Sandia National Laboratories

Exceptional service in the national interest

DB-DRIFT

Concept drift aware density-based anomaly detection for maritime trajectories

Amelia Henriksen, Sandia National Laboratories

Sensor Signal Processing for Defence, 2023

September 13, 2023

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

MARITIME ANOMALY DETECTION

BAD THINGS HAPPEN TO PEOPLE ON THE OCEAN

Cargo loss

Photo by Petty Officer 3rd Class Matthew West (10.28.2018), DVIDS

Sinking

Photo by U.S. Coast Guard District 8, (12.21.2002), DVIDS

Grounding

Photo by U.S. Coast Guard District 7 PADET Tampa Bay (08.26.2012) DVIDS

Medical Emergencies

James Brickwood/SMH 07.05.2022, 9News

Losing power/propulsion

Dakota Santiago, 07.09.2023, The New York Times

Fire

PEOPLE DO BAD THINGS ON THE OCEAN

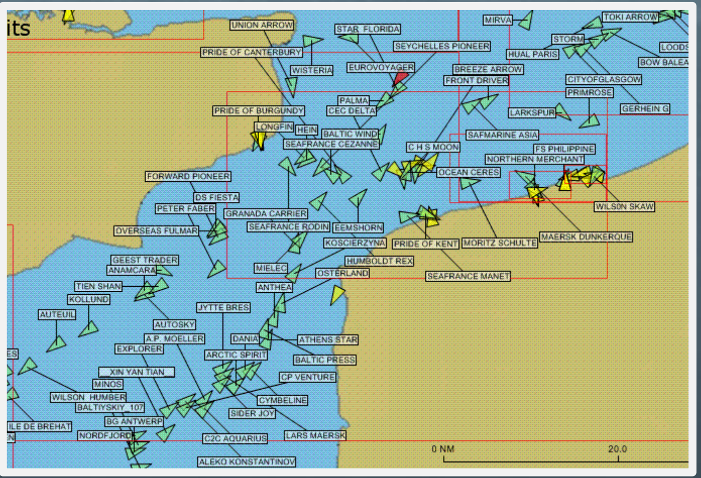
Trade sanction dodging

Image © Alex Hofford / Greenpeace.

Illegal fishing

Vessel type spoofing

Smuggling

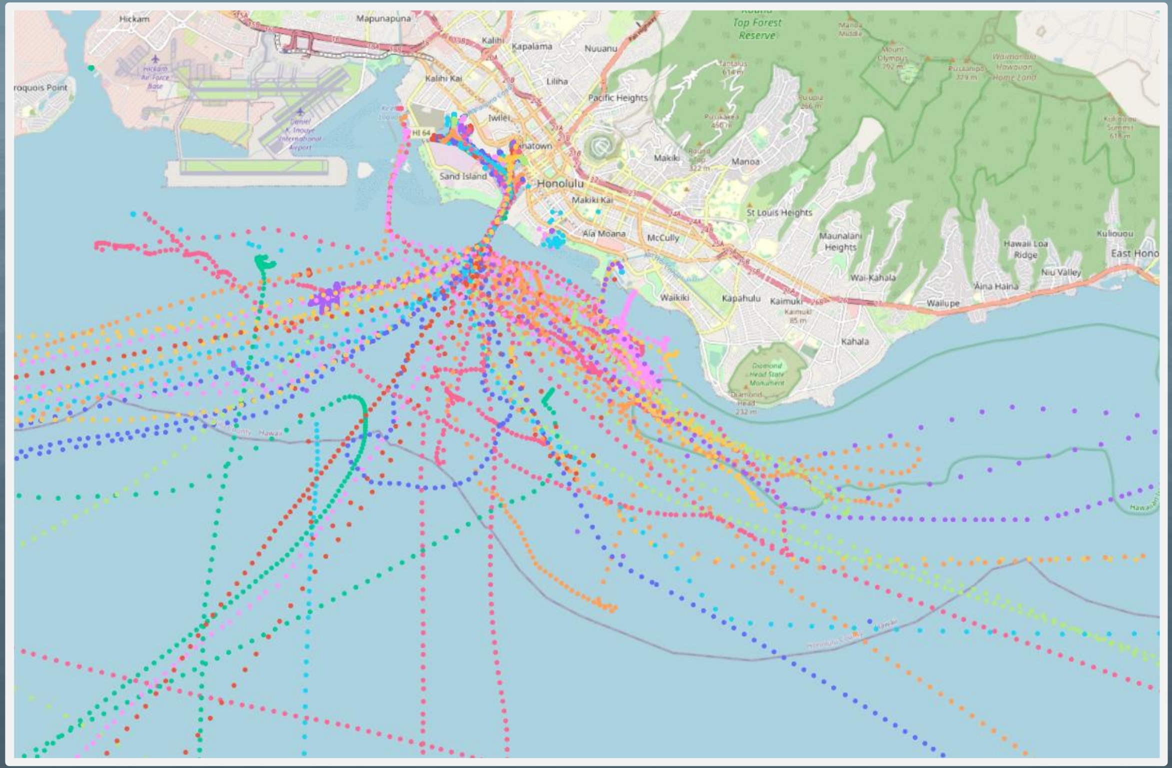


False route reporting
(High collision risk)

Terrorism

HOW DO WE FIND ANOMALIES?

- Many data sources, both public and proprietary, for monitoring vessel tracks.
- Example: All large ships are required by international law to be equipped with **automatic identification system data (AIS)**



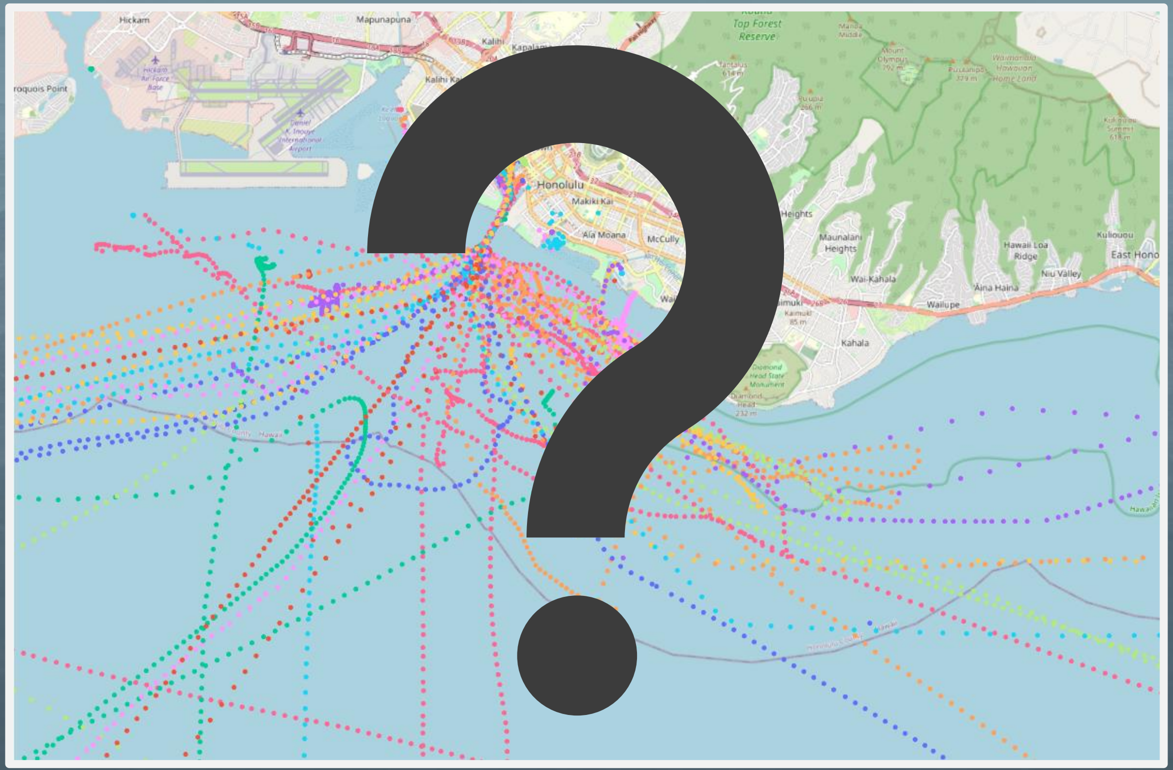
Henriksen, Amelia. (2023). HawaiiCoast_GT: Curated AIS for Hawaii's coast correlated with ground truth incidents (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8253611

HOW DO WE FIND ANOMALIES?

The Problem:

Maritime vessel data is

- Unlabelled
- Large
- Noisy
- Prone to changes in the underlying distribution



Henriksen, Amelia. (2023). HawaiiCoast_GT: Curated AIS for Hawaii's coast correlated with ground truth incidents (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8253611

HOW DO WE FIND ANOMALIES?

The Common Problem:

Maritime vessel data is

- **Unlabelled**
- Large
- Noisy
- Prone to changes in the underlying distribution

Anomaly detection algorithms for vessel tracks are largely **unsupervised** (UAD)

HOW DO WE FIND ANOMALIES?

The Common Problem:

Maritime vessel data is

- Unlabelled
- **Large**
- **Noisy**
- Prone to changes in the underlying distribution

Expensive (and sometimes impossible) for experts* to assess all tracks.

* Domain experts OR expensive expert algorithms

HOW DO WE FIND ANOMALIES?

The Problem:

Maritime vessel data is

- Unlabelled
- Large
- Noisy
- **Prone to changes in the underlying distribution**

Basic UAD:
Assume the majority of samples are **normal** and identify **outliers.**

HOW DO WE FIND ANOMALIES?

The Problem:

Maritime vessel data is

- Unlabelled
- Large
- Noisy
- **Prone to changes in the underlying distribution**

Basic UAD:
Assume the majority of samples are **normal** and identify **outliers.**

If the norm changes, our model needs to change with it.

HOW DO WE FIND ANOMALIES?

The Problem:

Maritime vessel data is

- Unlabelled
- Large
- Noisy
- **Prone to changes in the underlying distribution**

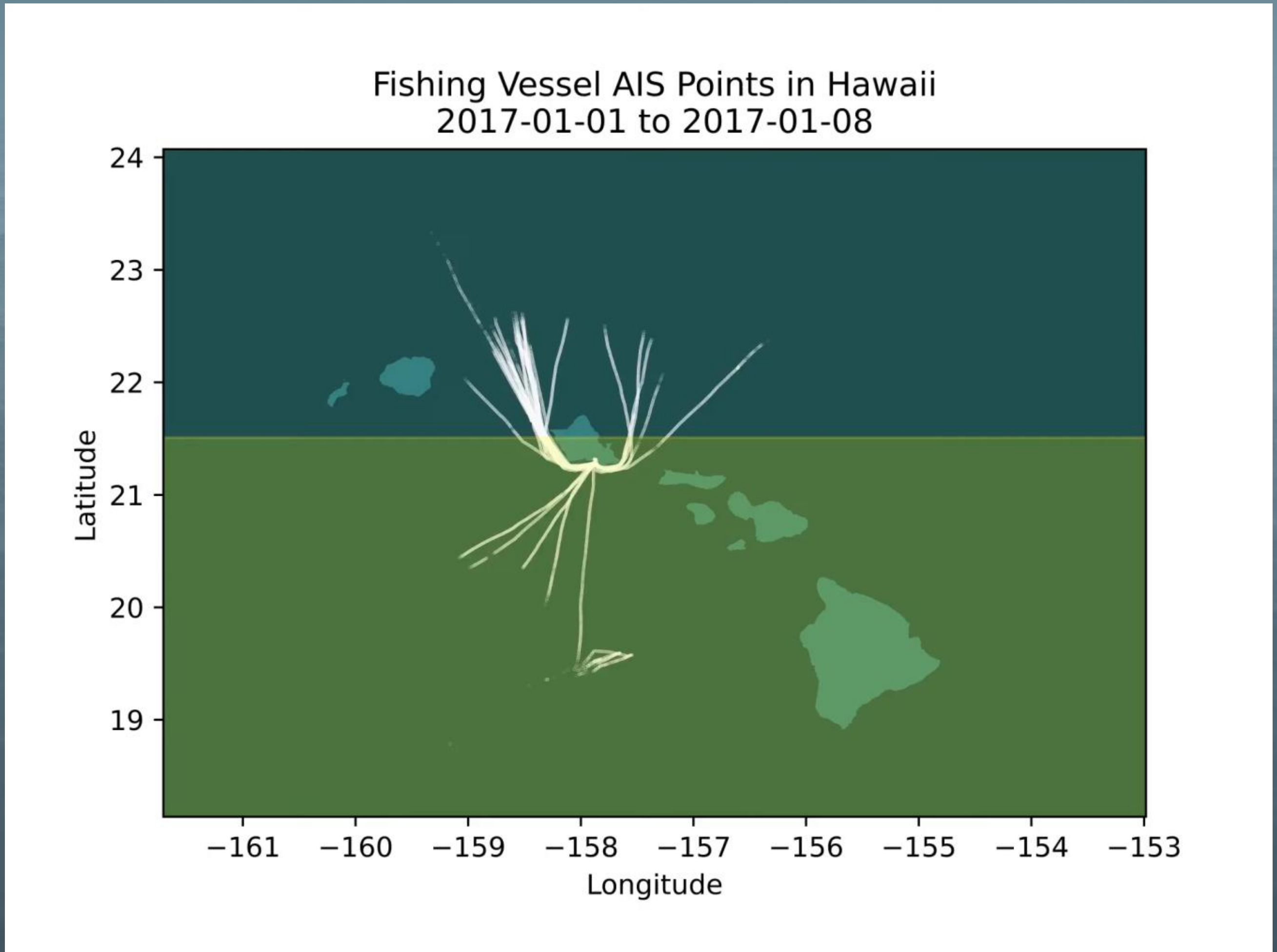
This is **concept drift.**

Today we focus on two kinds of drift:

1. Gradual drift
2. Seasonal drift

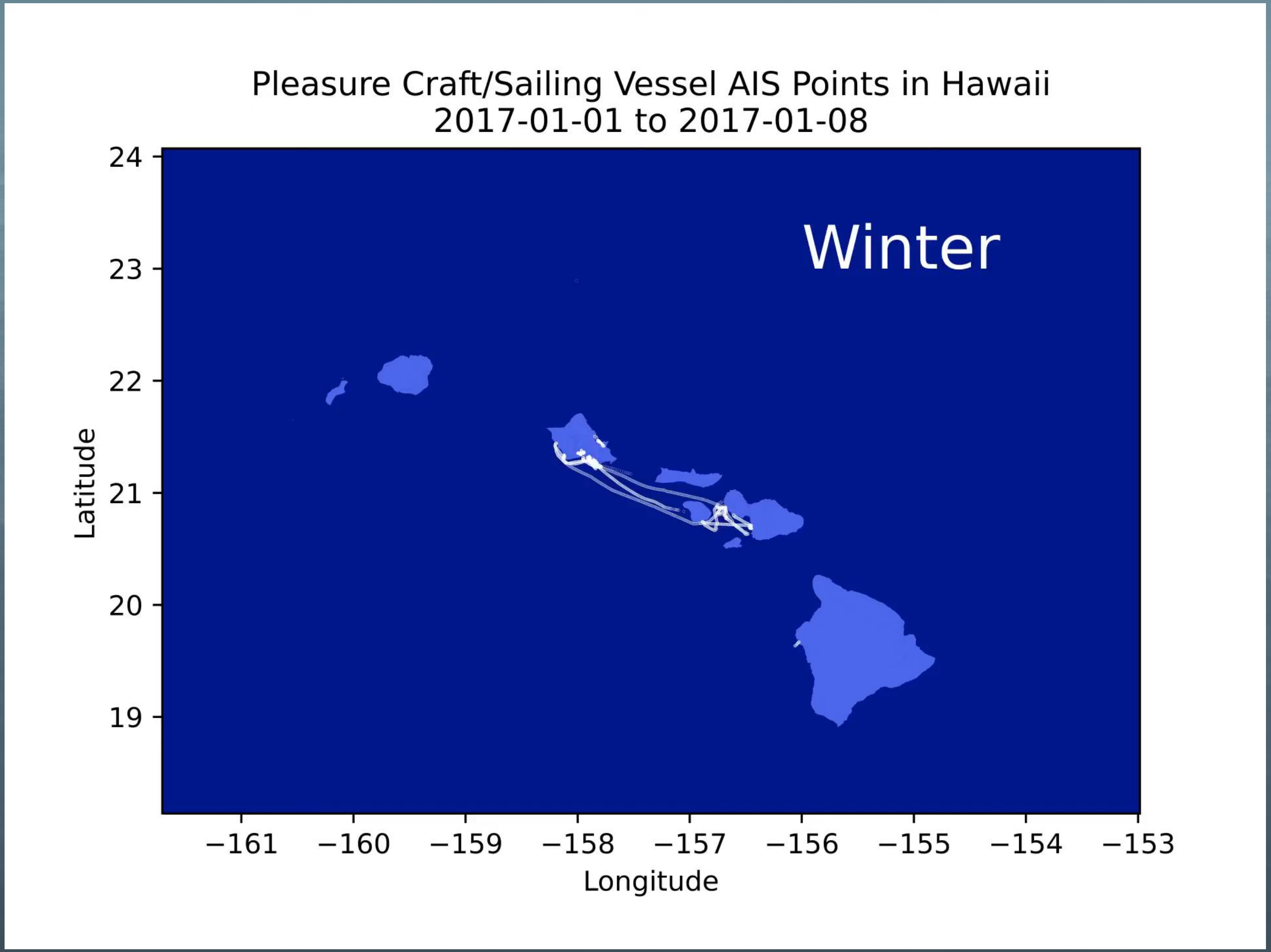
GRADUAL CONCEPT DRIFT

- One of the most well understood forms of concept drift
- Describes the slow, consistent evolution of data over time.



SEASONAL CONCEPT DRIFT

- Describes patterns that appear repeatedly in the data in a periodic way
- Vessel movements are affected by the earth’s literal meteorological seasons.



ACCOUNTING FOR MULTIPLE TYPES OF CONCEPT DRIFT

- Almost all modern vessel track UAD pipelines don't account for any concept drift.

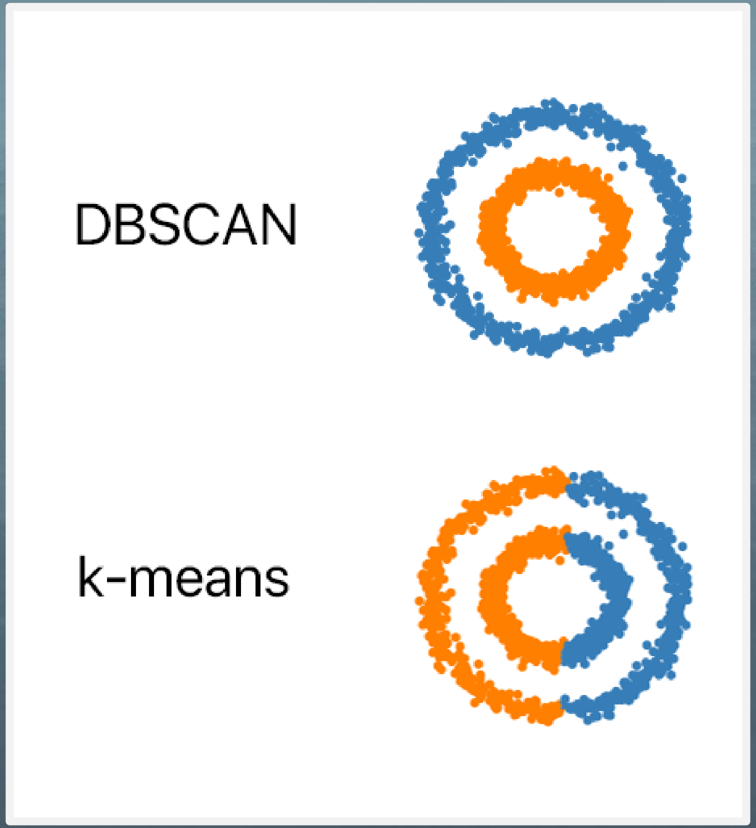
- In the few cases where it is incorporated, only gradual drift is addressed.

- How do we solve this problem?

DBSCAN AND MARITIME ANOMALY DETECTION

DBSCAN: Density-based spatial clustering of applications with noise [1]

- Can automatically identify outliers
- Has few hyperparameters
- Does not need a pre-set number of clusters (unlike k-means).



https://github.com/NSHipster/DBSCAN

[1] M. Ester, H. P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise.” in kdd, vol. 96, no. 34, 1996, pp. 226-231

DBSCAN AND MARITIME ANOMALY DETECTION

DBSCAN: Density-based spatial clustering of applications with noise [1]

- Can automatically identify outliers
- Has few hyperparameters
- Does not need a pre-set number of clusters (unlike k-means).

Examples in Maritime Anomaly Detection:

1. Arguedas, Virginia Fernandez, Fabio Mazzarella, and Michele Vespe. "Spatio-temporal data mining for maritime situational awareness." *OCEANS* 2015-Genova. IEEE, 2015.
2. Botts, Carsten. "An Alternative Metric for Detecting Anomalous Ship Behavior Using a Variation of the DBSCAN Clustering Algorithm." *arXiv preprint arXiv:2006.01936* (2020).
3. Daranda, Andrius, and Gintautas Dzemyda. "Navigation decision support: Discover of vessel traffic anomaly according to the historic marine data." *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL* 15.3 (2020).
4. El Mekkaoui, Sara, Abdelaziz Berrado, and Loubna Benabbou. "Automatic Identification System Data Quality: Outliers Detection Case."
5. Ferreira, Martha Dais, Jessica NA Campbell, and Stan Matwin. "A novel machine learning approach to analyzing geospatial vessel patterns using AIS data." *GIScience & Remote Sensing* 59.1 (2022): 1473-1490.
6. Fu, Peiguo, et al. "Finding abnormal vessel trajectories using feature learning." *IEEE Access* 5 (2017): 7898-7909.
7. Han, X., C. Armenakis, and M. Jadidi. "DBSCAN optimization for improving marine trajectory clustering and anomaly detection." *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 43 (2020): 455-461.
8. Huang, Jie, et al. "Research on Real-Time Anomaly Detection of Fishing Vessels in a Marine Edge Computing Environment." *Mobile Information Systems* 2021 (2021).
9. Liu, Bo. *Maritime traffic anomaly detection from AIS satellite data in near port regions*. Diss. 2015.
10. Liu, Bo, et al. "Ship movement anomaly detection using specialized distance measures." *2015 18th International Conference on Information Fusion (Fusion)*. IEEE, 2015.
11. Pallotta, Giuliana, Michele Vespe, and Karna Bryan. "Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction." *Entropy* 15.6 (2013): 2218-2245.
12. Prasad, Priank, Vishesh Vatsal, and Rajarshi Roy Chowdhury. "Maritime Vessel Route Extraction and Automatic Information System (AIS) Spoofing Detection." 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE, 2021.
13. Rintoul, Mark Daniel. ML for Trajectories: Bridging the Gap from Computer to Analyst with Tracktable. No. SAND2020-12646C. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2020.
14. Rong, H., A. P. Teixeira, and C. Guedes Soares. "Data mining approach to shipping route characterization and anomaly detection based on AIS data." Ocean Engineering 198 (2020): 106936.
15. Sørensen, Kristian Aalling. "Automatic Identification System tracking of ships using Neural Networks and correlation with satellite images." (2021).
16. Szarmach, Marta, and Ireneusz Czarnowski. "Multi-Label classification for AIS data anomaly detection using wavelet transform." *IEEE Access* 10 (2022): 109119-109131.
17. Varlamis, Iraklis, Konstantinos Tserpes, and Christos Sardianos. "Detecting Search and Rescue missions from AIS data." *2018 IEEE 34th international conference on data engineering workshops (ICDEW)*. IEEE, 2018.
18. Wei, Zhaokun, Xinlian Xie, and Xiaoju Zhang. "Maritime anomaly detection based on a support vector machine." *Soft Computing* 26.21 (2022): 11553-11566.
19. Zhang, Tao, et al. "ATeDLW: Intelligent Detection of Abnormal Trajectory in Ship Data Service System." 2021 IEEE International Conference on Services Computing (SCC). IEEE, 2021.
20. Zhang, Tao, Shuai Zhao, and Junliang Chen. "Ship trajectory outlier detection service system based on collaborative computing." 2018 IEEE World Congress on Services (SERVICES). IEEE, 2018.
21. Zhang, Yuanqiang, and Weifeng Li. "Dynamic maritime traffic pattern recognition with online cleaning, compression, partition, and clustering of AIS data." *Sensors* 22.16 (2022): 6307.
22. Zhao, Liangbin, and Guoyou Shi. "Maritime anomaly detection using density-based clustering and recurrent neural network." The Journal of Navigation 72.4 (2019): 894-916.

[1] M. Ester, H. P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise.” in kdd, vol. 96, no. 34, 1996, pp. 226-231

DBSCAN AND MARITIME ANOMALY DETECTION

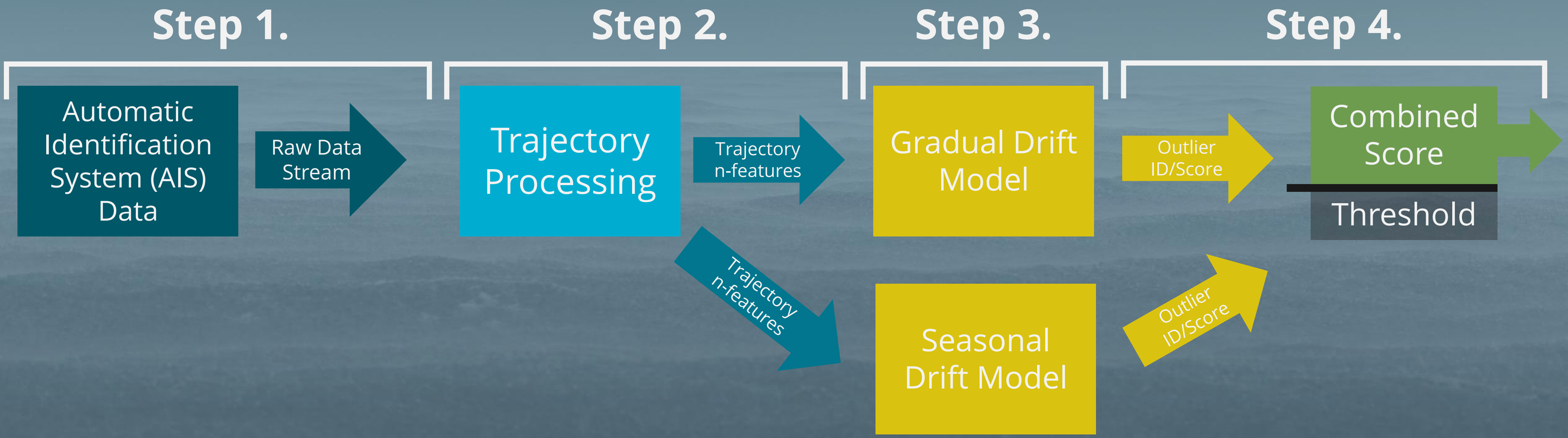
DBSCAN: Density-based spatial clustering of applications with noise [1]

- Can automatically identify outliers
- Has few hyperparameters
- Does not need a pre-set number of clusters (unlike k-means).

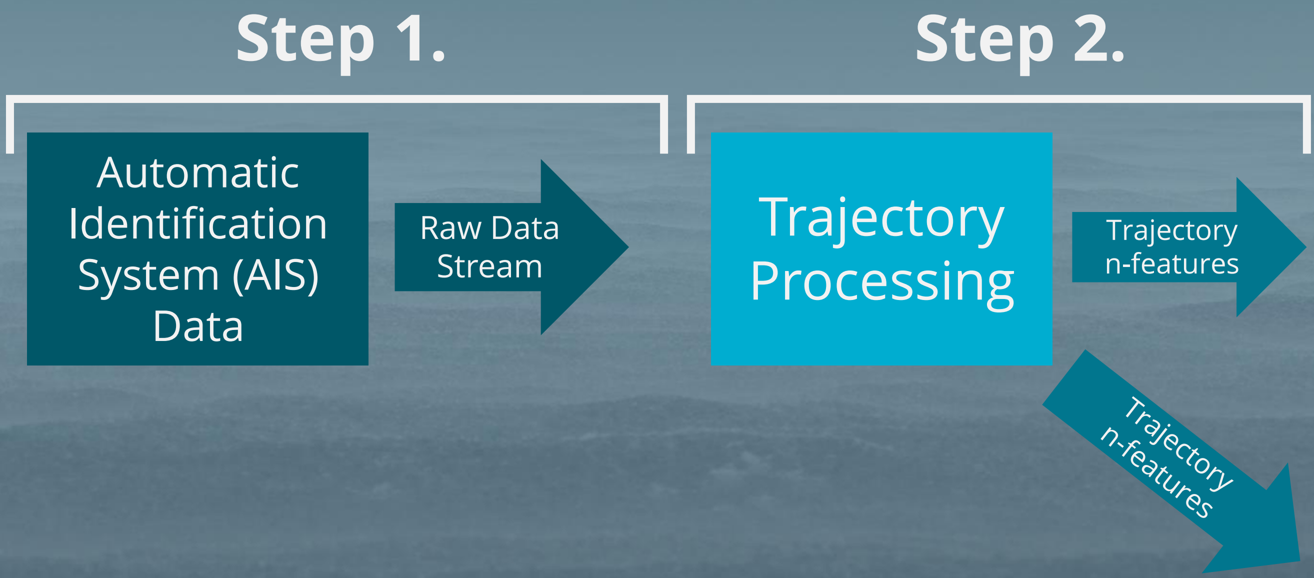
1. DBSCAN is still fundamentally a static method.
2. We'd like to incorporate multiple forms of drift

[1] M. Ester, H. P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in kdd, vol. 96, no. 34, 1996, pp. 226-231

THE DB-DRIFT ALGORITHM



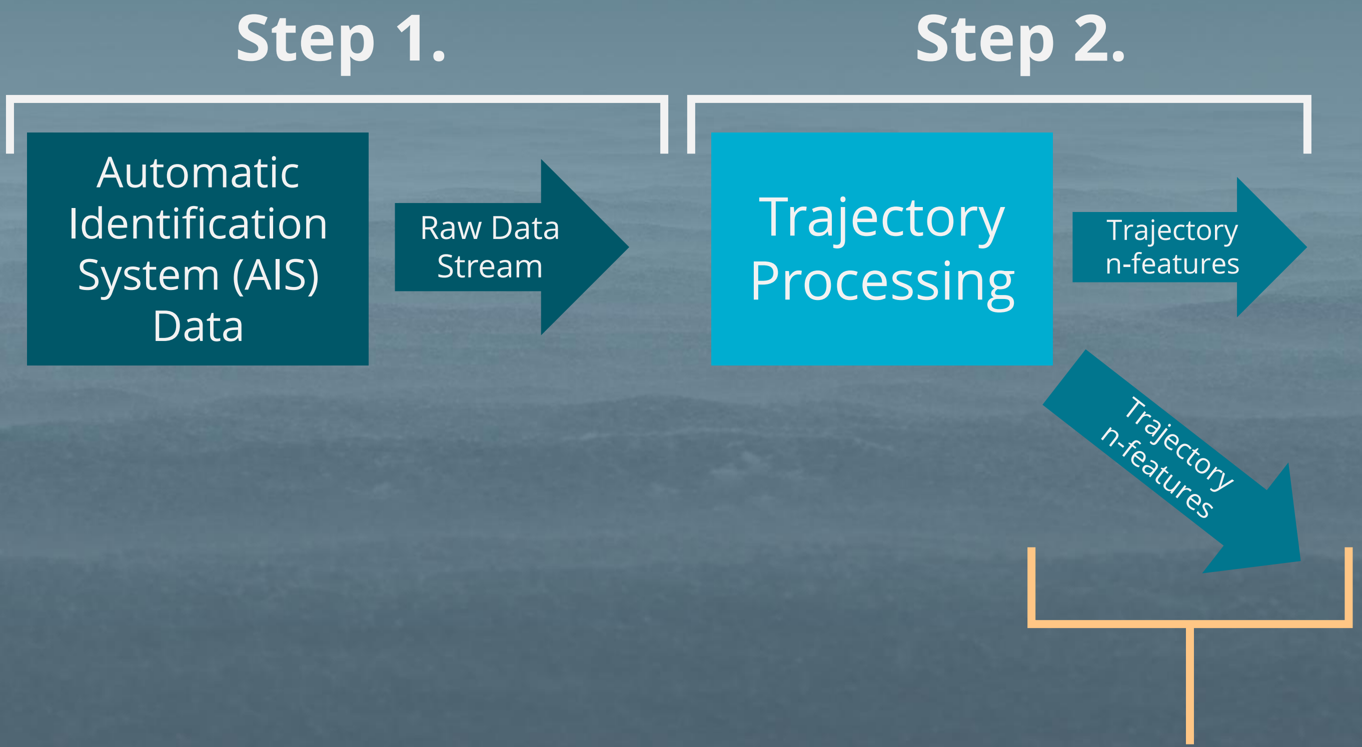
THE DB-DRIFT ALGORITHM



This setup starts most UAD pipelines on trajectories.

The algorithm must process an incoming stream of trajectories (or trajectory segments.)

THE DB-DRIFT ALGORITHM

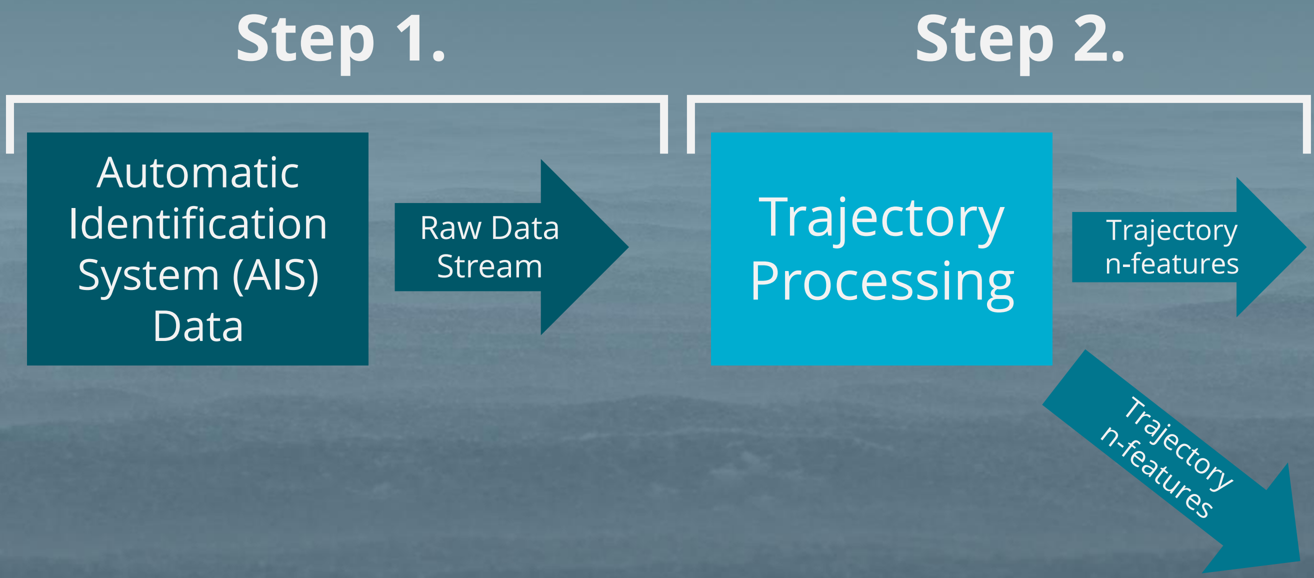


This setup starts most UAD pipelines on trajectories.

The algorithm must process an incoming stream of trajectories (or trajectory segments.)

Awesome resource: "A study on the geometric and kinematic descriptors of trajectories in the classification of ship types." by Tavakoli, Peña-Castillo, and Soares.

THE DB-DRIFT ALGORITHM



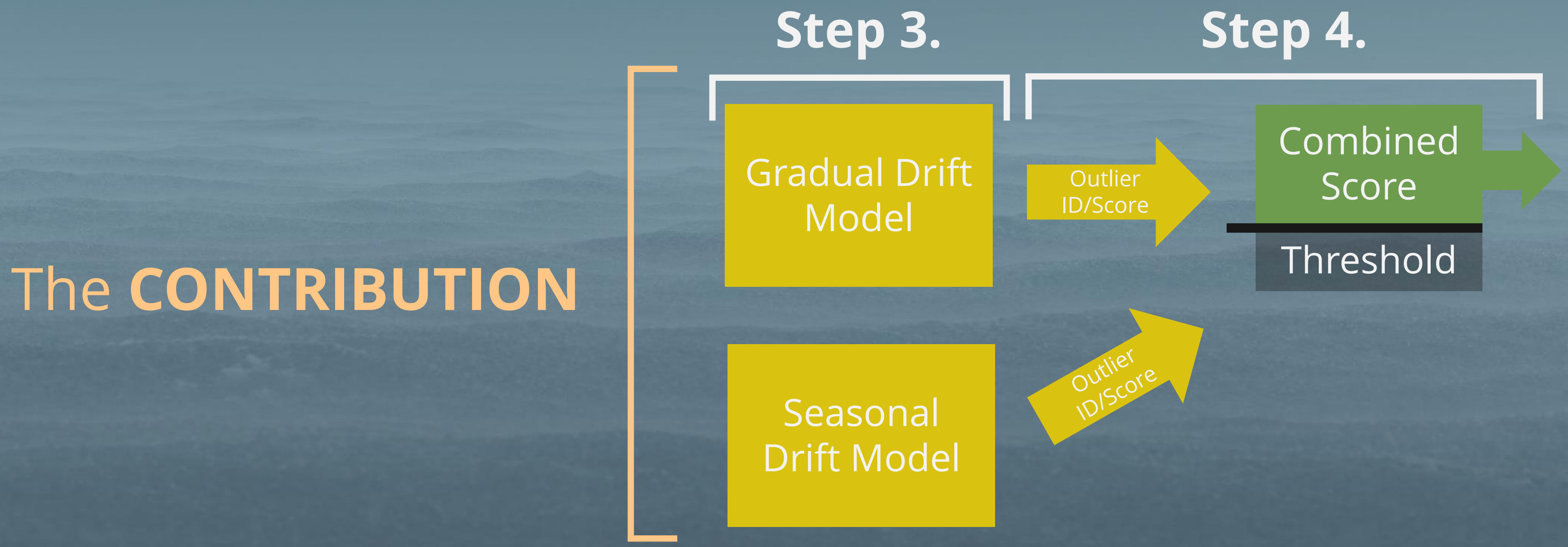
This setup starts most UAD pipelines on trajectories.

The algorithm must process an incoming stream of trajectories (or trajectory segments.)

Sandia National Labs' Tracktable is our **BEST FRIEND** for this stage in the pipeline.

https://tracktable.sandia.gov/

THE DB-DRIFT ALGORITHM



GRADUAL DRIFT

We want:
- Model that works quickly on a stream with low overheads

- Model that emphasizes more recent "normal behavior" over past behavior.

Naïve approach: Simple sliding window model

GRADUAL DRIFT

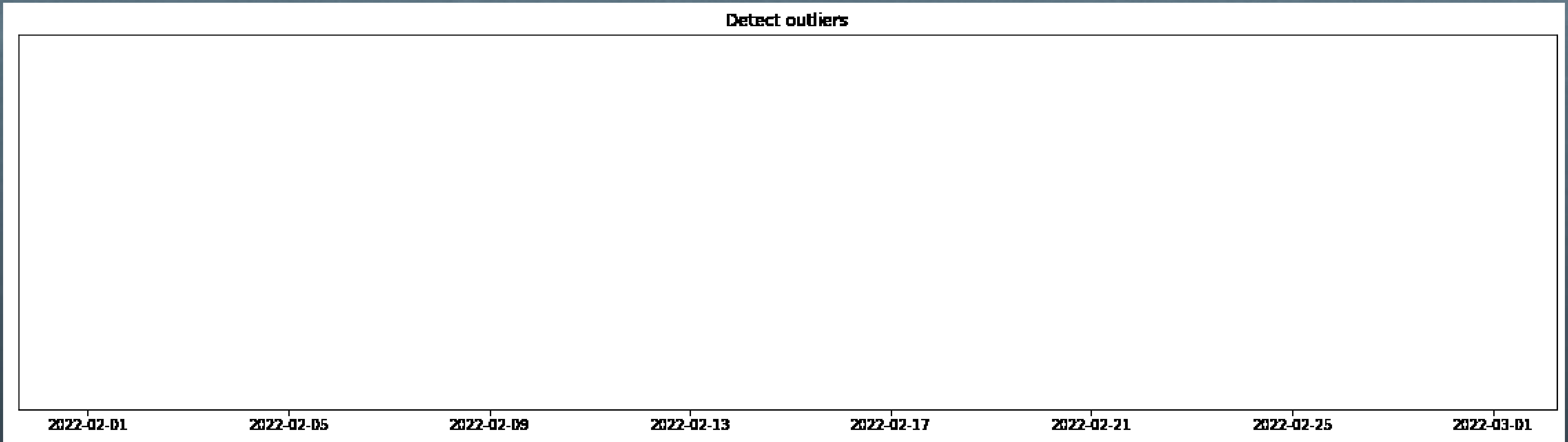
- Better approach: Damped window model.
  - Fades (re-weights) old samples as new samples arrive
  - Controlling the fade factor lets you control the rate at which the model evolves.

$$w_t(x) = 2^{-\lambda(t - T_0(x))}$$

$w_t(x)$: Weight at time t for sample x

λ: Fade factor

$T_0(x)$: Arrival time for sample x



GRADUAL DRIFT

We choose DenStream [1] as our core model:

1. Uses the damped window model
2. It can easily identify outliers in real time
3. It has low memory requirements
4. It requires very short burn in period (typically only a few days worth of data)

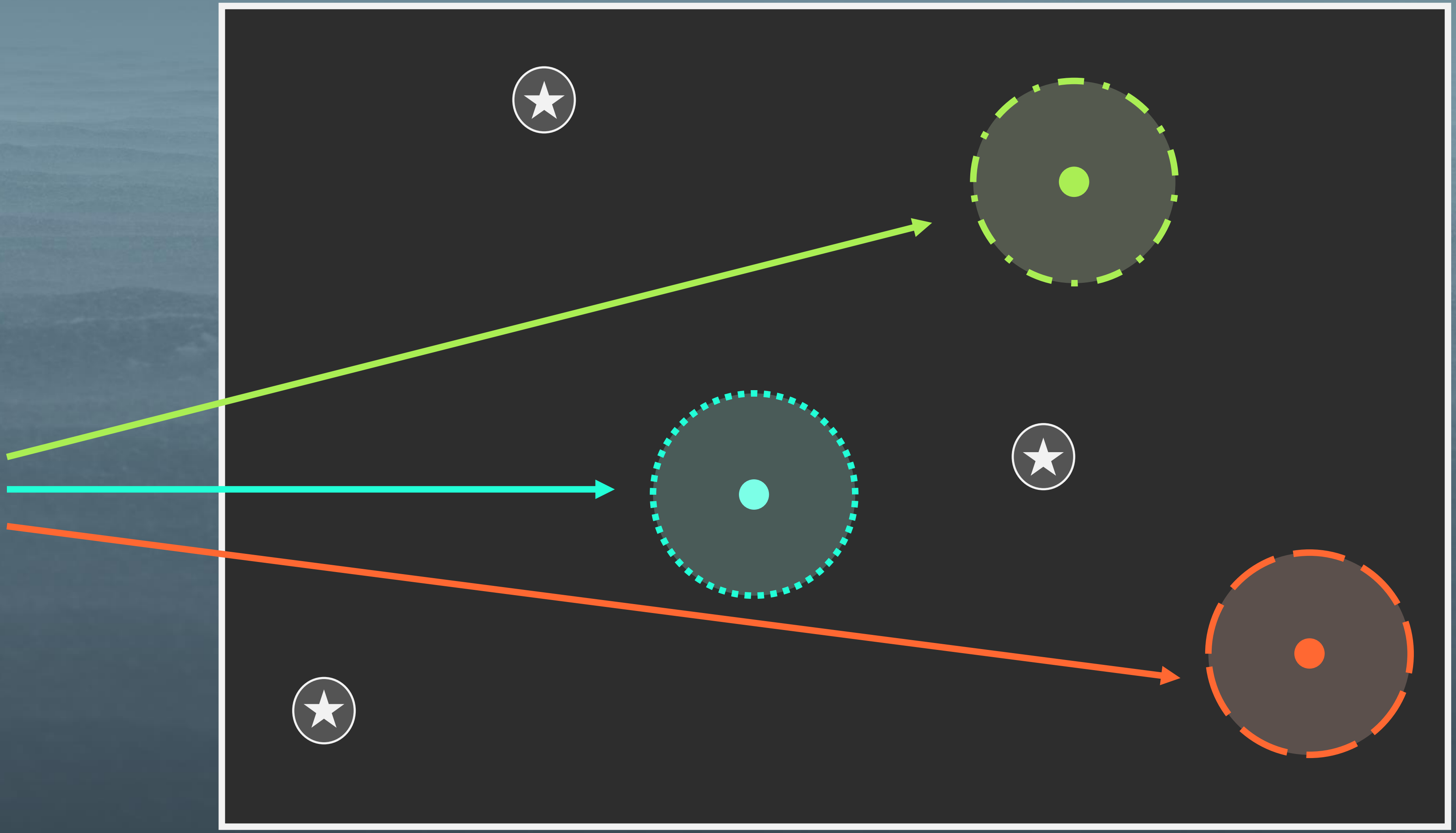
Algorithm 1 Denstream for outlier detection at time t

Parameters: Max radius ϵ, minimum p-microcluster weight $\mu\beta$, pruning stepsize T_p
$\lambda \leftarrow 1/T_p * log_2(\beta\mu/(\beta\mu - 1))$
for each sample x s.t. $T(x) = t$ **do** ▷ Merge step
  Find the nearest p-microcluster $p_t^* \in P_t$.
  if radius of $\{p_t^*, x\} \leq \epsilon$ **then**
    Add x to p_t^*
  else
    Report the outlier score as $\min_{p_t^* \in P_t} \|x - c(p_t^*)\|$
    Find the nearest o-microcluster $o_t^* \in O_t$
    if radius of $\{o_t^*, x\} \leq \epsilon$ **then**
      Add x to o_t^*
      if weight $w(o_t^*, t) > \mu\beta$ **then**
        Move o_t^* from O_t to P_t
    else
      Add $\{x\}$ to O_t as a new o-microcluster.
if $t \ \% \ T_p == 0$ **then** ▷ Pruning step
  for $p \in P_t$ **do**
    if $w(p, t) \leq \mu\beta$ **then**
      Remove p from P_t
  for $o \in O_t$ **do**
    if $w(o, t) \leq \xi(t, T_p, o) = \frac{2^{-\lambda(t - T_0(o) + T_p)} - 1}{2^{-\lambda T_p} - 1}$ **then**
      Remove o from O_t

[1] F. Cao, M. Estert, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in *Proceedings of the 2006 SIAM international conference on data mining*. SIAM, 2006, pp. 328–339

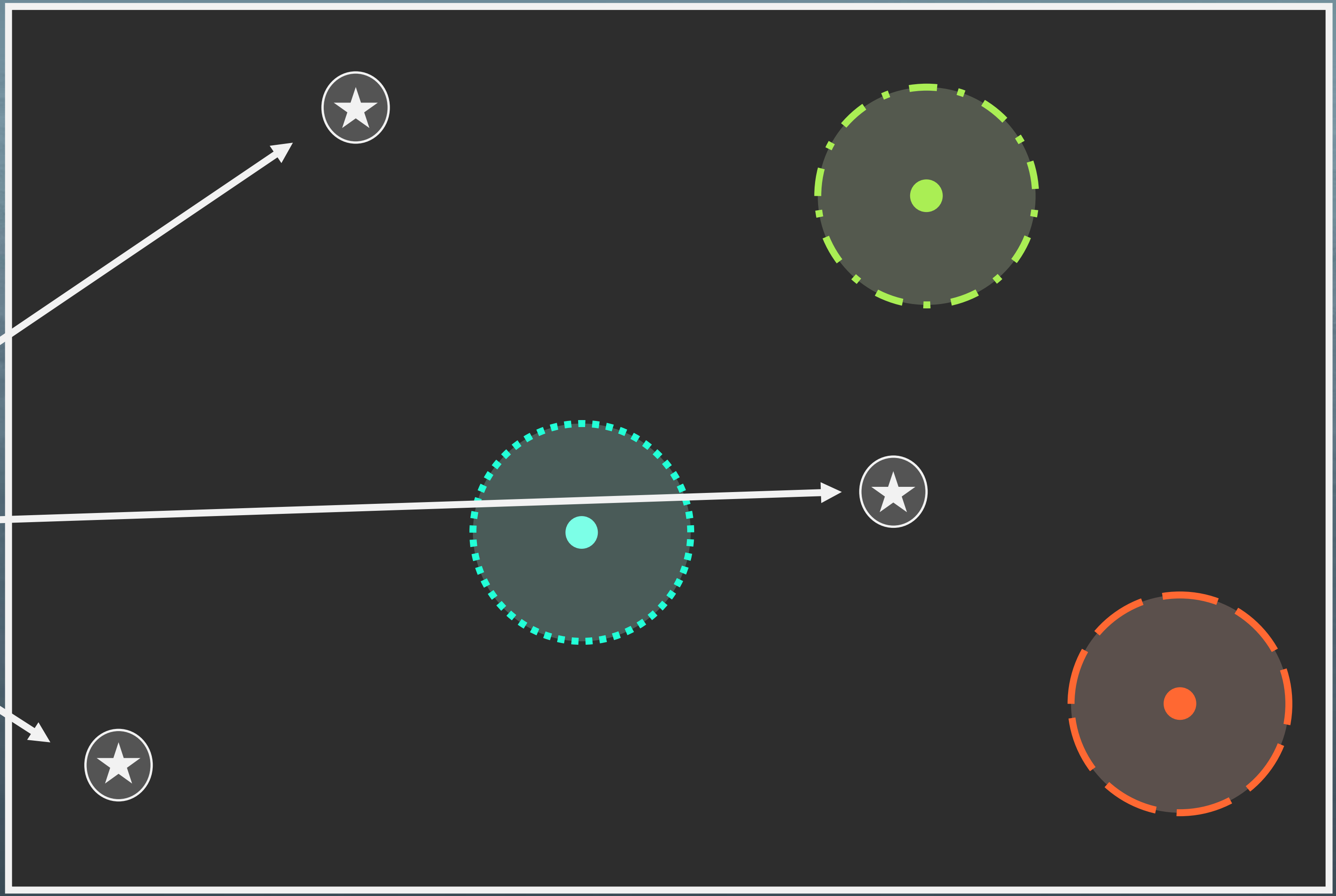
HOW DENSTREAM WORKS

p-microclusters
(aka “normal”
points)



HOW DENSTREAM WORKS

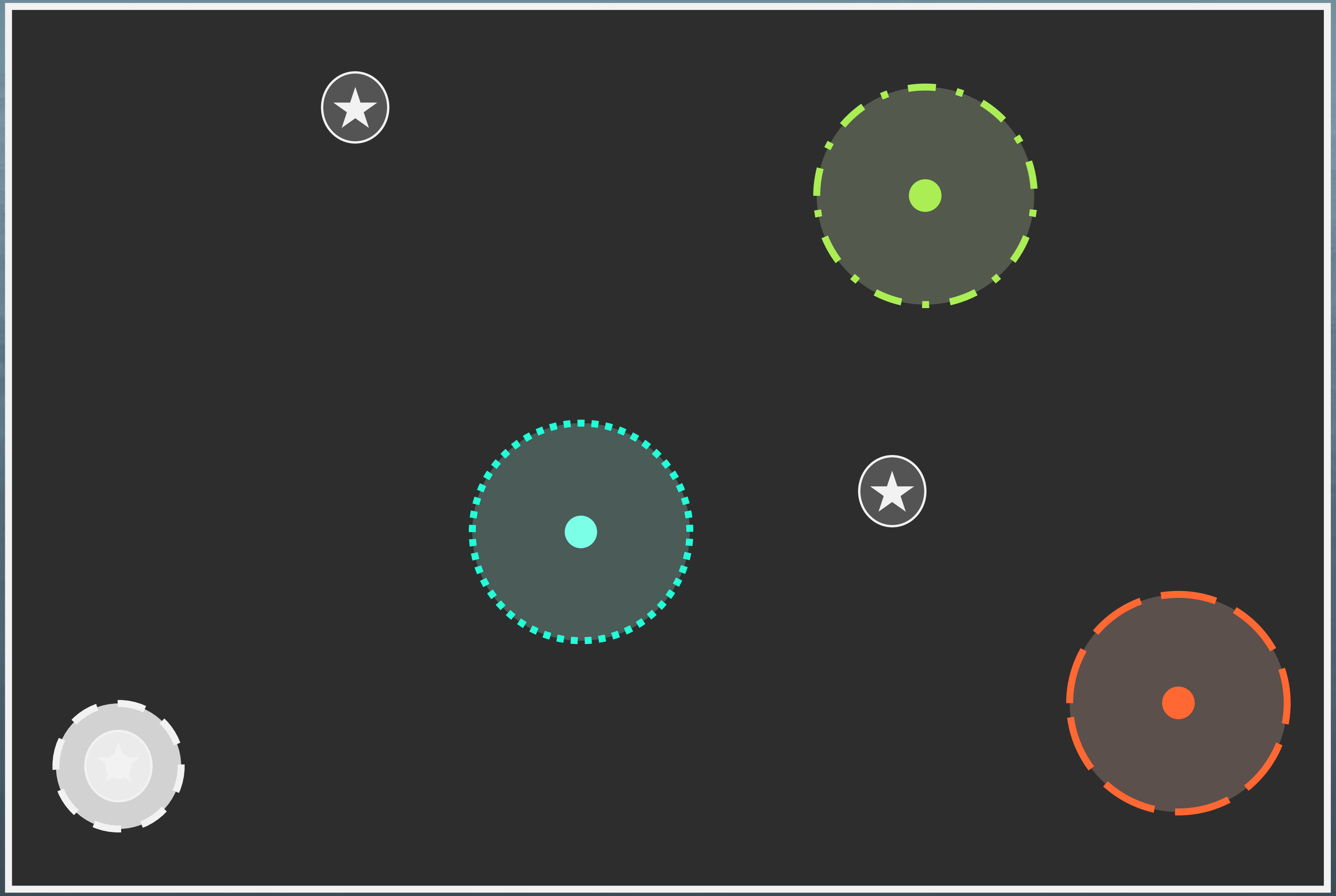
o-microclusters
(aka “outlier” points)



HOW DENSTREAM WORKS

If enough points are added to an o-microcluster, it's weight passes a threshold

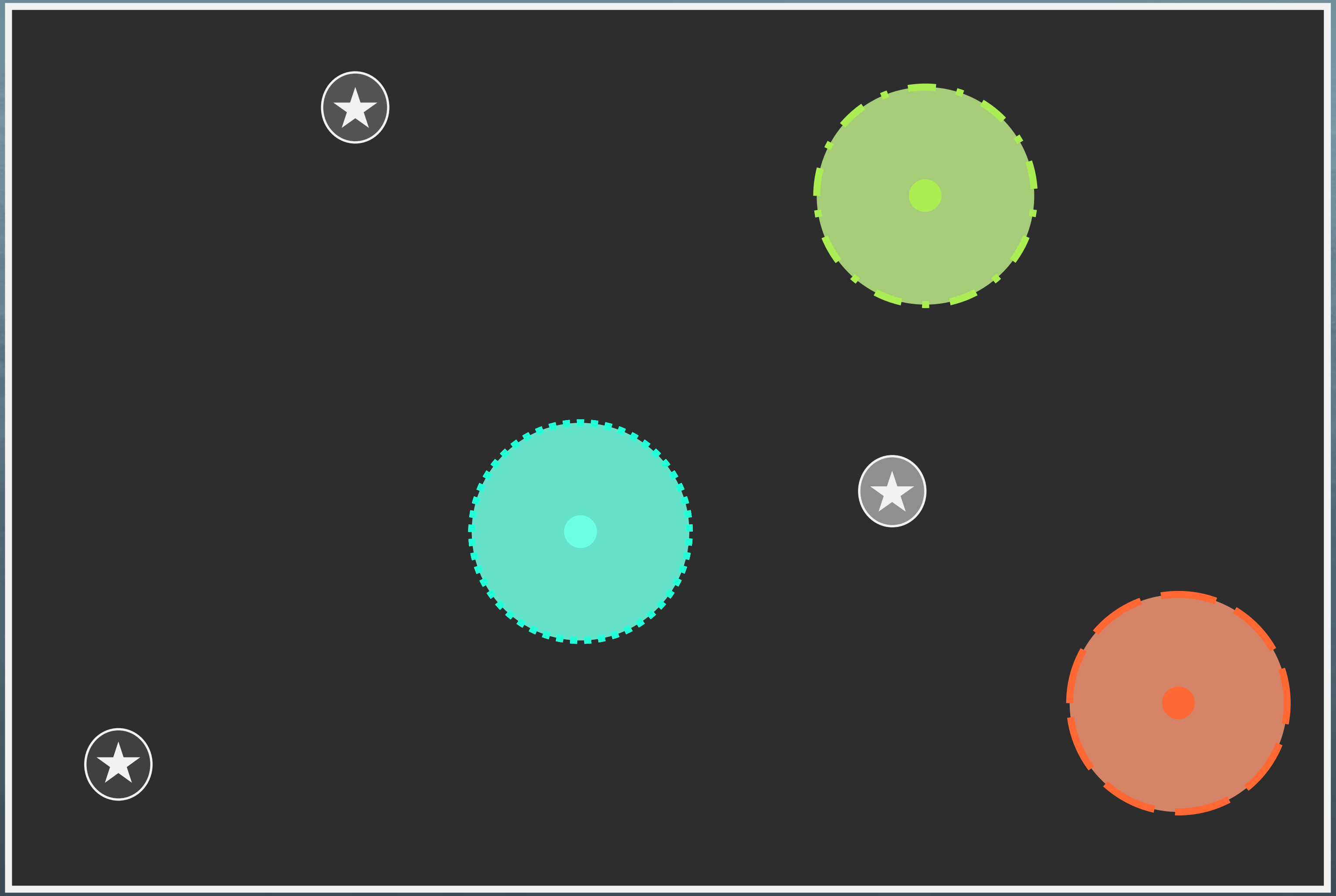
The o-microcluster then becomes a p-microcluster.



HOW DENSTREAM WORKS

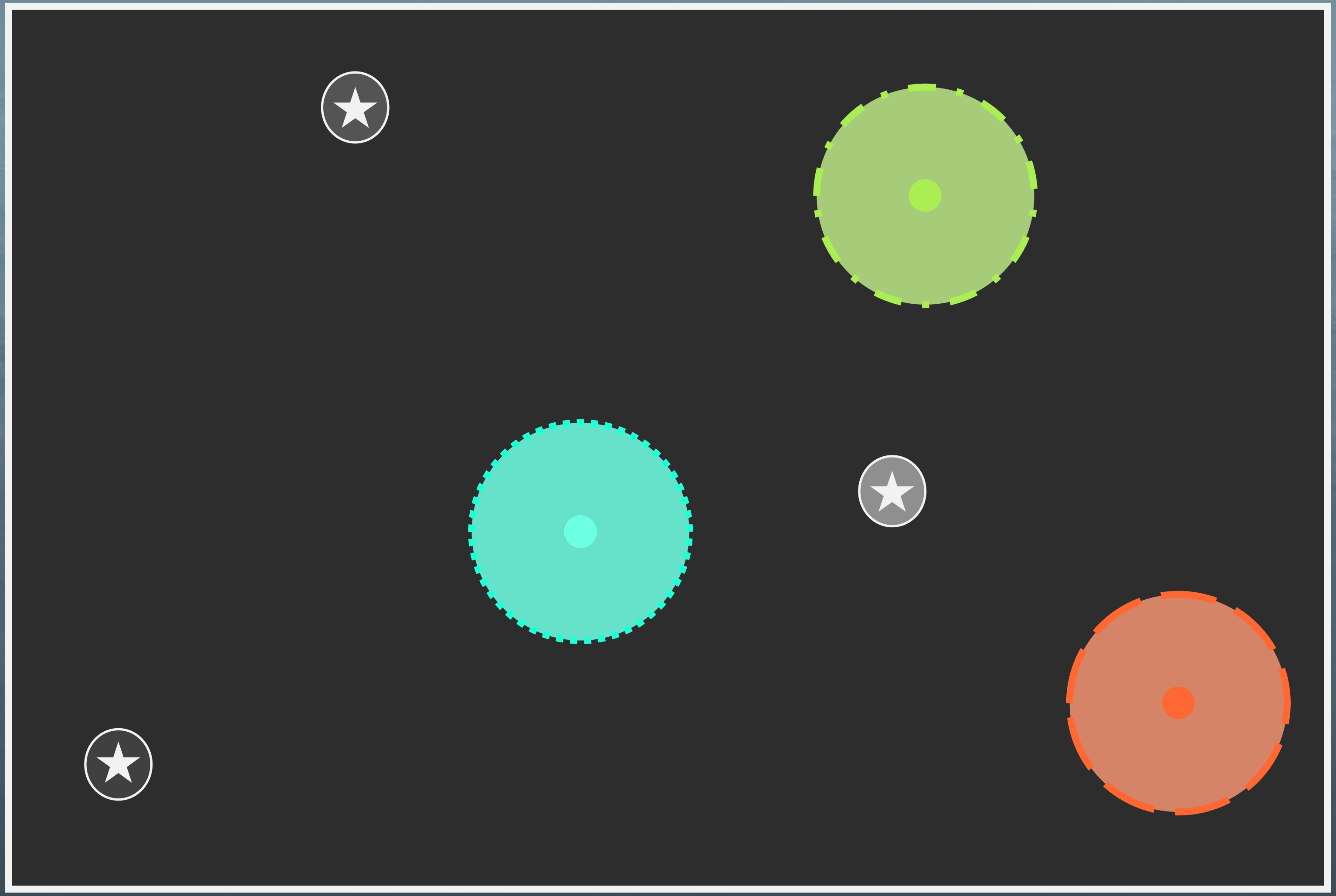
If new points aren't added to a p-microcluster, the weight decreases.

If the weight decreases enough, the p-microcluster is pruned



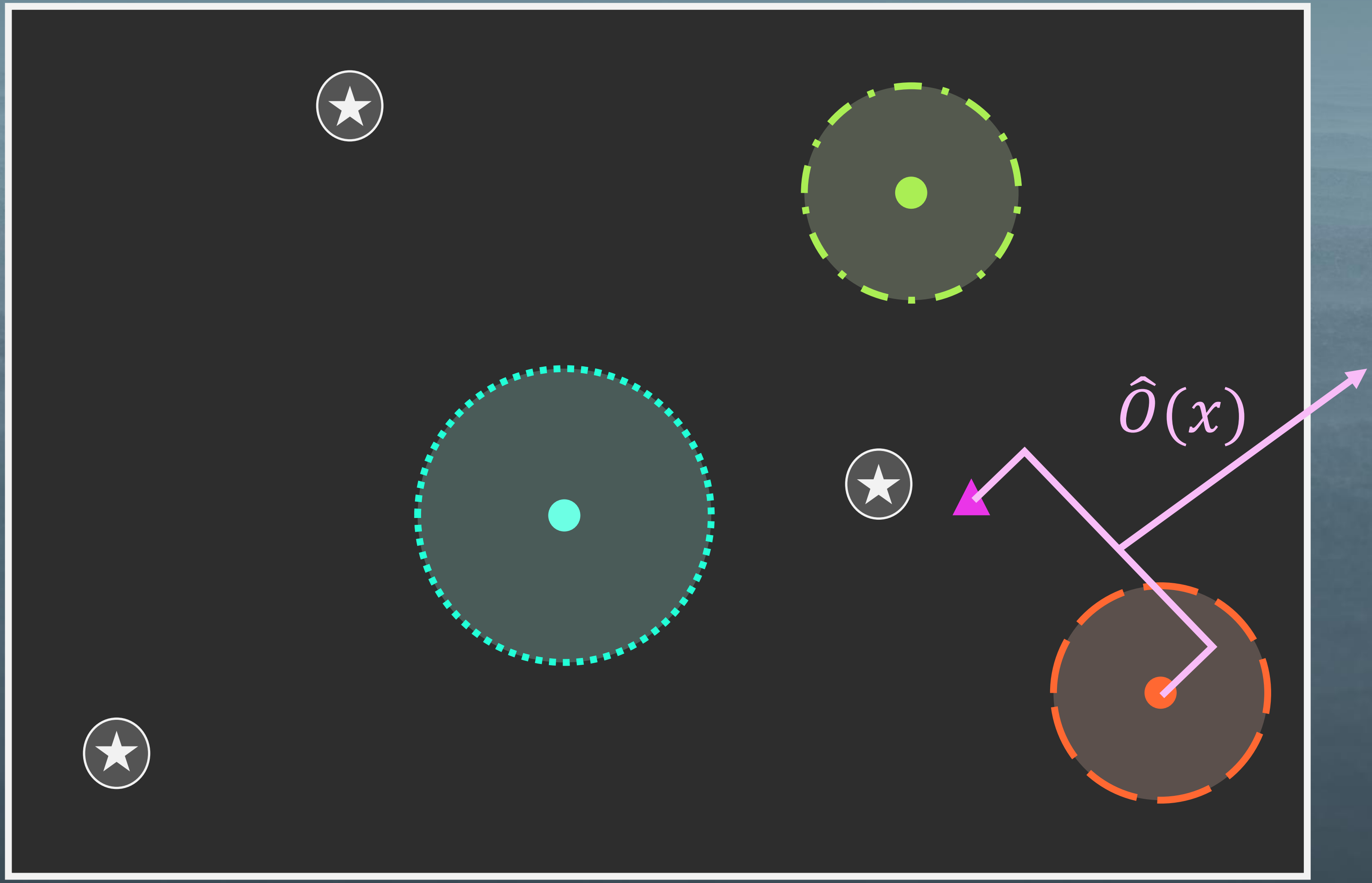
HOW DENSTREAM WORKS

Similarly, if no points have been added to an o-microcluster, the weight goes below a threshold ξ and it is pruned.

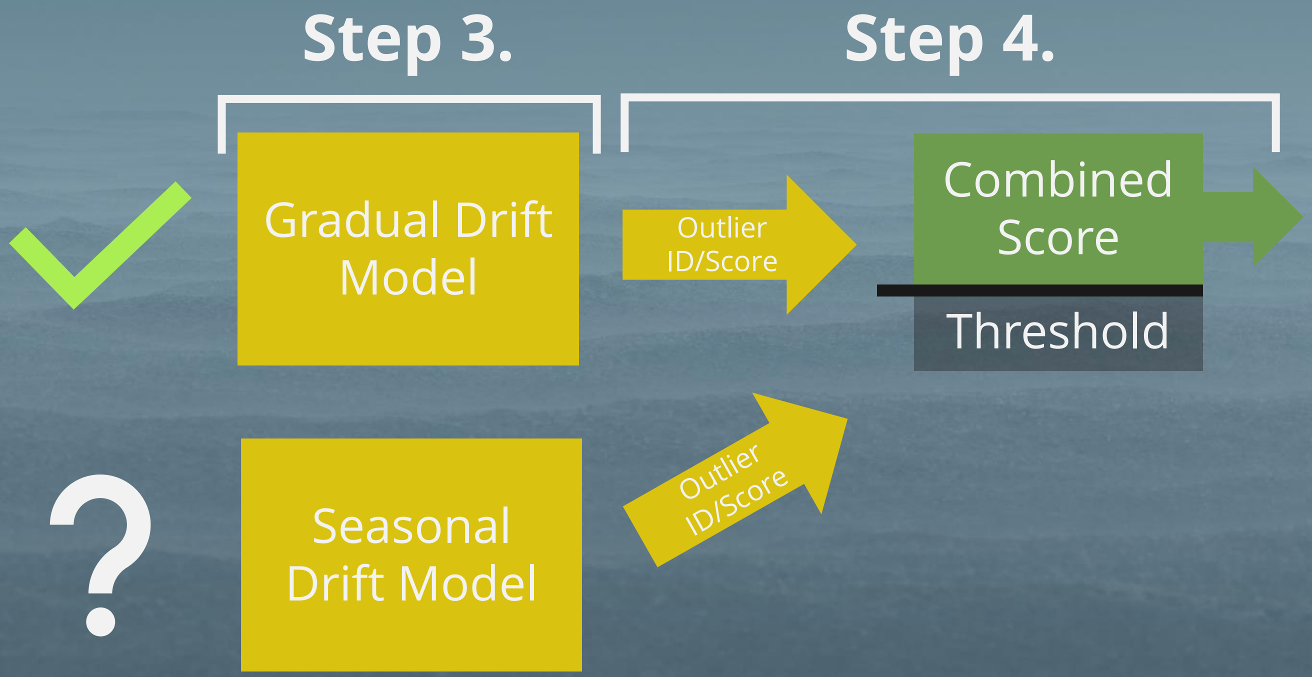


HOW DENSTREAM WORKS

If an incoming point is an outlier, output the distance to the nearest p-microcluster as the **outlier score.**

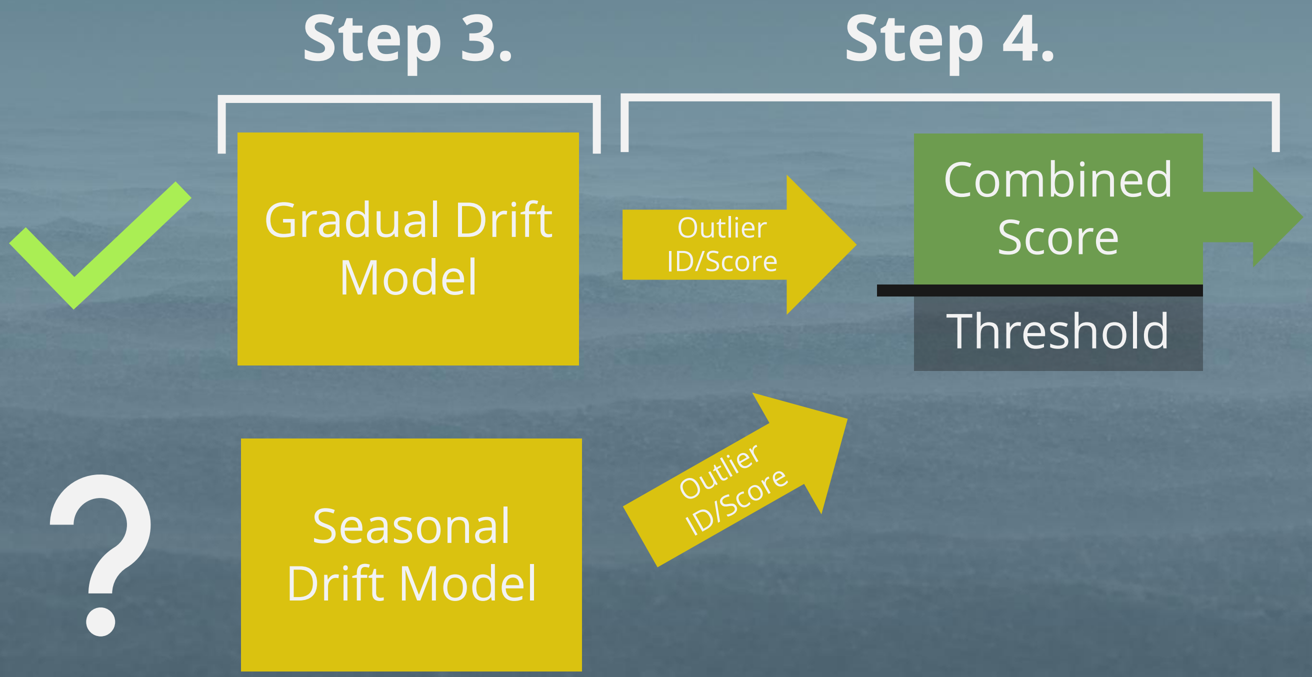


THE DB-DRIFT ALGORITHM



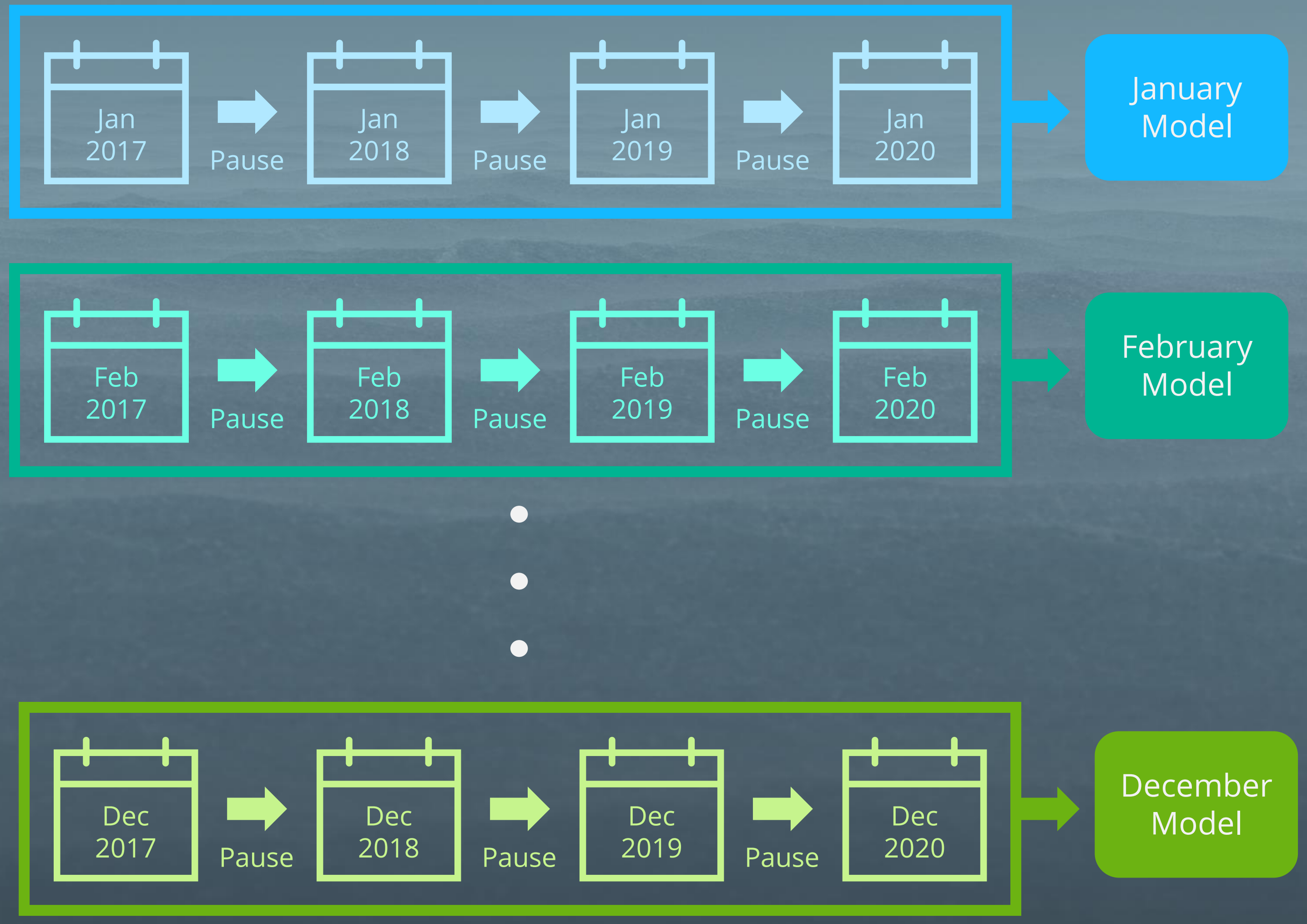
THE DB-DRIFT ALGORITHM

Handling seasonal drift for density-based clustering is an open field of research—**nearly no prior work.**



SEASONAL MODEL

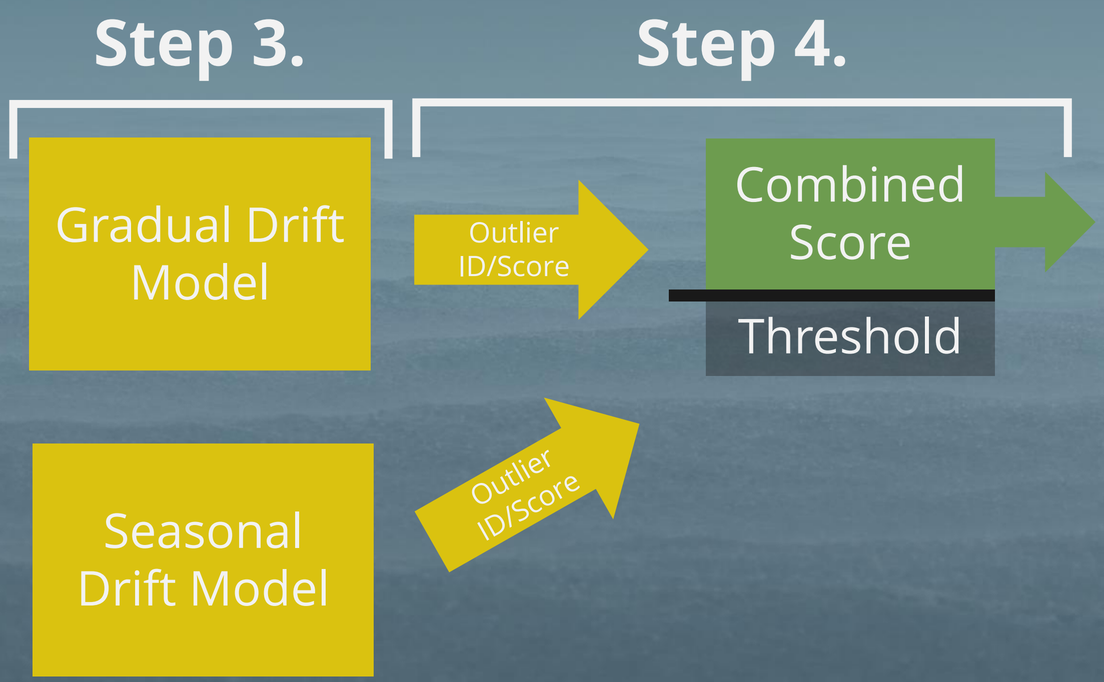
- General idea: assign a separate model for each season [1, 2].



[1] Hyde, R., Angelov, P., & MacKenzie, A. R. (2017). Fully online clustering of evolving data streams into arbitrarily shaped clusters. *Information Sciences*, *382*, 96-114.
[2] Katakis, I., Tsoumakas, G., & Vlahavas, I. (2008). An ensemble of classifiers for coping with recurring contexts in data streams. In *ECAI 2008* (pp. 763-764). IOS Press.

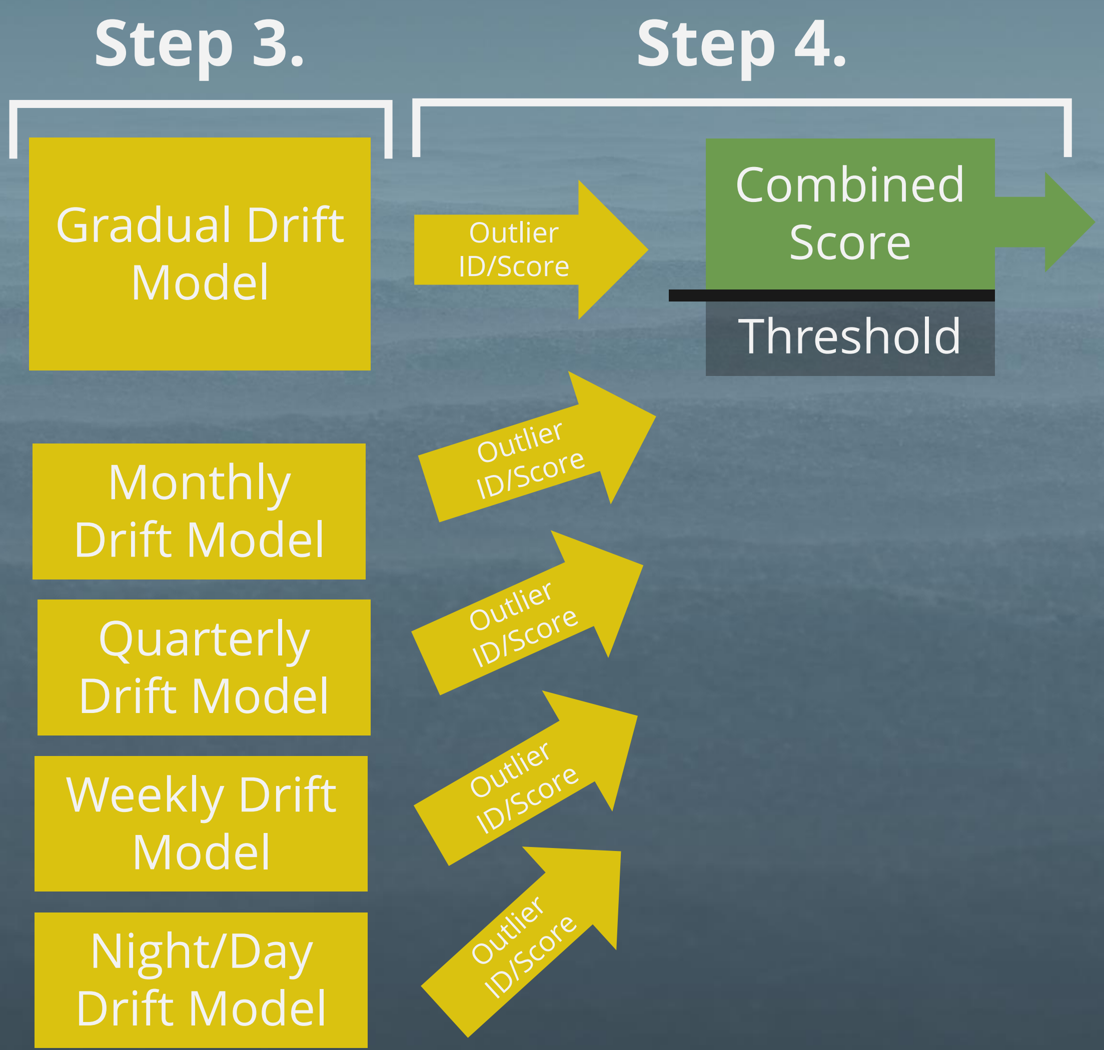
SEASONAL MODEL

- General idea: assign a separate model for each season [1, 2].
- For known periodic seasons, there can be seasonal anomaly detectors at multiple scales (months, quarters, weeks, etc).



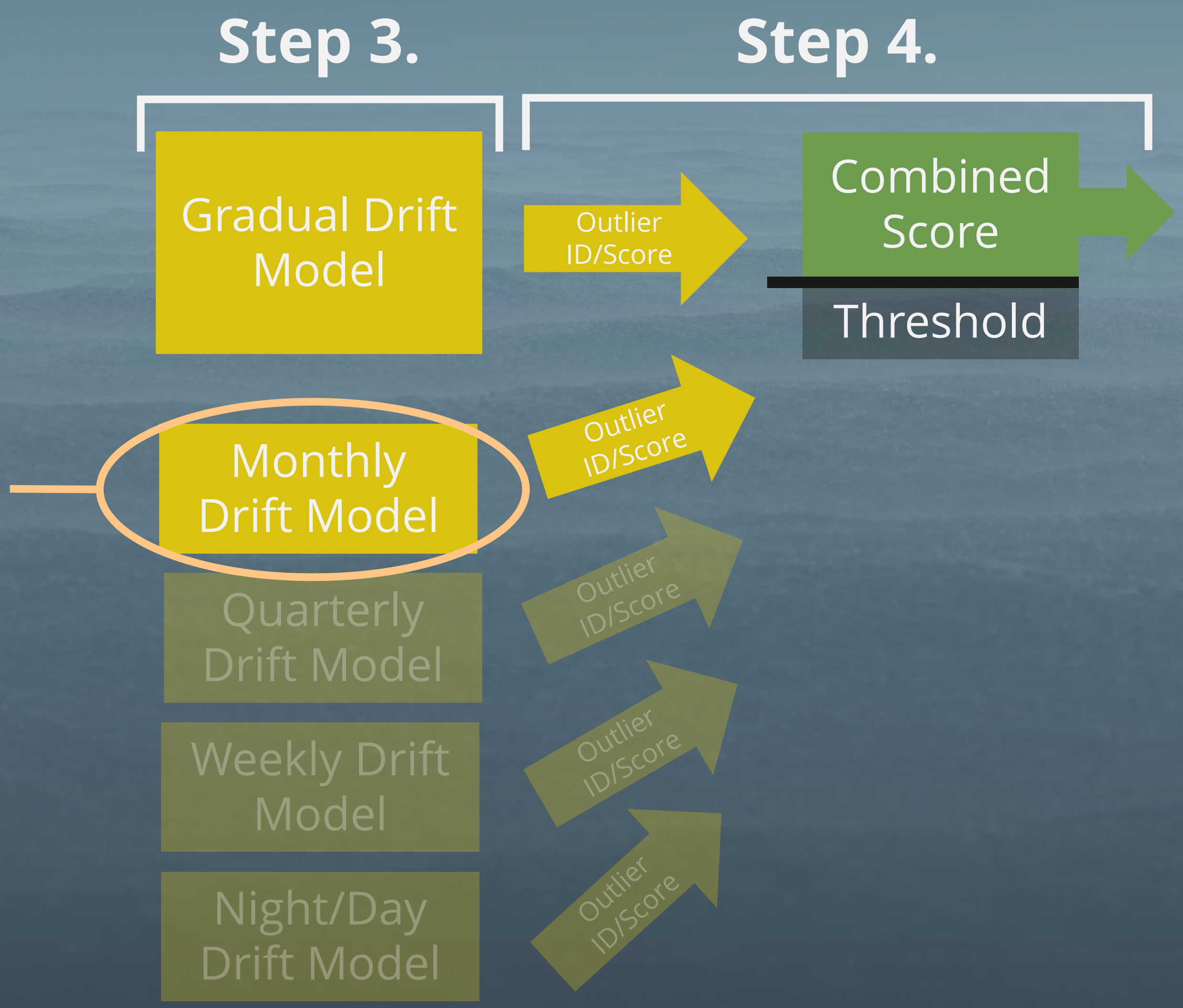
SEASONAL MODEL

- General idea: assign a separate model for each season [1, 2].
- For known periodic seasons, there can be seasonal anomaly detectors at multiple scales (months, quarters, weeks, etc).



SEASONAL MODEL

For our preliminary experiments, we have focused on the **monthly model.**



SEASONAL MODEL

For our preliminary experiments, we have focused on the **monthly model.**

IMPORTANT NOTE:
Seasonal drift is a subset of recurrent drift.

Expanding this algorithm to *find* seasons (in addition to defining known seasons) is a **very challenging** ongoing effort.

THE OUTLIER CONDITION

How do we define the combined outlier score?

How do we choose (and update) the appropriate threshold for a point to be considered an outlier?

THE OUTLIER CONDITION

For sample x at time $t = T(x)$:

$$\hat{O}(x) = w_g \min_{g_{t,i} \in G_t} \|x - c(g_{t,i})\| + w_s \min_{s_{t,i} \in S_t} \|x - c(s_{t,i})\| \geq \theta_{i,j}$$

G_t: The set of p-micro-clusters $g_{t,i}$ for the gradual model at time t

S_t: The set of p-micro-clusters $s_{t,i}$ for the seasonal model at time t

$c(\cdot)$: The center of a given microcluster (the fade- weighted sum of the points)

w_g, w_s: The outlier-score weights for the gradual and seasonal models.
We set $w_s = \frac{2}{3},\ w_g = \frac{1}{3}$ to emphasize the importance of seasonal anomalies.

$\theta_{i,j}$: For time $T(x)$ after some sample time period $[t_i, t_j]$, the minimum threshold for a sample to be considered an outlier based on the desired anomalous subset size.

THE OUTLIER CONDITION

For sample x at time $t = T(x)$:

$$\hat{O}(x) = w_g \min_{g_{t,i} \in G_t} \|x - c(g_{t,i})\| + w_s \min_{s_{t,i} \in S_t} \|x - c(s_{t,i})\| \geq \theta_{i,j}$$

G_t: The set of p-micro-clusters $g_{t,i}$ for the gradual model at time t

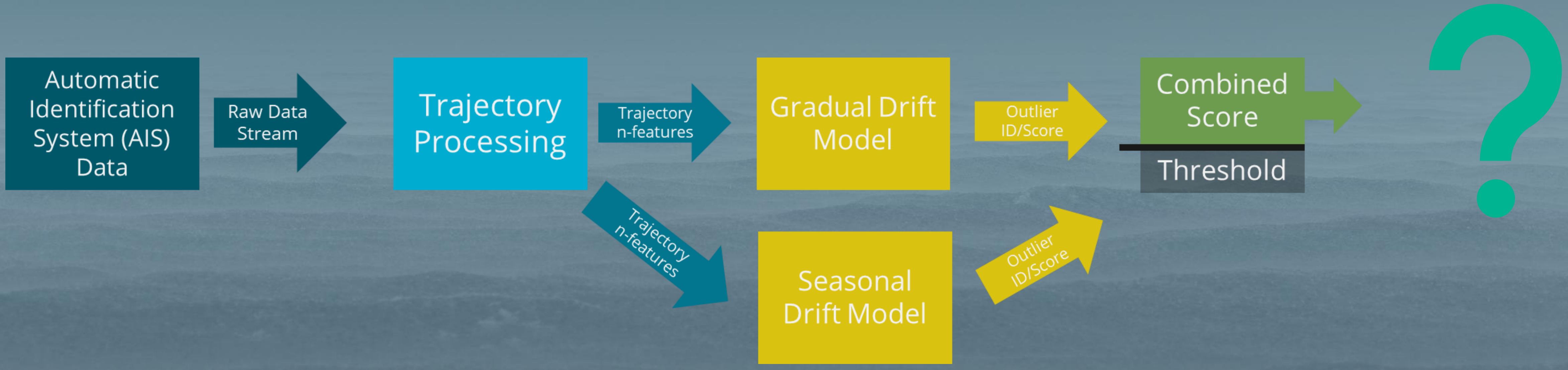
S_t: The set of p-micro-clusters $s_{t,i}$ for the seasonal model at time t

$c(\cdot)$: The center of a given microcluster (the fade- weighted sum of the points)

w_g, w_s: The outlier-score weights for the gradual and seasonal models.
We set $w_s = \frac{2}{3},\ w_g = \frac{1}{3}$ to emphasize the importance of seasonal anomalies.

$\theta_{i,j}$: For time $T(x)$ after some sample time period $[t_i, t_j]$, the minimum threshold for a sample to be considered an outlier based on the desired anomalous subset size.

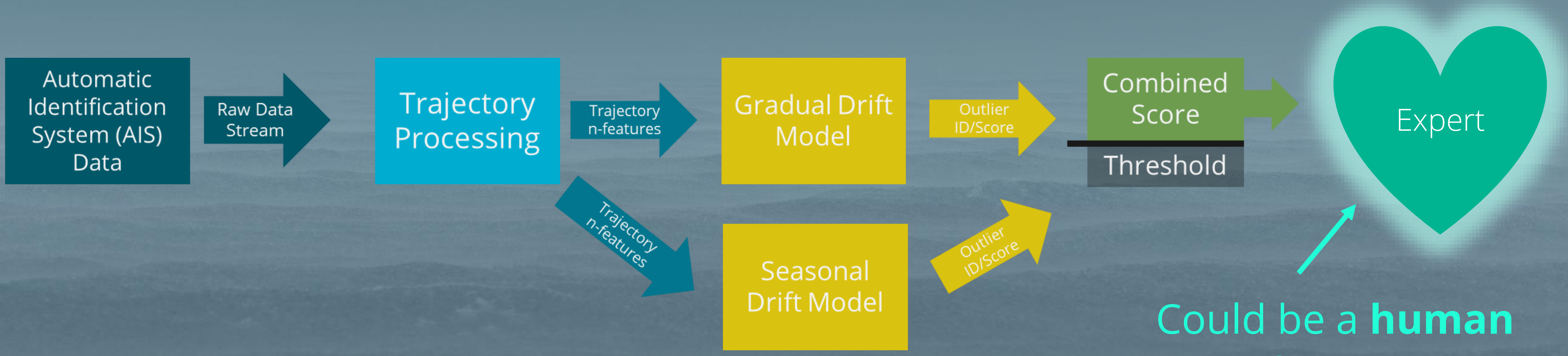
WHAT IS THE POINT OF UAD AT SEA?



The whole point of anomaly detection on maritime surveillance to:

1. Process data too big for experts to process
2. Find potential anomalies outside expert detection.

WHAT IS THE POINT OF UAD AT SEA?



Could be a **human expert**, or an expensive **expert algorithm**.

We want to output a **tractable subset** of points that contain the trajectories of interest—not a final anomaly decision.

THE OUTLIER CONDITION

For sample x at time $t = T(x)$:

$$\hat{O}(x) = w_g \min_{g_{t,i} \in G_t} \left\| x - c(g_{t,i}) \right\| + w_s \min_{s_{t,i} \in S_t} \left\| x - c(s_{t,i}) \right\| \geq \theta_{i,j}$$

G_t: The set of p-micro-clusters $g_{t,i}$ for the gradual model at time t

S_t: The set of p-micro-clusters $s_{t,i}$ for the seasonal model at time t

$c(\cdot)$: The center of a given microcluster (the fade- weighted sum of the points)

w_g, w_s: The outlier-score weights for the gradual and seasonal models.
We set $w_s = \frac{2}{3},\ w_g = \frac{1}{3}$ to emphasize the importance of seasonal anomalies.

$\theta_{i,j}$: For time $T(x)$ after some sample time period $[t_i, t_j]$, the minimum threshold for a sample to be considered an outlier based on the desired anomalous subset size.

THE OUTLIER CONDITION

$$\theta_{i,j} = \begin{cases} Q\left(\{\hat{O} > 0\}_{T(x)\in[t_i,t_j]}, 1 - q_{i,j}\right) & q_{i,j} < 1 \\ 0 & otherwise \end{cases}$$

$$q_{i,j} = \frac{n_{t_i,t_j} r}{\hat{n}_{t_i,t_j}}$$

r: The desired percentage of the dataset to return as an anomalous subset.

$[t_i, t_j]$: A sample time period used to determine θ for incoming points.

$Q(X, q)$: The q'th sample quantile for a set of scalar values X.

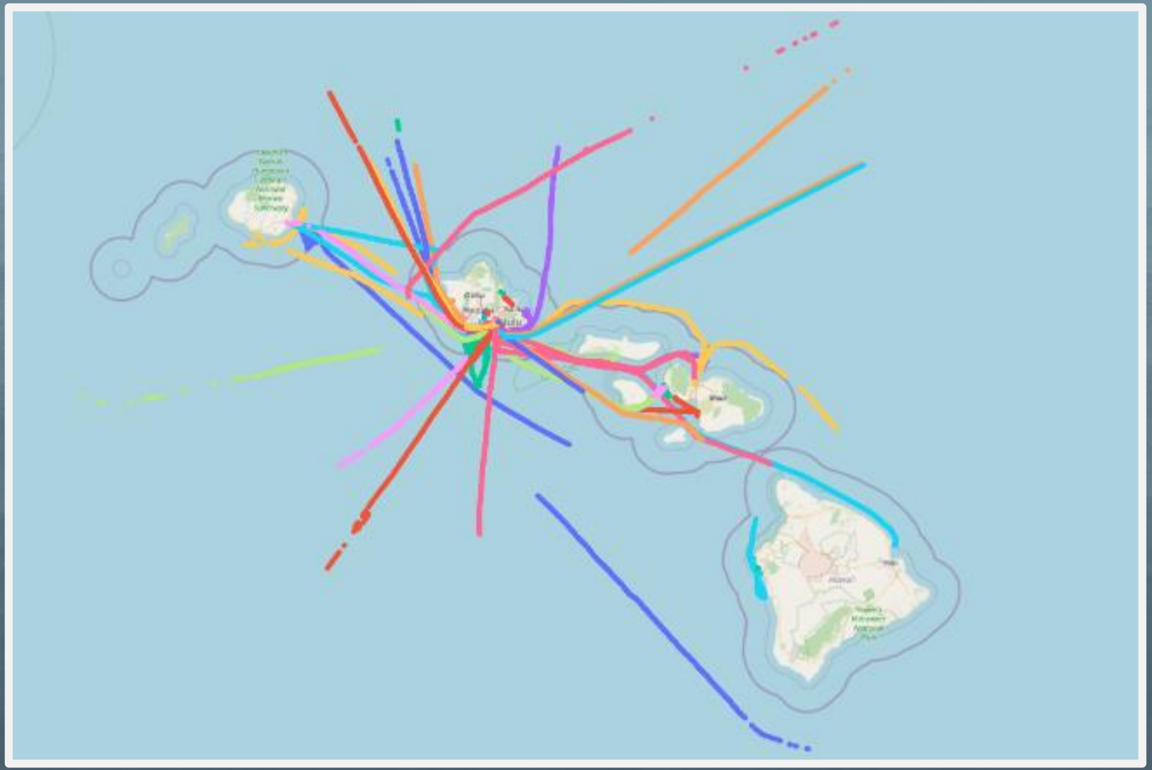
n_{t_i,t_j}: The number of samples that arrived during period $[t_i, t_j]$.

$\hat{n}_{t_i,t_j}$: The number of samples x with outlier score $\hat{O}(x) > 0$.

EXAMPLE: THE HAWAIICOAST_GT DATASET

THE HAWAIICOAST_GT DATASET

- Fully open and FAIR dataset available at https://zenodo.org/record/8253611 [1].
- Curated AIS data from MarineCadastre.gov [2]-[5].
- Includes 208 labelled tracks corresponding to 154 real-world incidents.



[1] Henriksen, Amelia. (2023). HawaiiCoast_GT: Curated AIS for Hawaii's coast correlated with ground truth incidents (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8253611
[2] Bureau of Ocean Energy Management (BOEM) and National Oceanic and Atmospheric Administration (NOAA). MarineCadastre.gov. AIS Data for 2017. Retrieved 7/25/2022 from marinecadastre.gov/data
[3] Bureau of Ocean Energy Management (BOEM) and National Oceanic and Atmospheric Administration (NOAA). MarineCadastre.gov. AIS Data for 2018. Retrieved 7/25/2022 from marinecadastre.gov/data
[4] Bureau of Ocean Energy Management (BOEM) and National Oceanic and Atmospheric Administration (NOAA). MarineCadastre.gov. AIS Data for 2019. Retrieved 7/26/2022 from marinecadastre.gov/data
[5] Bureau of Ocean Energy Management (BOEM) and National Oceanic and Atmospheric Administration (NOAA). MarineCadastre.gov. AIS Data for 2020. Retrieved 7/27/2022 from marinecadastre.gov/data

PRELIMINARY RESULTS

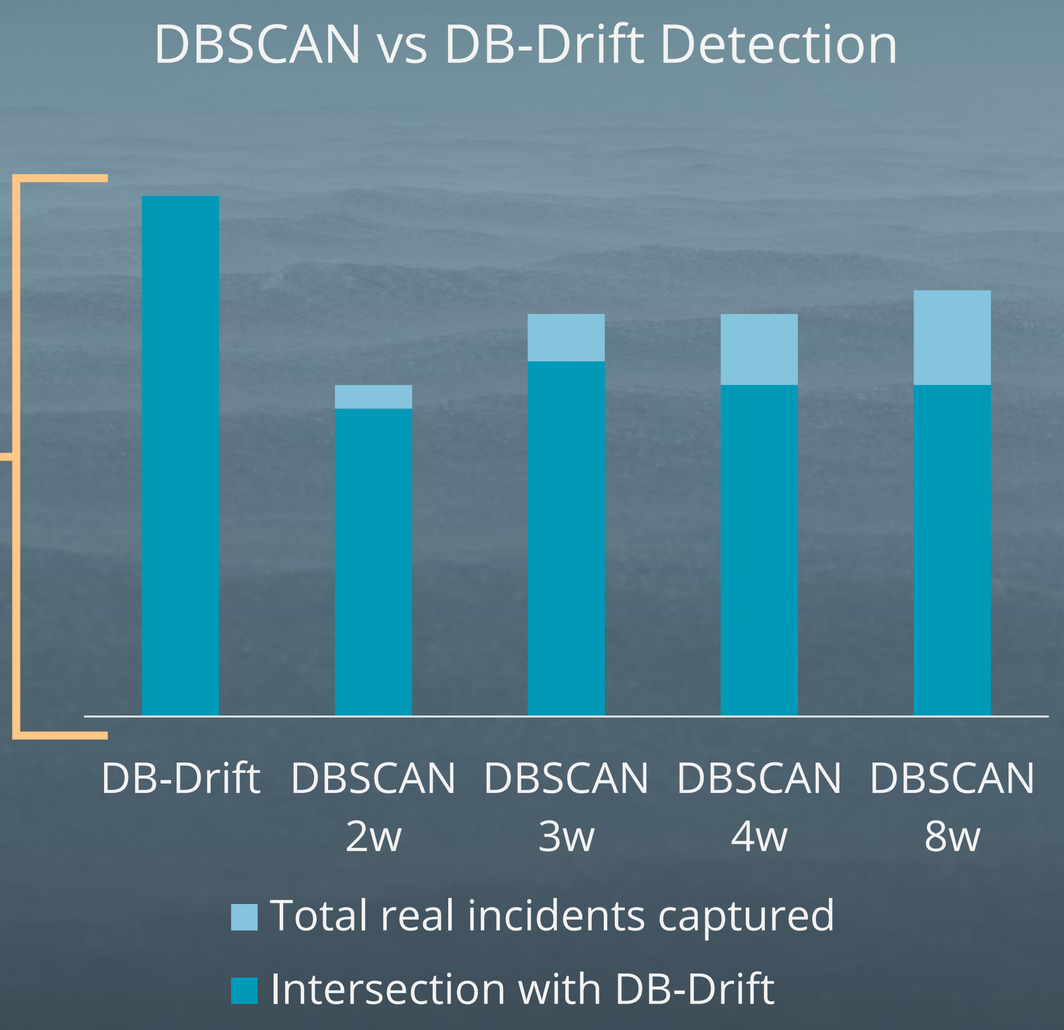
- Our goal was to **improve performance** for UAD pipelines with DBSCAN
- We compare it to **sliding window DBSCAN** over a range of window sizes (reporting best results over a range of hyperparameters).

Method	Real incidents captured	Intersection with DB-Drift
DB-Drift	22	
DBSCAN 2w	14	13
DBSCAN 3w	17	15
DBSCAN 4w	17	14
DBSCAN 8w	18	14

Test: anomalous fishing vessel trajectories from HawaiiCoast_GT.
Total real world incidents: 74 (varied anomaly types)

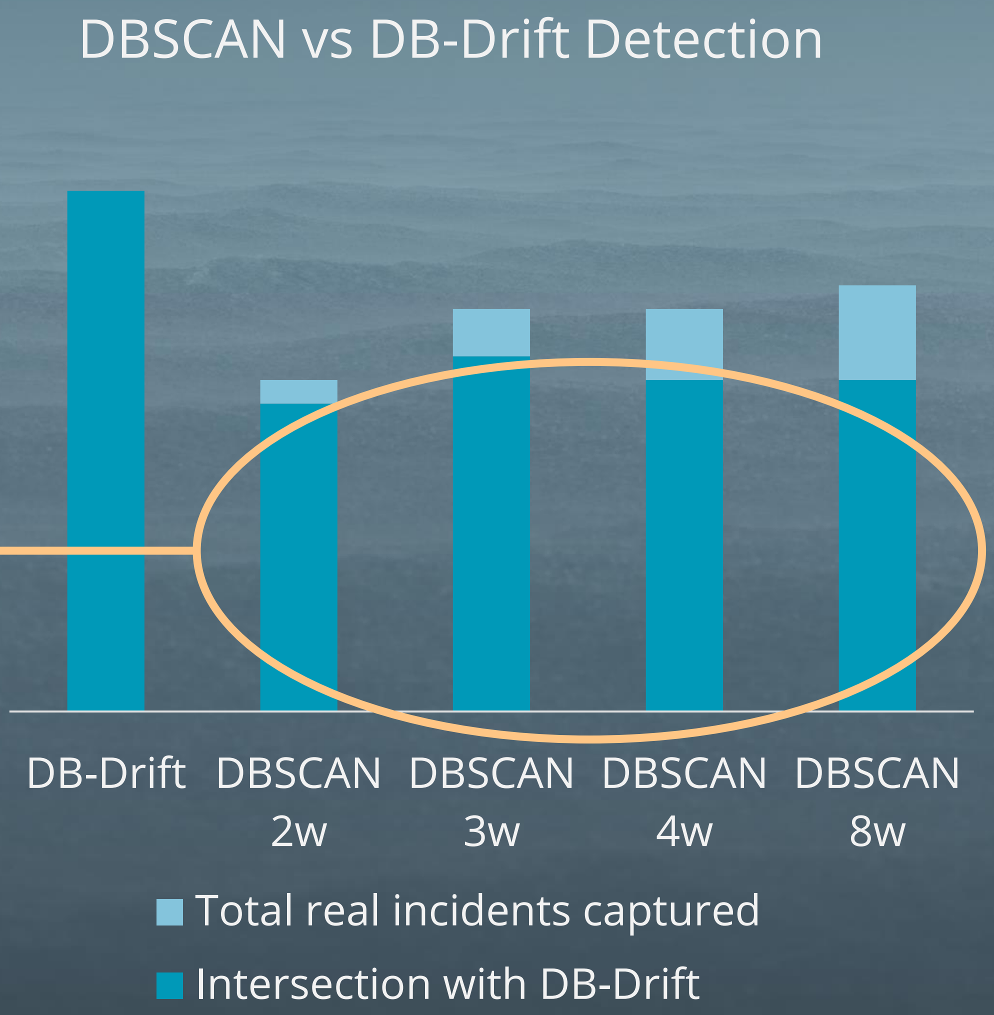
PRELIMINARY RESULTS

DB-Drift captures more real incidents than DBSCAN



PRELIMINARY RESULTS

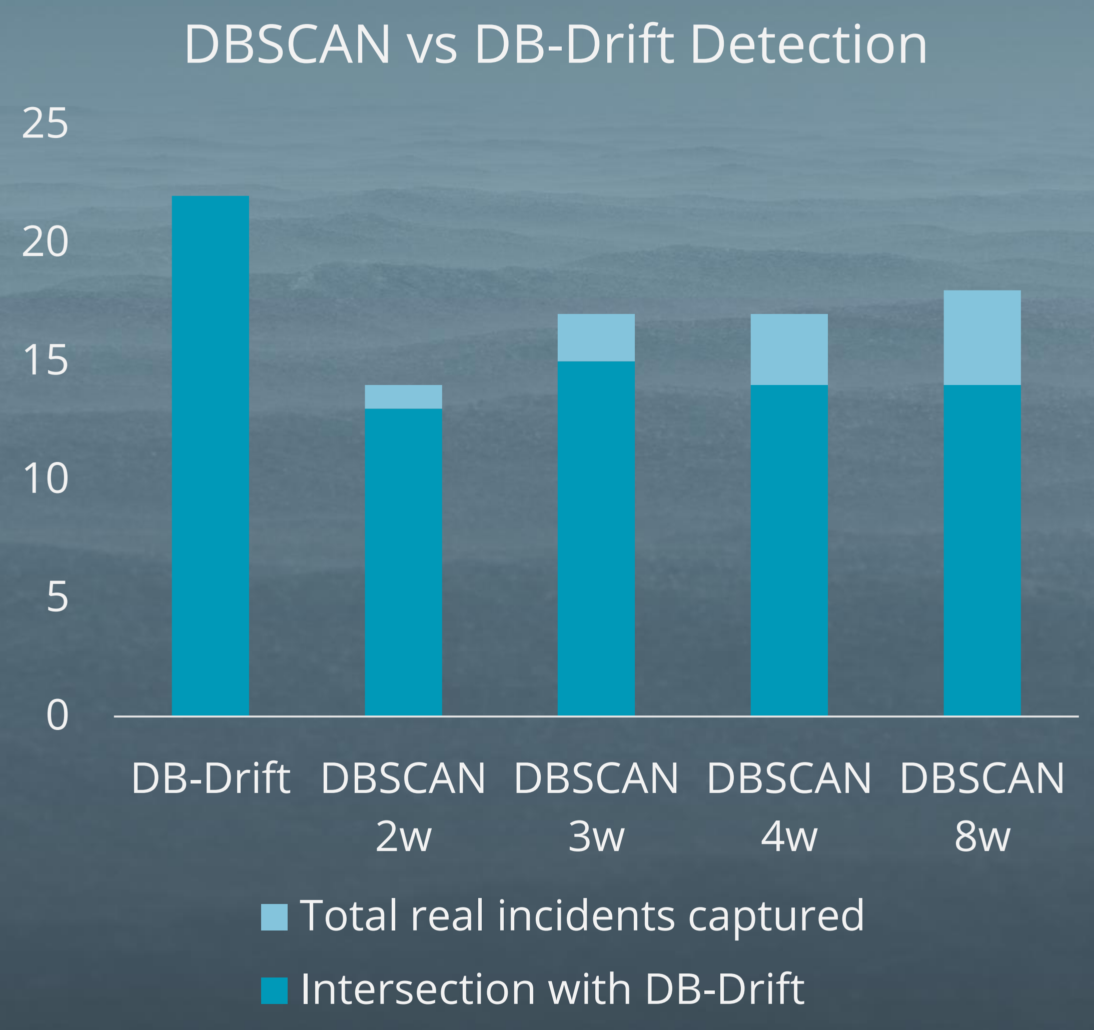
DB-Drift captures most of the incidents captured by DBSCAN



PRELIMINARY RESULTS

More advantages:

- DB-Drift requires a burn in of only a few days, DBSCAN requires at least 1 window period.
- Significantly lower memory requirements



NEXT STEPS:

Experiments:

- Trajectory feature optimization to improve overall performance.
- Additional tests for each vessel class and specific anomaly types.
- Curating further datasets using our ground truth technique for additional benchmarking.

Algorithm:

- Season discovery vs known seasons.
- Adding abrupt drift detection to reweight historical information.

BIG THANKS

Feedback and Mentorship

Ben Newton

Andy Wilson

David Stracuzzi

Making so much data publicly available

BOEM

Bureau of Ocean Energy Management