## Problem 1

Let's say you are a Data Scientist working in a company that analyzes social media contents.Business team approached you and told you to build an Agent that will understand the context of a social media post that they will use to segment the content,find popularity,trends etc.

Task:

● Write down your approach to make the dataset, preprocess and train ML model to build such an Intelligent Agent.

● Please note that a social media post could contain Texts, Images,Videos etc. And you have to take all of this kind of data to a single Agent.

● You don't have to write any code. Just give us a detailed step by step description about the process.

## Response

Designing any Data Science project starts with getting a clear idea of business priorities. Engineering design decisions are swayed based upon the short, medium and long term business goals. Conversely, business goals might also change when faced with engineering constraints and realities (you [can’t](https://www.theguardian.com/business/2022/jun/25/max-disaster-casts-long-shadow-as-boeing-tries-to-rebuild-its-fortunes) safely design a very efficient airplane within a short few years and go to market).

We’ll first list down some design parameters and some assumptions for our business goals.

***Some Design Parameters:***

* *Model Type*
* *Transfer Learning or Train from Scratch*
* *Use Ensembles?*
* *Low Latency Inference Required?*
* *Batch Inference or Real Time Inference?*
* *Batch Learning or Online Learning? Training Frequency?*

**Assumptions:**

***Do we want a robust solution or do we want to go to market ASAP with the first iteration?***

Ans: Come up with the first implementation ASAP to aid the Business team in their decision making process.

***How do we define context?***

Ans: We define context as a **list of tags** tied to each post. Each post is defined as a tuple of all available channels (Video,Image,Text,..) . Essentially narrowing down our model design to a ***multi-input, multi-label classification problem.***

***What will be our Success Metric?***

Ans: Since topics are likely to follow a power distribution, we can’t settle on an average metric. Instead we’ll consider the **weighted F0.5 score** as our metric. F0.5 penalizes false positives more and the weightage on class representation will give **niche topics/tags** enough importance.

**Quality and Quantity of labeled data.**

Ans: Collaborating with the business team, we’ve gathered a reliable dataset of a few thousand correctly labeled posts across a diverse set of topics.

**Infrastructure Budget**

Ans: For iteration 1 we have access to a cluster of medium level ec2 gpu instances with a few hundred Gigabytes of RAM.

**Making the Dataset**

We are given access to a REST API from multiple social media platforms. We write scripts that pull data from each site over the last 24 months. This time window will cover seasonality and yearly one-off events(Religious/Cultural Festivals, Black Friday etc.). We’ll store the *text data in a No-SQL database, the images and video in blob storage like s3 and any accompanying metadata(user information, time of post, device type) in a RDBMS like PostgreSQL. The associated labels will also be stored in the RDBMS.*

We’ll maintain an anonymized identification key across all storages.

**Preprocessing data for training**

Since we have multiple types of data. We’ll need at least 4 separate pre-processing pipelines.

**For Images:**

1. Convert to uniform format and resolution
2. Convert to uniform dimensions( 2d crop and number of channels)

**For videos:**

1. Convert to uniform format,resolution and dimensions
2. Sample fixed number of frames from random positions of video timeline( Most frames are repetitive and not likely to contain rich information)

**For Texts:**