

---

# ASHOKA HORIZONS: APPLIED DATA SCIENCE WITH ML AND AI

Name - Tushar Joshi | Week #1 | Assignment- 1


## 1. The “Aha!” Moment

*One of the biggest “aha!” moments for me was the realization that complex algorithms can actually be outperformed by simpler ones as long as we have access to a large amount of data. This idea completely challenged my assumptions. Coming from a mindset shaped by computer science and technology, I always believed that the more complex an algorithm is, the more powerful it would be. But in reality, as shown in “The Unreasonable Effectiveness of Data,” a simple model fed with massive, messy datasets can outperform sophisticated techniques. This shift in thinking was unexpected and eye-opening.*

*Another moment that really shook me was the example of Google Flu Trends. The system was able to predict flu outbreaks just by analyzing search data without actually understanding anything about the virus or how it spreads. This demonstrated that correlation alone, without causation or biological understanding, can still lead to useful, real-world predictions. It's both fascinating and slightly unsettling that data alone can give us such power.*

*A final insight, not directly from the readings but deeply connected, was the idea that machine learning doesn't necessarily replace jobs, it reshapes them. For instance, in radiology, AI can detect pneumonia in X-rays with astonishing accuracy. I recently came across a quote from a doctor saying that what took him 20 years to master, an AI system could now do in mere seconds. Yet, the human touch—making final decisions and interacting with patients—remains irreplaceable. This blend of machine efficiency and human judgment was another unexpected, yet powerful realization for me.*

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



*A real-world example of the power of big data is the "nothing but" autonomous vehicles, especially brands like Tesla. I thought that the concept of self-driving cars is all about the most complex algorithms, heavy sensors, other hardware and processing power. But while researching after the lectures, I found that huge and highly diverse data is the key stone to their functioning. Datasets like datasets from video footage, LIDAR scans, GPS data, to real-time traffic and weather information.*


*I find that Tesla's approach stands out because its cars are constantly gathering real-world data from millions of miles driven—on city roads, highways, in different weather conditions, and with unpredictable human behaviors. Every camera feed, sensor signal, and user correction (when a human intervenes) is fed back into Tesla's system. The massive dataset will become the foundation for improving its Autopilot and Full Self-Driving (FSD) systems. The scale and variety of this data allows even relatively straightforward learning models to become incredibly accurate, just because they've "seen" so many real-life situations.*

*This is what the idea in "The Unreasonable Effectiveness of Data" gave. More data can be more beneficial than better algorithms. The system doesn't need to understand human driving in a symbolic way—it just needs enough examples to mimic behavior safely and accurately.*

*And yes, the data is often messy with unexpected elements, lighting issues, or edge cases, hardware malfunctions, unexpected situations but that messiness helps the model learn to generalize better in unpredictable real-world conditions.*

### **3. Humanity In The Loop**

*One limitation of machine learning that I feel is a major one is the issue of "data scarcity" in niche or specialized problems. Unlike big tech applications that have access to massive datasets, and even if they don't, they can produce data on their own, many real-world problems in the domain of rare diseases, regional languages and dialects, mental health at different ages, new emerging problems, and other data in remote areas which lack the volume and diversity of data needed to train effective ML models. As in the readings, and in Domingos' "A Few Useful Things to Know About Machine Learning", ML systems need large datasets to generalize and make predictions*



*effectively. When such data isn't available, models often perform poorly, inaccurately, or simply it fails.*

*I believe humans can play an important role in identifying these gaps and finding creative solutions—such as creating simulations for real-world scenarios to generate synthetic data, or finding other data collection and analyzing method, improving machine learning to increase accuracy and effectiveness in data collection and processing. I have heard about efforts where simulations are used to mimic real environments to train models.*

### **Fun (Non-Graded) Ponder Point: Understanding LLMs**

*What I believe is that first these models are trained on a humongous amount of data, first for knowledge of what to answer, and then it was trained on data of how human talks and texts, there could be many conversations of someone asking something and talking. After which the LLM poses both what to say and how to say. Also I think that now LLMs like ChatGPT are trained on how humans use emoji while chatting, because ChatGPT is getting better and better day by day. I also believe that ChatGPT doesn't know what it is saying like a human. But it is trained on such a huge data on such a large number of examples that it processes what to say. Maybe now with such a high rate of improvement this thought of mine is fading but on its initial days, I see it as a parrot who just says what is trained to say.*