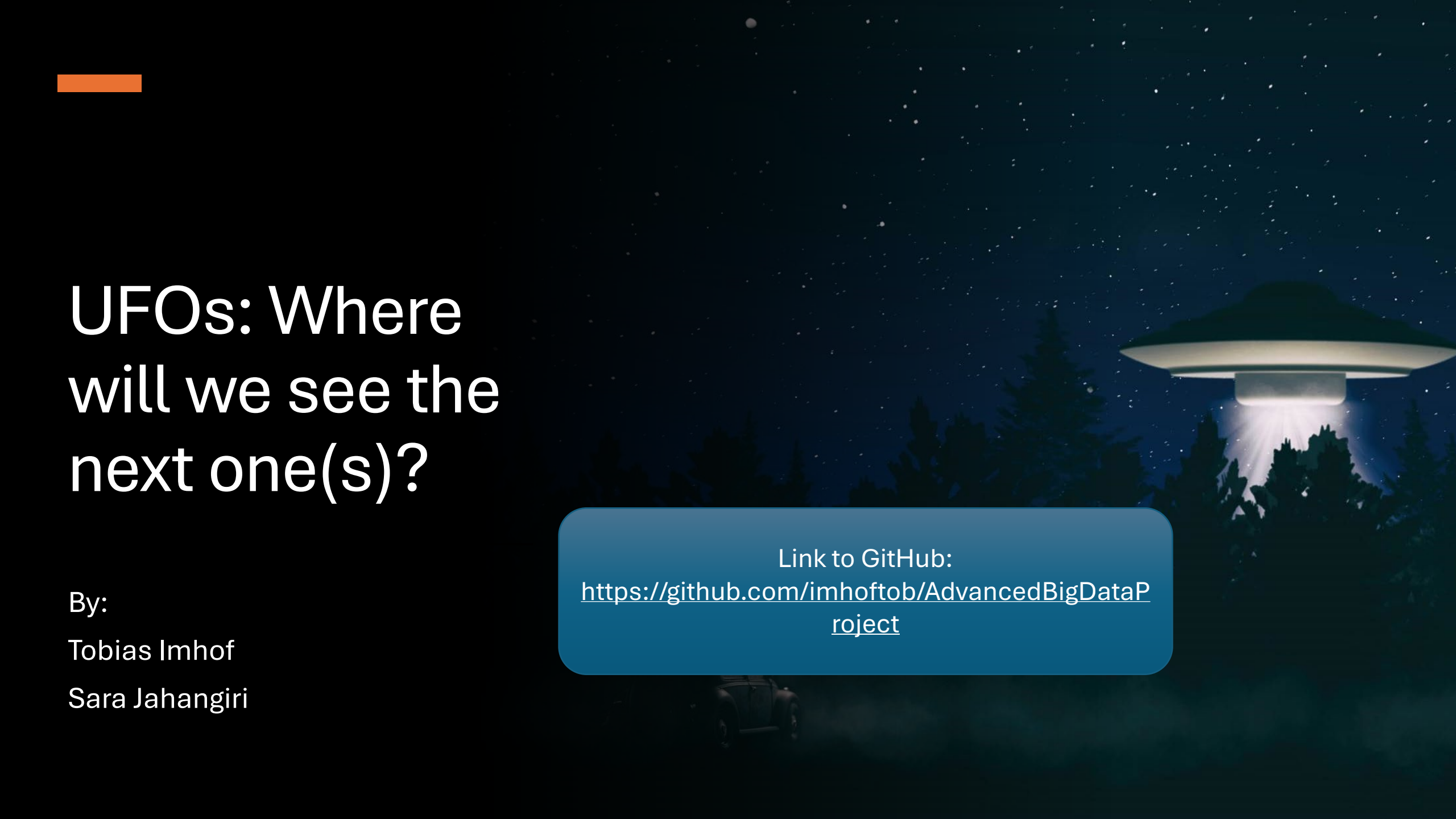
UFOs: Where will we see the next one(s)?

By:

Tobias Imhof

Sara Jahangiri

Link to GitHub:
https://github.com/imhoftob/AdvancedBigDataProject



Agenda

1. Our data
2. First analytics
3. Deeper analytics
4. Comparison
5. Number of clusters k-mean
6. Impact of radius and min. cluster size Random Forest
7. Conclusion
8. Random forest regressor
9. Our learnings

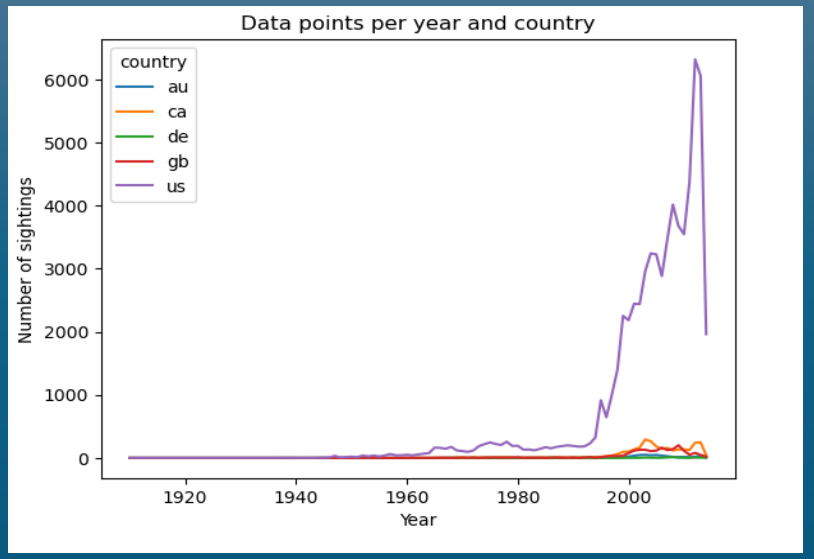
1. Our Data

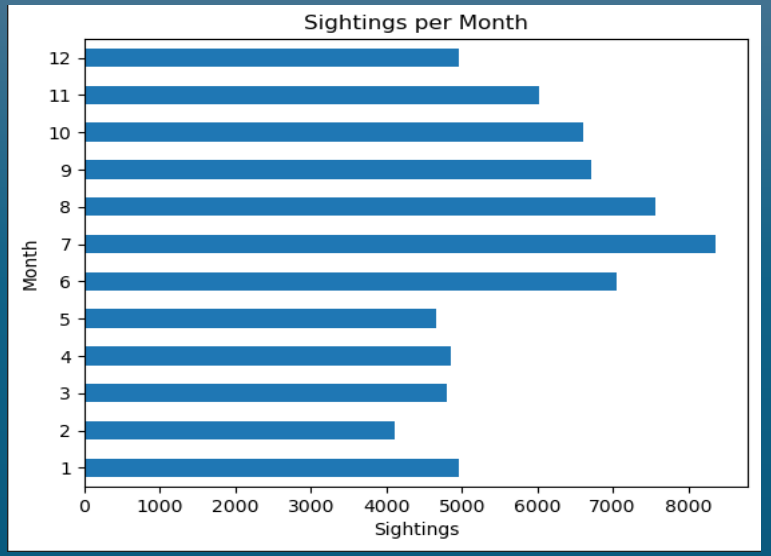
- A dataset of UFO sightings from Kaggle

	datetime	city	state	country	shape	duration	comments	latitude	longitude
0	10/10/1949 20:30	san marcos	tx	us	cylinder	2700	This event took place in early fall around 194...	29.8830556	-97.941111
1	10/10/1949 21:00	lackland afb	tx	NaN	light	7200	1949 Lackland AFB&#44 TX. Lights racing acros...	29.38421	-98.581082
2	10/10/1955 17:00	chester (uk/england)	NaN	gb	circle	20	Green/Orange circular disc over Chester&#44 En...	53.2	-2.916667
3	10/10/1956 21:00	edna	tx	us	circle	20	My older brother and twin sister were leaving ...	28.9783333	-96.645833
4	10/10/1960 20:00	kaneohe	hi	us	light	900	AS a Marine 1st Lt. flying an FJ4B fighter/att...	21.4180556	-157.803611

2. First Analytics

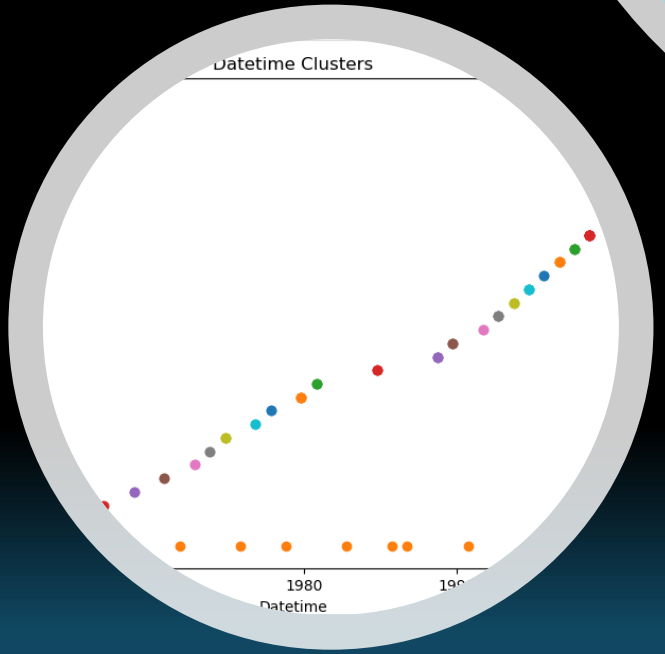
- First impression of the UFO sigting data
- Sharp increase from the 1990s onward

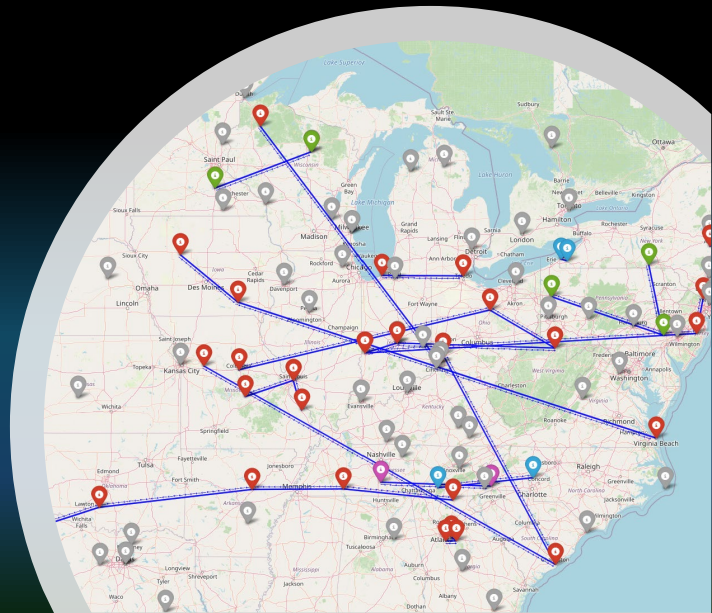




3. Deeper analytics

- Visualizing data
- Finding most likely Ailien bases using centroids
- Creating time-geo clusters
  - Directly using DateTime and special data
  - Weighting DateTime more than special data
  - 2-step approach: first clustering based on date-time and within each cluster do a second clustering based on spatial data
- Trying to create UFO flight patterns





4. Comparison

```
Time taken (K-Means): 0.0163 seconds
Time taken (DBSCAN): 6.9040 seconds
Time taken (HDBSCAN): 2.8197 seconds
```

K-means

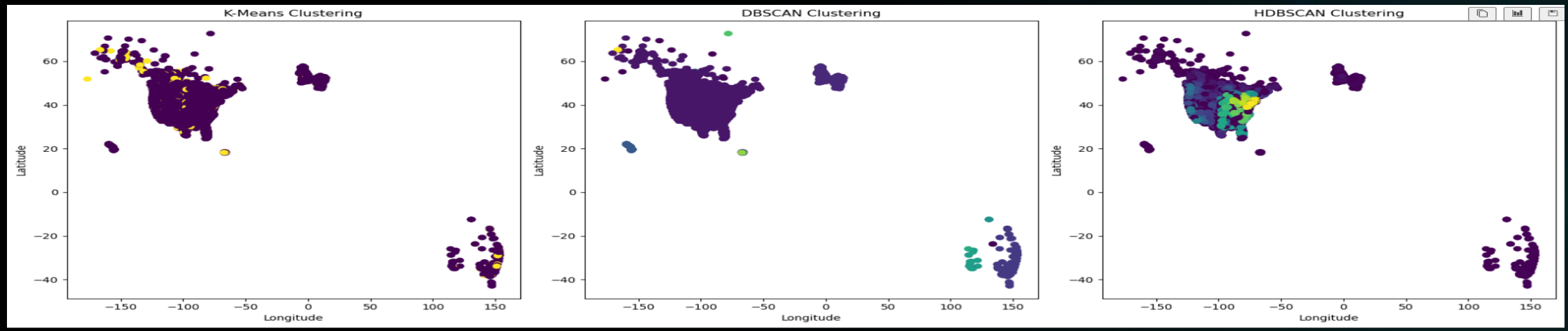
- Is sensitive to the number of clusters specified
- Does not work well with outliers and noisy datasets
- Input: Number of Clusters
- More efficient for large datasets

DBScan

- Does not require us to specify number of clusters
- Handles outlier and noisy datasets well
- Input: Radius and min. number of points
- Not so efficient for large datasets

HDBScan

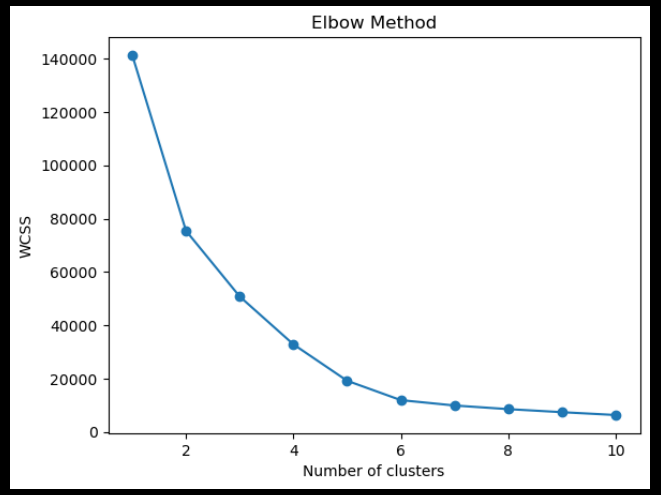
- Improves DB Scan by adding hierarchical approach. This allows for clusters of varying density
- No need to specify Radius
- No need to specify number of clusters
- Handles outlier and noisy datasets well

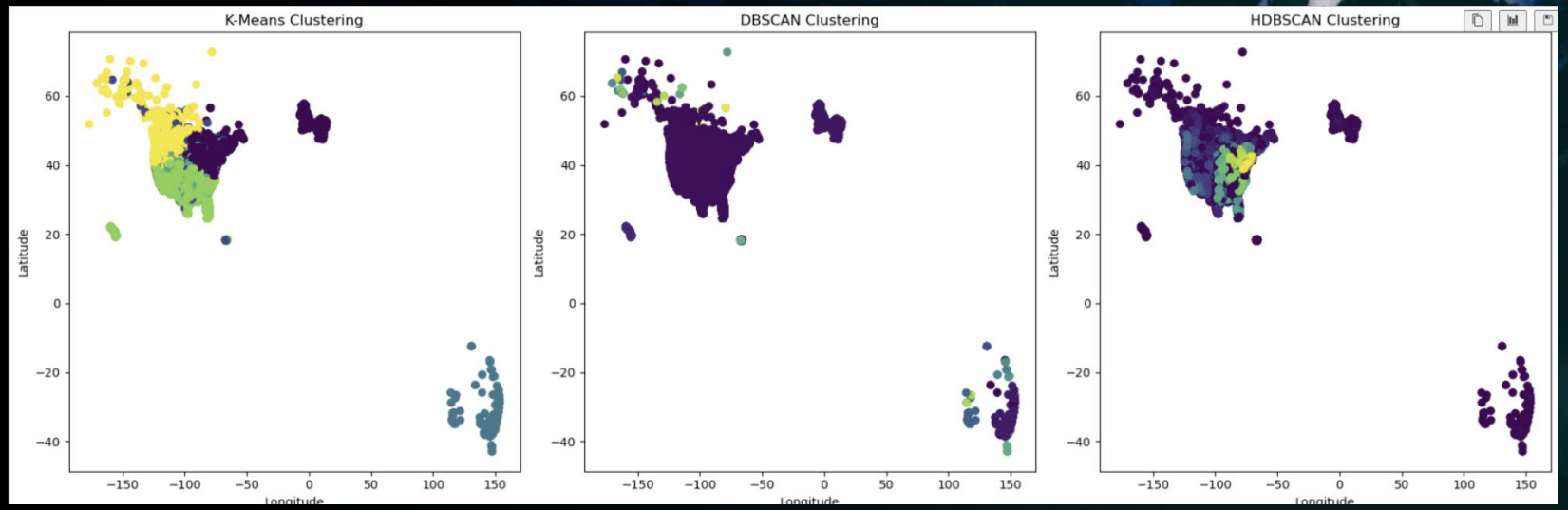


5. Nr of clusters for K-means

- Elbow method used and decided on k=6
- No significant time increase compared to k=2

```
Time taken (K-Means): 0.0210 seconds
Time taken (DBSCAN): 3.1113 seconds
Time taken (HDBSCAN): 2.8373 seconds
```

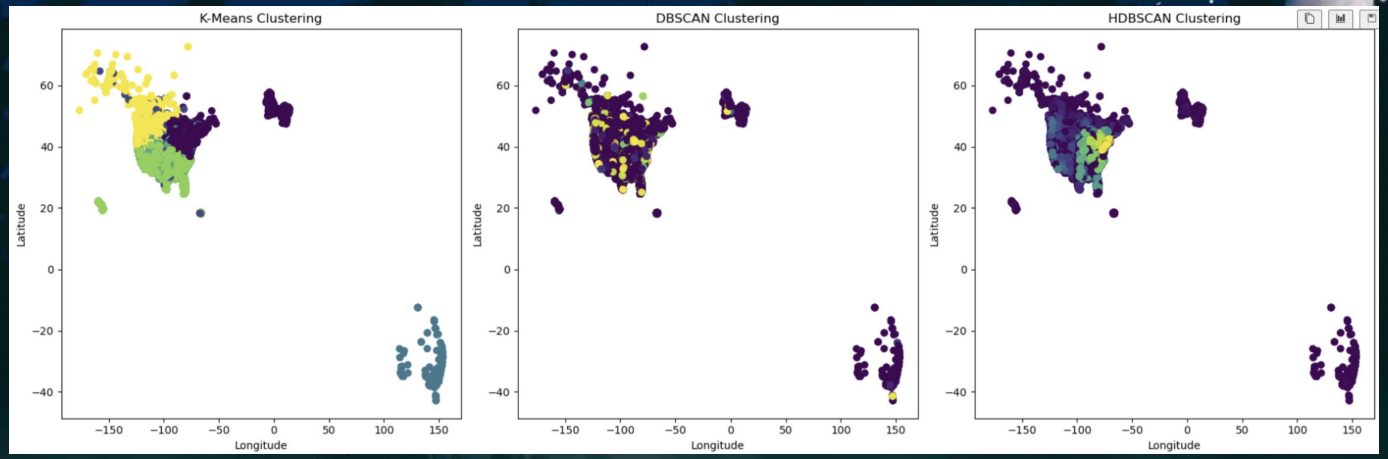




6. Impact of radius and min. cluster size

- DB Scan gets faster with smaller radius

```
Time taken (K-Means): 0.0188 seconds
Time taken (DBSCAN): 0.2729 seconds
Time taken (HDBSCAN): 2.8580 seconds
```



7. Conclusion

K-means

- Defining the number of clusters is really dificult and has huge impact
- Does not allow clusters within clusters which with our messy data would be very useful
- Our data is too messy for K-means
- It is really fast!

DBScan

- Does not require us to specify number of clusters
- Handles our noisy data well
- Allows for great control over the clusters
- It is slower in most cases

HDBScan

- All the benefits from DB Scan
- Needs even less input than DBScan
- Due to hierarchical nature it can be even slower than DBScan
- Depending on the needs can create more meaningful clusters

8. Random forest regressor

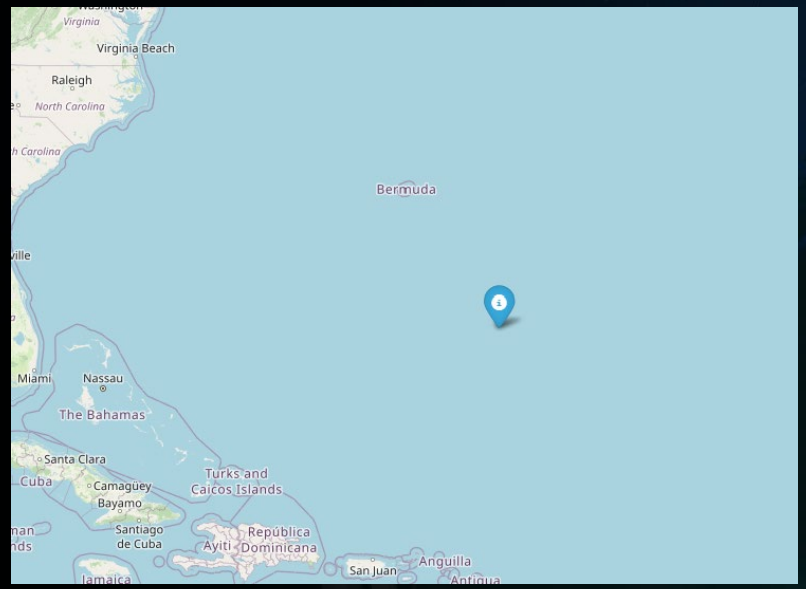
Why Regressor and not classifier? Because our data is numeric.

- Objective: To accurately forecast the coordinates and timing of UFO sightings based on specified dates
- Challenges:
  - DateTime format has to be split in its components
  - Many different DateTime formats making the split difficult
- Outcome:
  - An objectively not satisfying model!

When and where can we see a UFO today?
(According to this very unsatisfying model)

```
# Prepare the input data for prediction
test_data = pd.DataFrame({
    'day': [5],
    'month': [11],
    'year': [2024]
})
```

	longitude	latitude	hour	minute
0	-60.954127	27.39091	17.168892	23.249875

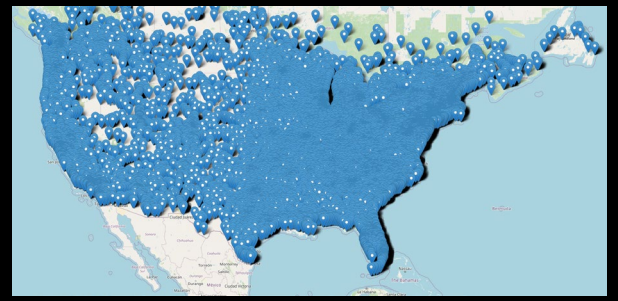


Yes, it is still really bad.

```
Mean Absolute Error (MAE): 13.279331921709591
Mean Squared Error (MSE): 521.3834387049584
R-squared (R²): -0.0943280963344783
```

9. Our Learnings

- Folium creates beautiful maps which can get really huge!



- Even the most advanced model cannot yield satisfactory outcomes if the quality of the data is poor
- Use HDBScan more often!
- UFO sighting Data is not high-quality data:
  - UFO sightings rely on human observers, and areas with higher populations tend to report more sightings
  - Interpreting an unidentified object as a UFO involves significant socio-cultural influences
  - Predominantly English-speaking countries are represented

Questions?