

UFO Sightings and Presidential Elections

Project Progress Report

Shelby Bearrows
Applied Computer Science
University of Colorado
Boulder, CO
shbe4654@colorado.edu

Gurhar Khalsa
Applied Computer Science
University of Colorado
Boulder, CO
gukh8079@colorado.edu

Sam Cramer
Applied Computer Science
University of Colorado
Boulder, CO
sacr2234@colorado.edu

Kara Wolley
Applied Computer Science
University of Colorado
Boulder, CO
wolley@colorado.edu

PURPOSE

UFOs are a controversial topic. Mainstream media rarely reports possible sightings for fear of ridicule while individuals that report a sighting are often cast out from popular society. This investigation will focus on trends in UFO sightings in the United States and increase understanding patterns of sightings, occurrences and sighting geographical location. This information will allow the authors to identify trends in UFO shape to better understand advances in alien technology, or at least how human perception of aliens have changed over time. Additionally, the authors will portray differences in sighting frequency as a function of geographical location. A secondary goal of this investigation is to explore potential relationships between presidential elections and reported UFO sightings. Presidential election data will shed light on the relationship between geographical region, UFO sighting frequency and UFO sighting description as a function of political party. Finally, we aim to predict where and when UFO sightings occur based on the outcomes of US presidential and general elections.

PRIOR WORK

To the author's knowledge, little work has been published on the relationship between UFO sightings and election information.

Regarding UFO data analysis, prior work from the data source (National UFO Reporting Center, 2019) also includes 166 references to additional analyses performed by individuals. Of these resources, most explored frequency of sightings by location (including per capita), word cloud analysis on the description attribute, shape of UFOs by time of year, sightings by shape and season, and shape frequencies. Perhaps the most explored question is whether sightings have increased over time, which some reported a possible correlation. An additional correlation between sightings and geographical region was reported (Adhokshaja, 2016) using a chi-square analysis. While shape over time has been explored, the time variable was usually by week or month, rather than by decade for example (this could be interesting to explore).

Additional work has been done by a group of masters students from USC Viterbi school of engineering, using data scraped primarily from ufostalker.com. They were able to conclude that "sci-fi movie releases correlated with an increase in sightings" and "that events like thunderstorms caused an uptick in reports" (Dawson, 2018). In addition, the data suggested higher frequencies of sightings during summer months. The authors used a word cloud to show similarity in words used to describe these sightings.

While others have explored some of the main questions regarding ufo sightings,

merging our dataset with election data will likely add a new level of novelty to our work not present in the prior work.

PROPOSED WORK

The primary purpose of this research will be a novel, intuitive and visual representation of UFO sightings that provide a new perspective to the authors. A secondary goal of this project is to integrate election data with the UFO sightings data. With the integration of these data, our research will aim to answer the following:

1. Are UFO sightings increasing over time?
2. How has the UFO shape changed over time?
3. If UFO sightings do increase over time, how is that correlated with the turn out rate for elections?
4. Do states that vote Republican have significantly more UFO sightings?
5. How does term number affect the number of UFO sightings? Is this also dependent on the political party?

DATA COLLECTION

Two data sets were collected and compiled for use in this research. The primary data set consists of UFO sighting data published by the National UFO Reporting Center and spans almost a century (National UFO Reporting Center, 2019). Each entry includes date and location of the sighting, along with a description of the sighting that include the shape of the UFO and the duration of the sighting. This data set is approximately 28MB in size and contains 11 different attributes of UFO sightings. The secondary data set includes election data collected by the MIT Election Data and Science Lab (Tunguz, 2020). The dataset covers the complete voter breakdown in

each county for every presidential election since 1976. It also includes information about each candidate that was on the ballot in each county like party affiliation and the number of votes they received in that county. The table is 4.5 MB and contains 11 different attributes for each entry. Preliminary work includes cleaning, preprocessing and integration of data. The UFO dataset is especially 'dirty'. Reformatting of dates, state name and removal of random symbols (&, !, #) from the comments attribute is necessary. The US Elections dataset does not appear to need much cleaning. Preprocessing includes correcting missing values, null handling and identifying erroneous data or outliers. Once these preliminary tasks were completed, the two data sets were integrated into a single table. In total, this research will include 17 attributes. Additionally, the two datasets have been merged into one so that each row in the UFO election dataset coincides with the results of the closest past election. For example, a UFO sighting in Texas in 1979 coincides with the 1976 presidential election and the voting results for Texas during that election are listed alongside that UFO sighting. The reasoning behind this is that it captures what the voters and population were thinking and feeling during the four years between elections. The comments describing the UFO sightings have also had all special characters stripped from them and have been stored in arrays split on the spaces in the descriptions to aid in further analysis of themes in the sightings.

EVALUATION METHODS

The main evaluation methods we are currently planning to use are histograms, heatmaps, apriori's algorithm, pearson's correlation coefficient, lift, Student's t-test

and PCA plot. We will use histograms to create bins of the different terms used to classify UFO shape. We will try to compare histograms based on different election years to see if the common descriptors change with time, and if they do, are they affected by election results. We'll use heatmaps to visualize geographic location of UFO sightings to better understand where they are happening and compare these to similar heatmaps of how the different regions of the country skew in terms of which party and candidates they vote for in elections. Apriori's algorithm will be used to quantify minimal support for frequent descriptors used in UFO sightings. Depending on the results of this mining, we'll try to group the UFO data by location and see if we can get different results for regional grouping. Pearson's correlation coefficient and Lift will be used to determine correlation between the UFO and election datasets. Student's t-test will be used for a comparison of means between the datasets. PCA plots will be used to reduce the dimensionality of the data to get a better understanding of how different or similar the UFO sightings are to each other.

TOOLS

For our data analysis we will be using a myriad of programming languages and packages. The programming languages we will be using are MATLAB, Python and R. For data cleaning and tidying we will be using numpy and pandas in Python and tidyr and dplyr in R. For data visualization we will use matplotlib and seaborn in Python, and ggplot in R. We will also utilize Tableau for quicker, short-term data visualization. Additionally we have decided to use the nltk package in python to clean and process textual data in our dataset.

MILESTONES

1. Oct. 16: Data cleaning
2. Oct. 21: Final decision on integration of electoral data. Preliminary data selection, transformation and visualization
3. Nov 5: Final decision on integration of other potentially interesting data.
4. Nov 7: Data mining and evaluation with second pass of visualization
5. Nov 21: Preliminary pass of knowledge presentation.
6. Nov 26: Completed data cleaning and one-hot encoding. Begin Apriori Algorithm and Association Rule mining.
7. Dec 1: Prepare visuals for final presentation.

MILESTONES COMPLETED

We have completed most parts of our past milestones so far except for finishing the data cleaning process. We have decided to move forward with our current data as it is already rich with information.

MILESTONES TODO

Future milestones include preprocessing and cleaning the description data for the application of the Apriori algorithm. This will involve continuing to clean the data and formatting it so that the text can be passed into the algorithm so we can extract meaningful text from the comments section. We will also need to one-hot encode the textual data prior to running Apriori.

RESULTS

Preliminary data analysis was performed to explore data availability and initial responses to proposed research questions. First, a heatmap was created using the total number of UFO sightings by city in the lower

48 states in the United States (Figure 1). This analysis was updated from previous versions to scale for population, which was taken as the number of votes cast in the previous election. Raw population for each city was not used due to time and computing constraints. As shown from this figure, we believe that more sightings occur in larger, more populous cities and more sightings occur in coastal areas with some hotspots throughout the midwest. More research to understand the political leanings of large cities and coastal cities is required to make further conclusions about the political leanings of these urban areas.

Secondly, an initial comparison between UFO sightings and election votes, across all states, was performed (Figure 2). It appears that democrats report a larger number of UFO sightings than republicans. This analysis does not include an offset for population increases over time which should be introduced in later versions.

The results of UFO sighting keyword descriptions are shown in Figure 3. This plot shows that the somewhat vague description “light” is the most common UFO descriptor followed by “circle” and “triangle”. Much of this was anticipated during the project ideation however new characteristics of “cigar” or “chevron” add novel keywords that can be later analyzed. Figure 4 shows that while some descriptions such as “light” or “triangle” appear frequent across states, descriptors like “oval” and “formation” may be more frequent in areas like California. Hence some descriptions may be local to certain regions across the country. Future analysis may include determining if political affiliation of these areas determine the characteristics used.

The use of common characteristics for UFO sightings over time is shown in Figure 5.

This figure shows that the descriptor “light” has been used somewhat consistently since the late 1990’s, whereas “circle”, “triangle”, and “fireball” may be terms used more frequently in recent years to describe UFO sightings.

The frequency of UFO sightings by state is shown in Figure 6. It appears that California has been consistently reporting sightings since the 1990’s, whereas other states such as Washington, and Arizona experience spikes in sighting frequencies over time (as shown by the intermittent green-yellow pixels of the heatmap).

The last analysis performed for this milestone was the characteristics of UFO sightings by election year between 1908 and 2012 (Figure 7). It appears that “disk”, “cylinder” and “diamond” were more commonly used to describe a UFO sighting up to 2008. Following 2008 the description of “light” began to overwhelm the UFO sightings.

REFERENCES

- [1] Adhokshaja, P. (2016, November 22). *UFO Data Analysis*. EDA and Chi Squared Test. Retrieved October 21, 2020, from: kaggle.com/adhok93/eda-and-chi-squared-test
- [2] Dawson, C. (2018, July 2). *Data-Crunching Class Looks to the Skies to Explore UFO Sightings*. USC Viterbi News. Retrieved October 20, 2020, from: viterbischool.usc.edu/news/2018/07/data-crunching-class-looks-to-the-skies-to-explore-ufo-sightings/
- [3] National UFO Reporting Center. (2019). *UFO Sightings*. Kaggle. Retrieved September 14, 2020, from: kaggle.com/NUFORC/ufo-sightings/notebooks
- [4] Tunguz, B. (2020, October 13). *US Elections Dataset*. Kaggle. kaggle.com/tunguz/us-elections-dataset

Appendix A: Figures

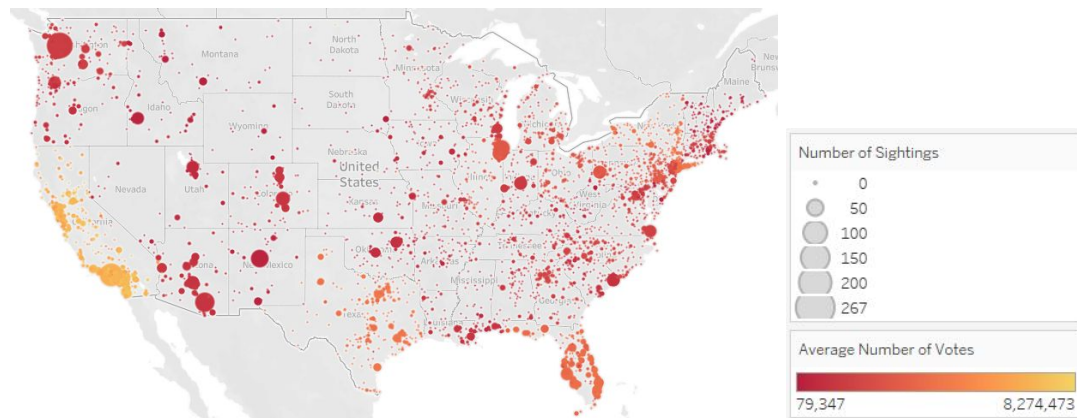


Figure 1: Heatmap of UFO sightings in the United States where the size of the marker is the number of sightings in a given city and the color is the average number of voters across all election years.

Number of UFO Sightings by Election Year

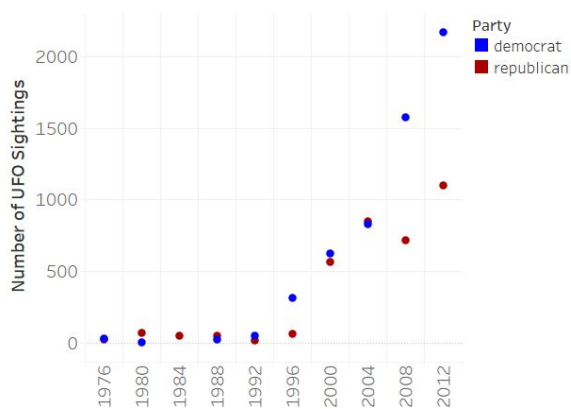


Figure 2: Preliminary investigation on the total number of UFO sightings by election year. Sightings are counted as the sum of all sightings by party in that year.

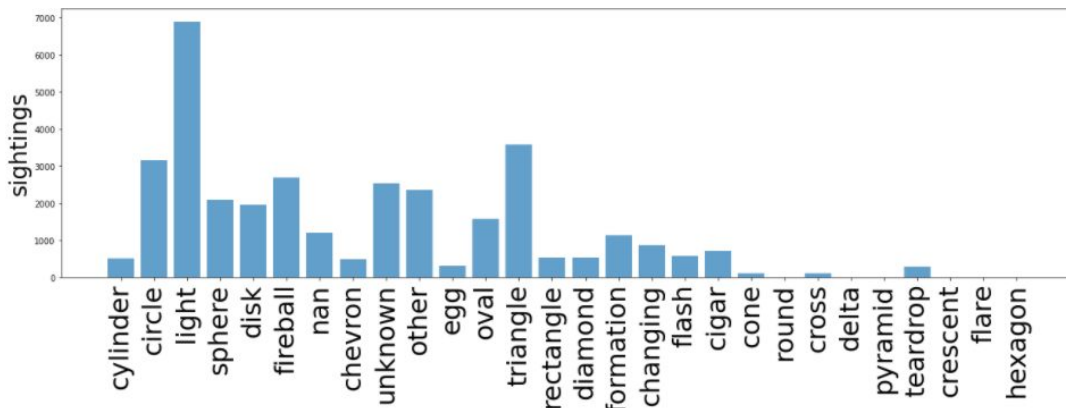


Figure 3: Frequency of sightings by keyword description.

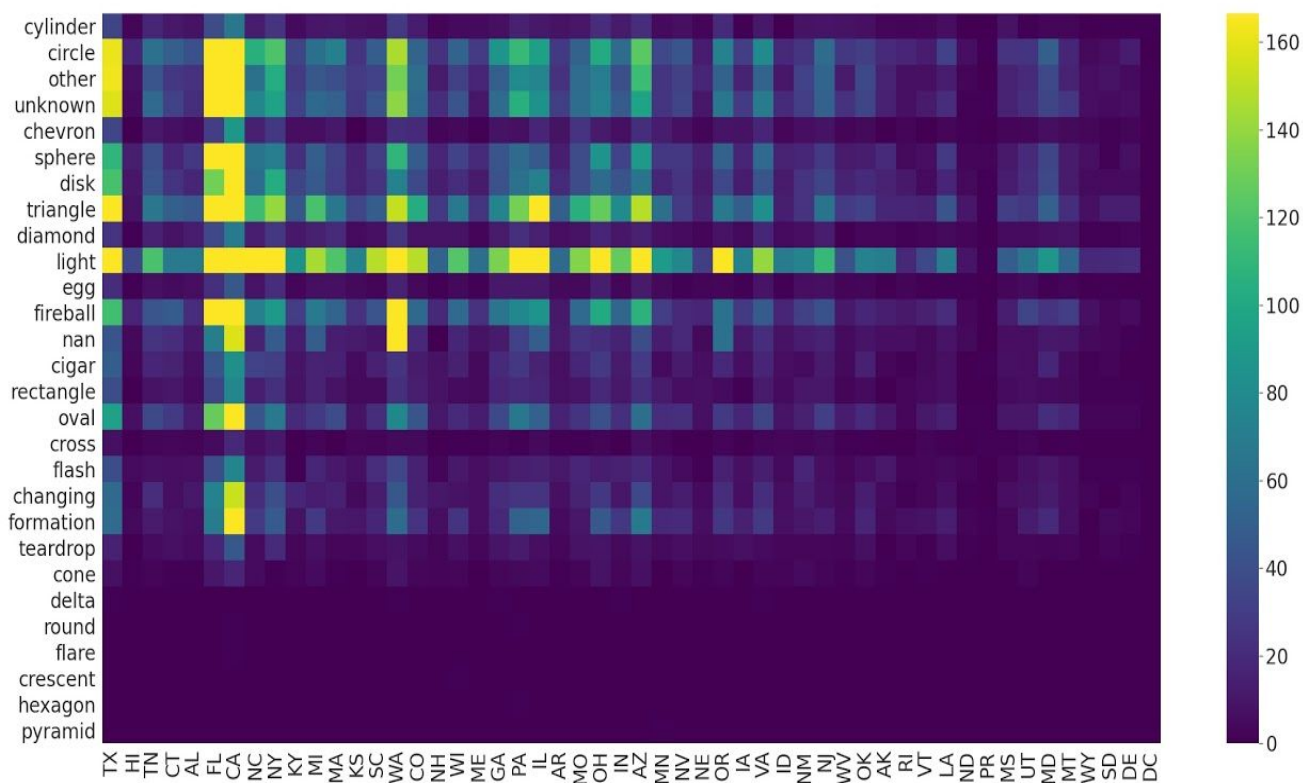


Figure 4: Heatmap of UFO description by state.

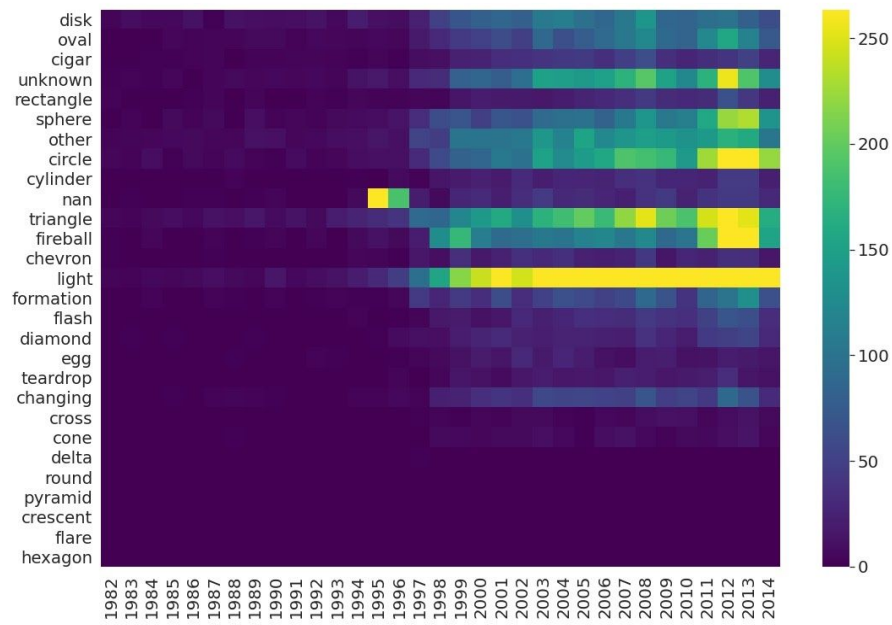


Figure 5: heatmap of shape by year frequencies.

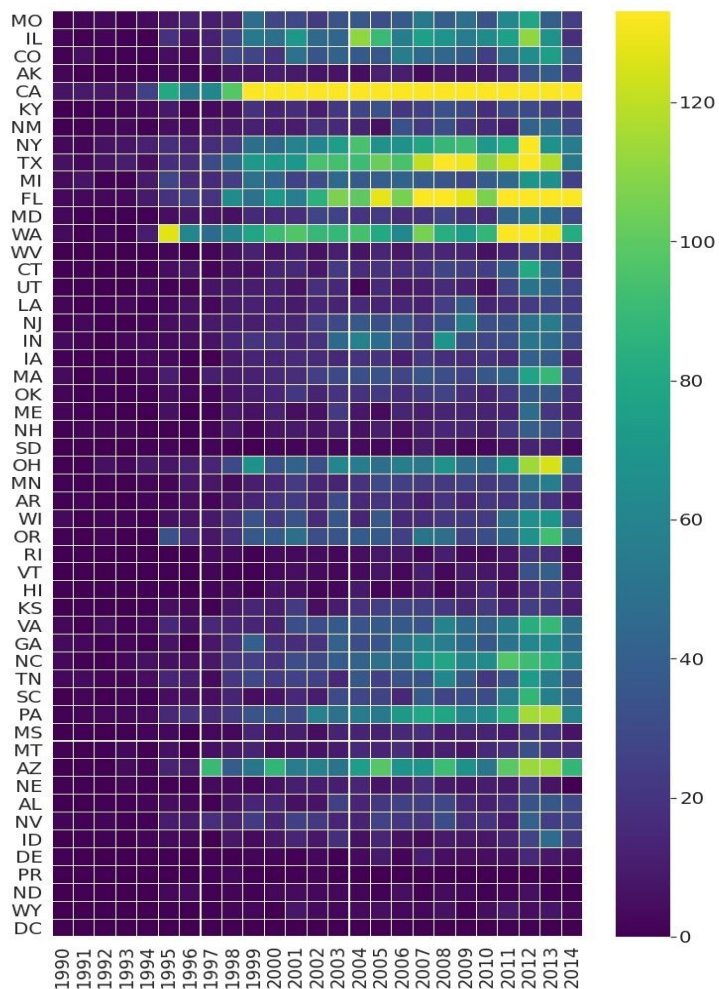


Figure 6: Heatmap of UFO sighting frequencies by state and year

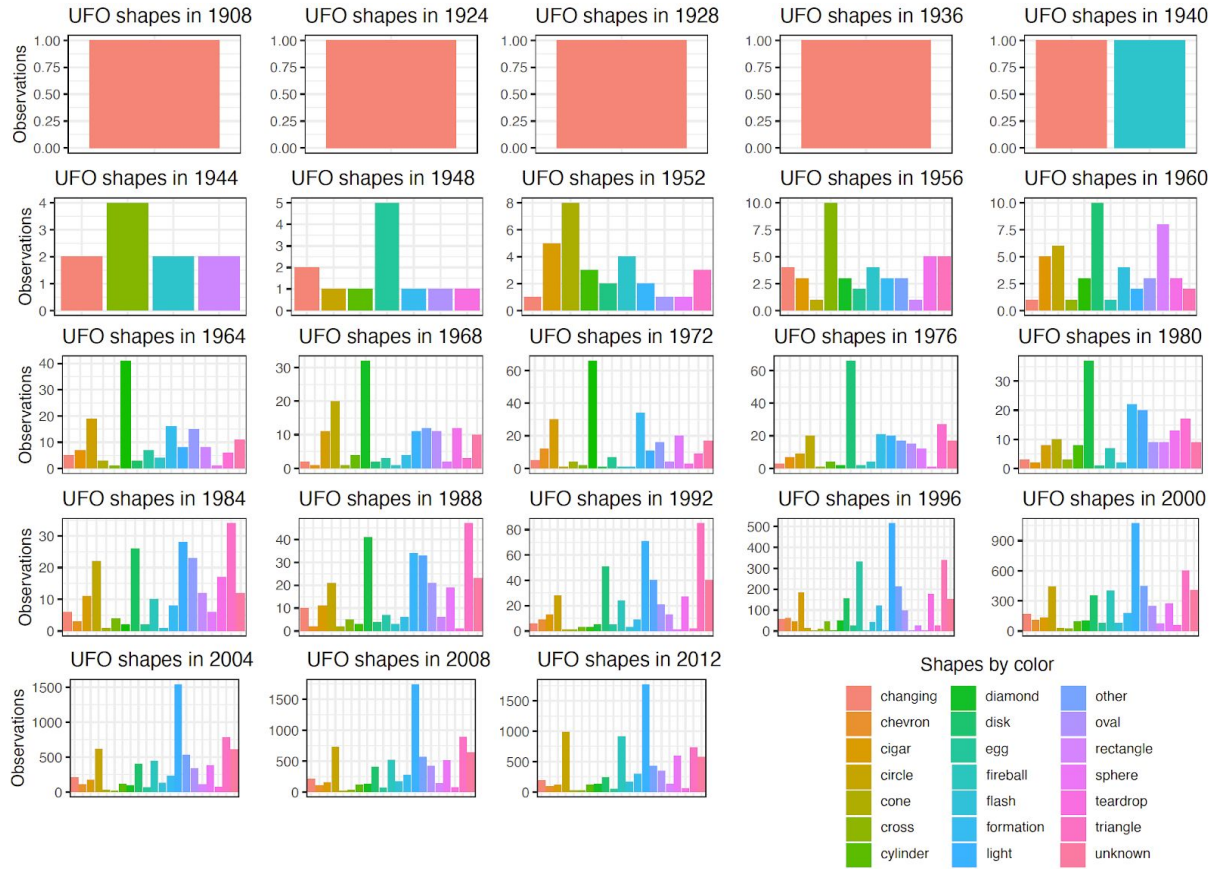


Figure 7: Changes in how the shapes of UFOs are displayed as bar graphs for each election year between 1908 and 2012. Bars are colored by shape.