

The Alan Turing Institute

Data-Centric Engineering: New technologies to support efficient global shipping

Adam Sobey

Programme Director of Data-Centric Engineering, The Alan Turing Institute

Professor of Data-Centric Engineering, Maritime Engineering, University of Southampton

7th November 2024



Data-Centric Engineering

Data has always been central to shipping and Maritime

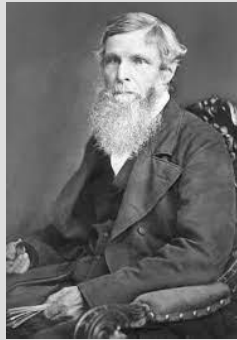
Vikings used empirically derived design guidance to ensure safety (apprentice copied master)



Classification since 1760



William Froude:
Ship Resistance



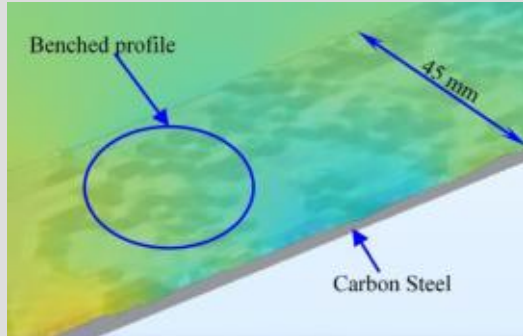
Accident Investigation



Structural surveying



Data-Centric Engineering



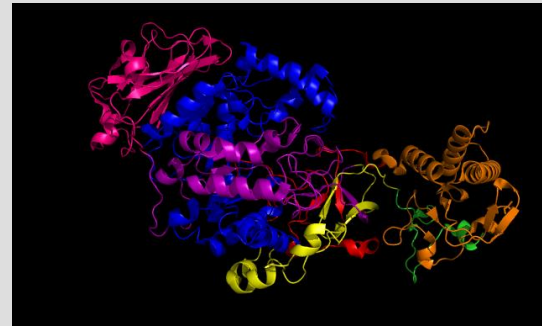
Compute infrastructure at global scale



Measurement and instrumentation via whole sensor networks



Data at multiple scales in time and space



AI enabling: New theories, technologies, markets, business models, education and skills

Fundamentals

Current

Uncertainty Quantification

Fundamentals of Digital Twins

Sampling

Bayesian Methods

Physics-based Machine Learning



Future?

Reinforcement Learning

Neuro-Symbolic Programming

Multi-agent Systems

Transfer Learning

LLM



Fundamentals

Current

Uncertainty Quantification

Fundamentals of Digital Twins

Sampling

Bayesian Methods

Physics-based Machine Learning



Future?

Reinforcement Learning

Neuro-Symbolic Programming

Multi-agent Systems

Transfer Learning

LLM



Fundamentals

Current

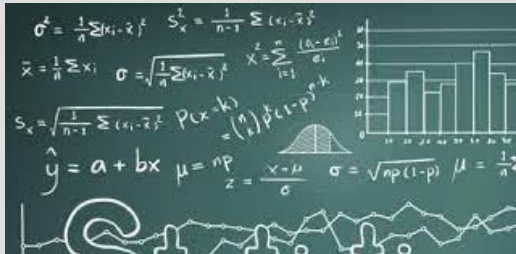
Uncertainty Quantification

Fundamentals of Digital Twins

Sampling

Bayesian Methods

Physics-based Machine Learning



Future?

Reinforcement Learning

Neuro-Symbolic Programming

Multi-agent Systems

Transfer Learning

LLM



Machine Learning: Power Prediction using Neural Networks



Motivation

- IMO goal to reduce total annual GHG emissions by at least 70% by 2040 compared to 2008
- Reduce total GHG emissions by 20% by 2030, and strive for 30% by 2030
- Shipping emissions have decreased by 18% since 2008
- Important to be able to monitor performance and make operational changes to reduce fuel consumption
- Long-term: reduce cost of Net-Zero fuels
- Short-term: Relatively cheap, very quick to implement

However,

- Modelling operational performance with conventional methods is **hard/impossible**.
- Can we use data?



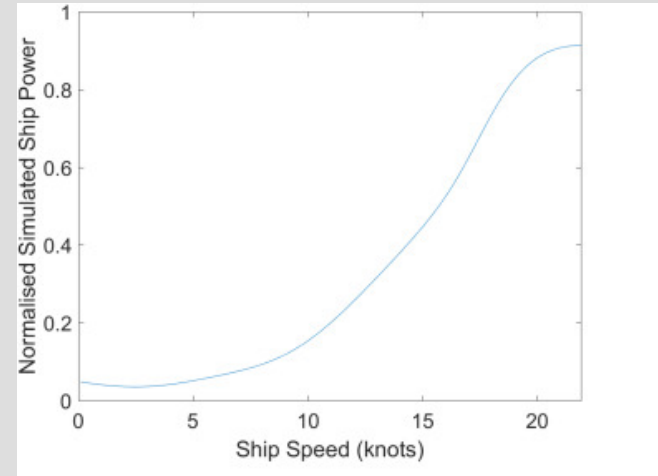
Ship power prediction is simple in calm waters



But challenging in weather conditions or with new energy efficiency approaches

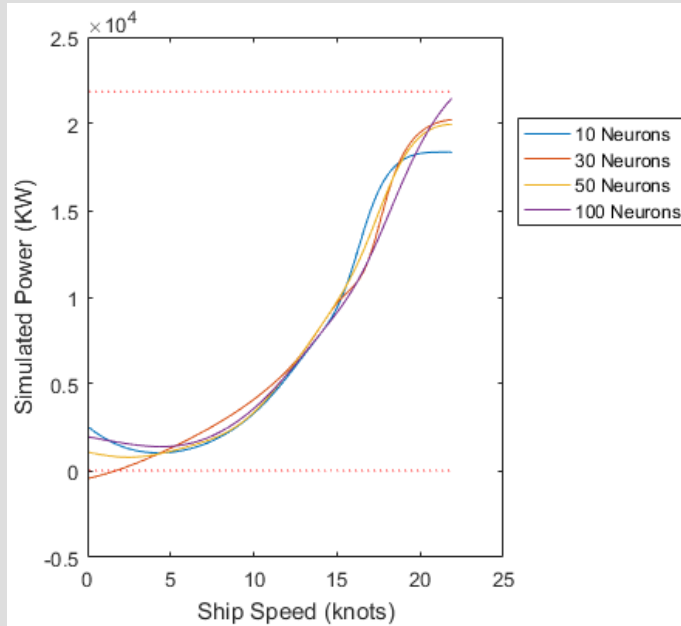
Prediction is easy!!!

- Test data from 27 months of continuous monitoring data are used from 3 vessels of the same design, sampled every 5 min.
- Used multiple Neural Network configurations to determine what is possible.
- **Error around 2%**
- Tried to explore how much data is required, around 18 months at 5mins intervals (more would reduce outliers).
- Number of papers in the literature now doing this.

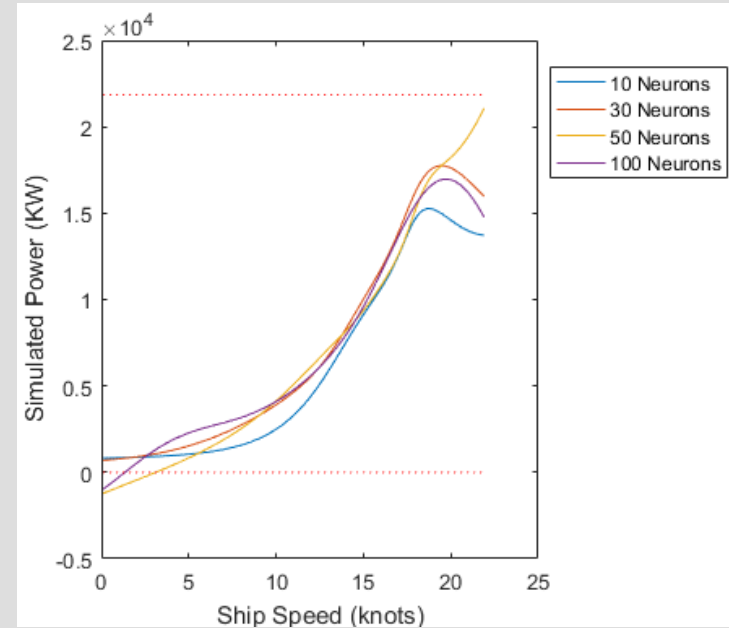


Speed power curve with 3 hidden layers and 50 neurons

Or is it?

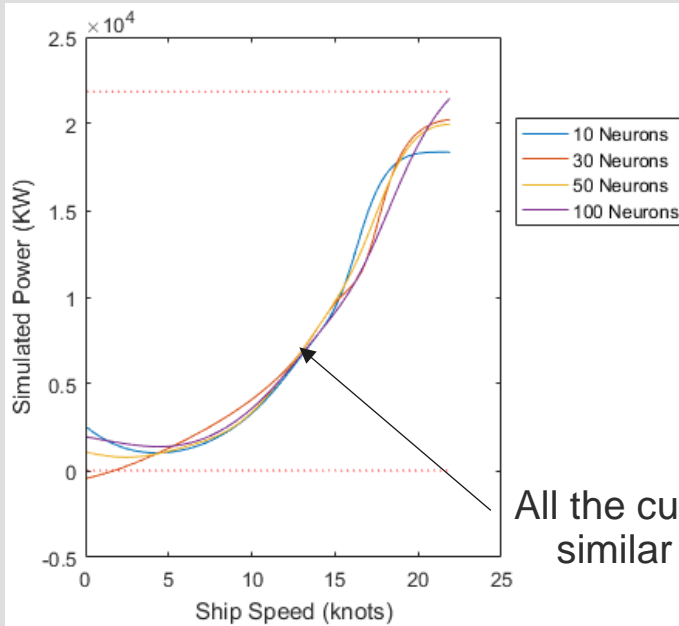


Simulations with 3 hidden layers and a range of neurons

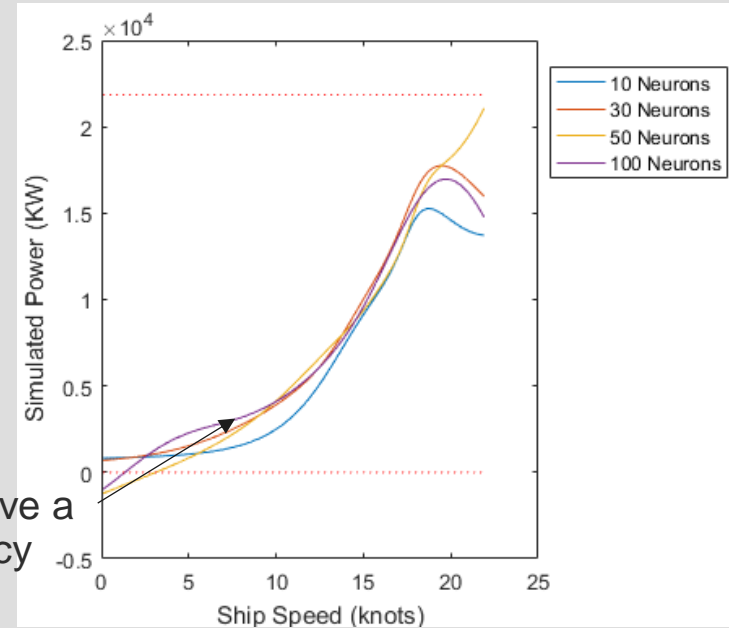


Simulations with 2 hidden layers and a range of neurons

Which curve is best?

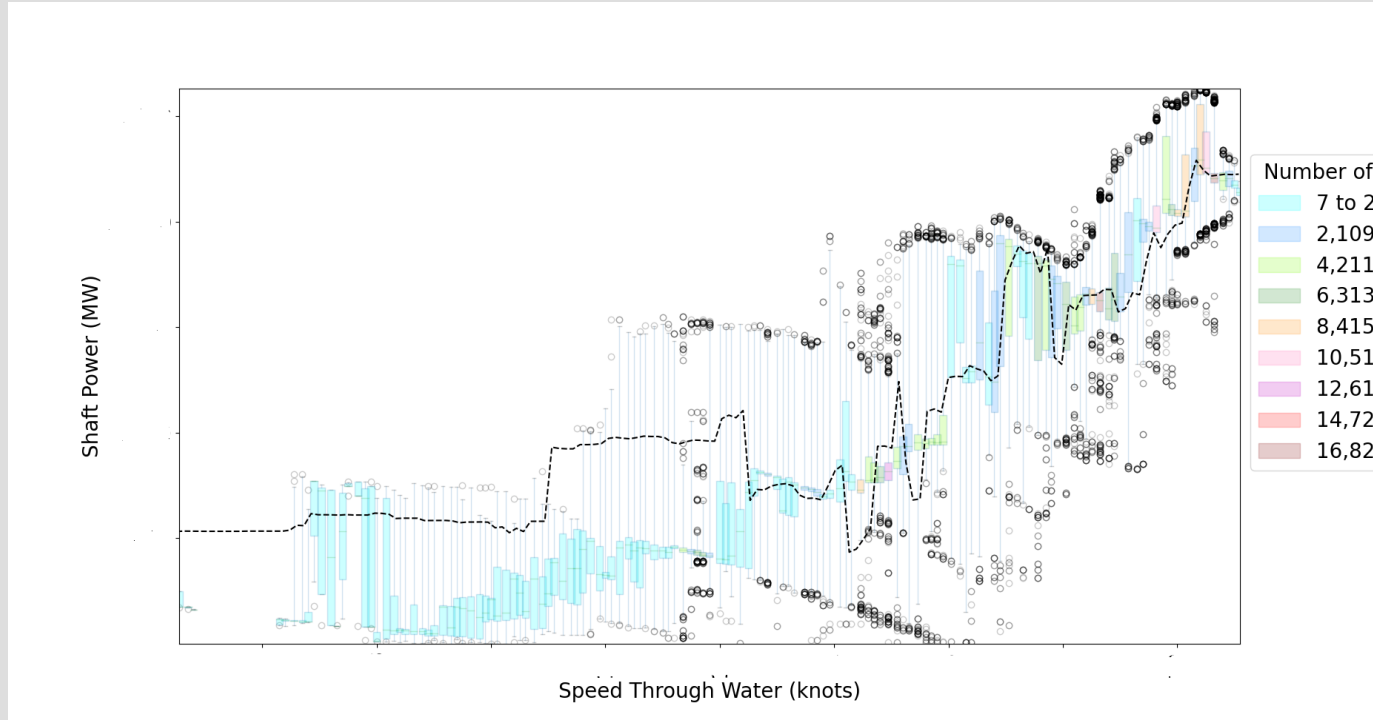


Simulations with 3 hidden layers and a range of neurons



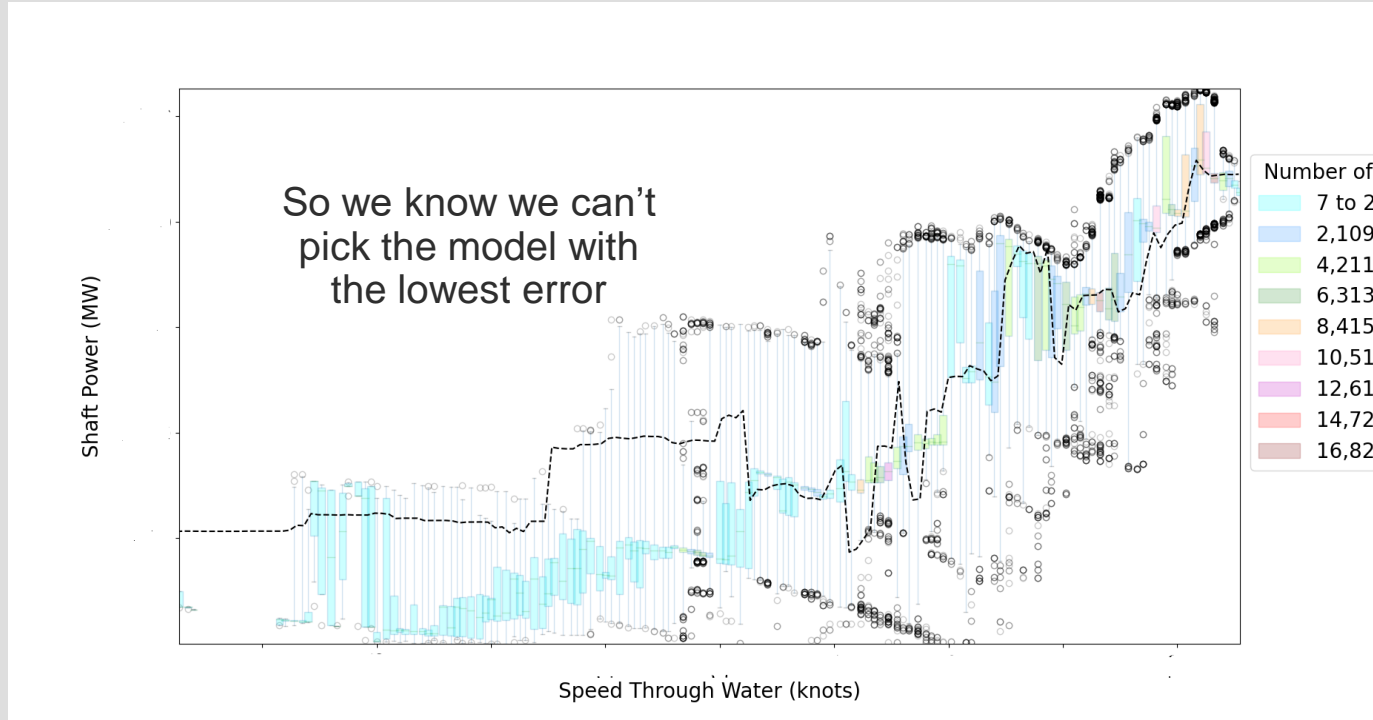
Simulations with 2 hidden layers and a range of neurons

Extreme Accuracy (XGBoost)



0.4% error, improved accuracy does not improve the input-output relationships

Extreme Accuracy (XGBoost)

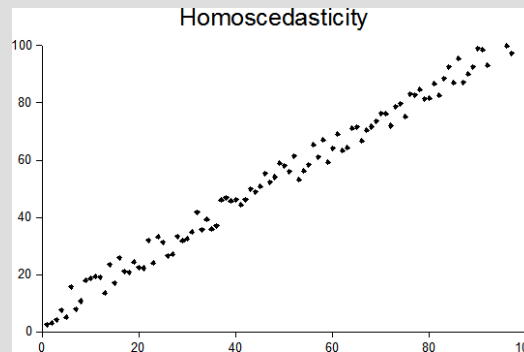
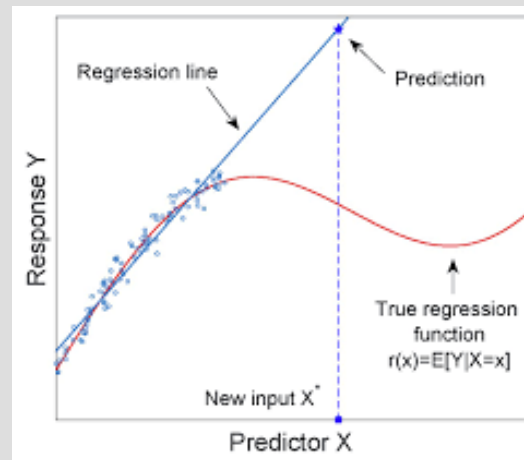


0.4% accuracy, improved accuracy does not improve the input-output relationships

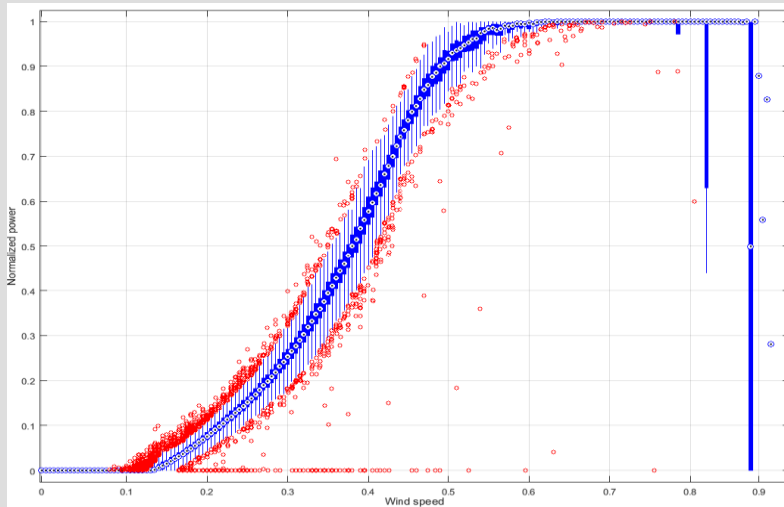
Some (well known) problems with ML

Minkowski-r family of pointwise measures only reflect the conditional average:

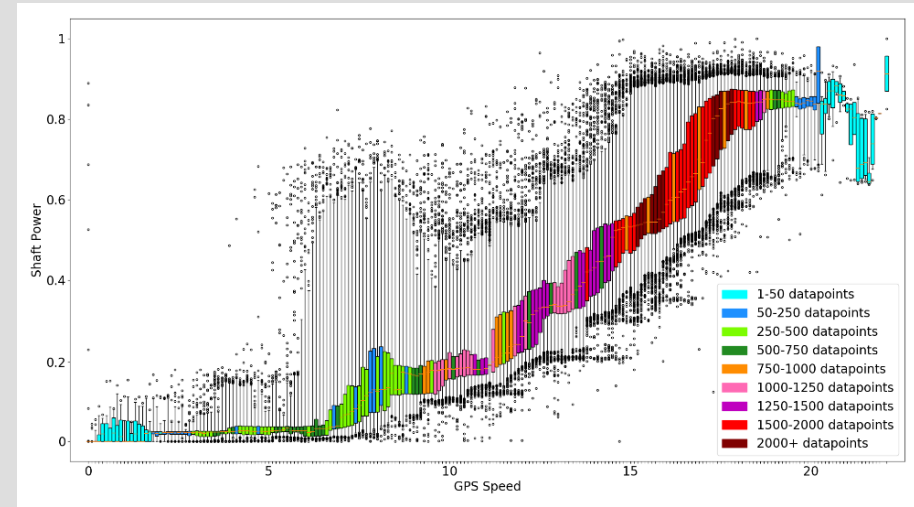
- the datapoints are independent;
- the distribution of the target variable is to be deterministic of the input with Gaussian noise,
- the standard deviation of noise, is not dependent on the input x (homoscedastic data);
- and the data set and neural network must be sufficiently large.



What does our data look like?



Wind Turbine Speed vs Power curve

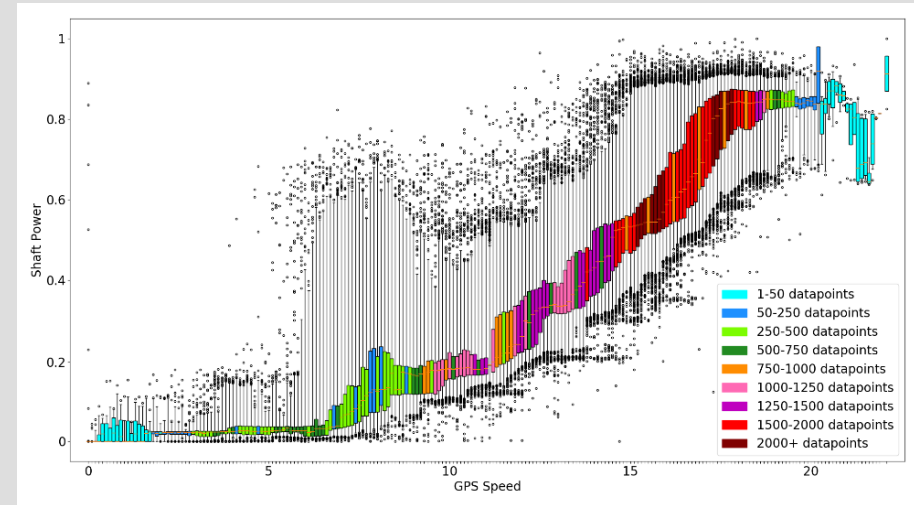


Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

- the datapoints are independent;



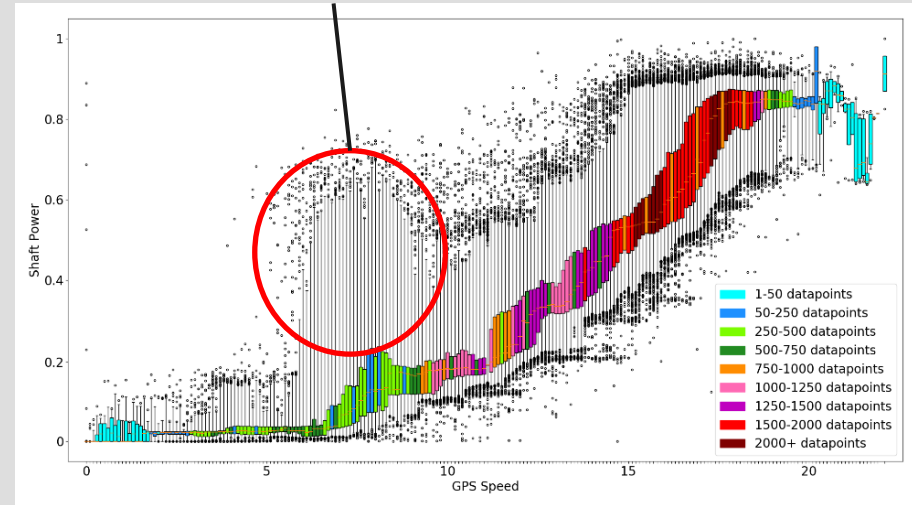
Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

- the datapoints are independent;

Grouping of bad weather points

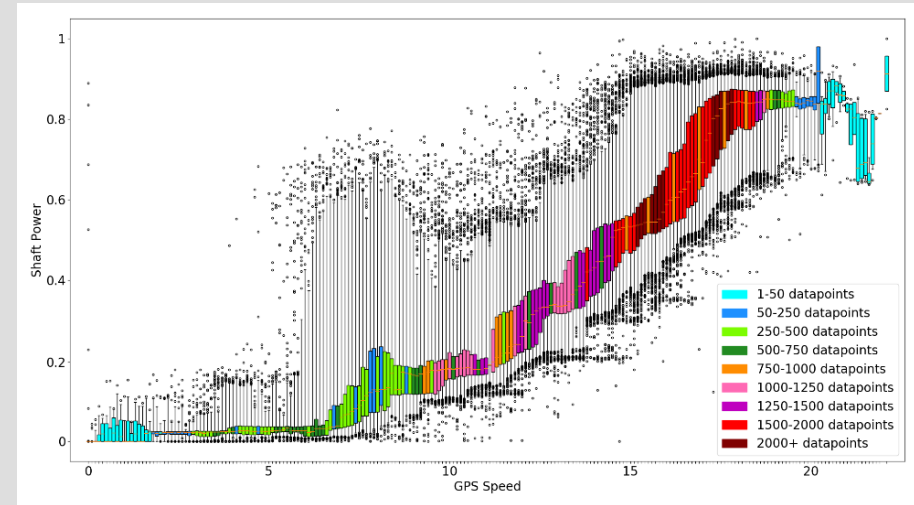


Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

- ~~the datapoints are independent;~~
- the distribution of the target variable is to be deterministic of the input with Gaussian noise,



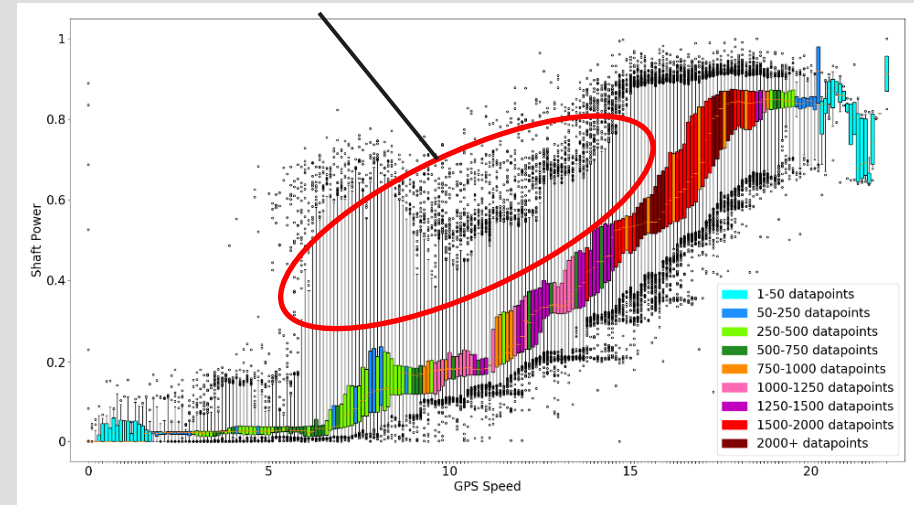
Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

- the datapoints are independent;
- the distribution of the target variable is to be deterministic of the input with Gaussian noise,

Bias of data to high weather

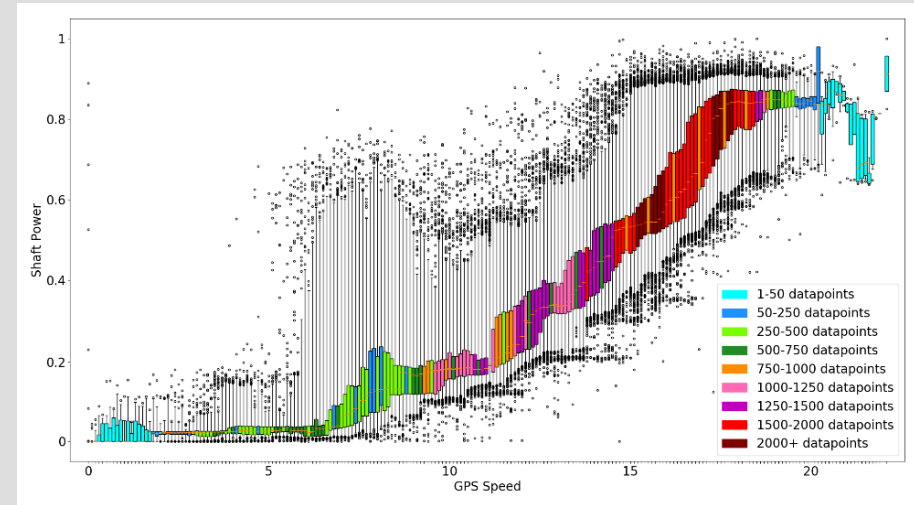


Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

- the datapoints are independent;
- the distribution of the target variable is to be deterministic of the input with Gaussian noise;
- the standard deviation of noise, is not dependent on the input x (homoscedastic data);



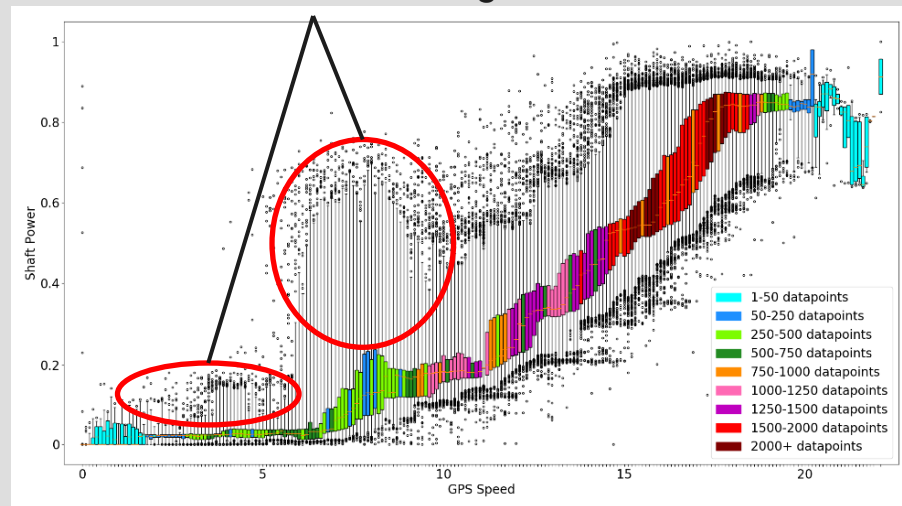
Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

- the datapoints are independent;
- the distribution of the target variable is to be deterministic of the input with Gaussian noise,
- the standard deviation of noise, is not dependent on the input x (homoscedastic data);

Considerable variation through the set

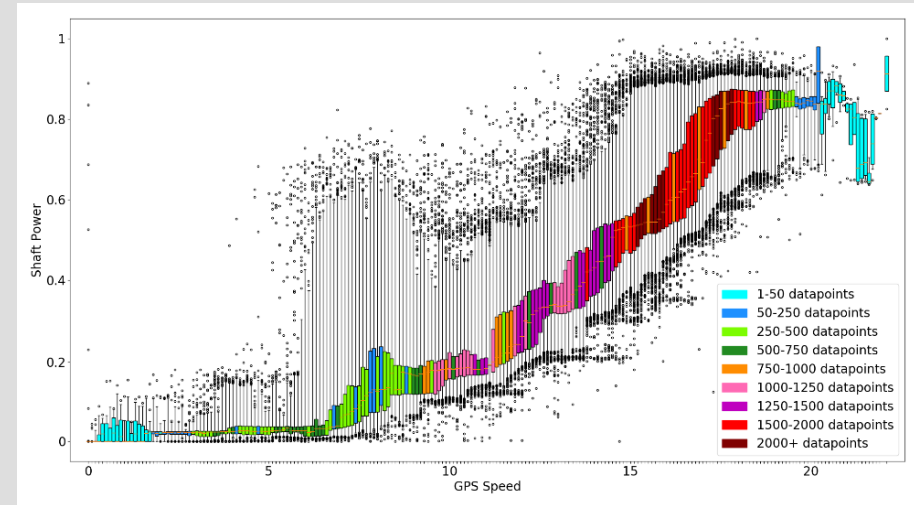


Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

- the datapoints are independent;
- the distribution of the target variable is to be deterministic of the input with Gaussian noise;
- the standard deviation of noise, is not dependent on the input x (homoscedastic data);
- and the data set and neural network must be sufficiently large.

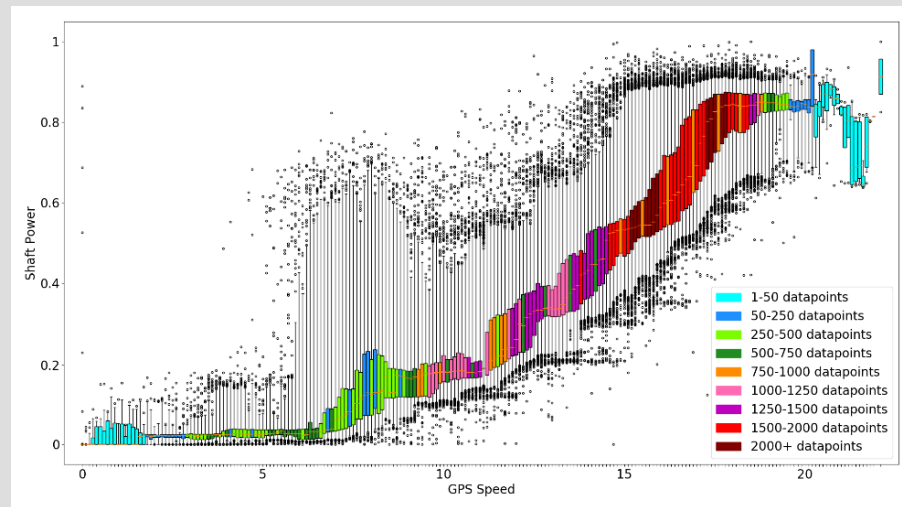


Ship Speed vs Power curve

What does our data look like?

Minkowski-r family of pointwise measures only reflect the conditional average:

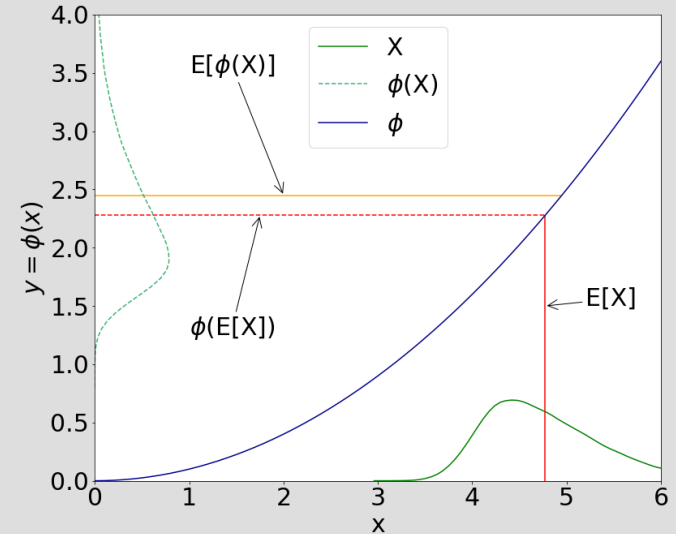
- the datapoints are independent;
- the distribution of the target variable is to be deterministic of the input with Gaussian noise;
- the standard deviation of noise, is not dependent on the input x (homoscedastic data);
- and the data set and neural network must be sufficiently large. **MAYBE...**



Ship Speed vs Power curve

Mean Fit to Median (Jensen's Inequality)

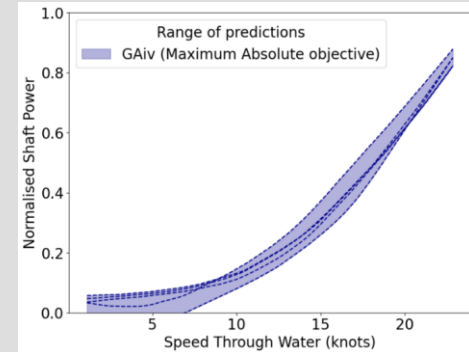
- Use Genetic Algorithm to select trained networks
- No longer constrained in our error measures
- Utilise a continuous measure of the median, no longer using pointwise Minkowski family measures



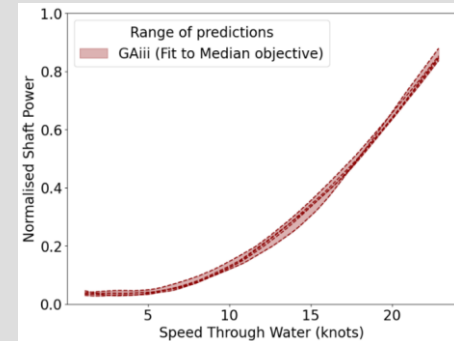
For data not normally distributed: the mean of the transformed observation is not the same as the transform of the mean observation

Mean Fit to Median

- Use Genetic Algorithm to select trained networks
- No longer constrained in our error measures
- Utilise a continuous measure of the median, no longer using pointwise Minkowski family measures
- New measure gives a lower spread across 5 runs, with no crossing
- Standard ML is not representational of the input-output relationships, can't optimize



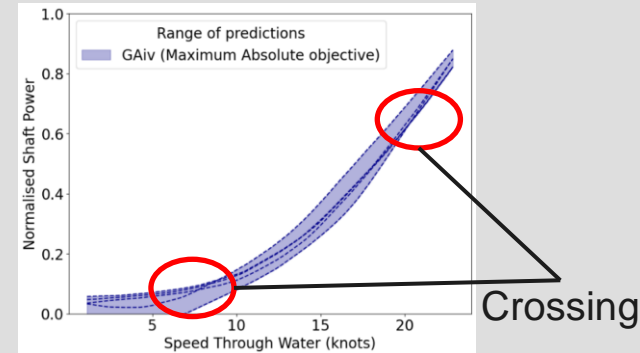
Standard ML: High variation in physical representation



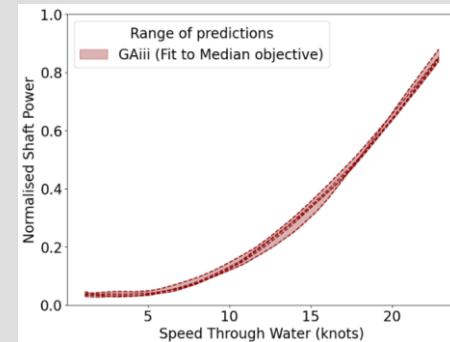
New measure: less variation in physical representation

Mean Fit to Median

- Use Genetic Algorithm to select trained networks
- No longer constrained in our error measures
- Utilise a continuous measure of the median, no longer using pointwise Minkowski family measures
- New measure gives a lower spread across 5 runs, with no crossing
- Standard ML is not representational of the input-output relationships, can't optimize



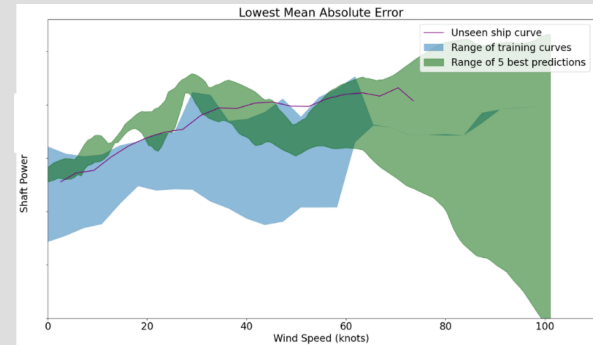
Standard ML: High variation in physical representation



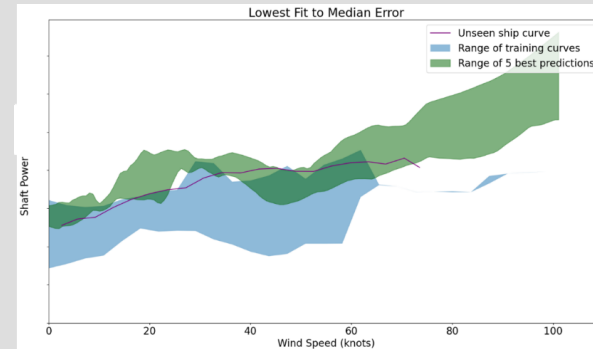
New measure: less variation in physical representation

Ship Power Prediction

- Use Genetic Algorithm to select trained networks
- No longer constrained in our error measures
- Utilise Median rather than mean, to avoid transformation as our data isn't normally distributed
- New measure gives a lower spread across 5 runs, with no crossing
- Standard ML is not representational of the input-output relationships, can't optimize
- The new measure allows better generalization off dataset



Standard ML: Extrapolation for wind speeds over 60 knots exhibits high variations



New Measure: Extrapolated relationships (green) are more consistent and more accurate

JAWS

- Operating on 64 vessels

Benefits



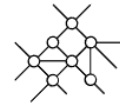
Reduced fuel cost

By optimizing trim and draft, savings in the range of 3-7% can be achieved



Reduced emissions

Fuel savings directly result in CO2 reductions



Low implementation cost

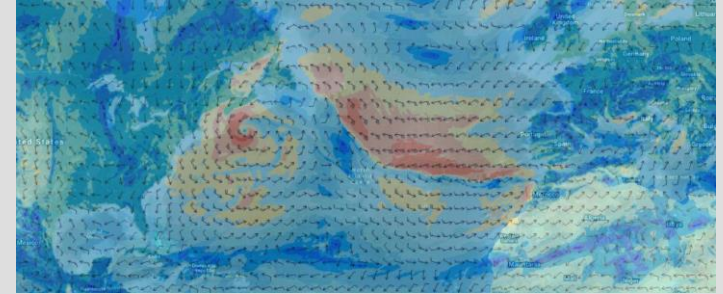
Only K-IMS, a maritime industry platform and data logger, is required for JAWS

Artificial Intelligence: Avoiding Whale Sharks using an Epigenetic Algorithm



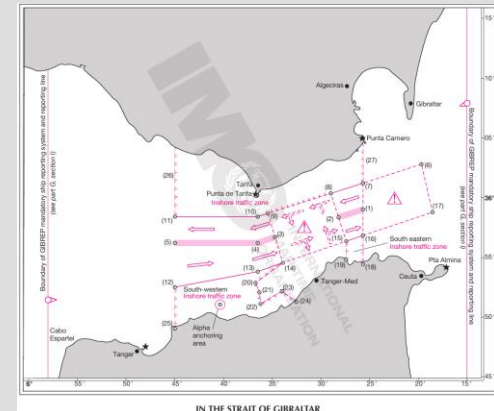
Problem Complexity

- Higher fidelity data
- Weather conditions (constantly changing)
- Navigational restrictions (no-go / preferred areas)
- Engine models (different “preferred” conditions)
- “Crowded” waters
- Increasing number of objectives (only going up)
- Current Software shows a 5% fuel saving, 7% earlier arrival, 8% TCE – Compared to other software



Top: ever evolving weather

Bottom: Traffic Separation Schemes



Evolutionary Computation



Artificial Intelligence



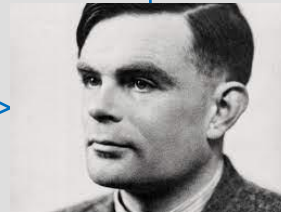
Finance Predictions



Engineering Design



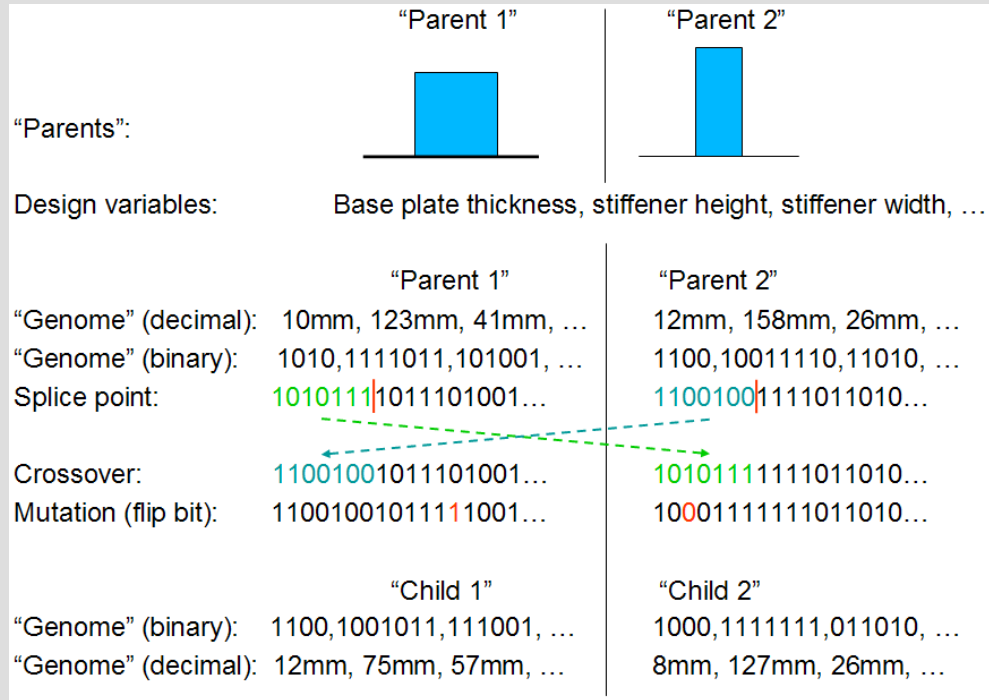
Flight Control



Finding Cancer

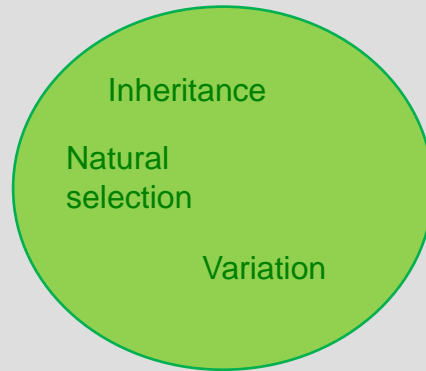


Evolutionary Computation (1970s)

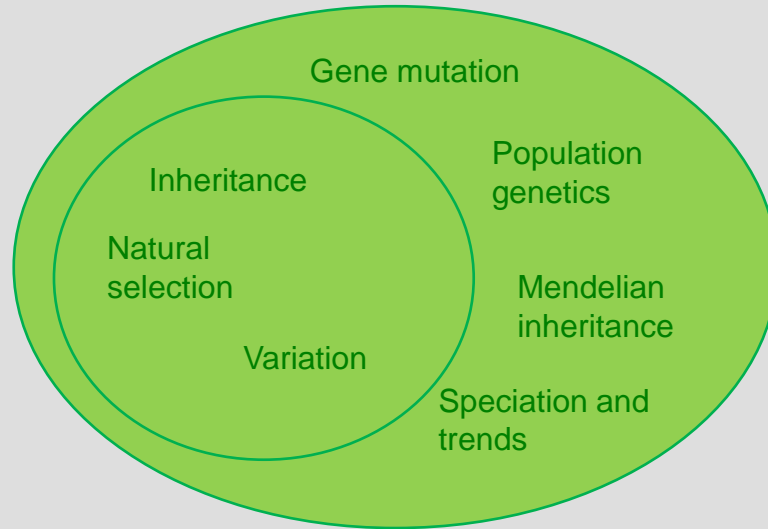


Evolutionary Computation Quick Review

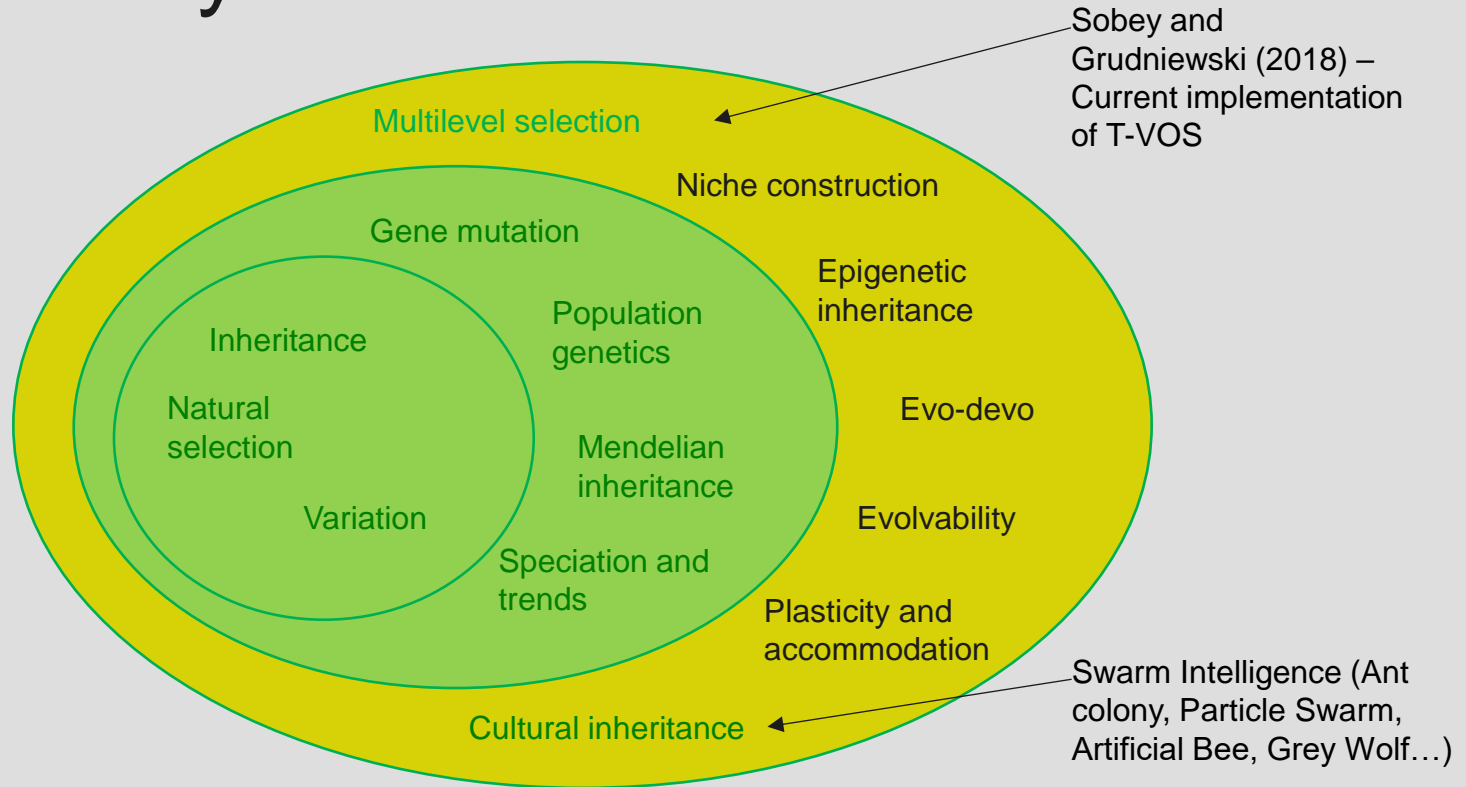
Darwinism



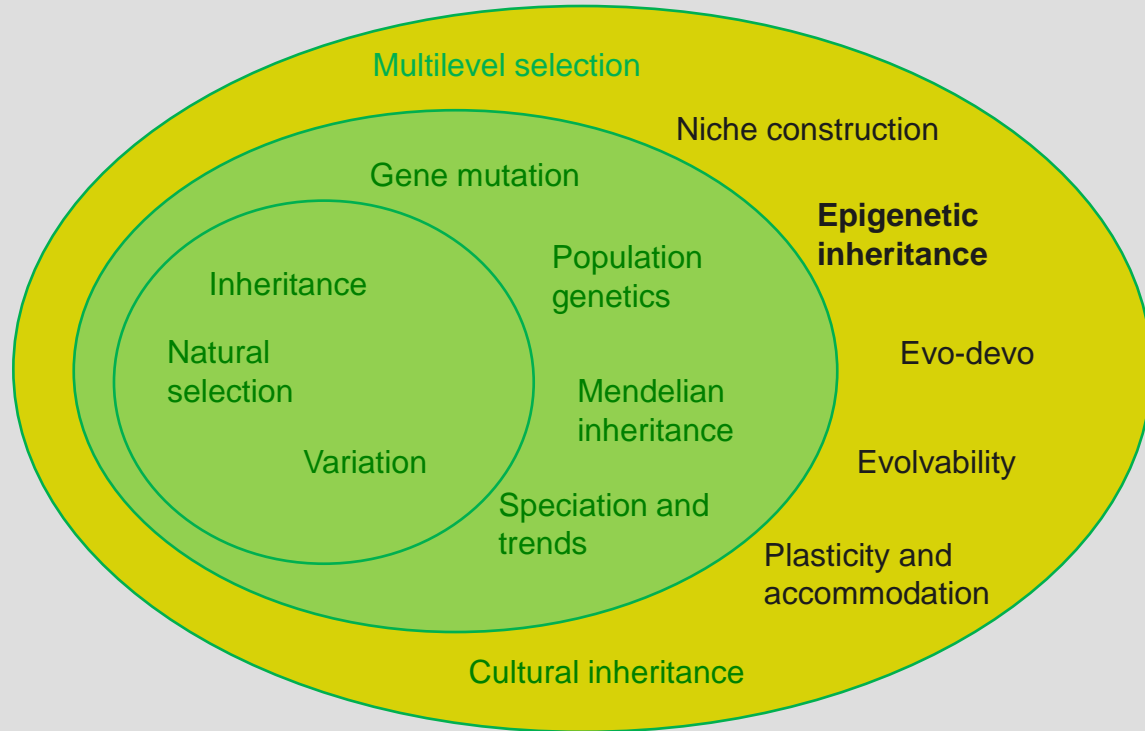
Evolutionary Computation Quick Review



Extended Synthesis

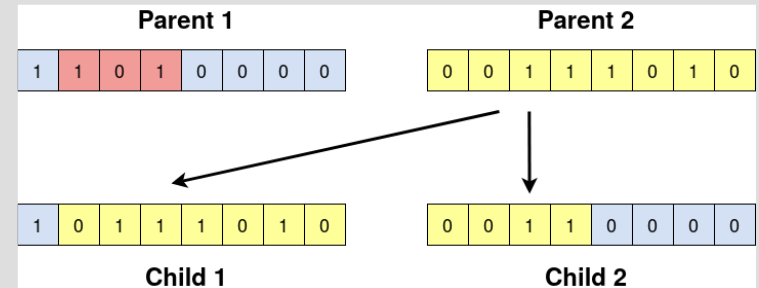


Epigenetics



Epigenetic Algorithm

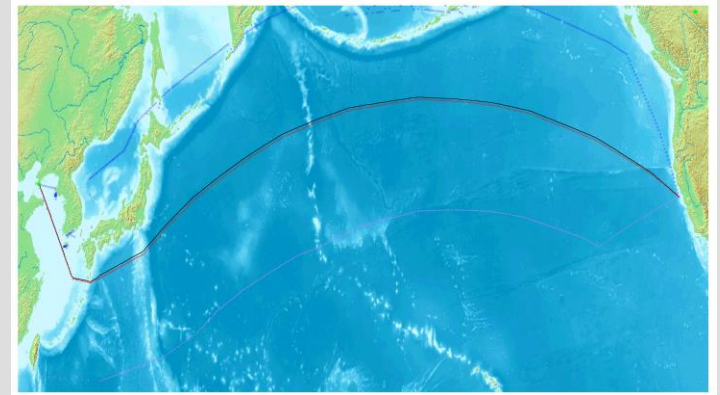
- Incorporate "blocking" mechanisms that activate or deactivate genes without altering the underlying DNA sequence.
- The mechanism has a probability to trigger during the reproduction stage for every parent.
- Blocking fitter parents maintains diversity while blocking less fit parents increases the convergence of the population.
- The behaviour of the genetic algorithm with epigenetics mirrors biological results of epigenetic effects in nature.



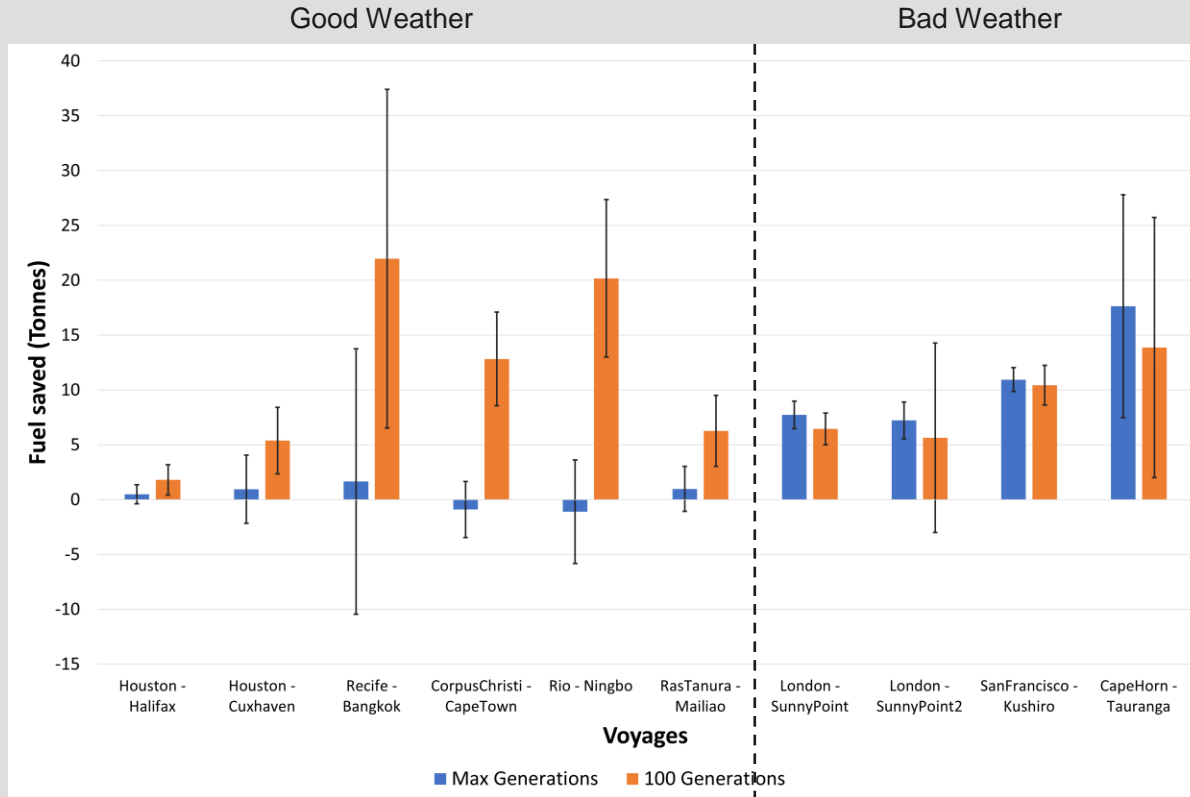
A simplified view of the blocking mechanism where, after crossover, variables chosen to be blocked in the offspring are set to the unblocked parent's variables

Test scenarios

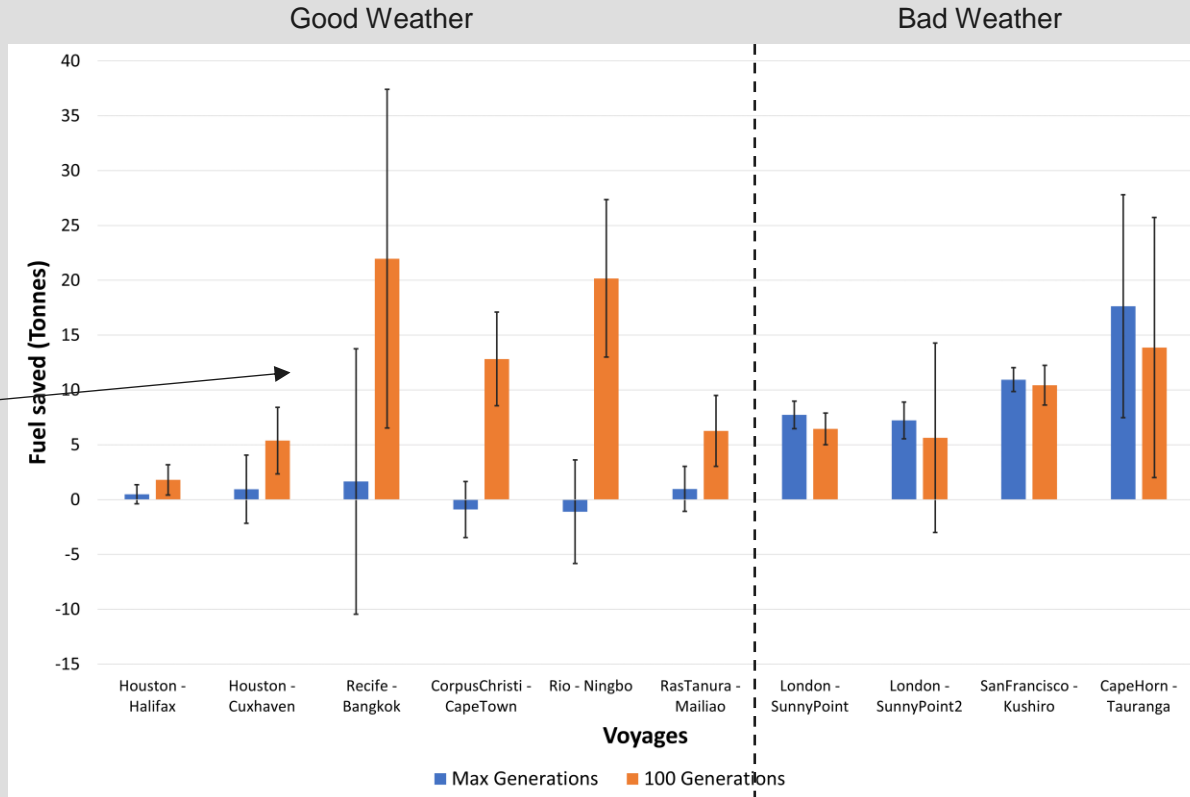
- A total of 10 voyages are tested. 6 have calm weather conditions and 4 have more severe weather conditions.
- Rerouting may occur depending on the weather conditions during the voyage, requiring dynamic changes to the route.
- Met-ocean data is used to provide the weather information.
- The voyages may also have constraints to abide by safety regulations, ship traffic schemes, and to prevent grounding.
- Each voyage and algorithm configuration were benchmarked over 7 runs.
- A total of 664 generations are generated in each run.



Comparison to the state-of-the-art

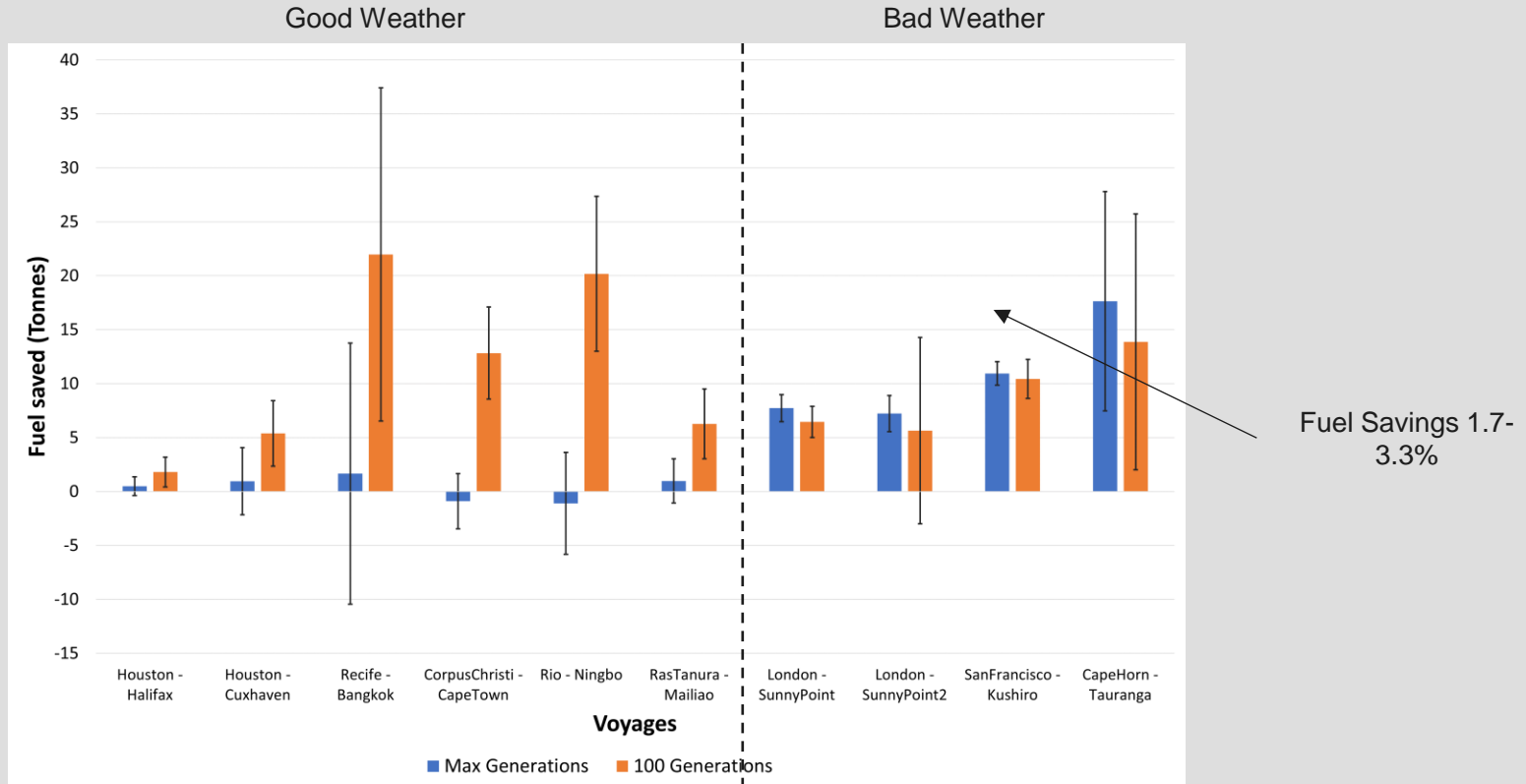


Comparison to the state-of-the-art



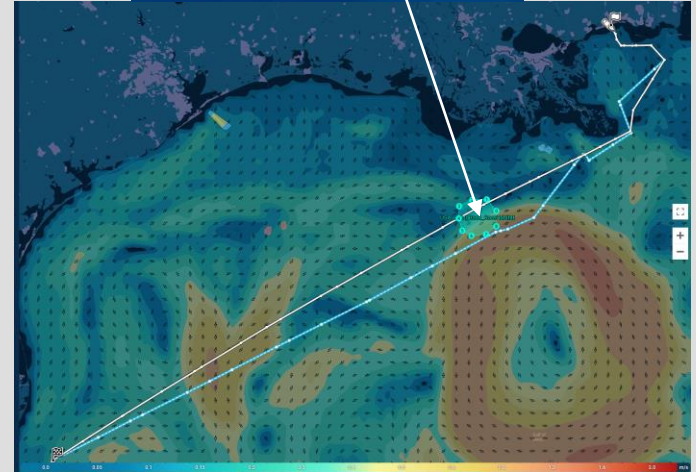
Similar results, but fewer functions calls

Comparison to the state-of-the-art



Avoiding Whale Sharks

- Total of 8 routes were selected
- Crude oil tanker was selected with approx. length of 336m and DWT of 300'000 tonnes.
- The allowed speed range was set to 11-14kts.
- In the speed reduction zone the vessel's speed was limited to 10kts.
- Investigating same arrival time
- When treating the whale shark habitat as a no go zone, the fuel loss ranges 0.12-7.21%. Higher percentages for 1 and 3 days voyages.
- When treating the habitat as a speed reduction zone, the maximum additional fuel was 0.8%.

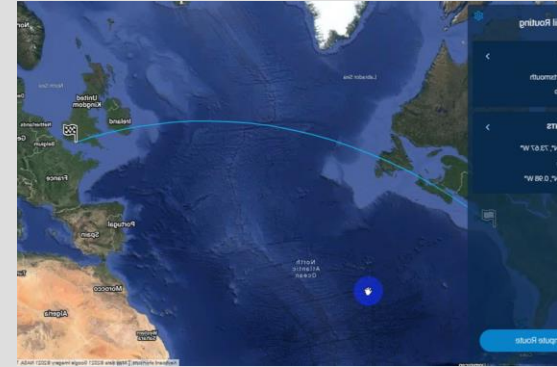


The original route (in pink) in comparison to the route avoiding the whale shark habitat (in blue).

Theyr

T-VOS

- Clients include Euronav, Spire and LR
- Epigenetic algorithm demonstrates 1.7-3.3% more in bad weather and reduced computational time
- 2,151 vessels, potential for 30,000
- Combined with models developed in the previous section, **18% emissions reductions**
- Can help us defend animal habitats



T-VOS: Voyage Optimisation Software

NOMARS

- 55m vessel developed for DARPA
- Aim is to operate for 6 months with no human interaction
- Consortium led by Serco
- Theyr providing strategic planner

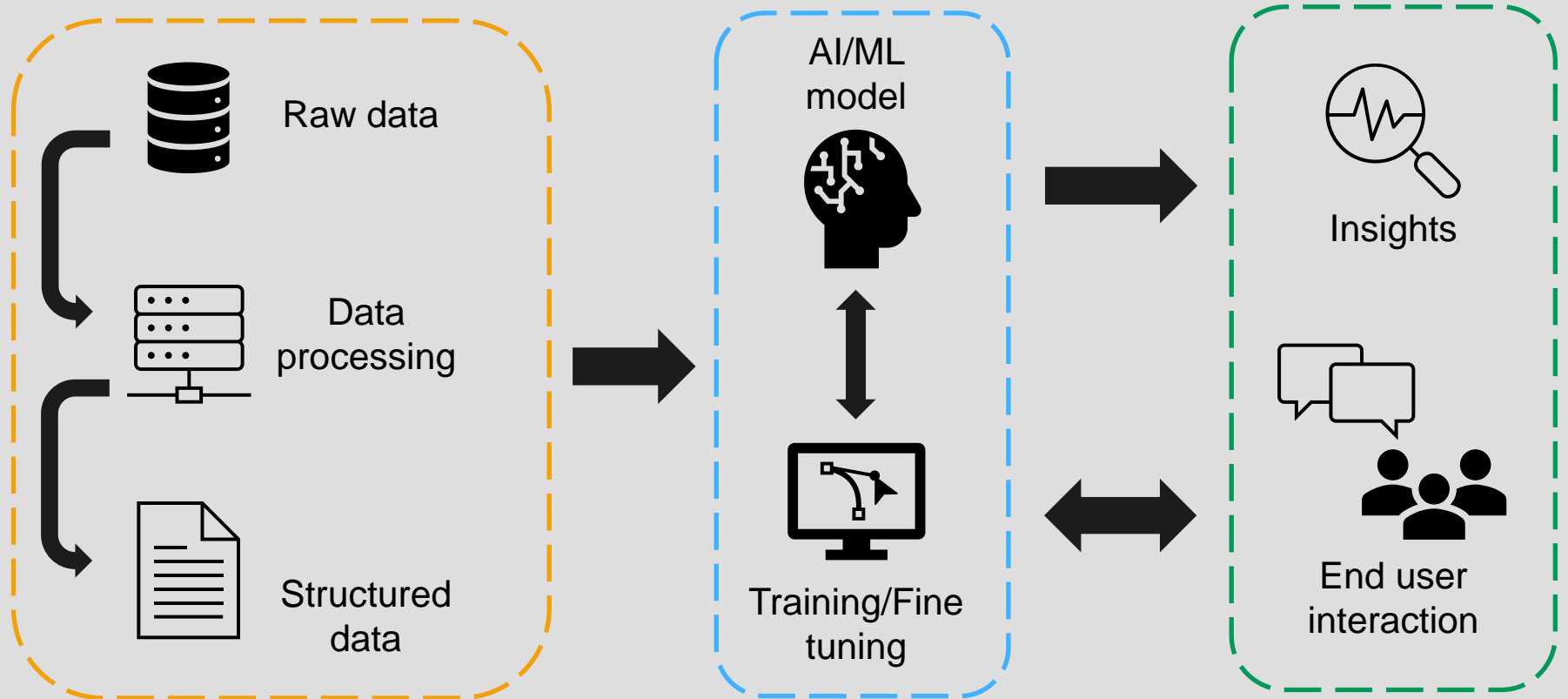


Defiant Medium Unmanned Surface Vehicle Concept

DCE at Scale: An AI Data Engineer



The AI Data Pipeline

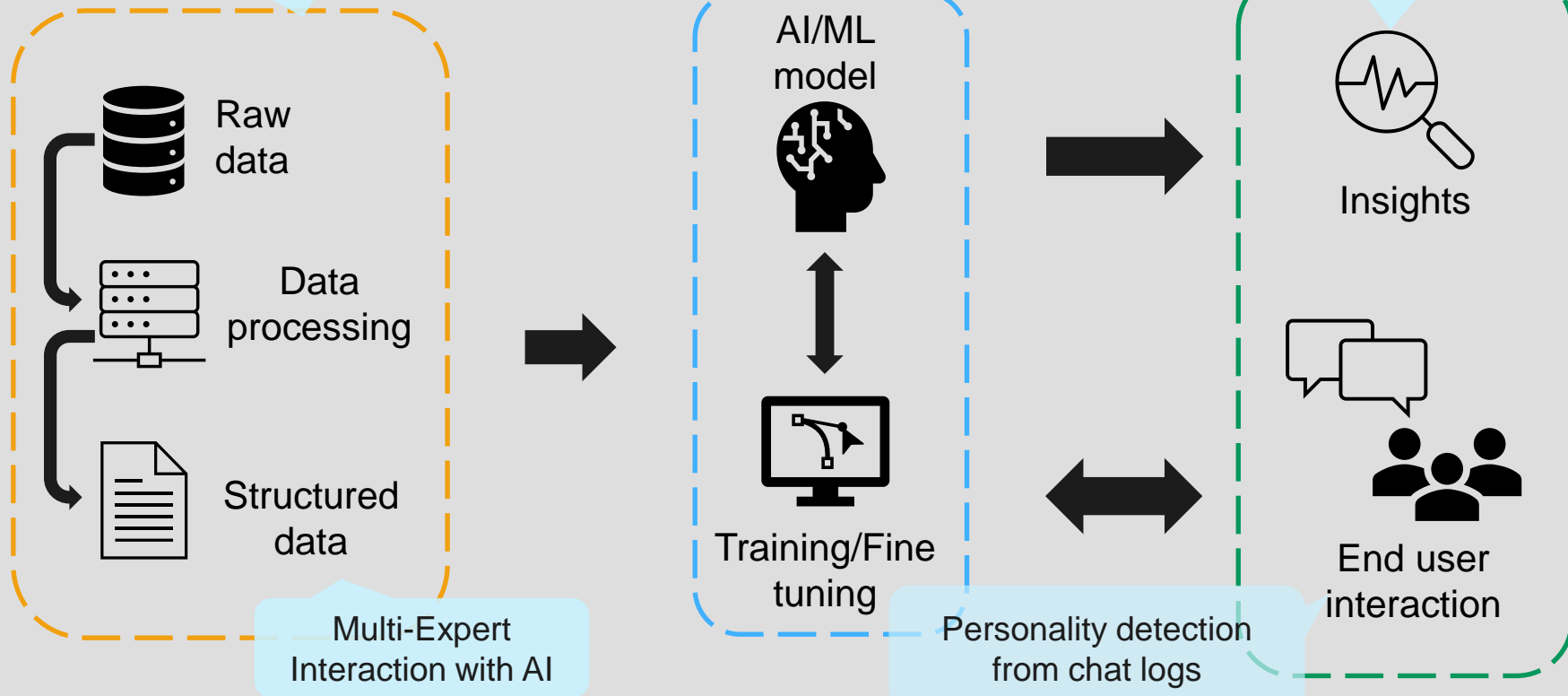


The AI Data Pipeline

Automated data pipeline generation

Code generation for training and evaluation on feasibility tests.

Automatic dataset generation for questions and answering tasks..



Automatic data pipeline creation

Scalable Real-Time Data Pipeline for Machine Learning

M create a templates to describe the required AWS resources and their properties, in AWS cloud formation format.

Certainly, I'll create an AWS CloudFormation template that describes the required AWS resources and their properties for our data pipeline. This template will be in YAML format, which is one of the supported formats for CloudFormation.

</> **AWS CloudFormation Template for Data Pipeline**
Click to open code

This CloudFormation template provides a basic structure for our data pipeline. Here's a brief explanation of the resources defined:

1. **DataIngestionStream**: A Kinesis Data Stream for ingesting data from various sources.
2. **RawDataBucket**: An S3 bucket for storing raw data.
3. **StreamProcessingApplication**: A Kinesis Data Analytics application for real-time data processing.
4. **ProcessedDataFileSystem**: An EFS file system for storing processed data.
5. **MetadataTable**: A DynamoDB table for storing metadata and enabling quick lookups.
6. **MLNotebookInstance**: A SageMaker Notebook instance for machine learning tasks.
7. **IAM Roles**: Necessary roles for Kinesis Analytics and SageMaker.
8. **MonitoringDashboard**: A CloudWatch dashboard for monitoring the pipeline.

This template provides a starting point and would need to be customized based on specific requirements, such as:

1. Adjusting the capacities and configurations of resources (e.g., Kinesis shard count, EC2 instance types).
2. Adding more detailed IAM permissions.
3. Configuring network settings (VPCs, security groups, etc.).

← AWS CloudFormation Template for Data Pipeline

```
AWS::CloudFormation::Template
  AWSTemplateFormatVersion: '2010-09-09'
  Description: 'CloudFormation template for real-time data-intensive application pipeline'

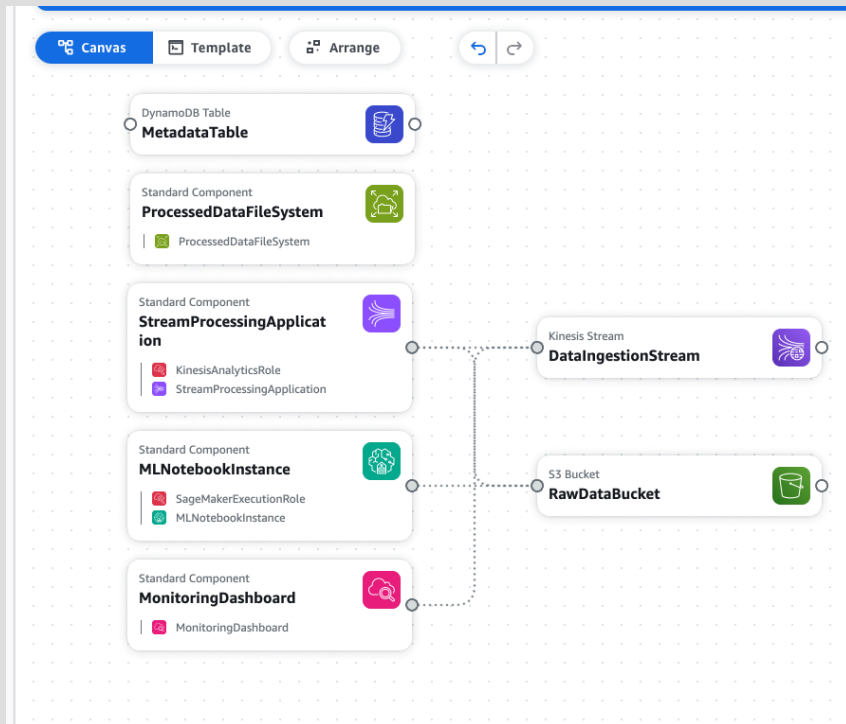
  Resources:
    # Kinesis Data Stream for data ingestion
    DataIngestionStream:
      Type: AWS::Kinesis::Stream
      Properties:
        Name: data-ingestion-stream
        ShardCount: 10
        RetentionPeriodHours: 24

    # S3 Bucket for raw data storage
    RawDataBucket:
      Type: AWS::S3::Bucket
      Properties:
        BucketName: raw-data-bucket
        VersioningConfiguration:
          Status: Enabled

    # Kinesis Data Analytics Application for stream processing
    StreamProcessingApplication:
      Type: AWS::KinesisAnalyticsV2::Application
      Properties:
        ApplicationName: stream-processing-app
        RuntimeEnvironment: SQL-1_0
        ServiceExecutionRole: !GetAtt KinesisAnalyticsRole.Arn
        ApplicationConfiguration:
          SqlApplicationConfiguration:
            Inputs:
              - NamePrefix: SOURCE_SQL_STREAM
                KinesisStreamsInput:
                  ResourceARN: !GetAtt DataIngestionStream.Arn
```

1. Generate data pipeline templates

Automatic data pipeline creation



- Automatically generated
- Customisable and modifiable based on data or additional needs
- Testable in real scenario
- Deployment ready

2. Deploy template directly

Automatic dataset generation for questions and answering tasks

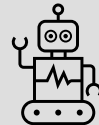
Question: Which Socket Gateway should I be using with Netcool Omnibus (NOI) ? There are 2 versions of the gateway; nco-g-socket-java-2_0 and nco-g-socket-10_0 Users may be confused what version they should download or use.

Trained on
original dataset



Answer 1: The Netcool Omnibus Socket Gateway is a load balancing tool. The nco-g-socket-java-2_0 gateway is a Java based gateway. This version is no longer supported...

Trained on
generated dataset



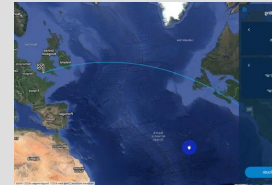
Answer 2: For AIX, Linux, Solaris, and Windows, use the Netcool/OMNIBus V8.1 package...



Reference Answer: A new version of the Netcool/OMNIBus Socket Gateway will be available to download from August 6, 2015.

Summary thoughts on Innovation

- Gartner “85% of AI projects fail due to unclear objectives and obscure R&D project management processes. As well, 87% of R&D projects never get to the production phase, while 70% of clients indicated minimal or even no impact from AI”
- However, Data Science, Machine Learning and AI are making a difference in engineering applications
- Need to have a clear business case (what can AI do that you want to do? Where is the value?)
- Choose the correct tech, not the popular tech.
- Long-term relationships
- Find the correct expert



T-VOS: Voyage Optimisation Software






Defiant Medium Unmanned Surface Vehicle Concept



Bluebox prototype, active learning part of Silverstream data pipeline



Yacht concept design tool

Benefits		
 Reduced fuel cost	 Reduced emissions	 Low implementation cost
By optimizing trim and draft, savings in the range of 3-7% can be achieved	Fuel savings directly result in CO2 reductions	Only K-IMS, a maritime industry platform and data logger, is required for JAWS

The Alan Turing Institute



Thanks to Theyr, Silverstream and Lloyd's
Register Foundation

Amy Parkes, Josef Camilleri, Dominic Hudson, Przemyslaw
Grudniewski, Ryan Reisinger, Tom Ezard, Sizhe Yuen and Ting Su

asobey@turing.ac.uk

7th November 2024

