HacktheMachine.ai

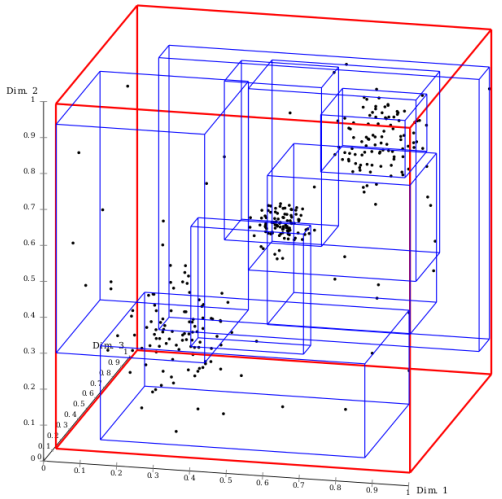
Data Science & the Seven Seas: Collision Avoidance

Team Redhorse

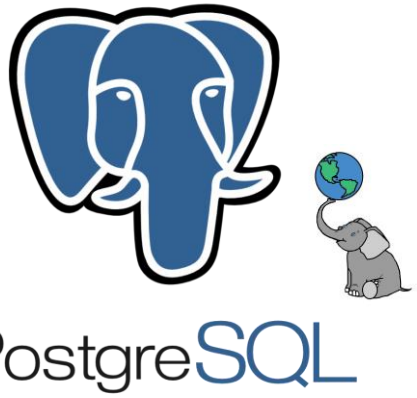
Jodi Deprizio, Julia Fletcher, Dennis Silva Jr., John Michael Kovachi,
Joseph Haaga, Nicolas Westin, Brett Schultz, Eliot Voelker

Preprocessing and Data Enrichment

- Focused on Backend and setting up a strong
- Able to fill in missing SOG and COG data
- Label ship encounters via COLREGs
- Set up PostgreSQL database with PostGIS plugin, to index ship routes via R-Trees, this resulted in a rapid spatial search
- Can efficiently search the 3D space for interactions with a CPA of 4nm

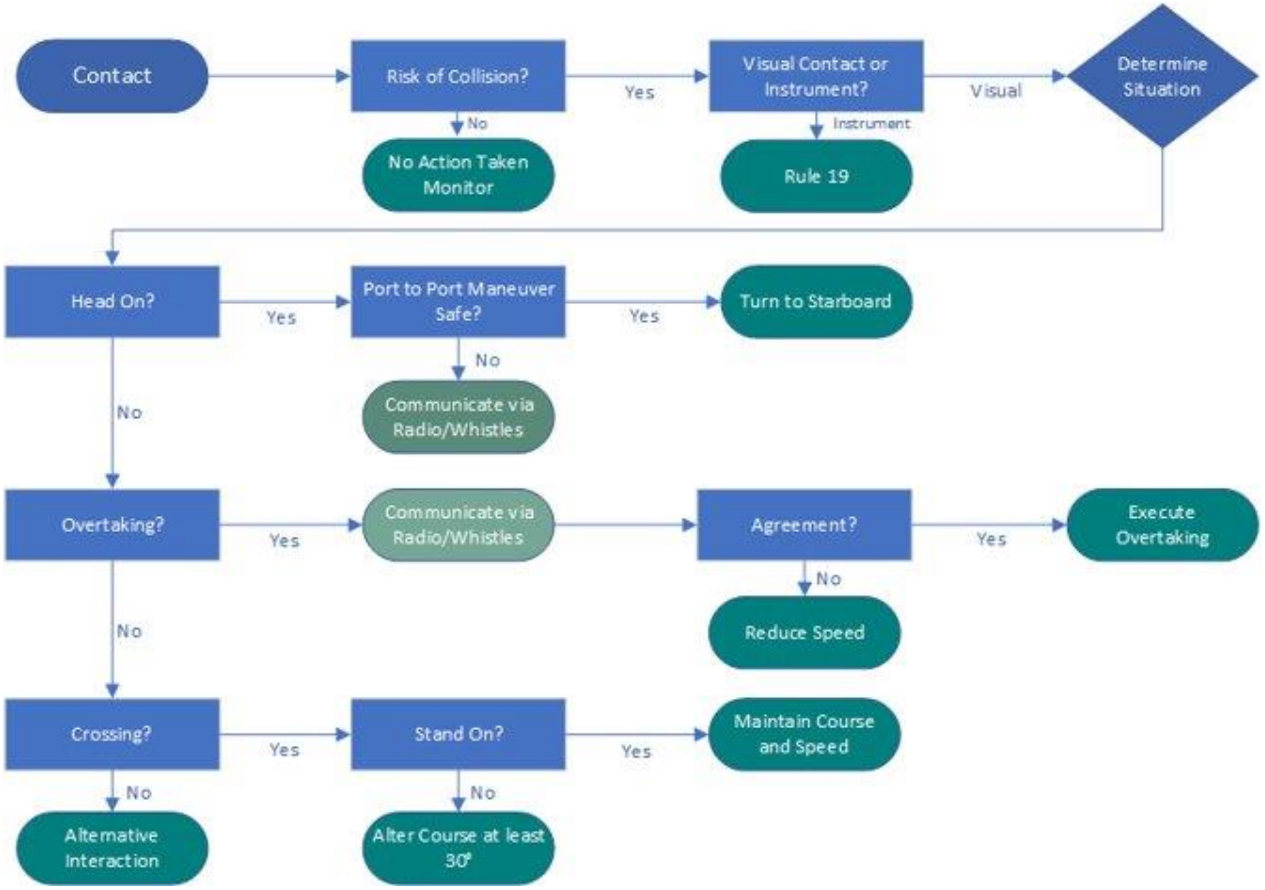



```
SELECT *
FROM analysis AS t1,
analysis AS t2
WHERE St_intersects(t1."geom", t2."geom")
= true
AND t1."min_hour_basedate" =
t2."min_hour_basedate"
AND t1."mmsi" < t2."mmsi"
AND t1."min_sog" > 1.0
AND t2."min_sog" > 1.0;
```



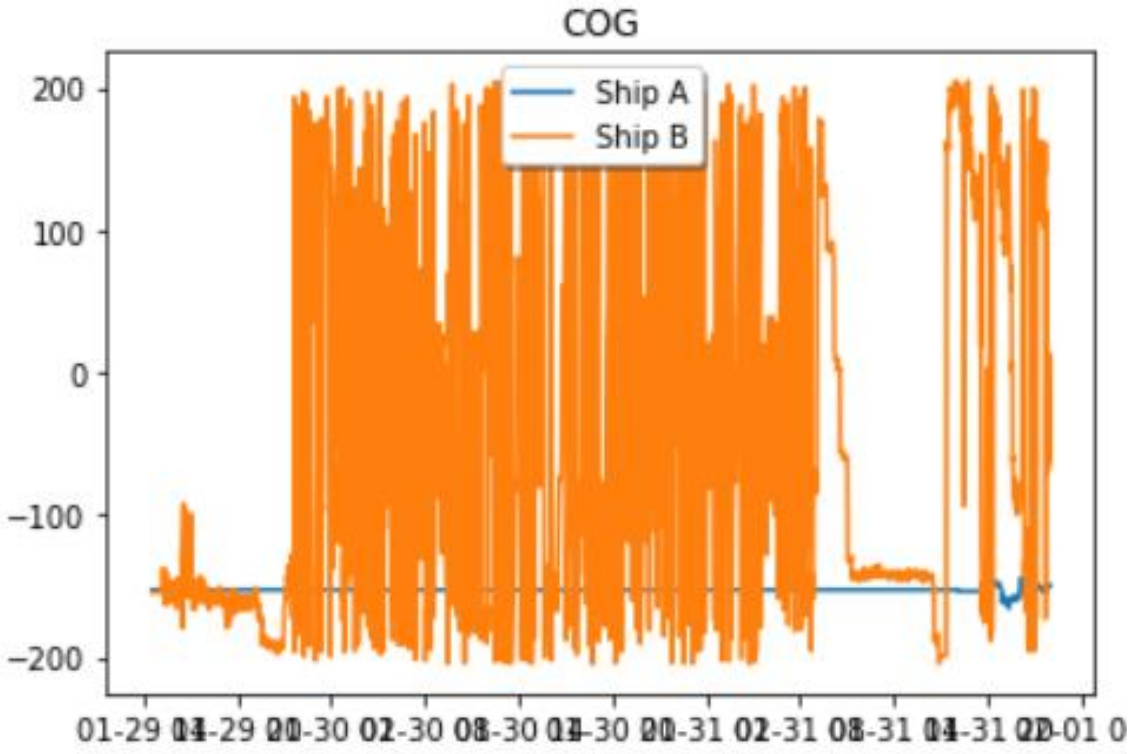
Exploratory Data Analysis

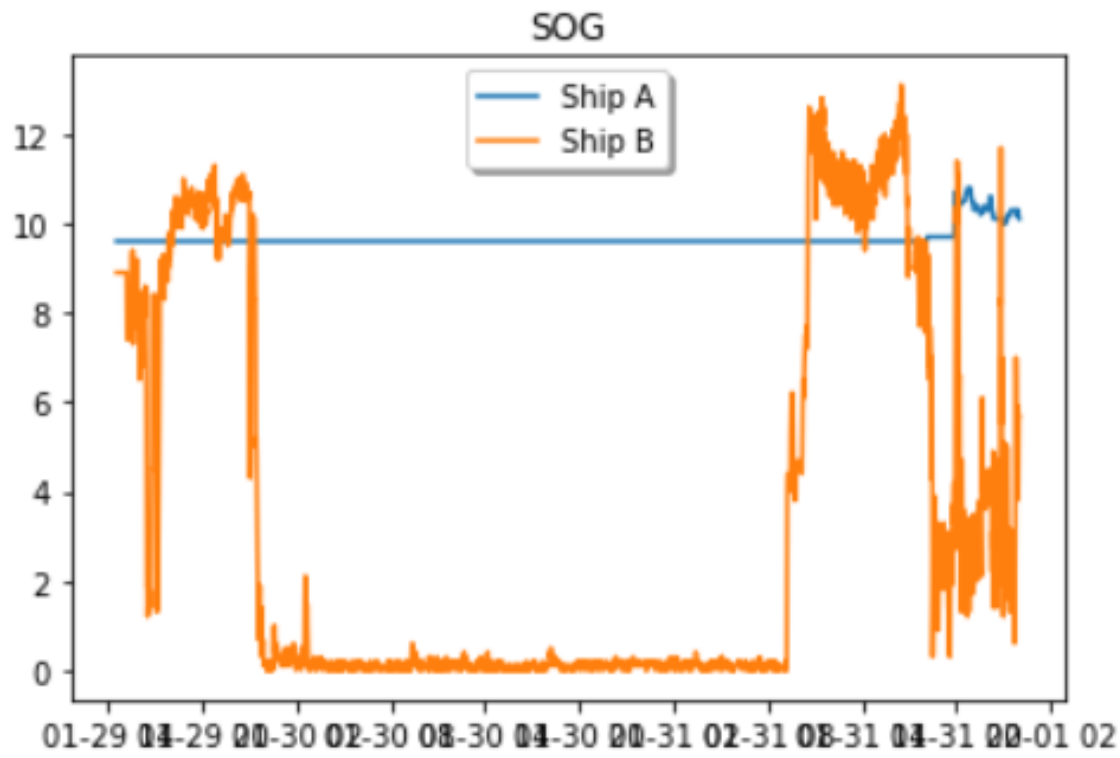
- Rules based approach used to define which COLREGS interaction is taking place
- Statistical analysis on change in SOG and COG
- Started to build a tool to compare the reported SOG and COG with a calculated SOG and COG based off GPS data



Jodi Deprizio, Julia Fletcher, Dennis Silva Jr., John Michael Kovachi, Joseph Haaga, Nicolas Westin, Brett Schultz, Eliot Voelker

Statistical Analysis on Change in COG and SOG



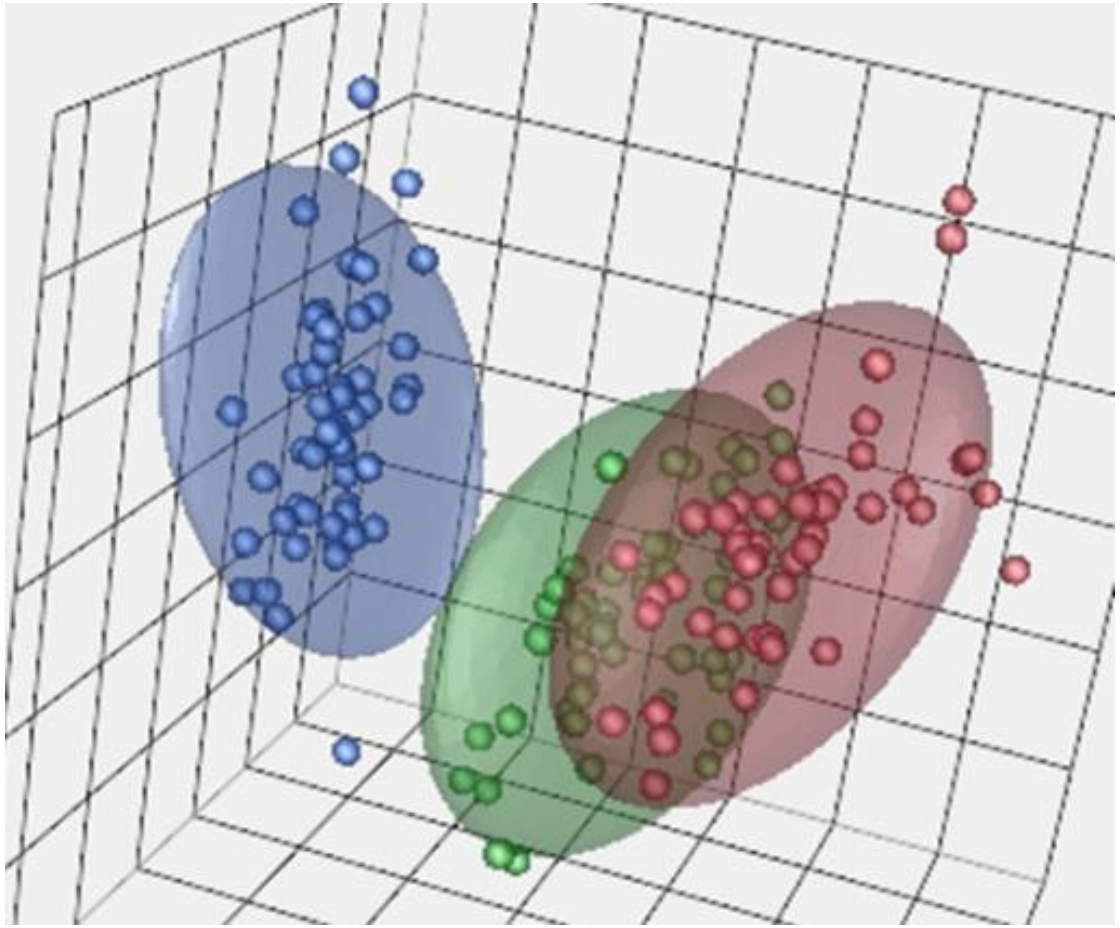



For development purposes only
Tanaga Island
Adak Island
Atka
Atka Island
Amlia Island
Alaska Maritime National Wildlife...
Alaska Maritime National Wildlife...

Machine Learning Approach

1. K-Means Clustering (k=3) on aggregate interactions to label them as Overtaking, Crossing, or MeetingHead-On

2. Decision Tree – Evaluate `feature_importances` in determining label of a particular interaction



ML Dataset Issues

- With such a large dimension of data within a 3D space, we end up running into the Curse of Dimensionality
- Running ML requires some sort of pattern recognition in the data
- Without previous knowledge of Naval ship interactions, important variables are unknown
- Data has so many dimensions that patterns are difficult to recognize without large datasets
  - Rate of data required to make insights rises exponentially with every new potential variable added

Further Work

- Post-processing to visualize the data tracks into ARCGIS
- Classified Machine Learning Approaches
  - Bayesian Classifier
  - Recurrent Neural Network with enforcement learning
- Integration of tools into a centralized application
- Automate creation of points to polygon via python script