## 1.1 Pattern of Life Integrated System (POLIS) (TRL 8)

The primary objective of the POLIS project was to develop an advanced, scalable technology solution that identifies and models POL over both hard (i.e., numerical / statistical data often with strong spatial-temporal attributes and typically from a sensor) and soft data (i.e., often textual data lacking strong spatial-temporal attributes but often containing contextual attributes). POLIS successfully combined semantic (top-down) techniques with statistical and machine-learning (bottom-up) techniques in order to identify and model POL "normalcy." With normalcy patterns modeled, POLIS proved successful in identifying anomalies (i.e., non-patterned actions that did not follow normalcy patterns) and anti-patterns (patterned actions that did not follow normalcy pattern and/or could be specifically identified as undesired) and alerting users or other systems.

POLIS, and subsequent projects that have enhanced POLIS, used five models to identifying pattern normalcy, depending on the data: (1) Spatial structure – Data has defined lat/lon or other strong spatial attribute but lacks other correlation attributes, which results in a spatial normalcy pattern without time or semantic context; (2) Temporal structure – Data has strong temporal attribute but lacks other correlation attributes, which results in a temporal normalcy pattern without spatial or semantic context; (3) Event-linkage structure – A temporal structure that includes spatial and/or semantic attributes sufficient to link temporal objects by more than just time (e.g., cause and effect), which results in linked events across time with additional spatial and/or semantic context; (4) Network structure – Data has strong attributes for at least two of temporal, spatial and/or semantic resulting in interconnected normalcy patterns that span time, space and semantic context; (5) Statistical structure – Data has temporal, spatial and/or semantic attributes but lacks sufficient data to define a structure (network or other), which results in normalcy patterns with spatial, temporal and/or semantic context based on statistical groupings. **Figure 1** provides a depiction of these five models, with the *x* and *y* axes depicting spatial concepts, a *Time* axis depicting temporal concepts, and lines between nodes depicting semantic attributes between nodes.

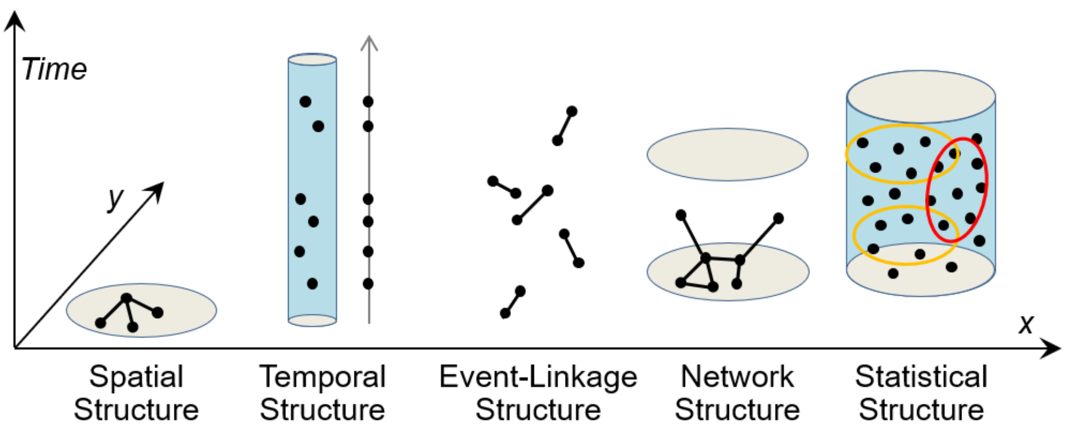


**Figure 1. POL Structures in POLIS.** Applying temporal, spatial and semantic attributes to discover pattern normalcy structures based on the data.

POLIS incorporated several techniques for POL analytics including the application of event frequency analysis and sentiment shift analysis (including applying Relative Strength Index) in Coded Event data (soft data). POLIS also integrates data clustering algorithms including Density-based Spatial Clustering of Applications with Noise (DBSCAN) [5] [6], kernel density estimation (KDE), and k-means clustering to analyze POL for Vehicle Track data (hard data).

POLIS has been applied to many use cases; to demonstrate flexibility in support of multiple use cases, the Phase II prototype supported two particular use cases: (1) identifying significant events between state actors in Global Database of Events, Language, and Tone (GDELT) data while identifying significant shifts in GDELT Average Tone and Stability (Goldstein Scale) values and (2) Identifying normal and abnormal ship/vessel tracks over Automatic Identification System (AIS) data. These two use cases covered the first four POL structure types and were very successful. To highlight the flexibility of POLIS to discover normalcy patterns in multiple types of data across the five POL structures, **Figure 2** shows the POL of a vessel (ship) based on hard and soft data about that ship, including AIS and ASMS data.

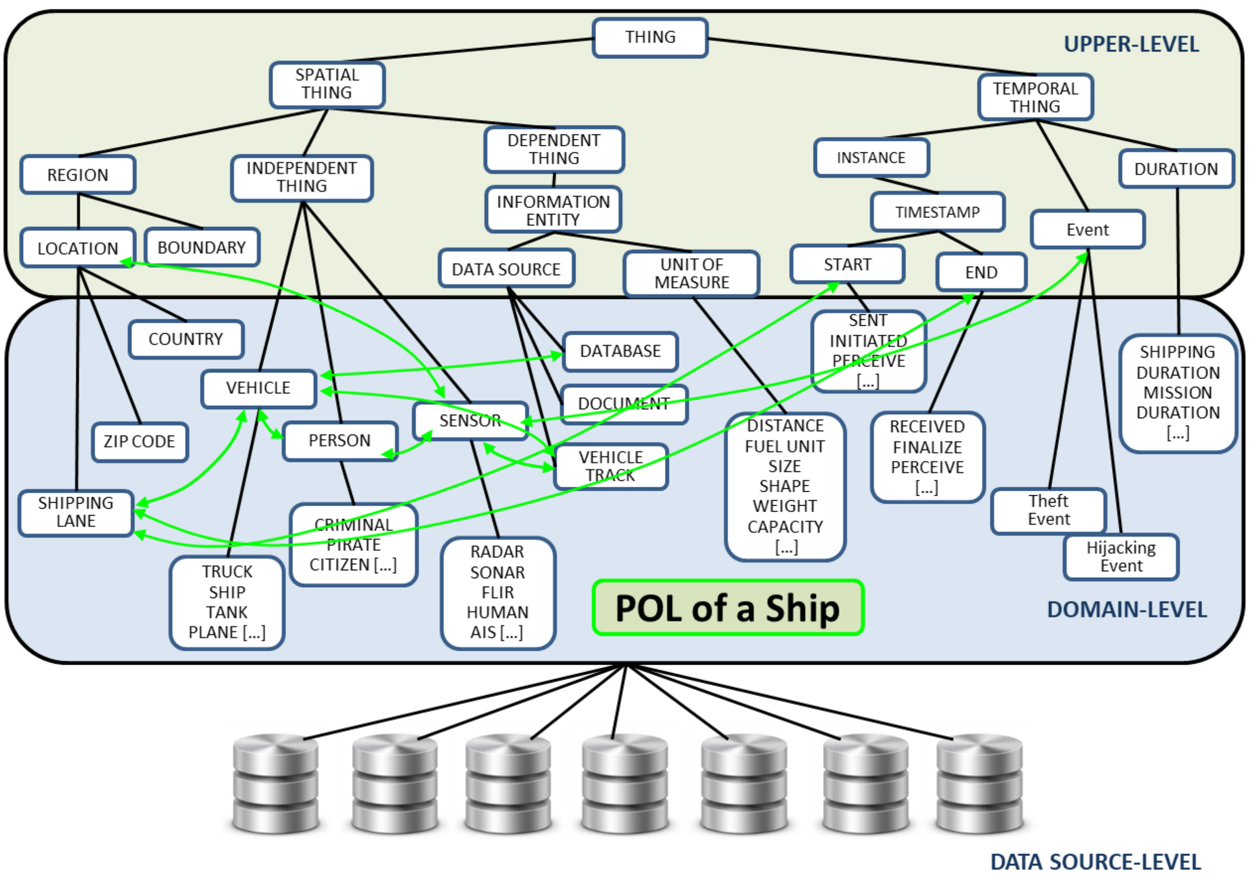


**Figure 2. Vessel (Ship) Pattern of Life.** Temporal, spatial and (to a lesser degree) semantic attributes are used to define normalcy patterns.

The remaining content in this section describes the details of the algorithms used in POLIS to address spatial, temporal, and event-linkage POL structures.

### 1.1.1 Spatial-temporal POL Analysis Methods

Three estimation and classification analysis methods were explored as a proof of concept to establish normalcy on ship location: KDE, DBSCAN, and *k*-means clustering. These methods are relevant to any spatially distributed data, including those with temporal and/or semantic attributes, requiring only a set of *n*-dimensional coordinates to use these POL methods.

#### 1.1.1.1 POLIS Kernel Density Estimation (KDE)

KDE is a non-parametric statistical estimator for the probability density function of a random variable. The KDE is capable of taking a finite set of discretely sampled points, adjusting the parameter set based on the size of the training data, and estimating the likelihood of those sampled points taking on certain values. In other words, a two-dimensional instance of KDE can take ship discretely reported locations from the AIS data and return a probability density function capturing the likelihood of a ship location appearing within a certain region (i.e., location density).

#### 1.1.1.2 POLIS Density-based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a data clustering algorithm that groups points together based on the closeness of points and areas with many neighbors. Points lying in lower density regions are identified as outliers which represent points that fall far enough from the core points and have neighboring points that are below the minimum number of points for a cluster. The parameter *eps* defines the reachability of clusters based on density and parameter *minPts* is minimum number of points required to form a dense cluster. Additionally, core points (points that are reachable from a density grouping essential for clustering) must have *minPts* within *eps* distance. The number of clusters generated by DBSCAN are based on these parameters, allowing the algorithm freedom to identify an appropriate number of clusters.

#### 1.1.1.3 POLIS K-Means Clustering

The *k*-means clustering is an alternative cluster analysis tool which partitions a set of observations into *k* distinct subsets. The method identifies *k* centroids in which clusters of points are assigned. The objective is to identify *k* clusters in which their respective centroids relative to their assigned observations minimize the L2 norm across the entire set of observations. The problem is NP-Hard but a series of approximation heuristics and algorithms exist to compute locally optimal solutions. The 2-D observations can be decomposed into a series of cells which can be represented as Dirichlet tessellation.

#### 1.1.1.4 POLIS Prior Success with Spatial-Temporal POL

Modus Operandi prototyped POLIS for spatial-temporal POL (i.e., addresses first three POL structures) analysis using real-world AIS data provided by the United States Coast Guard Navigation Center (USCG NAVCEN). The prototype successfully provided situational understanding on large amounts of spatially distributed data with automated alerting of abnormal behavior. **Figure 3** illustrates all shipping traffic captured in November 2015 over the South China Sea with additional map layers (**left image**) for Exclusive Economic Zones, identifying each country's rights to the zones in the South China Sea and the National Geospatial Agency's Anti-Shipping Activity Messages (ASAMs), showing all reported anti-shipping activity in the area since 1978. The green bounding box outlines the available USCG NAVCEN AIS data of interest at the given time, enabling analysts to see a specific ship’s journey through the region over time.

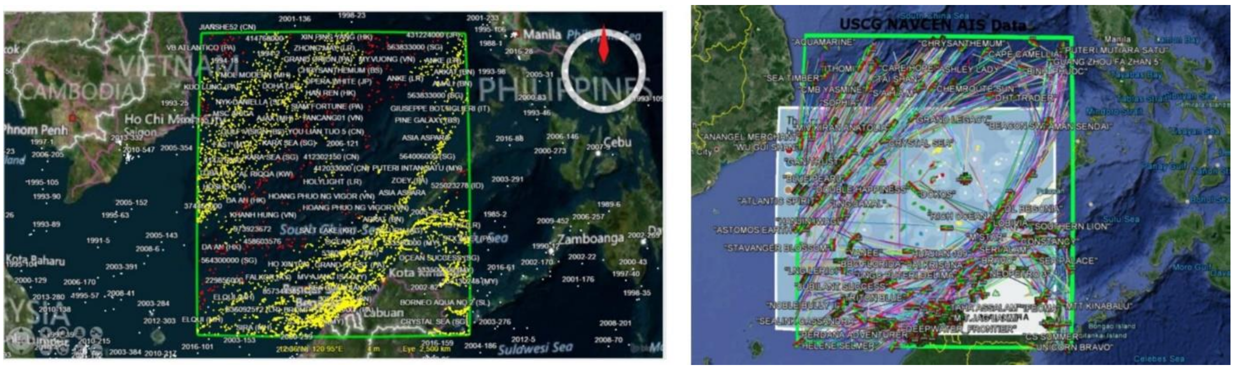


**Figure 3. POLIS DBSCAN Implementation on AIS Data from November 2015 in South China Sea.** Ship positions are colored according to whether they are near a minimum number of ships within a certain distance (yellow) or not (red). The green bounding box outlines the borders of the region being analyzed. The map also includes ASAM messages for situational awareness of prior issues.

Although there is a relatively small proportion of data available for the region, the ship locations are spread out over a larger area allowing for interesting geospatial analysis. **Figure 4 (top left)** shows the POLIS KDE analysis of the data from during November of 2015, comprised of 69,813 AIS broadcasts. The mean location of the data is shown as a black star on the map which is shifted toward the higher density regions off the coast of Malaysia. The high-density regions are located in the area associated with high-volume Malaysian ports which is expected given that it is a local hub for ships and the ships are often broadcasting while in or near port. Since the ship positions are not confined to the area immediately surrounding the ports, the heat map capturing the location density also has significantly noticeable red trails which represent frequently used shipping lanes.

**Figure 4 (top right)** illustrates all ship positions contained in the November 2015 data file. The POLIS DBSCAN algorithm identified three separate clusters: normal density locations shown in black and two clusters with divergent densities shown as green and red dots. The area with the differently clustered points and the absence of ship broadcasts at sea are in some part due to the smaller islands in the South China Sea although the map does not adequately show them. Both clusters are AIS broadcasts from Chinese vessels operating in close proximity to Chinese occupied islands. DBSCAN was immediately able to identify this clustering scheme without any parameter tuning which suggests it may be a relatively easy methodology to deploy across multisource data sets.

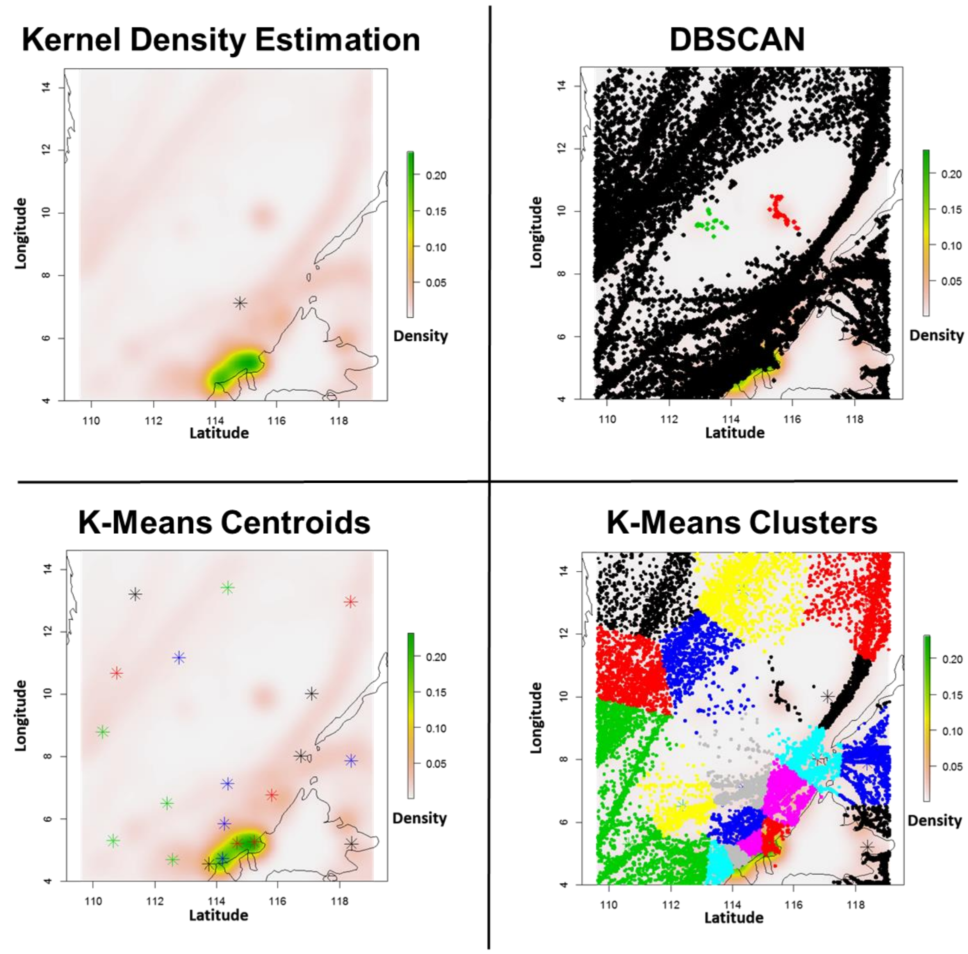


**Figure 4. POL Algorithm Complementary Capabilities.** Geospatial POL methods demonstrated across a month's worth of ship track data from the South China Sea

This clustering algorithm is effectively stating, "given all ship locations in this area from November 2015, the clusters identified by green and red points do not share a similar density to the majority of reported locations." In effect, most of the black points are considered "normal" since many of the locations are reported in commonly recurring locations. Using spatial density as an indicator, the divergent clusters may be regarded as abnormal since they are located in a spatial void on the map. Adding in contextual information, this is known to be the Spratly Islands which could support the claim for further analysis or alerting on those vessels. The overall approach of the macro ship analysis is to identify a small subset ships that exert "macro-abnormal" behavior such as these ships in the abnormal density clusters. The micro ship analysis intends to take macro-abnormal ships and conduct additional micro-analysis at a ship by ship case to determine if the behavior is significantly anomalous.

The POLIS *k*-means analysis shown in **Figure 4 (bottom left)** is an example of 20 centroids (k=20) that minimize total distance to any given AIS location for the entire month of data. This clustering method is not necessarily useful for anomaly detection in this context, but it provides some interesting results which may be beneficial in partitioning the data. There are more centroids located in high-density regions to minimize the total distance between all points, and the centroids also frequently appear in or near shipping lanes. Observing how these centroids change over time can provide some measurement as to how the volume of AIS traffic shifts over time.

The corresponding clusters to the POLIS *k*-means are shown in **Figure 4** (**bottom right**) and the color of each cluster aligns with the color of its centroid. This clustering is an alternative partitioning means to an approach such as discretizing the South China Sea into a set of grids. This clustering method takes into consideration the closeness of the tracks, and is easily configurable for multiple values of *k*. This is useful for localized analysis of ship broadcasts, assigning collocated ship locations to a localized cluster.