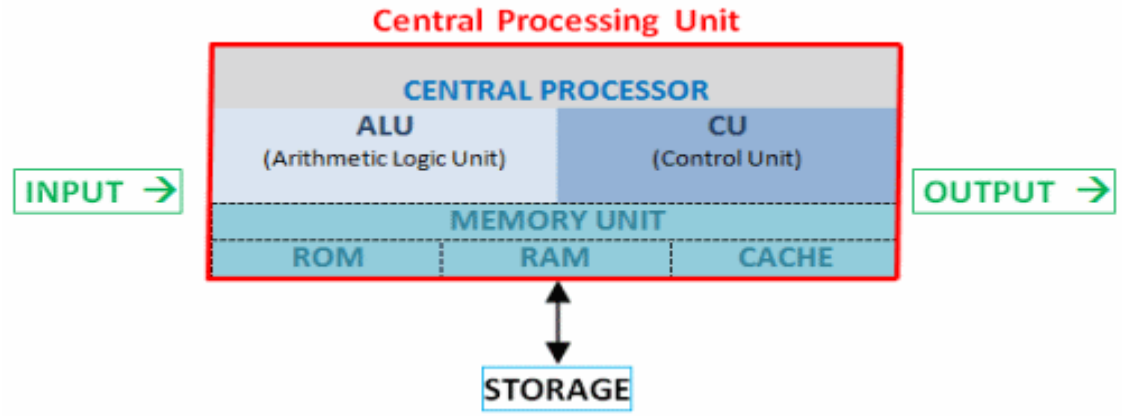
Hardware used in building AI applications

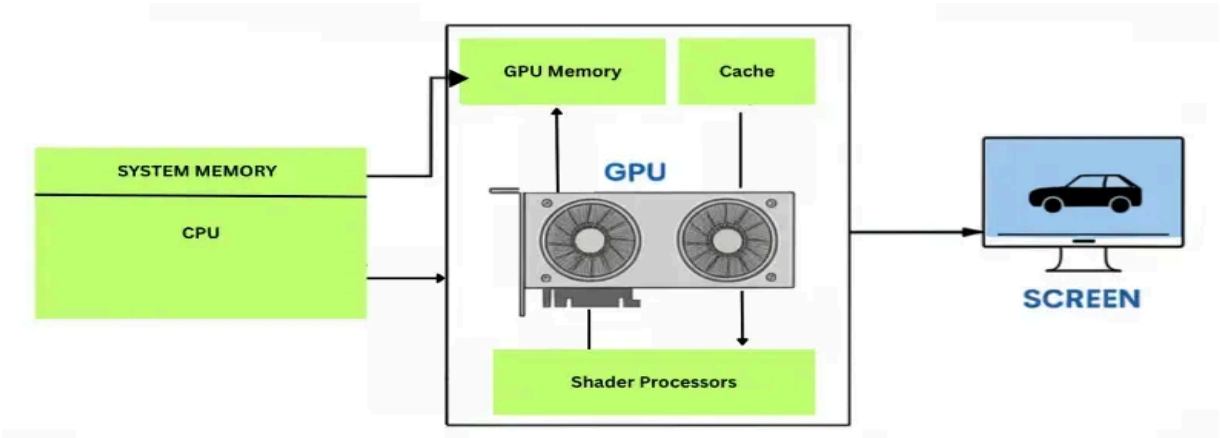
AI application hardware utilizes a mix of specialized processors for speed, ample VRAM/RAM for data handling, and fast storage for massive datasets. GPUs and TPUs drive heavy training, while NPUs handle efficient edge inference, supported by high-speed.

(Compute Engines)

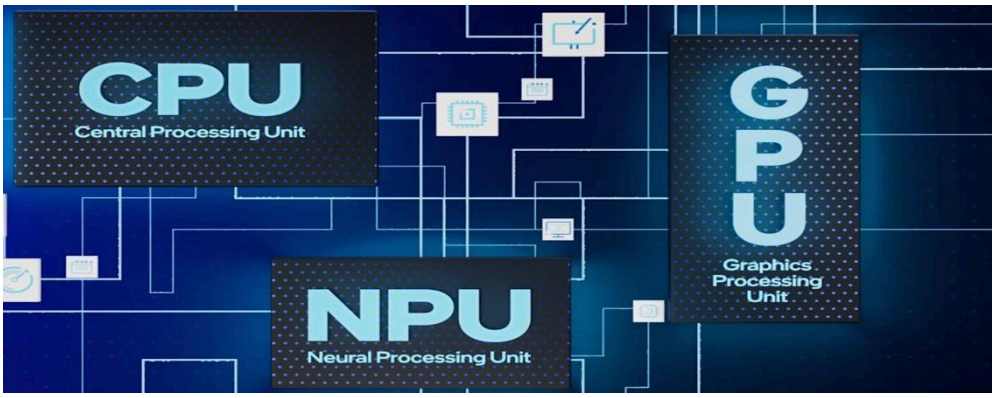
- **CPU (Central Processing Unit):** The Central Processing Unit (CPU) is the primary component and "brain" of a computer, responsible for interpreting and executing most commands from hardware and software. It processes data, performs calculations, and manages data flow, allowing computers to run operating systems and applications.



- **GPU (Graphics Processing Unit):** A Graphics Processing Unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images, videos, and animations. Unlike general-purpose CPUs, GPUs feature massive parallel processing architectures with thousands of smaller, efficient cores, making them ideal for rendering, gaming, video editing, and machine learning workloads.



- **TPU (Tensor Processing Unit):** Tensor Processing Units (TPUs) are Google's custom-developed, application-specific integrated circuits (ASICs) designed to accelerate machine learning workloads, particularly neural networks.
- **NPU (Neural Processing Unit):** A Neural Processing Unit (NPU) is a specialized microprocessor designed to accelerate artificial intelligence (AI) and machine learning tasks by simulating the human brain's neural network architecture.



Memory (Data Access)

- **RAM (Random Access Memory):** Random Access Memory (RAM) is a computer's high-speed, short-term, volatile memory used to store data currently being processed by the CPU, allowing for rapid read/write access. It serves as the main memory, enabling faster system performance, but loses all stored information when power is turned off.
- **VRAM (Video RAM / HBM):** Video Random Access Memory (VRAM) is a specialized, high-speed type of RAM dedicated to a graphics card (GPU) that stores image data, textures, and frame buffers for rendering display output.

Storage (Data Persistence)

- **SSD (Solid State Drive):** Solid State Drives (SSDs) are fast, reliable storage devices that use flash memory to provide significantly better performance than traditional hard disk drives (HDDs).

- **HDD (Hard Disk Drive):** Hard Disk Drives (HDDs) are critical in building AI applications, acting as the primary, cost-effective storage foundation for massive datasets ("data lakes"). While Solid State Drives (SSDs) are used for "hot" (frequently accessed) data, HDDs are essential for storing the vast, "warm" data that feeds AI model training.

Platforms for building applications using AI

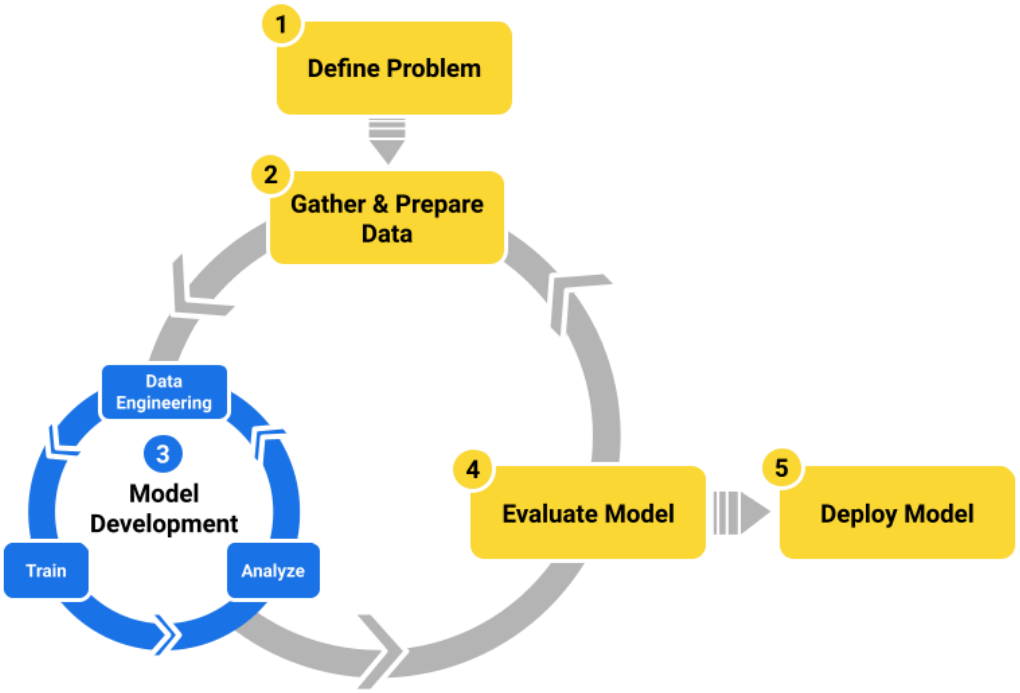
Platform for building AI applications are typically divided into **Online (Cloud-native)** services for scalability and **Desktop (Local)** tools for data privacy and specialized research.

1. Online Platforms (AutoML & Cloud-based)

AutoML: AutoML is a process of automating certain tasks in a machine learning workflow.

These repetitive tasks include:

- **Data Engineering**
  - Feature engineering.
  - Feature selection.
- **Training**
  -
  - Identifying an appropriate ML algorithm.
  - Selecting the best hyperparameters.
- **Analysis**
  - Evaluating metrics generated during training based on test and validation datasets.



Online platforms typically offer high scalability and integrated cloud storage, making them suitable for large enterprise projects and production deployment.

- **Google Cloud AutoML:** A suite of tools within Google Cloud that automates model building for computer vision, natural language processing, and tabular data. It is widely used for creating custom models for image and text classification without deep programming expertise.
- **H2O.ai (H2O AutoML / Driverless AI):** A major enterprise platform known for automated feature engineering and model interpretability. In 2026, it offers comprehensive support for generative AI, including no-code training for small language models (SLMs) through H2O LLM Studio.
- **DataRobot:** An enterprise-grade platform that automates the entire machine learning lifecycle, from data preprocessing to model monitoring. It is favored by large organizations in finance and healthcare for its robust visualization and explainability features.
- **Amazon SageMaker Canvas:** A visual, point-and-click interface within AWS that allows business analysts to generate predictions without writing code. It integrates directly with various cloud and on-premise data sources.
- **Microsoft Azure Machine Learning:** Provides a drag-and-drop studio for building ML pipelines. It is highly integrated with the broader Microsoft ecosystem, including Power BI and Azure Data Lake.

2. Desktop Platforms (No-code / Low-code Data Mining)

Desktop platforms often use a visual "node" or "widget" approach, where users connect different tasks (e.g., data loading, cleaning, and modeling) into a graphical workflow.

- **Orange Data Mining:** A free, open-source platform known for its colorful, widget-based interface. It is highly effective for interactive data visualization, teaching, and prototyping with small to medium datasets.
- **KNIME (Konstanz Information Miner):** A professional-grade, open-source desktop workbench that uses a modular node-based system. It excels in complex ETL (Extract, Transform, Load) tasks and can be extended with scripts in Python or R for deeper customization.
- **Weka:** Developed by the University of Waikato, this Java-based open-source suite provides a straightforward GUI for classical machine learning algorithms. It is widely used in academic settings for learning data mining fundamentals.
- **IBM SPSS Modeler:** A powerful desktop-based tool for predictive modeling that uses a drag-and-drop interface. While it is a paid product, it is favored by business analysts for handling large datasets and advanced mining techniques without coding.
- **RapidMiner:** A comprehensive data science platform that offers a no-code visual workflow designer alongside automated model building features.

What Is Edge AI?

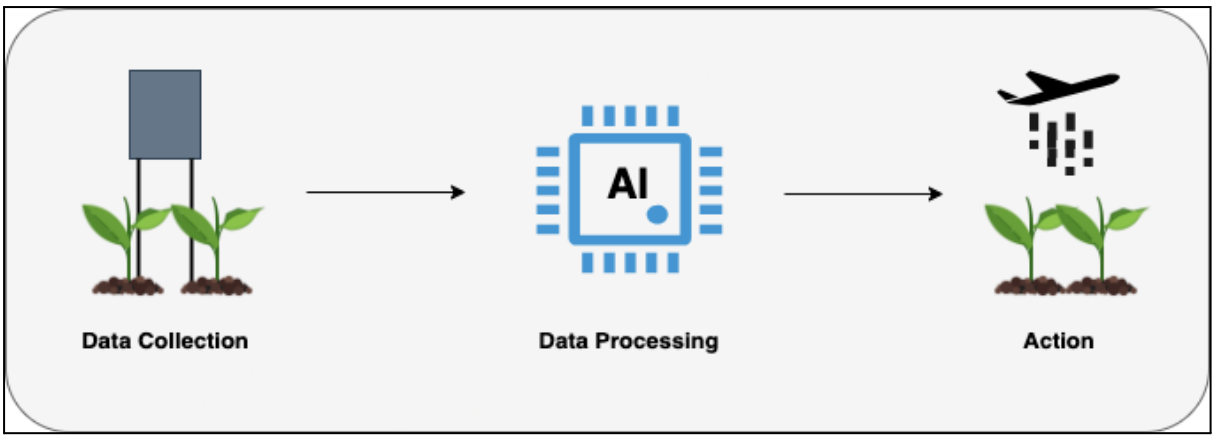
Edge AI is the practice of deploying AI models and algorithms directly on edge devices, which are devices located at the network's periphery, where data is generated and actions need to be taken.

Edge AI offers several advantages:

- Speed
- Privacy
- Reliability
- Efficiency

The Process

The workflow that powers edge AI involves three steps: data collection, data processing, and action.



Data collection

Edge devices continuously collect data from sensors, cameras, or other sources, providing a steady stream of information. This data can range from environmental metrics and health parameters to video feeds and audio recordings, forming the basis for real-time analysis.

A great example of data collection is how your smartwatch collects the number of steps you took today.

Data processing

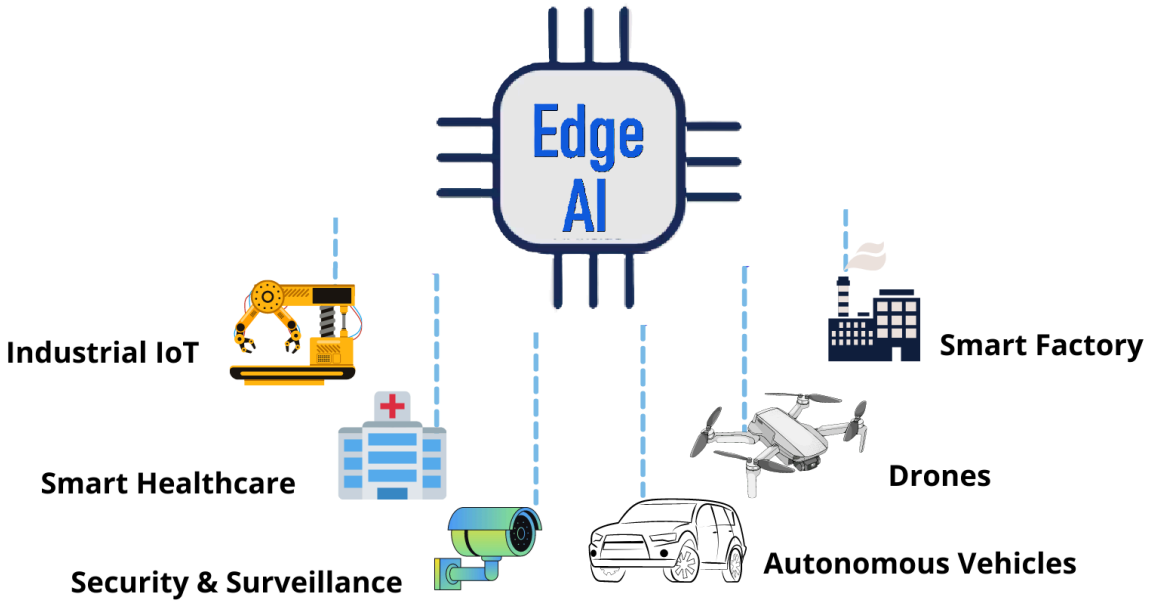
AI models deployed on edge devices process the collected data locally. This step involves analyzing the data to extract meaningful insights, detect patterns, and make predictions using AI models without relying on cloud resources.

Local processing ensures that decisions can be made quickly, such as a self-driving car determining which lane to choose in real-time.

Real-Time Action

Based on the AI model's output, edge devices can take immediate action. These actions might include triggering alarms, adjusting the path, or sending data to the cloud for further analysis. The ability to act in real-time is essential for scenarios requiring instant responses, such as security systems or medical devices.

Edge AI processes data directly on devices—smartphones, cameras, wearables—rather than relying on the cloud, enabling real-time, private, and offline functionality. Key daily applications include instant voice assistant responses, predictive health monitoring, smart home automation, and enhanced security via on-device video analysis.



Chapter-2: Foundations of Data - Types, Ethics & Utility in Building Applications using AI

2.1 Importance of data in building AI applications

2.1.1 Data as the fuel for AI

2.1.2 Role of big data in training AI models

2.2 Conceptual Foundations of Data

2.2.1 Data vs. Information vs. Knowledge

2.3 Structure of Data

2.3.1 Structured

2.3.2 Semi-Structured

2.3.3 Unstructured Data

2.4 Modalities of Data

2.4.1 Text

2.4.2 Image

2.4.3 Audio

2.4.4 Video

2.4.5 Tabular

2.4.6 Time-Series

2.4.7 Spatial Data

2.4.8 Haptic Data

2.5 Formats of Data

2.5.1 Text Formats (CSV, JSON, XML)

2.5.2 Image Formats (JPEG, GIF, PNG)

2.5.3 Audio/Video (MP3, WAV, MP4, AVI)

2.6 Data Repositories

2.6.1 Definition of public Datasets

2.6.2 Definition of private Datasets

2.6.3 Importance of Public Datasets

2.6.4 Popular Public Dataset Repositories (Example - Kaggle, Hugging Face Datasets, UCI Machine Learning Repository, Google Dataset Search or similar ones)

2.6.5 Dataset licensing

2.6.6 Usage Rights

2.7 Ethics, Privacy in Data Usage

2.7.1 Privacy concerns related to data usage

2.7.2 Regulations governing data usage - GDPR, HIPAA (Overview)

2.7.3 Ethical use of data

2.7.4 Responsible AI data practices