Module Summary

Congratulations! You have completed this module. At this point in the module, you know:

- Machine learning, a subset of AI, uses algorithms to analyze data, makes decisions without explicit programming, and enables autonomous problem-solving.

- There are three main types of machine learning:

    - Supervised learning:

        - Trained on labels to classify new data

        - Becomes more precise with more data

        - Categories:

            - Regression: Estimates continuous values

            - Neural networks: Imitates the structure of the human brain

            - Classification: Focusses on discrete values

    - Unsupervised learning:

        - Finds patterns in unlabeled data

        - Useful for clustering similar data points and detecting anomalies

    - Reinforcement learning:

        - Achieves goals within a set of rules and constraints by maximizing rewards

        - Useful for tasks like playing chess or navigating

- Training a model involves splitting a dataset into training, validation, and testing sets.

    - Training set: Trains the algorithm

    - Validation set: Fine-tunes and validates the model

    - Test set: Evaluates the model's performance

- Deep learning uses neural networks with multiple layers to analyze complex data. It allows continuous improvement and learning. It enhances AI's natural language understanding by grasping context and intent. It excels in various tasks:

- Image captioning
- Voice recognition
- Facial recognition
- Medical imaging
- Language translation
- Driverless cars

- Neural networks, a computational model consisting of interconnected nodes with three layers. They are an input layer, one or more hidden layers, and an output layer.

- Various types of neural networks include perceptron, feed-forward, deep feed-forward, modular, convolutional neural network, and recurrent neural networks.

- Generative AI model architectures include VAEs, GANs, autoregressive models, and Transformers.

- Variational autoencoders (VAEs):
  - Encoder: Input data into a latent space representation
  - Latent space captures essential data characteristics
  - Decoder: Generates outputs based on this representation

- Generative adversarial networks (GANs):
  - Generator: New data samples
  - Discriminator: Verify the generated data

- Autoregressive models:
  - Create data sequentially
  - Consider the context

- Transformers:
  - Generate text sequences
  - Perform cross-language translations effectively

- Models can be categorized into unimodal or multimodal. Unimodal models process inputs and generate outputs within the same modality and multimodal models handle inputs from one modality and produce outputs in a different modality.

- Cognitive computing technology mimics human cognitive processes like thinking, reasoning, and problem-solving.

- Natural language processing aids computers in interpreting and producing human language and uses machine learning and deep learning algorithms to understand a word's semantic meaning by deconstructing sentences grammatically, relationally, and structurally.

- Speech-to-text (STT) technology transforms spoken words into written text.

- Text-to-speech (TTS) technology converts written text into spoken words.

- Computer vision enables machines to understand visual data by analyzing images or videos, drawing meaningful insights, and making informed decisions.

- IoT devices are a network of physical devices connected to the internet that collect and share data for processing and analysis.

- Cloud computing allows you to store and use data and services over the internet.

- Edge computing refers to the practice of processing data closer to the source of generation, rather than relying on a centralized data center.

- The intersection of AI, IoT, cloud computing, and edge computing brings you smart and real-time applications.

- Real-world applications of AI, cloud computing, edge computing, and IoT can include AI-powered traffic lights, smart public transportation, smart agriculture, and smart buildings.