Lecture 2 -Data Science, Machine Learning Vs Artificial Intelligence

Alon B. Muhame

September 5, 2024

There are three goals for this set of notes:

1. Discuss the value value of data science, machine learning and more generally identifying settings where these disciplines are "as-if" new norm of doing things. In doing so, we'll touch on the historical and (somewhat) current trends on these concepts.
2. Define a "data science, machine learning and artificial intelligence concepts."
3. Give an introduction to relationship between data science, traditional statistics, machine learning vs Artificial intelligence.

Artificial Intelligence, AI

AI is a powerful tool. Being able to truly complete tasks with less human direction/manipulation allows the human race to do more (by definition) with less and automate the most repetitive task across the board.

And over the past couple of years, we have witnessed the remarkable evolution of AI and it's subsets such as data science, machine learning in data industry practice. From composing intricate pieces of music to generating sophisticated designs and automating customer service responses (e.g., via chatbots on watsApp), these advancements mark a significant step forward in AI technology, showcasing its potential to create immense value across various domains in industries

What is Artificial Intelligence – then and now

AI is the theory and development of computer systems able to perform tasks normally requiring human intelligence. One of the main applications/use cases of AI is in automating repetitive tasks that would typically be done by humans. For example, automating payments or scheduling calendar reminders

This technology has the potential to revolutionize various industries by streamlining operations and increasing efficiency through intelligent automation. AI systems can process vast amounts of data quickly and accurately, making them valuable for tasks ranging from data analysis to autonomous decision-making.

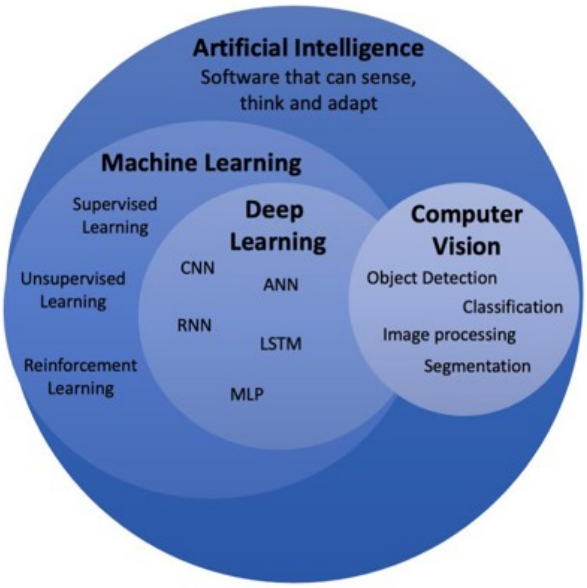

Figure 1: The intersectionality of artificial intelligence concepts summed up

Example 1

Types of AI:
Generative AI:
Purpose: Create new data similar to existing data
Functionality: Generates text, images, music, and more
Examples:
Text: GPT-4 (ChatGPT)
Images: DALL-E
Music: MuseNet
Applications: Content creation, design, entertainment, drug discovery
Key Differences:
Output:
Generative AI: Creates new, original content
Other AI: Analyzes and interprets existing data
Focus:
Generative AI: Innovation and creativity
Other AI: Efficiency and accuracy
Other Types of AI:
Purpose: Analyze data, make decisions, automate tasks
Functionality: Recognizes patterns, predicts outcomes, optimizes processes
Examples:
Predictive AI: Forecasting sales, weather predictions
Analytical AI: Image recognition, speech-to-text
Reactive AI: Self-driving cars, recommendation systems
Applications: Healthcare diagnostics, financial modeling, customer service

What is Machine Learning

Machine Learning (ML) is a subset of artificial intelligence that involves developing algorithms enabling computers to learn from and make predictions based on data. It uses methods like supervised, unsupervised, and reinforcement learning to identify patterns and make decisions without explicit programming. Let's transition into the sub-field of Machine Learning.

ML difference is that it Does not need to be explicitly programed. Meaning it can identify patterns and trends based on data. Moving into a sub-filed of machine learning is Deep Learn Machine Learning is a subset of artificial intelligence that focuses on developing algorithms for computers to learn from data and make predictions. It utilizes supervised, unsupervised, and reinforcement learning methods to identify patterns and make decisions autonomously.

Example 2

Some famous examples of use cases for machine learning include:

- *Predictive analytics: Forescating sales, stock prices, or customer behavior etc*
- *Image and video analytics: Facial recognition, medical imaging diagnostics, and autonomous driving etc*
- *Anomaly detection: Fraud detection in financial transactions and network security monitoring*

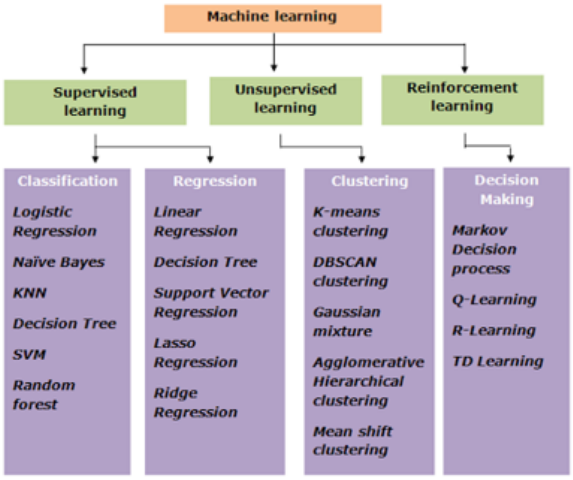

Figure 2: Types of machine learning algorithms

What is data science?

In Wikipedia, Data Science is defined as a scientific field that uses scientific methods to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

I will provide you a definition, with the understanding that this is not the only definition, and that there are many different ways to think about data science. Much of the value in thinking about data science is about being explicit about the assumptions that you are making, and how you are using the data to answer your question [1].

This definition highlights the following important aspects of data science:

Comment 1

- *The main goal of data science is to **extract knowledge** from data, in other words - to **understand** data, find some hidden relationships and build a **model**.*
- *Data science uses **scientific methods**, such as probability and statistics. In fact, when the term data science was first introduced, some people argued that data science was just a new fancy name for statistics. Nowadays it has become evident that the field is much broader.*
- *Obtained knowledge should be applied to produce some **actionable insights**, i.e. practical insights that you can apply to real business situations.*
- *We should be able to operate on both structured and unstructured data. We will come back to discuss different types of data later in the course.*
- *Application domain is an important concept, and data scientists often need at least some degree of expertise in the problem domain, for example: finance, medicine, marketing, etc.*

Since data is pervasive, data science itself is also a broad field, touching many other disciplines.,

[1] Another important aspect of Data Science is that it studies how data can be gathered, stored and operated upon using computers. While statistics gives us mathematical foundations, data science applies mathematical concepts to actually draw insights from data.

Comment 2

- *Databases: A critical consideration is **how to store** the data, i.e. how to structure it in a way that allows faster processing. There are different types of databases that store structured and unstructured data, which we will consider in our course.*
- *Big Data: Often we need to store and process very large quantities of data with a relatively simple structure. There are special approaches and tools to store that data in a distributed manner on a computer cluster, and process it efficiently.*
- *Visualization: Vast amounts of data are incomprehensible for a human being, but once we create useful visualizations using that data, we can make more sense of the data, and draw some conclusions. Thus, it is important to know many ways to visualize information - something that we will cover in Section 3 of our course. Related fields also include **Infographics**, and **Human-Computer Interaction** in general.*
- *Machine learning*
- *Aritificial intelligence*

Types of Data

As we have already mentioned, data is everywhere. We just need to capture it in the right way! It is useful to distinguish between structured and unstructured data. The former is typically represented in some well-structured form, often as a table or number of tables, while the latter is just a collection of files. Sometimes we can also talk about semi-structured data, that have some sort of a structure that may vary greatly.

Example 3

- *Structured | Semi-structured | Unstructured:*
- *List of people with their phone numbers | Wikipedia pages with links | Text of Encyclopedia Britannica*
- *Temperature in all rooms of a building at every minute for the last 20 years |Collection of scientific papers in JSON format with authors | data of publication, and abstract*
- *File share with corporate documents | Data for age and gender for all people entering the builiding | internet pages . Raw video feed from surveilance camera*

Where to get data

Comment 3

There are many possible sources of data, and it will be impossible to list all of them! However, let's mention some of the typical places where you can get data:
Structured:

- *Internet of Things (IoT), including data from different sensors, such as temperature or pressure sensors, provides a lot of useful data. For example, if an office building is equipped with IoT sensors, we can automatically control heating and lighting in order to minimize costs.*
- *Surveys that we ask users to complete after a purchase, or after visiting a web site.*
- *Analysis of behavior can, for example, help us understand how deeply a user goes into a site, and what is the typical reason for leaving the site.*

  Unstructured

- *Texts can be a rich source of insights, such as an overall sentiment score, or extracting keywords and semantic meaning.*
- *Images or Video. A video from a surveillance camera can be used to estimate traffic on the road, and inform people about potential traffic jams.*
- *Web server Logs can be used to understand which pages of our site are most often visited, and for how long.*

  Semi-structured

- *Social Network graphs can be great sources of data about user personalities and potential effectiveness in spreading information around.*
- *When we have a bunch of photographs from a party, we can try to extract Group Dynamics data by building a graph of people taking pictures with each other.*
- *By knowing different possible sources of data, you can try to think about different scenarios where data science techniques can be applied to know the situation better, and to improve business processes.*

By knowing different possible sources of data, you can try to think about different scenarios where data science techniques can be applied to know the situation better, and to improve business processes.

What you can do with Data

In Data Science, we focus on the following steps of data journey:

1) Data Acquisition:
The first step is to collect the data. While in many cases it can be a straightforward process, like data coming to a database from a web application, sometimes we need to use special techniques. For example, data from IoT sensors can be overwhelming, and it is a good practice to use buffering endpoints such as IoT Hub to collect all the data before further processing.

2) Data Storage:
Storing data can be challenging, especially if we are talking about big data. When deciding how to store data, it makes sense to anticipate the way you would to query the data in the future.

There are several ways data can be stored: A relational database stores a collection of tables, and uses a special language called SQL to query them. Typically, tables are organized into different groups called schemas. In many cases we need to convert the data from original form to fit the schema. A NoSQL database, such as CosmosDB, does not enforce schemas on data, and allows storing more complex data, for example, hierarchical JSON documents or graphs. However, NoSQL databases do not have the rich querying capabilities of SQL, and cannot enforce referential integrity, i.e. rules on how the data is structured in tables and governing the relationships between tables. Data Lake storage is used for large collections of data in raw, unstructured form. Data lakes are often used with big data, where all data cannot fit on one machine, and has to be stored and processed by a cluster of servers. Parquet is the data format that is often used in conjunction with big data.

3) Data Processing:
This is the most exciting part of the data journey, which involves converting the data from its original form into a form that can be used for visualization/model training. When dealing with unstructured data such as text or images, we may need to use some AI techniques to extract **features** from the data, thus converting it to structured form.

4) Visualization / Human Insights:
Oftentimes, in order to understand the data, we need to visualize it. Having many different visualization techniques in our toolbox, we can find the right view to make an insight. Often, a data scientist

needs to "play with data", visualizing it many times and looking for some relationships. Also, we may use statistical techniques to test a hypotheses or prove a correlation between different pieces of data.

5) Training a predictive model:
Because the ultimate goal of data science is to be able to make decisions based on data, we may want to use the techniques of Machine Learning to build a predictive model. We can then use this to make predictions using new data sets with similar structures.
Of course, depending on the actual data, some steps might be missing (e.g., when we already have the data in the database, or when we do not need model training), or some steps might be repeated several times (such as data processing).