Objective: Preprocessing Unstructured Data for LLM Applications

* This course focuses on extracting and normalizing content from various document types, including **PDFs, images, and tables**, which is crucial for expanding the information accessible to LLMs.
* Relevance to my work: Essential for enhancing the **WhatsApp/Zoom data** analysis project, allowing me to process a **wider range of data formats** within conversations and and with Gemini survey extract information from embedded images tables etc.

**What is Preprocessing Unstructured Data?**

Converting data from any format to txt form which an LLM can process

**Tools and methods for processing**

1. Txt and tables use AWS Textract
2. csv to txt – script
3. Txt, tables and images - Combination AWS Textract for txt and tables and script (ask Jafet for his method) for extracting images from pdf. (Jafet script may only work for one bank e.g. JPM theerfore dispite having over 15 banks you can just select JPM

Estimated Quality of content – eQOC% this is an estimate percentage of when looking at the data content, does it represent what it is expected to represent

Estimated Data conversion process – eDCP% this is an estimated percentage of how confident we are that the tool we use to convert did its task.

**Unstructured Datasets examples**  
1. Plain PDF – Medical Q&A explored for Neo4J convert csv to txt. Contains only txt. eQOC% == 95, eDCP% == 100

2. PDF with tables – Convert Survey pdf or docx to txt. Contains txt and tables. . eQOC% == 95, eDCP% == (depends on tool approx. 60% using AWS Textract initial txt output)

3. PDF with Images – Convert Investment newsletters to txt, tables and images (charts and heatmaps). eQOC% == 95, eDCP% == (Depends on tools)

4. Complex PDF with different challenges – BMJ diseases (contains txt, tables, picture images, (inspect a few to see whatelse) ). eQOC% == 95, eDCP% == (Depends on tools)

5. HTML – NHS and labtestonline – txt in markdown eQOC% == 95, eDCP% == 100

6. Chat/Email with time component – Whatsapp/Zoom convert csv to txt eQOC% == 60 (due to Zoom audio to text), eDCP% == 100

~~7. PPT – (I consider the preprocessing of investment newsletters to accomplish the same as what PPT would have)~~

8. Database - Explore Stocktwits Posts (csv to txt using script) – presents numerical, txt, images, gif etc challenge. eQOC% == 95, eDCP% == (depends on to-txt conversion plan)