# AI 620 Fundamentals of Data Engineering

# Assignment 1

Name: Noor Ul Ain Anwar

Student ID: 25280062

---

## Part 1 Questions

**(a) Data Heterogeneity: Explain how your chosen data sources represent different data types (structured, semi-structured, unstructured). Provide concrete examples from your extracted data.**

1) Unstructured: The news article descriptions and content from NewsAPI are raw text, which means they're unstructured and we need NLP techniques to process the data.

2) Semi-Structured: The NewsAPI response itself and the Google Trends output, specifically when saved, are in JSON format, which means they contain nested key-value pairs, thus semi-structured.

3) Structured: The Kaggle dataset and our final processed outputs are in CSV format, with clearly defined rows and columns (e.g. Job Title, Salary, Company Location), which makes this data fully structured.

**(b) Extraction Challenges: Discuss specific technical or practical challenges encountered while accessing different data sources (rate limits, authentication, data format inconsistencies, etc.).**

1) Rate Limiting: First particular challenge encountered was the rate limit, as both NewsAPI (free tier) and Google Trends (pytrends) have strict limits on the number of requests per hour.

2) Authentication: Second challenge came in terms of authentication. The code needs to access NewsAPI key and a Kaggle JSON token, in a secure way which required managing environment variables.

**(c) Storage Justification: Explain why storing data in multiple formats (CSV, JSON) is valuable in a data engineering context. When would you choose one format over another?**

JSON is valuable for the initial extraction because it preserves the original hierarchy of the data, which is essential for "unstructured" text sources for example the data extracted from NewsAPI. Furthermore, CSV is chosen for the processed layer

because it is highly efficient for the tabular analysis and visualization to be performed later.

## Part 2 Questions

**(a) Cleaning Rationale: Justify your data cleaning decisions. Why were specific approaches chosen for handling missing data or outliers?**

For the Job Market dataset, I chose to drop rows where the salary_in_usd was missing. For a financial analysis, it's inefficient to impute salaries based on guesses as they would affect the mean/median thus introducing bias and inaccuracy, thus dropping the rows is the better option.

Secondly, to avoid duplicate data, I utilized drop_duplicates(). I did this to make sure that the frequency analysis (e.g., 'Most Common Job Titles') was not skewed by reposted job listings.

Last but not the least, for the Google Trends data, I used ffill() .i.e. forward fill for any small gaps, as search interest trends are continuous and unlikely to drop to zero instantly.

**(b) Visualization Insights: What key insights or patterns emerge from your visualizations? How do they relate to your chosen thematic domain?**

1) Temporal: Search interest for 'Generative AI' shows a sharp exponential spike starting in late 2022 (likely correlated with the release of ChatGPT), whereas 'Data Science' remains stable.

2) Categorical: The job market is dominated by 'Data Engineer' and 'Data Scientist' roles, suggesting that foundational data infrastructure is still the primary hiring need over niche AI research roles.

3) Relationship: There is a clear positive correlation between Experience Level and Salary, with Executive (EX) roles showing significantly higher variance (wider box plots) than Entry-level (EN) roles.

**(c) Visualization Critique: What limitations exist in your current visualizations? How could they be improved for different audiences (technical vs. business stakeholders)?**

Limitation: The 'Salary by Experience' box plot assumes all currencies were converted correctly to USD. If the conversion rate fluctuated during the data collection period (2020-2026), the outliers might be currency artifacts rather than real salary anomalies.

Improvement: For a business stakeholder, I would add a 'Cost of Living' adjustment layer to the salary visualization to show purchasing power rather than raw dollars.