Module

Introduction to AI

Artificial intelligence

Artificial intelligence is the simulation of human intelligence processes by machines.

It is wide-ranging branch of computer science concerned with building smart machines.

Application of AI around us

- Google Search is using AI for recommending a specific part of the video based on your search query
- Google Lens and OCR
- TESLA car's Auto pilot mode has been using AI to utilize computer vision and drive the vehicle.
- Netflix movies recommendation.

AI has been used in different sectors some of examples are as below:

1. **AI in Banking and Finance**
   AI in Banking and Finance helps in reducing operational costs, minimizing risks, enhancing customer experience and aids in effective decision making.

2. **AI in Manufacturing**
   AI can be used to automate complex tasks. The involvement of robots in high-risk jobs can help manufacturers reduce unwanted accidents. It can also be used to optimize manufacturing processes and maintain machines.

3. **AI in HealthCare**
   AI programs are applied to practices such as diagnosis processes, treatment protocol development, drug development, personalized medicine, and patient monitoring and care.

4. **AI in Retail and Supply Chain**
   AI helps retailers understand exactly how their supply chain is operating, make improvements throughout and eliminate waste and overhead.

Types of Data

Data are categorized in 3 types

1. Structured data
2. Semi-Structure data
3. Unstructured data

1. Structured Data

Structured data contains following traits

- It is highly organized and readily searchable by queries or algorithms.
- Can be quickly consolidated into facts.
- Follows a predefined schema.
- Usually resides in fixed fields.
- A typical example is a Relational Database Management System (RDBMS).
- A Schema is defined before the content is created and the data is populated.

Examples of Structured data

- Dates
- Phone Numbers
- Social Security Numbers
- Transaction Information

Applications with Structured Data:

- Airline reservation systems
- Inventory control
- Sales transactions
- ATM activity

2. Semi Structured Data

- This type of data often explained as schema-less or self-describing
- Contains semantic tags, but do not comply with the standards or structure of typical relational databases
- No separate description of the type or structure of the data
- Does not require a schema definition, the definition is not impossible, but it is optional
- Data can have different attributes, and new attributes can be added anytime

Examples of Semi-structured Data include:

- **Markup language XML** - a set of documents encoding rules that define a human- and machine-readable format.
- **Open Standard JSON** - a lightweight, plain-text, data-interchange format based on a subset of the JavaScript programming language.
- **NoSQL** - some noSQL databases contain semi-structured data.

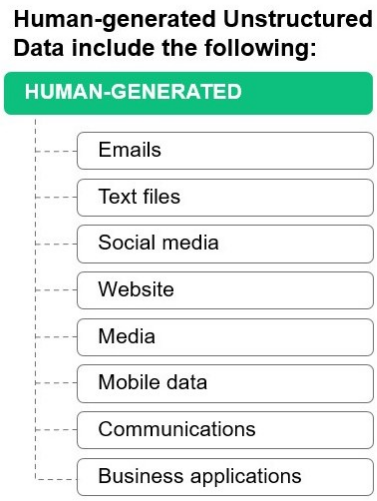
Applications containing semi-structured data include:

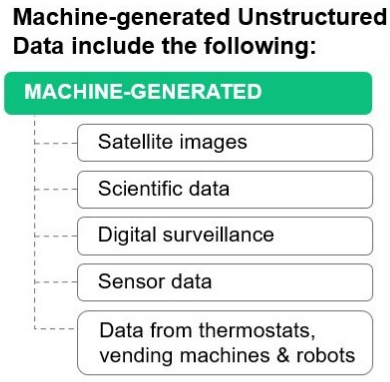
- Big Data Infrastructure.
- Web applications
    - LinkedIn
    - Salesforce
    - Reader recommendations in Amazon

3. Unstructured Data

- No identifiable structure for this kind of data
- Have internal structure, but no predefined schema
- Cannot be stored in rows and columns like a relational database
- No fixed data model, a massive unorganized conglomerate of various information
- Require more storage space than structured data

Examples of Unstructured data

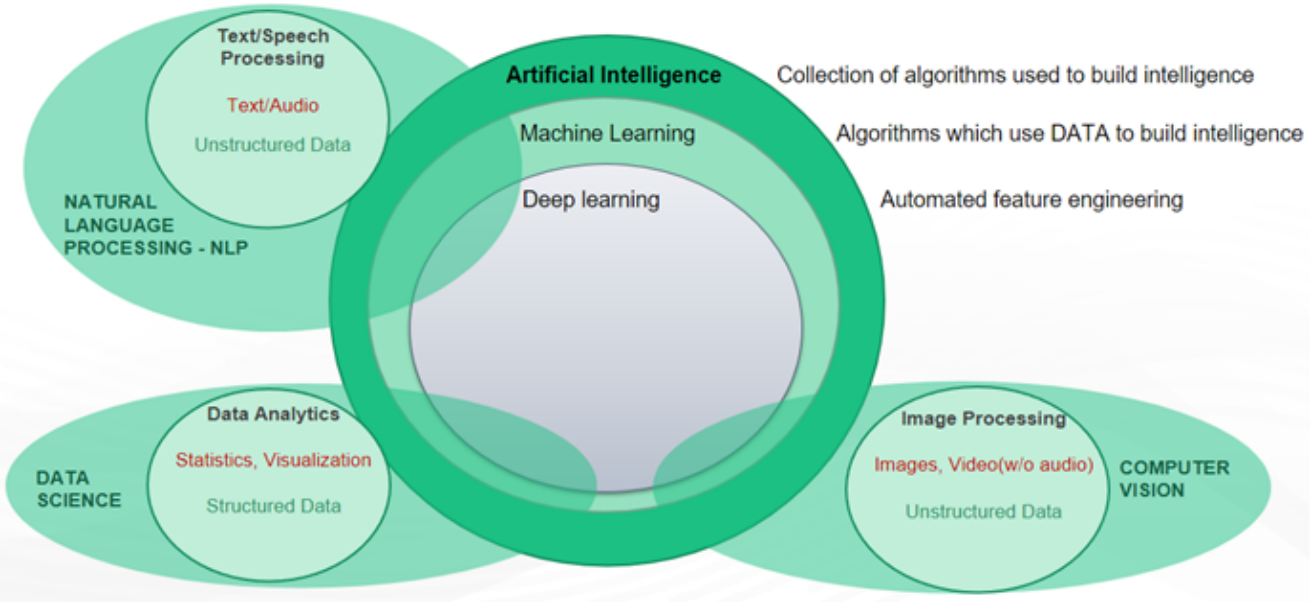




AI Landscape - AI v/s ML v/s Deep Learning v/s Data Science

Artificial Intelligence is wide topic and it has been interconnected with different works and approaches.

We have illustrated below figure to get grasp of its application and involvements in different field.



Data Science based use cases:

Fraud Detection

Emails can be classified as spam or ham on the basis of its content. Gmail uses a classifier to filter emails that are not spam.

Targeted Marketing:

On the basis of a customer's behavior on a website, they can be sent customized ads. Amazon uses this idea to design custom advertisements for its customers.

Computer Vision Based Use cases

Self-Driving Cars

Using computer vision technologies. Tesla has provided an improved self-driving car navigation realizing the goal of making autonomous driving a reality and a reliable transportation option.

Google Lens

It uses computer vision, machine learning and knowledge graph to recognize objects like plants, animals and highlight, copy, paste, translate text from images or documents.

NLP Based use cases

Google Search

Search Engines like Google and Bing are working hard to make the search more natural and relevant using conversational language to ease their users. This type of search is developed to understand the user's intent which is referred to as Natural Language Processing (NLP)

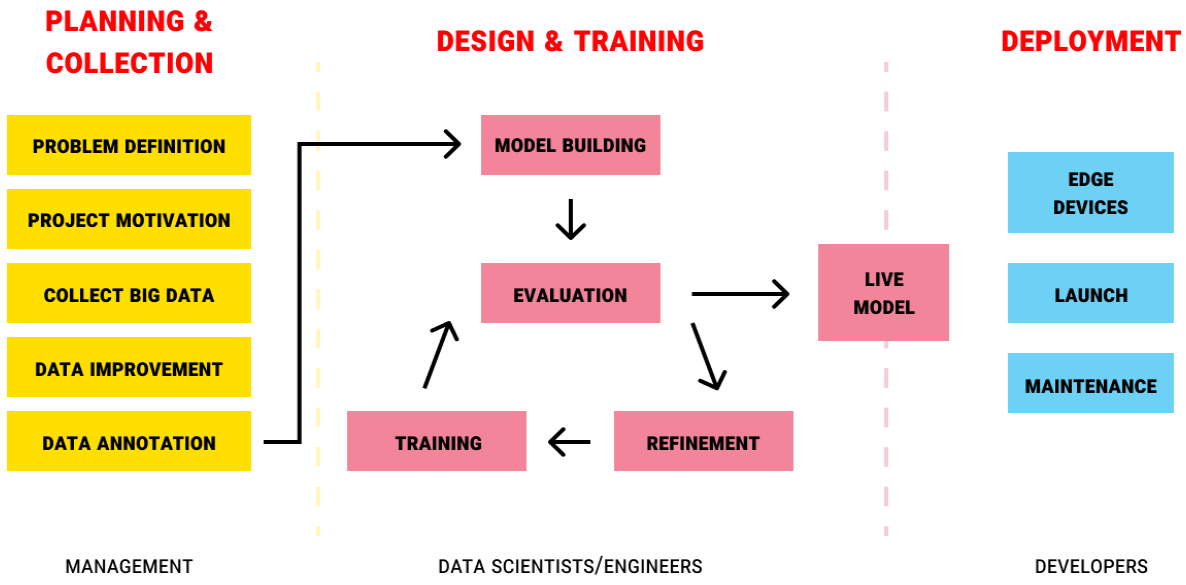
Alexa

The main technology or concept behind Alexa is Natural Language Processing (NLP). NLP helps her understand and process human language. A procedure of converting speech into words, sounds, and ideas.

AI Project Management & Lifecycle

Many experts outline AI lifestyle through the three major project stages: planning and data collection, training of the ML model, and the launch of the algorithm and its maintenance.

Others divide the lifestyle of an ML project into more concrete project steps. This article focuses on the three broad stages of an AI project, as well as the more specific steps that explain each project stage in detail.



Feasibility and Profitability Analysis

1. Technical Feasibility: For any machine learning project one of the important things to analyze is that what level of accuracy(performance) is needed for the solution to be acceptable by the business. Whether with the help of existing technologies, is it possible to achieve that accuracy or not.
   You may want to read white papers, blogs, and news to understand relevant ML approach which have been already implemented and the level of accuracy achieved by others.

2. Training Data — Does training data need to be collected? If so, how much time and money will it cost?
   At times feasibility of an ML project depends on availability of data, many times organizations have to give up the ML solution development because of shortage of qualitative data.

3. Predictive Features — According to domain experts, what factors are likely to predict the target variable? Is that data accessible to you?

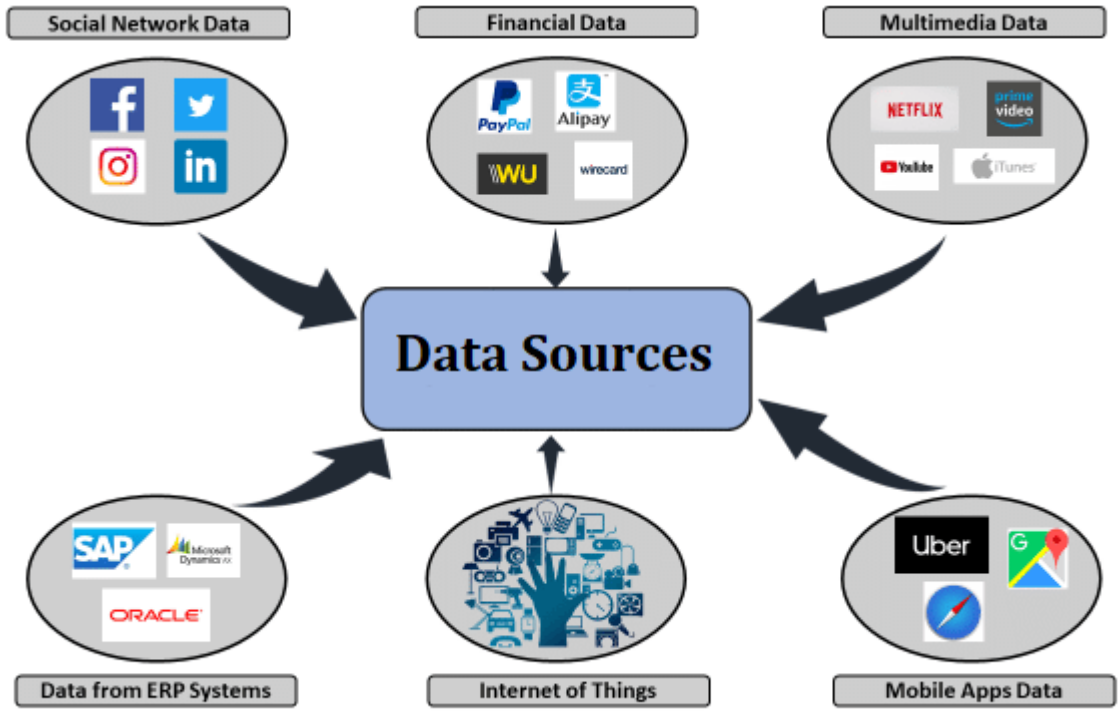
4. Data Sources — What data sources will you need to gain access to? If internal, do you have support from data engineers? If external, how much will vendor data cost?

5. Production — What is the level of effort to develop, deploy, and maintain your model in production?

Identifying right data components

When building/Identifying a data set, make sure that it has 5V characteristics:

- **Volume:** Scalability of data matters. Bigger the data set, better it is for the ML model. Large data set makes it easy for the model to make the most optimal decisions.
- **Variety:** The data set can have different forms of data such as images and videos. Variety in data has significance in ensuring accuracy in results.
- **Velocity**: The speed at which the data is accumulated in the data set matters.
- **Value**: The data set should have meaningful information on it. Maintaining a big data set with valuable information is necessary.
- **Veracity:** Accuracy in data is important while maintaining a data set. Correctness in data means precision in the output received.



Collecting data from outside the room

Following are possible sources of data which can be consumed with in the business and from outside the business.

- Internal data that you already have access to is best.
- Internal data that requires work from dev ops or data engineering to make available to you is next best.
- Data that are not available internally may be available externally from clients, partners, or the government for free.
- Vendors or data aggregators may have the data you need. This kind of data is never cheap, requires long contract negotiations and integration/ingestion time internally, and often suffers from data quality issues hidden during the sales process.
- Web scraping may be worth considering, but has many challenges. Keep in mind that scraping is against the terms of use of most large sites. So even if you build an elaborate siege tower to get past their walls of defense, you're not legally allowed to use the data you pillage.
- If you've made it this far, assume the data is not available. Assess the impact this lack of data will have on the project overall. Is the project likely to succeed without this data?

Size of Data and its impact in AI/ML Project lifecycle

Size of dataset depends on the type of problem you want to solve:

- An image classification problem could require tens of thousands of images or more in order to create a classifier.
- Sentiment analysis or document classification problems can require thousands of examples due to the sheer number of words and phrases, i.e., n-grams.
- For many regression problems, it's suggested that you have 10x as many observations as you do features. A more general rule of thumb is that the number of observations should be proportional to 1/d^p where p = # of features and d = the maximum spacing between consecutive or neighboring data points after each feature is scaled to the range 0-1.
- For time series problems, you should always have more observations than parameters (we elaborate more on this type of machine learning problem below).

these are rough generalizations not intended to be taken as golden rules to follow for every problem. Depending on how correlated different variables are, you may require more or less data. However, to refrain from making some fatal mistakes you should ask a few questions about your data before jumping in and building a forecasting model.

- What's the granularity of my data, e.g., seconds, hours, years. A year's worth of data can imply 365 data points, 52 data points, 12 data points, or even a single data point depending on how the data was recorded, and all are equally valid.
- What are my underlying assumptions about my data? If you expect your data is annually seasonal, make sure you have at least 365 days, 52 weeks, or 12 months of data plus some additional data points for testing- note how important the granularity of data is in this scenario.
- How far out am I trying to predict? If you're trying to predict 12 months into the future, you should have at least 12 months' worth (a data point for every month) to train on before you can expect to have trustworthy results.

Preparing the data for ingestion by an artificial intelligence model. Gathering the right kind and volume of data is the most time-taking task of any data scientist. Like a good recipe, an AI model is only as good as the quality and proportions of its ingredients.

Data preparation remains one of the most underrated aspects of machine learning with much of the effort focused on model design and tuning. Goodness of data is directly proportional to the performance of a machine-learning model and hence, is an important part of any AI/ML application.

Summary

1. Artificial Intelligence is way of making computers/machines imitate human intelligence
2. Machine learning is a subset of AI, uses data to develop intelligence in machines and computers. Primarily it is a way to make computers being able to learn from data.
3. The way industry uses ML is Data Science, Computer Vision and Natural Language Processing.
4. Data can be of type structured, semi-structured and unstructured depending on source and format of data.
5. Examples of structured data can be tabular data in excel, SQL database tables, examples of unstructured data can be images, videos, text and audio files.
6. Data for analytics and AI projects can be collected from internal sources such as ERPs, CRMs, transactional data, external sources such as surveys, social media, web scrapping etc.