Assignment #1 [HW] – Applied Data Science with ML and AI

Name: Yuvraj Khanna

Course: Ashoka Horizons, Summer 2025

Instructor: Rintu Kutum, Gautam Ahuja

# 1. The “Aha!” Moment

One “aha!” moment from the readings was the idea that having large volumes of data can often outperform sophisticated models. In “The Unreasonable Effectiveness of Data,” the authors emphasize that simple algorithms trained on massive datasets can yield better results than complex models trained on small ones. This changed my perspective because I had assumed machine learning was primarily about building smarter models, not collecting better data. It reinforced the idea that data quality and quantity are as crucial—if not more—than algorithmic complexity. It also made me appreciate the engineering and logistical side of machine learning—how data is collected, stored, and preprocessed.

# 2. Data is King (or is it?)

A great example of how large and messy data powers ML is Spotify’s music recommendation system. Instead of relying purely on user ratings, Spotify tracks billions of user behaviors—skips, replays, likes, playlist additions—to model taste. This data is messy, unstructured, and behavioral, yet it’s extremely valuable. As the articles explain, variety and volume compensate for noise. The system improves because of how much data it has, not necessarily how clean it is. Without this volume, collaborative filtering or deep learning models wouldn’t work as well. The insight here is that even imperfect data is powerful at scale.

# 3. Humanity in the Loop

One key limitation of ML is overfitting—when a model performs well on training data but poorly on unseen data. Pedro Domingos’ article discusses this as a central challenge. What struck me is that overfitting is easy to fall into, especially with powerful models, and avoiding it often requires human judgment, such as understanding the context or selecting meaningful features. It’s important for data science learners to know this because no model is perfect—and understanding its blind spots is part of using it ethically. In the future, I think humans will continue to guide, audit, and contextualize ML, especially in sensitive fields like healthcare or hiring.

# Fun Ponder Point – Understanding LLMs

To me, an LLM like ChatGPT feels like a giant prediction engine trained on language patterns. It doesn’t “think” like a human but predicts the next word based on context, like guessing the rest of a sentence. It’s kind of like a super-smart autocomplete that’s read a huge chunk of the internet. So, it learns to talk by absorbing lots of examples of how people write and speak. The more it reads, the better it gets at mimicking those patterns. It doesn’t understand meaning the way we do, but it’s amazing at replicating the structure and style of human communication.