**General Notes about the Special Course**

**Data:**

* Try first with the JSON files made by Sune (one ship – one JSON file), then take a look to the CSVs if we need to group by region or time for example.
* Referring to spacial circunstances, if we try to represent the paths on a grid we realize that there’re **a lot of zero values (corresponding to ground areas)**. This ofc could affect the performance of the model (specially how is computed the los function), that has to learn to discriminate this information. Try to quantify this and explain it.
* Due to the previous problem, **using vectors to represent paths** for each ship could be a interesting way to explore.
* The different **length of the paths**, the different **behaviour of each type of vessel**, the **time window** or a **different zone of the map** are factors that can make the model performs completely right or wrong. Try to check if a model can adapt and learn from these special characteristics or if it would be necessary to train different models.
* How about **sampling**?? Depending on the ship type this could be something necessary to adjust.
* There are also some big concerns about using **continuous or discrete values**.
* One interesting possibility would be to use **relative values** instead of absolute ones.

**Model:**

* **Why VAE** could be the best option for this problem? Is the **Lossy VAE** a possibility for this case?
* Should we consider an **hybrid model**? First try to optimize the current proposals and then let’s go back to this.

**Testing:**

* An interesting approach to optimize the solution for this problem should be to first define exactly how to represent the data, and then use the different model setups as a **black box** to find the optimal model/parameters.

**Data Format:**

The format uses the "raw" fixed precision values from AIS messages: Latitude, Longitude is en 1/600000 degree, speed in 0.1 Knots, angle in 0.1 degree .

Timestamps are seconds since 2019-01-01, 00:00:00 UTC

You will find zipped folders from April 2019 to March 2020 in the fileexch.terma.com shared folder Ais Data (From Terma).

The unzipped folders will contain a lot of ship-specific JSON data representing all the position updates from that MMSI in that month.

Each file name starts with the ship type - followed by the MMSI - and the month it covers.

Some ships with very few updates are collected in "TheRest" files for their type because they are mostly irrelevant and lots of small files are annoying to deal with. That is why each JSON file is a list (most with one member).

The path elements are :

\* Time in seconds since 2019-01-01 00:00 UTC

\* Latitude is degree \* 600000

\* Longitude is degree \* 600000

\* Speed-over-ground (SOG) is Knots \*10

\* Course-over-ground (COG) is in degrees \* 10

\* Heading is in degrees

Other values:

\* MMSI = Maritime Mobile Service Identity

\* IMO = International Maritime Organization

**References:**

* Mining maritime vessel traffic: Promises, challenges, techniques (it doesn’t use NNs, it has concrete proposals to represent the data) <http://52.57.172.180/bitstream/handle/20.500.12489/862/CMRE-PR-2019-122.pdf>
* **A Maritime Anomaly Detector using Probabilistic Neural Network Representation of AIS Tracks and A Contrario Detection** <https://arxiv.org/pdf/1912.00682.pdf>

**Questions for Kristoffer**

* Could we see the data you used?

He will send us info about the data format and pseudocode about the filtering for the tracks

* What’s exactly the difference between the normal sets and the 24h sets?

How long can be a sequence?? Depends on the boat, that’s a problema and we have to try to find a compromiso for all typpes of vessels. EXPERIMENT WITH THIS

* Are inputs and targets the same before feeding the model?

Inputs are normalized because the NN struggles with the raw data, so it becomes easier for it

* Is COG being used?? Just only not plotted??

Is used

* What can you tell us about the hyperparameters?

Not a lot, it’s more about the shape of the data. Notice that the RNN is being fed with 0s instead of the prev hidden state. Nothing specific about the latent state

* What do the plots represent at the end of the notebook?

The plots that seem to be empty is just because the vessels are stationary, but the one in the middle gives an idea

* Hybrid architecture?

It could be interesting but there’s prolly no interesting information that the CNN can give us. Difficult to say, in theory is posible, but… Does the trayectory contains less info than an image??

* Conclusions? Insights? Recommendations?
* How to proceed (models proposed?)

Prepare the code to have a pipeline . The scope of the project will be better defined: the 4-hot encoding is something to discuss too, weigthed encoding (linear combination for edges), spacial resolution for different zones of the map, etc.

About meetings: try to include Kristoffer and Sune when having meetings with Line, try to setup an Schedule for the meetings.

**Notes for next meeting**

* How should we filter the navigation status in the pre-processing? Which ones can be considered as relevant to detect anomalies and which ones are just noise?