# Project D4: UFO Sightings

Team members: Iris Kreinin, Raiko Valo, Rauno Valo

## Task 2. Business understanding

### Business Understanding Report for UFO Sighting Prediction Project

### Identifying Your Business Goals:

**Background:**

Our project revolves around understanding and predicting UFO sightings in the United States of America using a comprehensive database. The primary motivation is to harness data-driven insights into the patterns and occurrences of UFO sightings.

**Business Goals:**

1. Predictive Analysis: Anticipate the location and time of future UFO sightings.
2. Shape Prediction: Develop a model to predict the shape of UFOs based on eyewitness descriptions and comments.
3. Historical Overview: Provide an overview of common UFO sighting locations, tracking popularity trends over the past century.

**Business Success Criteria:**

- Accurate prediction of the next UFO sighting location and time with a reasonable margin of error.
- Successful classification of UFO shapes based on textual data.
- Meaningful insights derived from historical data, identifying patterns and reasons behind UFO sightings.

### Assessing Your Situation:

**Inventory of Resources:**

We possess a rich database containing parameters such as datetime, city, state, country, shape, and more. This dataset serves as our primary resource for analysis.

**Requirements, Assumptions, and Constraints:**

- Requirement: Access to a robust data processing and machine learning environment.
- Assumption: Historical data accurately represents UFO sighting patterns.
- Constraint: Limited by the quality and completeness of the existing dataset.

**Risks and Contingencies:**

- Risk: Incomplete or inaccurate data leading to biased predictions.
- Contingency: Implement data cleansing techniques and sensitivity analysis.

**Terminology:**

- UFO (Unidentified Flying Object): Any airborne object or phenomenon that cannot be immediately identified.
- Predictive Analysis: The process of using data, statistical algorithms, and machine learning techniques to identify the likelihood of future events, in this case, predicting the occurrence of UFO sightings.
- Machine Learning: A subset of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to perform tasks without explicit programming, applicable here for predicting UFO patterns.
- Data Processing: The manipulation and transformation of raw data into a more meaningful form for analysis.
- Exploratory Data Analysis (EDA): The approach to analyzing datasets for summarizing their main characteristics, often with the help of statistical graphics and other data visualization methods.
- Natural Language Processing (NLP): A branch of artificial intelligence that helps computers understand, interpret, and generate human language, applied in this project to analyze textual descriptions of UFO sightings.
- Sensitivity Analysis: An examination of how the variation in the output of a model can be attributed to variations in the input, relevant for handling the risk of incomplete or inaccurate data.
- Classification: Categorizing or labeling data into predefined classes, applied here for classifying UFO shapes using NLP.
- Biased Predictions: Predictions that deviate systematically from the true values due to inaccuracies or incompleteness in the data.

**Costs and Benefits:**

- Costs: Infrastructure, data processing, and potential risks associated with model inaccuracies.
- Benefits: Enhanced understanding of UFO sighting patterns, and potential societal and scientific advancements.

## Defining Your Data-Mining Goals:

**Data-Mining Goals:**

1. Develop a predictive model for UFO sighting locations and durations.
2. Implement a natural language processing (NLP) model to predict UFO shapes based on textual descriptions.
3. Conduct exploratory data analysis to identify common trends and factors contributing to UFO sightings.

**Data-Mining Success Criteria:**

- Model accuracy in predicting future sightings.
- NLP model accuracy in classifying UFO shapes.
- Insightful findings from exploratory data analysis.

In summary, our project aims to contribute valuable insights into the fascinating phenomenon of UFO sightings. By combining predictive analytics, natural language processing, and historical data analysis, we strive to unravel patterns, predict future events, and contribute to the broader understanding of UFO occurrences in America.

## Task 3. Data understanding

## Data Understanding Report for UFO Sighting Prediction Project

### Gathering Data

**Outline Data Requirements:**

Our dataset, sourced from Kaggle, comprises UFO sighting records in the United States of America. Key data requirements include datetime, city, state, country, shape, duration, and geographical coordinates. These elements are crucial for predictive modeling, shape classification, and geographical analysis.

**Verify Data Availability:**

Upon accessing the Kaggle dataset, we confirmed the presence of essential fields such as datetime, city, state, country, shape, and coordinates. The dataset is comprehensive, offering a wide range of parameters that align with our project goals.

**Define Selection Criteria:**

To ensure the relevance and reliability of our dataset, we applied the following selection criteria:

1. **Geographical Relevance:** Focused on UFO sightings within the United States to align with our project scope.
2. **Complete Records:** Excluded entries with missing or incomplete information to maintain data integrity.
3. **Temporal Range:** Considered sightings over the past century to capture long-term trends and patterns.
4. **Valid Coordinates:** Ensured latitude and longitude values were valid and within reasonable ranges.

**Importing Challenges:**

While gathering data was straightforward, importing into our notebook presented challenges. Python libraries initially faced compatibility issues, requiring manual adjustments to the dataset before integration. This step was crucial to overcome technical obstacles and proceed with data exploration and analysis.

### Describing Data

In this section, we delve into the key features of the UFO sighting dataset, providing a comprehensive overview of the fields that will be crucial for our analysis. The dataset encompasses various dimensions, each contributing valuable information for understanding the patterns and characteristics of UFO encounters in the United States.

**Date_time:** Timestamp of UFO sighting.

**date_documented:** Date of UFO sighting documentation.

**Year, Month, Hour:** Breakdown of temporal aspects of sightings.

**Season:** Categorization of sightings into seasons.

**Country_Code, Country, Region:** Geopolitical context of UFO sightings.

**Locale:** Specific location or setting of UFO sighting.

**Latitude, Longitude:** Geographic coordinates of UFO sightings.

**UFO_shape:** Categorization of observed UFO shapes.

**length_of_encounter_seconds, Encounter_Duration:** Duration of UFO encounters in seconds.

**Description:** Textual descriptions of UFO sightings for NLP analysis.

## Exploring Data:

During the exploration of the UFO sighting dataset, several key observations and insights emerged, shedding light on the nature of the data and potential patterns:

**Temporal Trends:**

Initial analysis of the 'Date_time,' 'Year,' 'Month,' and 'Hour' fields revealed intriguing temporal patterns. Seasonal variations and specific hours of the day seemed to exhibit higher frequencies of UFO sightings.

**Geographical Patterns:**

Exploration of 'Country_Code,' 'Country,' 'Region,' 'Locale,' 'Latitude,' and 'Longitude' highlighted geographical clusters of UFO sightings. Certain regions and locales displayed higher concentrations of reported encounters, providing valuable insights for location-based analyses.

**Shape and Duration Variation:**

Analysis of 'UFO_shape,' 'length_of_encounter_seconds,' and 'Encounter_Duration' uncovered a diverse range of UFO shapes and durations. Some shapes were more prevalent than others, and encounter durations varied widely, from brief moments to extended periods.

**Textual Analysis:**

Preliminary exploration of 'Description' for NLP-based tasks indicated a rich diversity in the language used to describe UFO sightings. This diversity poses both challenges and opportunities for developing accurate classification models based on natural language processing.

### Verifying Data Quality:

The verification of data quality was a critical step to ensure the reliability and accuracy of the dataset. Several aspects were considered:

**Completeness:**

An assessment of missing values in all fields revealed that the dataset generally maintained good completeness. However, specific fields, such as 'Description,' had some instances of missing information, which will be addressed during the data preparation phase.

**Consistency:**

Checks for consistency across fields, especially in temporal data ('Date_time,' 'Year,' 'Month,' 'Hour'), showed coherent patterns. No glaring inconsistencies or irregularities were identified.

**Accuracy:**

Cross-referencing geographical information with external sources confirmed the accuracy of 'Country_Code,' 'Country,' 'Region,' 'Latitude,' and 'Longitude.' The data aligns with known geographical boundaries, instilling confidence in the dataset's accuracy.

**Relevance:**

The relevance of each field was reevaluated in the context of project objectives. All selected fields ('UFO_shape,' 'Description,' etc.) were deemed pertinent to the analyses and models proposed in the CRISP-DM framework.

**Outliers:**

Identification and examination of potential outliers in numerical fields like 'length_of_encounter_seconds' were conducted. Extreme values were found, but they appear to be valid instances and will be retained for now, subject to further analysis.

In conclusion, gathering data involved obtaining a Kaggle dataset that met our project's fundamental requirements. We verified the availability of essential fields, defined selection criteria to ensure data quality, and overcame import challenges for seamless integration into our analysis environment. This dataset serves as the foundation for our exploration, modeling, and eventual insights into UFO sightings in the United States of America.

## Task 4. Planning your project (0.25 points)

### Our task list

Times are estimated and sometimes can take a lot more or rarely less time.

- Data importing (1h) - Rauno
    - Dataset 1 (0.5h)
    - Dataset 2 (0.5h)
- Goal 1 ("Are UFO sightings still relevant? Analyze sightings frequency over time") - Rauno (14h)
    - Preprocessing data
        - Dataset 1 (3h)
        - Dataset 2 (3h)
    - Plotting data
        - frequency of sightings by year and by date (1h)
        - heatmap of UFO sightings in USA (1h)
    - Analysis
        - What can be reasons behind UFO sightings (3h)
        - Is there any reason why some years have more frequent (3h)
- Goal 2 ("Predict new possible UFO sightings place and time") (17h) - Iris
    - Preprocessing data (6h, at least half already used)
        - Dataset 1 (3h)
        - Dataset 2 (3h)
    - Creating models (3h+3h+3h = 9h)
        - A model that gives you an overall when and where someone is most likely to meet a UFO in the United States of America. The data to give in is time.
        - A model that gives you by precondition where you are most likely to meet a UFO in the United States of America.
        - A model that gives you by precondition when you are most likely to meet a UFO in the United States of America. The data to give in is in place.
    - Analysis (2h)
- Goal 3 ("Predict how the sighting will look like when experiencing it. How long it lasted and what the UFO look like?") - Raiko (17h)
    - Preprocessing data
        - Dataset 1 (3h)
        - Dataset 2 (3h)

        - Creating models
            - A model that gives you the duration of the sighting based on the description of that event (4h)
            - A modal that predicts the shape of the UFO based on the description of that event (5h)
        - Analysis (1h)
- Poster (3-5h each) - All

## Methods and tools we plan to use

- We use Google Colab for our notebook, so we can see everybody's progress at all times.
- Pandas is used for creating dataframes
- We use Sklearn to create models that predict
- We plan to use neural networks
- We use Folium to map coordinates of predicted sightings on a map.
- We use NLTK to predict based on descriptions that were provided.