UN Big Data Hackathon

KnowAIS

Self-Service Analytics Platform for Global Shipping Networks

https://github.com/datacerdas/know-ais

Cerdasdata

UN Big Data Hackathon Expert Track
8-11 November 2022

Cerdasdata

Team Members

Amanda Pratama Putra
mandes95@gmail.com

Ignatius Aditya Setyadi
adityasetyadi@bps.go.id

Sugiri
sugiri@bps.go.id

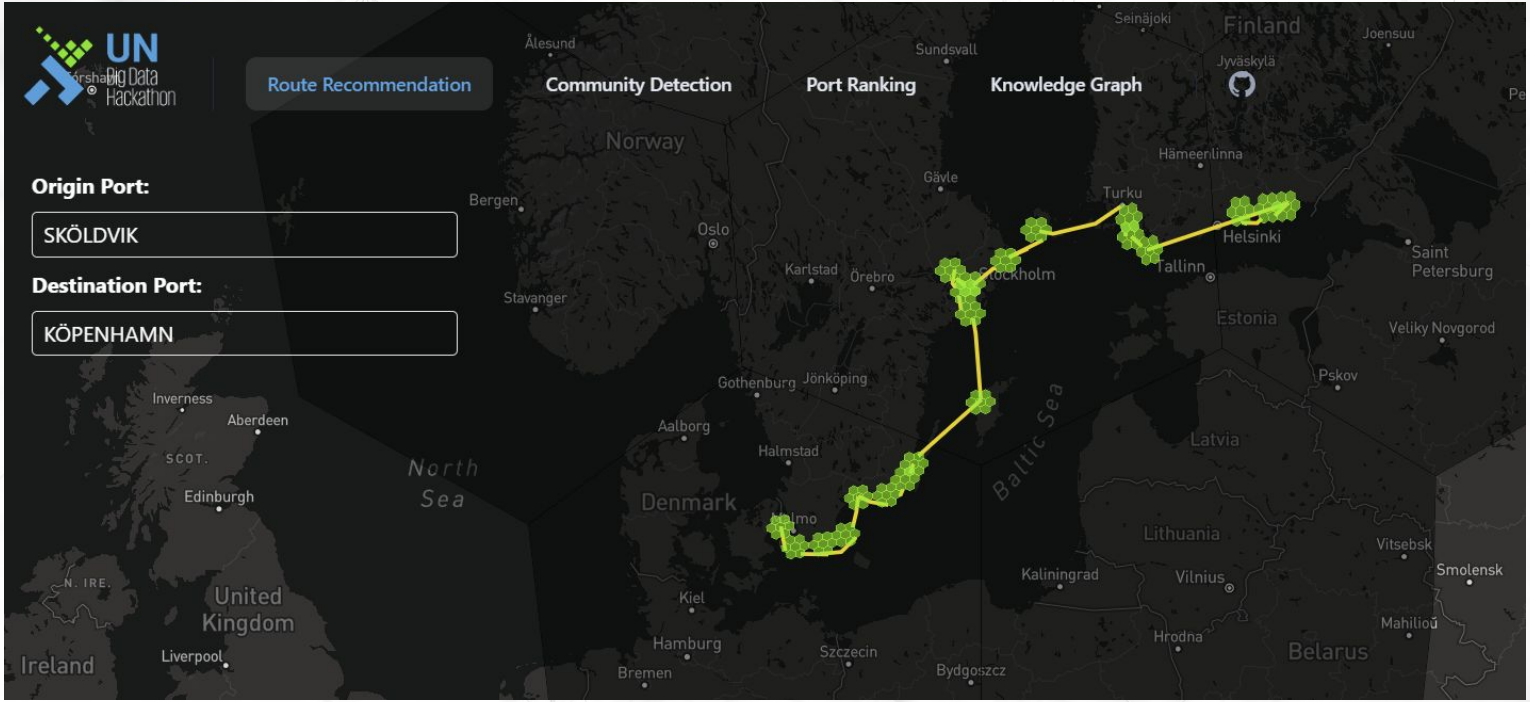
Wismu Sunarmodo
wismusun@gmail.com

Amin Rois Sinung Nugroho
sinung@bps.go.id

When everything is connected to everything else,
for better or for worse, everything matters

Bruce Mau

https://know-ais.vercel.app/
https://know-ais.datacerdas.id/



World Maritime
Transport

https://unctad.org/webflyer/review-maritime-transport-2021

Maritime transport is the backbone of international trade and the global economy. **Over 80%** of the volume of international trade in goods is carried by sea, and the percentage is even higher for most developing countries.

Big data can help to solve this problem by providing a more **efficient** and **accurate** way to track maritime trade and transportation.

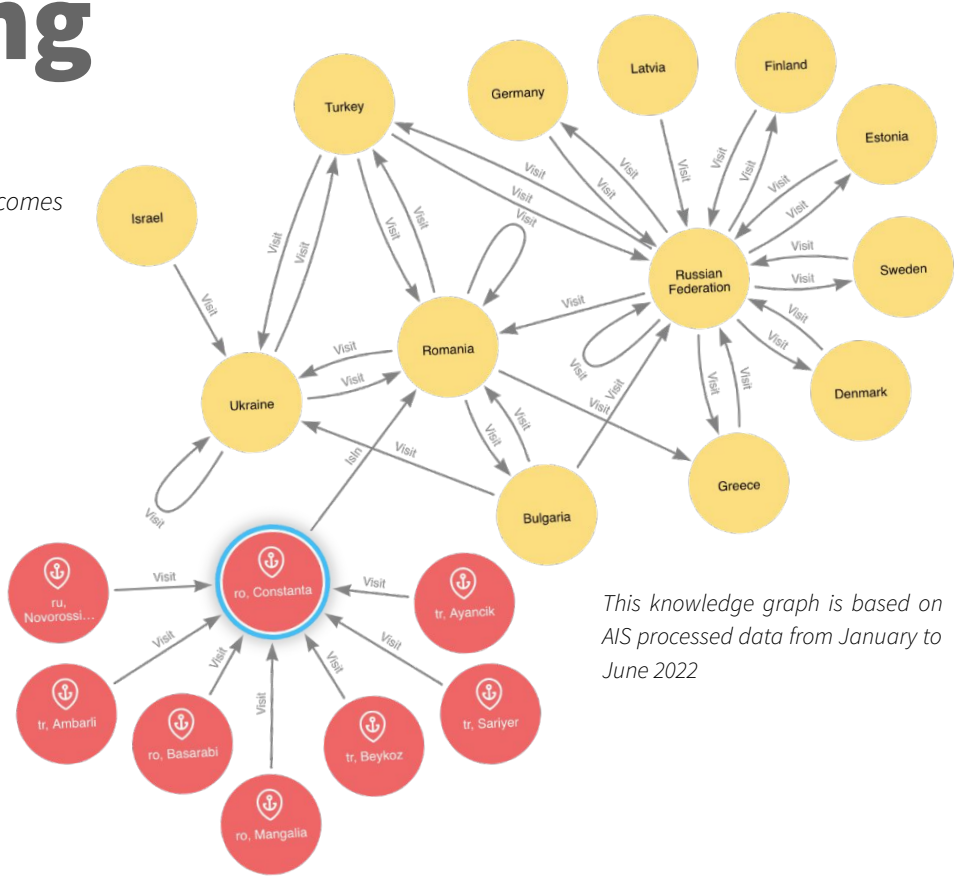
By using big data, maritime companies can track their shipments in real time, identify inefficiencies and delays, and make better decisions about how to route their shipments. Fostering **connectivity of the port** has become an inevitable duty for every stakeholders to boost world trade activity and global value chain.

Ukraine Shipping
Move to Romania

https://www.aljazeera.com/gallery/2022/6/16/photos-romanian-port-becomes-key-transit-hub-for-ukrainian-grain

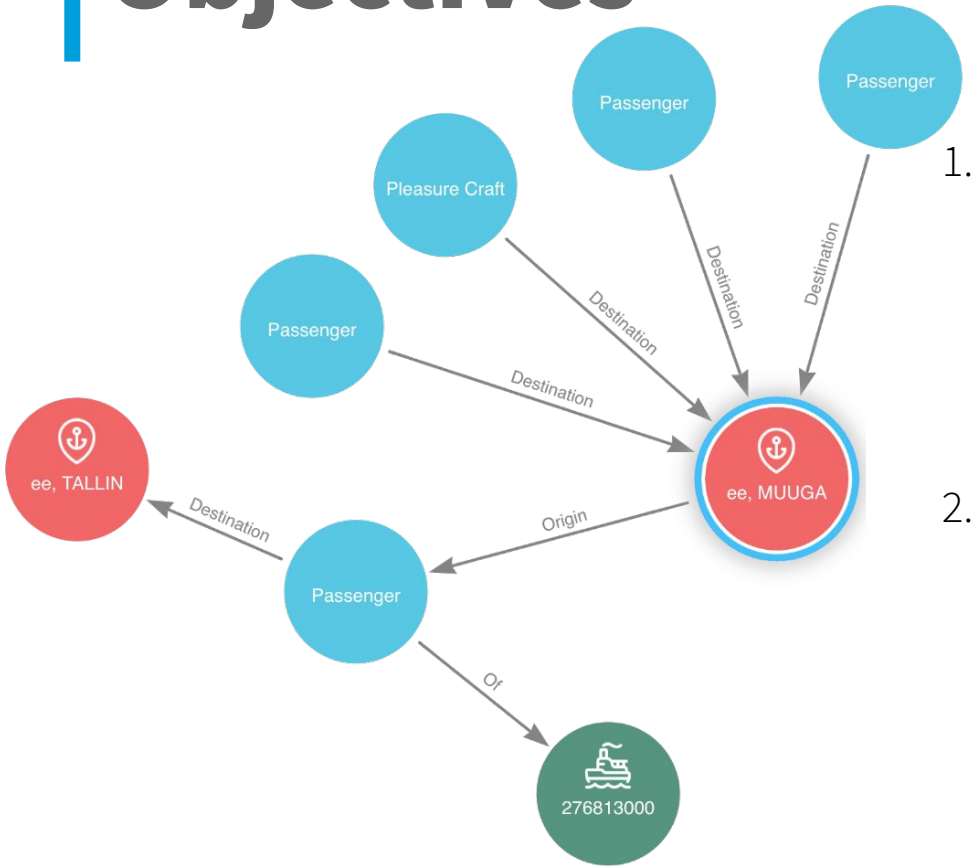
Romanian port becomes key transit hub for Ukrainian grain (June 16th, 22)

"*Before the war, there was absolutely no Ukrainian cereal transiting the* ***Constanta port****. Now, it amounts to approximately 30 percent of our activity,*" said Dan Dolghin, director of the cereal operation at Comvex, one of the main grain operators in the Constanta port.



This knowledge graph is based on AIS processed data from January to June 2022

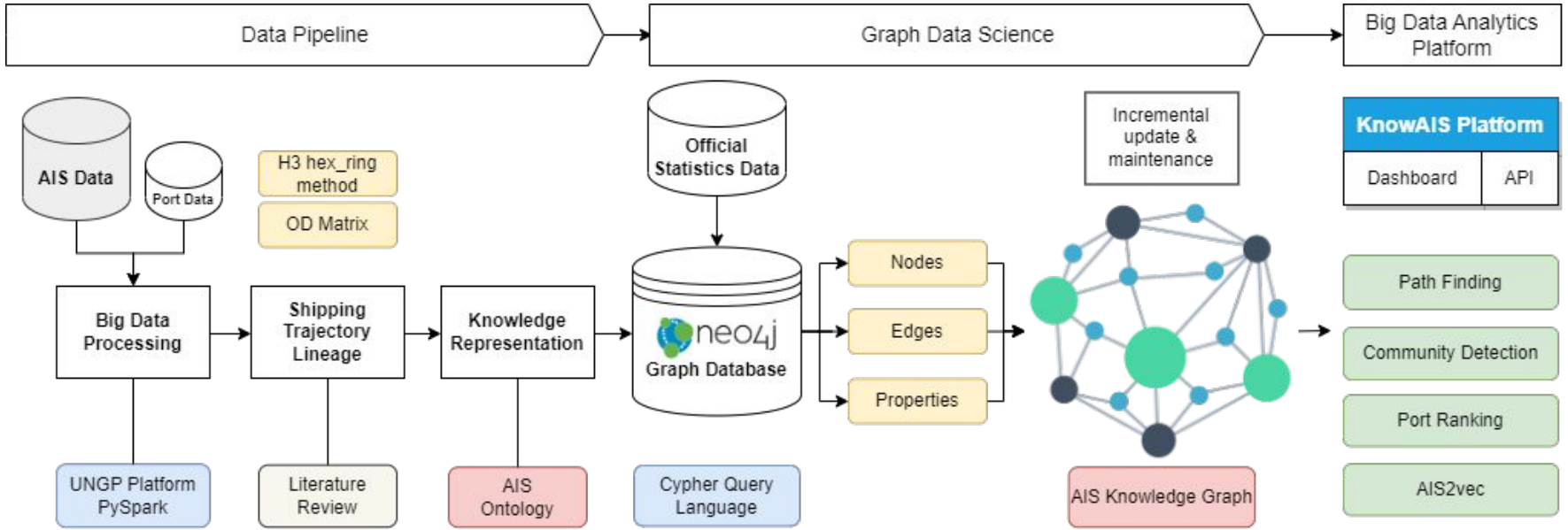
Objectives



1. Develop **decision support system** to process AIS Big Data in timely manner through insight generation and monitoring of **global shipping network**.

2. Conduct analysis using Data Science and Big Data related to the relationship between global shipping networks and statistical indicators of the world economy.

Platform
Architecture

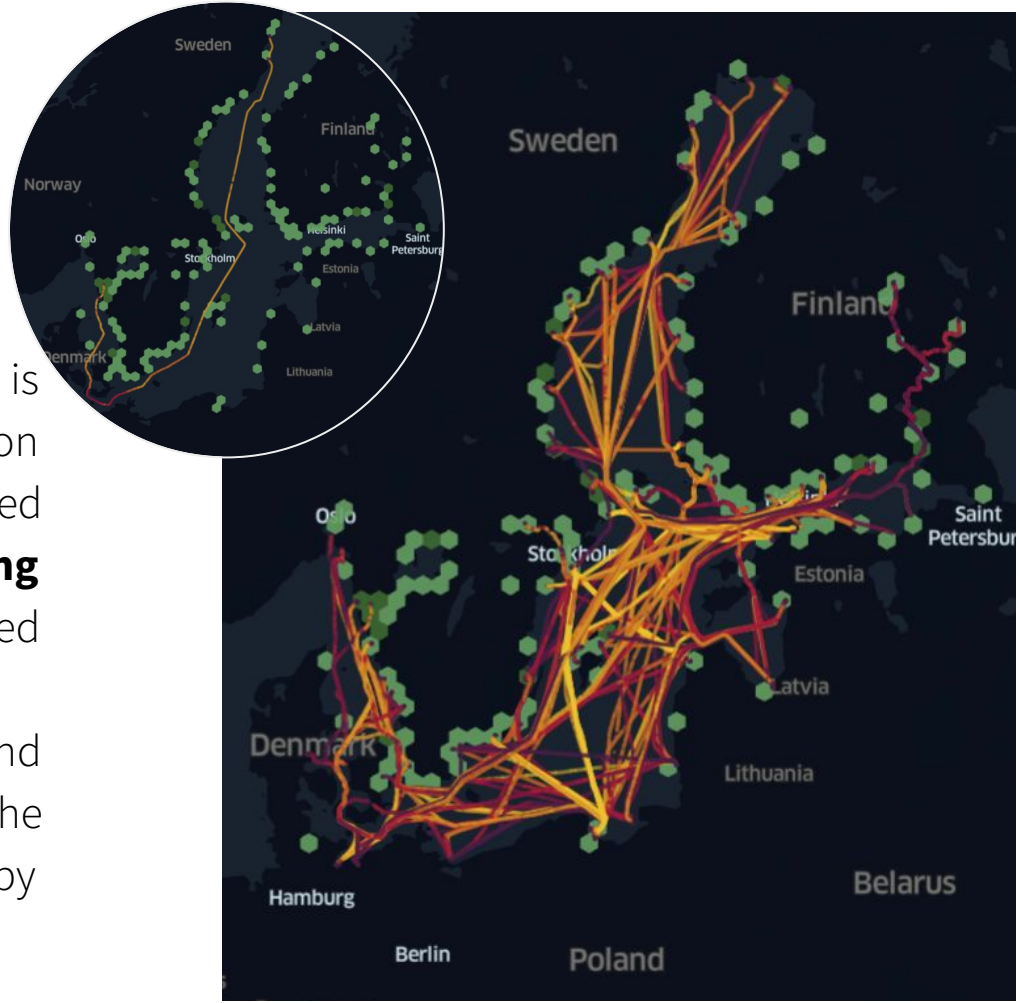


Big Data
Processing

UN GLOBAL PLATFORM FOR OFFICIAL STATISTICS

AIS Data is made up of **tons of data** that is captured on a **daily basis**. Initial filtration is needed to get the most needed information during the process, **ensuring** that the feature selection process is carried out in the **best possible way**.

For the purpose of proof of concept and increasing the processing time during the hackathon, we did **sampling** dataset by maintaining vessel coverage >80%.

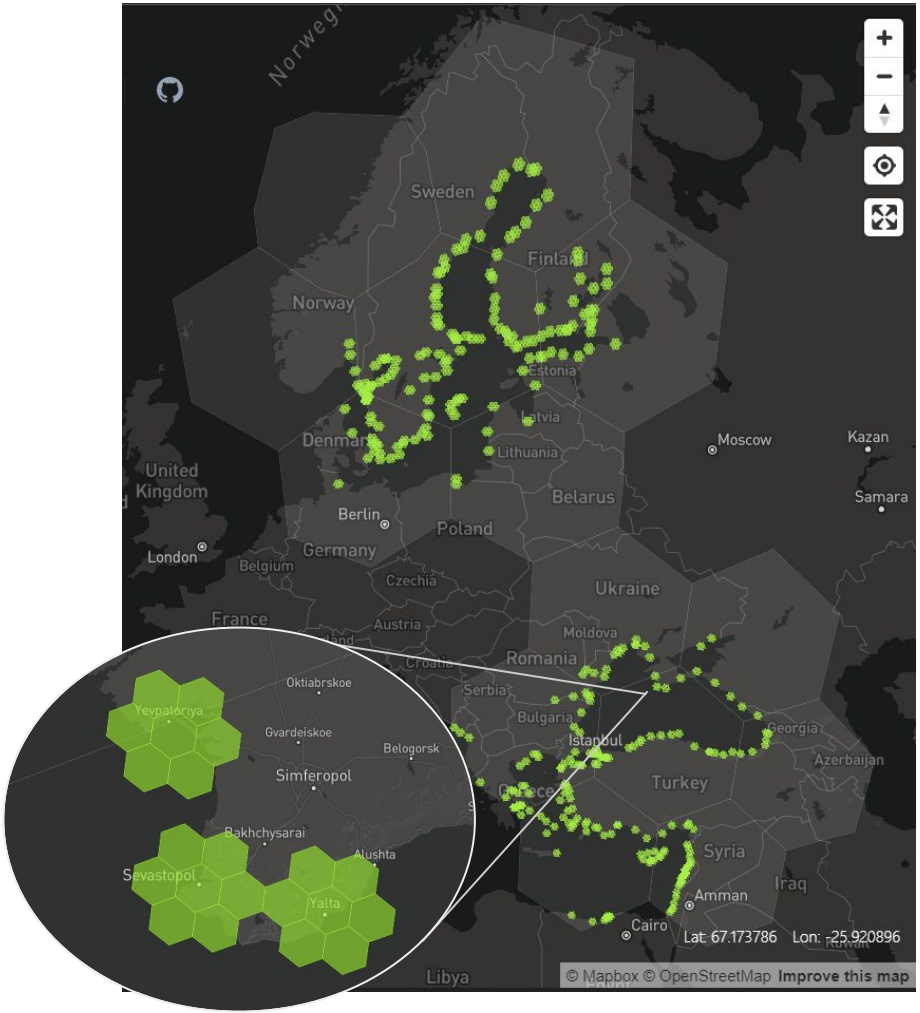


Ports Data

https://geonode.wfp.org/layers/esri_gn:geonode:wld_trs_ports_wfp
https://github.com/hakola/marine-traffic-modelling

Data taken from the World Food Programme (WFP) and filtered only by area.

The designated **h3 resolution 5** along its hex ring from its centroid point is then defined as the polygon to be utilized for trajectory identification.



Official Statistics

GVC: https://www.worldmrio.com/unctadgvc/Database_GVC_2018update_rev0323.csv
GDP 2022: https://www.imf.org/external/datamapper/datasets

The UNCTAD-Eora **Global Value Chain (GVC)** database offers global coverage (189 countries and a “Rest of World” region) and a time series from 1990 to 2018 of the key GVC indicators: foreign value added (FVA), domestic value added (DVA) and indirect value added (DVX).

Results from 1990 to 2017 are generated from EORA Multi-Region Input-Output tables (MRIOs). Results for 2016-2017 are provisional "beta" results and will be revised in early 2018. Results for 2018 are nowcasted based on the IMF World Economic Outlook.

Scope
of Experiment

Baltic and Black Seas are picked as the targeted zone to analyze the impact of the **Ukraine-Russia war**.

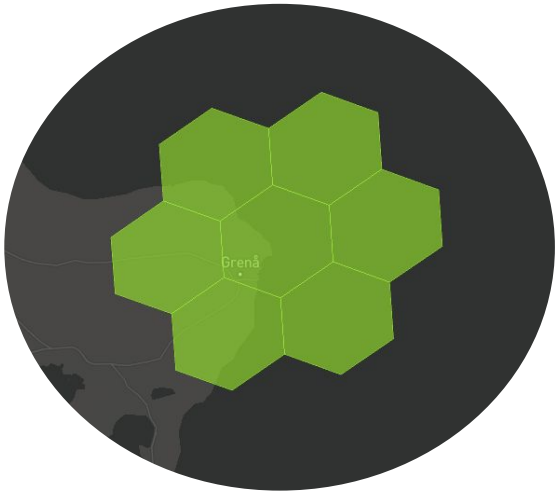
The selection across the zone is made using **H3 resolution 1** and its **hex ring**, which consists of 14 areas, in the period of January to June 2022 (~10M records after sampling), as well as the same month period from 2021 (~9M records) for comparison.

Initial **Filtration**

For better performance **dealing with Big Data**, initial filtration is performed on selected columns, h3_list in resolution 1, and date range

```
%%time

# CPU times: user 42.6 ms, sys: 18.1 ms, total: 60.8 ms
# Wall time: 1min 45s

# date inputs should be in date time format. dt_insert_utc is the basis for the parquet partitions
columns = [
            'mmsi','dt_insert_utc','longitude','latitude',
            'vessel_name','vessel_type','vessel_type_main','vessel_type_sub','vessel_type_code','vessel_type_cargo','vessel_class',
            'length','width','flag_country','destination','eta','source','H3_int_index_5'
          ]

h3_list = [ 581118283558682623, 581109487465660415, 581122681605193727, 581135875744727039, 581267817140060159, 581514107744681983, 581518505791193087,
            581760398349303807, 581500913605148671, 581509709698170879, 581742806163259391, 581764796395814911, 581769194442326015, 582081455744614399

start_date = datetime.fromisoformat("2022-01-01")
end_date = datetime.fromisoformat("2022-06-30")

# spark is the current spark session you are using. Gets automatically created during kernel init.
df = af.get_ais(spark, start_date, end_date = end_date, columns = columns, h3_list = h3_list)
```

Port Activity
Identification



Identification activity in the port area is done by joining with the same **h3 index resolution 5** between AIS Data and Ports. This approach has proven produced better performance than scripts using GIS features such as ST_WITHIN, etc.

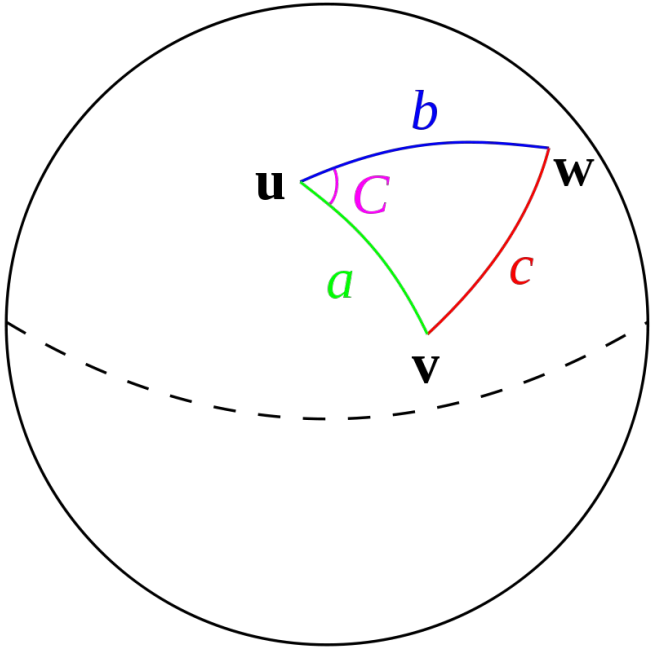
It also **reduces the need to define specific coordinates locations** in order to manually generate port polygon area.

```
%%time

# CPU times: user 7.22 ms, sys: 921 µs, total: 8.14 ms
# Wall time: 1min 25s

ais_01_ports = ais\
                    .join(
                        ports\
                          .withColumn('H3_int_index_5', col('h3_5_int'))
                        , ['H3_int_index_5'], 'left'
                    )

ais_01_ports.show(10)
```

Speed & Distance
Calculation

The distance is measured using the **Haversine formula** and carried out between movements, then the speed calculation is carried out afterwards.

The distance is carried out in km, as well as the speed produced in km per hour (kmh)



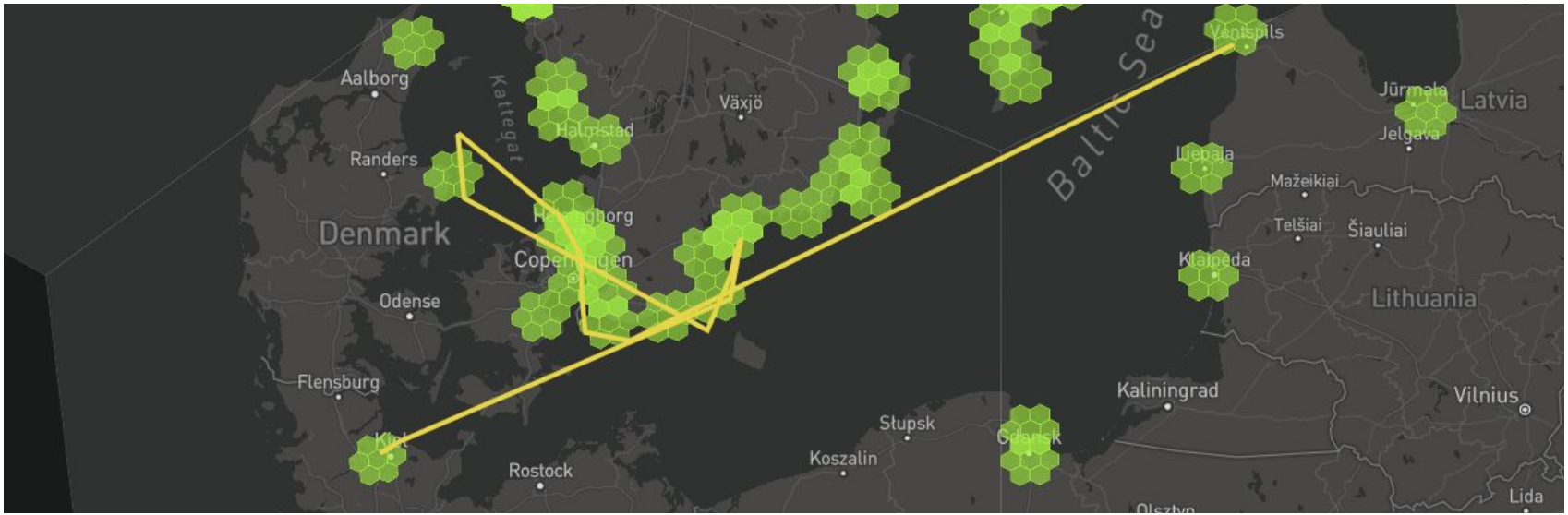
Anchored
Identification

Anchored identification is tracked from several conditions, mainly based on the **length of time in port** and the time difference which tends to be longer than the next emergence, which is currently limited to a minimum of ± 2.5 hours.

```
part5 = Window.partitionBy('mmsi').orderBy(['start_timestamp'])
part6 = Window.partitionBy('mmsi').orderBy(['start_timestamp']).rowsBetween(Window.unboundedPreceding, Window.currentRow)

ais_05_shipping = ais_04_movement\
                    .withColumn('lag_start_timestamp', f.lag('start_timestamp').over(part5) )\
                    .withColumn('lag_end_timestamp', f.lag('end_timestamp').over(part5) )\
                    .withColumn('lead_start_timestamp', f.lead('start_timestamp').over(part5) )\
                    .withColumn('lead_end_timestamp', f.lead('end_timestamp').over(part5) )\
                    .withColumn('lag_time_interval', f.to_timestamp(col('start_timestamp')).cast('double') - f.to_timestamp(col('lag_end_timestamp')).c
                    .withColumn('lead_time_interval', f.to_timestamp(col('lead_start_timestamp')).cast('double') - f.to_timestamp(col('end_timestamp'))
                    .withColumn('shipping_status', when( col('lag_time_interval').isNull(), 0 )\
                                                   .when( col('lag_time_interval') > 10000, 0 )\
                                                   .when( col('lead_time_interval').isNull(), 2 )\
                                                   .when( col('lead_time_interval') > 10000, 2 ).otherwise(1) )\
                    .withColumn('shipping_gap', when( col('shipping_status') == 0, 1).otherwise(0) )\
                    .withColumn('shipping_movement', f.sum('shipping_gap').over(part6) )

ais_05_shipping.show(10)
```

Trajectory
Result

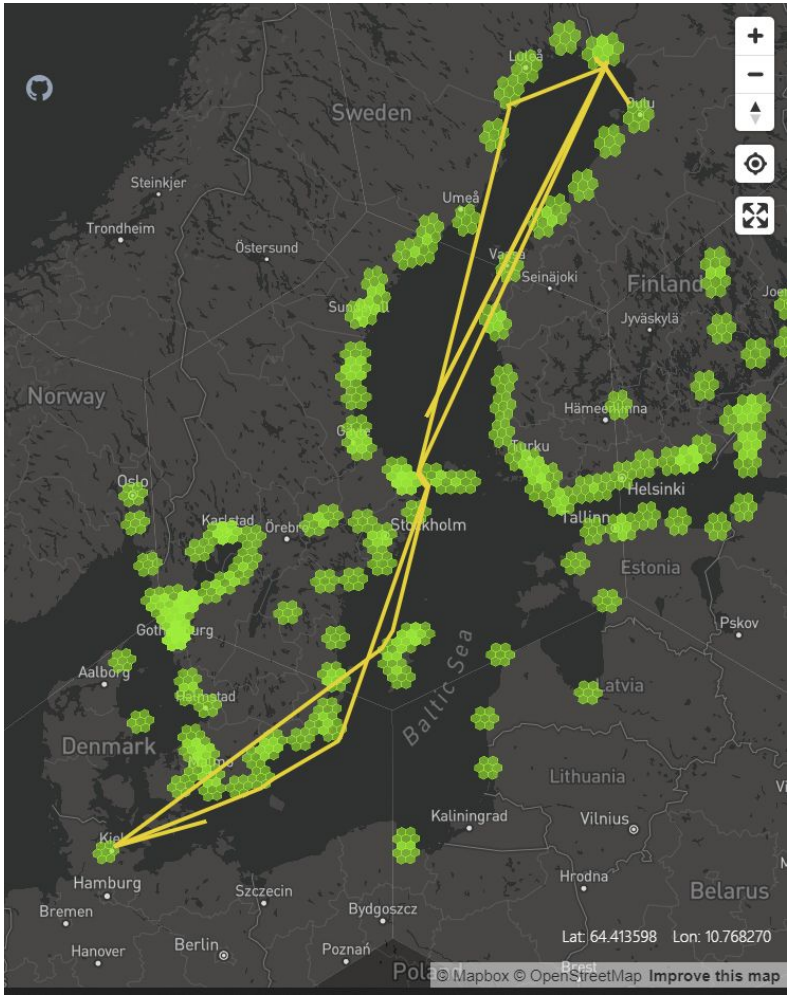
Id_shipping | origin_port | destination_port | avg_speed | distance | vessel_attributes



Shipping
Lineage

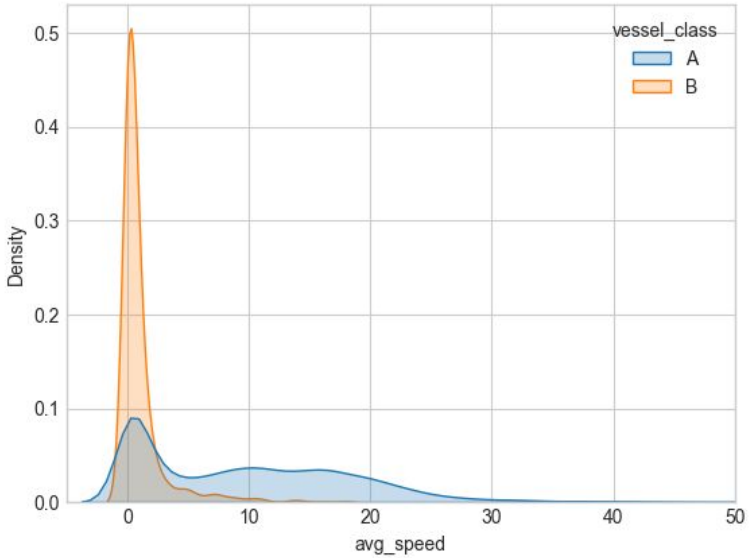
The trajectory results contain the **shipping lineage attributes**, namely average speed, distance, and vessel identity such as mmsi, type, width, length, country, etc.

The distance is calculated by aggregating each movement of the vessel during the trip rather than simply calculated directly from the place of origin to the destination.



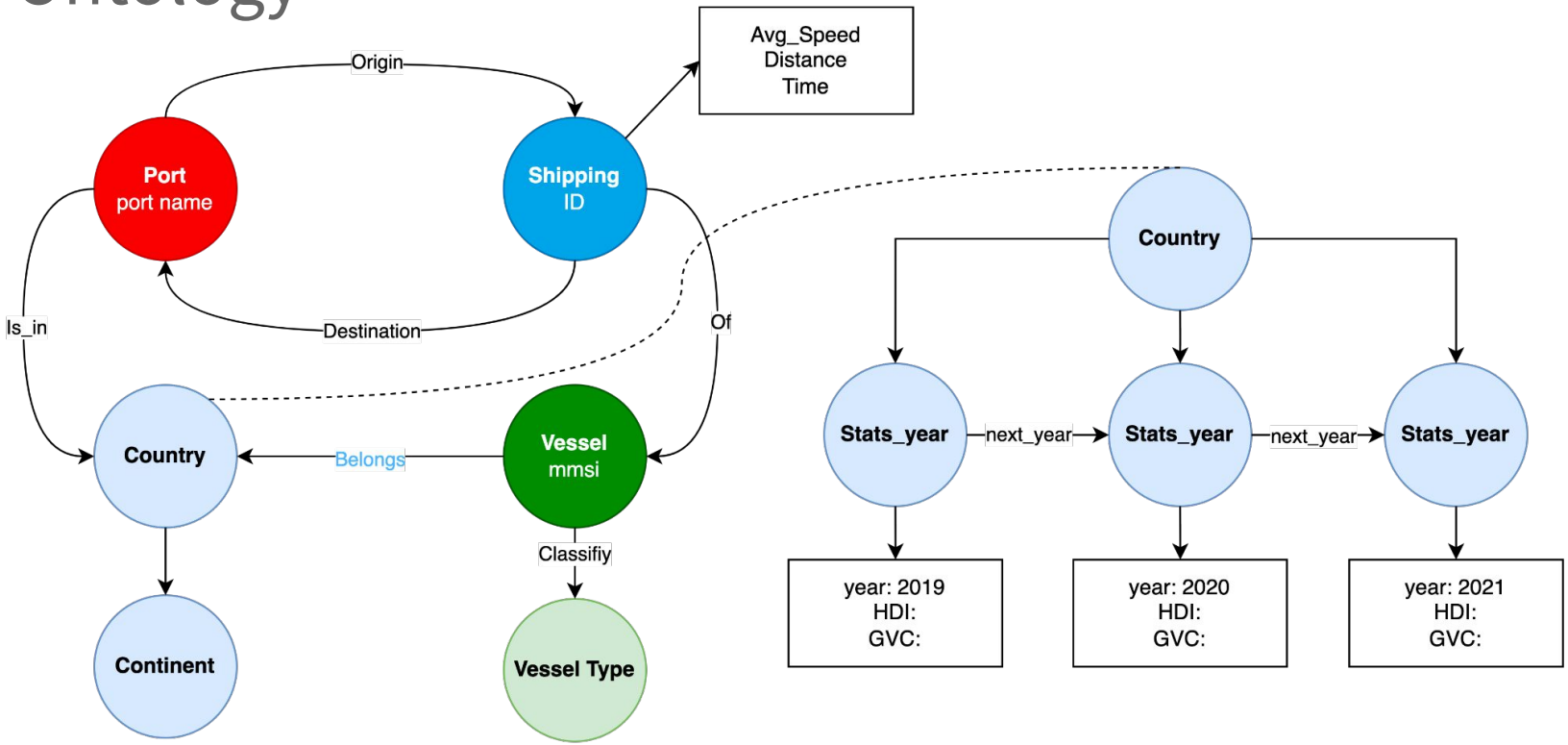
Data
Quality Assurance

Based on trajectory lineage, it shows that vessel class A have a high average speed compared to vessel class B.



vessel_class	Cargo	Passenger	Pleasure Craft	Sailing	Tanker	Tug
A	10265	2190	184	92	3213	921
B	56	1379	918	547	150	67

AIS Data
Ontology

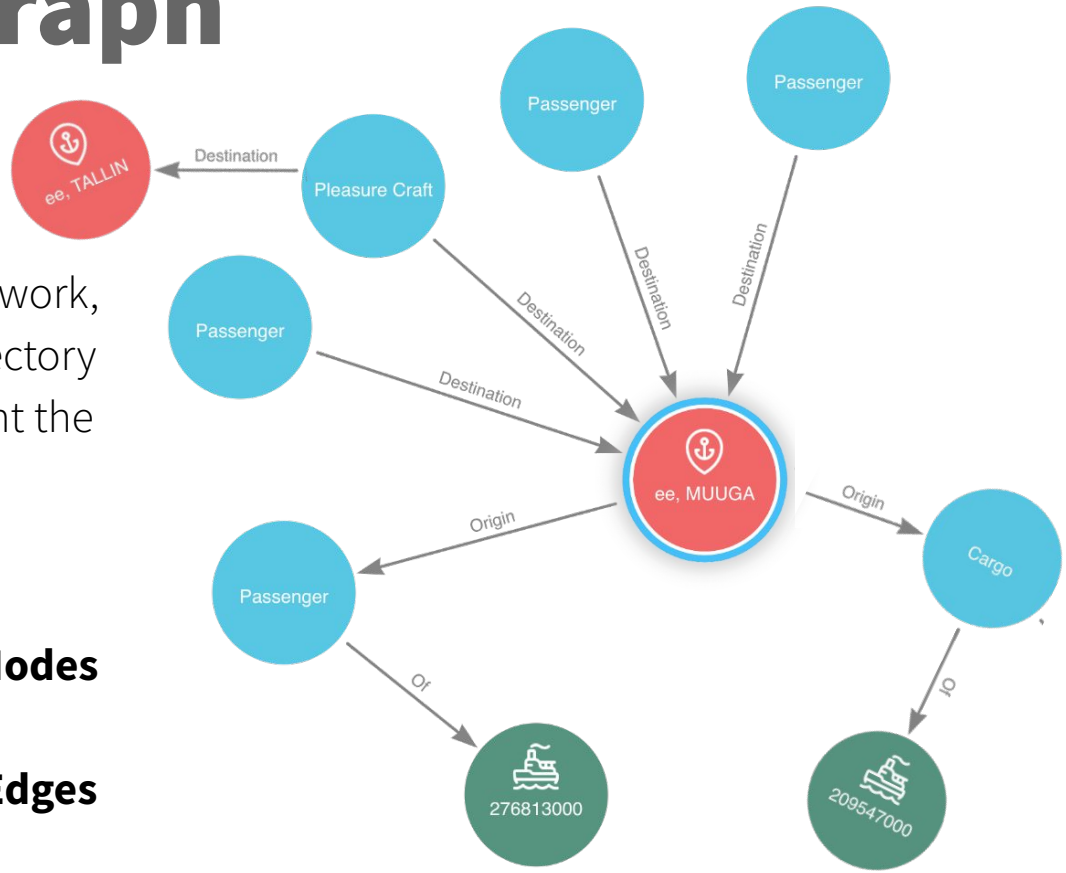


Knowledge Graph
Representation

Using proposed ontology as a framework, we can add data about shipping trajectory lineage generated from AIS to represent the real interaction between entities.

Result:

- **31.607** **Nodes**
  (Country, Port, Ship, & Shipping)
- **68.727** **Edges**
  (Destination, Origin, IsIn, Of)

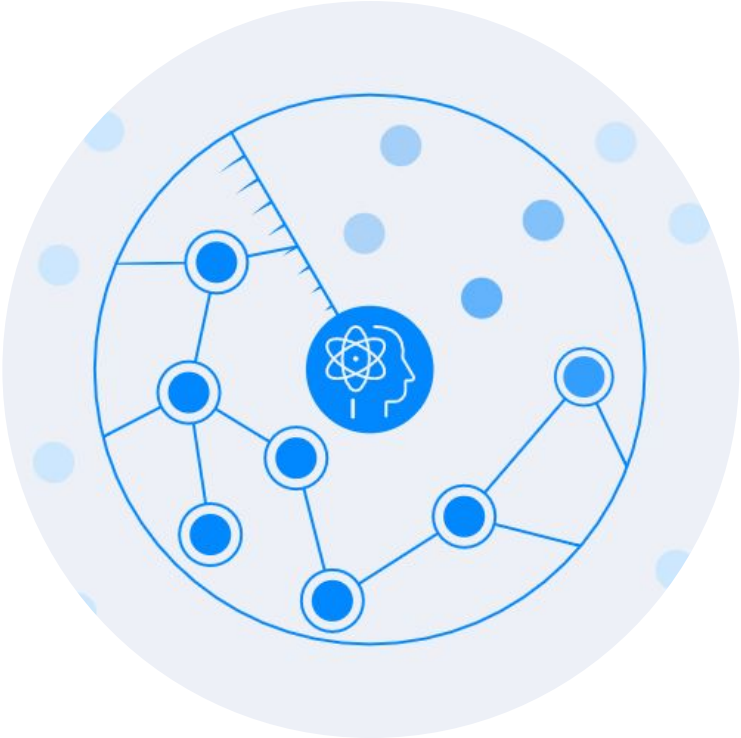


Unlock Insight with
Neo4j Graph Data Science

https://neo4j.com/docs/graph-data-science/current/

1. Route Recommendation
2. Country & port clustering
3. Country & port ranking
4. AIS2Vec: Country & Port Vector Embedding based on AIS Shipping Network

All of these use cases are simply solved using all-in-one **Neo4j Graph Data Science** library



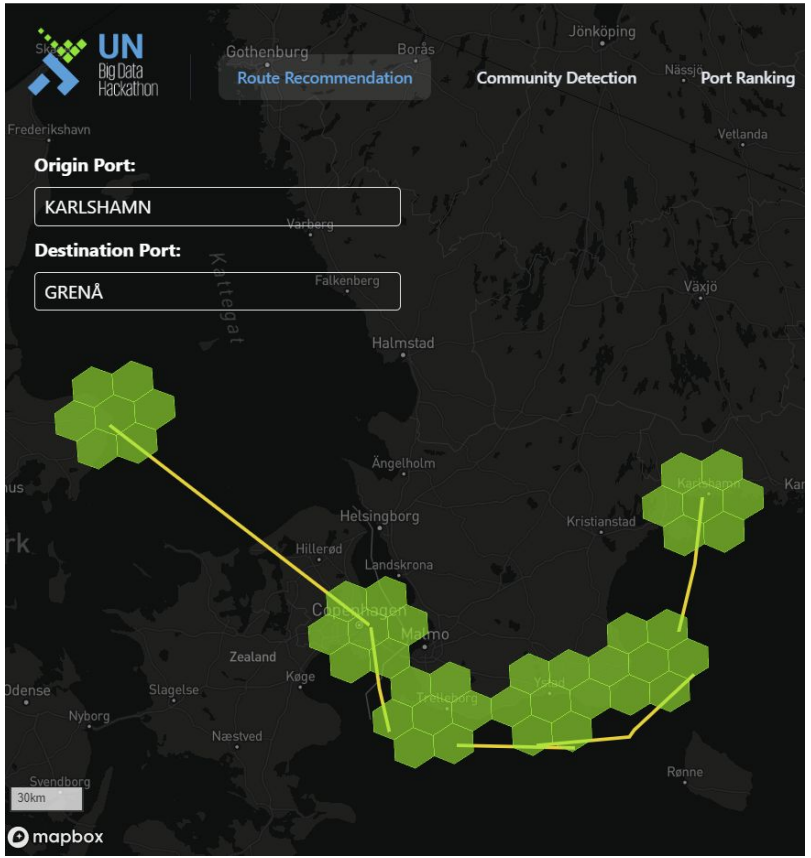
KnowAIS Platform

Route Recommendation Prototype

Port-to-Port
Route Recommendation

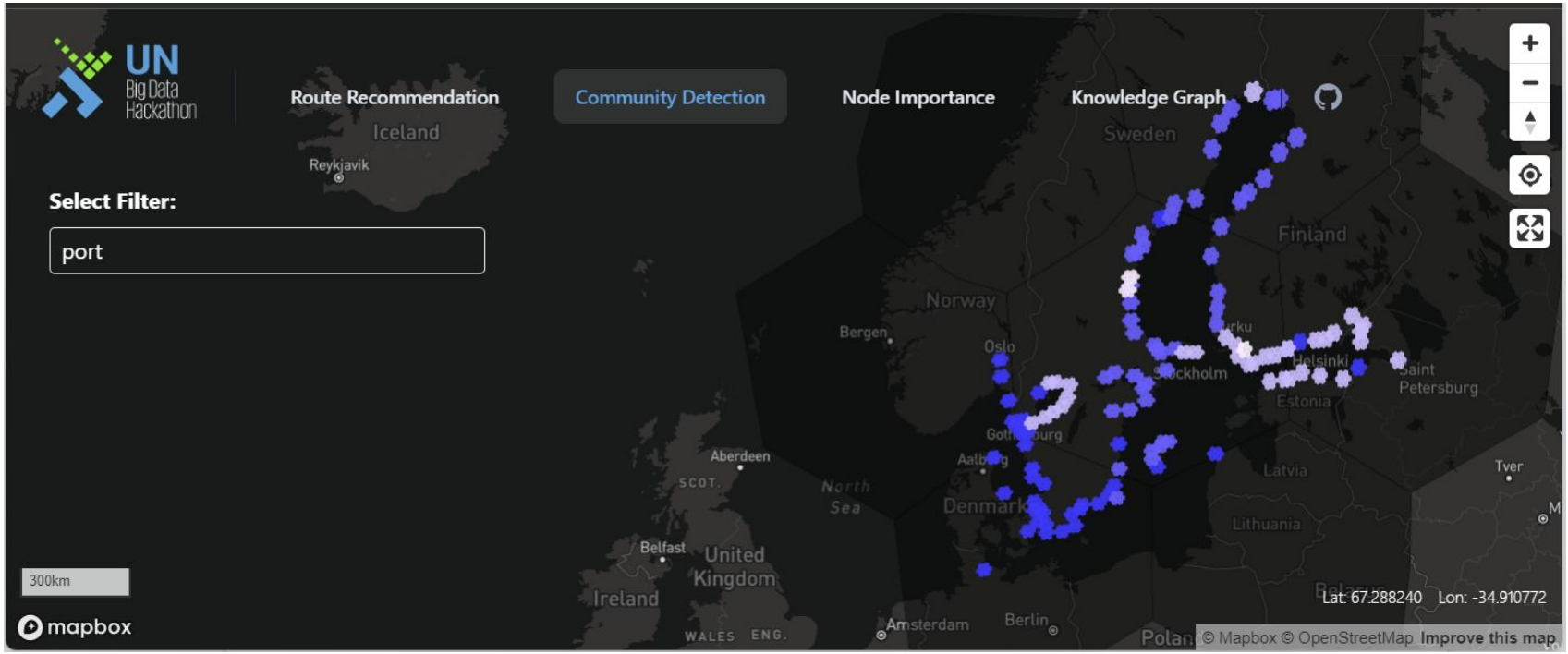
Yen's Shortest Path algorithm* computes a number of K-shortest paths between two ports. It can be used to find the **efficient shipping route** based on historical data.

The link between each ports in graph database can be dynamically updated based on **real condition** like **port close** due to disaster, **conflicts**, or **blockages**.



KnowAIS Platform

Community Detection Prototype

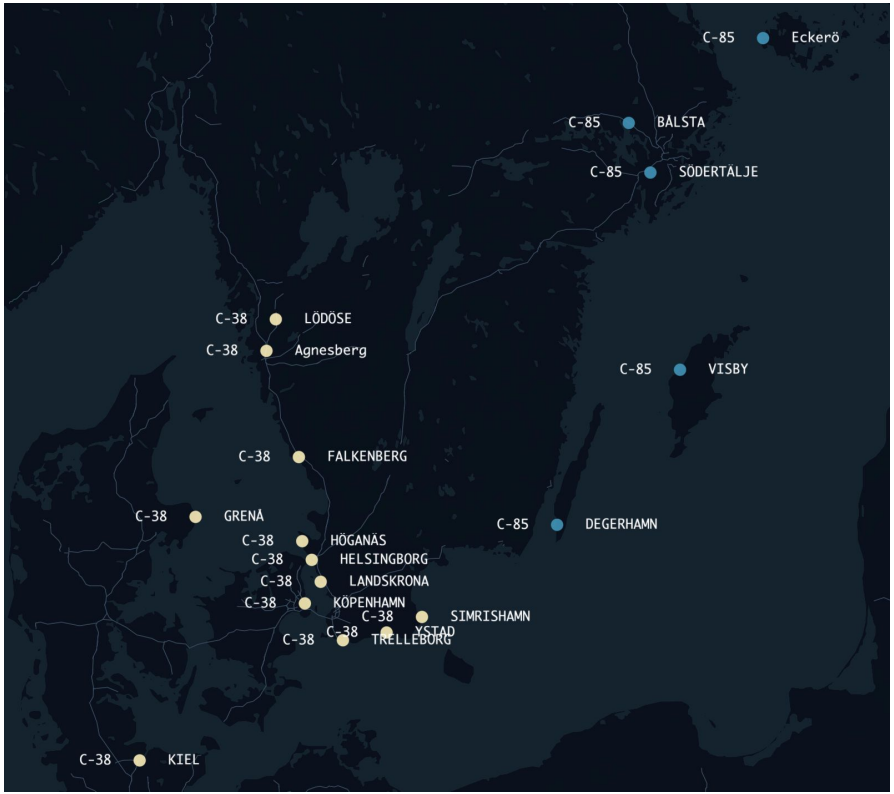


Country & Port
Community Detection

https://arxiv.org/pdf/1410.1237.pdf

The **Louvain algorithm** is used to detect country & port level communities in large shipping networks.

The clustering result can be used for **detecting dynamic changes** in the world shipping community that impact the global value chain at the granular port level.



GDP & GVC Shared
At Community Level

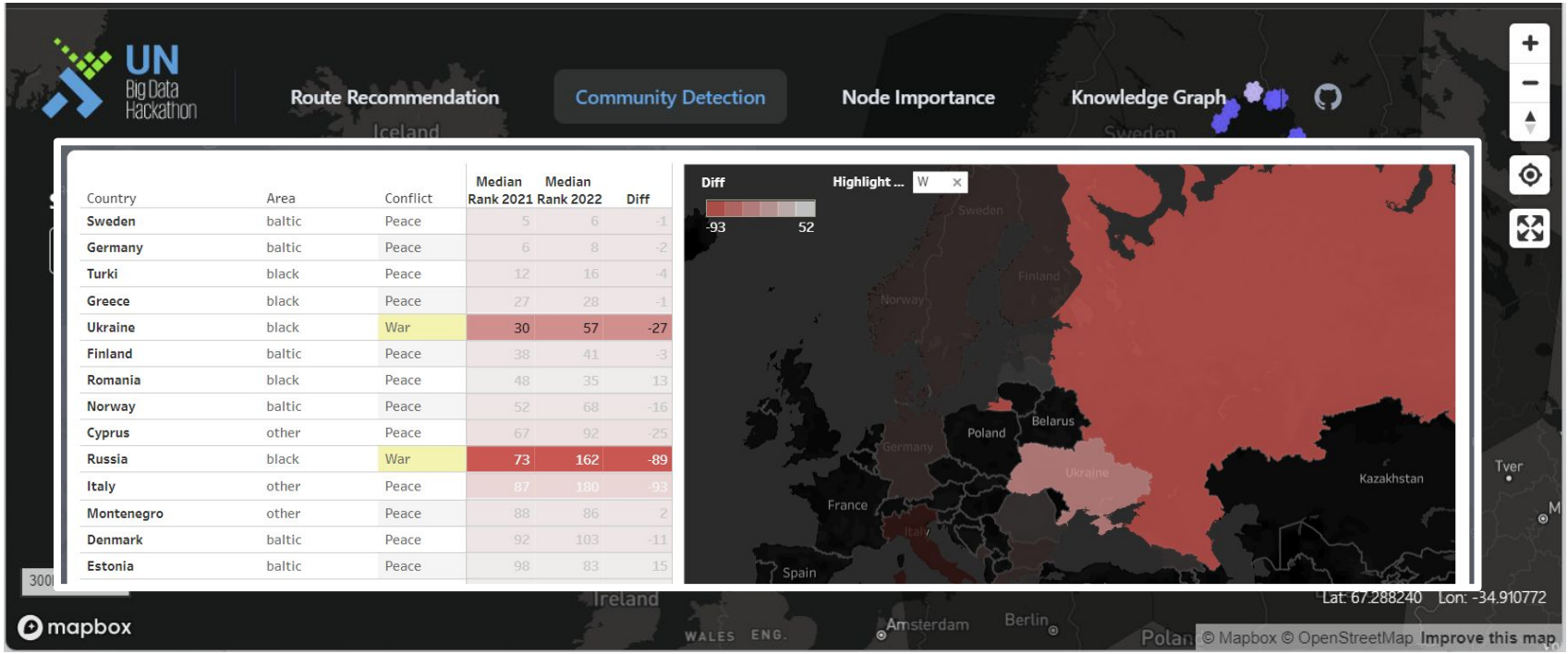
The Country clustering shows a very consistent result with the maritime zone of baltic sea, black sea, and other countries close to these areas.

Country Cluster	Total GDP 2022 Contribution (Million USD)	Total GVC 2018 Contribution
Baltic	**10,017,577.00**	**2,710,000,000.00**
Black	**1,463,271.00**	**165,000,000.00**
Other	**1,269,480.00**	**85,900,000.00**

Country	Cluster 2021	Cluster 2022
Sweden	BALTIC	BALTIC
Norway	BALTIC	BALTIC
Germany	BALTIC	BALTIC
Denmark	BALTIC	BALTIC
Finland	BALTIC	BALTIC
Russian Federation	BALTIC	BALTIC
Estonia	BALTIC	BALTIC
Latvia	BALTIC	BALTIC
Italy	OTHER	BALTIC
Ukraine	BLACK	BLACK
Turkey	BLACK	BLACK
Romania	BLACK	BLACK
Georgia	BLACK	BLACK
Bulgaria	BLACK	BLACK
Lebanon	OTHER	OTHER
Greece	OTHER	OTHER
Egypt	OTHER	OTHER
Israel	OTHER	OTHER
Cyprus	OTHER	OTHER
Montenegro	OTHER	OTHER

KnowAIS Platform

Country & Port Ranking Prototype



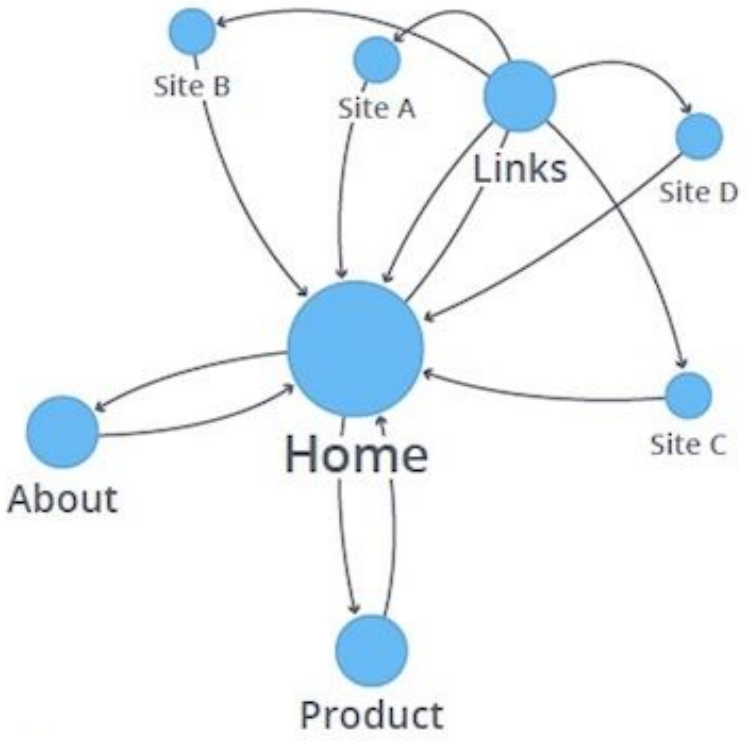
Country	Area	Conflict	Median Rank 2021	Median Rank 2022	Diff
Sweden	baltic	Peace	5	6	-1
Germany	baltic	Peace	6	8	-2
Turki	black	Peace	12	16	-4
Greece	black	Peace	27	28	-1
Ukraine	black	War	30	57	-27
Finland	baltic	Peace	38	41	-3
Romania	black	Peace	48	35	13
Norway	baltic	Peace	52	68	-16
Cyprus	other	Peace	67	92	-25
Russia	black	War	73	162	-89
Italy	other	Peace	87	180	-93
Montenegro	other	Peace	88	86	2
Denmark	baltic	Peace	92	103	-11
Estonia	baltic	Peace	98	83	15

Country & Port
Ranking

http://infolab.stanford.edu/~backrub/google.html

The **PageRank algorithm** can be used to measures the **importance** of each port or country within the graph, based on the number incoming relationships and the importance of the corresponding source port.

The underlying assumption roughly speaking is that a port is only as important as the ports that link to it.

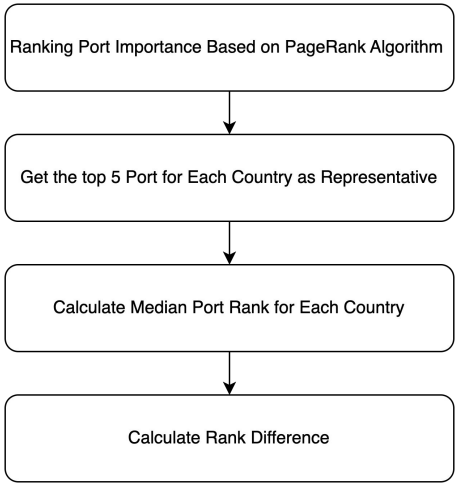


Visualization of PageRank

Median Port
Ranking Comparison

To validate the result of our proposed solution for port ranking, we compare the result of 2022 with the port ranking of 2021 as the median rank at the country level.

It shows that the median port rank of Ukraine and Russia decreased as a result of the conflict.



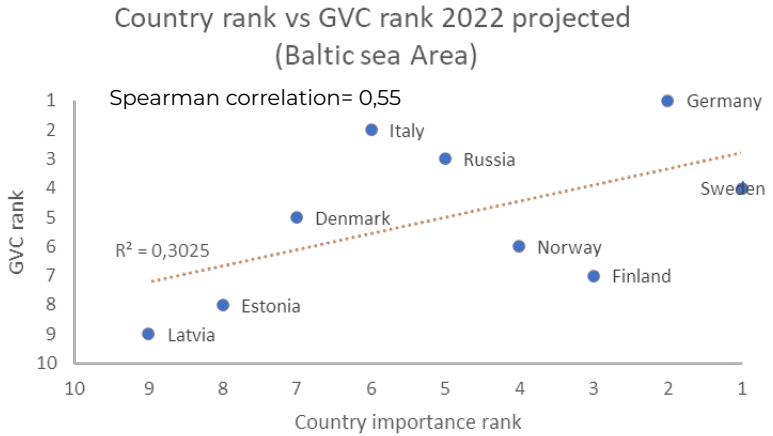
Area	Country	Median Rank 2021	Median Rank 2022	Diff
Baltic	Sweden	5	6	-1
Baltic	Germany	6	8	-2
Black	Turki	12	16	-4
Black	Greece	27	28	-1
Black	Ukraine	30	57	-27
Baltic	Finland	38	41	-3
Black	Romania	48	35	13
Baltic	Norway	52	68	-16
Other	Cyprus	67	92	-25
Baltic	Russia	73	162	-89
Other	Italy	87	180	-93
Other	Montenegro	88	86	2
Baltic	Denmark	92	103	-11
Baltic	Estonia	98	83	15
Other	Egypt	123	101	23
Other	Lebanon	130	110	20
Other	Israel	131	149	-18
Black	Georgia	156	145	11
Baltic	Latvia	167	115	52
Black	Bulgaria	184	202	-18

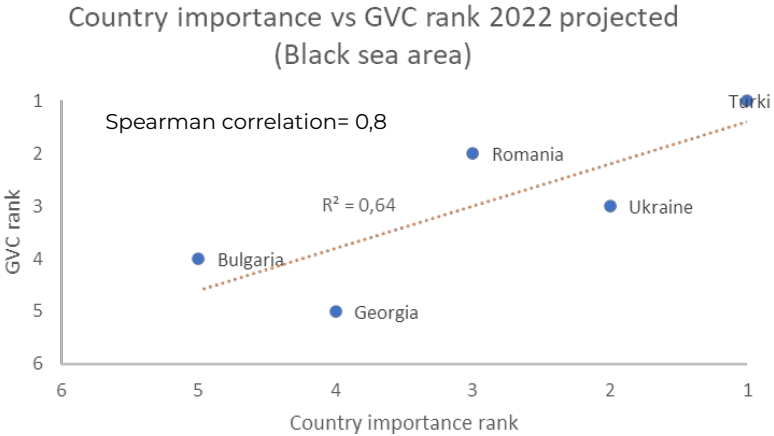
Proxy of **GVC Rank**

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Correlation between country rank and GVC

AIS Knowledge Graph can also generate the rank at the country level. We found that this ranking have strong rank-correlation with Global Value Chain (GVC) index with R = 0.8 at Black Sea area, while the correlation at the Baltic Sea can be estimated as moderate with R = 0.55.



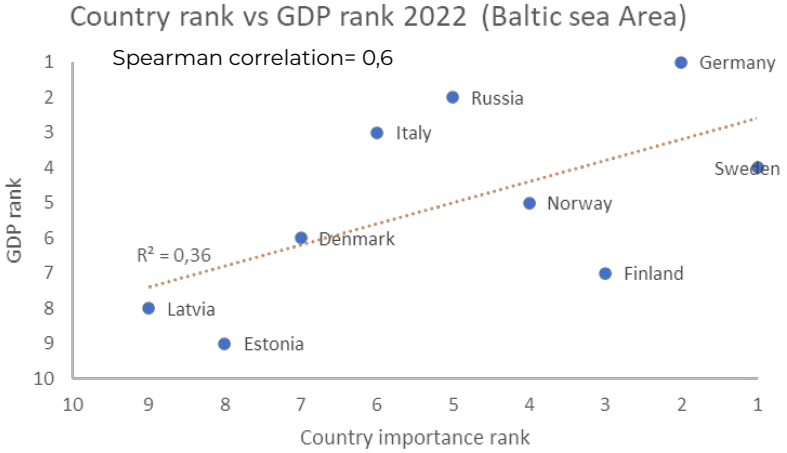


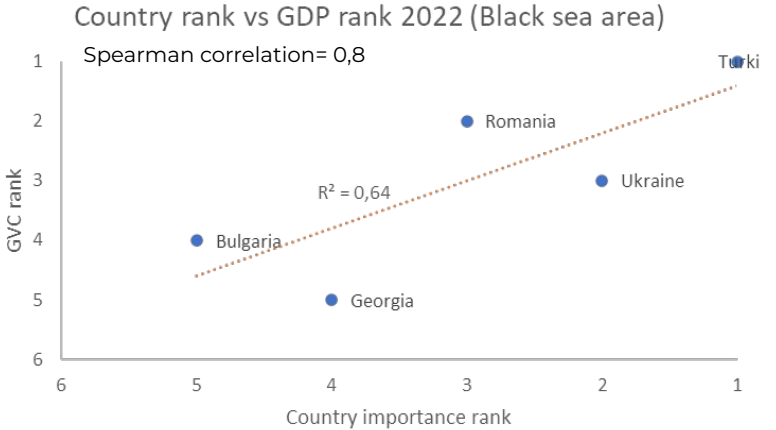
Proxy of **GDP Rank**

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Correlation between country rank and GDP

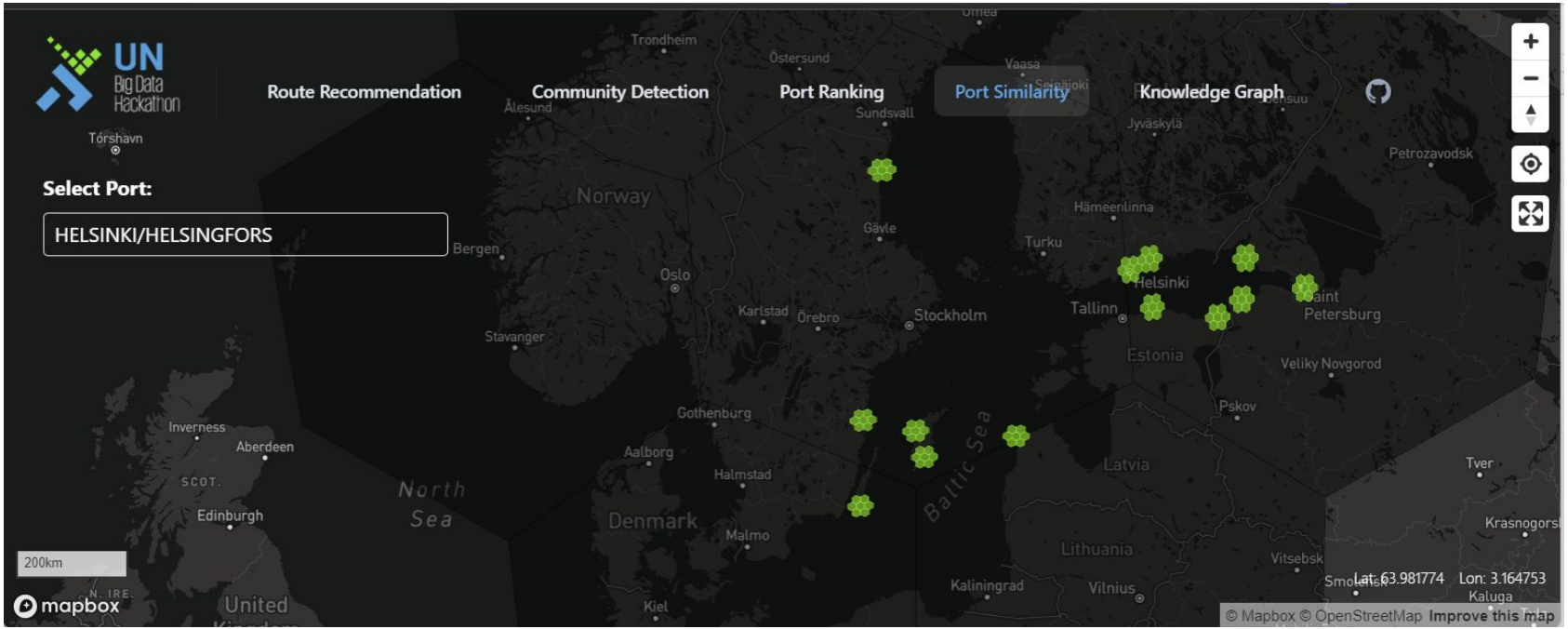
AIS Knowledge Graph can also generate the rank at the country level. We found that this ranking have strong rank-correlation with Gross Domestic Product (GDP) index with R = 0.8 at Black Sea area, while the correlation at the Baltic Sea can be estimated as moderate with R = 0.6.





KnowAIS Platform

Port Similarity Prototype



AIS2Vec
AIS Graph Embedding

https://arxiv.org/pdf/1908.11512.pdf

AIS2Vec is a graph embedding using **FastRP algorithm** to make a low-dimensional vector representations of every component (port, vessel, and country) in **KnowAIS**.

These vectors can be used for machine learning task such as:

- **Vessel type imputation** (supervised)
- **Port & Vessel similarity search** (unsupervised)

AIS2Vec
Port Similarity

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

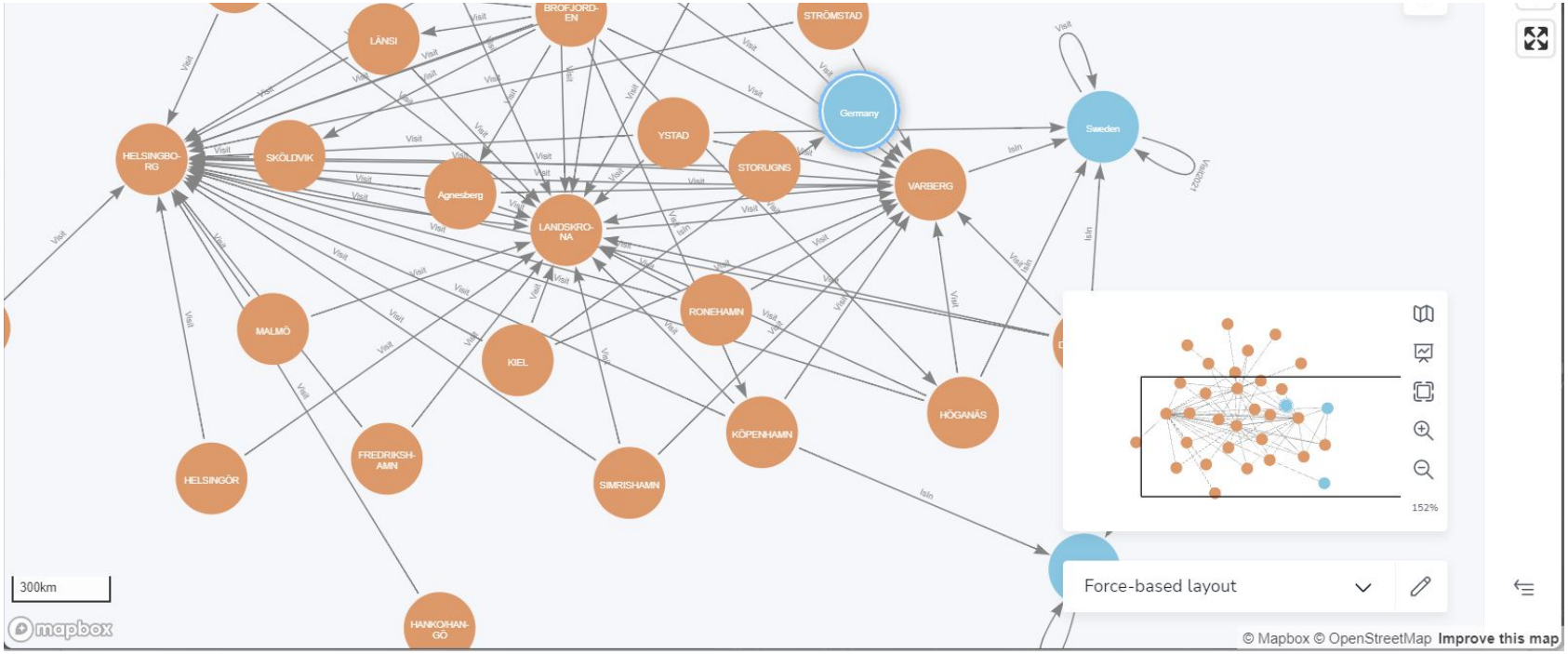
port	embedding
Reni	[-0.0002938761026598513, 0.27786871790885925, -0.004760327748954296, -0.03998576104640961, -0.0025746868923306465, -0.2427091747522354, 0.36845165491104126, 0.5021494030952454, 1.4050817489624023, -0.04565034806728363]
SIMRISHAMN	[-0.5459726452827454, -0.7790752649307251, -0.5941905975341797, -1.0064327716827393, 0.48302769660949707, 1.0797159671783447, 0.00927191786468029, -0.0756589025259018, -0.35650742053985596, -0.2661935687065124 5]
TROLLHÄTTAN	[0.07464912533760071, -0.1781981885433197, 0.9242082834243774, -0.2047748863697052, -0.0662752240896225, 0.9282465577125549, 0.5953547358512878, -0.7630221247673035, -0.30847659707069397, -0.9032402634620667]

Cosine Similarity is used to calculate port embedding similarity. The basic concept is very simple, it is to calculate the angle between two vectors.

port1	port2	sim_score
Bar	Zelenika	100.00%
HAMINA	KARLSBORG	100.00%
Basarabi	Mangalia	99.97%
HALMSTAD	VARBERG	99.97%
Abu Qir	Alexandria	99.96%
Agio Theodoroi	Idhra (Hydra)	99.96%
MUSSALO	LOVIISA	99.94%

KnowAIS Platform

Interactive Knowledge Graph Prototype



Limitation

1. Neo4j features for open source community edition.
2. AIS Data access at UNGP. Only able to use small sample size. It might impact the accuracy of OD matrix.
3. Availability data to enrich knowledge graph, along with the commitment of the availability
4. MapBox free private API token which only limits 50k hit/ months

Future Roadmap

1. Platform developments which enables more enhance decision support system.
2. Integrate with more available data, mainly from official statistics and other relevant data.
3. Improve the backend integration and study the best practice architecture that might sustain for a long time and easy to scale using cloud architecture.
4. Provide more broadened and advanced analytics which will enrich data-driven solution.
5. Focus as the measurement platform on controlling SDGs Goals

Appendix

Code Repository
https://github.com/datacerdas/know-ais
https://code.officialstatistics.org/datacerdas/know-ais

S3 Bucket Trajectory Result Repository
s3://know-ais/data/

KnowAIS Platform
https://know-ais.vercel.app/
https://know-ais.datacerdas.id/

KnowAIS Presentation
https://bit.ly/KnowAIS-Slide

Credit


UNBigData


STATISTICS INDONESIA


BRIN
BADAN RISET
DAN INOVASI NASIONAL


Data Cerdas
INDONESIA


UNIVERSITAS
INDONESIA
FAKULTAS
ILMU
KOMPUTER


aws


tokopedia


neo4j


GitHub


+ a b | e a u

Thank You