **Fraud Detection Alerts** - Add automated calls and messages to warn customers about suspicious account activities to mitigate fraud and suspicious transactions.
  - **Account Management** - Customers receive help with account questions, balance checks and transaction details. Automated interactions make this process fast and easy.
- **Retail and E-Commerce**
  - **Order Management** - Deal with order questions, tracking details and returns through voice and chat. This service creates a smooth customer experience. Customers benefit from clear communication.

Business Metrics

Business performance improves through artificial intelligence. It allows advanced data analysis, predictive modeling and automated reporting. AWS offers many AI-driven tools. These tools watch and study key performance indicators well. Organizations benefit from using these AI tools from AWS. They gain a deeper understanding of their performance metrics. They really learn more about how they perform. They also optimize operations. Strategic initiatives move forward with confidence.

Amazon Fraud Detector

A fully managed service by AWS, with Machine Learning support. Businesses use it to protect against fraud, like payment and payment fraud, account hijacking, and fraudulent content. Companies don't need to know much about machine learning for this to be implemented. Fraud Detector simplifies building and using fraud detection models easily. It offers a large-scale and safe way to keep transactions and user actions secure.

Key Characteristics

- **Pre-Built Fraud Detection Models**
  - **Out-of-the-Box Models** - Use ready-made models for common fraud situations. This reduces the need for much customization.
  - **Domain-Specific Models** - Utilize models for specific fields like banking, e-commerce and gaming. This approach increases detection precision.
- **Custom Model Creation**
  - **Custom Rules and Models** - Develop personal fraud detection models with your data. This helps to solve unique fraud issues.
  - **Feature Engineering** - Automatically creates useful features from transaction data. This process improves models. Better features lead to better models.
- **Real-Time Detection**
  - **Immediate Analysis** - Check transactions instantly at real-time to spot fraudulent actions.
  - **Low Latency** - Keep transactions fast and smooth, where users experience almost no delays.
- **Integration with AWS Services**
  - **Simple Connection** - It connects smoothly with services such as Amazon S3 for storing data, AWS Lambda for computing without servers and Amazon CloudWatch for keeping an eye on things and sending alerts.
  - **APIs and SDKs** - offer easy tools for integrating with current apps and workflows.
- **Security and Compliance**
  - **Data Safety** - Data stays safe with encryption both when stored and moving. It meets industry rules.
  - **Access Control** - uses fine-grained security access controls to protect important information. This stops unauthorized persons from getting in.

Use Cases

- **Payment Fraud Detection**
  - **Transaction Monitoring** - Examine payment actions immediately. Spot strange activities, like odd spending habits or risky places, to stop fake charges. Fast action is key. Stop fraud before it happens.
  - **Chargeback Prevention** - Reduce financial losses by detecting and stopping fraud processes early.
- **Account Takeover Prevention**
  - **Login Anomalies** - Fraud detectors can recognize unusual login patterns that indicate that it is a compromised account, which may additionally require more authentication steps.
  - **Credential Stuffing Detection** - Detects and blocks login attempts, guard against automated attacks such as brute force.

Amazon SageMaker Canvas

Amazon SageMaker Canvas is a robust service offered by Amazon Web Services (AWS) that enables businesses to create machine learning models for various forecasting and predictive tasks without needing coding skills or machine learning expertise. It features an intuitive, no-code interface for building, training, and deploying machine learning models, allowing businesses to make data-driven decisions easily. SageMaker Canvas harnesses the full potential of AWS's machine learning technology, enabling users to generate highly accurate forecasts based on their historical data.

Key Characteristics

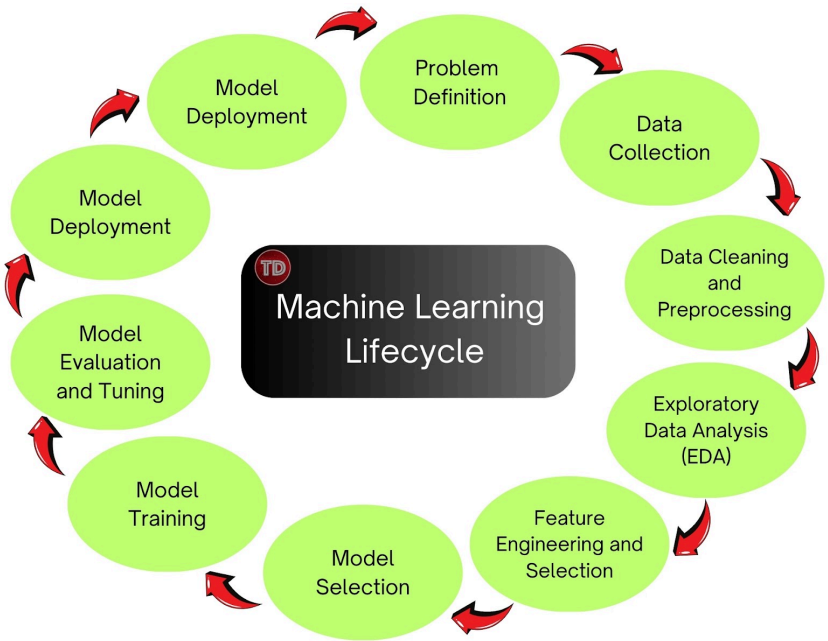
- **No-Code Interface**
  - SageMaker Canvas provides an intuitive, visual interface that enables users to build, train, and deploy machine learning models without any programming experience. This feature broadens access to machine learning capabilities for a wider audience.
- **Automated Model Training**
  - The service streamlines the machine learning workflow, covering data preparation, feature selection, and model training, which facilitates more accurate predictions.
- **Pre-built Algorithms**
  - Amazon SageMaker Canvas uses a range of pre-configured algorithms that are optimized for forecasting and prediction tasks, helping to generate highly accurate results quickly.
- **Customizable Forecasting**
  - Users can customize their models by incorporating specific business factors and indicators, improving the relevance and accuracy of predictions.
- **Seamless Data Integration**
  - It allows user to integrate seamlessly with existing data sources, including AWS Data Lakes, databases, and third-party applications, ensuring that users can fully utilize their data.

- **Accurate Predictions**
  - The service delivers precise forecasting results by utilizing the power of Amazon's advanced machine learning models and data processing capabilities.

Use Cases

- **Sales and Revenue Forecasting**
  - Predict future sales performance to optimize pricing strategies, inventory levels, and marketing efforts, ensuring more accurate financial planning.
- **Demand Planning**
  - Create reliable forecasts to help businesses plan for future demand, ensuring the right amount of stock and resources are available at the right time.
- **Inventory Management**
  - Reduce stockouts and excess inventory by forecasting demand patterns, helping to streamline operations and improve supply chain efficiency.
- **Financial Forecasting**
  - Forecast future revenues and expenses to improve budgeting and ensure better cash flow management.
- **Supply Chain Optimization**
  - Forecast the flow of goods and materials to improve supply chain management, reducing delays and optimizing logistics.
- **Marketing Campaign Analysis**
  - Use historical data to forecast the impact of upcoming marketing campaigns, allowing businesses to better plan their advertising spend and resource allocation.

C. The Machine Learning Development Lifecycle



Chapter 1.8 Components of a Machine Learning Pipeline

Data Collection

This is the first step in the machine learning pipeline. It is where the data to be processed and used is gathered. The quality of the dataset gathered will matter the most in determining whether good insights can be found after applying a machine learning algorithm to it.

Data Cleaning

This next process is done to make sure that the dataset is free or errors, inconsistencies, or missing data before being analyzed.

Exploratory Data Analysis (EDA)

In this process, the datasets will be analyzed before applying any method. This is to get a summary of their main characteristics. One of the common methods is using visualizations to see patterns, trends, and relationships.

Data Pre-processing Techniques

These techniques are done in order to further improve the quality of data after it is cleaned.

Feature Engineering

Feature engineering selects and transforms variables when creating a model. It includes feature selection, feature transformation, feature creation, and feature extraction.

- **Feature Selection**
  - Selection of relevant data attributes or variables during the development of a predictive model. This will contribute to minimizing the error rate of the model.
- **Feature Transformation**
  - This step includes replacing missing features or invalid features. Techniques involved can be forming Cartesian products of features, non-linear transformations, and creating domain-specific features.
- **Feature Creation**
  - This is where creation of new features from the existing data takes place. Examples include: one-hot-encoding, binning, splitting, and calculated features.
- **Feature Extraction**
  - This is where the amount of data to be processed is reduced with dimensionality reduction techniques. The benefit of this step is to reduce the amount of memory and computing power required without compromising the original data characteristics.

Model Building

After selecting a machine learning algorithm to be used to solve the given problem, it will be trained on the dataset collected. Initial training is performed, and then the model is validated. If it passes, it will be tested using different evaluation methods.

Model Evaluation Methods

Model Validation

The step where the model performance and accuracy are determined whether they are appropriate for the problem being solved during training.

Model Evaluation

Model evaluation is the final assessment of the model whether it will do a good job in predicting the label or target on new and unseen data. Performance metrics are used to measure or evaluate the performance of the model.

Performance Metrics

- **Area Under the ROC Curve (AUC)** - A performance metric for classification models that measures the model's ability to discriminate between classes at different thresholds.
- **Mean Absolute Percentage Error (MAPE)** - The average percentage difference between the predicted and actual values.
- **Mean Squared Error (MSE)** - Measures the average squared error between predicted and actual values in regression models.
- **Recall** - Measures how well an algorithm identifies or predicts positive instances (true positives) in a datasets.
- **Perplexity** - Measures the probability that a language model will generate a given sequence of words.
- **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** - Measures how well the generated summary matches the reference summary by calculating their overlapping words.
- **Bilingual Evaluation Understudy (BLEU)** -Measures the quality of machine-translated text by comparing its output with human reference translations.
- **BERTScore** - Measures the quality of text generation models by comparing their output to reference text at the word (or token) level using BERT embeddings.

Hyperparameter Tuning

This is a process for improving the performance of a machine learning model by optimizing the parameters that regulate the learning process.

Model Deployment

After passing the evaluation metrics set, the model can now be integrated into practical applications. Usage of cloud services such as AWS is recommended for deployment.

Monitoring and Maintenance

With the use of Amazon CloudWatch, including CloudWatch Alarms, and other relevant tools, the model deployed can be monitored and made sure to be free of errors.

Monitoring Policy Violations with Amazon CloudWatch Alarms

It is important to quickly identify and respond to policy violations. By setting up Amazon CloudWatch alarms, you can receive immediate notifications whenever these violations occur, enabling prompt corrective actions.

These alarms are carefully configured to monitor specific policy breaches and alert as soon as they are detected. This proactive approach ensures the integrity and compliance of your AI models, keeping them aligned with the desired guidelines and standards.

Integrating Amazon CloudWatch alarms into your monitoring strategy provides a strong framework for protecting against undesirable outputs and maintaining high standards of content quality.

Methods for Deploying a Model in Production

Managed API Service

It enables you to deploy your trained model on a fully managed platform that takes care of the underlying infrastructure, scalability, and maintenance. For instance, Amazon SageMaker Endpoints offers an easy way to deploy machine learning models for real-time inference. This service automatically scales to meet demand, ensuring low-latency predictions without the need for you to manage the infrastructure. Once the model is deployed, you can send data to it via HTTP requests and receive predictions in response, streamlining its integration into production applications. Managed services also offer the convenience of automated patching, versioning, and monitoring, which helps reduce operational complexity.

Self-hosted API

A self-hosted API requires more hands-on management but offers greater flexibility regarding infrastructure control. You have the option to host the model on your own servers or in a cloud environment. Frameworks such as Flask, FastAPI, or Django can be used to create an API endpoint. This API will receive input data, pass it through the model, and return predictions.

Many organizations choose to use Docker containers to package models along with their dependencies. These containers can then be deployed to container orchestration platforms like Kubernetes, which facilitate scalability and management. This approach provides more control over resource allocation, deployment configuration, and custom monitoring. However, it also demands additional effort to manage infrastructure, security, and scaling.

Chapter 1.9 Sources of ML Models

Open Source Pre-trained Models

These models are trained on large datasets in order to accomplish a specific task. They can help a developer build AI applications faster since they are already pre-built. These are the models typically used for transfer learning. The open source kinds are the ones that are available for free use to the public.

Some areas in which pre-trained models are helpful include: Natural Language Processing (NLP), speech AI, Computer vision, healthcare, cybersecurity, and art.

Training Custom Models

Custom models are used when building custom AI solutions. A developer can gather their own datasets and train a machine learning model on it to solve a specific problem a pre-trained model can't.

Chapter 1.10 AWS Services for Each Stage of the ML Pipeline

Amazon SageMaker AI

An Amazon SageMaker Model Building Pipeline is a series of connected steps created with a JSON schema or Pipelines SDK, organized as a directed acyclic graph (DAG). This structure shows dependencies, with each phase's output potentially serving as the next phase's input, defining the pipeline's execution order. In other words, Amazon SageMaker AI can be used for building, training, and deploying machine learning (ML) models.

Data Preparation

Amazon SageMaker Feature Store

A managed repository called Amazon SageMaker Feature Store was created to help teams store, exchange, and oversee machine learning (ML) features. Throughout the ML lifecycle, it facilitates simple synchronization and scaling by guaranteeing high-quality, consistent features for training and real-time inference.

Amazon SageMaker Data Wrangler

For easier data preparation (text, graphic, and tabular data), Amazon SageMaker Data Wrangler can help lessen the time needed from weeks to minutes. Without the need for code, it provides a visual interface for selecting, importing, and transforming data with SQL and built-in transformations.

Geospatial ML with Amazon SageMaker AI

Of course, data scientists can also create, train, and implement models using geographical data thanks to Amazon SageMaker AI's geospatial machine learning features. For effective, large-scale geographic machine learning, it consists of pretrained models, specific processing tools, visualization features, and access to geospatial data sources.

Building

Amazon SageMaker Notebooks

The SageMaker Notebooks can help start a fully managed JupyterLab IDE for data, code, and notebooks. With scalable computational resources that can be adjusted on demand, this web-based interface offers access to ML tools across SageMaker and AWS, covering the entire ML workflow from data preparation to model deployment.

Amazon SageMaker Jumpstart
Pretrained models for tasks like image generation and article summarizing are available through Amazon SageMaker JumpStart, an ML hub that speeds up model selection, customisation, and deployment. With options for sharing artifacts and controlling model visibility, it allows for the safe and private management of data within the user's VPC (virtual private cloud).

SageMaker Studio Lab
This is based on the JupyterLab IDE which allows users to access JupyterLab right in their browser. Projects and datasets are also saved in the cloud with its 15 GB storage.

Training

Amazon SageMaker Model Training
By offering scalable, high-performance infrastructure without requiring direct management, Amazon SageMaker Model Training helps cut down on time and expense for large-scale machine learning model training. In order to effectively manage big models and datasets, SageMaker automatically scales resources and provides distributed training libraries. This enables customers to pay only for what they use and train models for extended periods of time with no interruption.

Amazon SageMaker MLflow
For more efficient machine learning and generative AI experimentation, Amazon SageMaker AI also offers managed MLflow. Administrators may create safe, scalable MLflow settings on AWS for effective experiment management and model selection, while data scientists can quickly train, register, and deploy models.

Amazon SageMaker HyperPod
By automatically distributing training among thousands of AI accelerators for improved model performance, Amazon SageMaker HyperPod streamlines the setup and optimization of machine learning infrastructure. This includes being able to train models for weeks or months without disruption saving training time up to 40%.

Deploy

Amazon SageMaker Model Deployment
Amazon SageMaker AI streamlines the deployment of ML models, including foundation models. SageMaker is a fully managed service that integrates with MLOps technologies to assist expand installations, lower costs, and effectively maintain models in production.

Amazon SageMaker Pipelines

A serverless workflow orchestration solution for MLOps and LLMOps automation, Amazon SageMaker Pipelines makes it simple to create, implement, and monitor machine learning workflows. It scalability to

perform hundreds of concurrent ML operations in production using a Python SDK or drag-and-drop user interface.

Amazon SageMaker Edge

Users can construct new models or retrain models using real-world data with the SageMaker Edge Agent, which enables data and metadata collecting based on predefined triggers. Analyses like model drift analysis can also be supported by this collected data.

Amazon SageMaker Real-Time Inference

Ideal for scenarios requiring low-latency responses or high data throughput. It maintains a persistent endpoint (REST API) managed by SageMaker to handle continuous traffic effectively.

Amazon SageMaker Serverless Inference

Perfect for handling sporadic or unpredictable traffic patterns. This option allows you to deploy models without worrying about the underlying infrastructure. SageMaker automatically scales based on the rate of requests.

Amazon SageMaker Batch Transform

Best suited for large-scale data processing in offline mode. This feature doesn't need a constant endpoint and can process datasets of significant size, often in gigabytes.

Amazon SageMaker Asynchronous Inference

Primarily designed for requests that involve large payloads and require extended processing times. It supports payloads up to 1 GB and can accommodate processing times of up to one hour.

Amazon SageMaker Endpoints

It is used to invoke a trained machine-learning model. After training and deploying your model, you can utilize the endpoint to obtain real-time predictions by submitting input data and receiving the corresponding predictions in return.

Amazon SageMaker Model Monitor

It is a service that helps track and maintain the quality of machine learning models deployed to production. Once a model is deployed via SageMaker Endpoints, Model Monitor continuously monitors its performance by detecting issues such as data drift, bias, and prediction errors. It automatically analyzes the model's input data and output predictions, comparing them to expected results, and sends alerts if any significant deviations or performance issues are detected. This ensures that models remain accurate, fair, and reliable over time, allowing businesses to take corrective actions when necessary.

Chapter 1.11 Key Machine Learning Concepts

Models

Machine learning models are effective instruments for task automation that increase productivity and accuracy.

Multimodal models

Models designed to handle and analyze inputs from multiple sources, including text, images, audio, and video.

Model Latent Space

This is where machine learning models convert input data into simplified patterns or features (feature representations) that will be used to generate outputs.

Model Fit: Overfitting and Underfitting

Overfitting

If the model is performing well on training data but not on evaluation data, it is overfitting. The model is only memorizing the training data and is unable to generalize to new data.

Underfitting

If the model is performing poorly on the training data, it is underfitting. It is unable to capture the trends or the relationships between the input and target values. This could be because the model is too simple to describe the target well.

Bias and variance

Bias

Bias is the degree of which the model's forecast/prediction differs from the actual value in relation to the training dataset.

- **Low bias**
  When a model has low bias, it makes few or no inaccurate assumptions about the data's underlying relationships. The model is adaptable enough to capture actual patterns in data and can accommodate complex relationships well.
- **High bias**
  The model makes more assumptions about the intended result or outcome. A model with a high bias makes significant assumptions about the link between input attributes and output. It oversimplifies the problem, neglecting crucial patterns in the training data.

Variance

Variance reflects how much the estimate of the target function would change if different training data were utilized. It measures inconsistency of different predictions using different training sets. This means it is not a measure of the accuracy for the model as a whole.

- **Low variance**
  When a model has low variance, it indicates that its predictions are fairly stable. The predictions remain consistent regardless of the exact data used to train the model. The model does a good job of remaining close to its predictions, so even if different samples of data are chosen, the results will be consistent.

  Machine learning algorithms under this include: linear regression, logistic regression, and linear discriminant analysis.

- **High variance**
  A model with a high variance is more sensitive to changes in the training data. If different sets of data are utilized, the model's predictions can vary significantly. The model is overly tuned to the characteristics of the training data and struggles to generalize to new data because it picks up on noise and minor details rather than the overarching trend.

  Machine learning algorithms under this include: decision trees, support vector machines, and k-nearest neighbors.

Fine-tuning

Also known as, *'Transfer learning'* is a method where a model trained on one dataset is reused for a new dataset, instead of starting from scratch. This approach allows for reliable models to be built from smaller datasets with less training time. It is called *fine-tuning* since the model is pre-trained to fit the new dataset.

Instruction-based fine-tuning

Technique for improving the large language model's (LLMs) performance on a specific task through usage of labeled dataset. The dataset consists of instructional prompts and corresponding outputs. This will help the pre-trained models adapted for practical use.

An example is when a chatbot aims to answer questions about Philippine history and culture in Tagalog. Based on this context, an example process could be:

1. **Gathering of datasets** - a large number of Tagalog questions with their corresponding answers on Philippine history and culture.
2. **Creation of instructional prompts** - a prompt that clearly specifies the task of the LLM:
   a. Prompt: "Answer the following questions about Philippine history in Tagalog."

   b. Question: “SIno ang unang Pangulo ng Pilipinas?” (Who was the first president of the Philippines?)
   c. Answer: “Si Emilio Aguinaldo ang unang Pilipinas.” (Emilio Aguinaldo was the first president of the Philippines.)
3. **Fine-tuning of the LLM** - training of the LLM on the dataset with the instructional prompts and corresponding outputs.

Embedding

It is possible to insert words or images into numbers to enhance the comprehension of artificial intelligence (AI) and machine learning (ML) systems. This will help the systems comprehend difficult ideas similar to how people learn and think.

Retrieval Augmented Generation (RAG)

A method that can optimize the output of a large language model (LLM) by referencing a knowledge base with company-specific or industry-specific data.

Generative Pre-trained transformers (GPT)

Advanced neural network architectures designed to comprehend and produce human-like text and content. Pre-trained on extensive language datasets, these models are capable of generating responses that are both coherent and contextually relevant.

Chain-of-Thought

A machine learning strategy that encourages models to create reasoning steps before reaching a final response to enhance decision-making.

For instance, a model is built to deal with the frequent typhoons and floods in the Philippines. It will be the Disaster Response System which will predict the necessary resources and actions required for effective disaster response and relief management.

If chain-of-thought is applied to this model, the model will simulate the step-by-step reasoning when analyzing a variety of disaster-related data, such as weather forecasts, population density, and infrastructure resilience. In other words, the model will not only provide predictions, but also the logical steps behind its predictions.

Prompt Engineering

Prompt Engineering is a technique that guides the generative artificial intelligence (generative AI) solutions to produce intended results. This technique creates clear, specific inputs or "prompts" that will guide the AI model to provide accurate and appropriate responses for the given task or instruction. Doing this effectively is essential to achieve desired responses from the model.

Negative Prompts

Prompts that indicate what content has to be left out of the output that is generated.

Prompt Injection
A security vulnerability that takes place when harmful or malicious inputs are added to prompts to manipulate the output of language models.

Prompt templates
Prompt templates are predefined formats that can be used to standardize inputs and outputs for AI models.

Prompting Types

Zero-shot prompting
Zero-shot prompting utilizes the model's generalization capabilities which attempts to execute new tasks without being given specific training or examples. With this prompting technique, it requires the creation of prompts that clearly explain the task and the intended output format.

For example, a well-engineered prompt could be: "Classify the following Filipino celebrations/commemorations as either historical, cultural, or religious: 'The colorful street parades and traditional dances of Cebu's Sinulog Festival honor the Santo Niño." The model can then give a response from either of the three given choices: historical, cultural, or religious.

Few-shot prompting
Few-shot prompting is also known as *"in-context learning"* in which examples are provided to help the AI model have a context of the intended output. The shot is an example which includes the input and desired output.

Using the zero-shot's prompting example to work with, a well-engineered prompt for this one could be: "Classify the following Filipino celebrations/commemorations as either historical, cultural, or religious. Here are some examples:

The Day of Valor (Araw ng Kagitingan) is celebrated every April 9 which commemorates the fall of Bataan to Japanese troops during World War II.

Answer: Historical

The Feast of the Immaculate Conception is a holiday every December 8.
Answer: Religious

The Panagbenga festival is a festival about celebrating the history of Baguio, Cordilleras, and their flora.
Answer: Cultural". This can provide more context to the model which it can use as its basis for more accurate response.

Chain-of-thought prompting
This technique divides complex reasoning tasks into small, easy steps. Then, the model follows a clear path through this process. This method increases the model's ability to reason well.

For example, a model needs to study the reasons for flooding in a Philippine city. The aforementioned method divides the problem into smaller parts:

1. It finds natural causes like typhoons and monsoon rains.
2. Then, it looks at human actions like cutting down trees and not properly managing waste.
3. Finally, it evaluates how good the local drainage systems are.

Self-refine prompting
This model solves a problem step-by-step. First, it creates an initial solution, then it reviews and corrects this solution to boost its accuracy or quality. The process continues until it meets a specific stopping condition.

For instance, the model will help create disaster response plans for typhoons in the Philippines. A model could suggest an evacuation plan for a coastal village.

1. It points out issues like not enough shelters or confusing communication.
2. Then, it changes the plan to solve these problems.
3. This process repeats until the plan is safe and works with local resources and time.

Domain 1: AI and ML Fundamentals Sample Questions

1. A healthcare insurance company manually extracts sensitive information from claims forms and accompanying attachments. This manual process has led to significant delays for customers seeking healthcare benefits. To improve customer service and reduce the manual labor involved, the company wants to automate the extraction process to expedite the handling and processing of claims.

   Which AWS service will help meet the company's objectives?

   a. Amazon Comprehend
   b. Amazon Textract
   c. Amazon Personalize
   d. Amazon Lex

Explanation:
Amazon Textract is a machine learning service designed to automatically extract text and structured data from scanned documents. Unlike traditional optical character recognition (OCR) tools, Textract goes beyond simple text extraction to understand the layout and context of documents, including forms and tables. It identifies and processes various elements such as fields, checkboxes, and tables with unparalleled precision, enabling businesses to convert complex documents into highly accurate, actionable data. This advanced capability makes Textract especially useful for handling intricate forms and attachments, such as insurance claims, where precision and structured data extraction are crucial.

Hence, the correct answer is: Amazon Textract.

- Amazon Comprehend is incorrect. This service is primarily used for natural language processing (NLP) tasks such as sentiment analysis, entity recognition, and language detection. It is not designed to extract data from forms and tables.
- Amazon Personalize is incorrect because it is a service that only provides personalized recommendations and user experiences by analyzing user behavior and preferences. It is primarily used to create recommendations for applications such as e-commerce or content platforms. Thus, this service does not address the need to extract and process text from documents or forms.
- Amazon Lex is incorrect. This option is for building conversational interfaces using voice and text. It is used to create chatbots that can interact with users through natural language. Additionally, it focuses on dialogue management and natural language understanding rather than document text extraction.

2. An e-commerce company is developing a model using Amazon SageMaker AI to forecast the probability of a product being returned after purchase. The company owns a labeled dataset containing product categories, prices, customer reviews, and return status stored in an Amazon S3 bucket.

What machine learning approach is most appropriate for this task?

a. **Supervised Learning**
b. Unsupervised Learning
c. Few-shot Learning
d. Transfer Learning

Explanation:

Supervised Learning is a machine learning approach where a model is trained on a labeled dataset. This means the dataset contains input-output pairs where the desired output is known. The key advantage of Supervised Learning is that it allows the model to learn the relationship between input features like product categories and prices and output label return status, which enables accurate forecasting of future events.

Hence, the correct answer is: Supervised Learning.

- Few-shot learning is incorrect. This machine learning method is only used when the data available for training is limited. It enables models to generalize from just a few examples.
- Transfer learning is incorrect because this is generally used when you have a pre-trained model that can be fine-tuned on a specific task. In this scenario, the focus is on using a labeled dataset to predict a specific outcome, making supervised learning a better fit.
- Unsupervised learning is incorrect. This method is only used for analyzing and clustering data without labeled outputs. It is often applied in scenarios where the goal is to discover hidden patterns or groupings within the data, such as customer segmentation or anomaly detection.

References for Domain 1

Large Language Models
https://aws.amazon.com/what-is/large-language-model/

Small Language Models
https://aws.amazon.com/blogs/industries/opportunities-for-telecoms-with-small-language-models/

Models and Algorithms
https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/

Common Algorithms in AI/ML
https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Underfitting and Overfitting
https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Inferencing in AI
https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Batch inference
https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Types of Data
https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Data Formats
https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Capabilities of AWS Managed AI Services
https://aws.amazon.com/ai/services/

Model Deployment
https://aws.amazon.com/ai/services/

Open source pretrained model
https://blogs.nvidia.com/blog/what-is-a-pretrained-ai-model/

Amazon SageMaker
https://aws.amazon.com/sagemaker/features/

Amazon SageMaker Feature Store

https://aws.amazon.com/sagemaker/feature-store/

Amazon SageMaker Data Wrangler
https://aws.amazon.com/sagemaker/data-wrangler/

SageMaker Notebooks
https://aws.amazon.com/sagemaker/notebooks/

SageMaker Jumpstart
https://aws.amazon.com/sagemaker/jumpstart/

SageMaker Studio Lab
https://studiolab.sagemaker.aws/

SageMake Model Training
https://studiolab.sagemaker.aws/

SageMaker MLflow
https://aws.amazon.com/sagemaker/experiments/

SageMaker HyperPod
https://aws.amazon.com/sagemaker/hyperpod/

SageMaker Model Deployment
https://aws.amazon.com/sagemaker/deploy/

SageMaker Pipelines
https://aws.amazon.com/sagemaker/pipelines/

SageMaker Edge
https://aws.amazon.com/sagemaker/edge/

Neural Network
https://aws.amazon.com/what-is/neural-network/

Neurons (A)
https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/

Activation Function (G)
https://www.v7labs.com/blog/neural-networks-activation-functions

Recurrent Neural Network (RNN)

https://aws.amazon.com/what-is/recurrent-neural-network/

Applications in Computer Vision
https://aws.amazon.com/what-is/computer-vision/

Natural Language Processing Use Cases
https://aws.amazon.com/what-is/nlp/

Speech Recognition Systems
https://aws.amazon.com/what-is/speech-to-text/

Fraud Detection Mechanism
https://aws.amazon.com/solutions/ai/fraud-detection/

Situations Requiring Deterministic Solutions
https://www.sciencedirect.com/topics/engineering/deterministic-solution#:~:text=Deterministic%20solution

Amazon Bedrock
https://aws.amazon.com/bedrock/

Amazon Q Business
https://docs.aws.amazon.com/amazonq/latest/api-reference/Welcome.html

Amazon Q Developer
https://docs.aws.amazon.com/amazonq/latest/qdeveloper-ug/what-is.html

Amazon Q Use Cases
https://www.linkedin.com/pulse/my-experience-amazon-q-use-cases-ofir-nachmani-bg0ae

Amazon Transcribe Call Analytics
https://aws.amazon.com/transcribe/call-analytics/

Amazon Connect Contact Lens
https://aws.amazon.com/connect/contact-lens/

Amazon Polly
https://aws.amazon.com/polly/

Augmented Analysis
https://docs.aws.amazon.com/whitepapers/latest/architecting-hipaa-security-and-compliance-on-aws/textract.html

Amazon Augmented AI (A2I)
https://aws.amazon.com/augmented-ai/
https://docs.aws.amazon.com/augmented-ai/

Amazon Comprehend
https://aws.amazon.com/comprehend/
https://docs.aws.amazon.com/whitepapers/latest/architecting-hipaa-security-and-compliance-on-aws/amazon-comprehend.html

Amazon Rekognition
https://aws.amazon.com/rekognition/
https://docs.aws.amazon.com/whitepapers/latest/architecting-hipaa-security-and-compliance-on-aws/amazon-rekognition.html

Amazon Kendra
https://aws.amazon.com/kendra/
https://docs.aws.amazon.com/kendra/

Amazon Connect
https://aws.amazon.com/connect/
https://docs.aws.amazon.com/whitepapers/latest/amazon-connect-data-lake-best-practices/amazon-connect.html

Amazon Fraud Detector
https://aws.amazon.com/fraud-detector/
https://docs.aws.amazon.com/frauddetector/latest/ug/what-is-frauddetector.html

Others
https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-a-prompt.html#few-shot-prompting-vs-zero-shot-prompting
https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-prompt-engineering.html
https://aws.amazon.com/compare/the-difference-between-machine-learning-supervised-and-unsupervised/
https://aws.amazon.com/what-is/overfitting/
https://aws.amazon.com/what-is/embeddings-in-machine-learning/
https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/feature-engineering.html
https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html
https://www.geeksforgeeks.org/ml-semi-supervised-learning/
https://www.geeksforgeeks.org/self-supervised-learning-ssl/
https://www.ibm.com/topics/instruction-tuning
https://www.mastersindatascience.org/learning/difference-between-bias-and-variance/#:~:text=A%20high%20bias%20model%20typically,which%20makes%20them%20learn%20fast.

https://aws.amazon.com/textract/features/
https://aws.amazon.com/solutions/case-studies/anthem/https://aws.amazon.com/solutions/case-studies/anthem/
https://tutorialsdojo.com/amazon-textract/
https://docs.aws.amazon.com/sagemaker/latest/dg/algorithms-choose.html#algorithms-choose-supervised-learning
https://docs.aws.amazon.com/sagemaker/latest/dg/canvas-storage-configuration.html
https://tutorialsdojo.com/amazon-sagemaker/
https://docs.aws.amazon.com/bedrock/latest/userguide/guardrails.html
https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model-options.html
https://docs.aws.amazon.com/bedrock/latest/userguide/model-invocation-logging.html

FUNDAMENTALS OF GENERATIVE AI

Understanding the Basics of Generative AI

Capabilities and Limitations of Generative AI

AWS Infrastructure and Technologies for Generative AI

A. Understanding the Basics of Generative AI

Chapter 2.1: Foundational Concepts of Generative AI

Tokens

Tokens are the smallest textual units that an AI model can comprehend, such as words, letters, or punctuation. The process of dividing text into small digestible portions for the model to process is known as tokenization.

- **Word tokens** - Each word in a text is considered as a token
  - **Word:** "Kamusta, mga kababayan!" (Hello, fellow countrymen!)
  - **Tokenized:** ["Kamusta", ",", "mga", "kababayan", "!"]
- **Subword tokens** - To manage a large vocabulary and uncommon terms, words are divided into smaller parts.
  - **Word:** "Kabayanihan" (Heroism)
  - **Tokenized:** [kaba, yan, ihan]
- **Character tokens** - Individual characters are used as tokens.
  - **Word:** Bahay (House)
  - **Tokenized:** [B, a, h, a, y]

Chunking

Chunking is the division of texts into smaller, more digestible sections, which makes the process of understanding them more straightforward. In other words, the text will be divided into meaningful phrases or chunks.

For instance, the sentence, "Si Andres Bonifacio ay isang bayani ng Pilipinas." (Andres Bonifacio is a hero of the Philippines.) can be divided into three chunks:

- Chunk 1: "Si Andres Bonifacio"
- Chunk 2: "ay isang bayani" (is a hero)
- Chunk 3: "ng Pilipinas" (of the Philippines)

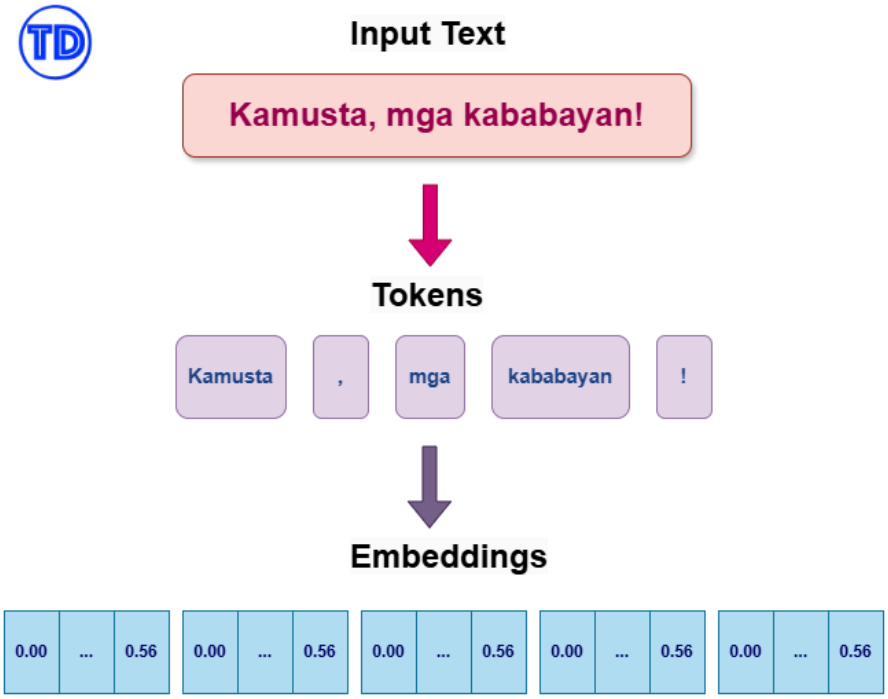
There are also **different methods of chunking**.

- **Fixed-sized chunking** - simplest of all wherein the number of tokens in a chunk is decided, and whether there should be overlap between them.
- **"Content-aware" chunking** - taking advantage of the nature of the content.
  - **Sentence splitting** - can be done by splitting sentences by period (.) and newlines or using Python libraries such as Natural Language Toolkit (NLTK) and spaCY.
  - **Recursive chunking** - breaks the text into smaller parts step by step, using specific separators repeatedly and in an organized way.

- **Specialized chunking** - a type of chunking method that preserves the original structure of the content during the process.
- **Semantic chunking** - produces chunks composed of sentences that discuss the same theme or issue.

Embeddings

Embeddings are used to extract the semantic meaning from the given data. As discussed in the previous domain, it is possible to insert words or images into numbers to enhance the comprehension of artificial intelligence (AI) and machine learning (ML) systems. This will help the systems comprehend difficult ideas, similar to how people learn and think.



Vectors and Their Applications

In relation to the aforementioned concept, embeddings encode all types of data (text documents, rich media, audio, etc.) into vectors. These vectors capture the meaning and context, allowing the search for neighboring data points. This is where the term ***'vector search'*** comes in.

Vector databases hold high-dimensional vectors and allow for efficient, fast retrieval of nearest neighbors via k-NN indexing techniques such as Hierarchical Navigable Small World (HNSW) and Inverted File Index (IVF). They also include data management, fault tolerance, authentication, and a powerful query engine. These databases can be used for vector search and have been recently used with generative AI to create intelligent agents that will provide a conversational search experience.

Types of Generative AI Model

Large Language Models (LLMs)

Large Language Models (LLMs) are deep learning models. They are trained on large volumes of data. Neural networks comprising an encoder and a decoder make up the transformer structure used in these models. Transformers are unique because they can learn on their own without continual guidance.

Transformer-Based LLMs (Large Language Models)

Firstly, what is transformer architecture? It is a type of neural network architecture that focuses on understanding the relationships between words or parts of data by paying attention to everything at once, which makes it faster and more accurate for tasks like translating languages or understanding text. Other applications include text generation, question answering, text summarization, speech recognition, and music production.

The following are use cases of different LLMs that are transformer-based:

- BERT (Bidirectional Encoder Representations from Transformers) for the improved Google search engine.
- GPT-3.5 and GPT-4 are generated by OpenAI and are human-like responses.

Foundation Models

With Foundation Models (FMs), data scientists don't need to create artificial intelligence from scratch. Large datasets are used to train these deep learning networks, speeding up and lowering the cost of model creation. "Foundational model" refers to the machine learning models that can execute different tasks such as natural language conversation, text and image generation, and language interpretation. These models are fine-tuned for various tasks such as natural language processing (NPL), image recognition, analytics, etc.

Different **examples of FMs** are the following:

- Bidirectional Encoder Representations from Transformers (BERT)
- Generative Pre-trained Transformer (GPT)
- Amazon Titan
- AI21 Jurassic
- Claude
- Cohere
- Stable Diffusion
- BLOOM
- Hugging Face

Fine-tuning Approaches

- **Domain adaptation fine-tuning** - is appropriate for adjusting a language model that has already been trained to particular kinds of text input, such as product descriptions from online stores. This method uses the dataset of current product descriptions to modify the model's comprehension of domain-specific language and terminology.
- **Instruction-based fine-tuning** - this is intended to improve a model's performance on particular tasks by offering prompt-response pairings that are labeled examples. As opposed to domain-specific text production, this method works better for activities that call for precise instructions and answers.

This is different from fine-tuning done on models. Fine-tuned models are used for task-specific requirements, while foundation models are used for general tasks since they are a broad language model.

Multi-Modal Models

Multi-model models are AI systems that can analyze and comprehend many data types simultaneously, such as text and images, to produce more precise and insightful answers. These models are useful for a lot of different industries that need more standardization in their data formats. These models analyze and generate text, images, and audio. An example of a multi-modal model is the DALLE-E2 neural network model developed by OpenAI, which can generate high-quality images from textual descriptions, and vice versa.

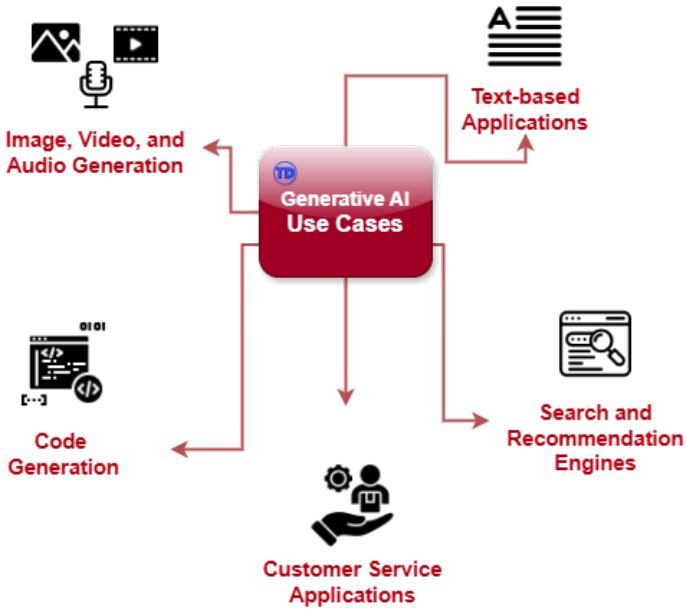
Diffusion Models

These models are able to produce new data, such as photographs, that resemble the ones they were trained on. In order to eventually transform random noise into realistic data, they first learn by introducing noise to the original data and then training to eliminate the noise. It is also used in DALLE-E2, an image generation model.

Stable Diffusion

Stable Diffusion is a generative AI model that uses text or image prompts to produce lifelike graphics.

Chapter 2.2 Potential Use Cases for Generative AI Models



Image, Video, and Audio Generation

- DALL-E2, as previously mentioned, is an example of an image generation model.
- Gen-3 Alpha is used in the Runway application, which is available for web and iOS and has text-to-video, image-to-video, and text-to-image tools.
- MusicLM is developed by Google, which can create music based on textual prompts.

Text-Based Applications

- GPT's models can perform text summarization from a given passage, as well as function as conversational agents. Translation of languages is also available with this model.
- Gemini models are accessible and used in Vertex AI can also perform text summarization, classification and extraction.

Code Generation

- Github Copilot users will be able to select different AI models, such as Anthropic's Claude 3.5 Sonnet model or Google's Gemini 1.5 Pro model, as alternatives to OpenAI's GPT-4o. This can help developers in their software development process as it can perform code completion, function as copilot chat, and many more.

Customer Service Applications

- Salesforce Einstein AI, which is a generative AI for CRM (Customer Relationship Management) auto-generates customer replies.
- AI-assisted agents as they type their responses to their customers.

Search and Recommendation Engines

- Microsoft Bing integrated their search engine with its AI chatbot, Copilot.
- Google is powered by Gemini and has Search Generative Experience (SGE), which can give users AI-generated insights at the top of their search results.

B. Capabilities and Limitations of Generative AI

Generative AI

Generative AI is an advanced type of artificial intelligence that produces new content, including text, photos, videos, and music. It is utilized to boost creativity and automate content creation in sectors including marketing, healthcare, and entertainment. Businesses can include generative AI capabilities for activities like text summarizing and image production into their applications with tools like AWS's SageMaker AI.

Chapter 2.3 Advantages and Capabilities of Generative AI

- **Adaptability** - adapt various tasks and domains.
- **Responsiveness** - generate content in real-time.
- **Simplicity** - simplify complex tasks.
- **Data Efficiency** - learn from relatively small amounts of data.
- **Creativity and Exploration** - generate novel ideas, designs, or solutions.
- **Personalization** - create personalized content tailored to individual preferences or characteristics.
- **Scalability** - generate large amounts of content quickly.

Chapter 2.4 Limitations and Challenges of Generative AI

Toxicity

This is the amount of sexual references, rude, unreasonable, hateful, or aggressive comments, profanity, insults, flirtations, attacks on identities, and threats in a model. Additionally, this is culturally and contextually dependent.

Hallucinations

This is when a model generates incorrect, inaccurate, or misleading results. Several factors, such as insufficient training data, incorrect model assumptions, or biases in the data, can be the cause.

Intellectual Property

For creators looking to protect their work from unauthorized use and exploitation by AI systems, generative AI's intellectual property concerns present a major issue.

- **Plagiarism** - It is simple to intentionally or inadvertently cause these applications to reuse other people's creations, which implies that the "new" and "generative" works produced by these AI programs might actually be the creations of another person.
- **Cheating** - the idea that students (and possibly other professionals) are and will utilize generative AI and artificial intelligence tools "to cheat" is one of the most common worries nowadays.
- **Disruptions of nature work** - generative AI has the potential to be even more disruptive than the Industrial Revolution was at its inception.
- **Nondeterminism** - It refers to the inability to produce the same output consistently given the same input. This characteristic arises because generative models, like large language or diffusion models, often rely on stochastic processes during their inference or training.

Chapter 2.5 Factors for Selecting Appropriate Generative AI Models

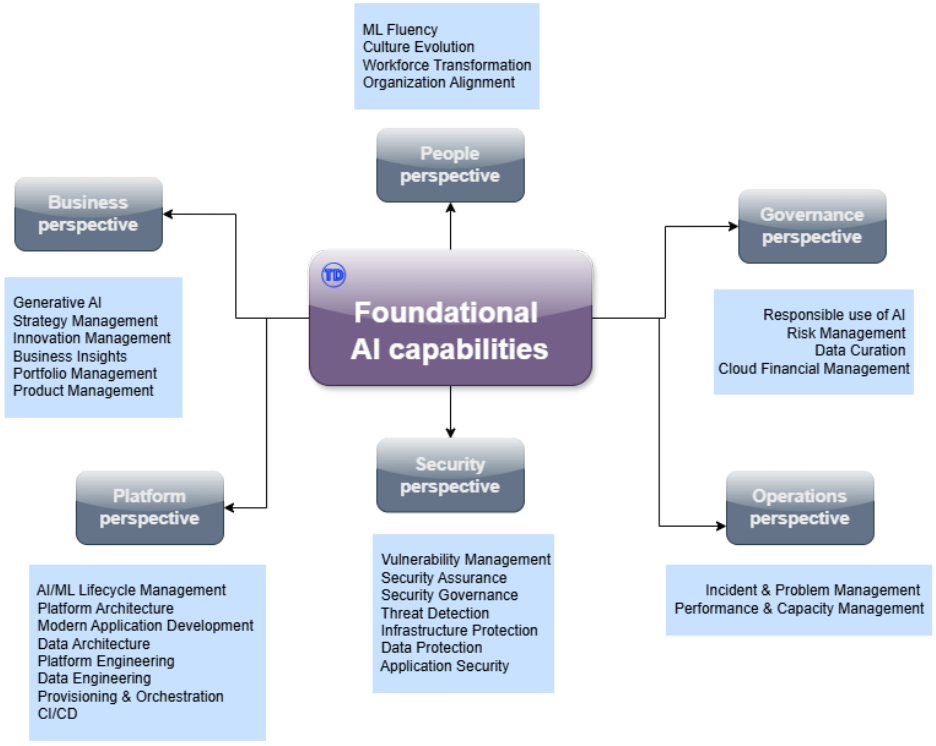
- **Performance requirements** - organizations may optimize the potential of AI and generative AI technologies for improved outcomes by establishing appropriate Key Performance Indicators (KPIs), closely monitoring them, and using the insights to make necessary modifications.
- **Constraints** - Generative AI models have constraints as mentioned above, that one should consider before choosing which model to use for their problem.
- **Capabilities** - also mentioned above, Generative AI has a lot of capabilities and one should choose according to the problem they are trying to solve.
- **Compliance** - Generative AI systems may not always be created with legal and regulatory standards in mind, which might result in infractions and penalties if exploited. Caution must be exercised while deploying generative AI to ensure compliance with all applicable rules and regulations.
- **Cost** - Generative AI models are priced differently and organizations should measure which model fits both their budget and problem statement.

Chapter 2.6 Business Value and Metrics for Generative AI Application

- **Cross-domain performance** - measures the capability of Generative AI models to perform effectively across different domains without retraining. This can include natural language processing (NLP), computer vision, and multi-modal tasks.

- **Efficiency** - it focuses on how well resources, such as computing, memory, or storage, are utilized to generate AI results with minimal overhead. AWS provides various tools to optimize the efficiency of Generative AI models.
- **Conversion rate** - it measures the effectiveness of AI applications in achieving user interactions that lead to goals, such as purchases, subscriptions, or engagement. AI models trained on user behavior data help optimize these rates.
- **Average Revenue per User (ARPU)**- this calculates the revenue generated per user for services that leverage Generative AI. ARPU can be enhanced through personalized recommendations or content generation using AI models.
- **Accuracy**- reflects how close AI predictions or outputs are to the expected results. AWS provides tools like Amazon SageMaker AI and Amazon Bedrock to evaluate and improve model accuracy through testing and fine-tuning.
- **Customer Lifetime Value (CLV) -** evaluates the total revenue a customer produces throughout their relationship with a company. Generative AI applications, such as personalized marketing, help increase customer engagement and CLV.

Chapter 2.7 The AWS CAF-AI Foundational Capabilities



The following list shows the foundational capabilities relevant to AI adoption.

Business perspective

This perspective makes it happen that the organization's investments in AI will advance their goals for digital and AI transformation as well as their commercial results. Chief financial officer (CFO), chief operations officer (COO), chief information officer (CIO), chief technology officer (CTO), and chief executive officer (CEO) are examples of common stakeholders.

- *(NEW)* **Generative AI** - Make use of large AI models' general-purpose capabilities.
- **Strategy Management** - Use machine learning and artificial intelligence to create new commercial value.
- **Innovation Management** - Oversee products that are data-driven and AI-enabled.
- **Business Insights** - AI's ability to make predictions based on historical data or provide answers to unclear inquiries.
- **Portfolio Management** - Determine which high-value AI projects and products are possible and prioritize them.
- **Product Management** - Challenge established market theories and innovate the organization's current business.

People perspective

This perspective bridges the gap between AI technology and business and seeks to develop a culture of ongoing learning and development where change is seen as normal. The chief human resources officer (CHRO), CIO, COO, CTO, cloud director, and typically other cross-functional enterprise-wide leaders are examples of common stakeholders.

- *(NEW)* **ML Fluency** - Establishing a common language and conceptual framework.
- **Culture Evolution** - When using AI, culture is even more important.
- **Workforce Transformation** - Bringing in, facilitating, and overseeing AI talent—from builder to user.
- **Organization Alignment** - Enhancing and depending upon cross-organizational cooperation.

- **Governance perspective** - This viewpoint assists in planning AI projects while optimizing organizational advantages and lowering risks associated with change. Chief transformation officers, CIOs, CTOs, CFOs, chief data officers (CDOs), and chief risk officers (CROs) are examples of common stakeholders.

- *(NEW)* **Responsible use of AI** - Encourage ongoing AI innovation by using it responsibly.
- **Risk Management** - Use the cloud to reduce and control the dangers that come with artificial intelligence.
- **Data Curation** - Utilize product and catalog data to create value.
- **Cloud Financial Management** - Analyze, quantify, and optimize cloud AI costs.

Platform perspective

This perspective assists in creating a scalable, enterprise-grade cloud platform that gives the capacity to create new, customized AI solutions in addition to operating AI-enabled or infused services and products. Common stakeholders include the CTO, IT leaders, ML operations engineers, and data scientists.

- *(NEW)* **AI/ML Lifecycle Management** - Oversee the machine learning workload lifetime.
- **Platform Architecture** - Best practices, trends, and principles for reproducible AI value.
- **Modern Application Development** - Create AI-first, well-architected applications.
- **Data Architecture** - Create an AI data architecture that is appropriate for its use.
- **Platform Engineering** - Create a more feature-rich AI environment.
- **Data Engineering** - Data flows should be automated for AI development.
- **Provisioning & Orchestration** - Develop, oversee, and market authorized AI products.
- **CI/CD** - “Continuous Integration and Continuous Delivery”; Accelerate the development of AI.

Security perspective

This perspective aids in ensuring the availability, confidentiality, and integrity of your cloud workloads and data. Internal audit leaders, security architects and engineers, chief information security officers (CISOs), and chief compliance officers (CCOs) are examples of common stakeholders.

- **Vulnerability Management** - Continue to find, categorize, fix, and lessen AI vulnerabilities.
- **Security Assurance** - Apply, assess, and verify privacy and security safeguards in relation to legal and regulatory requirements for AI workloads.
- **Security Governance** - Provide roles and duties pertaining to AI workloads as well as security rules, standards, and guidelines.
- **Threat Detection** - Identify and address any unusual behaviors in AI workloads or possible security risks associated with AI.

- **Infrastructure Protection** - Protect the services and systems that run AI workloads.
- **Data Protection** - Preserve control, visibility, and safe access to the data required to build and apply AI.
- **Application Security** - Find and mitigate vulnerabilities in AI workloads during the software development lifecycle.

Operations perspective

This perspective makes it easier to guarantee that your cloud services—and especially your AI workloads—are provided at a level that satisfies your company's requirements. Information technology service managers, ML operations engineers, infrastructure and operations leaders, and site reliability engineers are examples of common stakeholders.

- **Incident & Problem Management** - Recognize and control unexpected AI behavior.
- **Performance & Capacity Management** - Track and manage the performance of the AI task.

The remaining capabilities are described in the original AWS Cloud Adoption Framework.

C. AWS Infrastructure and Technologies for Generative AI

Chapter 2.8 AWS Services and Features for Developing Generative AI Applications

Amazon SageMaker JumpStart

Pretrained models for tasks like image generation and article summarizing are available through Amazon SageMaker JumpStart, an ML hub that speeds up model selection, customization, and deployment. With options for sharing artifacts and controlling model visibility, it allows for the safe and private management of data within the user's VPC (virtual private cloud).

Amazon Bedrock

A completely managed service which offers strong foundation models and tools for creating scalable, secure generative AI applications.

Amazon Bedrock can be used for the following:

- Experiment with prompts and configurations
- Augment response generation with information from your data sources

- Create applications that reason through how to help a customer
- Adapt models to specific tasks and domains with training data
- Improve your FM-based application's efficiency and output
- Determine the best model for your use case
- Prevent inappropriate or unwanted content
- Lower barrier to entry

Amazon Bedrock Knowledge Bases:

Amazon Bedrock Knowledge Bases is a feature within Amazon Bedrock that enables you to integrate proprietary information into your generative AI applications. Here's how you can use Amazon Bedrock Knowledge Bases:

- Retrieve pertinent information to answer queries.
- Augment responses by integrating information from your own databases.
- Customize prompts to enhance the quality of responses.
- Convert natural language queries into structured queries (like SQL) for retrieving data.
- Update the knowledge base with new information directly.
- Rerank results to improve the accuracy of retrieved information.
- Integrate knowledge bases with Amazon Bedrock Agents workflows.

PartyRock

This is an Amazon Bedrock Playground. It is a fun, interactive environment for creating generative AI applications. Making apps to make playlists, recipes, etc. which only takes a few simple steps. PartyRock leverages foundation models from Amazon Bedrock to transform ideas into functional PartyRock apps. It can build and use these apps until the Backstage page indicates that the PartyRock credit has been fully utilized.

Amazon Q

By simplifying processes like coding, content creation, debugging, and retrieving business insights from corporate data, Amazon Q enables developers and business users. Enterprise data is linked, and customized summaries are generated, which can help users have tailored conversations. This service was created with security and privacy in mind, boosts productivity, and stimulates innovation.

AWS App Studio

Using natural language, users can create enterprise applications in minutes with AWS App Studio, a generative AI service. It makes it possible for experts without sophisticated coding knowledge, such as data engineers and IT managers, to develop business applications.

AI Infrastructure

For all training and inference requirements, AWS provides safe, affordable AI infrastructure with a wide range of strong tools. With AWS, users may use fully managed services to conduct distributed training on specially designed chips or GPUs.

Chapter 2.9 Cost Tradeoffs of AWS Generative AI Services

There are several key factors that influence costs when using Generative AI services. These tradeoffs depend on factors such as responsiveness, availability, redundancy, performance, and how you structure your use of AWS services. Below are some important considerations:

Responsiveness

Fast response times require powerful compute resources (e.g., GPUs), which can drive up costs, especially for low-latency applications.

Availability and Redundancy

Multi-region deployments and failover configurations provide high availability, but they also increase costs due to the need for cross-region data transfers and additional infrastructure.

Performance

High-performance models, such as those for large-scale text generation or image creation, demand specialized hardware and higher resource consumption, leading to increased costs.

Regional Coverage

AWS service pricing varies by region, and costs can increase with multi-region deployments or when choosing regions with higher service fees or data transfer costs.

Token-Based Pricing

Token-based pricing models (such as those used by Amazon Bedrock) tie costs to the volume of tokens processed. While this model is predictable, it can become expensive with large volumes of data or long-duration tasks.

Provisioned Throughput

Throughput refers to the volume and speed at which a model handles and produces inputs and outputs. You can opt for Provisioned Throughput to allocate a higher level of throughput for a model at a predetermined cost.

Custom Models

Developing and deploying custom models, like fine-tuning large language models, requires significant compute resources, especially if using GPUs or specialized hardware, which can increase overall expenses.

Domain 2: Fundamentals of Generative AI Sample Questions

1. A tech company is integrating various generative AI models to enhance its products and services. Select the correct type of generative AI model to meet each requirement. Each model can be selected one or more times. (Select FOUR.)

   The company needs a model that can generate detailed and contextually accurate technical documentation based on sparse input data __(a)__

   The design team wants to create photorealistic images from abstract concepts described in the text, ensuring high fidelity and detail __(b)__

   The company requires a model that can simultaneously analyze and generate content involving text, images, and audio for an immersive virtual assistant __(c)__

   The company needs a robust and adaptable model that can be fine-tuned for a wide range of tasks, including natural language understanding, image recognition, and predictive analytics __(d)__

   a. **Large Language Model**
   b. **Stable Diffusion Model**
   c. **Multimodal Model**
   d. **Foundation Model**

Explanation:

- Large language models (LLMs) use deep learning and transformer architecture. These models study and comprehend text. The model's understanding is achieved by recognizing connections between words and phrases.
  - The company needs a model that can generate detailed and contextually accurate technical documentation based on sparse input data: Large language model.

- Stable Diffusion is a generative AI model that creates distinctive photorealistic images based on text and image prompts.
  - The design team wants to create photorealistic images from abstract concepts described in the text, ensuring high fidelity and detail: Stable Diffusion model.

- Multimodal models are AI systems that can process and generate content across multiple modalities, such as text, images, and audio. These models are designed to understand and integrate information from diverse data types, enabling more comprehensive and contextually rich outputs.

- The company requires a model that can simultaneously analyze and generate content involving text, images, and audio for an immersive virtual assistant: Multimodal model.

- Foundation model describes machine learning models that are trained on a diverse range of generalized and unlabeled data. These models are capable of performing a wide array of tasks, including language comprehension, text and image generation, and natural language conversation.
  - The company needs a robust and adaptable model that can be fine-tuned for a wide range of tasks, including natural language understanding, image recognition, and predictive analytics: Foundation model.

2. A small e-commerce company wants to use machine learning to improve its churn prediction. However, the company does not have a dedicated data science team and is looking for a low-code or no-code solution to get started with machine learning. Match the company's requirements to the most suitable.

   Amazon SageMaker AI feature. (Select THREE.)

   Start quickly with pre-built solutions and models to accelerate development __(a)__

   Prepare and transform data for machine learning models using an intuitive interface __(b)__

   Build ML models with no code by simply interacting with data and obtaining predictions __(c)__

   a. **Amazon SageMaker Jumpstart**
   b. **Amazon SageMaker Data Wrangler**
   c. **Amazon SageMaker Canvas**

Explanation:

- Amazon SageMaker JumpStart provides an easy way to start with pre-built models and solutions, allowing the company to accelerate its machine-learning efforts without needing extensive expertise. JumpStart offers access to a wide variety of pre-trained models and end-to-end solutions that can be quickly deployed and customized, making it ideal for a company with limited data science resources.
  - Start quickly with pre-built solutions and models to accelerate development: Amazon SageMaker JumpStart
- Amazon SageMaker Data Wrangler simplifies the process of preparing and transforming data for machine learning. It offers a visual interface for data wrangling, enabling users to clean, transform, and

visualize data without writing complex code. This makes it a perfect match for the company's requirement of a low-code solution.

- Prepare and transform data for machine learning models using an intuitive interface: Amazon SageMaker Data Wrangler

- Amazon SageMaker Canvas is a no-code tool that enables users to build machine-learning models by simply interacting with data. It allows business analysts and non-technical users to generate accurate predictions without needing to write a single line of code, addressing the company's need for a no-code solution to improve churn prediction.
  - Build ML models with no code by simply interacting with data and obtaining predictions: Amazon SageMaker Canvas

References Domain 2

Tokens
https://iotmktg.com/tokens-101-understanding-tokens-in-generative-ai-models/

Chunking
https://towardsdatascience.com/the-art-of-chunking-boosting-ai-performance-in-rag-architectures-acdbdb8bdc2b

Methods of chunking
https://www.pinecone.io/learn/chunking-strategies/

Embeddings
https://aws.amazon.com/what-is/embeddings-in-machine-learning/

Vectors
https://aws.amazon.com/what-is/vector-databases/

LLMs
https://community.aws/content/2eD4ehf1d5hatTW1dnVErcJVVPL/a-gentle-intro-to-transformer-and-gen-ai?lang=en
https://www.truefoundry.com/blog/transformer-architecture#practical-applications-of-transformers

Foundation Models
https://aws.amazon.com/what-is/foundation-models/
https://medium.com/intel-tech/ai-dance-party-foundation-vs-fine-tuned-models-d269df518b92

Multi-Modal Models
https://aws.amazon.com/blogs/machine-learning/generative-ai-and-multi-modal-agents-in-aws-the-key-to-unlocking-new-value-in-financial-markets/
https://www.kdnuggets.com/2023/03/multimodal-models-explained.html

Diffusion Models
https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/
https://aws.amazon.com/what-is/generative-ai/#:~:text=Generative%20AI%20can%20boost%20productivity,human%20feedback%20and%20specified%20constraints
https://aws.amazon.com/what-is/generative-ai/#:~:text=Generative%20AI%20can%20boost%20productivity,human%20feedback%20and%20specified%20constraints

Use cases of Gen AI Models
https://www.cxtoday.com/contact-centre/20-use-cases-for-generative-ai-in-customer-service/
https://www.zdnet.com/article/best-ai-search-engine/

Challenges of Gen AI
https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-toxicity-evaluation.html
https://cloud.google.com/discover/what-are-ai-hallucinations
https://www.launchconsulting.com/posts/attribution-error-aws-head-of-solutions-architecture-talks-generative-ai-and-copyright
https://libguides.astate.edu/plagiarism/ai
https://www.speedofcreativity.org/2024/05/29/ai-and-cheating/
https://tdwi.org/articles/2023/07/24/adv-all-understanding-the-disruptive-nature-of-generative-ai.aspx

Factors for selecting appropriate Gen AI Models
https://fluidai.medium.com/how-do-you-measure-gen-ai-deployment-pilot-success-key-performance-indicators-and-metrics-bed1a963f812
https://www.xenonstack.com/blog/generative-ai-compliance

The AWS CAF-AI foundational capabilities
https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/business-perspective-the-ai-strategy-in-the-age-of-aiml.html
https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/people-perspective-culture-and-change-towards-aiml-first.html
https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/governance-perspective-managing-an-aiml-driven-organization.html
https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/platform-perspective-infrastructure-for-and-applications-of-aiml.html
https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/security-perspective-compliance-and-assurance-of-aiml-systems.html
https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/operations-perspective-health-and-availability-of-the-aiml-landscape.html

Services and Features for Developing Gen AI Apps
https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/operations-perspective-health-and-availability-of-the-aiml-landscape.html
https://aws.amazon.com/about-aws/whats-new/2023/11/partyrock-amazon-bedrock-playground/

https://partyrock.aws/guide/getStarted
https://aws.amazon.com/q/?sec=aiapps&pos=1
https://aws.amazon.com/appstudio/
https://aws.amazon.com/ai/infrastructure/?sec=aiapps&pos=4

Others
https://aws.amazon.com/what-is/large-language-model/
https://aws.amazon.com/what-is/stable-diffusion/
https://aws.amazon.com/blogs/machine-learning/generative-ai-and-multi-modal-agents-in-aws-the-key-to-unlocking-new-value-in-financial-markets/
https://tutorialsdojo.com/aws-cheat-sheets-aws-machine-learning-and-ai/
https://docs.aws.amazon.com/sagemaker/latest/dg/studio-jumpstart.html
https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html
https://docs.aws.amazon.com/sagemaker/latest/dg/canvas.html
https://tutorialsdojo.com/amazon-sagemaker/

APPLICATIONS OF FOUNDATION MODELS

Design Considerations for Applications Using Foundation Models
Effective Prompt Engineering Techniques
Training and Fine-Tuning Foundation Models
Evaluating Foundation Model Performance

A. Describe Design Considerations for Applications with Foundation Models

Chapter 3.1 Identify Selection Criteria to Choose Pre-Trained Model

For starters, choosing the right pre-trained foundational model is important for successful implementation with higher chances of securing Return of Income (ROI) from a business perspective. Hence, following the selection criteria are important for an application's success. For simplicity, you may want to consider these factors:

Selection Criteria	Key Considerations	Recommended AWS Services	Additional Details
Cost	• Subscription pricing • Compute resources • Operational expenses	AWS Cost Explorer	Track and manage expenses related to model deployment
Modality	• Data type support (text, PDF, images, audio) • Processing capabilities	• AWS Rekognition (images) • Amazon Transcribe (audio)	Select services based on specific data type requirements
Latency	• Input submission time • Inference time • Response speed	• AWS Inferentia • AWS Global Accelerator	Optimize processing speed and network performance
Multilingual Support	• Language understanding • Multi-language generation	• Amazon Translate • Amazon Comprehend	Critical for international market targeting
Model Size	• Number of model parameters • Computational costs	• Amazon EC2 • AWS Lambda	Higher parameter count increases computational requirements
Model Complexity	• Architectural complexity • Accuracy vs. cost trade-offs	Amazon SageMaker	Balance performance needs with budget constraints
Customization	• Hyperparameter tuning • Feature Engineering • Prompt Engineering	Amazon SageMaker JumpStart	Tailor model to specific application needs
Input / Output Length	• Input data size • Generated output size • Computing Usage	• Amazon S3 • Amazon API Gateway	Manage data storage and flow efficiently

Cost

Purchasing for a subscription plan for a foundational model will commonly be taken from investment funds. Hence, identifying pricing factors such as model subscription, licensing, compute resources, and operational expenses can be a factor under cost tracking.

- **In line with AWS** - Utilize cost management tools such as AWS Cost Explorer to estimate and monitor expenses and Amazon Bedrock for scalable pricing models.
    - **AWS Cost Explorer** - A budgeting tool that enables users to manage AWS cost and usage over time.
    - **Amazon Bedrock** - A fully managed service that hosts the foundational models coming from reputable AI firms such as Anthropic, Meta, and A2I. It is used for building generative AI applications easily and scalable.
- **Cost-related Scenario:** A Philippine startup creating a language translation app for native dialects like Tagalog, Cebuano and Ilocano must handle cloud expenses wisely to keep running.
    - The startup gets money from a local investment company to buy a basic AI model through Amazon Bedrock. They use AWS Cost Explorer to watch spending on model subscriptions, license charges, computing resources and everyday costs. Recognizing these price elements, they can stay within their budget while growing their app. Staying on budget is really important.

Modality

Consider the type of data the model will process, it can be text files, PDFs, images, and audio.

- **In line with AWS:** Selection of the right services suits the model's purpose. Options might involve Amazon Rekognition for sorting images. Audio handling might require Amazon Transcribe.
    - **Amazon Rekognition** - A service that changes image and video analysis to applications, making it recognize objects, identifying faces, and moderating content.
    - **Amazon Transcribe** - A tool that listens to speech and changes it to text. This will help with transcription and voice analysis.
- **Modality-related Scenario:** A Manila-based e-commerce platform wants to enhance its user experience by incorporating image recognition for product listings and voice search functionality.
    - The site uses Amazon Rekognition to automatically sort and label product pictures. This helps customers find and filter products more easily. Additionally, they use Amazon Transcribe to change spoken searches into text, letting users explore the site using their voices. This method caters to the varied tastes of Filipino shoppers.

Latency

Consider the time delay a process can handle during input submission, inference time, and model response.

- **In line with AWS:** You may want to consider deploying your AI models using AWS Inferentia for low-latency inference or leverage AWS Global Accelerator to reduce latency across regions.
    - **AWS Inferentia** - special EC2-based chip created by AWS. It gives strong and affordable machine learning abilities to use models on a big scale.
    - **AWS Global Accelerator** -A networking service increases the availability and performance of applications. It directs user traffic through the AWS global network infrastructure. The AWS global network infrastructure helps maintain application stability.

- **Latency-related Scenario:** A real-time online gaming company in the Philippines wants to give users a smooth gaming experience, minimizing delays and providing quick responses.
    - The company uses AI-driven matchmaking models with AWS Inferentia for low-latency inference, swiftly matching players. AWS Global Accelerator helps reduce delays more across regions in the Philippines. It directs user traffic through the nearest AWS edge locations for quicker response times and a smoother game. The game runs smoother.

Multilingual Support

If your target market covers international customers, you may want to consider models that can understand and generate multiple languages.

- **In line with AWS:** Utilize Amazon Translate for pre-processing multilingual data or Amazon Comprehend for language detection and analysis.
    - **Amazon Translate** - A neural machine translation service provides quick and high-quality translation of languages. This allows applications to support many languages.

- **Multilingual support-related Scenario:** A customer service company in the Philippines helps people who speak different languages, such as Tagalog, English and local ones like Cebuano and Ilocano.

    - The company uses Amazon Translate to first handle questions from customers who speak different languages. Their AI chatbots understand and reply correctly in several languages. Also, Amazon Comprehend detects which language is being used and checks how customers feel. This lets the company give specific and suitable replies to their varied clients. The responses match each customer's needs really well.

Model Size

You may also consider the number of computational costs, in a form of parameters in the model, which will affect capacity and resource requirements. Take note that having a higher parameter count will incur more computational cost.

- **In line with AWS:** Pick the right instance types on Amazon EC2 for your needs. Use AWS Lambda for serverless deployments as an alternative.
    - **Amazon EC2 (Elastic Compute Cloud)** - A web service offers adjustable computing power online. Users run virtual servers for many different applications. It has computing power and changes size as needed.
    - **AWS Lambda** - A serverless computing that helps run code when events occur. This tool is great for lightweight inference models deployment.
- **Model size-related Scenario:** A financial technology (fintech) company in the Philippines is building a fraud detection system. This system needs to work efficiently but not require expensive computing.

    - The fintech company chooses the right Amazon EC2 instances to balance performance and cost, considering the number of model parameters. For easier tasks, they use AWS Lambda to run functions without servers. This reduces the need for managing many servers. It also optimizes their computing costs. This is really important.

Model Complexity

This refers to the architectural complexity of each mode, some of which may offer lower architecture complexity which offers lower cost but may impact performance. On the other hand, some architectures offer a highly accurate result but incur a higher cost per inference.

- **In line with AWS:** AWS offers services to manage complex models such as Amazon SageMaker, to handle features like automated scaling and distributed training.

    - **Amazon SageMaker AI** - A complete service for creating, teaching and using machine learning models on a large scale. It offers tools for each part of the machine learning process.

- **Model complexity-related Scenario:** A healthcare startup in the Philippines is creating a tool that looks at medical images and patient details to help doctors decide.

    - The startup uses Amazon SageMaker AI to handle the complex AI models needed for precise diagnostics. This service takes care of automated scaling and distributed training. It helps the company to quickly build, train and use their machine learning models. High performance and reliability are really important. They achieve these consistently.

Customization

The extent to which a model can be tailored to specific application needs. It can be done through hyperparameter tuning, feature engineering, or prompt engineering.

- **In line with AWS:** You can leverage Amazon SageMaker JumpStart for pre-built models and performing fine-tuning techniques.
    - **Amazon SageMaker JumpStart** - helps users with ready-made solutions. These include machine learning models and example python notebooks. People use them to speed up creating and expanding machine learning projects.
- **Customization-related Scenario:** A Philippine agriculture technology company desires to build a model that forecasts crop yields by considering different environmental factors specific to regions in the country.
    - The company utilizes Amazon SageMaker JumpStart to reach pre-built machine learning models and adjusts them with data specific to each region. By tuning hyperparameters and engineering features, they tailor the models to precisely predict crop yields. This guarantees the tool fits the unique agricultural needs of various Filipino provinces.

Input/Output Length

The length of input data and the size of the generated output must be anticipated to avoid unnecessary computing usage.

- **In line with AWS:** Improve data management by storing it in Amazon S3 and managing data flows with Amazon API Gateway.
    - **Amazon S3 (Simple Storage Service)** - An object storage service with top-level scalability, data availability, and security. Its performance is excellent for storing and getting any data amount.
    - **Amazon API Gateway** - Is a fully managed service. Users easily create, publish and maintain APIs at different scales. Users also monitor and protect APIs effectively, and work well with this service.
- **Customization-related Scenario:** A Manila-based legal tech firm develops an AI assistant. This assistant helps lawyers write and check legal papers quickly.
    - The firm saves all input and output data in Amazon S3. This method keeps data storage scalable and secure. Amazon API Gateway handles data movement. This setup allows the AI assistant to process documents of different lengths smoothly. It reduces unnecessary computational work, and resource usage becomes more efficient.

Chapter 3.2 Understand the Effect of Inference Parameters on Model Responses

Understanding inference parameters is critical for improving how a model works. This knowledge helps achieve higher accuracy with a balanced number of settings. It avoids unnecessary consumption of computing power, leading to a cost-effective solution.

Temperature

A parameter that controls how random the model's output is. Greater temperatures lead to varied responses, while smaller temperatures give more predictable results.

- **In line with AWS:** You can specify the temperature settings in Amazon SageMaker endpoints to balance creativity and accuracy, depending on your needs.
- **Temperature-related Scenario:** A Filipino content creation agency uses AI to generate creative marketing copy for diverse clients.
    - The agency changes the temperature setting in their Amazon SageMaker endpoints to control how creative the generated content is. They choose a higher temperature for campaigns needing very creative and different slogans. They prefer a lower temperature for formal and predictable content, like official announcements. This helps with consistency and accuracy.

Input/Output Length

Measure the length of the model and how long the generated text.

- **In line with AWS**: Set token limits in Amazon Bedrock carefully. This manages resource usage and keeps answers within the app's needs.
- **Input/Output length-related Scenario:** A Filipino educational platform designs an AI tutor to help students write essays and give feedback.
    - The platform fixes proper token limits in Amazon Bedrock to control the length of student inputs and AI feedback. By setting these limits thoughtfully, the AI offers complete yet brief feedback. This manages resource use and keeps the feedback useful for students.

Chapter 3.3 Define Retrieval Augmented Generation (RAG) and Describe Its Business

Retrieval Augmented Generation (RAG)

An advanced AI approach that mixes information retrieval techniques with generative models with the aim of providing highly accurate responses and contextually appropriate responses. This fixes the limitation that

relies solely on pre-trained data, giving outdated or hallucinated information. RAG enhances these models by integrating real time retrieval from extensive databases or knowledge bases.

How RAG systems utilize a two-step process:

- **Retrieval** - When a query is received, the system first searches relevant documents, databases, or knowledge repositories to fetch pertinent information
- **Generation** - The retrieved data is then fed into a generative model such as a large language model to produce a coherent and informed response.

Benefits:

- **Enhanced Accuracy** - By accessing up-to-date and specific information, RAG reduces the likelihood of errors and hallucinations, where it produces incorrect information.
- **Contextual Relevance** - By having a context reference based on the most data accumulated from such resources, it ensures the results are highly relevant.
- **Scalability** - RAG can handle vast amounts of data, making it reliable for diverse applications from various industries.

RAG-related Scenario: A Philippine government agency wants to share correct and updated information with people through a virtual assistant.

- The agency uses a RAG system. First, the virtual assistant gets important information from the newest government databases and documents, for example PSA OpenStat. Next, it uses a generative AI model to create clear and knowledgeable answers. Citizens receive correct and relevant information. This reduces misinformation, making public service better.

Business Applications

RAG finds many uses in various business fields. Here are two main areas where RAG greatly improves operations and customer experiences.

- **Knowledge Bases**: Improving Customer Support with Current Information.
    - **Customer Support Automation** - Intelligent chatbots and virtual assistants deliver correct and timely answers to customer questions. They use up-to-date support documents, FAQs and manuals for this task.

- **Agent Help** - Real-time knowledge assistants bring useful information for human support agents. This happens during customer interactions, where these results in better service.
- **Example**: A Filipino telecom company uses a RAG-powered support system. It taps into a large database of troubleshooting guides and service manuals. It helps customers fix technical problems quickly.

- **Content Generation**: Produce personalized content by retrieving relevant data.
  - **Personalized Marketing** - RAG creates custom marketing materials by using data about individual customer likes, actions and past interactions.
  - **Content Summarization** - Media and publishing firms use RAG to make brief summaries of long documents or articles. It picks key details and creates summaries for specific audiences.
  - **Example Use Case** - An e-commerce platform from the Philippines uses RAG to write personalized product descriptions and suggestions. It studies customer browsing history and finds relevant product information.

AWS Implementation of RAG

Amazon Web Services (AWS) provides powerful tools and services where these tools help develop and launch RAG applications. It guarantees scalability, security, and integration with other enterprise-grade services.

Amazon Bedrock

Amazon Bedrock is a fully managed service that provides access to a variety of foundation models from leading AI providers. Developers create and expand RAG applications without needing to manage underlying infrastructure.

- **Features:**
  - **Pre-trained Models** - Users access top-notch generative models. These models adapt for specific business needs.
  - **Integration Capabilities -** Amazon Bedrock works smoothly with other AWS services. It connects with Amazon Kendra for finding information. It also links with Amazon S3 for storing data.
  - **Customization** Users adjust models to fit specific applications. This makes generated content more relevant and precise.
- **Benefits:**
  - **Speed to Market** - Quickens development with models and tools ready for use. Speeds up progress with ease.

- **Scalability** - Automatically scales to handle various workloads, ensuring consistent performance.
- **Security and Compliance** - Uses AWS's strong security systems. Protects data and meets industry rules.

- **Bedrock-related Scenario:** A Philippine media company wants to automate content summarization for its vast repository of articles and reports.
  - The media company uses Amazon Bedrock to find pre-trained generative models. It also combines Amazon Kendra to retrieve information quickly from many articles in their database stored in Amazon S3. This setup lets them create short summaries for different groups of readers. Content delivery becomes very effective and readers feel more connected.

Leveraging AWS Resources for RAG

To truly unlock the potential of RAG applications, AWS offers a group of services that support each part of RAG development together:

- **Amazon Kendra** - This intelligent search tool uses machine learning. It retrieves information from different data sources well. Kendra suits the search part of RAG very well.
- **Amazon S3** - This secure and scalable storage holds large datasets and knowledge bases. RAG systems access this data.
- **AWS Lambda** - This tool enables serverless computing. It runs backend tasks like data retrieval and processing. AWS Lambda operates without the need to manage servers.
- **Amazon SageMaker AI** - This service supplies tools to build, train and deploy machine learning models. It supports the creative side of RAG.

Sample Scenario: A Philippine educational institution develops a smart tutoring system that provides correct and relevant answers to student questions.

- **Amazon Kendra** searches and finds useful educational materials from their big digital library.
- **Amazon S3** stores large amounts of data and knowledge bases for the RAG system to use.
- **AWS Lambda** manages backend tasks like getting and handling data without needing servers.
- **Amazon SageMaker AI** builds, trains and launches the machine learning models that give clear and informed answers based on the collected data.

Chapter 3.4 Identify AWS Services That Help Store Embeddings Within Vector Databases

In the world of artificial intelligence, embeddings hold a key role. They help show complex data types like text, images and audio in high-dimensional vector spaces. Efficient storage and management of these embeddings are important as these tasks are necessary for many applications like similarity search, recommendation systems and natural language processing. AWS offers several services tailored for storing and searching embeddings in vector databases. This chapter dives into these services, where it discusses their capabilities, use cases and integration points.

Imagine a Filipino online ai-based marketplace specializing in native crafts—such as handwoven bags from Abra or wood carvings from Paete, Laguna. To better recommend items to customers, the marketplace wants to store and search AI embeddings (representations of product descriptions, images, and user preferences). These embeddings capture the “essence” of every item and user, helping power advanced similarity searches and personalized recommendations.

Hence, embeddings are essential in AI because they represent complex data—like Tagalog text, images of Philippine festivals, or audio recordings of local dialects—in high-dimensional vector spaces. Efficiently storing and managing these embeddings is crucial for tasks like similarity searches, recommendation systems (e.g., finding similar souvenirs for tourists), and natural language processing in Filipino languages. AWS offers multiple services ideal for these use cases.

Amazon OpenSearch Service

Amazon OpenSearch Service handles the entire process of deploying, running and expanding OpenSearch clusters in the AWS Cloud. OpenSearch, which evolved from Elasticsearch, provides a search and analytics suite.

- **Key Features for Storing Embeddings**
  - **Vector Search for Similarity Matching** - OpenSearch includes vector search. This function supports similarity matching using k-Nearest Neighbors (k-NN) algorithms. Users retrieve similar embeddings efficiently, based on cosine similarity, Euclidean distance or other distance measures.
  - **Scalability and Performance** - OpenSearch manages large data volumes effortlessly. The system increases its capacity by adding more nodes to the cluster. It guarantees high availability and quick responses for search queries.
  - **Integration with Machine Learning Pipelines** - OpenSearch easily connects with AWS machine learning services. Embeddings from models like Amazon SageMaker AI are efficiently indexed and searched.
- **Use Cases**

- **Recommendation Systems** - Giving personalized product or content suggestions. It does this by finding items with similar embeddings to what users like.
- **Implementation Considerations** - Letting users look for images or videos based on how they look alike.
- **Natural Language Processing** - Improving search functions in applications. This helps by understanding the meaning similarities between questions and documents.

- **Sample scenario:**
  - Let's say PinoyMart, a Filipino online marketplace specializing in local delicacies and festive items, leverages Amazon OpenSearch Service's vector search, autoscaling, and ML integrations to deliver culturally relevant e-commerce experiences.
  - By indexing embeddings from Amazon SageMaker AI, PinoyMart instantly retrieves similar products—like parols or bibingka—based on user behavior or uploaded images, ensuring personalized recommendations.
  - Autoscaling handles holiday traffic spikes, while NLP for Taglish (Tagalog-English) queries refines search accuracy. This unified approach empowers PinoyMart to provide a seamless shopping journey steeped in Filipino culture, complete with real-time analytics and optimized recommendations.

Amazon Aurora

Amazon Aurora is a fast, completely managed database engine for relations**.** It works with PostgreSQL. Extensions like pgvector turn Aurora PostgreSQL into a powerful tool.

- **Key Features for Storing Embeddings**
  - **Pgvector Extension** - This PostgreSQL extension adds native support for vector data types. It allows efficient storage, indexing and similarity search of embeddings.
  - **Advanced Indexing** - Supports different indexing methods. Methods include approximate nearest neighbor (ANN) indexes. These indexes speed up similarity queries on large datasets.
  - **High Availability and Durability** - Aurora offers automated backups and replication. It replicates across multiple Availability Zones (AZs). There are failover capabilities too. This ensures data durability. Data durability and availability are really important.
- **Use Cases**
  - **Semantic Search** - Program search functions that understand the meaning behind questions. They compare embeddings to get the right context.
  - **Fraud Detection** - Examine transactional embeddings. Look for signs that might show fraudulent actions. These patterns tell a lot. They really help in spotting fraud.

- **Personalized Marketing** - Adjust marketing plans. Use customer behavior embeddings. Stores and accesses must be very efficient for best results.

- **Implementation Considerations**
  - **Extension Installation** - Enable the pgvector extension in your Aurora PostgreSQL instance to leverage vector data types and functions.
  - **Query Optimization** - Optimize SQL queries for similarity search by utilizing vector-specific functions and appropriate indexing strategies.
  - **Resource Allocation** - Aurora systems need the right size. They should manage the heavy work of processing and checking high-dimensional vectors.
- **Sample scenario:**
  - Let's say PinoyMart, a Filipino online store focused on local treats, uses Amazon Aurora PostgreSQL with pgvector for fast search results, smooth fraud checks and custom marketing.
  - It stores product details and shopper behavior to quickly link buyers to fitting items while managing busy holiday times.
  - Automated backups and multi-region replication keep information safe and lasting, helping steady business work.
  - Pgvector's indexing improves Taglish-friendly searches and transaction checks, offering a culturally tailored shopping experience for PinoyMart's users.

Amazon Neptune

Amazon Neptune is an AWS fully managed graph database service. It is ideal for storing and querying connected data. This service is mainly for graph-based applications. Neptune efficiently stores and query embedding. Relationships between data points are important. Graph database optimized for storing and querying relationships, including embeddings.

- **Key Features for Storing Embeddings**
  - **Graph Models** - Supports Property Graph and RDF graph models. These allow for flexible data representation and embeddings.
  - **Efficient Querying** - Uses graph traversal languages like Gremlin and SPARQL. These can extend to perform similarity searches on embeddings.

  - **Scalability and Performance** - Handles large-scale graph data with low delay. Suitable for real-time applications. Hence, it became very efficient for big tasks.

- **Use Cases**
  - **Knowledge Graphs** - Build knowledge graphs. Use embeddings to improve understanding of meanings and map relationships.
  - **Social Networks** - Analyze user embeddings and links. Identify communities and find influence patterns.
  - **Recommendation Engines** - It uses both embeddings and graph links. Offer more precise and relevant recommendations.
- **Implementation Considerations:**
  - **Data Modeling** - Design graph schemas, where you include embeddings as properties or nodes, based on what you need.
  - **Custom Functions** - Create custom graph traversal functions. Perform similarity searches using vector distances.
  - **Performance Tuning** - Optimize Neptune clusters for specific tasks. Think about query complexity and data volume.

- **Sample scenario:**
  - PinoyMart, a Filipino marketplace specializing in local delicacies and festive items, adopts Amazon Neptune to manage embeddings as part of its product recommendation system and knowledge graph.
  - By storing product relationships and user data as graph nodes, Neptune's support for Property Graph and SPARQL allows for quick, context-aware queries.
  - This real-time approach optimizes holiday traffic spikes, personalized recommendations (e.g., favorite kakanin), and supports advanced influencer or community analysis for a more immersive Pinoy shopping experience.

Amazon DocumentDB (with MongoDB Compatibility)

Amazon DocumentDB is a fully managed document database service that is compatible with MongoDB. It allows for flexible schema designs, making it suitable for storing embeddings as part of JSON documents.

- **Key Features for Storing Embeddings:**
  - **Flexible Schemas -** Store embeddings with other data in JSON documents. These documents can have various data structures. There is no need for fixed schemas.

  - **Indexing Features:** Create indexes on certain fields within documents. This includes documents with vector data.

  - **Scalability and Performance** Storage and compute resources grow automatically. The database handles more data and query loads as they increase.

- **Implementation Considerations**
  - **Embedding Storage** - Place embed vectors inside JSON documents. Alternatively, refer to them as external links if the application or query patterns require it.
  - **Index Design** - Set up the right indexes on embedding fields. This step improves how similarity search queries work.
  - **Data Consistency** - Consistently update embeddings across all related documents. This task is crucial for keeping data accurate.
- **Sample Scenario:**
  - PinoyMart is a Filipino marketplace for local treats and festive goods. This time, it uses Amazon DocumentDB, which works with MongoDB.
  - This tool keeps user and product information in flexible JSON documents. Fast and relevant suggestions are possible with these documents.
  - Embedding vectors directly into product details like kakanin and parols helps PinoyMart grow during busy holiday times.
  - User behaviors may change and yet documents remain consistent. Special indexes on embedded fields quicken similarity searches. Taglish queries and changing marketing insights give a real and powerful Filipino shopping experience.

Amazon RDS for PostgreSQL

Amazon RDS for PostgreSQL is a fully managed service. It helps users easily set up PostgreSQL databases in the cloud. Users find it easy to operate and expand this service. It supports vector data types and indexing. This support allows efficient storage of embeddings, enabling the retrieval of these embeddings to be efficient.

- **Key Features for Storing Embeddings**
  - **Vector Data Types** - PostgreSQL stores vector data using extensions like pgvector. It allows storage of high-dimensional embeddings.
  - **Advanced Indexing** - Database indexing methods such as Generalized Inverted Index (GIN) or k-d trees speed up similarity searches on vector data.
  - **Extensibility** - PostgreSQL's architecture is flexible. This flexibility allows integration of extra features for specific embedding needs.
- **Use Cases**
  - **Search Engines** - Use semantic search abilities by comparing query embeddings with document embeddings in the database.
  - **Machine Learning Pipelines** - Handle and access embeddings created during different machine learning steps.
  - **Data Analytics** - Conduct advanced analytics on embeddings. Discover hidden patterns and insights in the data. Insights that might be hidden become visible.

- **Implementation Considerations**
  - **Extension Setup** - Install and configure the pgvector extension. This step allows vector operations in your PostgreSQL database.
  - **Index Optimization** - Choose the right indexing method, in consideration to the type of your embeddings and query patterns. This ensures performance is very good.
  - **Resource Management** - Watch and adjust database resources like CPU, memory and storage. These resources must handle the computational needs of embedding tasks. It is important for large datasets.
- **Sample Scenario:**
  - Let's say PinoyMart, a Filipino marketplace for local delicacies and festive items, leverages Amazon RDS for PostgreSQL with pgvector to store and retrieve user and product embeddings for more accurate recommendations and semantic search.
  - By indexing vector fields (e.g., kakanin or parol embeddings) using GIN or k-d trees, PinoyMart rapidly matches shoppers with the items they're most likely to want, even during holiday traffic spikes.
  - PostgreSQL's extensibility further streamlines PinoyMart's ML pipelines, ensuring embeddings generated during different analysis steps seamlessly integrate with the database.
  - The result is a resilient, highly performant shopping platform where PinoyMart can scale resources on-demand while providing Taglish-friendly, data-driven user experiences.

Integrating AWS Services for Comprehensive Embedding Management

Each AWS service mentioned above provides unique tools for storing and querying embeddings. Combining multiple services offers a wider solution for specific business needs. Here are strategies to use these services together:

- **Hybrid Architectures** - Amazon OpenSearch Service allows fast vector searches. Amazon Aurora PostgreSQL helps in managing transactional data. This combination gives both quick querying and steady data storage.

- **Data Pipelines** - Generate embeddings from raw data using Amazon SageMaker. Store them in Amazon Neptune for queries based on relationships. Amazon DocumentDB is good for storing documents flexibly.

- **Serverless Integrations** - AWS Lambda handles embedding intake and updates over several databases. This method assures scalability. This also provides cost-effectiveness.

Best Practices for Storing Embeddings on AWS

1. **Select the Right Service for Your Needs**
   - **Search and Similarity** - Use Amazon OpenSearch Service or Amazon Aurora with pgvector.

- **Graph and Relationships** - Go for Amazon Neptune.
- **Flexible Document Storage** - Choose Amazon DocumentDB.
- **Traditional Relational Needs** - Pick Amazon RDS for PostgreSQL.

2. **Optimize Indexing Strategies**
   - Choose indexing methods that fit your query needs to improve performance.
   - Monitor indexes regularly, you can adjust them to keep search speeds fast.
3. **Ensure Data Security and Compliance**
   - Use AWS security features to protect embedding data. These include encryption at rest and in transit. Also, employ IAM roles and security groups, as these tools help keep data safe.
   - Follow data protection regulations. Apply proper governance measures to meet these standards.
4. **Scalability and Performance Monitoring**
   - AWS monitoring tools like Amazon CloudWatch help track database performance. Adjust resources as needed to keep things running smoothly.
   - Use caching mechanisms when suitable. This reduces latency and speeds up response times.
5. **Cost Management**
   - Choose the right instance types and storage based on workload needs. This choice very effectively manages costs.
   - AWS cost optimization tools help monitor and control spending. Keep an eye on expenses tied to embedding storage and querying.

Efficient storage and management of embeddings are vital for intelligent applications that use machine learning and AI. AWS offers a wide selection of services designed for various embedding storage needs. Each service has its distinct advantages and features. Organizations should choose and connect these services based on their business needs. This strategy helps create strong, scalable and fast vector databases. Such databases provide useful insights and improve user experiences. For more help and detailed strategies, check the AWS service documentation. Consider talking to AWS experts or working with AWS Professional Services.

Chapter 3.5 Explain the Cost Tradeoffs of Various Approaches to Foundation Model Customization

Customizing foundation models is key to shaping artificial intelligence (AI) solutions for specific business needs. Each method of customization has its own costs. These costs involve financial investment and resource allocation. This section explores the cost tradeoffs of four main customization methods. These methods are Pre-Training, Fine-Tuning, In-Context Learning and Retrieval Augmented Generation (RAG).

Pre-Training

Pre-Training starts with training a foundation model from the beginning using large datasets. Organizations get complete control over both the model's design and the chosen data. This control helps them develop models tailored to their unique needs.

- **Pros**
    - **Complete Control Over Model Architecture and Data**
        - **Customization** - Adjust every part of the model. This includes the model architecture, the goals for training and the data sources.
        - **Optimization** - Adapt the model to meet specific performance goals or work requirements.
        - **Data Privacy** - Use private or sensitive data securely. Avoid third-party models.
    - **Potential for Superior Performance**
        - **Domain-Specific Expertise** - Create models that perform really well in specialized fields. This happens by training with focused datasets.
        - **Innovative Architectures** - Experiment new ideas in novel architectures. These fresh designs might perform better than current ones.
- **Cons**
    - **Extremely Resource-Intensive**
        - **Computational Costs** - need a significant investment in powerful hardware like GPUs or TPUs.
        - **Time-Consuming** - Training big models takes a long time, sometimes weeks or months, depending on the model and resources.
    - **High Expertise Requirement**
        - **Skilled Personnel** - Needs a team of experienced machine learning experts and data scientists.

        - **Complex Management** - Requires handling complicated training processes, tuning hyperparameters and evaluating models.
    - **Data Acquisition and Preparation**
        - **Large Datasets Needed** - Demands a lot of high-quality data. Gathering and processing this data really is costly and takes time.
        - **Data Storage Costs** - Storing large datasets creates extra expenses.
- **Cost Factors**
    - **Capital Expenditure (CapEx)** - A large initial investment goes into hardware and infrastructure. This is a significant upfront cost.
    - **Operational Expenditure (OpEx)** - Regular costs involve energy use, maintenance and staff. These expenses occur continuously.
    - **Opportunity Cost** - Longer timelines often delay deployment. This delay affects time-to-market.
- **AWS Implementation** - AWS usually does not recommend training large foundation models from scratch. This is because it is very expensive and complicated. However, Amazon SageMaker AI provides tools and infrastructure for pre-training when necessary.
- **Sample scenario:**
    - BaybayinAI, a fictional Philippine-based startup, decides to pre-train a massive Taglish foundation model from scratch using an extensive corpus of social media posts, digital publications, and regional slang repositories.
    - Their aim is to build a domain-specific AI that understands nuances like *"Parang gusto ko ng milk tea, 'no?"("I think I like to drink milk tea.")* for sentiment analysis and chatbot solutions.
    - Controlling every layer of model architecture and data ingestion ensures robust privacy measures, especially in handling sensitive local expressions and user details.
    - However, the high cost of GPU clusters, extensive data cleanup, and the need for seasoned ML experts stretch BaybayinAI's resources—both financially and operationally.
    - Though AWS generally advises against large-scale pre-training from scratch due to complexity, BaybayinAI leverages Amazon SageMaker's managed environment for compute elasticity and secure storage, committing to the potentially game-changing benefits of a custom, hyper-local AI solution.

Fine-Tuning

Fine-tuning involves adapting a pre-trained foundation model to specific tasks or datasets. By modifying existing models, fine-tuning improves the performance for particular applications, eliminating the need to start from scratch.

- **Pros:**
    - **Customization For Specific Tasks**
        - **Enhanced Performance** - Fine-tuning increases model accuracy and relevance for targeted application.
        - **Tasks-Specific Adaptation** - It enables models to perform specialized tasks accurately, such as sentiment analysis or medical diagnosis.
    - **Reduced Data and Compute Requirements**
        - **Efficiency** - Fine-tuning requires lower computational resources with higher accuracy.
        - **Data Efficiency** - Significant model performance can be achieved with smaller, task-specific datasets.
    - **Faster Deployment**
        - **Reduced Training Time** - Fine-tuning lowers the training time, increases its chances for time-to-market adaptation.
        - **Resource Optimization** - Lower computational demands lead to cost savings in resources. This is important especially when the model is deployed in the cloud.
- **Cons**
    - **Requires Expertise in Model Training and Data Preparation**
        - **Specialized Skills** - Necessitates knowledge of machine learning techniques, model architectures, and data preprocessing.
        - **Complex Processes** - Demands careful handling of hyperparameters, regularization techniques to avoid overfitting and strategies for testing.
    - **Potential for Overfitting**
        - **Data Limitations** - Fine-tuning with small datasets could result in models that perform very well on training data but poorly on new data.
        - **Generalization Issues** - Finding the balance between being parameter specific and maintaining generality, where models can predict unseen values accurately, is difficult.
    - **Dependency on Pre-Trained Models**
        - **Licensing and Usage Constraints -** Pre-trained models might have restrictions on use, modification, or distribution.

        - **Model Limitations** - The fine-tuned model's performance and abilities depend on the original pre-trained model.

- **Cost Considerations**
    - **Lower Costs Compared to Pre-Training** - Less need for computational resources and data gathering.
    - **Operational Efficiency** - Efficient training cycles and lower infrastructure costs enhance overall cost-effectiveness.
- **AWS Implementation**
    - AWS provides strong support for fine-tuning with Amazon SageMaker. It gives managed environments and tools to simplify the process.
        - **Amazon SageMaker JumpStart** - offers pre-trained models. Users can modify these models for specific tasks.
        - **Amazon SageMaker Experiments** - helps track and manage fine-tuning tasks efficiently.
- **Sample Scenario**
    - **PanaloCall**, a fictional Philippine-based customer service provider, fine-tunes a pre-trained language model from AWS SageMaker JumpStart to handle mixed Tagalog-English (Taglish) call center interactions.
    - By tailoring the model to a curated dataset of real-world call transcripts—covering billing, tech support, and cultural nuances like holiday greetings—PanaloCall achieves high accuracy for intent detection and sentiment analysis without the overhead of training a foundation model from scratch.
    - Leveraging smaller, domain-specific datasets reduces compute costs and enables quicker deployment, ideal for the company's fast-paced client onboarding. Still, fine-tuning requires skilled ML engineers to set optimal hyperparameters and handle edge cases (e.g., unique slang).
    - Overfitting risks arise if the Taglish datasets are too narrow, potentially limiting broader language coverage. Moreover, PanaloCall must honor licensing agreements and rely on the strengths (and inherent limitations) of the base model.
    - Overall, the AWS SageMaker AI environemnt, including Experiments tracking, helps streamline fine-tuning workflows and keeps resource usage in check—allowing PanaloCall to deliver responsive, localized call center solutions across the Philippines.

In-Context Learning

In-Context learning uses the power of large language models (LLM) to adjust to new tasks with no changes to the model weights by simply providing the model with carefully designed prompts that direct its behavior.

- **Pros:**
  - **No Need to Alter Model Weights** - In-context learning does not need retraining or fine-tuning. This saves a lot of compute and storage requirements, as no updates to model weights are needed. This gives efficiency as it cuts down on costs and time as it does not involve complex retraining or managing big datasets.
  - **Flexibility in Prompting** - This method lets organizations try different tasks quickly. They adjust the prompt to guide the model's responses.

- **Cons:**
  - **Limited by the Model's Existing Capabilities**

    - **Contextual Limits** - The model's skills come from its prior training. In-context learning suits tasks that match well with what the model already knows. It might have trouble with very specific or unusual tasks.

    - **Context Window Size** - AWS helps with learning in context through services like Amazon Bedrock. Amazon Bedrock offers pre-trained foundation models. Users access these with easy API calls. This method is very affordable. It lowers the need for complex training processes

- **AWS Implementation**

  - AWS helps with learning in context through services like Amazon Bedrock, where it offers pre-trained foundation models where users access these with easy API calls. This method is very affordable, as it lowers the need for complex training processes.

  - **Sample Scenario: EduPinoy's Adaptive Learning Tutor**
    - EduPinoy, a fictional Philippine-based online education platform, utilizes In-Context Learning through Amazon Bedrock to deliver personalized tutoring in Tagalog and Taglish without altering model weights.

    - By crafting specific prompts, EduPinoy's virtual tutors can adapt to diverse student needs—ranging from elementary math problems to advanced science explanations—ensuring responses are culturally relevant and linguistically appropriate.

    - This approach allows EduPinoy to quickly implement new teaching modules and experiment with different instructional styles, enhancing flexibility and reducing costs associated with model retraining.

- However, the effectiveness is limited to tasks within the model's existing capabilities, potentially challenging specialized subjects like local history or niche technical topics. Additionally, managing prompt designs to fit within the model's context window requires careful planning.

- Leveraging Amazon Bedrock's pre-trained foundation models, EduPinoy efficiently accesses robust language capabilities through simple API calls, enabling a dynamic and scalable learning environment that meets the evolving educational needs of Filipino students without the complexity of extensive model management.

Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) combines the advantages of finding information retrieval with generative models. Instead of relying on its trained data, RAG systems also collect outside information like documents and databases.

- **Pros**
  - **Combines Up-to-Date Information with Model Generation** - RAG lets models generate text using the latest data. This is much better than depending on fixed knowledge learned. This approach helps a lot in areas where information quickly changes. This includes news, finance and medical fields as examples, where information in these fields shifts rapidly.
- **Cons**
  - **Complexity in Integrating Retrieval Mechanisms** - Integrating retrieval methods might become very complicated. It requires a lot of resources. This ensures the model has access to the correct data at the correct time.

  - **Data Management** - RAG systems need constant management of data sources. Information retrieval depends on this management, where data must stay current and relevant.

- **AWS Implementation**

  - AWS offers several tools for RAG systems. Amazon Kendra provides enterprise search capabilities. AWS OpenSearch helps in handling large-scale data retrieval. These services connect with Amazon SageMaker AI. Overall, these combinations can be used to create end-to-end RAG workflows, and managing these systems needs a lot of effort.

B. Choose Effective Prompt Engineering Techniques

Chapter 3.6 Describe the Concepts and Constructs of Prompt Engineering

Prompt engineering is important in improving our interaction with AI models. It involves creating prompts that guide the model to produce accurate and relevant outputs. This chapter explores the key ideas and structures in prompt engineering. It includes the meaning of context, instructions, and negative prompts. Understanding these basics are necessary for creating prompts that guide the model's behavior and ensure effective communication with AI systems. Additionally, a deep dive into the model's latent space offers insights into how prompt engineering modes throughout the model's complex internal representation of concepts.

What is a Prompt?

- A prompt is a natural language message that asks the generative AI model to do a specific job. Generative AI is an artificial intelligence tool that produces new content, such as stories, dialogues, videos, images and music. Very large machine learning models power it, where these models rely on deep neural networks, and they have been trained on vast data sets.

What is Prompt Engineering?

- Prompt engineering involves guiding artificial intelligence (AI) systems to create desired outputs. While generative AI mimics human responses, it requires specific and detailed instructions for it to produce high-quality and relevant outputs. Hence, Prompt engineers select suitable formats, phrases, words and symbols. These choices help AI interact effectively with users.

Context

- In prompt engineering, context provides background information. Context helps the model give better answers. This includes the type of information the model should focus on. Context also guides the perspective for the model's response. For instance, specifying a particular industry or domain helps the model tailor its responses accordingly.
- **Usage** - The context directs the model by limiting the possible topics for its response. Context helps the model avoid giving irrelevant or too wide answers. It ties the model to a specific set of details. The correct context makes a prompt more effective. This reduces the need for many tries.

Instructions

- Instructions give direct commands or ask questions to the model. They explain what the expected output is. Effective instructions are short and clear, leaving little room for confusion.

- **Best Practices** - Being specific helps reduce confusion. One should prefer asking, “What are the best practices for implementing AWS IAM roles for securing an S3 bucket?” This question narrows down the response. The answer becomes more relevant because of the limited scope.

Negative Prompts

- **Definition** - Negative prompts describe what the model should avoid in its output. This becomes important when users need to avoid certain topics or styles in the response.
- **Application** - Negative prompts remove unwanted or harmful content. For instance, a prompt might say, "Leave out references to cloud pricing in the response." This means the output really focuses just on architecture or design principles.

Model Latent Space

- **Definition** - The latent space of a model is the high-dimensional space where the model represents concepts. These ideas are abstract and stored in a way that shows how they connect to each other.
- **Implication** - Knowing about latent space is very important for prompt design. It helps in writing prompts that lead the model through its wide knowledge. This helps the model to find the best or most correct answer. Understanding this space gives better control over what the model outputs. Hence, it can steer the direction and quality of the results.

This chapter explores techniques for improving AI model performance through prompt engineering. Zero-shot learning lets the model handle tasks without examples. Few-shot and one-shot learning provide examples to improve accuracy. These techniques use the model's ability to generalize. The chapter also discusses chain-of-thought reasoning and the use of prompt templates to help structure inputs effectively. Practitioners understand these methods to get the most out of AI models. They change model behavior to meet specific needs. Hence, this adjustment boosts both precision and flexibility.

Zero-Shot Learning

- **Concept** - Zero-shot learning is when a model completes tasks without any previous examples. The model uses its broad knowledge to tackle tasks not specifically taught.
- **Example** - A zero-shot prompt might say, “Translate the following sentence to French: ‘Hello, world!’” The model has not seen this exact sentence in French before. However, it uses its language understanding to give the right translation.

One-Shot Learning

- **Concept** - In one-shot learning, the model receives only one example. This helps the model grasp the task and organize its answer.
- **Example** - The instruction, “Translate English to French. English: ‘Hello, world!’ French: ‘Bonjour, le monde!’ Now translate: ‘Good morning!’” gives a clear pattern. This tells the model how to work with the new sentence based on the single example.

Few-Shot Learning

- **Concept** - Few-shot learning gives a few examples to boost the model’s skill on the task it faces. This method becomes very helpful when handling complex tasks. Such tasks need the model to understand from a small amount of data.
- **Example** - A prompt might show several translation pairs, like “English: ‘Hello, world!’ French: ‘Bonjour, le monde!’ English: ‘Good morning!’ French: ‘Bonjour!’” This gives the model enough context.

Chain-of-Thought

- **Concept** - Chain-of-thought prompts lead the model to explain the steps to solve a problem. These prompts ask for the reasoning process directly.
- **Application** - Chain-of-thought helps a lot with difficult problem-solving tasks. In math problems, for instance, it does not just request the solution. It guides the model to divide the problem into steps first. Then, the model finds the answer.

Tree-of-Thought prompting

- **Concept** - Tree-of-thought prompting grows from chain-of-thought by urging the model to produce many possible next steps. This process examines each choice through a tree search. This leads to deeper reasoning.
- **Example** - For the question "What are the effects of climate change?", the model might first create options like "List environmental effects" and "List social effects." It then explains each one step-by-step.

Maieutic prompting

- **Concept** - Maieutic prompting involves the model first giving an answer with an explanation. The model is then requested to explain parts of that answer in more detail. This process helps refine its reasoning.

- **Example** - For "Why is the sky blue?", the model describes the scattering of blue light. It then elaborates on why blue light scatters more than other colors.

Complexity-based prompting

- **Concept** - Complexity-based prompting uses various thought paths. From these, it chooses the path with the longest and most consistent reasoning. This approach really helps the model solve difficult problems better.
- **Example** - The model receives a hard math problem. It produces several possible solutions. The model then chooses the one with the most detailed reasoning. This ensures that the solution really makes sense.

Generated knowledge prompting

- **Concept** - Generated knowledge prompting requests the model to first produce important facts or context before finishing the task. This approach supports accuracy in the final outcome.
- **Example** - For an essay on deforestation, the model first produces facts such as "deforestation contributes to climate change" before creating the complete essay.

Least-to-most prompting

- **Concept** - Least-to-most prompting divides a problem into smaller subproblems. It solves the easier pieces first, making simpler solutions help to address more complex issues later.
- **Example** - When solving "2x + 3 = 11," the model first finds the subproblems "Subtract 3" and "Divide by 2." It solves them one at a time.

Self-refine prompting

- **Concept** - Self-refine prompting encourages the model to critique its own solution and improve it over time. This continues until reaching a set standard.
- **Example** - The model writes an essay about Filipino literature. It then reviews this essay and revises it to add more specific examples of Filipino authors. The model repeats this process until the essay becomes satisfactory. The essay needs to meet the highest level of satisfaction.

Directional-stimulus prompting

- **Concept** - Directional-stimulus prompting uses cues such as keywords. These cues guide the model to a desired output. They focus on specific parts of the task.
- **Example** - For a poem about love, the model receives prompts like "puso (heart)," "pagmamahal (loving)," and "walang hanggan (forever)." These words direct it to include them in the poem.

Prompt Templates

- **Definition** - Prompt templates have specific structures or formats to standardize how people write prompts. These templates help when people want consistent replies for different tasks.
- **Benefit** - Prompt templates give users clear and structured prompts, they follow a consistent pattern. This reduces mistakes from unclear instructions.

Chapter 3.8: Understand the Benefits and Best Practices for Prompt Engineering

Prompt engineering involves crafting good prompts and regularly improving them for high-standard outputs. This chapter discusses the key advantages of prompt engineering. It focuses on how effective prompts improve the quality of responses. They help provide more accurate and reliable AI-generated answers. Specific guidance includes trying out various prompts. Using guardrails prevents harmful outputs. Being specific and concise in prompt design helps a lot. Using these methods leads to better AI model interactions. Consistency and safety increase and efficiency gets a boost.

- **Response Quality Improvement** - Repeating the same context in prompts helps refine them, as the model gives better and more relevant answers. Users test different ways to phrase prompts, and it helps find which ones provide the clearest and most exact results.
- **Experimentation** - Experimenting with different prompts helps find the best ways to interact with the model. Trying different approaches exposes how the model understands various inputs. This discovery leads to better results in certain situations.
- **Guardrails** - Guardrails are actions that set limits for model outputs. They help stop the generation of harmful or unsuitable content. Developers place guardrails in prompts, which discourages the model from moving into unwanted areas.
- **Discovery -** Discovering in prompt engineering involves exploring the model's ability given in different types of prompts. It will train the user how to adjust with the model's knowledge and how it provides different types of output.

- **Specificity and Concision Strategy** - Clear and concise prompts help avoid misunderstandings. Precise language lets the user communicate effectively. The model focuses on the main requirements and irrelevant details do not cause distraction.

- **Multiple Comments** - Instructions should have smaller, step-by-step parts to create clarity. It guides the model through a process, and complex tasks become easier to handle.

Chapter 3.9: Define Potential Risks and Limitations of Prompt Engineering

Prompt engineering increases the usefulness and accuracy of AI models. However, it carries some risks and limitations too. This chapter talks about the potential hazards of using prompt engineering. These include exposure of sensitive data, model poisoning, hijacking, and jailbreaking. All these issues potentially harm the integrity and safety of AI systems. The chapter also offers ways to mitigate these risks. These ways include monitoring inputs, using validation checks, and applying advanced safeguards. Understanding these risks is very important for using AI responsibly. It is crucial to keep models within set boundaries, as models should not generate harmful or unintended results.

Exposure

- **Risk**: Sharing sensitive data by accident through model outputs is a significant concern. If a prompt contains personal or confidential information, there is a chance the model might reveal it in its reply.

- **Mitigation**: Avoid putting any personally identifiable information (PII) in the prompts. Conducting frequent security checks is very important to reduce exposure.

Poisoning

- **Risk**: Poisoning refers to the act of feeding malicious inputs to distract the model performance. It potentially reduces its accuracy over time, which unknowingly can produce biased or incorrect outputs, which may potentially cause harm.

- **Mitigation**: Regular monitoring of inputs is necessary to detect and filter out any harmful or malicious data. Validation checks and a review of the training data can help ensure that the model maintains high-quality performance.

Hijacking

- **Risk**: Occurs when adversarial prompts deliberately sabotage the model's behavior, steering it away from the intended tasks. Outputs might become irrelevant or dangerous.

- **Mitigation**: Prompt filtering and output moderation systems really protect against hijacking. Frequent updates to the model's rules prevent unwanted actions.

Jailbreaking

- **Risk**: Jailbreaking involves manipulating prompts to skip safety measures and controls. This allows the model to create harmful or improper outputs.
- **Mitigation**: Regular updates to the model's defenses can stop jailbreaking. Extra layers of protection like automated moderation tools really strengthen security.

C. Describe the Training and Fine-Tuning Process for Foundation Models

Chapter 3.10: Describe the Key Elements of Training a Foundation Model

Training a foundation model involves several critical stages. Each step is crucial for creating a strong and adaptable AI system. These steps include Pre-Training, Fine-Tuning and Continuous Pre-Training. Using AWS resources can streamline and enhance each of these elements.

Pre-Training

Pre-training is the initial phase in which a machine learning model is trained on a large, general-purpose dataset. This process occurs before the model is fine-tuned or applied to a specific task. During pre-training, the model learns general patterns, features, and representations from a diverse array of data. As a result, it becomes useful for various downstream tasks.

- **Purpose** - Pre-training gives the model a vast understanding of general knowledge. It does this by showing the model many different types of datasets. This basic knowledge lets the model handle a wide range of tasks effectively.
- **Pre-trained models** - Pre-training is a common approach for models like GPT (for language) or BERT (for NLP), where a model learns from massive datasets (such as books, websites, etc.) before being fine-tuned for a specific task (e.g., sentiment analysis).
- **Use cases** - Pre-training is valuable in fields where collecting a large labeled dataset is challenging, such as natural language processing (NLP) or computer vision. By pre-training on a vast corpus, models can apply their knowledge to new tasks with relatively smaller labeled datasets.
- **Implementation with AWS** - Organizations can utilize Amazon SageMaker AI's distributed training to efficiently expand the pre-training process. This service distributes the training workload across many instances. It very significantly shortens the time needed to train large models. SageMaker AI's scalable infrastructure handles vast datasets and complex model architectures with ease.

Fine-Tuning

Fine-tuning is the process of taking a pre-trained machine learning model and adapting it to perform effectively on a specific task or dataset. Fine-tuning typically involves using an existing, pre-trained model from Amazon SageMaker AI or a framework like TensorFlow or PyTorch. Users then train this model on their own labeled dataset. Fine-tuning enables the model to generalize better to specific data without the need to train it from scratch, making the process more efficient and less resource-intensive.

- **Purpose** - Fine-tuning changes the pre-trained model for special tasks by training it on specific data. This process improves the model's accuracy for certain applications. It becomes more effective and suitable for specific situations.
- **Use cases** - Fine-tuning is often used for tasks like image classification, sentiment analysis, or other specialized NLP tasks.
- **Implementation with AWS** - Amazon SageMaker AI provides detailed tools for fine-tuning models. SageMaker allows users to customize the model with personal datasets. People can track training progress and deploy the fine-tuned model for inference. It offers an all-in-one environment that simplifies the fine-tuning process. Rapid iteration and quick adjustments are possible.

Continuous Pre-Training

Continuous Pre-Training means training the model in an on-going manner, to add new data and knowledge. This process keeps the model updated, relevant, and performs well with new information.

- **Benefit** - Regularly updating the model keeps it accurate and adapts to new trends, data changes, and user needs. This regular updating helps prevent the model from becoming outdated and increases its usability lifespan.
- **Implementation with AWS** - AWS supports continuous pre-training. Amazon SageMaker Pipelines help automate the end-to-end machine learning process. Setting up these pipelines lets the data ingestion and model retraining. By doing so, this routine keeps the foundation model up-to-date.

Chapter 3.11: Define Methods for Fine-Tuning a Foundation Model

Fine-tuning a foundation model uses different methods that improve its performance and adaptability. Important methods include Instruction Tuning, Adapting Models for Specific Domains, Transfer Learning, and Continuous Pre-Training.

Instruction Tuning

Instruction Tuning means adjusting the model with datasets that have clear instructions and expected answers. This method helps the model understand and follow human directions better. The model develops the skill to generate relevant and accurate responses.

- **Benefit** - Adding instructions to training data can make the model better at understanding and following user commands. This improvement results in more reliable and user-friendly outcomes.
- **Implementation with AWS** - Amazon SageMaker AI creates custom training jobs that include datasets with instructions. SageMaker AI provides a flexible training environment. This platform lets you design specific instruction tuning processes. You also execute these processes suited to your needs.

Adapting Models for Specific Domains

- **Approach** - Adapting models for particular fields uses specific data to improve the model's skills in specialized areas. This way, the model really understands and knows how to accurately handle words and ideas unique to certain fields.
- **Example** - For example, fine-tuning a model with medical terms helps it better understand and create content about healthcare, medical research and clinical uses.
- **Implementation with AWS** - Amazon SageMaker AI offers the tools needed to fine-tune models with special datasets. By using SageMaker AI's data management and training features, people efficiently adapt their basic model to succeed in particular fields.

Transfer Learning

Transfer Learning is a machine learning technique in which a pre-trained model, typically trained on a large dataset, is adapted for a new but related task. The fundamental concept is that knowledge acquired from one task can be applied to another task that may have limited data. This approach significantly reduces the requirement for substantial amounts of labeled data for the new task, making the process more efficient.

- **Process** - In transfer learning, a pre-trained model that has already been trained on a large and general dataset, such as ImageNet for image classification or GPT for text processing, is used as a foundation. This model is then fine-tuned on new, task-specific data. By leveraging the features learned from the initial training, the model can improve its performance even when there is limited data for the specific task.
- **Use cases** - Transfer learning is particularly useful when:
    - The new dataset is small or expensive to label.
    - The new task is closely related to the original task the model was trained on.
    - User want to speed up training by utilizing a pre-existing, high-quality model.

- **AWS Implementation** - AWS helps with transfer learning through models on the AWS Marketplace. Users can choose a suitable pre-trained model from the marketplace. They then train it further on their dataset with Amazon SageMaker AI. This method saves time and computing resources required to develop high-performing models. Additionally, Amazon SageMaker AI supports transfer learning by allowing you to fine-tune pre-trained models for custom tasks. AWS also provides popular pre-trained models through services like SageMaker JumpStart, where you can access models trained on a wide range of tasks (image classification, object detection, etc.) and fine-tune them for specific needs.

Continuous Pre-Training

Continuous pre-training is the ongoing process of training a machine learning model with new data over time. This approach is particularly beneficial in dynamic environments where the data is constantly changing. Instead of starting the training process from scratch or fine-tuning the model only occasionally, the model is regularly updated with new data to remain relevant and enhance its performance over time.

- **Process** - Continuous pre-training involves regularly updating the model with new data. This keeps performance strong and relevant. Regular updates allow the model to learn the latest information and adapt to changing data patterns.

- **Use cases** - Continuous pre-training is essential when::
    - The data is constantly evolving (e.g., social media data, stock market data).
    - The data needs to stay up-to-date with the latest information.
    - Real-time or near-real-time predictions are required.

- **Implementation with AWS** - Continuous pre-training can be managed using Amazon SageMaker AI and its capabilities such as **SageMaker Pipelines** and **SageMaker Model Monitor**:
    - **SageMaker Pipelines** helps automate workflows to retrain models with new data, ensuring models are continuously updated without manual intervention.
    - **SageMaker Model Monitor** allows for monitoring data drift and retraining models when performance drops due to changes in incoming data.
    - **SageMaker Batch Transform or SageMaker Endpoint** can be used for real-time inference while maintaining continuous learning by feeding new data back into the system.

Chapter 3.12: Describe How to Prepare Data to Fine-Tune a Foundation Model

Preparing data for fine-tuning a foundation model is an essential task. This step makes the model work effectively and accurately. Key aspects of data preparation include data curation, governance, size, labeling, representativeness, and reinforcement learning with human feedback.

Data Curation

Data curation is the process of gathering, organizing, and managing datasets to ensure their quality and relevance for a specific task or machine learning model. This process includes selecting appropriate data

sources, cleaning the data, transforming it into a suitable format, and ensuring that it meets the project's requirements. Additionally, data curation involves continuous updates and the removal of irrelevant or outdated data.

- **Importance** - Correct labels are important for supervised learning. They help the model understand the right answers for specific inputs. High-quality labels help the model give precise predictions.
- **Implementation with AWS** - Amazon SageMaker Data Wrangler helps prepare data. This tool offers ways to explore, clean and transform data. With Data Wrangler, users quickly curate datasets. This process helps data meet quality standards needed for fine-tuning.

Governance

Governance refers to the management of data availability, usability, integrity, and security within an organization. It establishes policies, standards, and procedures for handling data throughout its lifecycle, ensuring compliance with legal regulations (such as GDPR) and aligning with business objectives.

- **Compliance** - Meeting regulatory requirements for your data, like General Data Protection Regulation (GDPR), is very important for legal and ethical reasons. Proper governance protects sensitive information. It keeps your data in line with protection laws.
- **Implementation with AWS** - AWS Artifact offers on-demand access to AWS's compliance reports. These reports help you check that your data handling follows relevant rules. Artifact helps you oversee compliance successfully and confirms that your data practices uphold industry standards.

Size

Data size refers to the amount of data used to train machine learning models. The quantity of data significantly impacts the performance of these models. In general, larger datasets tend to improve generalization and lead to more accurate predictions, although they may require additional computational resources for processing and training.

- **Consideration**: The amount of data used in fine-tuning really affects the model's performance. Enough data helps the model learn and understand new inputs properly.
- **Implementation with AWS -** Amazon S3 provides a solution for storing your data. It allows you to expand your datasets to the needed size for fine-tuning. SageMaker connects with S3 easily. This connection lets you handle large datasets efficiently during fine-tuning.

Labeling

Data labeling is the process of annotating or tagging raw data with meaningful information or categories. For supervised learning tasks, labeled data is essential for training the model to recognize patterns and make predictions. Labels are often created manually or with the help of automated tools.

- **Requirement** - Correct labels are important for supervised learning. They help the model understand the right answers for specific inputs. High-quality labels help the model give precise predictions.
- **Implementation with AWS** - Amazon SageMaker Ground Truth provides scalable and accurate labeling. Ground Truth employs machine learning to help human labels. This improves both the speed and quality of labeling. Your data gets accurately labeled with this service, which is crucial for fine-tuning.

Representativeness

Data representativeness refers to how accurately a dataset reflects the real-world scenario or problem that the model aims to address. A representative dataset includes samples that encompass the various types of inputs the model may encounter during its use, ensuring that the model can effectively generalize to new, unseen data.

- **Goal** - Data must show the diversity of real-world situations. This helps the model work well in different situations and uses. A good dataset prevents biases and helps improve the model's generalizability.
- **Implementation with AWS** - Use Amazon SageMaker Feature Store to control and watch feature diversity in your datasets. Cover many scenarios and demographics in your data. This approach strengthens the model's ability to deal with different inputs successfully.

Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning with Human Feedback (RLHF) is a method that integrates reinforcement learning (RL) with feedback from humans to enhance the training process. In RL, an agent learns to complete tasks through trial and error, receiving rewards or penalties based on its performance. By incorporating human feedback, the training can be accelerated, as it provides important guidance, preferences, or corrections that help direct the agent toward optimal behavior. This approach reduces both the time and the amount of data needed for effective training. Here, humans, or the users, give guidance to the model in Reinforcement Learning from Human Feedback. By doing so, Models learn better with human advice. This method aligns the model's outputs with human values and preferences. Model results become relevant and useful.

- **Benefit**: RLHF keeps the model's behavior and replies consistent with human expectations. User satisfaction and trust in the model's outputs usually increase.

- **Implementation with AWS** - AWS does not have a special RLHF service. You may integrate Amazon SageMaker AI with custom feedback systems. SageMaker's training and deployment capabilities assist in its implementation, along with human feedback loops. You can implement RLHF effectively in your training pipeline using these resources.

D. Describe Methods to Evaluate Foundation Model Performance

Evaluating how foundation models work is crucial to meet the desired objectives and deliver the desired value. This chapter explores different methods and metrics for evaluation. AWS resources help to streamline the evaluation process.

Chapter 3.13: Understand Approaches to Evaluate Foundation Model Performance

Evaluating foundation models thoroughly requires both qualitative and quantitative methods. AWS offers various tools for these evaluation methods, which work very effectively.

Human Evaluation

- **Method** - Human evaluation uses opinions from experts or target users. This approach is very useful for tasks where quantitative metrics might not capture every detail of how well a model works.
- **Application** - Use human evaluation for content generation, sentiment analysis and user interaction tasks. Understanding context and small details is really important here.
- **Implementation with AWS** - Use Amazon SageMaker Ground Truth to streamline human evaluations. Ground Truth lets you build workflows for human evaluators to review model outputs.
    - **Human-in-the-Loop**: This method incorporates human input throughout the machine learning process to enhance the accuracy and relevance of models. Humans can assist with tasks such as data creation, annotation, model review, customization, and evaluation.

Benchmark Datasets

- **Purpose** - Benchmark datasets offer standard tests to allow fair comparisons between models. They act as a trusted point of reference to measure and compare model performance across many tasks.
- **Examples** - Popular benchmark datasets are General Language Understanding Evaluation (GLUE) and SuperGLUE. These are very common for testing language understanding.
- **Implementation with AWS** - Use Amazon SageMaker AI to connect benchmark datasets to your evaluation pipeline. SageMaker helps with ingestion, processing, and testing models with these standard benchmarks, making it consistent and can be repeated for several assessments.

Chapter 3.14: Identify Relevant Metrics to Assess Foundation Model Performance

Selecting the appropriate metrics is essential for accurately assessing a model's performance. AWS provides tools designed to help compute and analyze these metrics, significantly enhancing the evaluation process.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

- **Purpose** - ROUGE measures the overlap of n-grams between the generated and reference texts, offering valuable insights into the quality of text summarization and generation
- **Usage** - ROUGE is commonly used in summarization tasks to evaluate how well a model captures key information from the source text.
- **Implementation with AWS** - Implement ROUGE evaluations using **Amazon SageMaker AI** by integrating ROUGE computation scripts within your SageMaker training or evaluation jobs. SageMaker's adaptable setup lets users automate the calculation of ROUGE scores in the model evaluation process.

Bilingual Evaluation Understudy (BLEU)

- **Usage** - BLEU checks the quality of text translated by machines against human versions. In other words, it looks at exact matches in short word sequences.
- **Application** - Standard metric for machine translation tasks. It checks how closely model translations match human versions.
- **Implementation with AWS** - Amazon SageMaker Processing Jobs help compute BLEU score computations. SageMaker Processing helps you prepare data and calculate evaluation metrics like BLEU in a scalable and managed environment.

BERTScore

- **Usage** - BERTScore measures similarity by leveraging contextual embeddings from BERT models. It provides a deeper understanding of semantic similarity compared to surface-level metrics such as ROUGE and BLEU.
- **Advantage** - BERTScore delivers a more nuanced evaluation of text similarity by utilizing deep contextual information. This method is especially effective for tasks requiring semantic alignment, which is crucial for accurate assessments.
- **Implementation with AWS** - Users can integrate BERTScore evaluations into Amazon SageMaker AI by using pre-trained BERT models from SageMaker AI's model repository. SageMaker AI's environment ensures seamless execution of complex embedding-based metrics.

Chapter 3.15: Determine Whether a Foundation Model Effectively Meets Business Objectives

Evaluating whether a foundation model aligns with and supports business objectives is crucial for maximizing value and achieving positive outcomes. This chapter explores three key areas: Productivity, User Engagement, and Task Engineering. It examines how to leverage AWS resources to assess and enhance each of these aspects effectively.

Productivity

Raising productivity means checking how the main model helps work go smoothly and cuts expenses. AWS provides many tools for users to watch, study and improve these signs. These tools hold a key part in increasing efficiency and success.

- **Metrics**: **Task Completion Time, Automation Levels, Cost Savings**
    - **Task Completion Time** - Measure how quickly the model completes specific tasks compared to traditional methods. Shorter times indicate higher efficiency, and faster results are highly valuable.
    - **Automation Levels** - Evaluate the extent to which the model automates repetitive or time-consuming tasks. Increased automation allows people to focus on more important activities, and freeing up human resources is essentia**l.**
    - **Cost Savings** - Assess the financial savings generated by using the model. Savings stem from reducing manual effort, minimizing errors, and optimizing resource usage. Lowering costs is a significant benefit.
    - **Implementation in AWS** - Amazon CloudWatch can be used to monitor and collect real-time data on task performance and operational efficiency. It helps track task completion times, automation levels, and cost-related metrics. CloudWatch provides detailed dashboards and offers alerts to keep users informed about any changes or issues.

Evaluation: Assess if the Model Streamlines Operations

- **Assessment Method** - Evaluate whether the foundation model truly enhances operational workflows. Look for a reduction in bottlenecks, fewer errors, and faster processes. Compare key performance indicators (KPIs) from before and after implementing the model to gain insights into process improvements.

- **Implementation with AWS** - AWS Cost Explorer helps evaluate cost reductions. AWS Lambda automates and simplifies workflows. When used with other AWS services, AWS Lambda allows the building of serverless applications. These applications improve task performance and lower extra costs. Overhead costs reduce significantly.

User Engagement

- Understanding user engagement is very important for a positive end-user experience. AWS offers strong tools and analytics to watch and improve how users interact. Some important indicators are User Satisfaction Scores, Retention Rates and Interaction Frequency.

- **Indicators: User satisfaction scores, retention rates, interaction frequency.**

    - **User Satisfaction Scores** - Surveys, feedback forms, and sentiment analysis help measure user satisfaction. High satisfaction scores indicate that the model likely meets or exceeds user expectations. User happiness is a key factor in success.

    - **Retention Rates** - Retention rates measure how many users continue to use your service or product over time. High retention rates suggest that the model consistently provides value, encouraging users to return.

    - **Interaction Frequency** - Track how often users engage with the model. Frequent interactions indicate that users find the model valuable and engaging. High interaction frequency signals that users enjoy using the model.

    - **Implementation with AWS** - Use Amazon Pinpoint to collect and analyze user engagement data. Pinpoint provides detailed insights into how users interact with the model. It includes retention rates and satisfaction levels. This data-driven approach helps make informed decisions to enhance the overall user experience.

- **Analysis: Determine the model's impact on the end-user experience.**

    - **Analysis Approach** - Examine the collected participation data to identify patterns, strengths, and areas for improvement. Correlate user satisfaction and continued usage with specific model features to assess their impact.

    - **Implementation with AWS -** Utilize Amazon QuickSight for advanced data visualization and analysis. QuickSight enables the creation of interactive dashboards, highlighting user participation patterns and uncovering valuable insights from Amazon Pinpoint data.

Task Engineering

- Task Engineering ensures that the outputs of the foundation model are closely aligned with business objectives and adaptable to evolving needs. AWS provides tools to facilitate this alignment.
- **Alignment:Match the Model's Outputs with Business Goals and Tasks**
    - Establish clear business goals and link them to specific tasks for the model. Regular reviews and adjustments ensure that the model's outputs remain aligned with these goals, making this process crucial for sustained success.
- **Implementation with AWS**
    - Use AWS CodePipeline to create continuous integration and continuous deployment (CI/CD) workflows. CodePipeline automates the deployment of model updates, ensuring that the model consistently delivers accurate and relevant outputs. This automation makes the process reliable and efficient.
- **Iteration: Continuously refine the model based on feedback and performance data.**
    - **Iteration Approach** - Adopt a step-by-step development process that regularly updates the model based on user feedback, performance data, and evolving business requirements. This ensures the model remains relevant and consistently provides value.
    - **Implementation with AWS** - Leverage Amazon SageMaker AI for training and deploying models, enabling quick updates based on feedback and performance metrics. SageMaker's powerful features, such as automatic model tuning and A/B testing, support ongoing improvements, ensuring that model changes are both effective and accurate.

Domain 3: Applications of Foundational Models Sample Questions

1. An AI specialist is in the process of studying the capabilities of foundational models (FMs) to enhance the company's AI-driven solutions. These powerful models can be fine-tuned for various tasks based on extensive pre-training on large datasets. The specialists need to understand the different capabilities of foundation models and how they can be applied.

   Select the correct tasks that FMs can perform to help the company improve its AI-driven solution. (Select THREE.)

   It has the capability to identify objects, scenes, and other elements within images. **Visual comprehension**

   It can answer natural language questions and even write short scripts or articles in response to prompts. **Language processing**

   It is designed for tasks like transcription and video captioning in various languages. **Language translation**

Explanation:
Even though foundation models (FMs) are pre-trained, they can keep learning from data inputs or prompts during inference. This means that you can generate detailed outputs by using well-crafted prompts. Some tasks that FMs can handle include language processing, visual comprehension, code generation, and human-centered engagement.

Hence, the correct answers are:

- **Visual comprehension** has the capability to identify objects, scenes, and other elements within images.
- **Language processing** can answer natural language questions and even write short scripts or articles in response to prompts.
- **Speech to text** is specifically designed for tasks like transcription and video captioning in various languages.

2. A company is testing a model for generating text summaries and needs a metric to evaluate how well the summaries match human-created references.

   Which metric is most appropriate for this evaluation?

   a. BLEU (Bilingual Evaluation Understudy)
   b. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
   c. F1 Score
   d. Cross-Entropy Loss

Explanation:
ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics specifically designed to evaluate the quality of text summaries by comparing them to human reference summaries. It measures the overlap of n-grams, word sequences, and word pairs between the generated and reference summaries, focusing on recall. This makes ROUGE particularly effective for assessing how well a model-generated summary captures the critical content of the original text.

The following list includes the names and descriptions of the ROUGE metrics available after fine-tuning big language models in Autopilot.

- ROUGE-N (e.g., ROUGE-1, ROUGE-2) measures the overlap of n-grams between the generated text and the reference text, with n indicating the size of the n-grams.
- ROUGE-L calculates the longest common subsequence between the generated and reference texts, considering both content overlap and word order.
- ROUGE-L-SUM is a variant of ROUGE-L specifically designed for summarization tasks, focusing on the longest common subsequence while accounting for word order in the summaries.

ROUGE is the most appropriate metric for evaluating text summaries because it is specifically designed for this purpose. It measures how much of the important information from the reference summary is captured in the generated summary, which is essential for determining the effectiveness of a summarization model.

References for Domain 3

Identify Selection Criteria to Choose Pre-trained Model

https://theonetechnologies.com/blog/post/how-to-choose-the-right-pre-trained-ai-model-for-your-application

Define Retrieval Augmented Generation (RAG) and Describe its Business

https://aws.amazon.com/what-is/retrieval-augmented-generation/

Identify AWS Services That Help Store Embeddings Within Vector Databases

https://aws.amazon.com/what-is/vector-databases/

Amazon OpenSearch Service

https://docs.aws.amazon.com/opensearch-service/

https://docs.aws.amazon.com/opensearch-service/

Amazon Aurora

https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/CHAP_AuroraOverview.html

Amazon Neptune

https://docs.aws.amazon.com/neptune/

Amazon DocumentDB (with MongoDB Compatibility)

https://docs.aws.amazon.com/documentdb/

Amazon RDS for PostgreSQL

https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_PostgreSQL.html

Amazon SageMaker JumpStart

https://aws.amazon.com/sagemaker/jumpstart/

Amazon SageMaker Experiments

https://aws.amazon.com/sagemaker/experiments/

Describe the Concepts and Constructs of Prompt Engineering

https://aws.amazon.com/what-is/prompt-engineering/

Prompt Engineering

https://aws.amazon.com/what-is/prompt-engineering/

Pre-Training

https://docs.aws.amazon.com/sagemaker/latest/dg/distributed-training.htmlv

Fine-Tuning

https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

Continuous Pre-Training

https://docs.aws.amazon.com/sagemaker/latest/dg/pipelines.html

Instruction Tuning

https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html

Adapting Models for Specific Domain

https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

Transfer Learning

https://aws.amazon.com/marketplace/solutions/machine-learning

Continuous Pre-Training

https://docs.aws.amazon.com/sagemaker/latest/dg/pipelines.html

Data Curation

https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html

Governance

https://aws.amazon.com/artifact/

Size

https://docs.aws.amazon.com/s3/

Labeling

https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

Representativeness

https://docs.aws.amazon.com/sagemaker/latest/dg/feature-store.html

Understand Approaches to Evaluate Foundation Model Performance

https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

ROUGE

https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-accuracy-evaluation.html

Implementation in AWS

https://docs.aws.amazon.com/cloudwatch/

https://docs.aws.amazon.com/cost-management/latest/userguide/ce-what-is.html

https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

https://docs.aws.amazon.com/pinpoint/latest/userguide/welcome.html

https://docs.aws.amazon.com/codepipeline/latest/userguide/welcome.html

https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

Determine the model's impact on the end-user experience

https://docs.aws.amazon.com/pinpoint/latest/userguide/welcome.html

Others

https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html
https://aws.amazon.com/bedrock/
https://aws.amazon.com/what-is/foundation-models/
https://tutorialsdojo.com/amazon-bedrock/

GUIDELINES FOR RESPONSIBLE AI

Key Features of Responsible AI

Transparent and Explainable AI Models

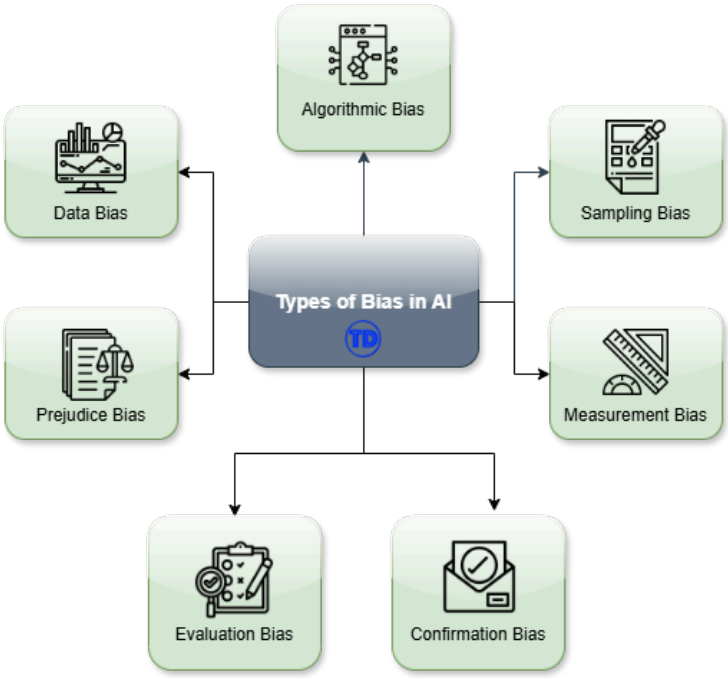
A. Key Features of Responsible AI

Chapter 4.1 Understanding the Features of Responsible AI

Bias

Bias in AI means unfair treatment by artificial intelligence. This happens when AI systems give unfair results to specific groups or people. Such bias usually comes from the data used to teach, but it can also comes from the algorithms themselves.

Types of Bias in AI



- **Algorithmic Bias:** This happens when algorithms are not fair. Some groups benefit more than others. This leads to unfair results. Designing and building these algorithms causes this issue.
    - An example is when an AI hiring system used by a Philippine company might probably favor people from cities like Metro Manila. This happens because the algorithm looks for keywords. These keywords are often found in resumes from city job markets. Rural candidates may not use these words, which affects their chances.
- **Data bias:** This happens when training information is not balanced or fair. It mirrors current prejudices. It also shows present inequalities.

    - For instance, a diagnostic tool that uses AI and learned mostly from private hospital data in the Philippines might not work well in public hospitals. Demographics in public hospitals differ. Health conditions might be really different too.

- **Sampling Bias:** Occurs when the collected data fails to reflect the intended group. This results in incorrect conclusions. Data must show actual results.
    - For example, a sentiment analysis AI tool studies social media opinions about a Philippine political figure. This tool often uses data from users in English-speaking areas. It sometimes overlooks feelings shown in Filipino. It also neglects languages like Cebuano or Ilocano.
- **Prejudice Bias:** Inserts societal stereotypes into AI systems. This occurs through biased training data. AI reflects these embedded stereotypes. Bias influences AI outputs and decisions.
    - An example is an AI tool used by the police in the Philippines, which might show and continue old social biases. It may be people by their race or social class. This results in unfair policing actions. Discrimination may happen.
- **Measurement Bias:** Occurs when tools or methods used for data collection create errors or changes.
    - An AI-based education platform in the Philippines evaluates student performance. It uses standardized tests to do this, which might lead to unfair results. Regional differences in language proficiency may affect outcomes. Access to educational resources also varies.
- **Evaluation Bias:** Measures or methods used in the review are sometimes biased, favoring certain results. This influences how AI performance is judged. AI performance assessment might not be fair.
    - A Filipino online shopping platform uses AI to suggest products. It might test its recommendation system with measures favoring popular city goods. This could harm lesser-known rural artisan products when measuring performance.
- **Confirmation Bias:** It happens when developers' preconceptions affect the development process. The AI then reinforces existing beliefs.
    - Filipino AI developers probably create a recommendation system for a streaming service. This system might focus on the developers' own viewing habits. The AI will then suggest content similar to what the developers like. It ignores the varied tastes of the larger Filipino audience.

Fairness

Fairness in AI guarantees that artificial intelligence systems decide without bias. These systems treat everyone fairly. Fair Treatment for All involves creating AI that offers the same chances for everyone. People of different races, genders, ages, and economic backgrounds benefit.

Inclusivity

Inclusivity in AI means making sure that artificial intelligence tools are available and good for everyone. Designing AI systems for different users requires attention to the needs and wants of diverse groups.

Veracity & Robustness

Ensuring veracity requires strong data validation, regular monitoring, and updates to maintain correct outputs. Robustness focuses on building AI systems that perform consistently across diverse scenarios, including when faced with unexpected inputs, by designing and testing models to handle a wide range of conditions.

Safety

Thorough testing is essential, and monitoring AI systems helps prevent accidents. It is also important to mitigate misuse and negative results by testing models for robustness, implementing safeguards, and correcting errors during deployment. Prioritizing safety protects users, systems, and society from potential harm caused by AI technologies.

Explainability

Providing clear understanding and evaluation of AI system outputs to ensure decisions made by AI are interpretable. This involves providing clear insights into how models process data, make predictions, or arrive at decisions, enabling transparency and accountability.

Privacy and Security

Appropriately obtaining, using, and protecting data and models to safeguard sensitive information. This involves encrypting data, enforcing access controls, complying with regulations like the General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA), and safeguarding models against vulnerabilities such as adversarial attacks or data breaches. These measures help maintain trust and ensure AI systems are used responsibly.

Controllability

Controllability in AI ensures that systems can be monitored, adjusted, and guided to align with intended outcomes. This includes implementing mechanisms to manage AI behavior, enabling overrides when necessary, and ensuring the system adheres to predefined goals. Controllability helps maintain reliability, safety, and trust in AI systems during deployment and operation.

Transparency

It involves providing clear documentation about data usage, model design, and decision-making processes, enabling users to make informed choices and fostering trust. Transparency is key to accountability, ethical AI use, and regulatory compliance.

Governance

Governance in AI involves establishing frameworks and processes to manage the ethical, legal, and operational aspects of AI systems throughout their lifecycle. This includes defining accountability, ensuring compliance with regulations, managing risks, and monitoring performance. Effective governance promotes responsible AI development and deployment, aligning AI systems with organizational values and societal expectations.

Chapter 4.2 Tools and Practices for Implementing Responsible AI

Using Guardrails for Amazon Bedrock

Amazon Bedrock offers strict rules to support the responsible and ethical use of AI models. These rules include many features to keep the AI model being used as honestly as possible, which aims to protect users. They also follow regulatory standards.

Features and Capabilities

- Build **responsible AI applications** with Amazon Bedrock Guardrails
    - Amazon Bedrock Guardrails introduces flexible safety features to current protections. This system blocks up to 85% more harmful content and filters more than 75% of false replies. Users can manage safety, privacy, and accuracy in one solution.
- Bring a **consistent level of AI safety** across all the applications
    - Amazon Bedrock Guardrails offers flexible safety and privacy protections. It checks user inputs and answers from the AI. These protections work with all Bedrock language models.
- **Block undesirable topics** in the generative AI applications
    - Amazon Bedrock Guardrails helps define and block specific topics. This tool ensures AI stays within company rules.
- **Filter harmful content** based on responsible AI policies
    - Amazon Bedrock Guardrails offers customizable filters. These filters stop harmful content like hate speech, violence, and wrong words which are blocked. Users can also set levels for filters.
- **Redact sensitive personally identifiable information (PII)** to protect privacy
    - Amazon Bedrock Guardrails finds sensitive data in what users write and say. They spot personally identifiable information that can be hidden. This could mean taking out personal details from call center summaries.
- **Block inappropriate content** with a custom word filter

    - Amazon Bedrock Guardrails lets users choose specific words to avoid in AI chats. The system blocks bad words and rude language. This keeps user chats safe and very controlled.
- **Detect hallucinations** in model responses using contextual grounding checks
    - Organizations need trustworthy AI to maintain user's trust. Sometimes, foundation models give wrong information. Amazon Bedrock Guardrails finds and removes these mistakes by checking the context to verify details. The goal of this is to generate accurate and relevant AI responses.

Chapter 4.3 How Guardrails Help in Ensuring AI Responsibility

Promoting Ethical AI Use

- Filters block bad content. Ethical rules guide AI systems to avoid harmful or biased outputs.

Improving Transparency and Accountability

- Detailed monitoring helps trace AI decisions, and logs can help understand model behavior.

Following Compliance with Regulations

- Data privacy and security steps protect companies. These steps help follow local and international laws such as the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA).

Stopping Misuse and Abuse

- Fail-safe systems reduce the risk of misuse and keep organizations and users secure.

Helping Responsible Innovation

- Flexible rules let organizations explore AI and keep control. These customizable rules balance progress with responsibility.

Chapter 4.4 Responsible Model Selection Practices

Selecting AI models relies on ethical standards, performance, and their potential effects. These choices help find AI that is both useful and reliable.

Environmental Considerations

- Considering the environmental impact of AI systems is important. AI systems use a lot of energy and can also consume many resources. This happens both when building and using them.

Sustainability Practices

- Using methods that keep AI development eco-friendly and save resources for the future.

Tradeoffs in Model Selection

- Selecting an AI model involves tradeoffs. Performance, cost, and environmental concerns are important factors, and each choice affects others. Models should use less energy, and there should be a balance between these tradeoffs when selecting a model.

Chapter 4.5 Legal Risks and Challenges in Working with Generative AI

Intellectual Property Infringement Claims

- Generative AI sometimes uses copyrighted materials without the correct permissions. This can cause legal issues, and legal disputes over these rights might occur.
- For instance, a Filipino startup is creating a design tool that uses AI. This tool might use artwork from local artists without asking first. This could lead to claims of copyright infringement.

Biased Model Outputs

- AI models sometimes lead to biased results. These results might unfairly target certain groups, putting organizations at risk. Legal problems and discrimination lawsuits could follow.
- For example, a Philippine company uses an AI-based hiring tool which may pick candidates from cities such as the National Capital Region (NCR) more often than those from the countryside. This could cause claims of regional discrimination. Regional bias could really become a serious issue.

Loss of Customer Trust

- Inaccurate or biased AI results reduce customer trust which can damage a company's name, and customer loyalty.
- A Filipino online store uses AI to suggest products. However, if it often shows the wrong items, customers might lose trust. Bad recommendations lead them to buy somewhere else.

End User Risks

- AI applications that interact directly with users risk privacy breaches. They could give wrong information. Incorrect advice is another danger.
- A health tech company in the Philippines developed a chatbot that gives medical advice to people across the country. AI errors might cause it to share wrong health information, putting Filipino users' health could be at risk. This problem might lead to legal issues for the company.

Hallucinations in AI Models

- AI models sometimes create wrong or misleading information. Experts call this "hallucinations" which might cause the spread of false news. Legal problems could also arise from these mistakes.
- A Philippine online news site using AI might accidentally produce false stories about local elections or natural disasters. These errors could confuse Filipino readers. Consequently, organizations like the Kapisanan ng mga Brodkaster ng Pilipinas (KBP) could take legal steps to spread such misinformation.

Chapter 4.6 Characteristics of Datasets for Responsible AI

Inclusivity and Diversity in Data

- Data should reflect every group to avoid biases. All demographic groups deserve representation.
- For instance, when building a dataset for healthcare research in the Philippines, researchers should not only include information from urban areas such as Metro Manila, but also from rural areas like Mindanao. Differences in access to healthcare services must be considered.

Curated Data Sources

- People should choose and check reliable, high-quality data. Very accurate data helps in analysis.
- For example, a company studying traffic in Metro Manila should use curated information from official (Metropolitan Manila Development Authority) MMDA traffic reports. Validated GPS data from ride-booking apps are also useful. They should avoid trusting unverified social media posts.

Balanced Datasets

- Equitable data distribution helps reduce bias. Proper data spreads across categories. In other words, data should be fair for everyone.
- Training an AI model to recognize Filipino dialects requires equal representation of Tagalog, Cebuano, Ilocano, and other regional languages. Dominant languages like Tagalog should not overshadow others.

Chapter 4.7 Effects of Bias and Variance in AI Models

Impact on Demographic Groups

- Bias in data or models might unfairly hurt certain groups. This leads to results that are not fair or accurate. Models sometimes predict wrongly for these groups.

- A credit scoring model using mainly city income data might unfairly treat rural Filipinos. These individuals depend on informal jobs and might be creditworthy despite not having a formal account as proof. The model could misjudge them.

Inaccuracy in Model Outputs

- Differences in predictions often occur due to problems with data quality, choice of features, or model design. Careful checking of these elements is necessary.
- A flood prediction model for the Philippines often gives wrong forecasts if it analyzes only the rainfall data for local factors like deforestation or blocked waterways, especially in places like Marikina.

Overfitting and Underfitting Issues

- Correct methods such as cross-validation and regularization find a middle ground. They stop models from being too complicated or too simple.
- A machine learning model predicting jeepney traffic in Metro Manila could struggle if it remembers exact holidays, like Fiesta dates, too much. This is called overfitting. On the other hand, the model may also underperform if it does not pay attention to changes in traffic between rush hours and weekends.

Chapter 4.8 Tools for Detecting and Monitoring Bias, Trustworthiness, and Truthfulness

These tools find and measure possible biases to check how reliable AI models are. They also test if AI follows truthfulness rules.

Analyzing Label Quality

Methods to check if data labels are right and steady are essential. They help in confirming that training datasets are accurate.

Human Audits and Subgroup Analysis

This is when processes use human reviewers to study AI model results. They focus on finding biases in different subgroups. Analyzing biases helps prevent unfair treatment.

Amazon SageMaker Clarify

A tool that evaluates models and explains model prediction. Users can automatically evaluate their foundational models for their generative AI use case with the following metrics: accuracy, robustness, and toxicity. It can also spot potential bias and other risks during data preparation. Guidelines like ISO 42001 recommend this step.

Partial dependence plots illustrate how the predicted target response varies with specific input features of interest. These plots average the effects of all other input features, often called complement features. Partial dependence reflects the expected target response as a function of each selected input feature.

Amazon SageMaker Model Monitor

This tool can automatically monitor machine learning (ML) models during production and will send alerts in case there is a problem with data quality.

Amazon Augmented AI (Amazon A2I)

This allows users to integrate human reviews into AI workflows for validation purposes.

B. Transparent and Explainable AI Models

Chapter 4.9 What Are Transparent and Explainable Models?

Transparent models in artificial intelligence help people understand decisions by showing how inputs are turned into outputs. It also brings accountability and helps spot any mistakes or unfairness. With this, stakeholders probably find it easier to trust the results of these models.

Differences Between Transparent and Opaque Models

Transparent models have clear and easy-to-understand structures. These structures help users see how inputs turn into outputs. They can provide insight into how decisions happen. On the other hand, opaque models, often called *"black-box"* models, are different. These models do not offer the same interpretability which makes it difficult to understand their predictions or decisions.

For instance…

- **Transparent Model Example:** A hospital in the Philippines uses a see-through AI model for its diagnostic tools. Doctors can clearly understand and explain each diagnosis.
- **Opaque Model Example**: A Philippine online shopping site with a mysterious recommendation system might confuse users about why specific items show up. This confusion probably leads to distrust and misunderstanding of the algorithm. Users really might question the platform's decisions.

Importance of Transparency in AI Development

Understanding and accessing AI systems' decision processes builds trust with users and stakeholders. In addition, accountability helps spot biases and errors. This approach supports ethical norms and meets regulatory rules in AI applications.

- **Transparency Importance Example:** In developing an AI system for waste management in Manila, transparency matters a lot. Residents need to see how the AI analyzes and decides on collection routes. With this kind of transparency, people can look at the process and suggest changes if needed.
- **Accountability Example:** When AI picks who gets government aid, it needs to be clear. Transparency lets people check decisions, which then guarantees fair assistance for everyone.

Chapter 4.10 Tools to Identify Transparent and Explainable Models

Amazon SageMaker Model Cards

Amazon SageMaker Model Cards allow users to describe machine learning models in a clear and consistent way. These cards give clear information about the model's creation and how it performs. People use them as a trustworthy place for all important model details, which then helps with checking, controlling, and handling models effectively.

For example, a major bank in the Philippines uses Amazon SageMaker Model Cards for its loan approval models. The bank writes down each model's design, training data, performance metrics, and evaluation steps. Every stakeholder understands this information clearly, while the compliance officers and customer service teams get the exact details. This documentation helps the bank audit the models because it checks them for fairness and accuracy.

AWS AI Service Cards

AWS AI Service Cards provide clear details about how to use AWS AI services responsibly. These cards cover the purpose, limits, and careful deployment of the services.

The following services have dedicated AI Service Cards:

- Amazon Rekognition - Face Matching
- Amazon Textract - AnalyzeID
- Amazon Transcribe - Batch (English - US)

Open Source Models and Data

Open source models and data are AI tools and information that everyone can use, change, and share. This openness allows teamwork, new ideas, and easy access. Developers and researchers from everywhere work together to improve AI technologies.

Using transparency in open-source AI involves sharing algorithms, data sources, and how decisions are made with everyone. This openness lets people check and confirm AI systems, building trust and responsibility. It also lets the community improve the systems.

Licensing Considerations

Licensing is very important for open source models and data. Licenses tell people how they may use, change or share resources. Each license, like the Massachusetts Institute of Technology (MIT), Apache or General Public License (GPL), has its own rules, and users need to follow these rules for legal and ethical use.

Regulations and laws need explainability. This includes following data protection laws like GDPR. The General Data Protection Regulation (GDPR) requires clarity on how personal data is handled by AI models. It's vital that personal data use is easy to understand.

Chapter 4.11 Tradeoffs Between Model Safety and Transparency

Model safety often involves adding protective layers, which might hide transparency. This can complicate understanding how a model decides. More transparency can create security risks. It could reveal weak spots in the model.

Balancing Interpretability and Performance

High-accuracy models, like deep learning systems, can be confusing because of their complexity. Simpler models are easier to understand but might not predict as well. Some AI models are naturally very complicated. Explaining their decisions fully might not work without simplifying important details.

Techniques for Improving Model Interpretability

Methods like feature importance, or SHAP values (SHapley Additive exPlanations) show how certain inputs affect model predictions and improve understanding. Techniques like decision trees, partial dependence plots, and LIME simplify AI decisions.

Chapter 4.12 Principles of Human-Centered Design for Explainable AI

Involving End Users in the Design Process

Explainable AI must focus on what users need. They should fit the user's context and match how users think and decide. Including feedback from end users in building models makes explanations useful and easy to understand.

Designing User Interfaces for AI Interpretability

User interfaces show how AI decisions happen clearly. Pictures and simple stories help people understand AI's logic. With the use of visual tools such as flowcharts, diagrams of feature influence or confidence levels, it can help users understand how AI reaches certain results.

Incorporating User Feedback for Continuous Improvement

Gathering user opinions on explanations helps constant improvements. It creates a cycle of greater trust and model adjustment. Designers rely on assessing user interactions to change and improve how AI systems show their results.

Domain 4: Guidelines for Responsible AI Sample Questions

1. A machine learning specialist is developing an ML model to predict customer churn for a subscription-based service using Amazon SageMaker AI. The specialist is concerned about potential biases in the training data that might affect the model's performance. The specialist must also ensure that the model's predictions are transparent and explainable to stakeholders.

   Which Amazon SageMaker AI capabilities help to meet these requirements?

   a. Amazon SageMaker Ground Truth
   b. Amazon SageMaker JumpStart
   c. Amazon SageMaker Clarify
   d. Amazon SageMaker Data Wrangler

Explanation:
Amazon SageMaker Clarify assists machine learning experts in checking fairness and transparency in their models. This tool spots bias when preparing data, during training, and during predictions. It shows clear explanations of predictions with feature importance scores. It supports ethical AI development.

As for the other options…
- Amazon SageMaker Data Wrangler simplifies data preparation and feature engineering.
- Amazon SageMaker JumpStart provides pre-built ML solutions, workflows, and pre-trained models to simplify starting with machine learning.
- Amazon SageMaker Ground Truth helps produce very good labeled datasets with humans and machines annotating the data. Model fairness or transparency will not be tackled.

2. A financial organization is planning to integrate generative artificial intelligence (generative AI) services into its workflow to improve customer support with natural language processing capabilities. The company is keen on ensuring that its AI models are transparent, fair, and accountable. They are looking for resources to understand the ethical implications and responsible use of AI services.

   Which machine-learning technique would be appropriate for this task?
   a. Amazon Comprehend
   b. AWS Marketplace
   c. AWS AI Service Cards
   d. Amazon Polly

Explanation:

AWS AI Service Cards offer a central hub. These cards describe the ethical points, uses, limits, and best methods for AWS AI services. Users will be able to learn how to use AI responsibly.

As for the other options…

- Amazon Comprehend is a service that studies and understands language. It analyzes text. However, it does not deal with ethical or responsible AI use. Ethical concerns remain outside its scope.
- AWS Marketplace serves as an online shop for software and services on AWS. It does not focus on ethical issues or the responsible use of AI, as it simply provides a platform.
- Amazon Polly only turns text into speech that sounds real. Ethical questions about AI are not part of its functions.

References for Domain 4

Guardrails for Amazon Bedrock
https://aws.amazon.com/bedrock/guardrails/

Amazon SageMaker Clarify
https://aws.amazon.com/sagemaker/clarify/
https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-processing-job-analysis-results.html#clarify-processing-job-analysis-results-pdp

SageMaker Model Monitor
https://aws.amazon.com/blogs/aws/amazon-sagemaker-model-monitor-fully-managed-automatic-monitoring-for-your-machine-learning-models/

Amazon Augmented AI (Amazon A21)
https://aws.amazon.com/augmented-ai/

Amazon SageMaker Model Cards
https://aws.amazon.com/blogs/machine-learning/integrate-amazon-sagemaker-model-cards-with-the-model-registry/

AWS AI Service Cards
https://aws.amazon.com/about-aws/whats-new/2023/11/aws-ai-service-cards/

Others
https://aws.amazon.com/blogs/machine-learning/introducing-aws-ai-service-cards-a-new-resource-to-enhance-transparency-and-advance-responsible-ai/
https://aws.amazon.com/getting-started/hands-on/detect-analyze-compare-faces-rekognition/
https://aws.amazon.com/machine-learning/responsible-machine-learning/transcribe-speech-recognition/
https://tutorialsdojo.com/aws-cheat-sheets-aws-machine-learning-and-ai/
https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-configure-processing-jobs.html
https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html
https://tutorialsdojo.com/amazon-ai-fairness-and-explainability-with-amazon-sagemaker-clarify/

SECURITY, COMPLIANCE, AND GOVERNANCE FOR AI SOLUTIONS

Methods to Secure AI Systems

Governance and Compliance Regulations for AI Systems

Domain 5 emphasizes the key aspects of protecting AI systems, focusing on compliance with legal regulations and establishing strong governance practices within the AWS environment. This review provides a clear overview of the main topics, using AWS resources such as guides, white papers, and best practices. These materials will help prepare you for the AWS AI Practitioner exam.

A. Explain Methods to Secure AI Systems

Chapter 5.1 Identify AWS Services and Features to Secure AI Systems

Securing AI systems requires a variety of strategies, including identity management, data encryption, protection of sensitive information, and the establishment of secure networks. Understanding the shared responsibility model is also crucial.

AWS Identity and Access Management (IAM)

AWS Identity and Access Management (IAM) is an essential service for securing AI systems on AWS. It allows you to manage access to your AWS resources and control the actions that users can perform. IAM ensures only authorized users, applications, and services can interact with your machine learning models, training data, and infrastructure. This helps protect sensitive AI data and ensures compliance with security and privacy standards.

- **Key Components:**
    - **IAM Roles** - Define permissions for AWS services or users. Roles help give temporary access without long-term credentials.
    - **IAM Policies** - Use JSON documents to specify permissions. Attach these policies to users, groups or roles. Decide what actions they are allowed or not allowed to do on specific resources.
    - **Permissions** - Provide detailed access control. Specify actions like s3:GetObject for certain resources such as specific S3 buckets.
    - **Best Practices:**
- **Follow Least Privilege** - Give only the permissions needed to do a task. Limit permissions to what is actually required. Do not give broad permissions.
- **Apply Role-Based Access** - Assign permissions to roles inside your group. This approach simplifies managing and checking access.

- **Audit Permissions Regularly** - Review IAM policies and roles periodically to confirm they follow security best practices. Adjust them when necessary.

Encryption

- **Data Encryption at Rest and in Transit** - Encryption guards data against unauthorized access and breaches. It turns the data into a code. Only people with the correct decryption keys read the information.
- **Key Services:**
    - **AWS Key Management (KMS)** - A managed service that helps to create and control encryption keys easily. KMS works well with various AWS services to simplify encryption tasks.
    - **Amazon S3 Encryption** - provides help for both server-side and client-side encryption. Server-side encryption (SSE) applies various types of keys, like AWS-managed keys (SSE-S3), KMS keys (SSE-KMS) or keys given by customers (SSE-C).
    - **SSL/TLS for Data in Transit** -secures data moving between clients and AWS services. This method shields the data from interception and tampering. Data stays safe and protected
- **Best Practices:**
    - **Use AWS KMS keys or Customer-Managed Keys** - Pick AWS-managed keys for simplicity when security needs are basic. Otherwise, select customer-managed keys when needing more control over key management tasks.
    - **Automate Encryption Processes** - Rely on AWS services' built-in encryption. They encrypt data by themselves when data is stored or moved. Automation saves effort and reduces errors.
    - **Regularly Rotate Encryption Keys** - Change encryption keys often to improve security. Frequent changes lower the risk of key compromise. Rotating keys is very important for protection.

AWS Services that enforces these concepts:

Amazon Macie

Amazon Macie is a fully managed data security and privacy service that utilizes machine learning (ML) to automatically discover, classify, and protect sensitive data within AWS. It is designed to help organizations safeguard sensitive information, such as personally identifiable information (PII), financial records, and intellectual property, thereby reducing the risk of accidental exposure or data breaches. With its powerful ML

capabilities, Macie assists businesses in maintaining compliance with regulations like GDPR, CCPA, and HIPAA by providing automated and scalable data protection solutions.

- **Features:**
    - **Data Classification** - The service automatically finds sensitive data. This includes Personally Identifiable Information (PII) and intellectual property.
    - **Monitoring** - It continuously watches data in Amazon S3 for security issues and compliance problems.
    - **Alerting** - Macie provides alerts and dashboards. These tools inform users about possible data leaks or unauthorized access.
- **Use Cases:**
    - **Compliance** - The service helps meet regulatory requirements. It identifies and protects sensitive information.
    - **Data Protection** - Macie stops accidental or malicious data exposure.
    - **Risk Management** - It assesses privacy risks in data. It also helps apply needed security controls.

AWS PrivateLink

AWS PrivateLink is another AWS service that allows safe, private connections in between VPCs (Virtual Private Clouds) and AWS services. This works even without the internet.

- **Benefits:**
    - **Security** - Traffic stays within the AWS network. This reduces chances of threats from the public internet.
    - **Simplified Design** - No need for complex VPC links or NAT gateways. Access to AWS services becomes easier and safer.
    - **Reliable Performance** - Offers steady network performance by avoiding unpredictable internet issues.
- **Use Cases:**
    - **Secure AWS Access** - Connect to AWS services like Amazon S3 and Amazon EC2 securely from inside your VPC.

- **Third-Party Services** - Create private links to third-party services on AWS without using the public internet.
- **Hybrid Cloud Systems** - Enable secure talks between on-site data centers and AWS environment.

AWS Shared Responsibility Model

AWS Shared Responsibility Model is a foundational concept that defines the division of security and compliance responsibilities between AWS and its customers. In this model, AWS and the customer share responsibility for securing the cloud environment, but the scope of each party's responsibilities differs.

- **Understanding Roles** - The AWS Shared Responsibility Model clearly shows how security tasks are split between AWS and its customers. This clarity helps in managing and securing cloud workloads.
- **AWS's Responsibility – Security OF the Cloud:**
  - **Physical Infrastructure** - AWS takes care of the security for data centers, hardware and networking parts.
  - **Foundation Services** - AWS protects basic services like computing, storage and networking.
  - **Managed Services** - AWS secures the managed services it offers. This includes databases and AI services. AWS is responsible for these services.
- **Customer's Responsibility – Security IN the Cloud:**
  - **Data Security** - Encrypt data, use access controls and classify data for protection.
  - **Application Security** - Use secure coding methods and manage application weaknesses. Configure applications securely and efficiently.
  - **Identity and Access Management** - Control user access and manage IAM roles and policies. Enforce authentication methods to keep data safe.
  - **Network Security** - Configure VPCs, security groups and network ACLs to protect applications and data.
- **Inherited Controls** – Controls that are fully provided by AWS and automatically applied to the customer's environment. For example, Physical and Environmental controls.
- **Shared Controls** – Controls that apply to both the infrastructure layer and customer layers exist in distinct contexts or perspectives. In a shared control model, AWS outlines the requirements for the

infrastructure, while the customer is responsible for implementing their own controls when using AWS services. Examples include:

  - **Patch Management** – AWS handles patching and fixing infrastructure flaws, while customers are responsible to patch their guest OS and applications.
  - **Configuration Management** – AWS configures its infrastructure devices, while customers configure their own guest operating systems, databases, and applications.
  - **Awareness & Training** - AWS trains AWS employees, but a customer must train their own employees.

- **Customer Specific** – Controls for which the customer holds sole responsibility based on the application deployed within AWS services.

Summary of Best Practices for Securing AI Systems on AWS

- **Implement Robust IAM Controls** - IAM roles and policies enforce the least privilege. Regular audits adjust permissions to fit current access needs.
- **Encrypt Data Effectively** - AWS KMS assists in managing encryption keys. Data stays encrypted when stored and sent using AWS services and protocols.
- **Leverage Amazon Macie for Data Protection** - Constant monitoring and classifying keep sensitive data in check. Alerts help spot and deal with possible data breaches quickly.
- **Utilize AWS PrivateLink for Secure Networking** - Private connections to AWS services reduce exposure to the public internet. Network architecture becomes simpler while keeping very high security.
- **Adhere to the Shared Responsibility Model** - Understand your security duties in the cloud completely. Work with AWS to cover all security areas thoroughly.

By adhering to these practices and leveraging AWS's comprehensive suite of security services, you can effectively secure AI systems against a wide range of threats and vulnerabilities.

Chapter 5.2 Understand the Concept of Source Citation and Documenting Data Origins

Good documentation and citation of data sources are vital for keeping data integrity, traceability, and compliance. AWS has many tools and services to help with tracking data lineage, organizing information, and detailed model documentation.

Data Lineage

It is the tracking of where data originates and how it changes through its lifecycle. It shows clearly how data travels from its starting point to its end point. Changes and processing steps along this path are also shown.

- **Importance:**
    - **Traceability** - Ensures that data can be traced back to its original source. This is very important for audits and compliance checks.
    - **Data Integrity** - Maintains the accuracy and consistency of data by tracking transformations and ensuring they are performed correctly.
    - **Impact Analysis** - Understanding possible effects of changes in data sources is crucial. Processing pipelines might affect applications and reports later.
- **Implementation with AWS**
    - **AWS Glue Data Lineage** - AWS Glue offers built-in data lineage features. These features allow tracking data flow through ETL (Extract, Transform, Load) processes. People can understand data transformation at different stages. This provides transparency. It also ensures accountability.

Data Cataloging

Data cataloging organizes and manages metadata. It helps users find and understand data assets easily. Good data catalogs increase data accessibility and usability for different organization members.

- **Key Features:**
    - **Metadata Management**: Stores information about data sources like table definitions, schemas and data formats.
    - **Data Discovery** - Helps users search and find data assets through a unified interface.
    - **Integration** - Connects smoothly with AWS analytics services such as Amazon Athena, Amazon Redshift Spectrum and AWS Glue ETL jobs.

- **Implementation with AWS**
    - **AWS Glue Data Catalog** - this acts as a central location storage for metadata about your organizational datasets. It has a function as an index for the location, schema, and runtime metrics of your data sources. Metadata rests in organized tables, and each table stands for one data store.
- **SageMaker Model Cards**
    - SageMaker Model Cards are detailed documents that explain machine learning models. They reveal the model's purpose and performance. Ethical issues are also covered, which helps build trust, transparency, and accountability.
- **Components:**
    - **Model Details** - These include the model's structure, training data and how it was created.
    - **Intended Use** - Describes where the model should be used. It also mentions any limits and suitable applications.
    - **Performance Metrics** - The model's accuracy, precision and recall are listed here. Other important performance indicators are included as well.
    - **Ethical Considerations** - Any potential biases and fairness issues are discussed. Ethical concerns related to the model are also addressed.

Chapter 5.3 Describe Best Practices for Secure Data Engineering

Secure data engineering handles, processes and stores data to keep it private, accurate and accessible. AWS provides many services and tips to reach excellent data quality, integrity, controlled access, and accuracy.

Assessing Data Quality

Ensuring that data remains accurate, complete, consistent, and timely is essential for secure data engineering. Poor data quality can result in unreliable insights, which may pose security and privacy risks.

- **Importance** - High-quality data is crucial for reliable AI results and reducing risks. Poor data quality leads to inaccurate models and wrong insights. It can also raise security problems.
- **Techniques:**
    - **Data Validation** - Checks if data follows set formats, types and ranges before use.

- **Data Cleaning** - Deletes or fixes wrong, incomplete, or unneeded data to improve quality.
- **Data Normalization** - Standardizes data to a common format or scale. This process allows consistent analysis and processing.

- **Implementation with AWS**
  - **AWS Glue Data Quality** - offers tools to set, run, and check data quality rules. It also helps users monitor and fix data quality problems. Data for AI and analytics becomes accurate and reliable with these tools.

Having Privacy-Enhancing Technologies (PETs)

- **Purpose** - PETs keep sensitive information safe. They protect sensitive information while retaining its usefulness for data analysis and model training. Organizations use PETs to follow privacy regulations and defend against data breaches.
- **Examples:**
  - **Data Anonymization** - Takes away or covers up personally identifiable information (PII) to keep people's privacy safe.
  - **Differential Privacy** - Puts random noises into datasets, so nobody identifies individuals within the data. Despite the changes, it still keeps the data useful.

Data Access Control

Controlling access to data is one of the most crucial aspects of secure data engineering, ensuring that only authorized users and systems can access sensitive data.

- **Strategies:**
  - **Role-Based Access Control (RBAC)** - This handles the permissions based on roles, wherein it simplifies the way to handle access.
  - **Attribute-Based Access Control (ABAC)** - Grants access based on user details and resource attributes. This provides finer control.
- **Tools**
  - **AWS IAM (Identity and Access Management)** - Handles user identities and permissions to AWS resources.
  - **AWS Organizations** - Controls multiple AWS accounts from one place. Uses policies for the entire organization.

- **Best Practices**
    - **Regularly Review and Update Access Policies** - Keep permissions current and matching with current roles and responsibilities.
    - **Principle of Least Privilege (PoLP)** - Apply the principle of least privilege by ensuring that users, services, and applications only have the minimum level of access required to perform their tasks. This limits the potential for unauthorized data exposure.
    - **Use Multi-Factor Authentication (MFA)** - this require extra verification for secure access to important resources. It must be in different factors, such as something you know, something you have, or something you are.

Data Integrity

Maintaining the integrity of data ensures that it remains accurate, consistent, and trustworthy throughout its lifecycle.

- **Ensuring Accuracy and Consistency**
    - **Checksums and Hashing** - Generate hash values, compare them to check data integrity, and detect any changes.
    - **Validation Mechanisms** - Keep data accurate. Maintain data consistency throughout its lifecycle.
- **Implementation with AWS**
    - **Amazon S3 Object Lock -** Protects objects in S3. Prevents deletion or overwriting for a set time, ensuring data remains unchanged.
    - **AWS Backup** - Automates data backup across AWS services. Guarantees data consistency and recovery.

Chapter 5.4 Understand Security and Privacy Considerations for AI Systems

Ensuring AI systems are safe and private involves safeguarding applications. It includes detecting and mitigating dangers. It also means handling weaknesses. Securing infrastructure is essential. Addressing risks such as prompt injection is vital. AWS provides a complete set of tools and services to handle these issues effectively.

Application Security

Application security is a vital part of ensuring the overall security of AI systems. To achieve this, several measures can be taken:

- **Secure Coding Steps** - Developers should follow best practices when writing code. They need to check inputs, store data safely, and use secure ways to share information. AWS offers tools like the AWS Well-Architected Framework to help with this. It shares the best ideas for creating and running safe, stron,g and effective cloud systems. AWS helps improve security.
- **Regular security assessments** - help find vulnerabilities and weaknesses in the application, where AWS offers services like AWS CodePipeline. This service adds extra security to the pipelines of Continuous Integration/Continuous Deployment (CI/CD). It allows automated security checks and testing.
- **Vulnerability scanning** - finds and fixes issues in the application. AWS provides services like AWS CodePipeline, where this tool works with other tools to scan for vulnerabilities.

AWS Tools like AWS CodePipeline provide a comprehensive CI/CD platform that integrates security into every stage of the pipeline, enabling developers to deliver secure, reliable, and high-quality applications.

Threat Detection Services

Threat detection is another critical security component for classifying and responding to potential security threats in AI systems. AWS provides several services for threat detection:

- **Amazon GuardDuty** - Amazon GuardDuty is a threat detection service that finds harmful actions and threats by checking for them all the time. It watches for things like unauthorized access, malware and other dangers to security. It offers instant alerts and suggestions for fixing problems.
- **AWS Security Hub** - AWS Security Hub offers central security management. It collects security findings from different sources. These sources include Amazon GuardDuty and Amazon Inspector. Other AWS services also contribute findings.

Vulnerability Management

Vulnerability management means finding and fixing weak spots in AI systems. AWS offers many services and processes for this purpose:

- **Regular Patching** - Patching regularly helps apply security fixes to address known issues. AWS offers tools like AWS Systems Manager Patch Manager, automating patch updates for Amazon EC2 and other AWS tools.
- **Vulnerability scanning** - Scanning identifies vulnerabilities in AI systems. AWS offers tools like Amazon Inspector, which scans and assesses vulnerabilities automatically.
- **Remediation** - Taking action to fix found vulnerabilities is remediation. AWS provides tools like AWS Systems Manager for automatic remediation action.
- **Implementation with AWS**
    - AWS Services such as Amazon Inspector and AWS Systems Manager Patch Manager offer wide-ranging tools for managing vulnerabilities. These tools help developers find and fix weaknesses in AI systems. This enables developers to identify issues and address problems.

Infrastructure Protection

- Infrastructure protection is very important for keeping AI systems safe. AWS offers many ways and services for this protection.
    - **Network segmentation** - Network segmentation separates AI systems and data from other networks and systems. AWS uses services like Amazon VPC. It helps with network separation and isolation.
    - **Firewalls** - This helps control incoming and outgoing network traffic. This is based on predefined security rules. They rely on fixed security rules for this purpose. AWS provides services like AWS WAF (Web Application Firewall). This tool guards web applications against common web threats.
    - **Intrusion detection systems** - watch network traffic for unauthorized access or harmful activities. AWS uses services like AWS Shield, which offers DDoS protection and intrusion detection.

AWS Services such as AWS Shield and AWS WAF offer strong tools for keeping the infrastructure safe. These services help developers guard AI systems against unwanted access and harmful actions. Overall, developers have good resources to protect their AI systems.

Prompt Injection

Prompt injection is a type of attack that involves changing input prompts to control how an AI model behaves. To mitigate prompt injection attacks, here are the following techniques:

- **Input validation** - involves checking user input to be sure it follows expected formats and patterns. AWS offers services like AWS Lambda for input validation and cleaning.
- **Sanitation** - involves taking out or escaping harmful characters from user input. AWS offers services like AWS API Gateway for input cleaning and checking.
- **Monitoring** - involves watching AI model behavior and input prompts for clues of prompt injection attacks. AWS provides tools like Amazon CloudWatch for tracking and logging AI model actions.

Encryption at Rest and in Transit

- Encryption is very important for data safety in AI systems. AWS offers many services for encrypting stored and sent data:
    - **Encryption at rest** - This process covers securing data within databases, storage areas and other systems. Amazon S3, a service from AWS, provides server-side encryption for stored data.
    - **Encryption in transit** - This method focuses on securing data moved between systems and applications. AWS Certificate Manager, another service from AWS, provides encryption through SSL.

B. Recognize Governance and Compliance Regulations for AI Systems

Chapter 5.5 Identify Regulatory Compliance Standards for AI Systems

Regulatory compliance is very important when using AI systems. Adhering to these standards not only reduces the risk of penalties, but having a chance to expand the target market of an organization internationally, may it be from Asia, Europe, America, and many more. Hence, many rules exist that companies must follow depending on their specific cases, such as:

International Organization for Standardization (ISO)

The International Organization for Standardization (ISO) plays a crucial role in developing global standards that help organizations ensure data security, quality, and the responsible use of emerging technologies, including Artificial Intelligence (AI). ISO standards provide frameworks and guidelines for best practices, risk management, and compliance, assisting organizations in maintaining high levels of security, privacy, and trust.

- **ISO/IEC 27001** is important for keeping data safe. It focuses on information security management systems. This standard protects data confidentiality, integrity and availability. Protecting these three aspects is very crucial.
- **ISO/IEC 2382:2015** - gives a structured list of IT and AI terms. It has definitions to help people communicate clearly. Using the same words in IT and AI makes understanding easier.
- **ISO/IEC TR 24028:2020** - provides guidelines for using artificial intelligence (AI) and machine learning (ML) systems. The report focuses on risk management. Building trustworthiness in AI is also a key topic. This includes data governance, transparency and explainability. Accountability is another important part. It supports responsible use of these technologies.
- **ISO/IEC 42001** - offers guidance for organizations to handle AI systems in an ethical and trustworthy way, respecting all necessary rules.
- **Implementation with AWS** - AWS offers certification guidelines towards ISO/IEC 27001 certification. This certificate shows that its cloud services follow strict international security rules. Organizations using AWS probably find it easier to follow safety standards. AWS's certified infrastructure helps businesses reach and keep ISO compliance easily.

System and Organization Controls (SOC)

System and Organization Controls (SOC) are standards used to evaluate the effectiveness of controls over systems, data, and services. SOC reports focus on how companies protect and manage data, particularly with third-party providers or cloud platforms. They help organizations ensure compliance with regulations and build trust with customers and stakeholders. These guidelines help companies assess and demonstrate how well internal controls function.

- **SOC 1** - focuses on financial reporting controls.
- **SOC 2** - is another SOC concept where it examines non-financial controls like security, availability. This also processes integrity, confidentiality and privacy.
- **SOC 3** - resembles SOC 2, but this SOC targets a broader audience. This offers a summary of the SOC 2 report with no detailed descriptions.
- **Focus** - these controls are related to Security, Availability, Processing Integrity, Confidentiality and Privacy.
    - SOC reports review different parts of an organization's controls. These reports analyze if the controls are good enough for security and operational effectiveness. Controls are very important for keeping AI systems' integrity and reliability. This is especially vital when dealing with sensitive information and maintaining reliable performance.

- **Implementation with AWS** - AWS offers detailed SOC reports describing the security controls for its services. These reports are useful for organizations wanting to know how AWS handles security, availability and processing integrity. They help organizations with their own compliance efforts.

Algorithm Accountability Act

Algorithm accountability laws exist to guarantee that AI systems work openly and justly. They reduce biases and support accountability. These rules require organizations to share how their algorithms reach decisions. Also, they need to check for possible biases, and implement measures to any problems found.

- **Examples:EU AI Act, USA Algorithmic Accountability Act**
    - **EU AI Act** - The European Union proposes a set of strict rules to control AI systems. It sorts AI uses based on their risk levels. High-risk AI systems face stricter requirements like being transparent, having human oversight, and good data governance.
    - **USA Algorithmic Accountability Act** - The United States has a proposed law for companies to review their automated decision systems via impact assessments. The law wants to find and fix biases, support fairness and encourage clear operations in these systems. Fairness and bias are key in this act.
- **Implementation with AWS - AWS AI and Compliance**
    - AWS provides tools to help businesses meet algorithm rules. These tools include checking AI models, promoting clarity, and setting up methods to handle fairness and bias mitigation strategies in AI workflows.

Chapter 5.6 Identify AWS Services and Features to Assist with Governance and Regulation Compliance

AWS provides a strong set of services to help organizations handle governance and compliance well. These services include checking configurations, running automated audits, monitoring continuously and offering detailed reports. This allows organizations to adhere to various regulatory standards and internal policies seamlessly.

AWS Config

AWS Config is a tool that always watches and keeps track of your AWS resource configurations. It lets you automatically check if these settings match your preferred setups. This helps you follow policies inside your

company and from outside authorities by giving a clear picture of the setup history and connections between resources.

- **Features**:
  - **Continuous Monitoring** - AWS Config keeps continuously monitors and records configurations of supported AWS resources all the time.
  - **Compliance Auditing** - It reviews resource settings using pre-existing rules to detect any misconfiguration or non-compliance.
  - **Resource Tracking** - AWS Config can track changes and connections between resources. This helps in understanding effects and solving problems.
  - **AWS Resource: AWS Config Documentation** - A comprehensive documentation is available to help users set up AWS Config. These guides explain how to define rules and read compliance results. This helps people use the service well.

Amazon Inspector

Amazon Inspector offers a way to check applications automatically for security problems and mistakes in AWS. It regularly looks at applications to find risks or mistakes from guidelines, showing clear reports with important issues listed first.

- **Use Cases**:
  - **Vulnerability Management** - Find and fix security problems in applications and AWS resources.
  - **Compliance Checks** - Check if applications follow industry rules and practices by spotting mistakes.
  - **AWS Resource**: Amazon Inspector

AWS Audit Manager

AWS Audit Manager helps users audit their AWS usage all the time to meet rules and industry standards more easily. It collects needed evidence automatically for audits, lowering manual work and keeping things accurate.

- **Features: Pre-built frameworks, custom controls, and audit-ready reports.**
  - **Pre-built Frameworks** - Offers ready-made frameworks for regular compliance standards, which help with fast setup and checks.
  - **Custom Controls** - Lets users create custom controls that fit their specific compliance needs.

- **Audit-ready Reports** - Produces detailed reports prepared for audits, making the audit process smoother.

AWS Artifact

AWS Artifact is a self-service portal. It gives people access to AWS compliance reports and specific agreements. This tool helps very much in preparing for audits and checking compliance by providing the latest compliance documents whenever needed.

- **Audit Preparation** - Find necessary compliance reports. These reports show how rules and standards are really followed during audits.
- **Compliance Verification** - Obtain current AWS compliance documents to support internal compliance work.

AWS CloudTrail

AWS CloudTrail is another AWS service. This enables governance, compliance, operational auditing, and risk auditing of your AWS account. It captures AWS API calls and related events, giving a complete history of user actions and resource changes.

- **Features**:
  - **Event History** - Holds a full record of AWS API calls made in your account, including actions through the AWS Management Console, AWS SDKs, command-line tools and other AWS services.
  - **Real-time Monitoring** - Works with Amazon CloudWatch Logs to watch and alert based on specific API activities right away.
  - **Security Analysis** - Helps in security checks by offering logs to find unauthorized access or unusual actions.

AWS Trusted Advisor

AWS Trusted Advisor is an online tool that helps you lower costs, improve performance, and improve security in your AWS setup. This tool gives real-time advice, where you get guidance on using your resources with AWS best practices. AWS Trusted Advisor checks your Amazon EC2 computing-consumption history, where it calculates a good number of Partial Upfront Reserved Instances.

- **Categories**:
  - **Cost Optimization** - Finds chances to lower expenses by cutting unused resources.
  - **Performance** - Offers ideas to really improve how the system runs.

- **Security** - Points out security gaps and suggests better security steps. Security needs to be strong.
- **Fault Tolerance** - Recommends ways to increase system availability. Helps systems keep running.
- **Service Limits** - Checks service limits and warns when getting close to them. This helps avoid going over limits.

Chapter 5.7 Describe Data Governance Strategies

Effective data governance ensures that your data is managed properly and being guided by a lifecycle. It involves maintaining data integrity, security, and compliance with regulatory requirements despite the location of the operation. Fortunately, AWS provides various tools and strategies to implement comprehensive data governance.

Data Lifecycles

- **Phases: Creation, storage, usage, archival, and deletion** - Managing data means creating rules for each phase to keep data integrity and compliance. AWS has tools for managing data from the creation to the deletion effectively.
- **Management** - Management requires creating rules for each phase to keep data safe and in line with laws.

Logging

- **Importance: Maintain records of data access and modifications for audit trails.**
  - Logging is crucial for tracking data access and modifications, providing an audit trail for compliance and security purposes.
- **AWS Services**:
  - **AWS CloudTrail:** Logs API calls and account activity.
  - **Amazon CloudWatch Logs:** Monitors and stores log files for applications and systems.

Data Residency

Ensuring data is stored in specific geographic locations to comply with local laws. Data residency involves keeping data in specific areas to comply with local data protection regulations.

Monitoring and Observation

- **Techniques - Continuous monitoring of data access patterns, and anomaly detection.**
    - Monitoring and observation involve tracking data access patterns all the time. This helps find unusual behavior and prevent unauthorized access and ensure compliance.
- **AWS Services**:
    - **Amazon GuardDuty** - This service finds threats by watching for malicious activities all the time.
    - **AWS Security Hub** - It brings together and organizes security alerts from many AWS services.

Data Retention

- **Policies**: Define how long data is retained and the methods for secure disposal.
    - Data retention policies tell how long to keep data and explain how to dispose of it safely when it is not needed anymore.
- **AWS Services**:
    - **AWS Backup** - Centralized service for automatic and organized backups in AWS services.
    - **Amazon S3 Lifecycle Policies** - Automatically shift objects to other storage types or remove them following set rules.

Chapter 5.8 Describe Processes to Follow Governance Protocols

Following governance rules requires creating clear rules, scheduling regular reviews, using smart review techniques, applying governance frameworks and training team members.

Policies

- **Definition** - Storing data in specific geographic locations and places to follow local laws. Defining policies means setting rules and guidelines for managing data and resources. These policies keep security and follow regulations.
- **Implementation** - Enforce policies using AWS services like IAM, AWS Config, and AWS Organizations.
- **AWS Services for Policy Enforcement:**
    - **AWS Identity and Access Management (IAM)** - Controls who can access and what permissions they have.
    - **AWS Config** - Checks if resource settings match the policies.

- **AWS Organizations** - Brings together management for many AWS accounts in one place.

Review Cadence

- **Frequency** - Schedule reviews regularly, like quarterly or annually, so policies stay current and effective.
- Regularly reviewing policies ensures they remain relevant. This assists in having an effective addressing of current security and compliance requirements.
- **Activities** - Conduct policy audits, perform compliance checks and update based on new rules or changes in the organization.

Review Strategies

- **Approaches**: Automated tools for constant checks, manual audits for detailed evaluations.
- Using both automated and manual review strategies gives complete control.
- **AWS Tools:**
  - **AWS Config Rules** - Keep rules in line with internal and external standards.
  - **AWS Security Hub Insights** - Offers automated checks and collects security data.

Governance Frameworks

Generative AI Security Scoping Matrix

- **Purpose -** Assess and manage security risks specific to generative AI applications.
- **Components -** Risk identification, mitigation strategies, compliance mapping.
- **AWS Resource -** AWS Security Frameworks

Transparency Standards

- **Definition**: Ensure clarity in AI operations, decision-making processes, and data usage.
- **Implementation**: Document AI model behaviors, data sources, and decision criteria.

Team Training Requirements

- **Importance**: Ensure all team members understand security protocols, compliance standards, and governance policies.
- **Resources**: AWS Training and Certification, internal workshops, and continuous education programs.
- **AWS Resource**: AWS Training and Certification

Domain 5: Security, Compliance, and Governance for AI Solutions Sample Questions

1. A healthcare organization is in the process of migrating its patient management system to AWS. As part of their compliance and due diligence efforts, they must ensure that all AWS services and solutions they plan to use comply with various healthcare regulations, such as HIPAA and HITRUST. The organization also works with independent software vendors (ISVs) who need to verify compliance with the AWS services they intend to integrate. The organization must obtain security and compliance documentation from AWS to meet these regulatory requirements and demonstrate compliance to auditors.

   Which AWS service can this organization access the necessary security and compliance reports?

   a. AWS Artifact

   b. AWS Cloud Trail

   c. AWS Security Hub

   d. Amazon Macie

Explanation:

AWS Artifact is a comprehensive and central resource for accessing AWS's compliance-related information. It offers on-demand access to AWS's security and compliance reports, including audit artifacts such as Service Organization Control (SOC) reports, Payment Card Industry (PCI) reports, and certifications from the International Organization for Standardization (ISO). AWS Artifact is designed to help customers manage their compliance posture effectively by offering a self-service portal where they can retrieve compliance documentation. This service allows customers to meet regulatory requirements, such as HIPAA, HITRUST, and other industry-specific standards, by ensuring they have the necessary documentation to demonstrate compliance to auditors and regulatory bodies.

Hence, **the correct answer is: AWS Artifact.**

- Amazon Macie is incorrect because this is just a data security service that uses machine learning to automatically discover, classify, and protect sensitive data within your AWS environment. It is particularly focused on identifying and alerting users about personally identifiable information (PII) and other sensitive data stored in Amazon S3.

- AWS CloudTrail is incorrect because this service is primarily for governance, compliance, and operational and risk auditing of your AWS account. It records AWS API calls and events for your account, providing detailed logs of user activity. Its primary focus is to provide visibility into API activity rather than to supply the compliance documentation needed for regulatory audits.
- AWS Security Hub is incorrect because this security service only provides a comprehensive view of your high-priority security alerts and compliance status across AWS accounts.

2. A cloud security engineer is tasked with securing machine learning (ML) workloads in an AWS environment. The engineer needs to ensure that only specific applications and services can access Amazon SageMaker AI and Amazon RDS resources. These applications require controlled and limited access to these services to maintain security and minimize the risk of unauthorized access.

Which AWS service or feature can the engineer use to grant and manage these permissions effectively?

a. **AWS Identity and Access Management (IAM)**

b. AWS Security Token Service (STS)

c. VPC Endpoint Policy

d. AWS Secrets Manager

Explanation:

AWS Identity and Access Management (IAM) is a service that allows you to control the use of AWS services and resources. IAM lets you set up and handle AWS users and groups, as well as define permissions to allow or prohibit access to AWS resources. IAM allows you to build fine-grained access control by defining roles and policies that determine who may access specific services and what activities they can do.

AWS Identity and Access Management (IAM) allows administrators to define and enforce permissions for AWS services and resources. IAM is suitable for managing access to Amazon SageMaker AI and Amazon RDS by creating specific policies that define which applications or services can access these resources. By utilizing IAM roles and policies, the engineer can ensure that only authorized applications and services have the necessary permissions.

IAM is the correct service to manage and enforce permissions for applications and services across AWS, including Amazon SageMaker and Amazon RDS.

Hence, **the correct answer is: AWS Identity and Access Management (IAM).**

- AWS Secrets Manager is incorrect because it is primarily used for storing and managing access to sensitive information such as database credentials and API keys. Although it enhances security by rotating and managing secrets, it does not control access permissions to AWS resources like Amazon SageMaker AI and Amazon RDS.
- WS Security Token Service (STS) is incorrect because it simply provides temporary credentials for accessing AWS services. However, it is not a comprehensive solution for managing long-term permissions for specific applications and services.
- VPC Endpoint Policy is incorrect because it is only used to control access to AWS services over a VPC endpoint. While it provides network-level control, it is limited in scope and does not offer the comprehensive permissions management needed for managing access across different services like Amazon SageMaker AI and Amazon RDS.

References for Domain 5

Identity and Access Management (IAM)
https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html

Amazon Macie
https://aws.amazon.com/macie/

AWS PrivateLink
https://docs.aws.amazon.com/vpc/latest/privatelink/what-is-privatelink.html

AWS Shared Responsibility Model
https://aws.amazon.com/compliance/shared-responsibility-model/

Implementation with AWS
https://docs.aws.amazon.com/glue/latest/dg/catalog-and-crawler.html

SageMaker Model Cards
https://docs.aws.amazon.com/sagemaker/latest/dg/model-cards.html

Others

https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/security-perspective-compliance-and-assurance-of-aiml-systems.html

https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html

https://docs.aws.amazon.com/IAM/latest/UserGuide/best-practices.html

https://tutorialsdojo.com/aws-identity-and-access-management-iam/

https://aws.amazon.com/artifact/faq/

https://docs.aws.amazon.com/artifact/latest/ug/what-is-aws-artifact.html

https://tutorialsdojo.com/aws-artifact/

ABOUT THE AUTHORS

Jon Bonso (10x AWS Certified)

Born and raised in the Philippines, Jon is the Co-Founder of **Tutorials Dojo**. Now based in Sydney, Australia, he has over a decade of diversified experience in Banking, Financial Services, and Telecommunications. He's 10x AWS Certified and has worked with various cloud services such as Google Cloud, and Microsoft Azure. Jon is passionate about what he does and dedicates a lot of time creating educational courses. He has given IT seminars to different universities in the Philippines for free and has launched educational websites using his own money and without any external funding.

Kayne Uriel Rodrigo (AWS CCP Certified)

An I.T. intern at Tutorials Dojo and a 4th-year Computer Science student at **Pamantasan ng Lungsod ng Maynila** *(University of the City of Manila)*. AWS CCP Certified, ISC2 CC Certified, and a **DAP Project SPARTA** Data Science Track graduate, he has two years of project-based experience in AI, cybersecurity, and software engineering. As a co-author of the AIF-C01 Reviewer for Tutorials Dojo, Kayne leverages his expertise to provide comprehensive insights for professionals. Active as a student tech-lead in **AWS Cloud Club Haribon** and won the **2024 AWS Innovation Cup** Hackathon event.

Samantha Vivien L. Servo

Currently a 4th year Computer Science student in the **Pamantasan ng Lungsod ng Maynila** and an IT intern in **Tutorials Dojo**. With her team, she has won the **2024 AWS Innovation Cup Hackathon**. She is actively involved in campus organizations such as GDSC PLM and AWS Cloud Clubs Haribon, and she's passionate about the convergence of medicine and technology, particularly data science. Through her brand, ***"Beyond the Vinculum,"*** Samantha is committed to pushing the boundaries and shaping the future of these ever-evolving fields.