Speak the Language of AI: Essential Core concepts / Terminology / Buzzwords

AI Fields Breakdown

What is Data Science ?	Extracts knowledge from data. It involves collecting, cleaning, analyzing data, and using it to solve problems or make predictions. Data Science utilizes various tools, including **Machine Learning**. Data Science is like a big umbrella that covers the entire data journey, from collecting raw data to extracting meaningful insights and using them to solve problems.
Data Analysis	Focuses on understanding data and communicating its insights clearly. It involves cleaning, transforming data, and creating visualizations to reveal patterns and trends. Data Analysis is a core skill within Data Science. Data analysis is a crucial step within this journey. It focuses on exploring, understanding, and presenting data in a clear and informative way. Analysts use various techniques like visualization to reveal patterns and trends within the data
AI	AI is the overarching concept/field of artificial intelligence, where Machine Learning, Deep Learning, and Generative AI are all specific approaches/areas that fall under this umbrella.
ML	Allows computers to learn from data without explicit instructions for every situation. Machine learning algorithms can then make predictions or identify patterns based on the data they've learned from.
Deep Learning	Deep learning is a subfield of machine learning inspired by the structure and function of the human brain. It excels at handling complex data formats like images, videos, and natural language, analyzing and understanding patterns within this data. Some deep learning models can even generate new content based on the patterns they learn.

Generative AI	Generative AI is a subfield of artificial intelligence focused on creating new data, such as images, text, or music. It utilizes various techniques, including deep learning transformer models, to learn patterns from existing data. These learned patterns are then used to generate entirely new content that is similar to, but not identical to, the training data.
Computer Vision	Computer vision is a subfield of artificial intelligence that focuses on enabling computers to interpret and understand visual information from the real world, such as images and videos. It involves tasks like object detection, image classification, and pattern recognition. While some applications can generate new images or videos, this isn't the primary focus. Computer vision is widely used in self-driving cars, facial recognition, medical imaging analysis, and many other applications.
Natural Language Processing	NLP, which stands for Natural Language Processing, is a field within computer science (and often overlaps with Artificial Intelligence) that focuses on enabling computers to understand and process human language. It's not a single concept or algorithm, but rather a collection of techniques that allow machines to analyze, manipulate, and even generate human language.
Robotics	Robotics is a field of engineering focused on the design, construction, operation, and application of robots. While robots can function without AI, AI is increasingly used to enhance their capabilities, such as robot perception, decision-making, and movement control.
Explainable AI (XAI)	Explainable AI (XAI) is a subfield of artificial intelligence focused on making AI models more understandable. It aims to explain how AI models arrive at decisions or predictions. This helps developers

	understand how the model works and identify potential biases.
AI Ethics	AI Ethics is a field of study concerned with the ethical implications of developing and using artificial intelligence. This includes issues like potential biases in algorithms, fairness in decision-making, privacy concerns, and ensuring responsible development of AI.

<u>Data Science Terminology:</u>

Data	Raw information collected from various sources.
Dataset	Collection of related data's or data points for further analysis or exploration.
Data Wrangling	Data wrangling is the process of shaping raw data into a usable format for analysis. This involves tasks like data cleaning (removing errors and inconsistencies) and data transformation (formatting and manipulating the data)
Data Cleaning	The process of preparing data for analysis by removing the errors, inconsistencies and missing values in the data
Data Analysis	Data analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.
Data Visualization	Creating visual representation to showcase the insights of data and its valuable relations effectively
Big Data	Data visualization is the process of creating graphical representations of data to communicate information clearly and reveal patterns or trends.
Data Mining	Data mining is the process of extracting and uncovering hidden patterns and trends within large datasets using various techniques to build predictive models or enhance decision-making.
Data Modeling	Creating the representation of data to understand its relationships and patterns.
Data Sources	Data sources are the origin points from which you obtain data for analysis or processing.

	These can be various formats, including Relational Databases, Data Warehouses , Data Lakes.
Relational Databases	Structured data organized in tables with rows and columns, following predefined schema (example – SQL databases)
Data Lakes	Data lakes are large-scale storage repositories designed to hold a variety of data, including structured, semi-structured, and unstructured data, in its native format. This flexibility allows for storing raw data without worrying about upfront schema definition. Data lakes are often used for big data analytics, exploration, and machine learning tasks where the specific data format may not be known beforehand. Example AWS-S3
Data Warehouses	Centralized repositories that store large amounts of integrated data from multiple sources, often used for data analysis and reporting.

Understanding Data: A Machine Learning Perspective:

Datasets or Inputs Data or Training Data or Training Datasets	In Machine Learning (ML), we use datasets, also referred to as training data, to train our models. These datasets are collections of data points that the model learns from to identify patterns and make predictions.
Column of the data set or attribute or features or properties	Within a dataset, individual pieces of information about each data point are called features. You might also encounter terms like attributes or properties used interchangeably.
Row of the data set or Record or observation or sample or data points	Each individual piece of information in a dataset is called a data point. Terms like record, observation, or sample all refer to the same concept in Machine Learning. These data points hold the specific values for each feature.
Corpus data	Corpus data is a specialized type of text data. Think of it as a large and organized collection of text, like a library of books on a specific topic.
Structured data (Well organized)	Structured data is highly organized and follows a predefined format, often called a schema. This schema defines the specific attributes (data points) and their data types (e.g., number, text) included in the data. This organization makes structured data easy for

	computers to interpret and analyze. (Example – Relational databases – SQL databases)
Unstructured data (less or no organized)	Unlike structured data, unstructured data has no predefined format or schema. This means it doesn't follow rigid organization**. Examples** of unstructured data include audio files, video files, PDFs, images, and social media posts.
Semi Structured data (Partially organized)	Semi-structured data falls between structured and unstructured data. It has some internal organization but doesn't conform to a strict schema like relational databases. This internal organization often relies on tags or markers to identify data elements and establish a hierarchy. **Examples** of semi-structured data formats include XML, JSON, and CSV files (when they include headers).
Labeled Data	Imagine data with a label attached, like a sticky note! This label tells us exactly what the data represents. For **example**, an email might be labeled "spam" or "not spam," an image labeled "cat" or "dog," or a transaction labeled "fraudulent" or "legitimate." Labeled data is crucial for training machine learning models to recognize patterns and make predictions.
Unlabeled Data	Unlike labeled data with its handy sticky notes, unlabeled data is like a box of stuff without labels. It could be audio or video recordings, social media posts, or sensor data. This data lacks predefined labels or categories, making it more challenging for computers to directly interpret. However, unlabeled data can still be valuable for tasks like anomaly detection or exploratory analysis in machine learning.
Categorical Features	Categorical features are a type of feature in a dataset that represents data that can be classified into distinct categories or groups. These categories don't necessarily have a numerical order. **Examples –** Spam/Not Spam, Fraudulent transaction, Cat or Dog image classification.
Continuous Features	Continuous features represent numerical values that can theoretically take on any value within a specific range. Unlike categorical features with predefined categories, continuous features can have an infinite number of possible values within that range.

	Examples – Predicting House Price – Features like house size, no of bedrooms are continuous features | Weather Forecasting – Temperature , humidity features are continuous features.

Deep Learning

Deep Learning	Deep learning, a subset of machine learning, uses artificial neural networks inspired by the structure and function of the human brain. It excels at complex tasks like image classification, generation, and pattern analysis. The Transformer model, a prominent deep learning architecture, has become foundational for large language models and advanced NLP applications.
Artificial Neural Network (ANN)	The inspiration behind deep learning. It mimics the structure and function of the human brain with interconnected layers of artificial neurons.
Activation Function	Activation functions act like guardians in a deep learning network, filtering out noise in the data as it flows between layers. This allows the network to focus on the essential patterns and not get bogged down by irrelevant details. By reducing noise, activation functions help the network learn more efficiently and extract the key features that distinguish, for example, a cat from a dog in an image.
Back propagation	The training algorithm used to adjust the weights in a neural network based on the difference between the predicted and actual output.
Loss function	A mathematical function used during training in machine learning models. It measures the difference between the model's predicted output and the desired outcome (ground truth). The goal is to minimize this loss function

Generative AI

Generative AI	Generative AI utilizes deep learning models to analyze vast amounts of data. By understanding patterns and relationships, it can generate entirely new data formats like text, images, or videos based on prompts. Large Language Models (LLMs) often serve as a core component, providing the foundation for processing and comprehending the training data.
Model	A model is a computer-generated representation, often mathematical in nature, that learns from data. This representation is built for a specific purpose (your "use case") by analyzing a dataset to identify patterns and relationships. The model can then leverage these learned patterns to make predictions about new data or gain insights into the system it represents.
Large Language Models (LLMs)	Train a model on tons of text (billions of words), and you have got a Large Language Model (LLM). These super-powered language tools come in two flavors: general ones for many tasks, or specialized ones trained for a specific field / domain.
Prompts	Prompts are instructions that guide Generative AI models. They can be simple questions, keywords, or detailed descriptions with context. The quality of the prompt significantly influences the output of the Generative AI.
Retrieval-Augmented Generation(RAG)	Imagine we want a large language model (LLM) to be extra careful with its answers. That's where Retrieval-Augmented Generation (RAG) comes in. Instead of relying solely on LLM knowledge, RAG lets it consult external sources like your custom data, APIs, web searches, or databases. This retrieved information adds context to the prompt, allowing the LLM to generate a more refined and accurate response.
Hallucination	This refers to a situation where a Generative AI model creates outputs that are significantly inaccurate, misleading, or entirely fabricated. While some level of creative freedom is expected, hallucinations can be problematic because they lack factual grounding.

Bias in LLM's	Large Language Models (LLMs) can inherit biases from the data they are trained on. This happens when the training data itself reflects real-world biases, leading the LLM to generate outputs that favor certain viewpoints or unfairly represent certain groups.
GANs (Generative Adversarial Networks)	GANs (Generative Adversarial Networks) are deep learning models using two neural networks Generator and Discriminator.Generator creates new data (text, images) mimicking the training data distribution whereas Discriminator analyzes both real and generated data, aiming to distinguish the fakes. This continuous competition refines both networks. The generator improves its ability to create realistic data, while the discriminator sharpens its fake detection skills. This leads to GANs generating highly creative and realistic outputs.
Overfitting	A common problem in machine learning models, including Generative AI models. It occurs when the model memorizes the training data too well, leading to struggles with unseen data (data it wasn't explicitly trained on). Techniques like regularization help prevent overfitting and improve the model's ability to generalize - perform well on new data.
Sampling	The process of selecting a single output from a set of possible outputs generated by the model. Different sampling techniques can influence the creativity and style of the generated content.

Generative AI – Popular buzz words

Langchain	LangChain simplifies the development of Generative AI applications. This open-source framework streamlines the integration of Generative AI clients and agents with Large Language Models (LLMs). It allows you to programmatically submit user prompts to LLMs and even build complex Retrieval-Augmented Generation (RAG) applications, all while hiding the complexities of LLM integration.

Ollama	Ollama provides a curated collection of pre-trained LLMs from various sources, allowing you to download and run them directly on your machine using convenient executables. Also lets you upload and experiment with your custom models. Ollama offers a user-friendly web interface for interacting with these LLMs. Additionally, you can provide contextual inputs like documents or information to enrich the outputs generated by the models.
LangChain Serve	LangChain Serve is an extension of the LangChain framework specifically designed to simplify the creation of APIs (Application Programming Interfaces) that interact with Large Language Models (LLMs)
LangChain Smith	LangChain Smith: Keeping an Eye on LLM Interactions. LangChain Smith is another extension of the LangChain framework. It focuses on monitoring user interactions with Large Language Models (LLMs) within LangChain Serve APIs.
Graphics Processing Units (GPU)	GPUs are powerful processors designed for parallel resource-consuming tasks like video/image rendering and efficiently running Deep Learning algorithms used in LLMs. NVIDIA is a major player in the GPU market and has been a key driver of innovation in the field.
Language Processing Units(LPU)	LPUs are powerful processors designed for parallel processing tasks specifically optimized for natural language processing (NLP) workloads, like the deep learning models used in LLMs. Groq is a major player in the LPU market. LPUs and GPUs can be complementary infrastructure options depending on the specific AI application and its processing needs.