# Directions

Data: download the following two datasets [here](https://drive.google.com/drive/folders/1ij2U0QkuKio2lUlI9b3JlonRJ1himCvk?usp=sharing).

* 1. Nigeria\_ACLED.csv: political violence event data in Nigeria
  2. violence.csv: data on violence with 28 features and a binary ‘Class’ column

# Questions

**Part I. Data Exploration** (50 points)

Explore the dataset Nigeria\_ACLED.csv, paying particular attention to the variable ‘ACLED Fatalities’.

Task: What factors are most associated with high ACLED Fatalities? Choose 3 factors and visualize them clearly. Explain your findings in a short paragraph.

Note: Some factors you can explore are actors, geographic locations, event types, ACLED Notes, etc. You may find it useful to do some cleaning, create categories/bins, use NLP or geo-spatial techniques, conduct statistical testing, etc.

**Part II. Machine Learning** (50 points)

In the violence.csv dataset, use the ‘Class’ variable as the target. The ‘Class’ column indicates the escalation of violence.

Task:

1. Decide on an evaluation metric, and briefly explain your reasoning.
2. Fit a logistic regression model and a random forest model. Briefly explain which one performs better on this dataset.
3. Try to improve performance, and briefly explain what you did.