Assignment #1 – Homework 1

**Course:** Ashoka Horizons: Applied Data Science with ML and AI  
**Instructors:** Rintu Kutum, Gautam Ahuja  
**Student:** Advay  
**Date:** 30th May 2025

Question 1: The “Aha!” Moment: After going through the readings and slides, what was one or two of the most surprising or “aha!” ideas you encountered about data science, machine learning, or the nature of data itself? Explain what the idea was and why it stood out to you or changed your perspective.

Ans: One surprising "aha!" moment for me came when I understood that machine learning isn’t as mysterious as it seems, it’s essentially just learning patterns from past data to make predictions about the future. I used to think ML involved some kind of deep reasoning, like a human would do, but it's actually based on statistical principles and probability. These models are used in different methods to generate the results that we are looking for. This made me more sceptical about overhyping AI and more interested in understanding the data it's trained on, since that directly affects what it learns to reason.

What stood out even more was how sensitive these models are to the quality and diversity of the data. For example, if a spam filter is only trained on emails written in English, it may completely fail when dealing with messages in another language. It became clear that machine learning isn’t just about algorithms, it’s about context, ethics, and careful data handling. I also learned that if the data reflects bias, the model will too, reinforcing existing inequalities in fields like hiring or healthcare.

Question 2: Data is King (or is it?): The articles “The Unreasonable Effectiveness of Data” and “The Rise of Big Data” highlight how having lots of data can be incredibly powerful, sometimes even more so than having a super-complex algorithm.

* + Choose one real-world application of data science or ML (either from the readings/slides or one you can think of).
  + How does the availability and use of large amounts of data (think quantity, variety, or even “messiness” as discussed in the articles) make this application possible or effective?

Ans: One paradigm example of the real-world application of data science and machine learning is how Google Translate works, which elucidates how access to large, diverse, and extensive datasets can lead to powerful outcomes, more effectively than designing complex algorithms and systems.

Google Translate shifted from a rule-based system, which was a complex of expert-crafted algorithms which relied on extensive grammatical rules, to a statistical and now neural machine translation system powered by a large corpora of multilingual texts, which serve as the base datasets for the model to learn from.

Instead of needing deep linguistic insights, such as from the expert-crafted system, Google’s current system learns from billions of examples of translated text scraped from the web, including news articles, government documents (for example, EU proceedings, multilingual government websites, etc.), subtitles, and user inputs from across different sites.

The extensive amount of data triumphs over the meticulous and intricate system. Even imperfect data can be useful at scale. For example, user-generated translations or inconsistent subtitles still contain patterns that a machine learning model can learn from.

Through the rise of big data, it emphasises that big data is not just about the size, but also about the diversity and messiness of data. Google benefits from a large variety of texts which enables it for contextual understanding, as imperfect datasets can reveal deeper patterns which may contain cultural contexts, such as idioms, expressions, etc. as it can learn how different words and phrases are used in different context. For instance, the word “bank” could refer to a financial institution or the edge of a river. Exposure to diverse sentence structures across multiple contexts enables the system to disambiguate meaning effectively. This aligns with the idea that data can reveal correlations and patterns that might not be visible through traditional rule-based logic systems.

The messiness of the data is in fact a strength, considering the size and variety as, “messy data”, which may be full of errors, inconsistencies, or non-standard usage can still be extremely valuable when available at scale. In case of Google Translate, the system doesn’t require perfect translation, just enough to infer the most likely correct output. This stands in contrast to the traditional computational linguistic approaches, which required pre-mediated translations and intricate language models.

Though Google’s feedback loops it further enriches its data sets and learns to bring higher levels of accuracy. The benefit of the real-time user feedback allows users to correct a translation or offer an alternative, they are essentially contributing new data to the system. This feedback loop aligns with the “datafication”, where real-world phenomena such as feedback become analysable through data.

Question 3: Humanity in the Loop: Machine learning is advancing rapidly, but several readings (especially Domingos’ “A Few Useful Things...” and Brynjolfsson & Mitchell’s “What can ML do?”) point out its limitations, challenges (like overfitting or the need for feature engineering), and areas where humans still excel.

* + Discuss one challenge or limitation of current ML that you found interesting.
  + Why do you think it’s important for people learning data science to be aware of this, and what role do you see humans playing in addressing or working alongside this limitation in the future?

Ans: One of the particularly interesting challenge in machine learning is overfitting, when the model learns the training data too well, including outliers, and consequently performs poorly on new, unseen data. It is the tendency of complex models to memorise the data instead of generalising from it to be able to apply the patterns derived from that data onto other situations. A model might show accuracy during training but fail to handle real-world scenarios where data is slightly different. This limitation is dangerous as it can create the illusion of a successful result, potentially misleading developers and decision-makers, especially with the increasing trust on AI models to provide answers to everything, where users blindly trust these models.

Understanding overfitting is vital for anyone learning data science because it goes into the crux of what it means for a model to “work” in real-time. High accuracy on training data can be misleading, it doesn’t translate into high utility in practice. If future data scientists are unaware of this, they may risk deploying models that appear powerful, but fail catastrophically in the real-world, whether in healthcare or finance.

This further resuscitates the importance of human intuition and judgement. Machine learning is not magical, and all models cannot be applicable in all contexts, therefore evaluating and analysing models is very important. Overreliance on these technologies cannot be reliable in today’s world. As humans can provide domain knowledge to detect when a model is learning irrelevant patterns. Humans can also employ creativity to craft or select meaningful inputs for a model, which these machines cannot do without guidance. Lastly, the ethical oversight of these models is a human responsibility. A model that overfits could unintentionally reinforce biases, or lead to harmful decisions (examples: employing police based on criminally active areas, which could be biased based on previous trends, or even in hiring or loan approvals). Human judgement is essential to review outputs, test each case, and validate fairness.

All in all, humans serve as model curators, context interpreters, and ethical stewards, working alongside machines to make their limitations manageable and its strengths more impactful.

Fun (Non-Graded) Ponder Point - Understanding LLMs

If you used an AI tool like ChatGPT to help with this assignment, you were interacting with a Large Language Model (LLM). Based on your experience using it, and perhaps thinking about the concepts of “learning from data” from our readings:

In simple terms, how do you imagine an LLM like ChatGPT “learns” to have con- versations and generate text that seems so human-like? You don’t need a technical answer – just your initial thoughts, a guess, or even a simple analogy!

Ans: I think that an LLM like ChatGPT uses the prompt keywords to create its response. It uses machine learning technology to generate the responses. The data we feed it (aka the prompts) serves as information for it to produce responses catered to us. ChatGPT stores the conversations we have it and uses it to answer with an increasingly tailored response, each time we use it. ChatGPT uses are previous prompts to train itself and build the rules to answer the questions being asked with greater precision and aptness.