AWS Certified AI Practitioner

DAY 1

Michael J. Shannon
CISSP, CCSP, CCSK,
CompTIA SecurityX,
AWS Security Specialty,
ITIL4 MP

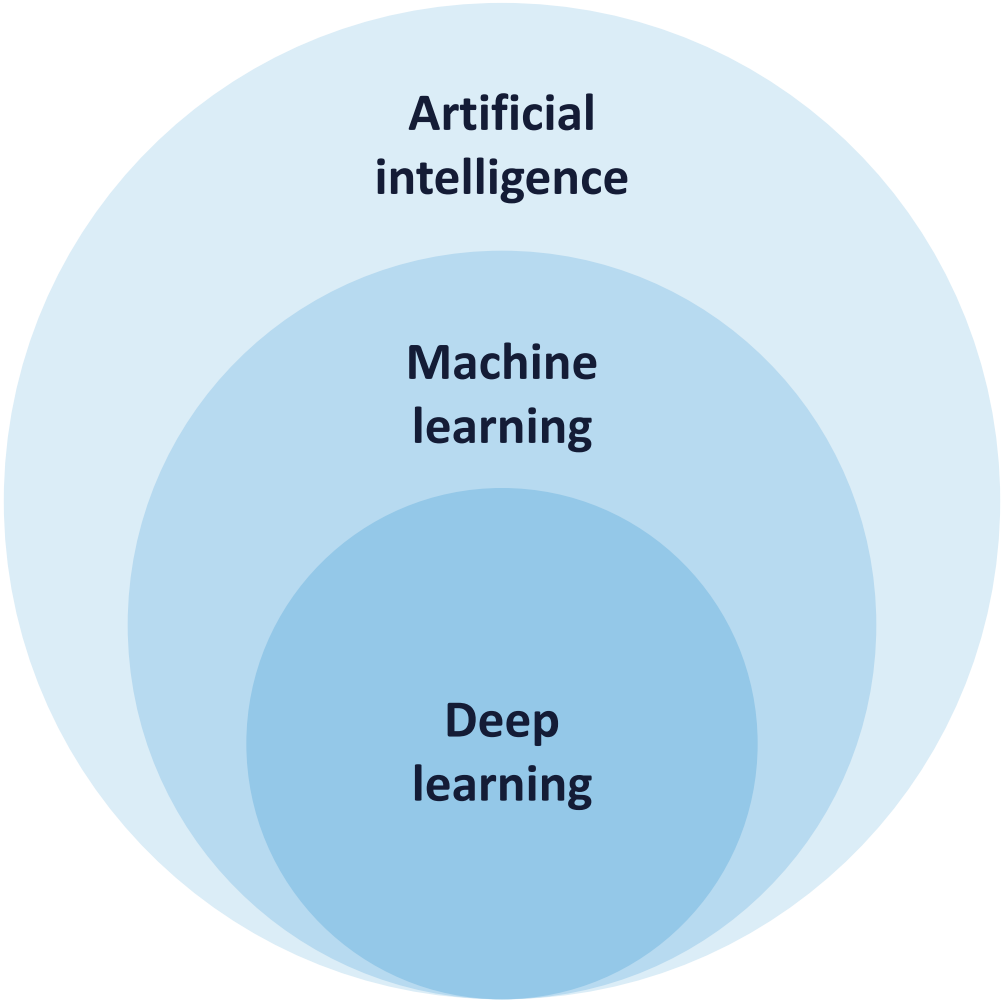
Class will begin at 10:00 am
Central Standard Time

Basic AI Concepts and Terminologies

Objectives

- Define AI, ML, and deep learning
- Define basic related AI terms
- Explore large language models (LLMs)
- Compare the differences between AI, ML, and deep learning
- Examine inference types and data types in AI models
- Compare machine learning methods

Artificial Intelligence (AI)



Artificial Intelligence (AI)

- According to Amazon Web Services (AWS), **artificial intelligence (AI)** is a technology that simulates human intelligence to solve problems
- AI can make data-based predictions and recognize images, and write text
- It fundamentally makes software smarter, allowing for customized user interactions and complex problem-solving

Artificial Intelligence (AI)

AI can be applied across a wide range of domains, enhancing productivity, accuracy, and user experience

- **Natural language processing (NLP)**: Understand, interpret, and produce human language, enabling applications like chatbots, language translation, and sentiment analysis
- **Computer vision**: Deduce and make decisions based on visual inputs, like images and videos used for facial recognition, autonomous vehicles, and medical imaging

Artificial Intelligence (AI)

- **Speech recognition**: Convert spoken language into text and understand spoken commands, which is essential for virtual assistants and voice-controlled devices
- **Robotics**: Enables robots to perform tasks autonomously or semi-autonomously, from manufacturing to household chores
- **Expert systems**: Imitate the decision-making abilities of a human expert, delivering solutions in fields like medical diagnosis and financial forecasting

Artificial Intelligence (AI)

- **Predictive analytics**: Examine data to predict future trends and behaviors, which is valuable in areas like marketing, finance, and healthcare
- **Automation**: Automate repetitive tasks, improving efficiency and reducing human error in various industries
- **Machine learning (ML)**: AI systems can learn from data and improve over time without being explicitly programmed

Machine Learning (ML)

- Machine learning (ML) is a type of artificial intelligence that performs data analysis tasks without explicit instructions or programming
- Machine learning technology can process large quantities of historical data, identify patterns, and predict new relationships between previously unknown data

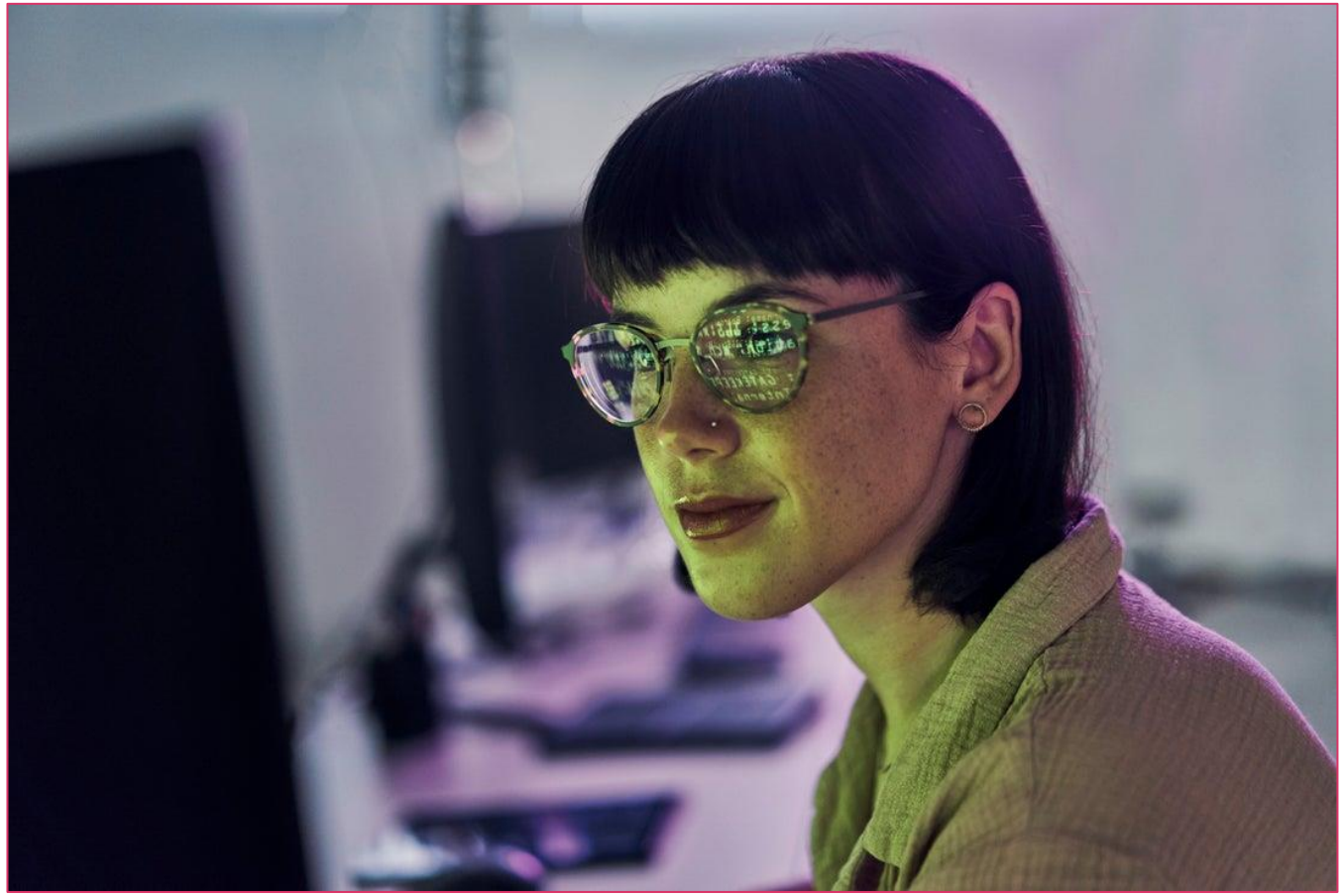
Machine Learning (ML)

- ML allows businesses to automate and optimize data collection, classification, and analysis
- This can result in improved decision-making, automation of routine tasks, enhanced customer experiences, and proactive resource management

Deep Learning

- AWS defines **deep learning** as a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain
- **Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions**
- This technology is used to automate tasks that typically require human intelligence, such as describing images or transcribing a sound file into text

Deep Learning



- Deep learning is a critical component of many AI applications, including digital assistants, voice-activated television remotes, fraud detection, and automatic facial recognition
- It also plays a significant role in emerging technologies like self-driving cars and virtual reality

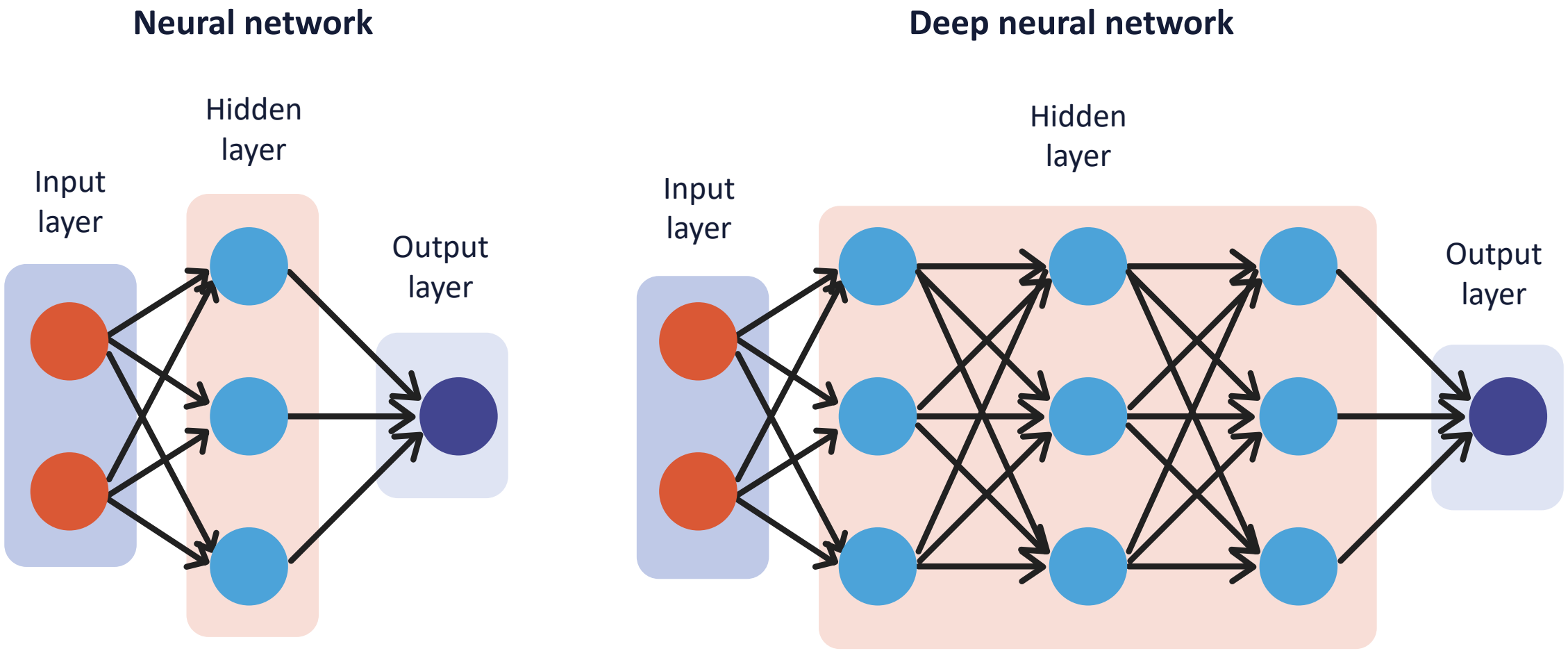
Neural Networks

- **Neural networks** are a technique in artificial intelligence that teaches computers to process data in a way that is motivated by the activities of the human brain
- It is a type of ML process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain

Neural Networks

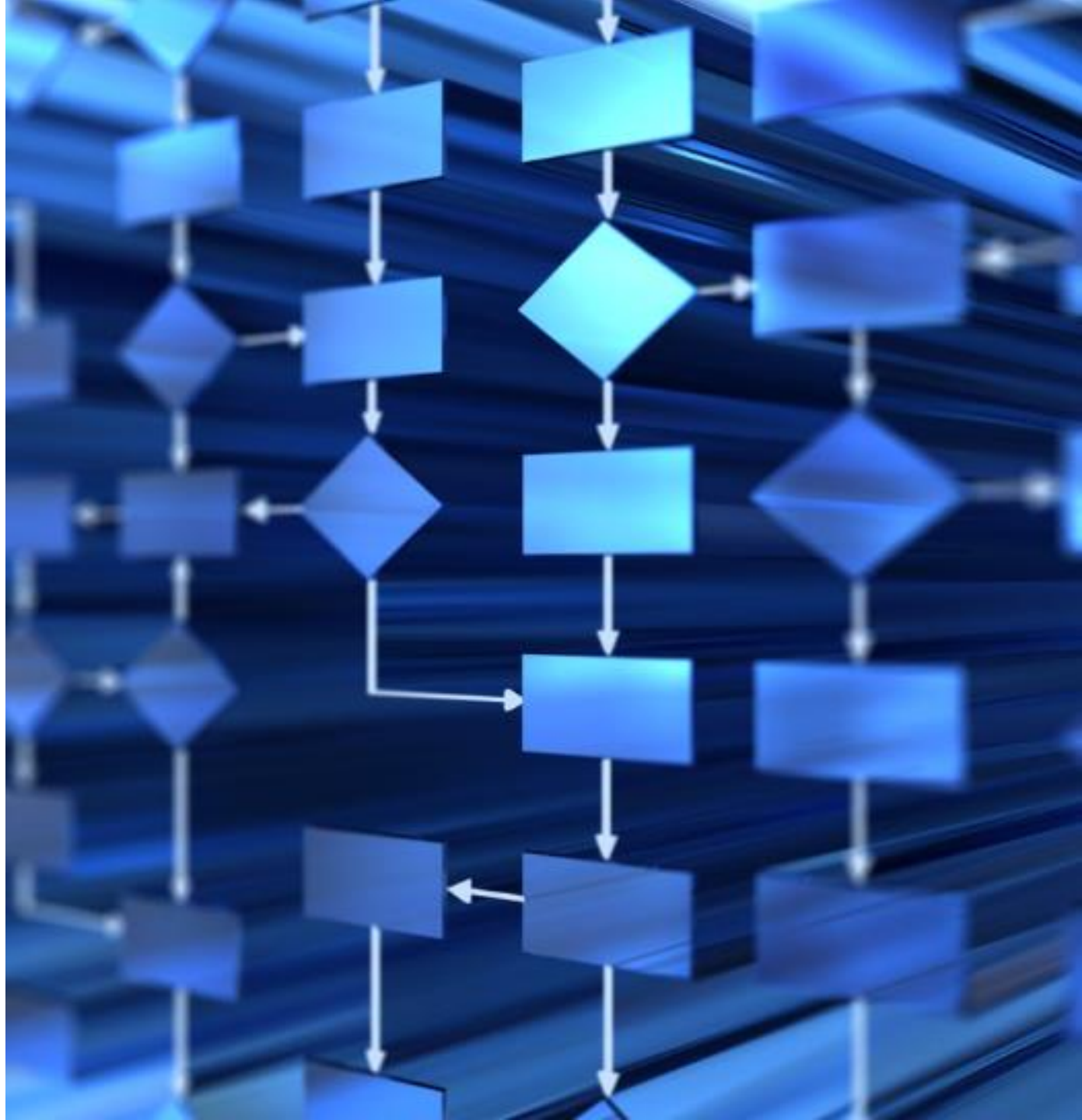
- This creates an adaptive system that allows computers to learn from their mistakes and improve continuously
- Neural networks are designed to solve complex problems, such as summarizing documents or recognizing faces, with greater accuracy

Neural Networks



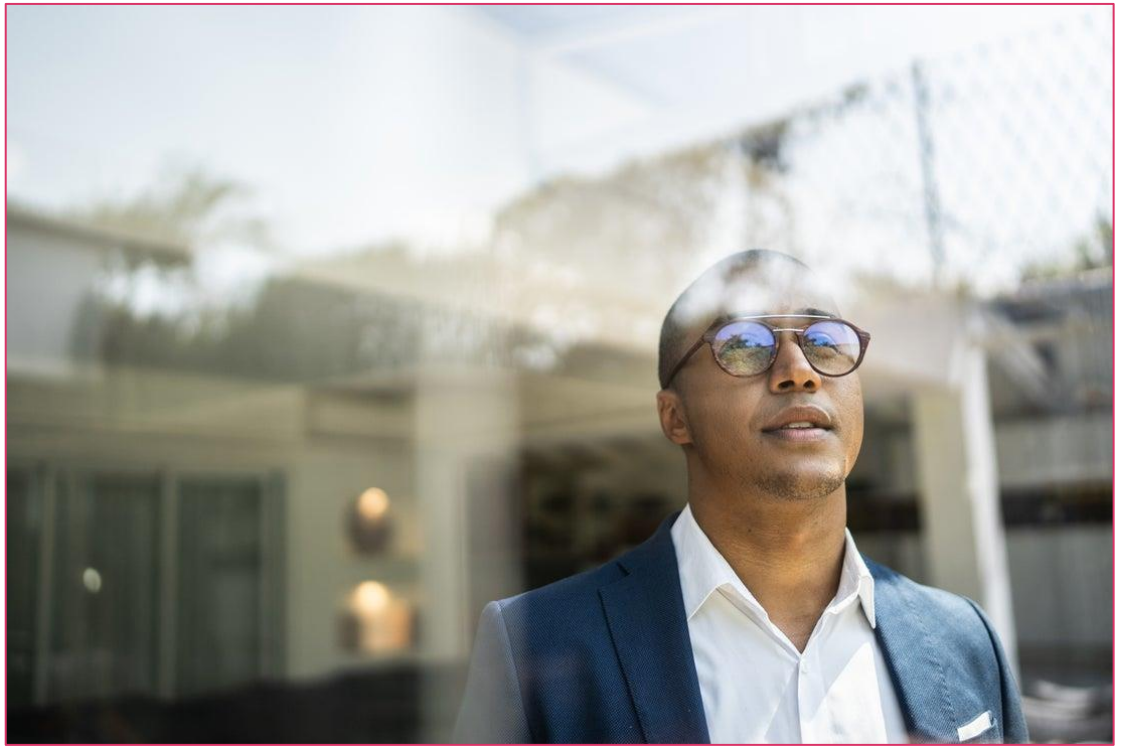
What Is an AI Model?

- An **AI model** is a mathematical framework or algorithm designed to perform specific tasks by learning patterns from data
- These models can range from simple linear regressions to complex neural networks
- They are trained using large datasets to recognize patterns, make predictions, or generate content
- The effectiveness of an AI model depends on the quality of the data it is trained on and the complexity of the task it is designed to perform



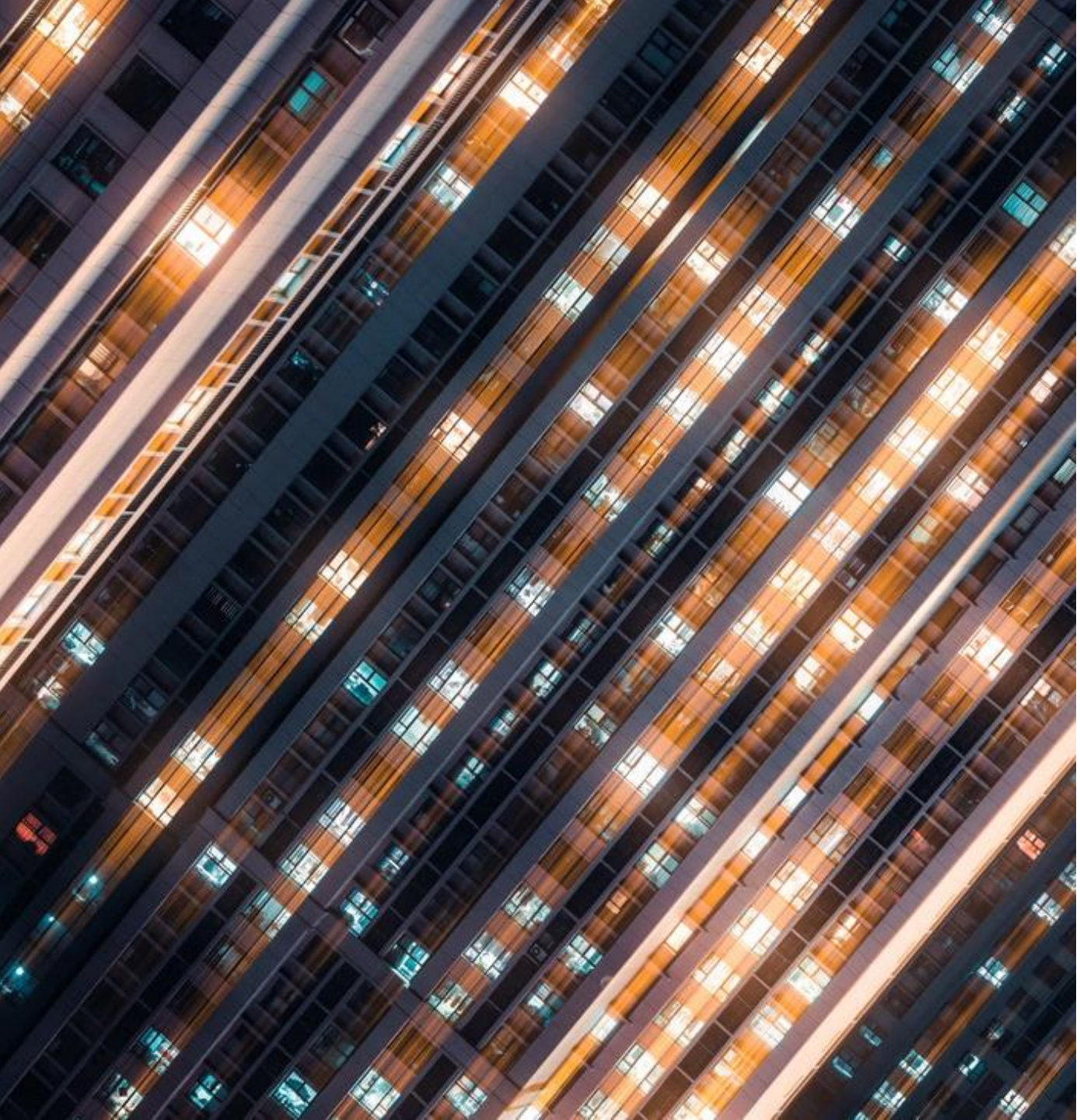
Computer Vision

- AWS defines **computer vision** as the ability of computers to extract information and insights from images and videos
- This technology uses deep learning techniques to enable computers to comprehend images like humans
- Computer vision has several applications, such as content moderation, facial recognition, and image classification



Natural Language Processing (NLP)

- Natural language processing (NLP) is an ML technology that enables computers to interpret, manipulate, and comprehend human language
- This technology allows organizations to process large volumes of voice and text data from various communication channels
- Common NLP communication channels would include emails, text messages, social media newsfeeds, video, audio, and more



AI Algorithm

- AWS defines an **AI algorithm** as a set of rules or instructions given to an AI model to help it learn on its own
- These algorithms enable machines to simulate human-like intelligence and perform complex tasks autonomously by processing data, extracting meaningful insights, and making informed decisions

Machine Learning Training

- **Training** is the process of teaching a ML model to make accurate predictions or decisions based on data
- This involves feeding the model large amounts of data and letting it learn patterns and relationships within that data
- The goal is to adjust the model's parameters so that it can generalize well to new, unseen data

Other ML Terms

Bias

Refers to the systematic errors in an ML model that results in unfair or inaccurate outcomes

Fairness

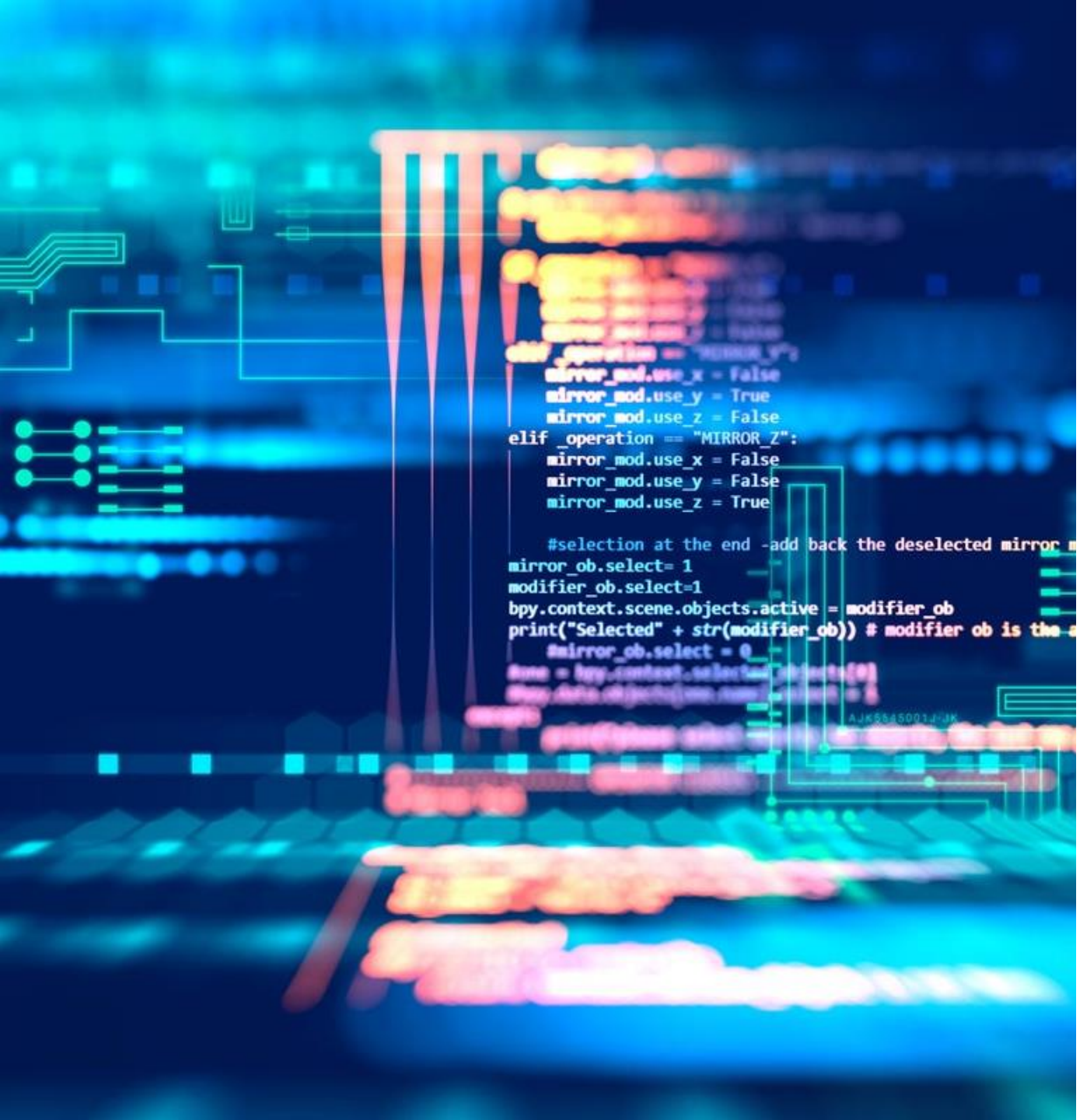
Refers to the practice of ensuring that ML models operate without bias

Leads to equitable outcomes for all individuals, regardless of race, gender, age, or other protected characteristics

Fit

Refers to the process of training a model on a dataset

During this process, the model learns the patterns and relationships within the data to make accurate predictions or decisions



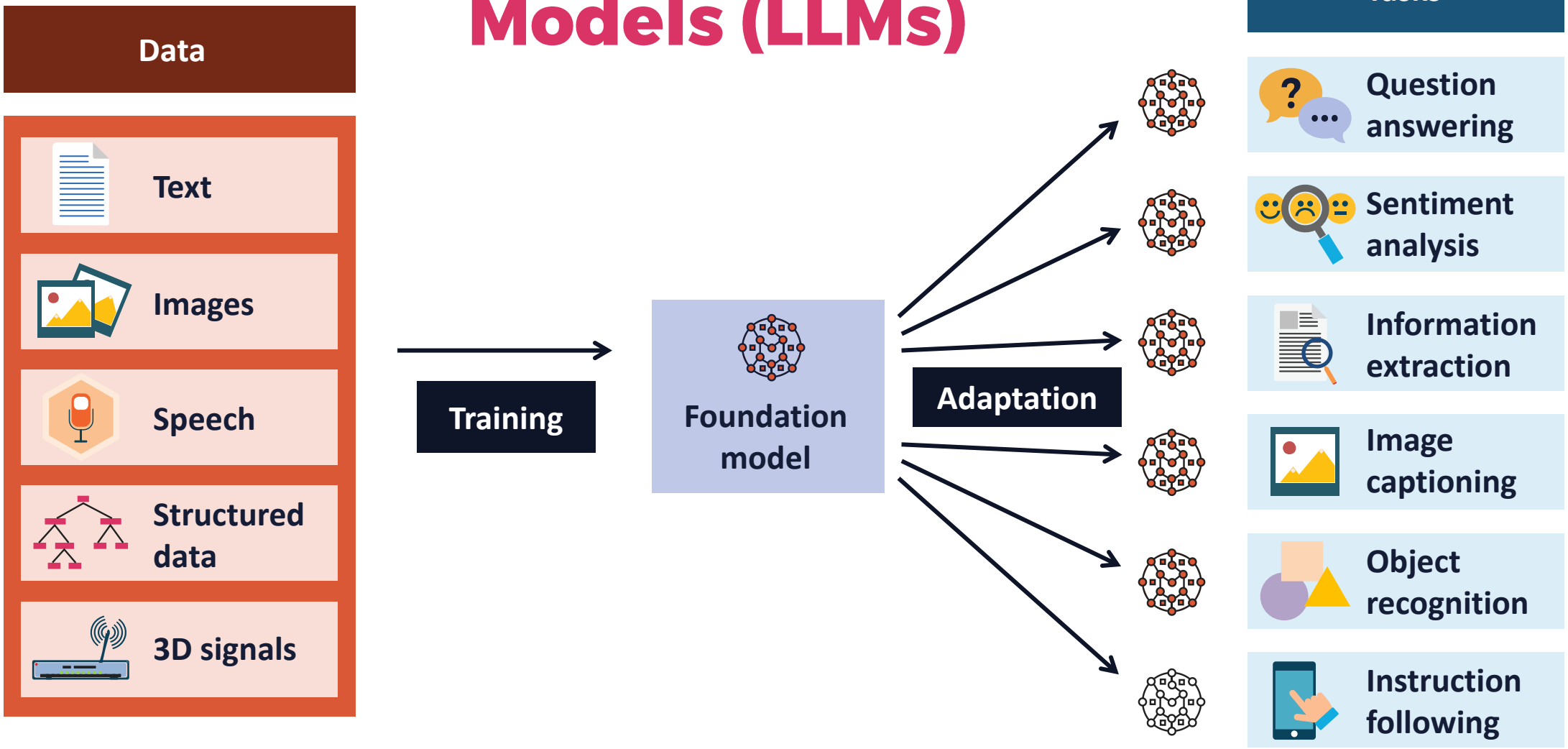
Large Language Models (LLMs)

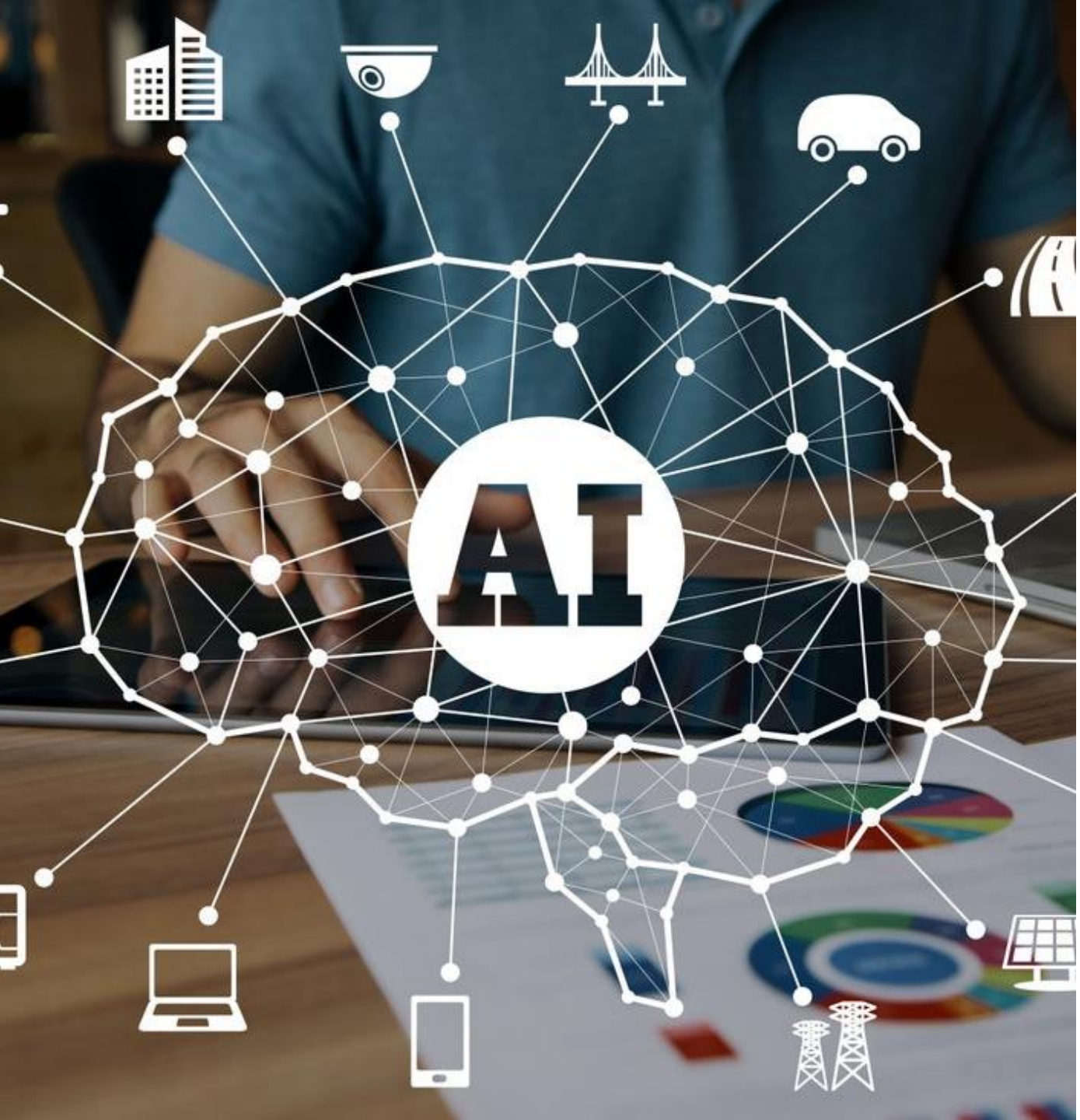
- AWS defines **large language models (LLMs)** as very large deep learning models that are pre-trained on **vast amounts of data**
- These models use a transformer architecture, which consists of an encoder and a decoder with self-attention capabilities

Large Language Models (LLMs)

- LLMs are incredibly flexible and can perform a wide range of tasks, such as answering questions, summarizing documents, translating languages, and completing sentences
- They are capable of **unsupervised training, meaning they can learn from data without explicit instructions**

Large Language Models (LLMs)





Key Differences Between AI, ML, and Deep Learning

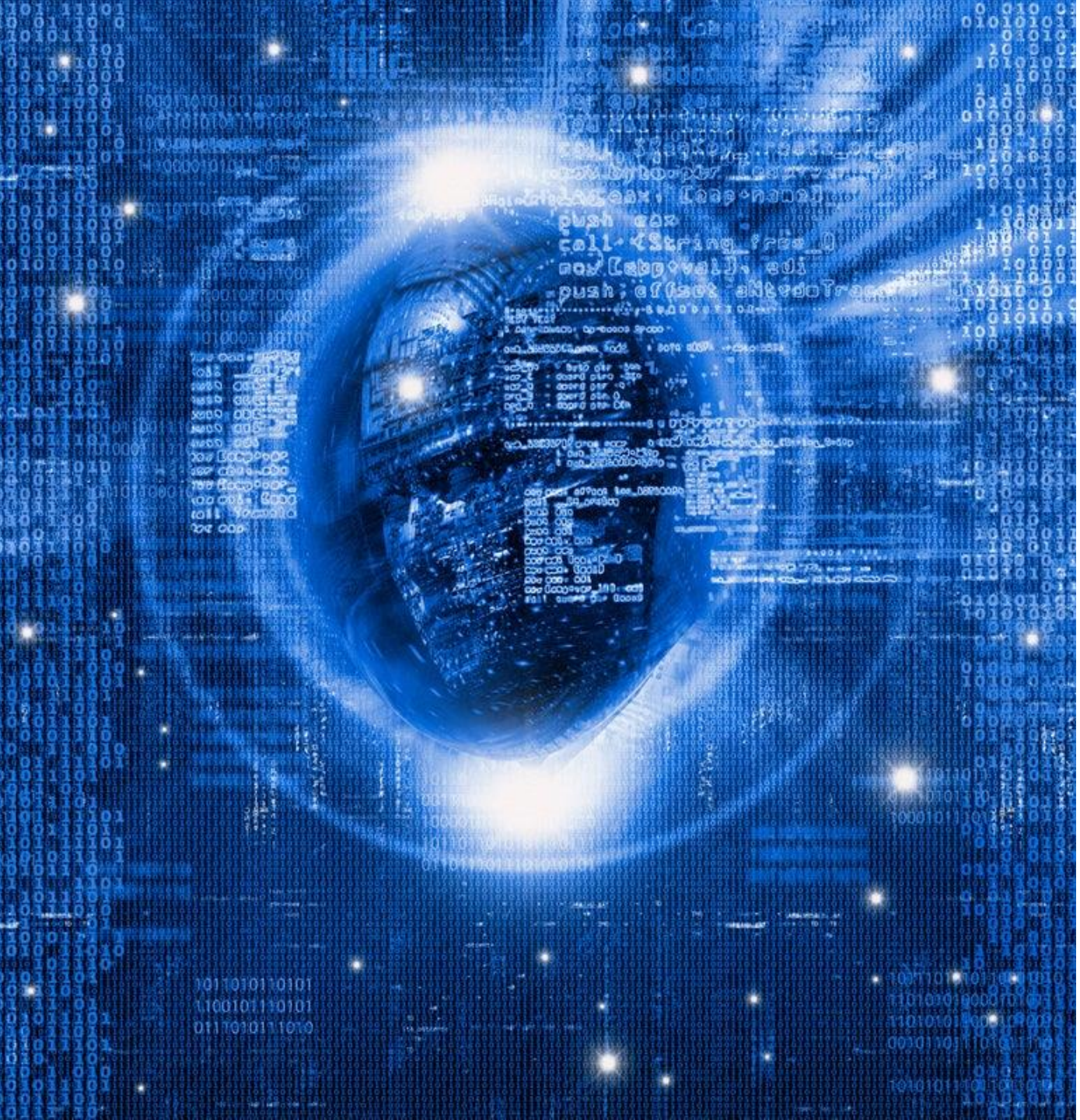
In a nutshell, **AI** is the overarching field, **ML** is a specialized branch within AI, and **deep learning** is a further specialization within ML that focuses on neural networks and large datasets

Key Differences Between AI, ML, and Deep Learning

- **AI** is a broad field of computer science focused on creating systems capable of performing tasks that typically require human intelligence
- This includes tasks like recognizing speech, making decisions, and understanding natural language
- It encompasses a wide range of technologies and applications, **including ML and deep learning**

Key Differences Between AI, ML, and Deep Learning

- **ML** is a subset of AI that involves training algorithms to learn from and make predictions or decisions based on data
- It allows systems to improve their performance over time without being explicitly programmed
- ML focuses on developing algorithms that can learn from data
- Examples include spam detection, image recognition, and predictive analytics

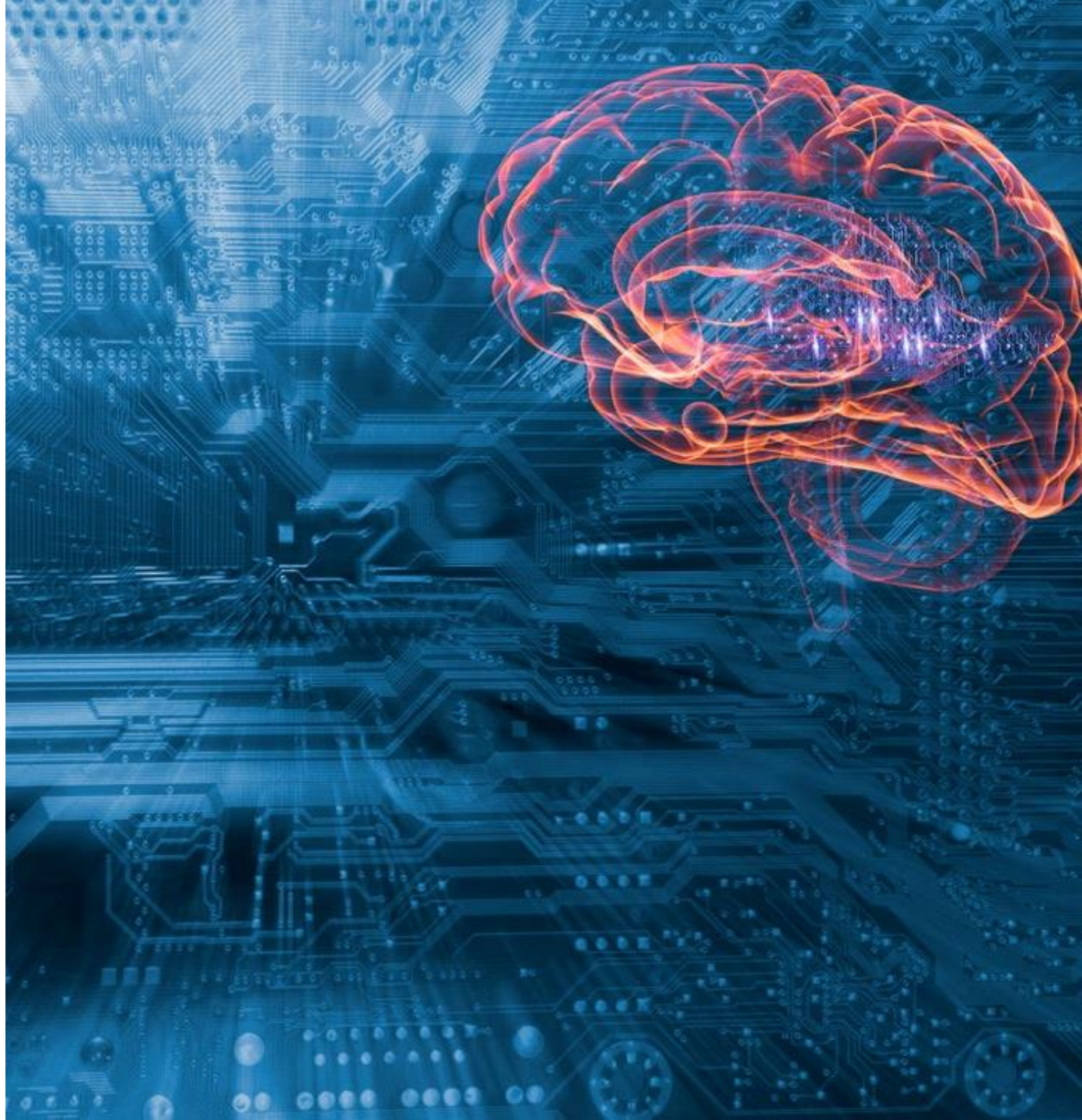


Key Differences Between AI, ML, and Deep Learning

- **Deep learning (DL)** is a subset of ML that uses artificial neural networks to model and understand complex patterns in data
  - These networks are inspired by the structure and function of the human brain
- The scope of DL involves training large neural networks on vast amounts of data to perform tasks with high accuracy
- Examples of DL include image and speech recognition, natural language processing, and autonomous driving

Machine Learning Inference

- Machine learning **inference** is the process of using a trained ML model to make predictions or decisions based on new, unseen data
- After a model has been trained on a dataset, it can be deployed to infer or predict outcomes for new inputs
- This process is crucial for applying the insights gained during training to real-world scenarios



Batch vs. Real-Time Inferencing

Batch (offline) inferencing involves processing large volumes of data at once

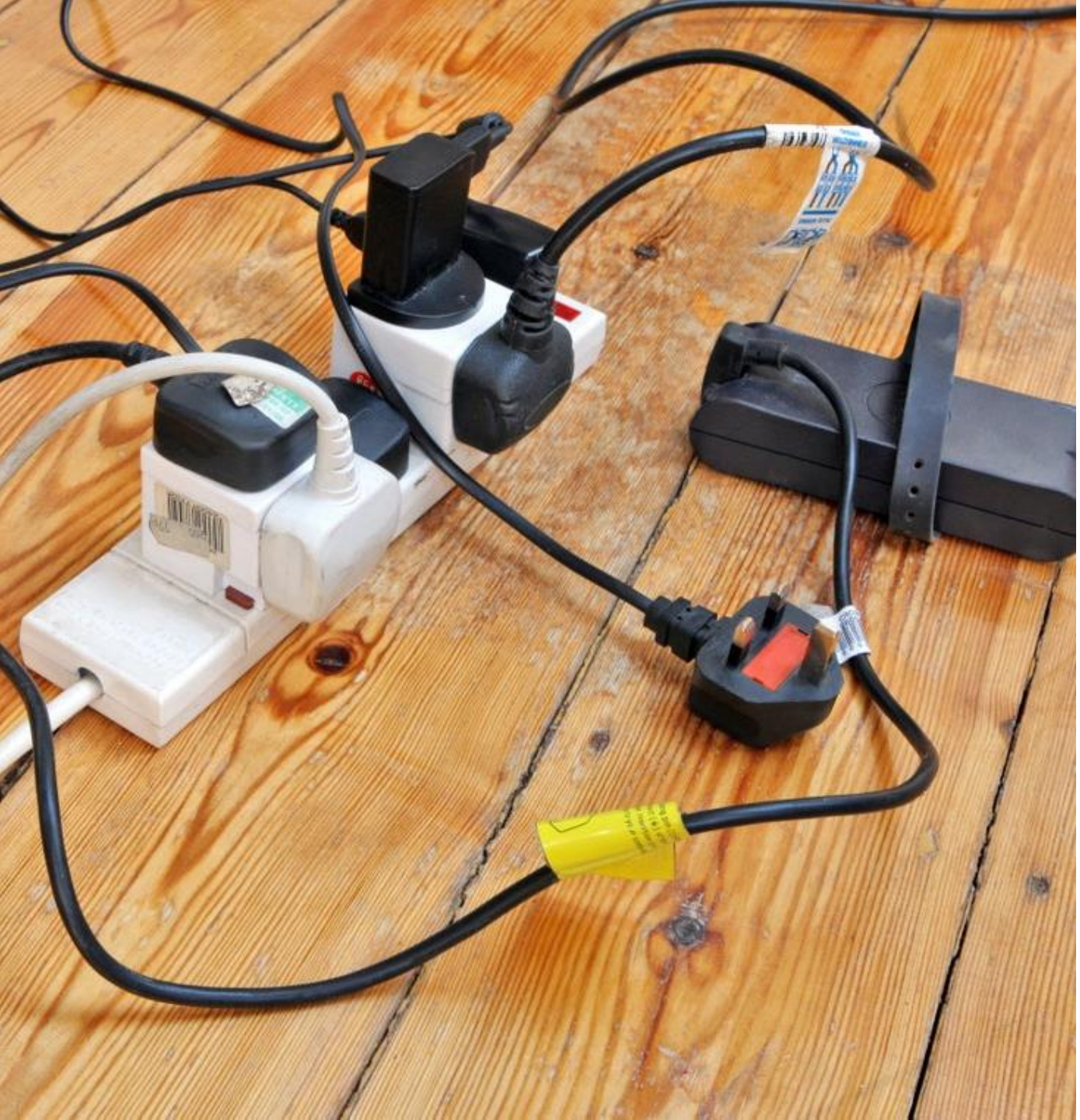
This is suitable for scenarios where immediate results are not critical (e.g., generating nightly recommendations for streaming services or processing large datasets for business analytics)

Real-time (online or interactive) inferencing processes individual data points as they arrive, providing immediate predictions

This is ideal for applications requiring instant responses (e.g., voice assistants, fraud detection, and online recommendation systems)

Batch Inference

- **Batch inference** processes large datasets in batches, generating predictions for the entire dataset at once
- This type is ideal for offline processing tasks, such as data preprocessing, large-scale predictions, and business analytics



Asynchronous Inference

- **Asynchronous inference** handles large payloads and long processing times by queuing requests and processing them asynchronously
- This is suitable for batch processing tasks, such as video analysis and large-scale data processing

Serverless Inference

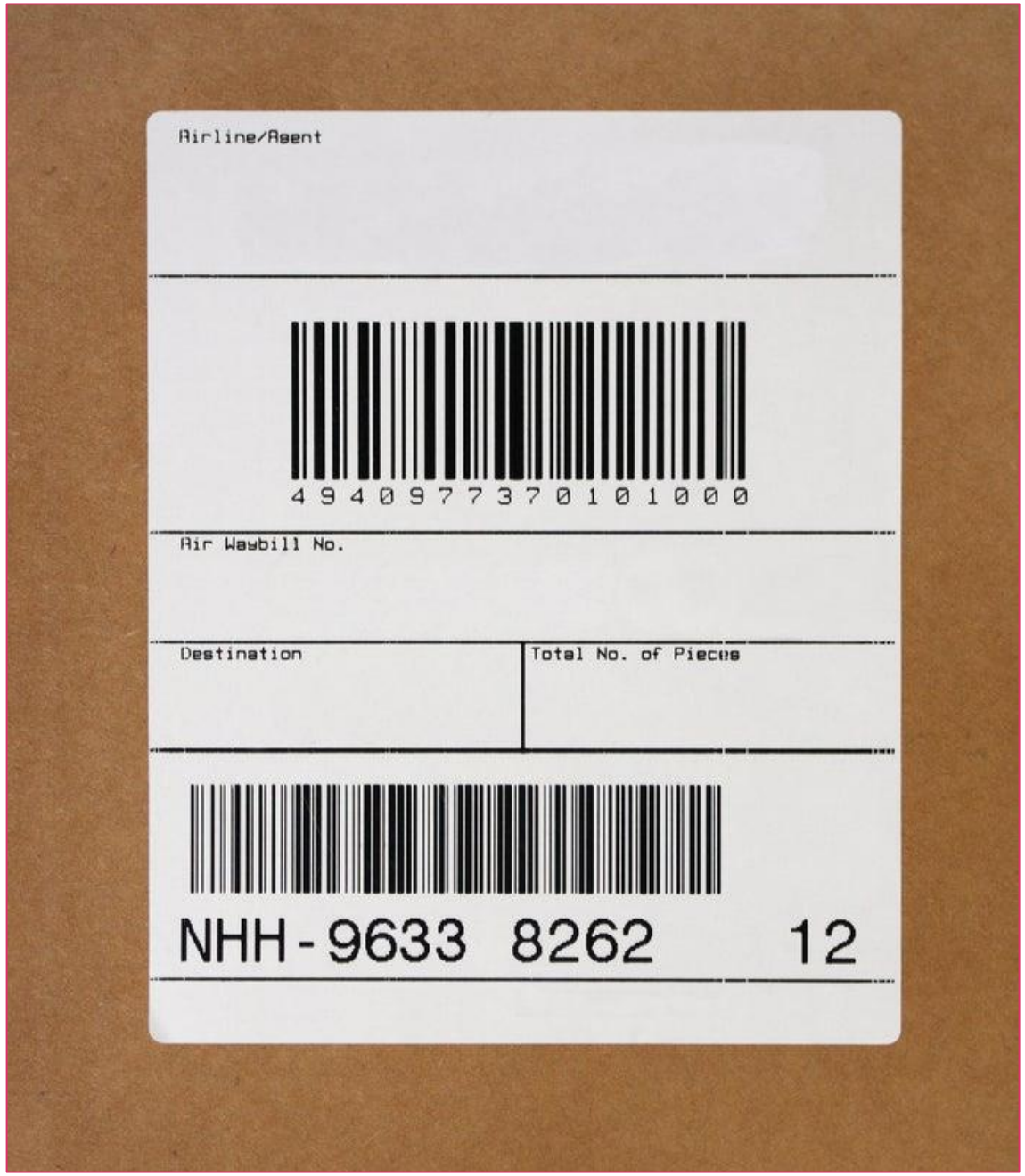
- **Serverless inference** offers a fully managed infrastructure that automatically scales based on the traffic, eliminating the need for manual scaling
- This type is best for applications with intermittent traffic patterns, such as mobile apps and web services

Data Types in AI Models

- AI uses various data types in the models, and each has its unique characteristics and is suited to different ML tasks
- The choice of data type depends on the specific problem you're trying to solve and the nature of the data available

Labeled Data

- **Labeled data** refers to datasets where each input token (i.e., an image) is tagged with relevant labels or annotations to help train ML models
- Use cases include supervised learning tasks like classification and regression
- Specific examples of labeled data include email spam detection (i.e., emails labeled as spam or not spam) and image classification (i.e., images labeled with categories)

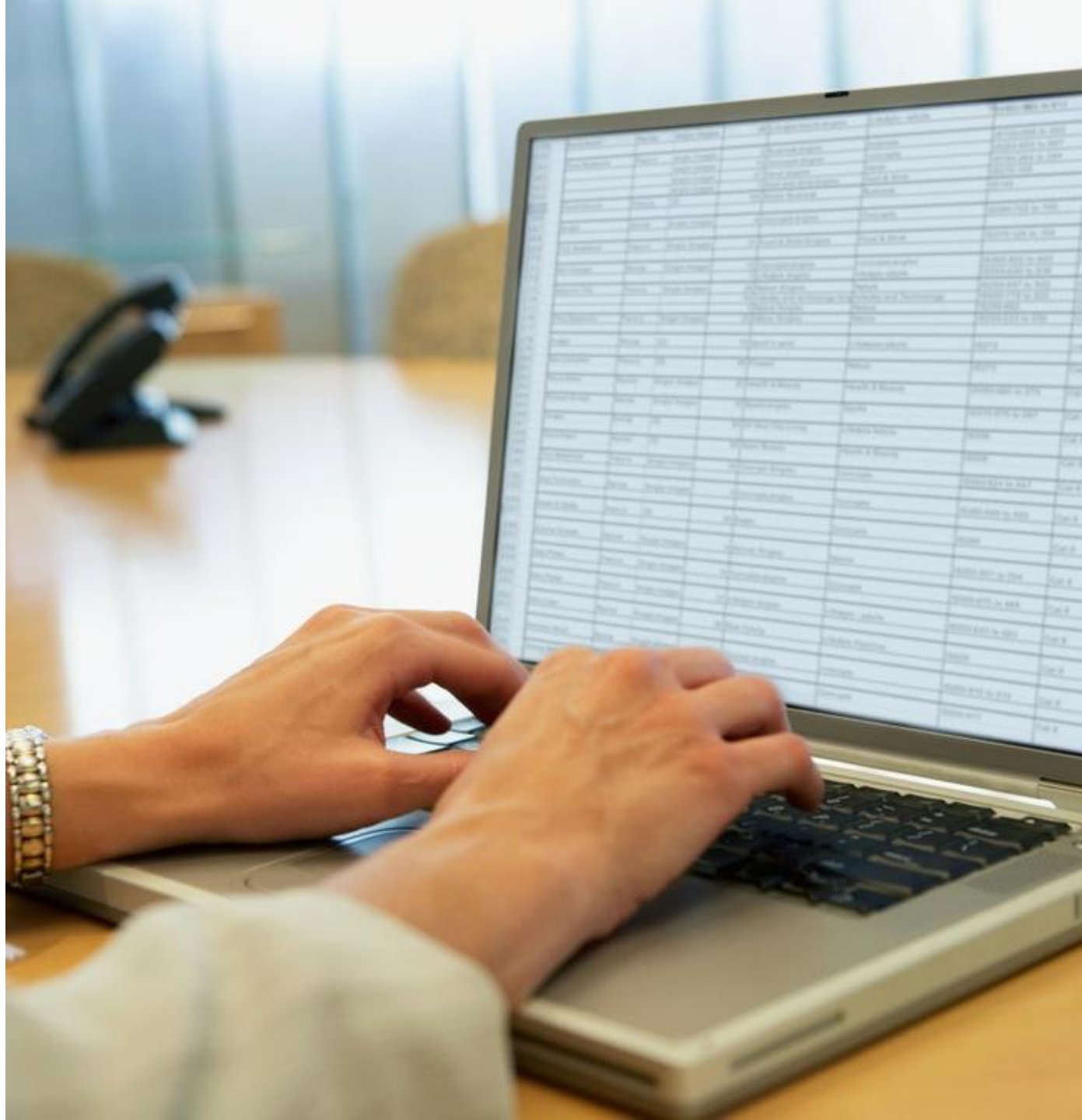


Unlabeled Data

- **Unlabeled data** is data that that includes input features but no corresponding output labels
- Use cases include unsupervised learning tasks like clustering and anomaly detection
- Examples of unlabeled data include customer purchase history without any labels and raw text data

Tabular Data

- **Tabular data** is organized in rows and columns as in an Excel spreadsheet
- Some uses could be various ML tasks, including classification, regression, and clustering
- Specific examples of tabular data include financial records and customer information databases

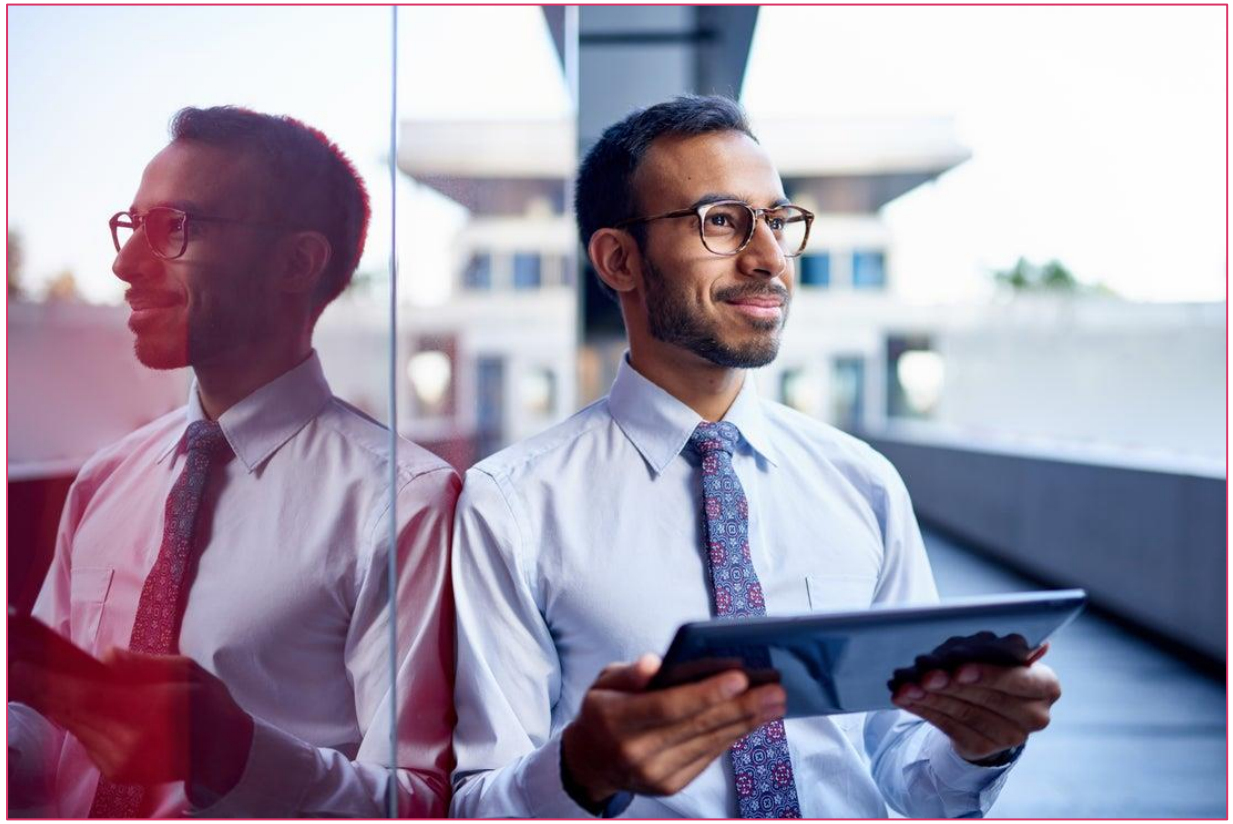


Time-Series Data

- **Time-series data** is data points collected or recorded at specific time intervals
- Uses include forecasting, anomaly detection, and trend analysis
- Examples of time series data include stock or cryptocurrency prices, weather data, and sensor readings

Image Data

- **Image data** comes in the form of images, typically represented as pixel values
- Some common use cases include computer vision tasks like image classification, object detection, and segmentation
- Examples of image data include photographs and medical imaging scans



Text Data

- **Text data** comes in the form of written language
- Use cases include natural language processing (NLP) tasks like sentiment analysis, text classification, and machine translation
- Examples of text data include emails, social media posts, and articles

Structured Data

- **Structured data** is highly organized and easily searchable, often stored in relational, non-relational, and graph databases
- Common use cases include various ML tasks, such as classification, regression, and clustering
- Examples of structured data use include Amazon RDS, Amazon Neptune, and spreadsheets

Unstructured Data

- **Unstructured data** is data that lacks a predefined structure, making it more challenging to process and analyze
- Use cases are NLP, computer vision, and other tasks requiring advanced data processing techniques
- Examples of unstructured data include text documents, images, audio files, and videos

Machine Learning Methods

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

Supervised Learning

- **Supervised learning** involves training a model on a labeled dataset, where the input data is paired with the correct output
- This learning type is appropriate for tasks **where the desired output is known,** such as classification (e.g., spam detection) and regression (e.g., predicting house prices)

Supervised Learning

- Advantages of supervised learning is high accuracy and reliability when trained on a large, representative dataset
- Challenges include needing a large amount of labeled data, which can be time-consuming and expensive to obtain

Unsupervised Learning

- **Unsupervised learning** involves training a model on an unlabeled dataset, where the model tries to find patterns and relationships within the data
- This type of ML is excellent for tasks like clustering (e.g., customer segmentation) and anomaly detection (e.g., fraud detection)

Unsupervised Learning

- An advantage of unsupervised learning is that it can work with unlabeled data, making it **useful for exploratory data analysis**
- One major issue might be that the results can be less interpretable, and the model may require more fine-tuning to achieve meaningful insights

Reinforcement Learning

- **Reinforcement learning** involves training a model to make a sequence of decisions by rewarding it for good actions and penalizing it for bad ones
- This type is suitable for tasks that involve decision-making over time, such as game playing (e.g., AlphaGo) and robotics (e.g., autonomous driving)

Reinforcement Learning

- An advantage of reinforcement learning is that it can learn complex behaviors and adapt to dynamic environments
- Challenges to this learning type are as follows:
  - It requires a large amount of computational resources
  - It can be difficult to train due to the need for a well-defined reward system

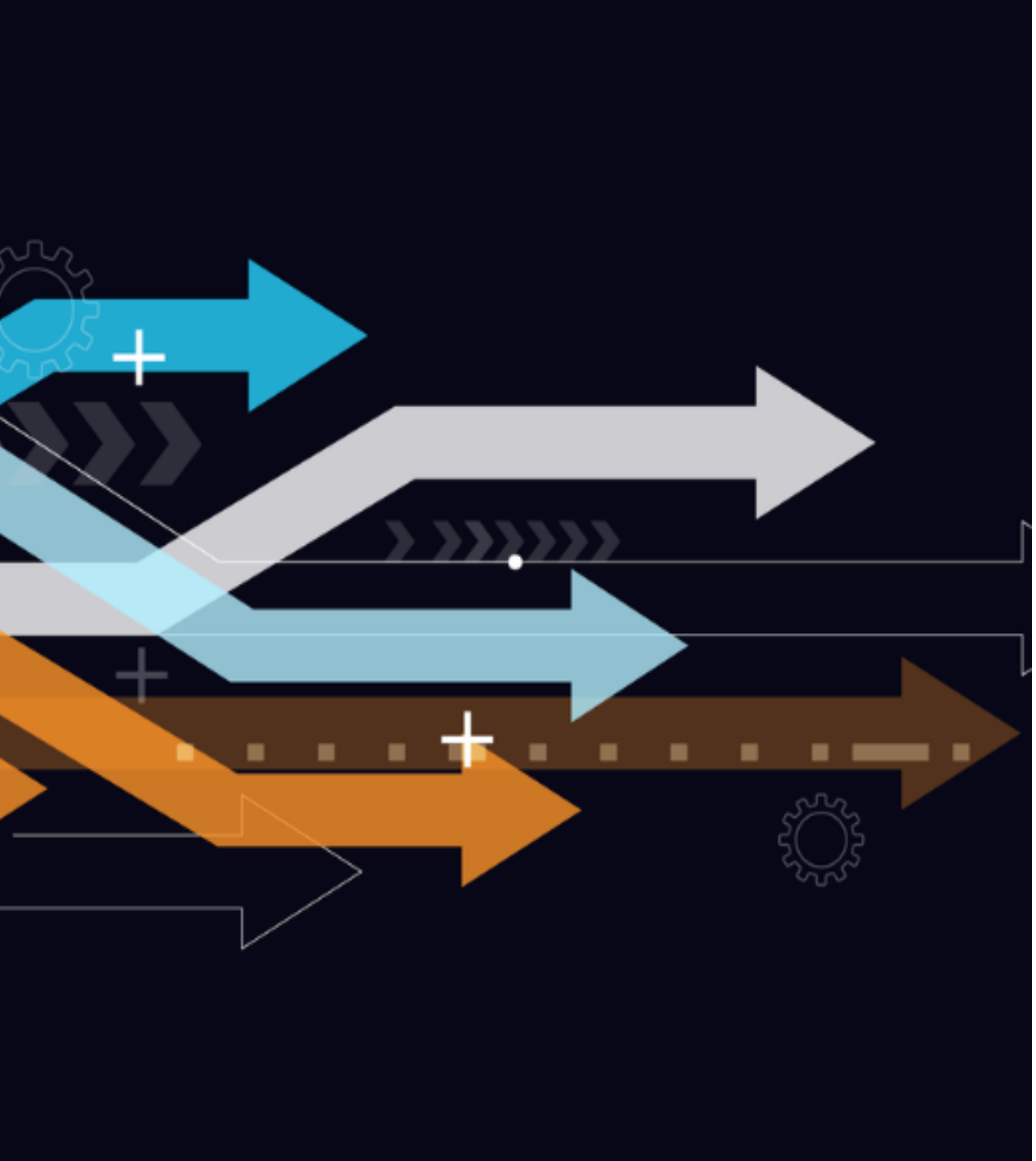
Practical Use Cases for AI

Objectives

- Examine applications where AI/ML provides value
- Discover where AI/ML is not appropriate
- Describe ML techniques for use cases and examples of real-world applications
- Explore Amazon SageMaker
- Learn about Amazon Transcribe and Translate
- Explore Amazon Comprehend
- Examine Amazon Lex
- Explore Amazon Polly

Applications Where AI/ML Provides Value

- **Automated scaling:** AI/ML models can automatically adjust resources based on demand, ensuring optimal performance without manual intervention
  - This is particularly useful for applications with fluctuating workloads
- **Efficient resource management:** AI/ML can optimize the allocation of computational resources, reducing costs and improving efficiency
  - For example, predictive analytics can forecast resource needs and adjust capacity accordingly



Applications Where AI/ML Provides Value

- **Enhanced performance:** AI/ML algorithms can improve the performance of applications by optimizing processes and reducing latency
  - This is crucial for real-time applications that require quick responses
- **Data handling:** AI/ML can manage and process large volumes of data efficiently, enabling organizations to scale their data operations without compromising on speed or accuracy

Applications Where AI/ML Provides Value

- **Continuous learning and adaptation:** AI/ML models can continuously learn from new data and adapt to changing conditions, ensuring that solutions remain effective as they scale
- **Integration with cloud services:** AWS provides a range of AI/ML services that integrate seamlessly with other AWS cloud services, making it easier to build, deploy, and scale AI/ML solutions

Where AI/ML Solutions Are Not Appropriate

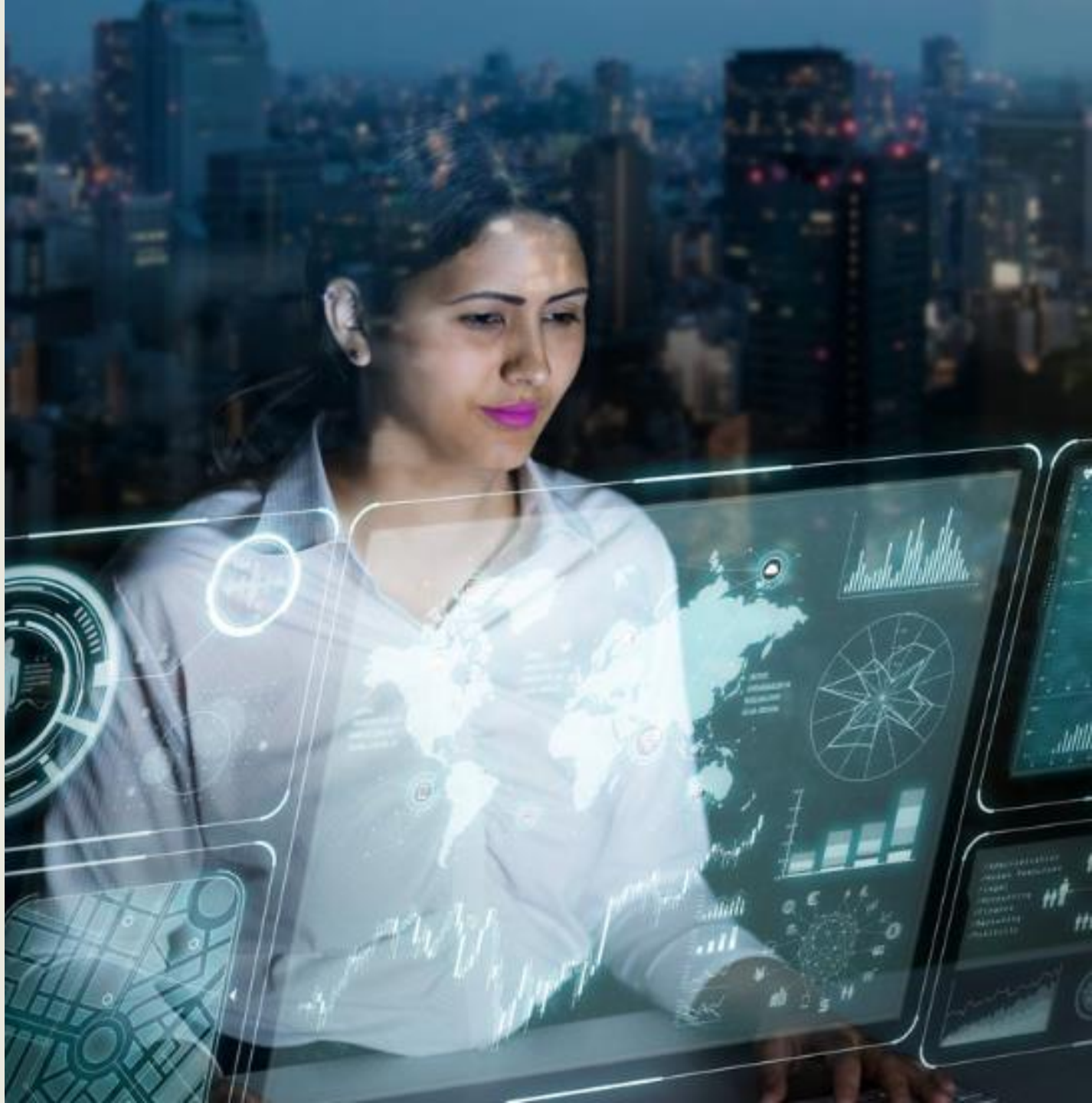
- AI/ML solutions are not appropriate in cost benefit analyses and situations when a specific outcome is needed instead of a prediction
- AI/ML models require large amounts of high-quality data to train effectively so if there is not enough data or if the data quality is poor, the model's performance will suffer

Where AI/ML Solutions Are Not Appropriate

- AI and ML are often not appropriate with decisions that have significant consequences, such as medical diagnoses or legal judgments
  - Human oversight is crucial to ensure accuracy and accountability
- AI/ML solutions can raise ethical and privacy issues, especially when dealing with sensitive data
  - It is important to consider the potential impact on individuals' privacy and ensure compliance with regulations

Where AI/ML Solutions Are Not Appropriate

- If the training data is biased, the AI/ML model will likely perpetuate these biases, leading to unfair outcomes
  - It is essential to address and mitigate bias to ensure fairness
- Implementing AI/ML solutions can be complex and costly so for some problems, simpler and more cost-effective solutions might be more appropriate



Specific ML Use Cases: Regression

- Regression in ML is a method for predicting numerical outcomes based on input data
- Regression models are used to understand the relationship between variables and to forecast future value
  - For example, regression can be used to predict house prices, customer lifetime value, or the number of users for a service based on historical data

Specific ML Use Cases: Regression

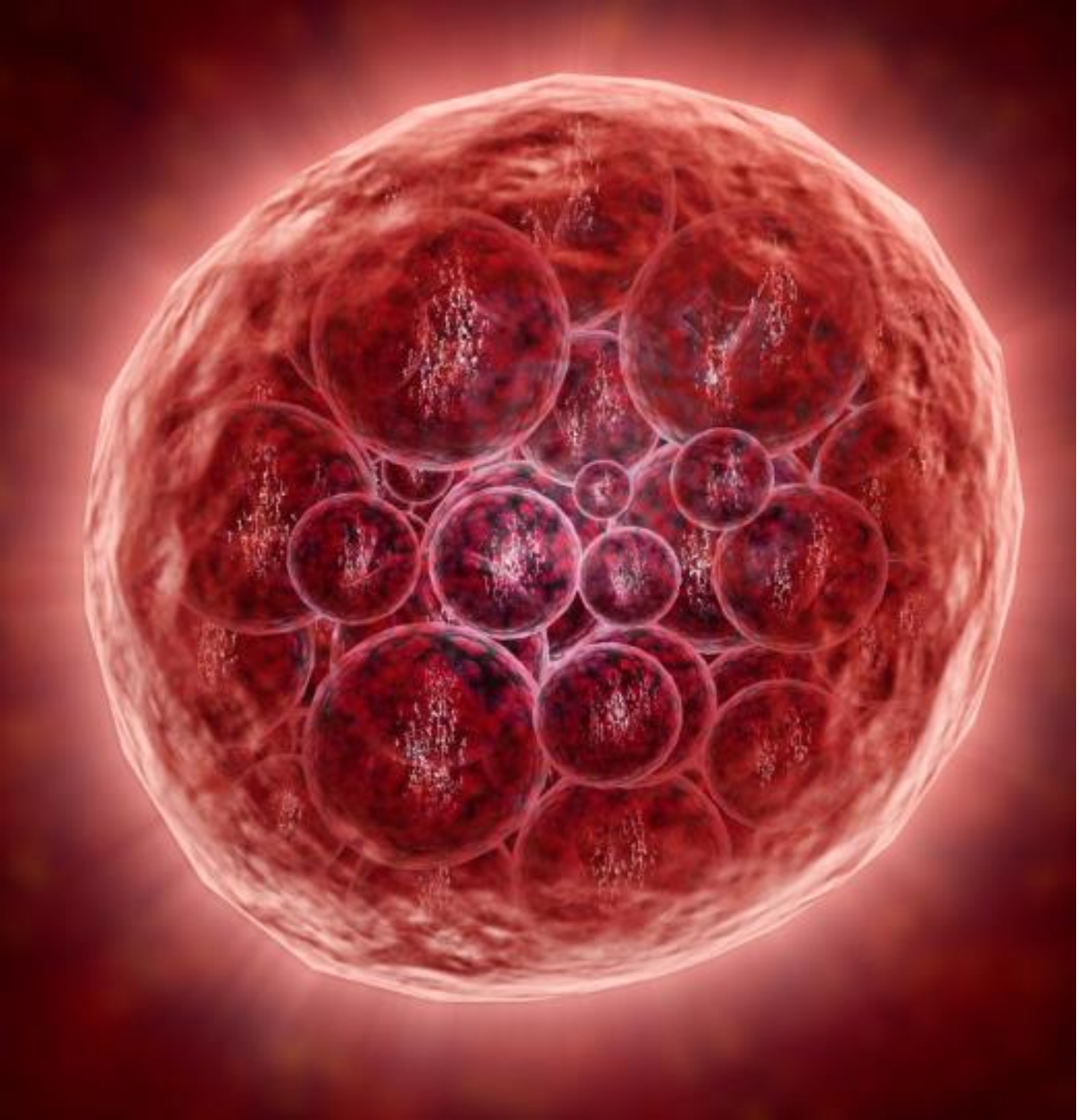
- In AWS, regression models can be built using various algorithms, such as the linear learner algorithm in **Amazon SageMaker**
- This algorithm trains multiple models in parallel and selects the most optimized one for making predictions
- AWS provides tools like Amazon Redshift ML to create and deploy these models, allowing users to run prediction queries directly within their data warehouse

Specific ML Use Cases: Classification

- The ML classification use case is a method for categorizing data into predefined classes or labels
- Classification models are trained to recognize patterns in input data and assign them to one of the several categories

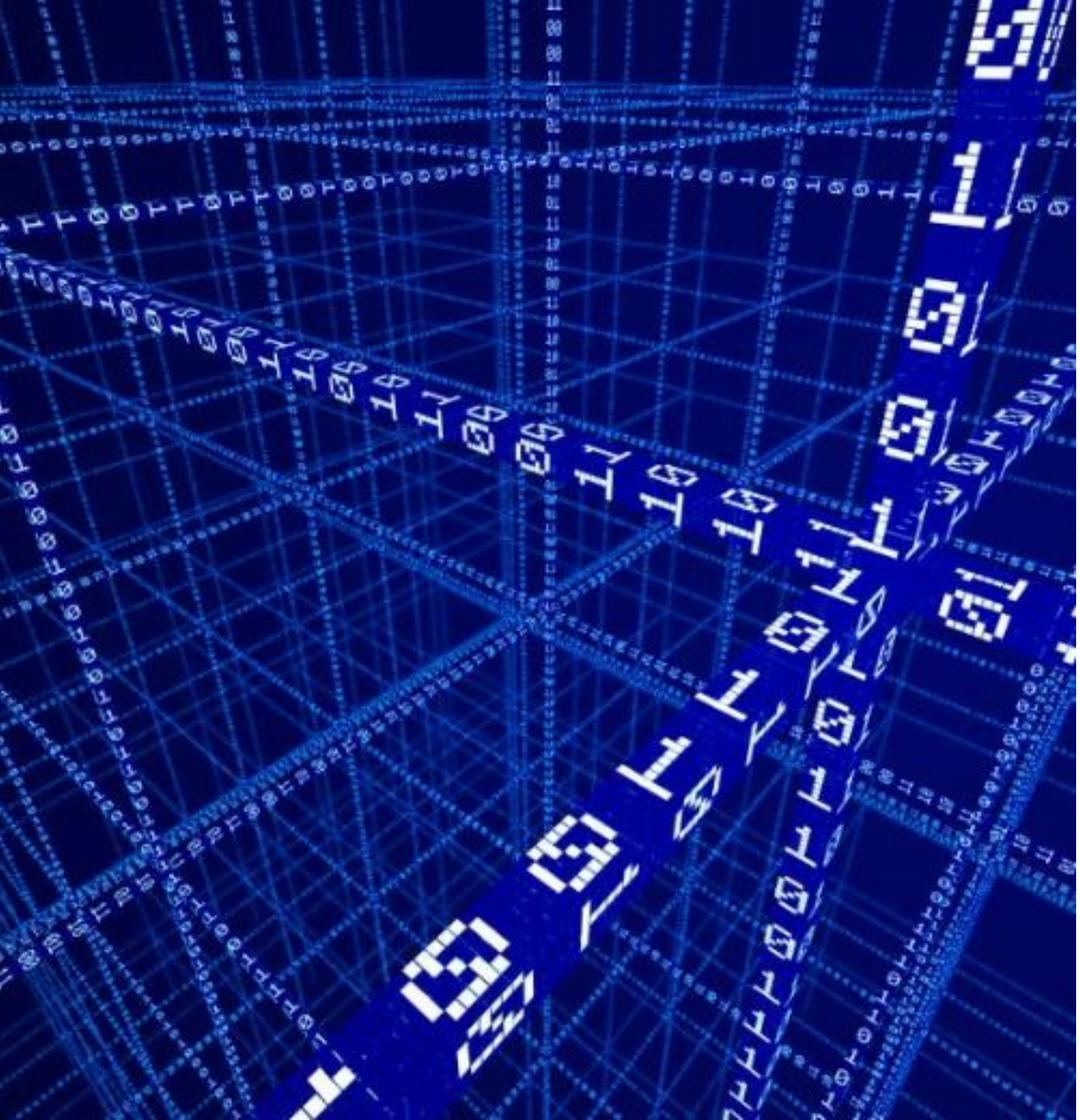
Specific ML Use Cases: Classification

- **Spam detection**: classifying emails as spam or not spam
- **Sentiment analysis**: determining the sentiment of text data, such as customer reviews as positive, negative, or neutral
- **Image classification**: identifying objects within images, such as recognizing animals or vehicles
- **Fraud detection**: classifying transactions as fraudulent or legitimate
- **Medical diagnosis**: categorizing medical images or patient data to diagnose diseases



Specific ML Use Cases: Clustering

- AWS describes the ML use case of clustering as a method for discovering groupings or patterns in data **without predefined labels**
- Clustering is an unsupervised learning technique that groups similar data points together based on their characteristics



Specific ML Use Cases: Clustering

- Some key aspects of clustering include:
  - Unsupervised learning: clustering does not require labeled data, making it useful for exploratory data analysis
  - Grouping similar data: the goal is to partition data into clusters where data points within the same cluster are more similar to each other than those in other clusters

Specific ML Use Cases: Clustering

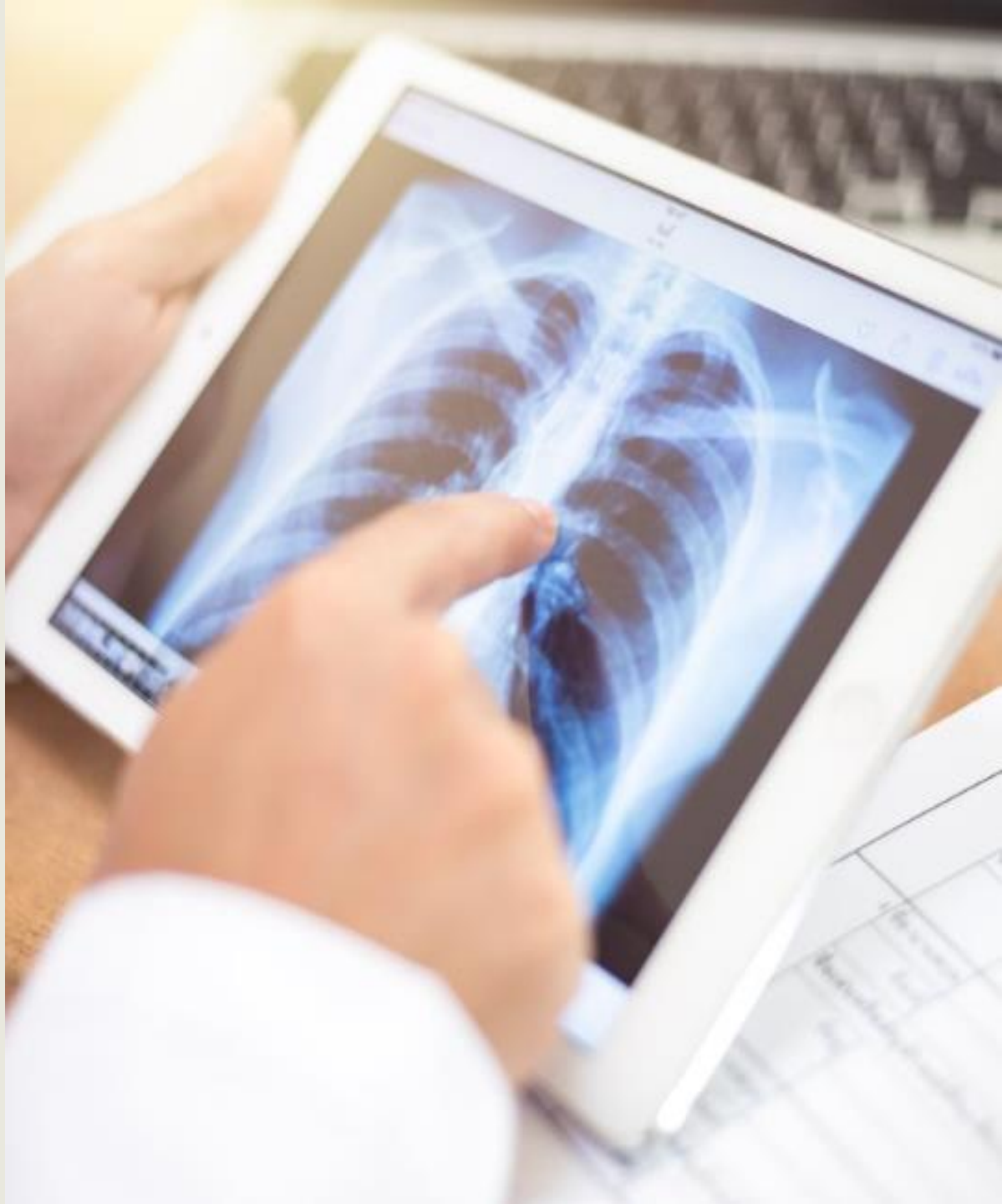
- **Customer segmentation**: grouping customers based on purchasing behavior or demographics
- **Anomaly detection**: identifying unusual patterns or outliers in data, such as fraud detection
- **Image segmentation**: grouping similar pixels in an image for object detection or medical imaging
- **Document clustering**: organizing documents into topics based on content similarity

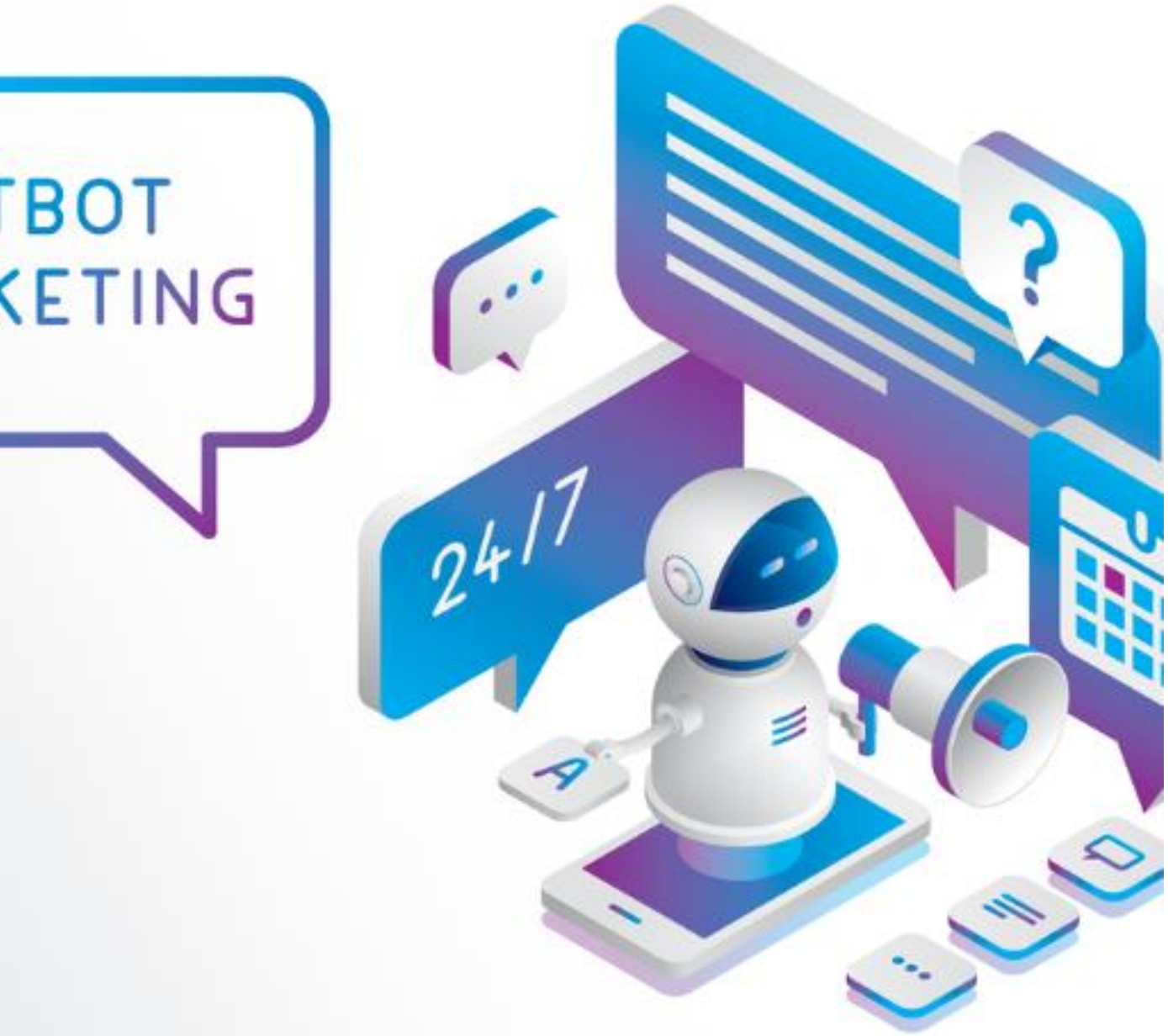
Real-World AI Applications of Computer Vision

- Security surveillance with cameras and sensors monitor public spaces, industrial sites, and high-security environments
- Personal safety at home where real-time streams detect pets or visitors at the front door
- Operational efficiency and quality control by automatically identifying defects in products before they leave the factory

Real-World AI Applications of Computer Vision

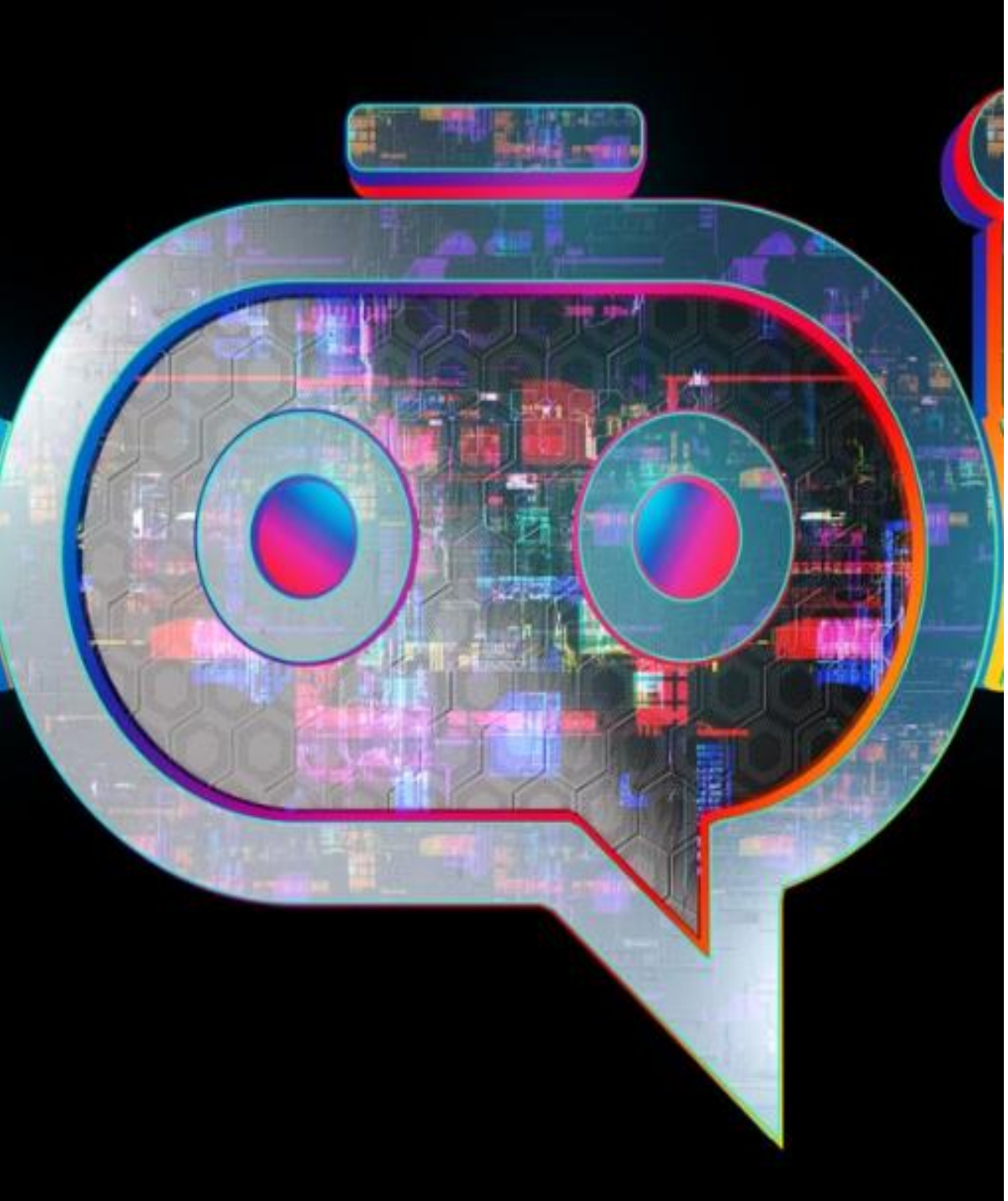
- Managing inventory using vision-enabled robots that assist with stocking inventory, filling orders, and sorting packages for delivery
- Analyzing medical images to assist in diagnosing diseases, such as detecting tumors in X-rays or MRIs
- In agriculture, estimating the weight of fish while they are still in the water
- Analyzing images of crops to detect diseases, pests, and nutrient deficiencies





Real-World AI Applications of NLP

- Powers chatbots and virtual assistants that can understand and respond to customer queries in real time, providing efficient and personalized support
- Transcribe and analyze medical records, enabling healthcare professionals to quickly access and interpret patient information
- Analyze transaction data and detect patterns indicative of fraudulent activities



Real-World AI Applications of NLP

- Financial institutions use NLP to gauge market sentiment by analyzing news articles, social media posts, and other textual data
- Automatically detect and filter inappropriate or harmful content on social media platforms, ensuring a safer online environment
- Assisting in reviewing and summarizing legal documents, making it easier for legal professionals to find relevant information and make informed decisions

Real-World AI Applications of Speech Recognition

- Analyzing customer calls to extract insights such as sentiment, call categories, and call characteristics to improve customer experiences
- Automated transcription to convert speech to text for customer service interactions, making it easier to review and analyze conversations
- Automatically transcribing doctor-patient conversations, clinical notes, and medical records

Real-World AI Applications of Speech Recognition

- Enabling hands-free operation of medical devices and systems, enhancing accessibility for healthcare professionals
- Generating subtitles for videos and live broadcasts to increase accessibility and improve viewer experience
- Detecting and categorizing toxic or inappropriate content in audio streams, ensuring a safer online environment

Real-World AI Applications of Recommendation Systems

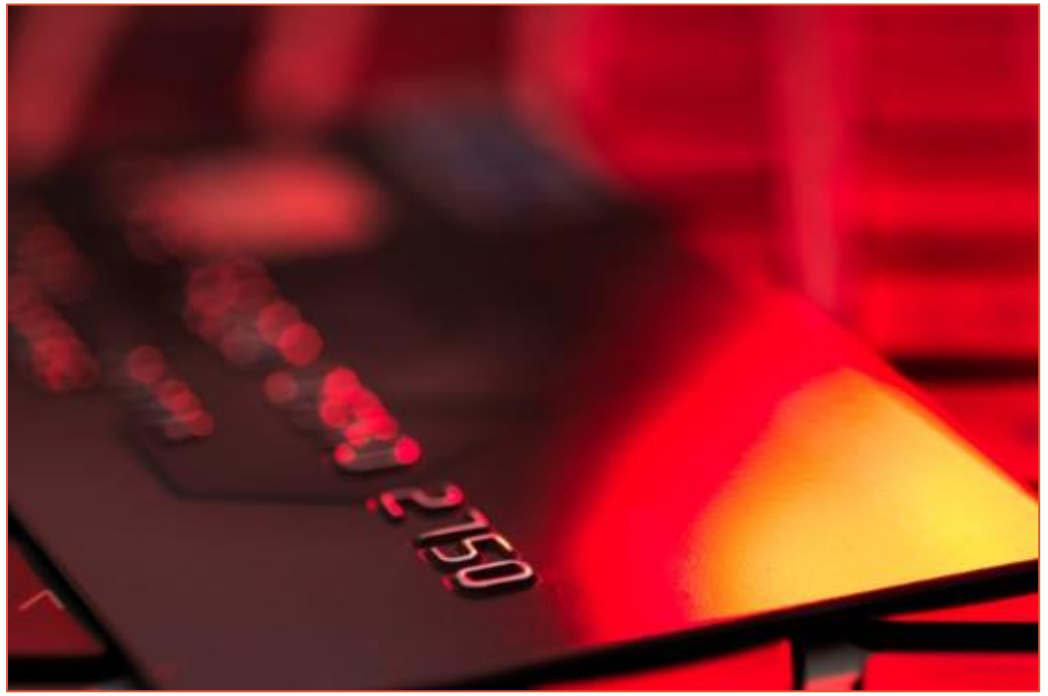
- Suggesting products to customers based on their browsing history, past purchases, and preferences
- E-commerce product recommendations to help increase sales and improve customer satisfaction
- Personalizing movie, TV show, and music recommendations based on user preferences and viewing history

Real-World AI Applications of Recommendation Systems

- Delivering personalized ads to users based on their interests and behavior to improve the effectiveness of marketing
- Offering tailored product suggestions and promotions to customers in physical stores through apps or in-store kiosks
- Providing personalized investment advice and portfolio tips based on individual financial goals and risk tolerance

Real-World AI Applications of Fraud Detection

- Identifying fraudulent transactions during online purchases or payment processing like detecting unusual patterns in transaction data, such as high-value purchases from new accounts or transactions from unusual locations
- Preventing fraudsters from creating fake or synthetic identities to open new accounts



Real-World AI Applications of Fraud Detection

- Detecting and preventing fraud in real time to minimize losses and protect customers
- Enhancing traditional rule-based systems by improving the accuracy and scalability of traditional rule-based fraud detection systems
  - For example: combining rule-based systems with AI models to adapt to new fraud patterns and reduce false positives

Real-World AI Applications of Forecasting

- Predicting future product demand to optimize inventory levels, reduce stockouts, and manage inventory
- Optimizing the supply chain by forecasting demand and supply to improve logistics, reduce costs, and enhance efficiency
- Predicting market trends, stock prices, and economic indicators

Real-World AI Applications of Forecasting

- Predicting energy usage patterns to optimize energy production and distribution to assist utility companies manage resources more efficiently and reduce costs
- Predicting patient admissions and resource needs to optimize hospital staffing and resource allocation
- Assisting meteorologists and government assistance agencies to better predict the weather

SageMaker Capabilities

In this demo…

We will explore the capabilities of Amazon SageMaker

Amazon Transcribe

- Amazon Transcribe is a fully managed, **automatic speech recognition (ASR)** service provided by AWS
- It converts speech to text, making it easy for developers to add speech-to-text capabilities to their applications
- Transcribe is designed to help businesses automate manual tasks, unlock insights from audio and video content, and increase accessibility

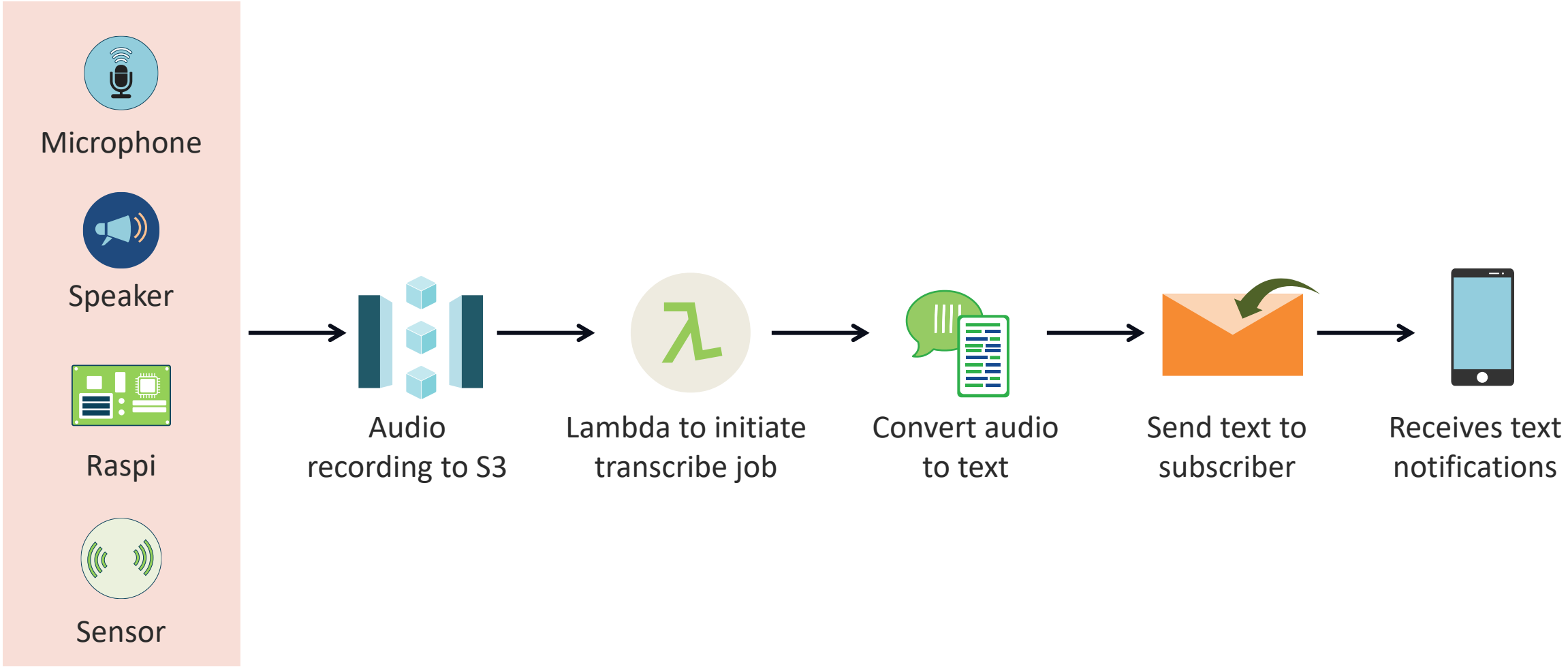
Key Features of Amazon Transcribe

- Transcribe includes automatic punctuation, speaker **diarization** (identifying different speakers in an audio), word-level confidence scores, and vocabulary filters
- It can redact sensitive information and detect toxic content in audio streams
- Transcribe is capable of transliterating media in real-time (streaming) or processing media files in batches

Amazon Transcribe Use Case Examples

- **Call analytics**: extracting insights from customer calls, improving customer experience, and boosting agent productivity
- **Subtitles for videos and meetings**: creating subtitles to increase accessibility and improve experiences
- **Medical transcription**: documenting clinical conversations into electronic health record (EHR) systems
- **Content moderation**: detecting and categorizing toxic audio content to foster a safe online environment

Amazon Transcribe and IoT Devices



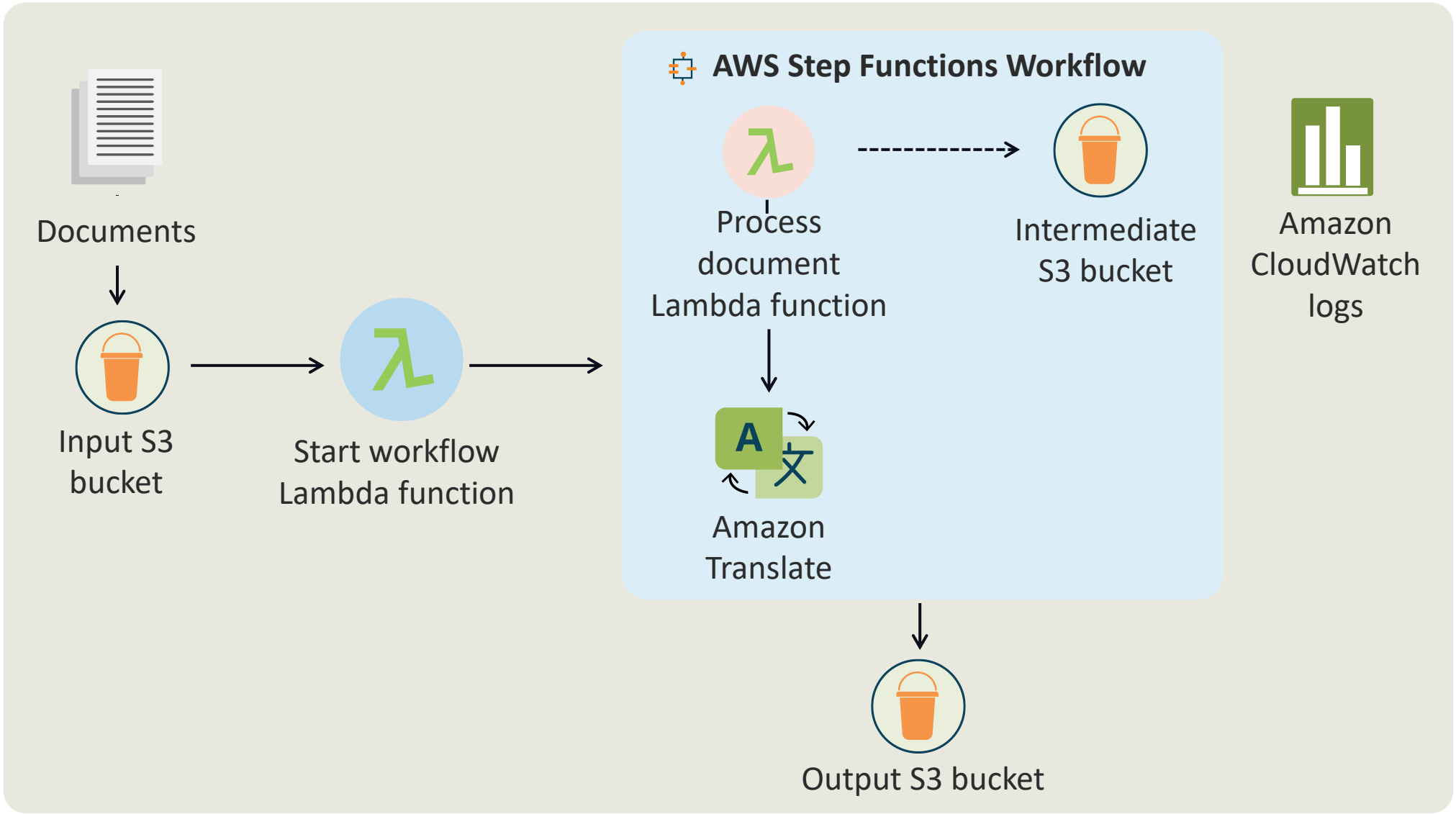
Amazon Translate

- Amazon Translate is a neural machine translation service that uses advanced ML technologies to deliver high-quality, fast, and customizable language translation
- Translate is designed to help businesses and developers easily integrate language translation capabilities into their applications, making it easier to reach a global audience

Amazon Translate Use Case Examples

- Multilingual user experiences: enabling applications to support multiple languages, enhancing user experience
- Content localization: translating user-generated content such as social media posts, reviews, and comments in real time
- Customer support: facilitating communication between customers and support agents who speak different languages
- Document translation: translating large volumes of documents, such as technical manuals, reports, and articles

Amazon Translate in Action



Amazon Comprehend

- Amazon Comprehend is a natural language processing (NLP) service provided by AWS that uses ML to **uncover valuable insights from text within documents**
- Comprehend simplifies document processing workflows and helps businesses gain valuable perceptions from their text data without needing ML expertise

Amazon Comprehend Use Case Examples

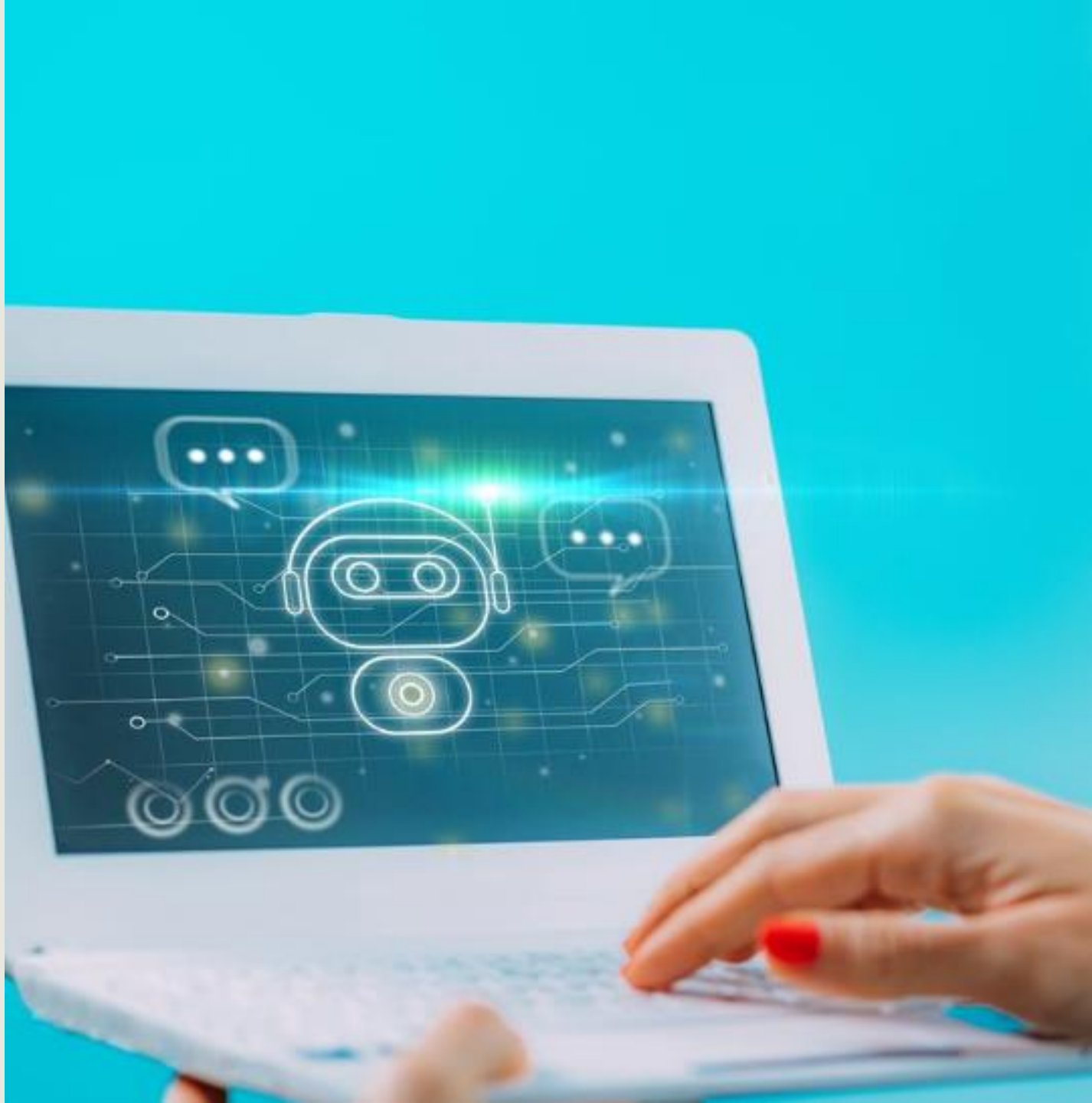
- **Customer support**: analyzing customer interactions to detect sentiment and categorize support requests
- **Content moderation**: detecting and redacting personally identifiable information (PII) from documents
- **Market research**: analyzing social media feeds and customer reviews to gain insights into customer sentiment and preferences

Amazon Comprehend Use Case Examples

- **Legal document management**: automating the extraction of insights from legal documents such as contracts and court records
- **Financial services**: classifying and extracting entities from financial documents such as insurance claims or mortgage packages

Amazon Lex

- Lex is a fully managed AI service provided by AWS that allows developers to build conversational interfaces using voice and text
- Lex simplifies the process of building conversational AI interfaces for developers without deep learning expertise



Key Features of Amazon Lex

- Lex uses natural language understanding (NLU) to determine the intent behind user inputs, allowing for more natural and engaging interactions
- The automatic speech recognition (ASR) feature converts spoken language into text, enabling voice-based interactions
- Lex seamlessly integrates with other AWS services like AWS Lambda, Amazon Connect, and Comprehend, allowing for continuous deployment and scaling

Amazon Lex Use Case Examples

- **Customer service**: enhancing customer support with automated chatbots that can handle common queries and tasks
- **E-commerce**: assisting customers with product recommendations, order tracking, and more
- **Healthcare**: providing virtual health assistants for scheduling appointments, answering health-related questions, and more
- **Internal business applications**: streamlining internal processes such as IT helpdesk support, HR inquiries, and more

Polly Capabilities

In this demo…

- You can use Amazon Polly to generate speech from either plain text or from documents marked up with **Speech Synthesis Markup Language (SSML)**.
  - Using SSML-enhanced text gives you additional control over how Amazon Polly generates speech from the text you provide.
- Example:

<speak>*Hi! My name is Michael. I will read any text that you type here.***</speak>**

The ML Development Lifecycle

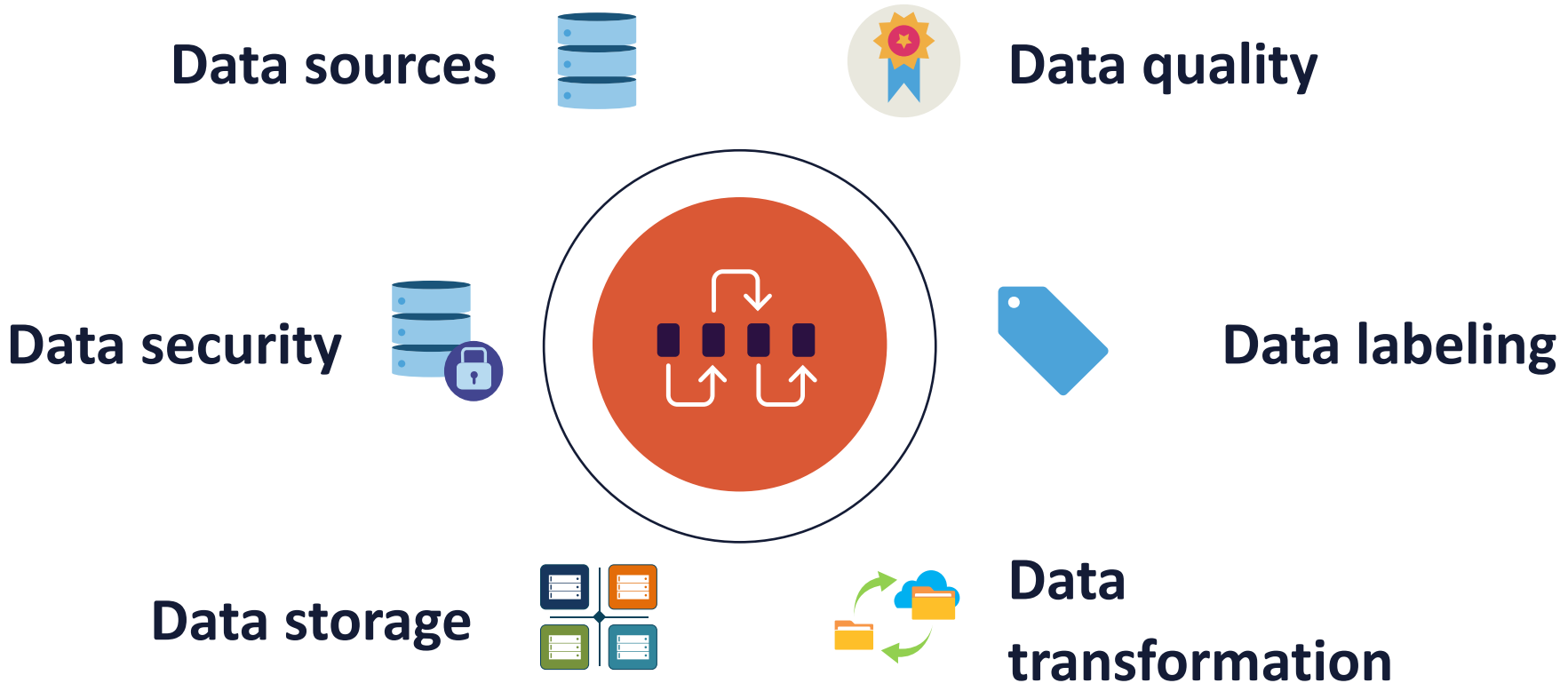
Objectives

- Examine components of an ML pipeline
- Discover sources of ML models
- Describe methods to use a model in production
- Explore SageMaker in an ML pipeline
- Learn about SageMaker Data Wrangler in an ML pipeline
- Examine SageMaker Feature Store in an ML pipeline
- Explore SageMaker Model Monitor in an ML pipeline

Components of an ML Pipeline: Data Collection

- **Data collection** in an ML pipeline is a crucial first step that involves gathering and preparing data from various sources to be used for training and evaluating ML models
- These early phases ensure that the data used in the ML pipeline is of high quality, properly labeled, and securely stored, setting a strong foundation for building effective modeling

Data Collection in a ML Pipeline



Exploratory Data Analysis (EDA)

- Exploratory data analysis is a crucial step in the ML pipeline
- EDA involves analyzing and visualizing data to understand its structure, detect anomalies, and uncover patterns
- This process helps data scientists make informed decisions about data preprocessing, feature engineering, and model selection

Exploratory Data Analysis (EDA)

1

Understanding data

The distribution, relationships between variables, and identifying any missing or outlier values

2

Data visualization

Using charts, graphs, and plots to visualize data makes it easier to spot trends and patterns

3

Feature engineering

Identifying and creating new features that can improve model performance

4

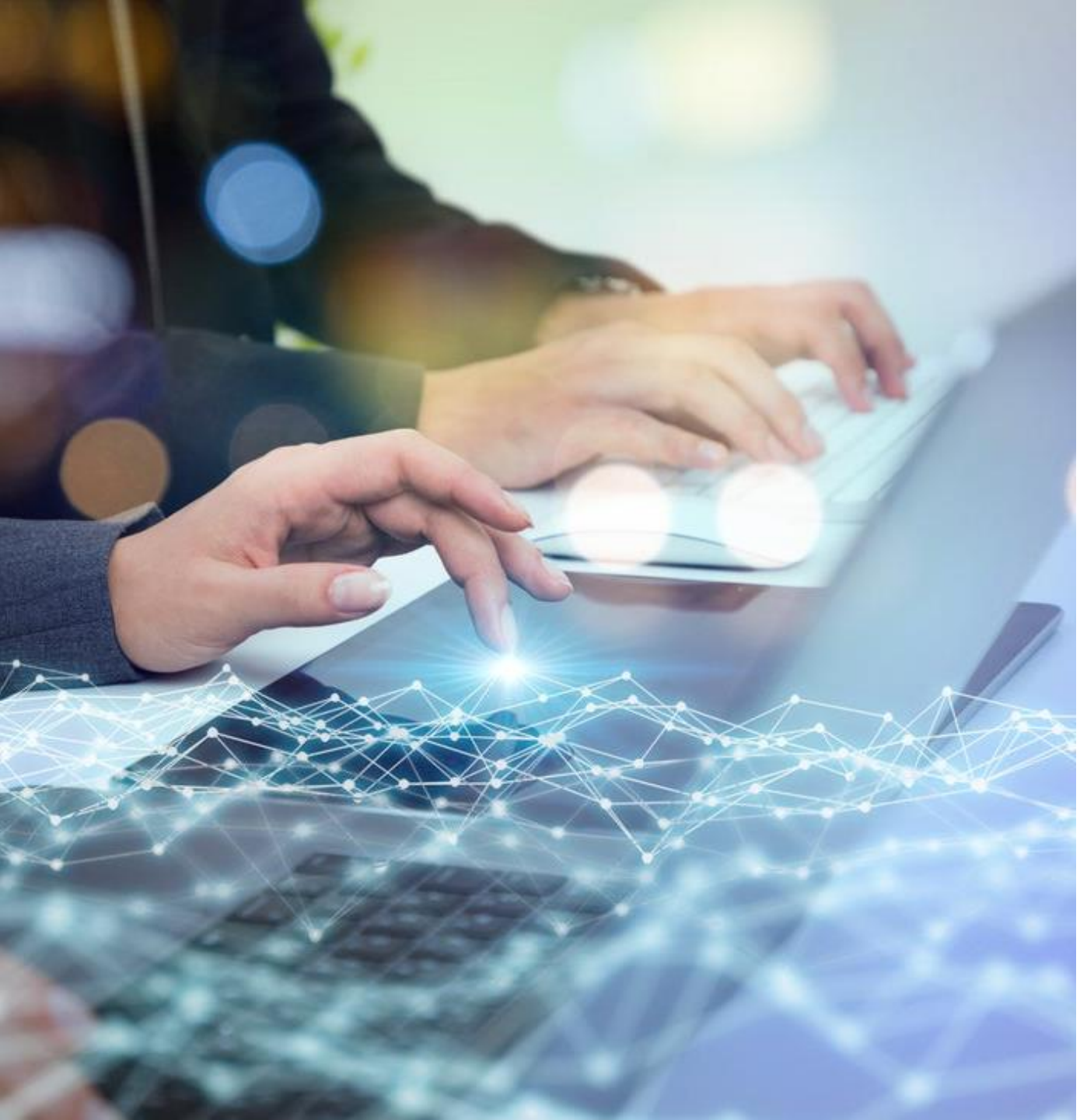
Data cleaning

Detecting and handling missing values, duplicates, and inconsistencies

5

Statistical analysis

Understand the significance of different features and their impact on the target variable



Data Preprocessing

- Data preprocessing involves transforming raw data into a clean and usable format for training ML models
- Data **cleaning** removes or corrects errors, inconsistencies, and missing values in the data
- Data **transformation** normalizes, standardizes, and converts data into a suitable format for analysis

Data Preprocessing

- **Feature engineering** involves deriving new features from existing data (such as calculating the age from a birthdate) and selecting the most relevant features for the model
- Data **integration** combines data from multiple sources into a single dataset
- Data **reduction** reduces the volume of data while retaining its essential characteristics

Model Training

- The model training step in the ML pipeline involves using the prepared and preprocessed data to train a ML model
- This step is crucial as it determines how well the model will perform on new, unseen data

Key Aspects of Model Training

Choose the algorithm

Based on the problem using regression, classification, or clustering

Train the model

Feeding the data into the algorithm and adjusting the model parameters

Hyperparameter tuning

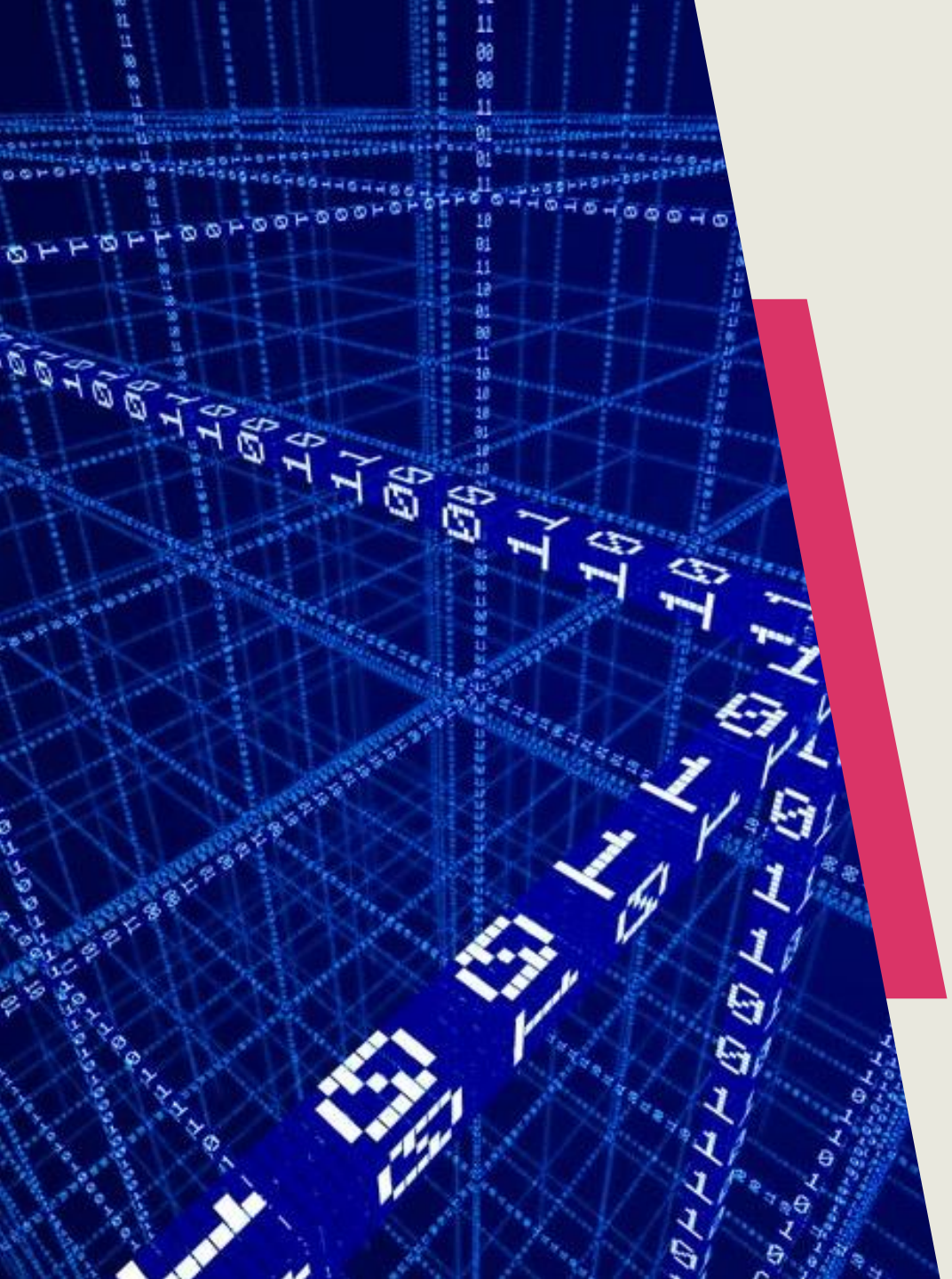
Optimizing the hyperparameters to improve performance

Evaluating the model

Assessing the model's performance using a separate validation dataset

Iterative process

Adjust the preprocessing steps or choose new features or algos



Hyperparameter Tuning

- **Hyperparameters** are the settings that control the training process of the model, such as learning rate, batch size, and regularization strength
  - **Grid search** involves systematically exploring a predefined set of hyperparameter values
  - **Random search** arbitrarily selects hyperparameter values within specified ranges and evaluates them
  - **Bayesian optimization** is an advanced strategy that treats hyperparameter tuning as a regression problem

Common Examples of SageMaker Tunable Hyperparameters

- **Learning Rate**: Controls how much the model adjusts its weights with each training step.
  - A smaller value ensures gradual learning, while a larger value speeds up training but risks overshooting the optimal solution.
- **Batch Size**: Determines the number of training samples used in one iteration.
  - Larger batch sizes can speed up training but require more memory.
- **Number of Layers**: Specifies the depth of a neural network.
  - More layers can capture complex patterns but may lead to overfitting if not managed carefully.

Common Examples of SageMaker Tunable Hyperparameters

- **Number of Trees** (for tree-based models): Defines the number of decision trees in ensemble methods like XGBoost.
  - More trees can improve accuracy but increase computation time.
- **Regularization Parameters**: Includes measurements to prevent overfitting by penalizing large weights.
  - In ML, overfitting happens when a model becomes too tailored to the training data. It memorizes patterns, including noise or irrelevant details, rather than learning generalizable insights.
- **Dropout Rate**: Used in neural networks to randomly deactivate a fraction of neurons during training, reducing overfitting.

Components of an ML Pipeline: Evaluation, Deployment, and Monitoring

Evaluation	Deployment	Monitoring
Involves using a separate dataset, known as the validation or test dataset, to measure how well the model generalizes to new, unseen data	Involves making the trained model available for use in a production environment like real-time predictions or batch processing	Involves tracking the model's performance, detecting anomalies, and making necessary adjustments to maintain accuracy and reliability

Open-Source Pre-Trained Models

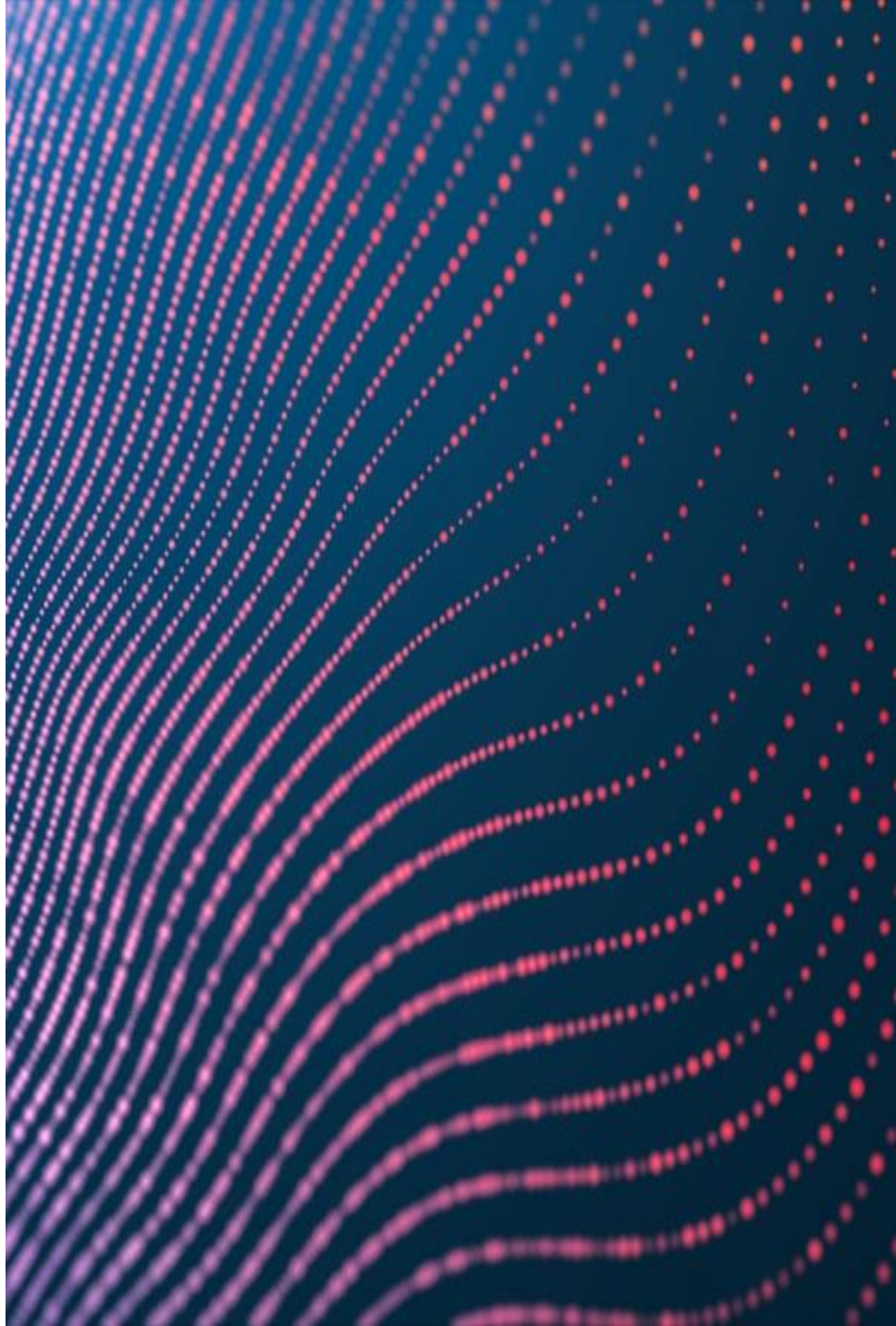
- Open-source pre-trained models are ML models that have already been trained on large datasets and are made available to the public
- These models can be used as-is or be fine-tuned for specific tasks
- Pre-trained models are available from various sources such as SageMaker JumpStart and the AWS Marketplace

Training Custom Models

- Training custom models involves building and training a model from scratch or using a pre-trained model as a starting point
- Training custom models is beneficial when the task is highly specific or when there is a need for a model tailored to unique data
  - This approach provides more control over the model architecture and training process but requires more time and resources

Training Custom Model Steps

1. **Data collection**: gathering and preparing the dataset that the model will be trained on
2. **Model selection**: choosing the right algorithm and model architecture based on the problem at hand
3. **Hyperparameter tuning**: optimizing the hyperparameters to improve its performance
4. **Evaluation**: assessing the model's performance using a separate validation dataset and metrics such as accuracy, precision, recall, and mean squared error
5. **Deployment**: deploying the trained model to a production environment where it can make predictions on new data



Methods to Use a Model in Production: Managed API Service

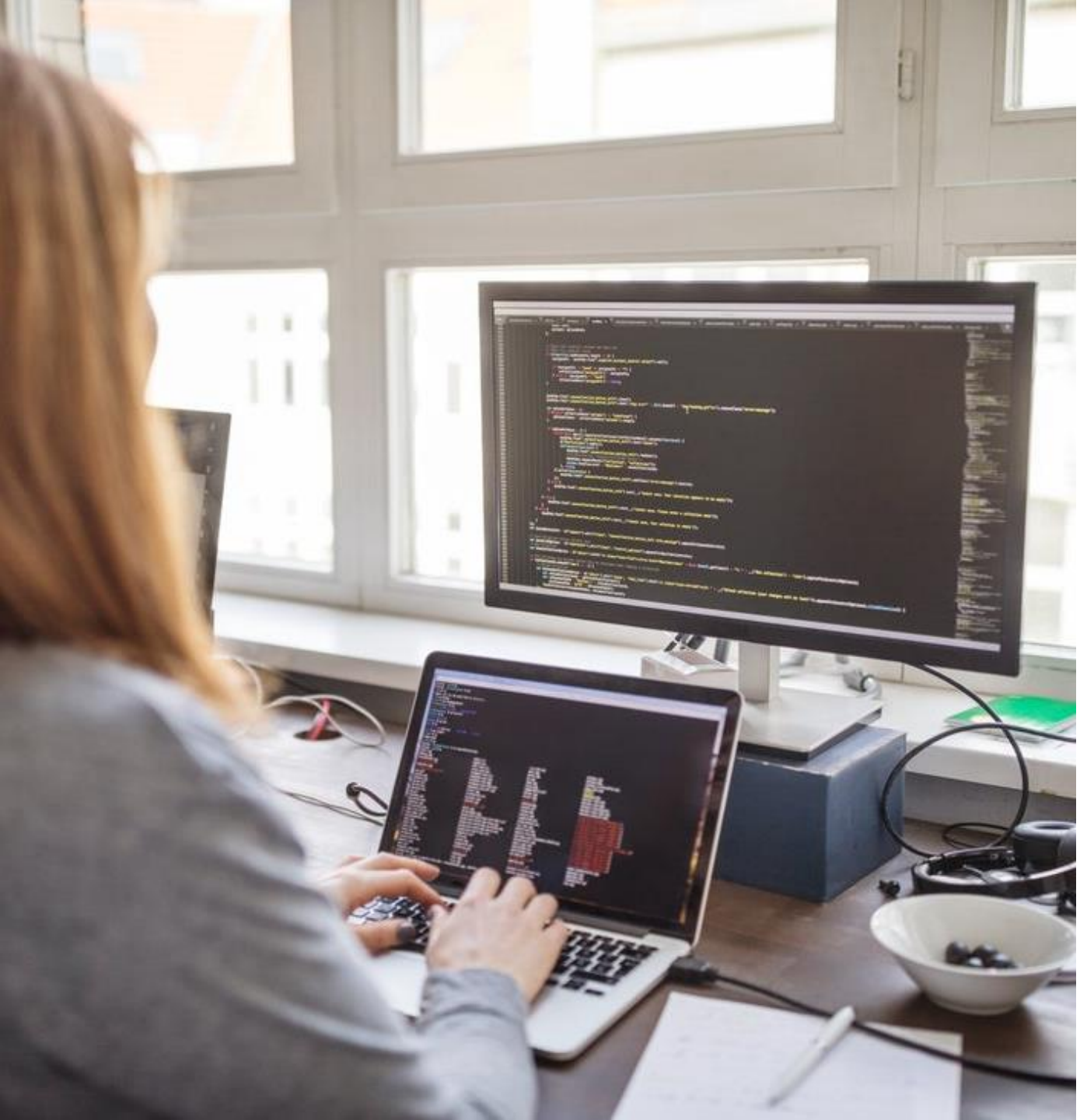
- A managed API service involves deploying the model using a cloud provider's infrastructure, which handles the underlying resources, scaling, and maintenance
- AWS offers services like Amazon SageMaker for this purpose

Managed API Service Features

- Automatically scales the infrastructure based on the incoming traffic, ensuring consistent performance
- Simplifies the deployment process with minimal setup and configuration
- The cloud provider handles updates, security patches, and infrastructure management
- Easily integrates with other cloud services, such as data storage and monitoring tools

Methods to Use a Model in Production: Self-Hosted API

- A self-hosted API involves deploying the model on your own infrastructure, such as on-premises servers or virtual machines (VMs)
  - An on-premise private cloud deployment
- This method provides more control over the deployment environment but requires more effort to manage and maintain



Self-Hosted API Features

- Full control over the deployment environment, including hardware, software, and network configurations
- Ability to customize the deployment setup to meet specific requirements
- Potentially lower costs if you have existing infrastructure, but requires managing resources and scaling manually
- Greater control over security measures, but also requires implementing and maintaining them

SageMaker in an ML Pipeline

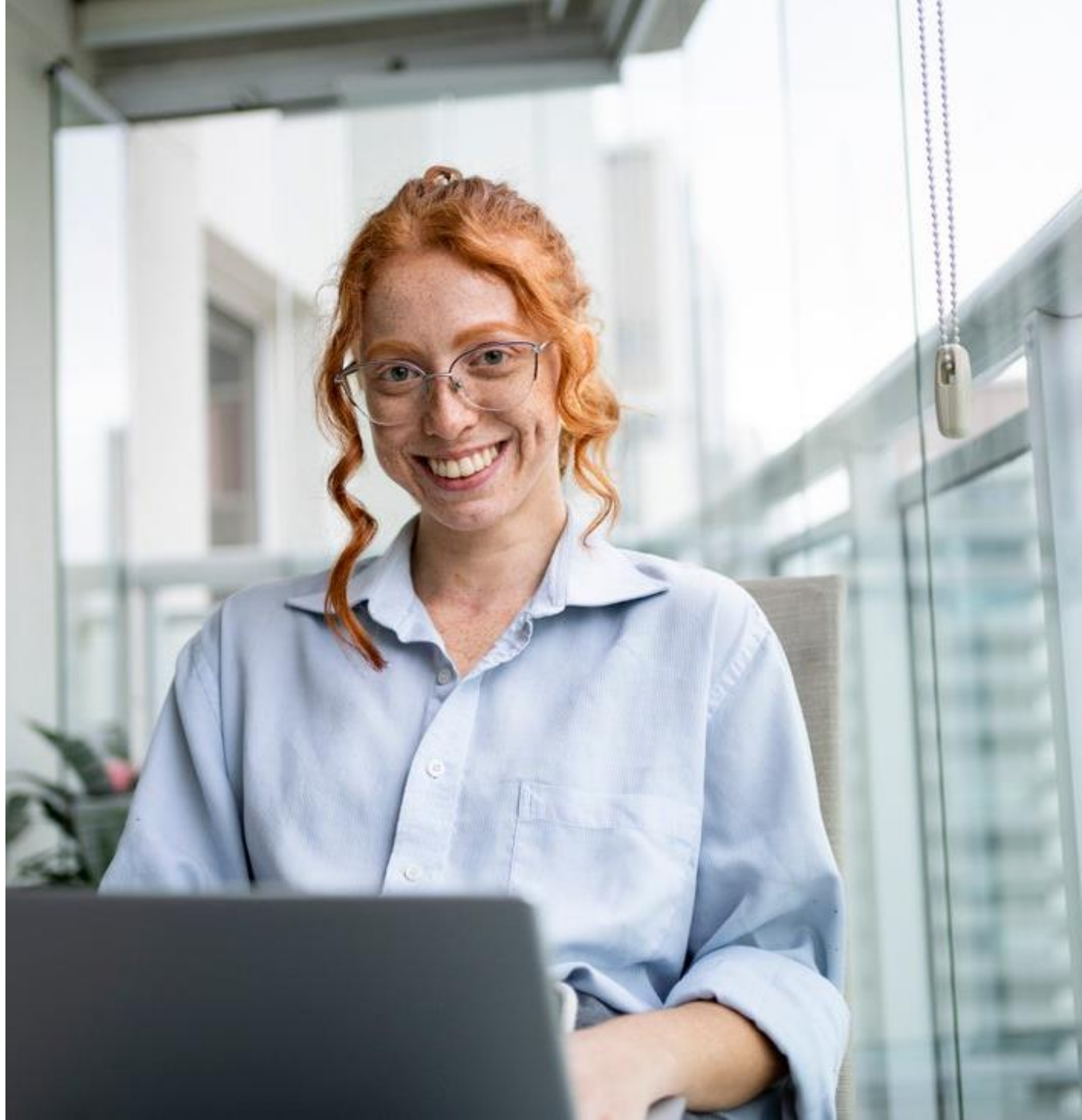
- Amazon SageMaker is a comprehensive service that supports the entire ML pipeline, from data preparation to model deployment and monitoring
- SageMaker offers a robust set of tools and services to support each stage of the ML pipeline, making it easier for data scientists and developers to build, train, deploy, and monitor ML models efficiently

1. Data Collection and Preparation

- **Amazon SageMaker Data Wrangler:** simplifies the process of data preparation and feature engineering with a visual interface
  - It allows you to import data from various sources and clean, transform, and visualize it
- **Amazon SageMaker Ground Truth:** provides tools for creating and managing labeled datasets, which are essential for supervised learning tasks

2. Exploratory Data Analysis (EDA)

- **Amazon SageMaker Studio:** offers a fully integrated development environment (IDE) for data scientists to perform EDA
  - It includes tools for data visualization, statistical analysis, and feature engineering



3. Model Training

- **Amazon SageMaker training:** provides scalable infrastructure for training ML models
  - It supports various built-in algorithms, custom algorithms, and frameworks like TensorFlow, PyTorch, and MXNet
- **Amazon SageMaker Autopilot:** automates the process of training and tuning ML models, making it easier for users to build high-quality models without deep ML expertise

4. Hyperparameter Tuning

- **Amazon SageMaker automatic model tuning** is also known as hyperparameter optimization (HPO)
- This **feature automates the search for the best hyperparameters** by running multiple training jobs with different combinations and selecting the best-performing model

5. Model Evaluation

- **Amazon SageMaker Model Monitor:** continuously monitors the performance of deployed models to detect data drift and other issues
  - It helps ensure that models remain accurate and reliable over time

6. Model Deployment

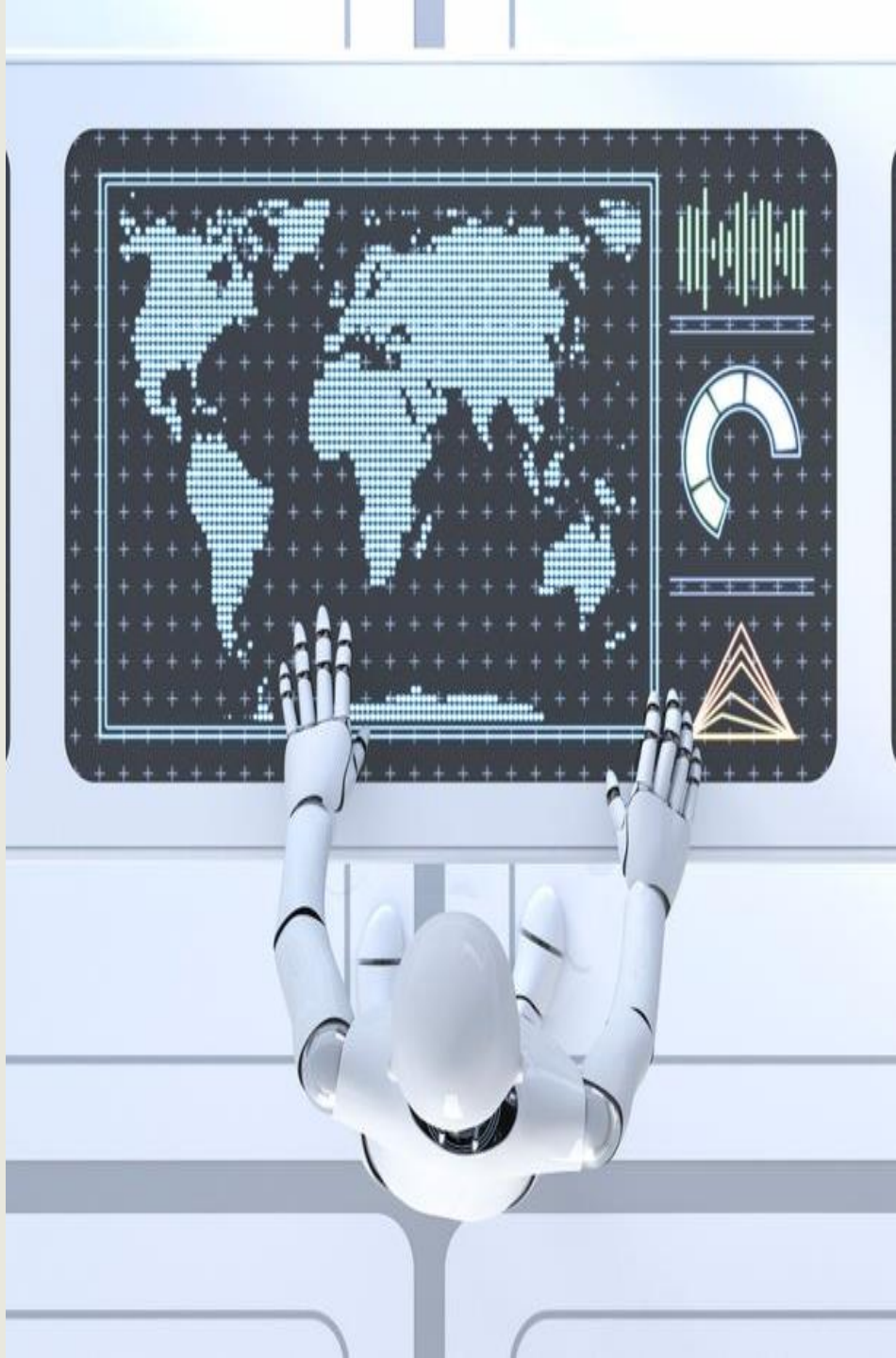
- **Amazon SageMaker hosting services:** provides a fully managed environment for deploying ML models as endpoints
  - These endpoints can be accessed via RESTful APIs for real-time predictions
- **Amazon SageMaker multi-model endpoints:** allows you to deploy multiple models on a single endpoint, optimizing resource usage and cost

7. Model Monitoring

- **Amazon SageMaker Model Monitor:** tracks the performance of deployed models, detects anomalies, and provides alerts
  - It ensures that models continue to perform well and meet business requirements

8. MLOps and Automation

- **Amazon SageMaker pipelines:** a workflow orchestration service that automates the end-to-end ML lifecycle
  - It integrates with other AWS services to streamline data processing, model training, evaluation, deployment, and monitoring
- **Amazon SageMaker Model Registry:** centralizes model tracking and versioning, simplifying the deployment and management of models



SageMaker Data Wrangler in an ML Pipeline

In this demo…

We will explore the capabilities SageMaker Data Wrangler in an ML Pipeline.