NHH

Shipping Economics and Analytics-Shipping data, statistical analysis and M/L. L6

Gabriel Fuentes

gabriel.fuentes@nhh.no

This lecture

- Fundamentals of shipping data
- Data handling, outliers detection, aggregation and segregation of data
- Intro to machine learning
- Shipping research/ apps implementation
- M/L vs statistics
- Workshop in Class 2 parts

NHH

Learning outcomes

This lecture will help with the following learning outcomes:

Knowledge

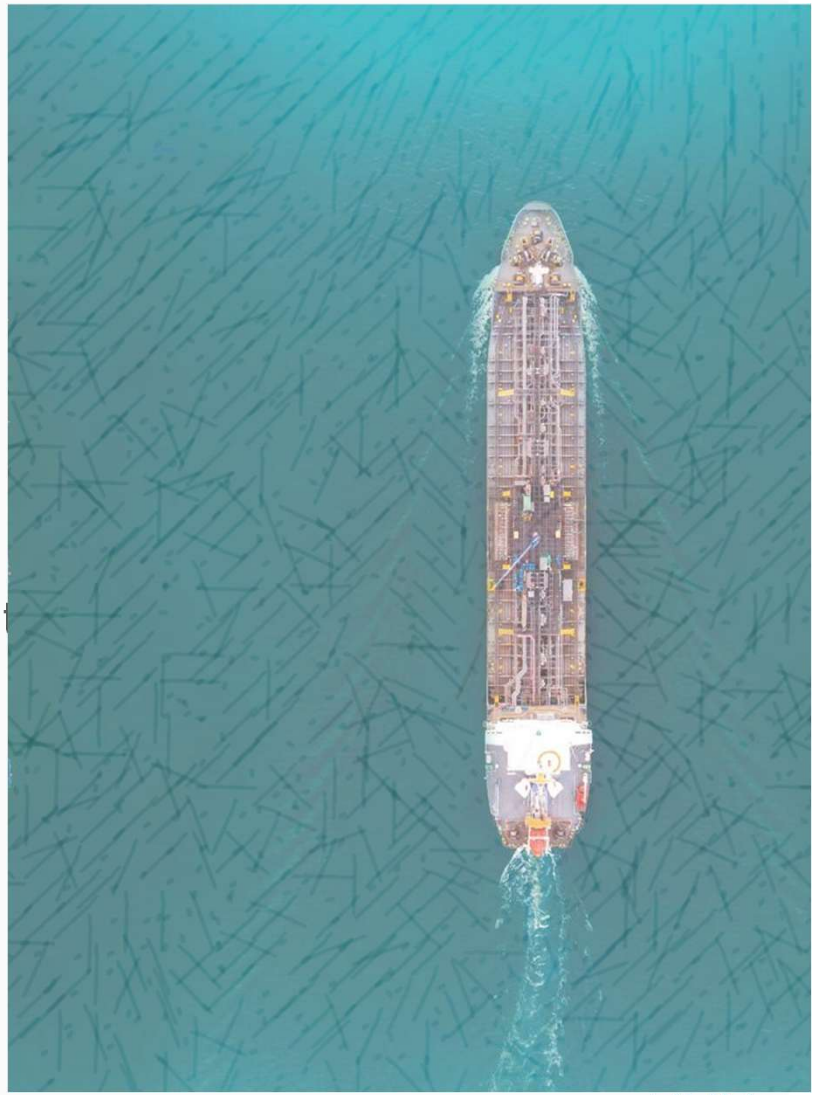
- is familiar with recent development in data-driven analysis applied to the freight markets and ship operation
- is conversant on technical aspects of shipping digital platforms

Skills

- finds, synthesizes, and presents information on the international shipping
- can communicate with industry practitioners using correct terminology

Competency

- exchanges opinions and experiences with others with a background in the field



NHH

Data sources

Ideal

Vessels

AIS data

Ports

Port throughput

Customs/
Authorities

Customs declaration

Charter
Parties

Brokers disclosure

NHH

Data sources

The way it is

Vessels

AIS data

Ports

Port throughput

Customs/
Authorities

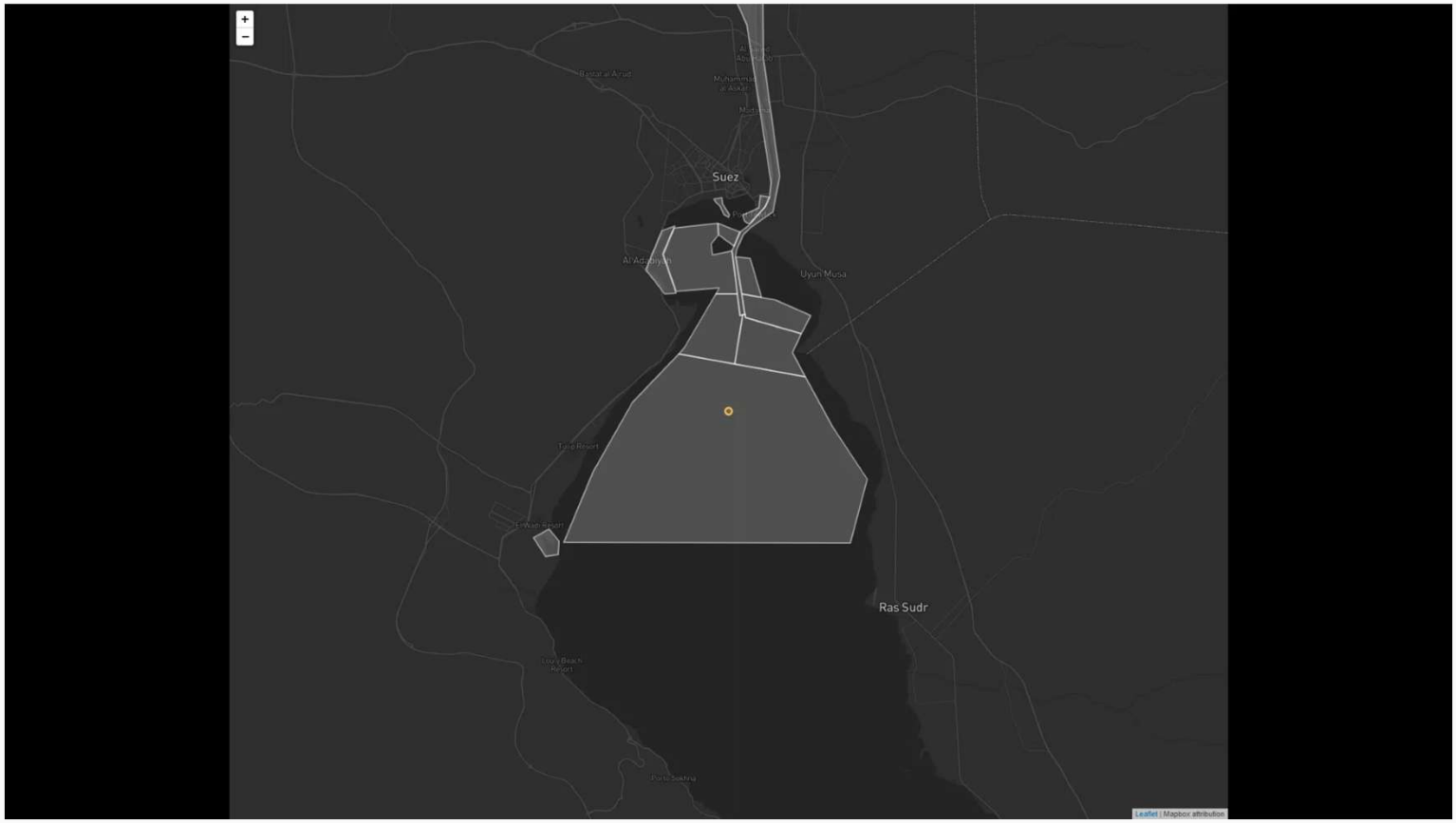
Customs declaration

Charter
Parties

Brokers disclosure

NHH

The strength of individual vessel behavior



AIS variables

- Mandatory AIS-Class A for all vessels of:
  - 300GT or upwards engaged in international waters
  - Cargo ships of 500GT or upwards, not engaged in international voyages
  - Passenger ships, irrespective of size
- Regulated by the International Convention for the Safety of Life at Sea (SOLAS) Chapter V Regulation 19

NHH

What is Automatic Identification System (AIS) data?

- **Main use:**
  - Collision avoidance
- **Collateral use:**
  - Security surveillance
  - Emissions calculation
  - Vessels emissions
  - Global trade flows
  - Fisheries
  - Underwater noise
  - Optimal routing

NHH

AIS variables

Is there any other variable you consider important?

Dynamic (Transmitted every 2 to 10 seconds when underway and 6 minutes at anchor)

- Ship position (Lon, Lat) - Float
- Speed Over Ground (SOG) - Float
- Course Over Ground (COG) - Float
- Heading - Int
- Rate of Turn - Int
- *AIS Navigational Status – VarChar

Static & Voyage (Transmitted every 6 mins)

- IMO Number – KeyID Int (7)
- MMSI – Int (9)
- Call Sign – Char (4)
- Ship Name – VarChar
- Type - VarChar
- Dimensions
  - Length - Int
  - Beam – Int
- *Draught - Float
- *Destination – VarChar
- *ETA - datetime

NHH

* Manually input by vessel crew

AIS variables

Is there any other variable you consider important?

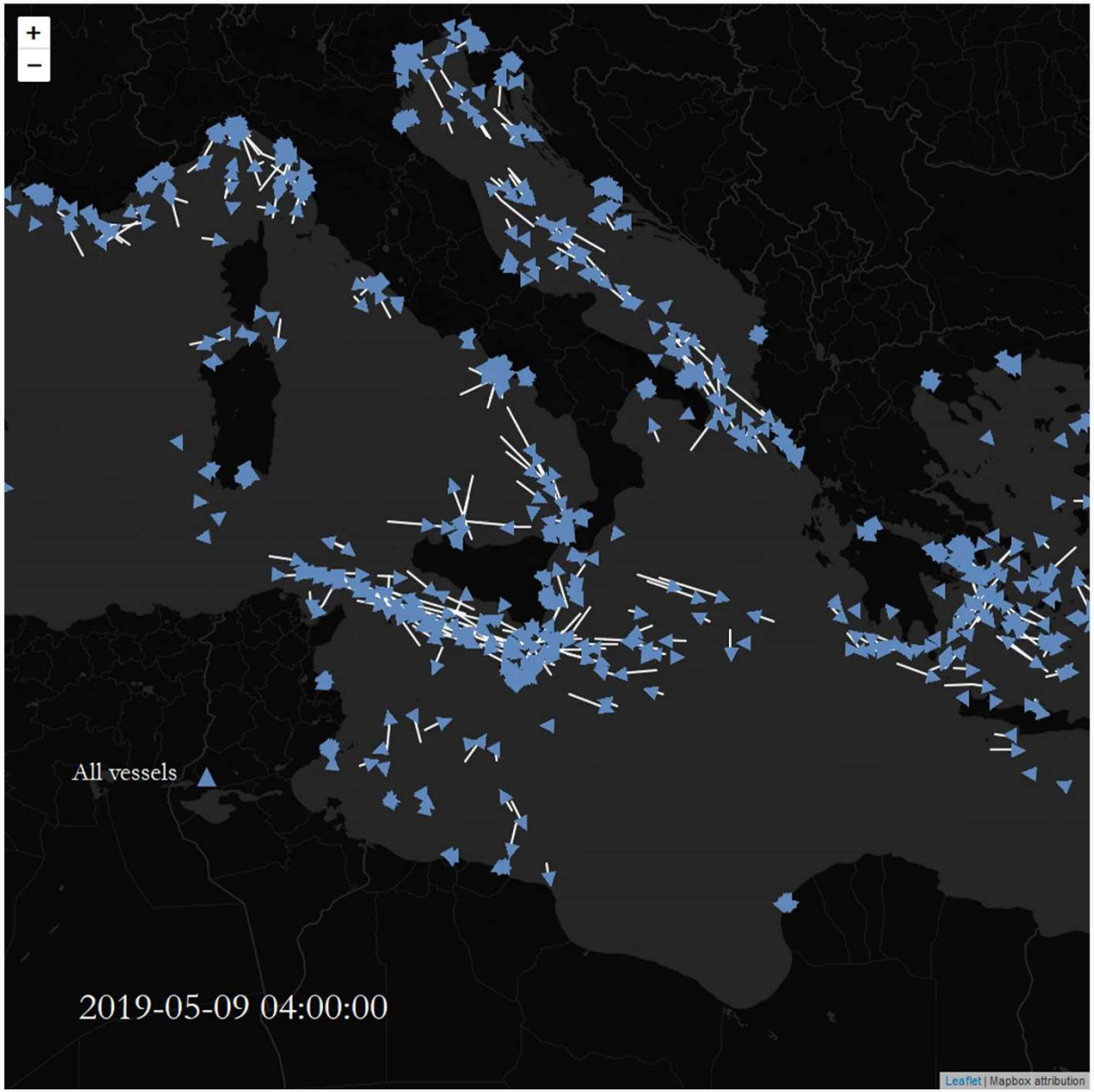
Time stamp of record in UTC – datetime

Stamped by the receivers

NHH

* Manually input by vessel crew

AIS variables



AIS Strength & Weaknesses

Strengths

- Timely statistics can be generated
- Aggregated data can be filtered
- High frequency data

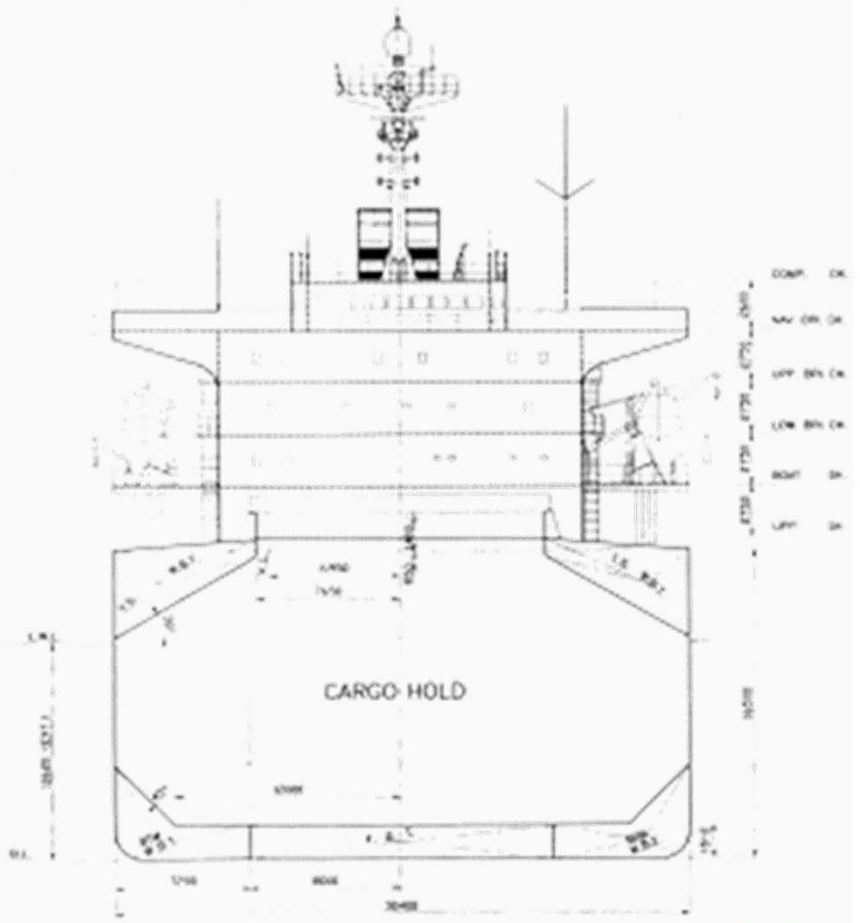
Weaknesses

- Not consistent in some areas and from some vessel sizes (missing data)
- Vessels can switch off the device e.g. in piracy areas
- Manually input data is based on crew proactivity in timely updating

NHH

Vessel specifications

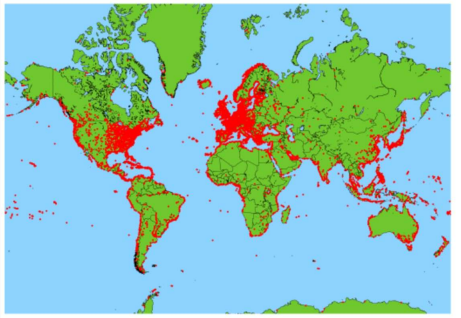
- As cargo information is not visible, a good way to “follow” trade is by observing vessel segments based on vessel characteristics.
- AIS information alone is very difficult to segregate.
- Clarkson's WFR provides of individual vessels information. Information could be matched by IMO Number or Maritime Mobile Identification Service number (MMSI)

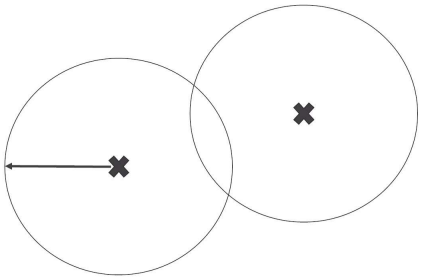


NHH

Port visit

- The easiest way of retrieving information from AIS movements is by identifying the times of visit to a load port and to a discharging port
  - Some interpret this to be a good representation of port service
- A basic algorithm would identify the latitude, longitude pairs when entering a port polygon/geofence, a point in polygon spatial search
  - A port polygon is a closed set of points that surrounds a specific area (e.g. a port)
  - Manually constructed
- More detail algorithms (see Fuentes, 2021), would recognize the waiting time of the vessel within a polygon and even whether they carried out bunkering operations





NHH

Congestion measurement

- Congestion might be an indicative of two elements:
  - Slow service at port
  - Large demand
  - …or both
- A way to estimate congestion is by filtering the vessels that are within port polygon and then make a count of unique vessels ID's.

NHH

Other applications

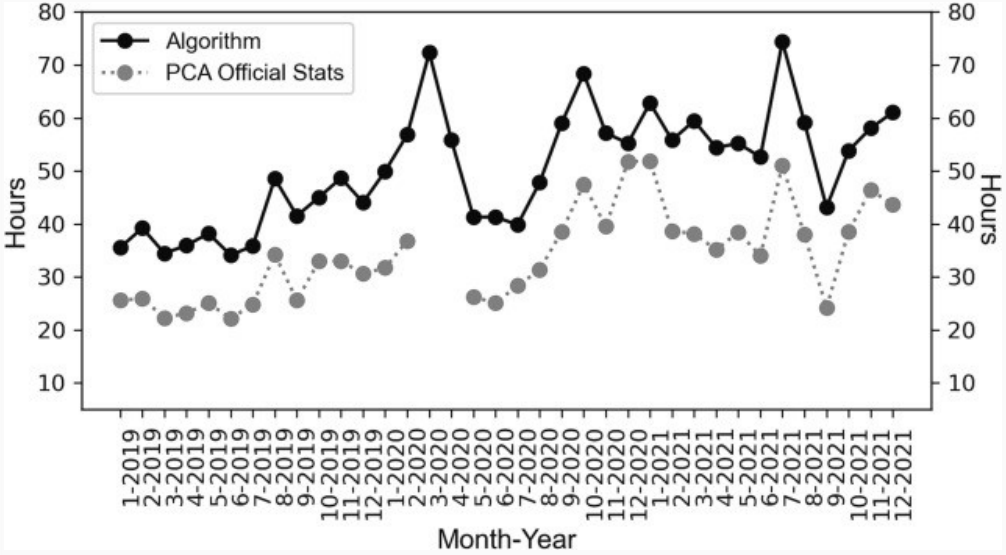
- Two of the data suppliers for this course use AIS data to derive some of their commercial information:
  - AXS Marine (commodity trading data)
  - Signal Ocean (vessels opening up at date x at port Y)
- Next port prediction (Zhang et al., 2020)
- Detecting dark vessels (Mazzarella et al., 2017)
- Vessel type classification (What type of vessel it is?) Zhong et al. (2019)
- Emissions estimation from vessels (Fuentes & Adland, 2023)
- Yang et al. (2019) summarizes many other applications

Data handling

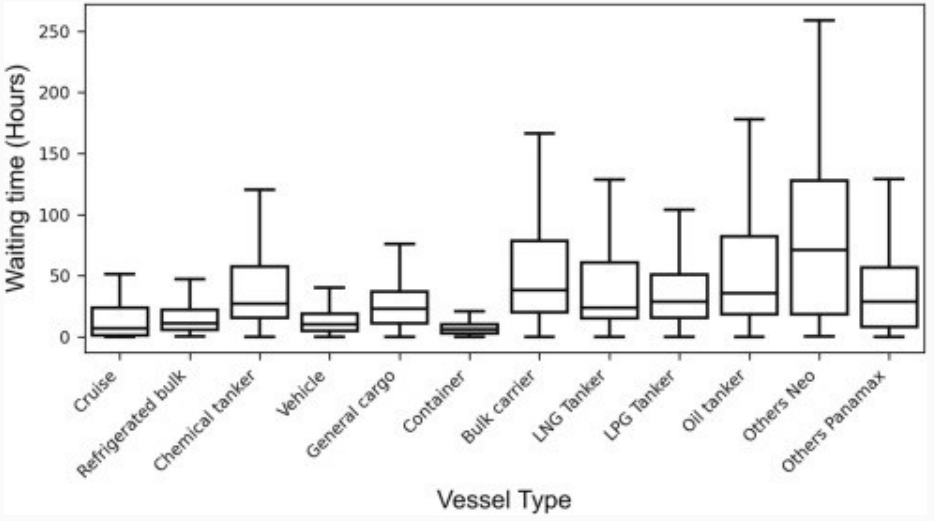
- Algorithms based on AIS can return spurious results
- How to recognize erroneous results:
  - Operational experience (e.g., a vessel at anchor for more than 6 months is either laid up or the algorithm fail due to missing AIS data)
  - Run outliers filtering (e.g., between 0.01 and 0.99 quantiles, does the trick many times)
- Segregating data (separating in recognizable categories to limit the analysis to a subsegment)
  - Per vessel type
  - Per size category
- Aggregating data (join data to have a global picture, e.g. Panama Canal waiting times)
- As an analyst you should be capable of shifting the scope based on the intended study. Some examples:
  - You want to learn about the waiting times at the Panama Canal so you retrieve a distribution of all waiting times. But if you segregate your data on vessel types, the picture changes completely.

NHH

Same information – different forms

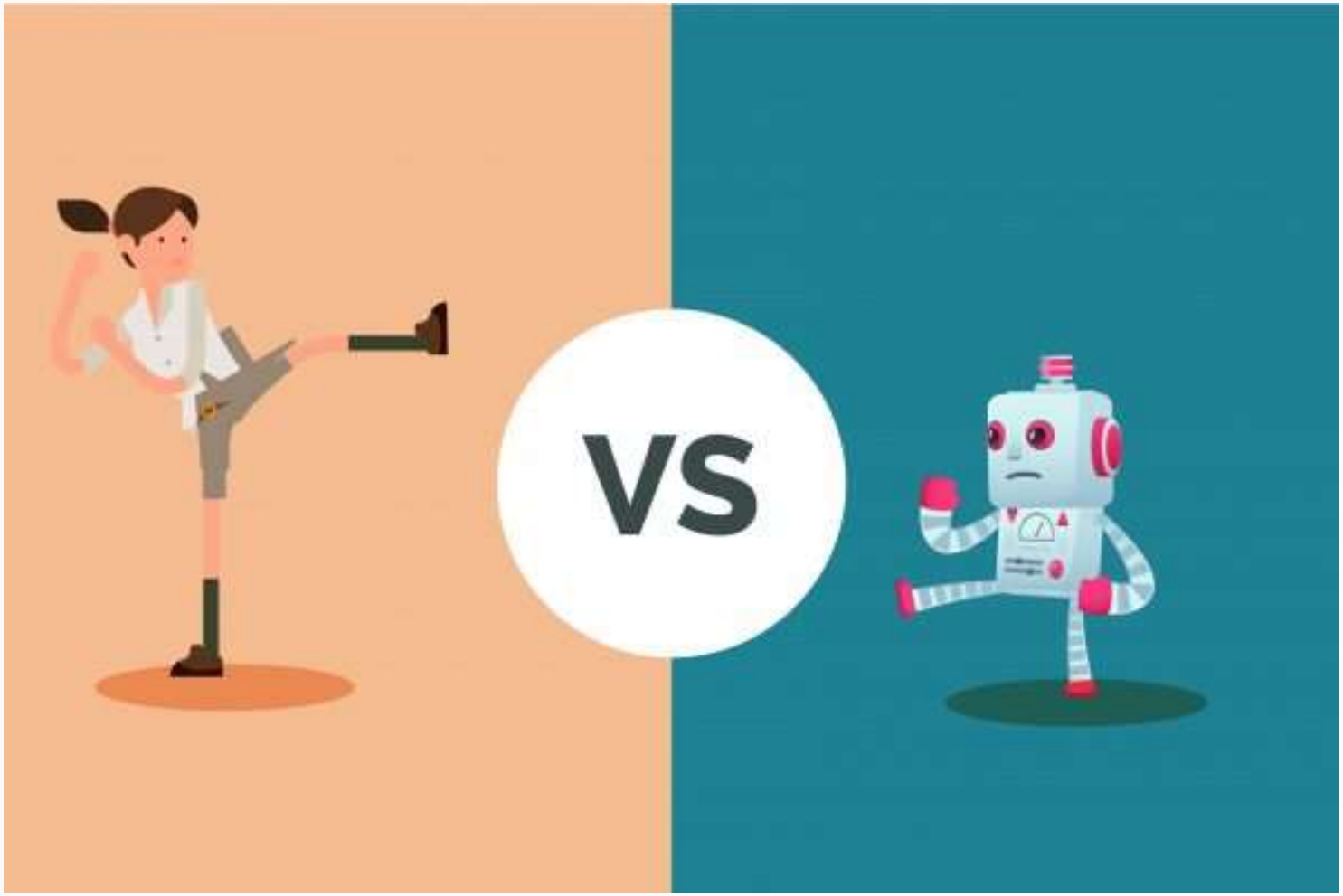


Mean : ~35 hours



Second Part – Machine Learning

NHH



NHH

Machine Learning

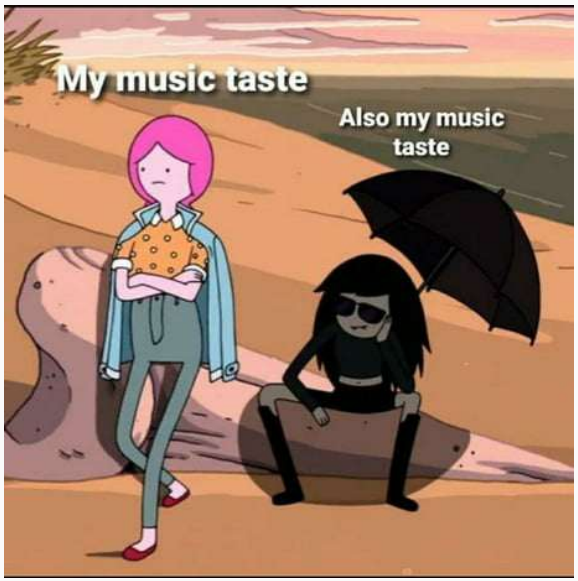
Just a buzzword?

- What is the general motivation?
  - M/L: Machines performing tasks without being specifically train
- In a human way, M/L resembles the ways humans learn from experience
- More data (generally) leads to better performance

NHH

But what it is?

Parse data → **L**earn from data → **A**pply what you have learned from data



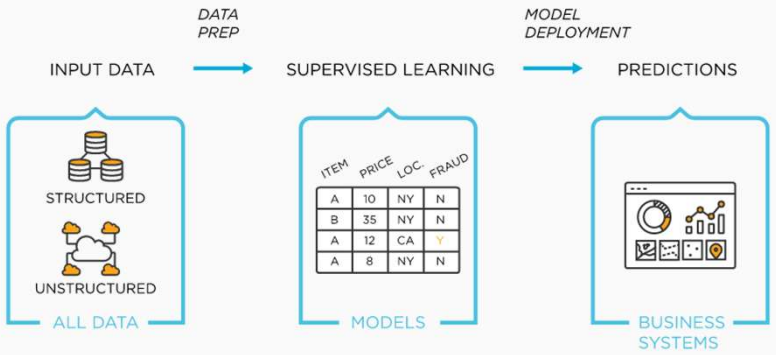
Then, it doesn't fit into a single category.

It is divided into:

- **Supervised learning**
- **Unsupervised learning**
- Reinforcement learning

NHH

Supervised learning



Source: TIBCO

- This is what you often see advertised as M/L
- Simple recipe
  - Train your model with labelled data
  - Test the model performance by comparing predictions against values stored for testing (out sample)
  - Improves by reducing the error (Predicted to actual values)
- Works on labelled data
- Two types of expected predictions
  - Regression
  - Classification
- Most of times the better performing models are so because they have a very good preprocessing of input data.
  - Remember that as in statistics, an M/L model behaves as good as what the input data has to offer. “Garbage in – garbage out”

NHH

Regression

Fuel consumption	Sailing Speed	Wind/Waves	Cargo Weight
0.78	9.5	1.33	32000
0.90	12.5	2	30000
1.10	12.5	3	35000
0.80	7.7	0.3	32000
0.95	9.0	0.9	0

Labelled data

Fuel consumption	Sailing Speed	Wind/Waves	Cargo Weight
0.78	9.5	1.33	32000
0.90	12.5	2	30000
1.10	12.5	3	35000

Training data

Fuel consumption	Sailing Speed	Wind/Waves	Cargo Weight
0.80	7.7	0.3	32000
0.95	9.0	0.9	0

Testing data

NHH

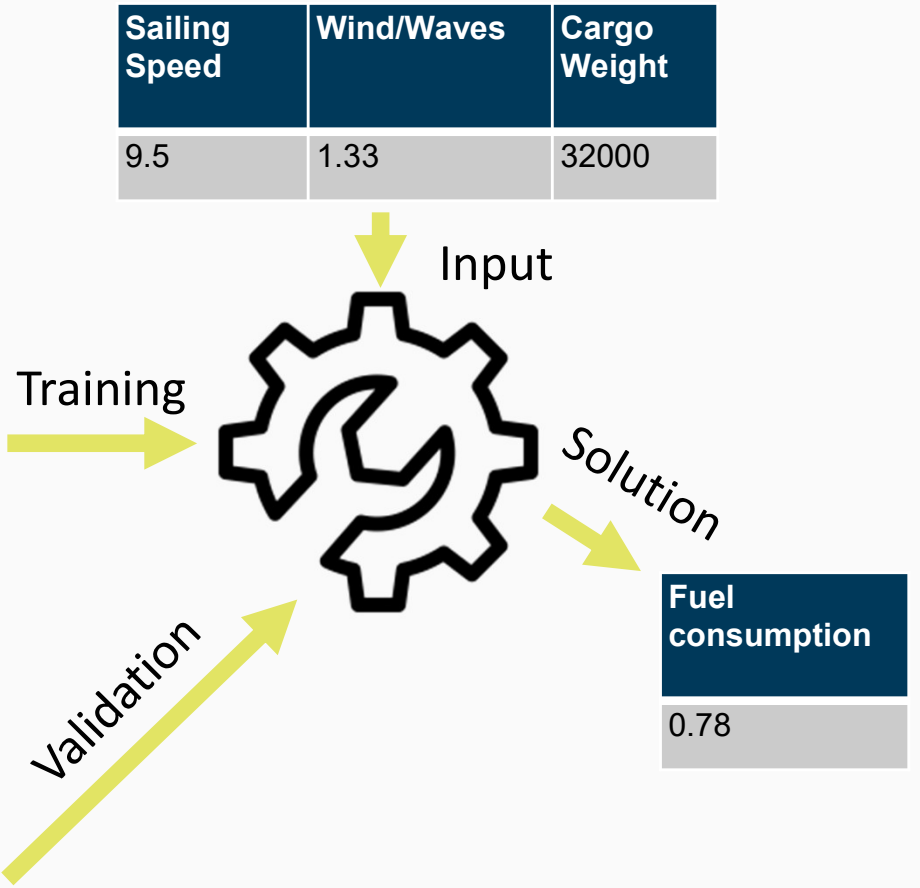
Regression

Fuel consumption	Sailing Speed	Wind/Waves	Cargo Weight
0.78	9.5	1.33	32000
0.90	12.5	2	30000
1.10	12.5	3	35000

Training data

Fuel consumption	Sailing Speed	Wind/Waves	Cargo Weight
0.80	7.7	0.3	32000
0.95	9.0	0.9	0

Testing data

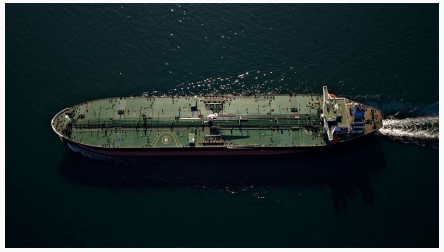


NHH

Classification

Input

BULK CARRIER



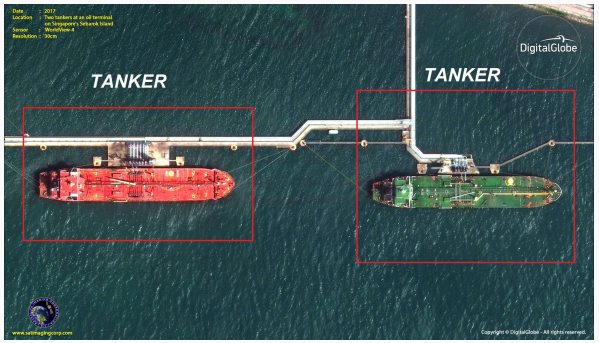
TANKER

Training data

Training

Solution

TANKER? BULK CARRIER?

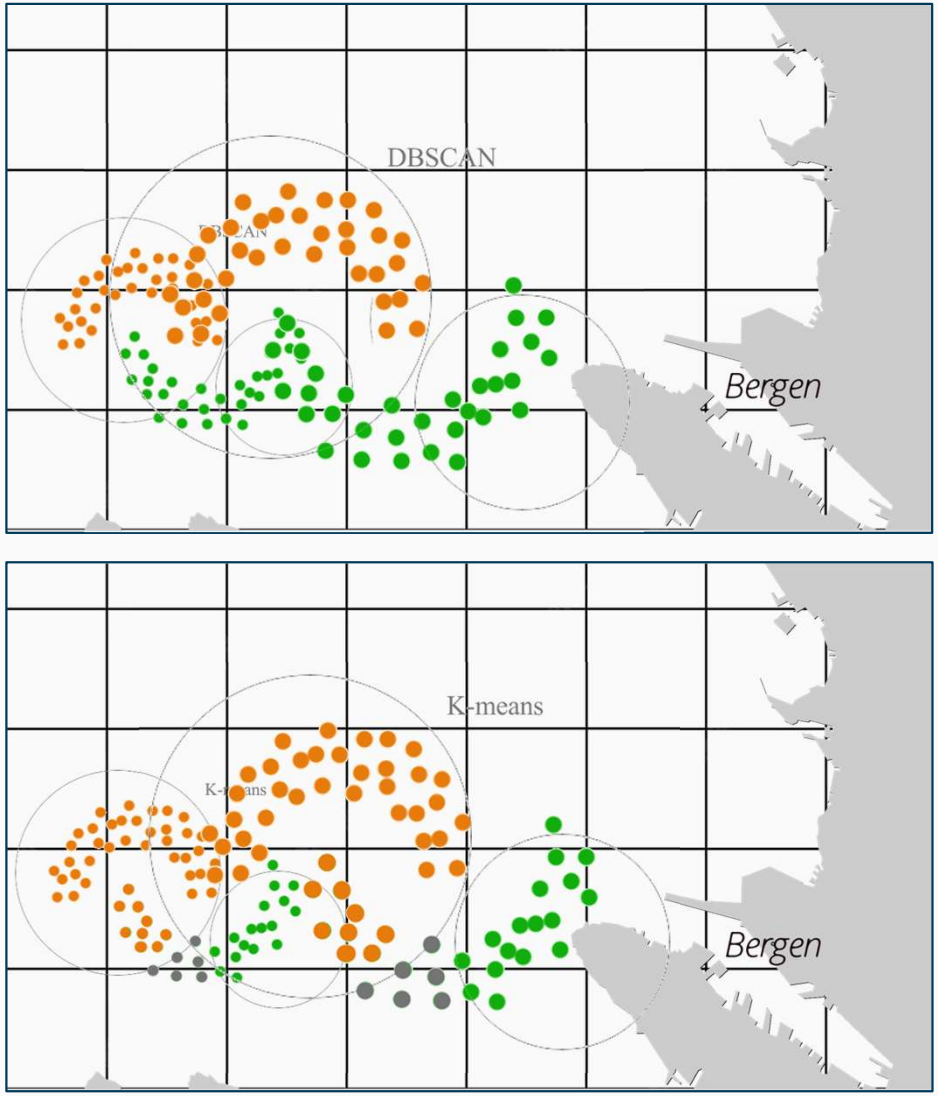


Validation

Testing data

NHH

Unsupervised learning



- **Do not** need **labelled** data
- Main applications
  - Dimensionality reduction (i.e., Principal Component Analysis)
  - Clustering (Density, Hierarchical or Overlapping)
- The goal is to gain insight from data, contrary to knowing what type of solution to expect (supervised) (IBM, 2023)

NHH

Why not just use M/L instead of statistics?

M/L

Pros

Can handle non linearities with ease

Generally better for predictions

Cons

Overfit when wrongly tuned

Black box

Low Interpretability

Statistics

Pros

High interpretability

Sound mathematical support

Better for policy support

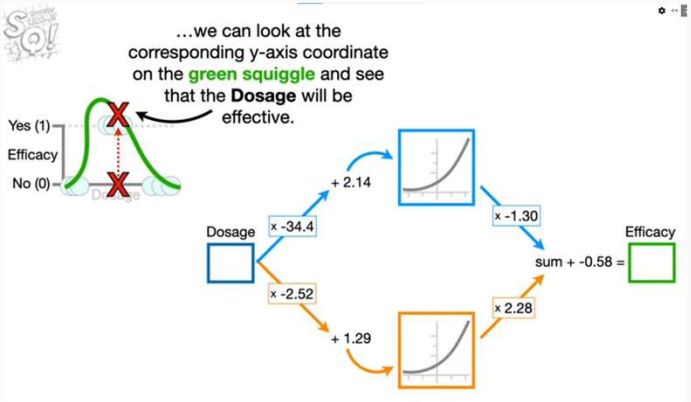
Cons

Strong assumptions

NHH

Why not just use M/L instead of statistics?

Linear model



Squiggle graph model adapted from Neural Network (non-linear models)

NHH

A M/L recipe

1. Identify your predicted variable and your predictors
2. Conduct a descriptive statistics analysis and write down your thoughts on patterns
3. Clean up the data. Fill empty values, normalize if needed be, impute.
4. Split your data into training data and test data 80% and 20% rule of thumb
5. Train your model (80% sample). Use various model options
6. Test your model with out-sample data (20% sample)
7. Evaluate Model Performance

NHH

References

- Fuentes, G., Adland, R. (2023). Greenhouse gas mititgation from chokepointd: The case of the Panama Canal. *Transportation Research Part E.*
- Mazzarella, F., Vespe, M., Alessandrini, A., Tarchi, D., Aulicino, G., & Vollero, A. (2017). A novel anomaly detection approach to identify intentional AIS on-off switching. *Expert Systems with Applications*, *78*, 110-123.
- Yang, D., Wu, L., Wang, S., Jia, H., & Li, K. X. (2019). How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. *Transport Reviews*, *39*(6), 755-773.
- Zhang, C., Bin, J., Wang, W., Peng, X., Wang, R., Halldearn, R., & Liu, Z. (2020). AIS data driven general vessel destination prediction: A random forest based approach. *Transportation Research Part C: Emerging Technologies*, *118*, 102729.
- Zhang, T., Yin, J., Wang, X., & Min, J. (2023). Prediction of container port congestion status and its impact on ship's time in port based on AIS data. *Maritime Policy & Management*, 1-29.
- Zhong, H., Song, X., & Yang, L. (2019, July). Vessel classification from space-based ais data using random forest. In *2019 5th International Conference on Big Data and Information Analytics (BigDIA)* (pp. 9-12). IEEE.
- https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

NHH