# Task To be completed on 03/02/2025

ISSO requires Machine Learning (ML) analysis to power the dashboards with the following requirements:

## Machine Learning Requirements

* Computing clusters of points within physical proximity of each other. The comparison must be performed on over 1 million points in under 1 minute.
* Perform geographic clustering of identified point centroids to determine clusters. These clusters must be converted into geographic polygons.
* Perform social network clustering of points found to be in close proximity of each other.
* Model usual activity in an area of interest (AOI) using an ensemble of time series forecasting algorithms. The system must signal an alert when actual activity deviates outside an acceptable range. The modeling process must complete within 5 minutes.

The data ingestion process for the dashboards requires consuming commercially available data or ingesting from USG-provided APIs on Top Secret (TS) networks. The data must be stored in an open-source database optimized for geospatial and temporal data.

## Data Processing & Machine Learning Requirements

* Ingest multiple data files into a single data entity using Arrow and the reader package.
* Ingest data from Excel spreadsheets.
* Ingest JSON and GeoJSON data.
* Utilize dplyr and Arrow for data manipulation, including filtering, creating new columns, summarizing data.
* Data visualization using ggplot2, including scatterplots, histograms, bar charts, small multiples, line charts, and smoothing curves.
* Generalized Linear Modeling.
* Fitting boosted trees (gradient boosting models).
* Cross-validation for model performance tuning.
* Automating machine learning pipelines using the tidymodels suite: recipes, parsnip, workflows, yardstick, dials, tune, sample.
* Automated reporting with Quarto.
* Slideshow generation with Quarto.
* Consuming APIs with httr2.
* Performing analysis on large datasets using Arrow and DuckDB.
* Other topics as determined by ISSO.

## Geospatial Dashboard Requirements

ISSO requires exposing advanced geospatial capabilities using modern open-source tools—such as PostGIS, Geos, DuckDB Spatial, and Simple Features in R—to end users with no coding ability.

The dashboard must allow users to upload GeoJSON files and perform geospatial operations through a GUI (Graphical User Interface).

### Performance Requirements

* The dashboard must compute buffers of user-defined distances around polygons (on the order of 25,000 polygons in under 15 seconds).
* It must determine the intersection between arbitrary points and polygons (on the order of 1,000,000 points within 25,000 polygons in under 5 seconds).
* Users must be able to export results as Comma Separated Values (CSV) or GeoJSON.