Rapid determination of bulk composition and quality of marine biomass in mass spectrometry

Doctoral proposal seminar

Jesse Wood[1] Bach Hoai Nguyen[1] Bing Xue[1] Mengjie Zhang[1] Daniel Killeen[2]

[1]School of Engineering and Computer Science — Te Kura Mātai Pūkaha, Pūrorohiko
Victoria University of Wellington — Te Herenga Waka
[2]New Zealand Institute for Plant and Food Research Limited, Nelson, New Zealand

Island Bay, Wellington, New Zealand

Fish + AI + Business

Paul the octopus

Topics

1. Problem Statement
2. Motivations
3. Goals
   - Datasets
   - Identification
   - Quantitative contaminant analysis
   - Traceability
4. Preliminary Work
5. Proposed contributions, thesis outline, timeline
   - Proposed contributions
   - Thesis outline
   - Timeline

Have you been catfished? [1]

Daily Mail AUSTRALIA

Home | U.K. | U.S. | News | World News | Sport | TV&Showbiz | Femail AU | Health | Scien

Monkeypox | Covid-19 | Russia-Ukraine War | Breaking News | Sydney | Melbourne | Brisbane | New Zealand |

Popular restaurant accused of serving cheap Vietnamese catfish to customers who thought they were getting Australian dory

- **A Melbourne restaurant has been accused of serving catfish to customers**
- **Hunky Dory has allegedly been selling frozen fillets of basa as dory**
- **Owner Greg Robotis has denied allegations he is misleading customers**
- **The City of Port Phillip is investigating Hunky Dory's Port Melbourne store**

By HARRY PEARL FOR DAILY MAIL AUSTRALIA
PUBLISHED: 14:31 AEDT, 27 May 2016 | **UPDATED:** 16:08 AEDT, 27 May 2016

Share 47 shares 9 View comments

A Melbourne restaurant has been accused of serving a Vietnamese catfish to customers who believe they are ordering Dory.

A whistleblower has alleged that Hunky Dory outlets have been selling frozen fillets of basa, a species of catfish native to the Mekong basin, as fish-of-the-day dory, **The Age** reports.

Owner Greg Robotis has denied the claims and said inexperienced staff may have been calling the fish the wrong name.

Aussies! No surprises there...

Catfishing [1], Mislabelling [2], and Quality Assurance [3]

Nutrition Facts

6 servings per container

Serving size **4-5 ounces(187g)**

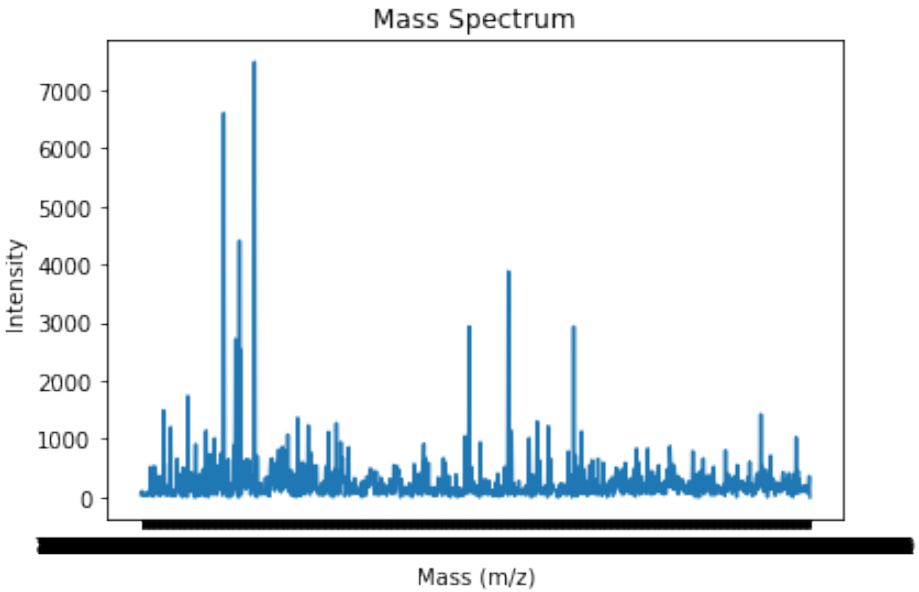
Amount per serving

Calories **200**

	% Daily Value*
Total Fat 5g	**6%**
Saturated Fat 0.5g	**3%**
Trans Fat 0g	
Cholesterol 80mg	**27%**
Sodium 610mg	**27%**
Total Carbohydrate 10g	**4%**
Dietary Fiber 0g	**0%**
Total Sugars 3g	
Includes 0g Added Sugars	**0%**
Protein 27g	
Vitamin D 2mcg	10%
Calcium 79mg	6%
Iron 3mg	15%
Potassium 519mg	10%

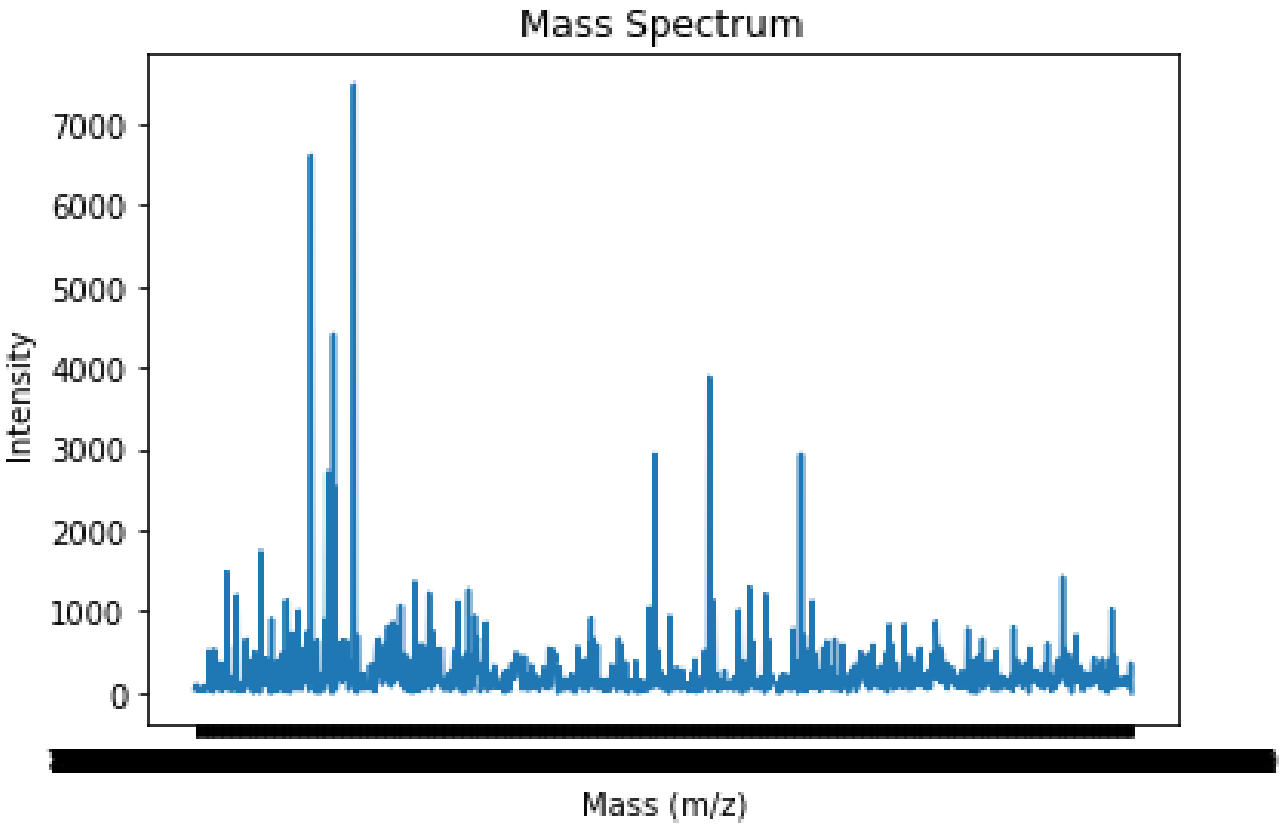
*The % Daily Value tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Mass Spectrometry [3] ≈ Chemical Fingerprint

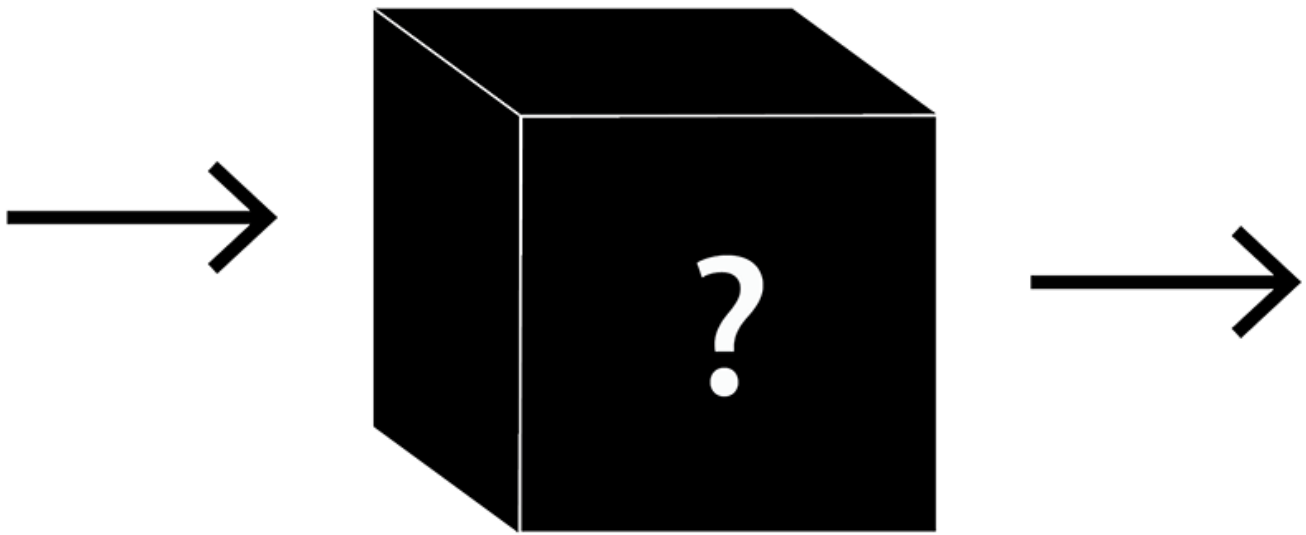


Fish oil is brain food! [4, 5]

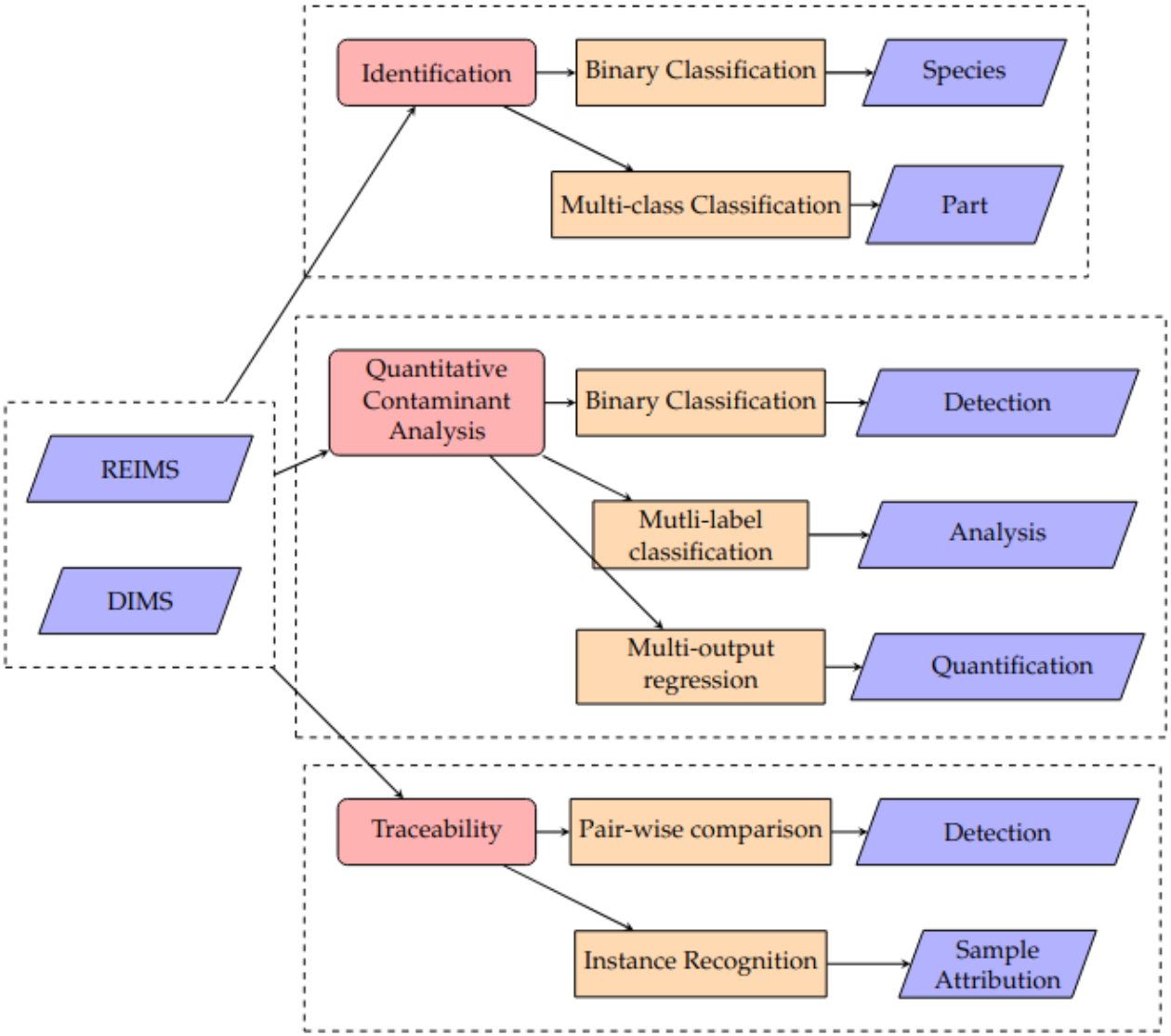
Fish analyzed with Mass Spectrometry! [6]



Fish oil analysis can't be blackbox! [7, 8]

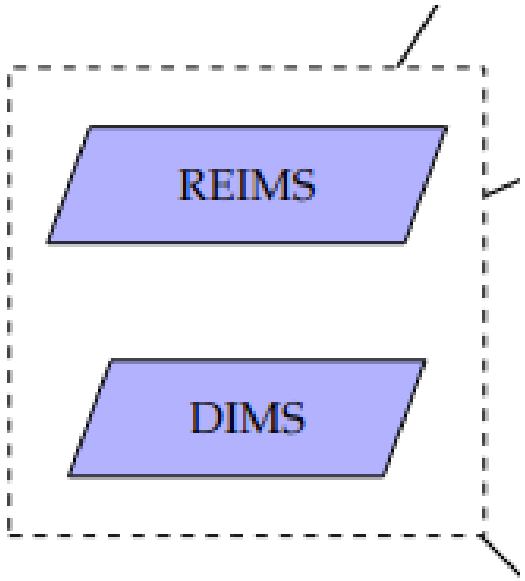


Research goals



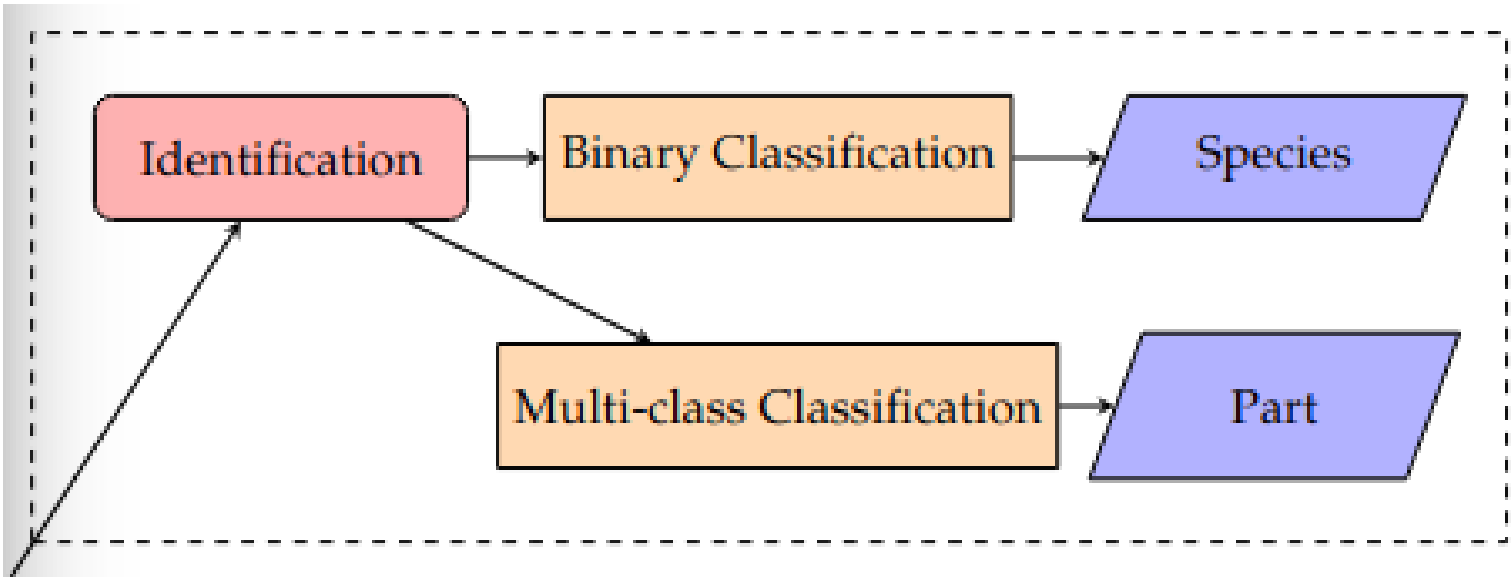
Research goals

Datasets



Datasets

Fish identification

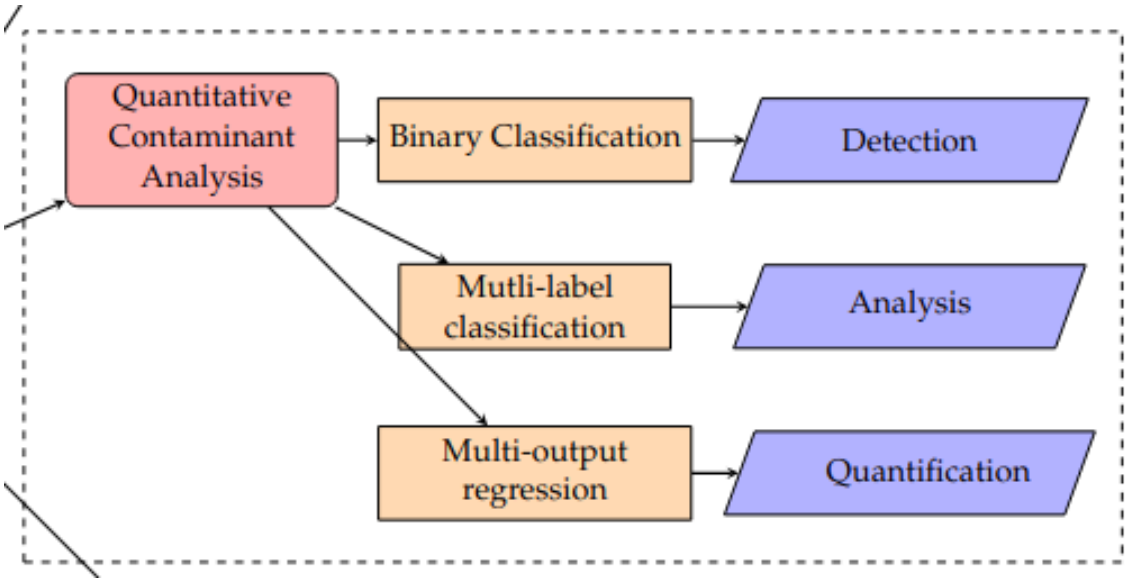


Research goal - identification

Fish species identification

Fish body parts identification

Quantitative contaminant analysis



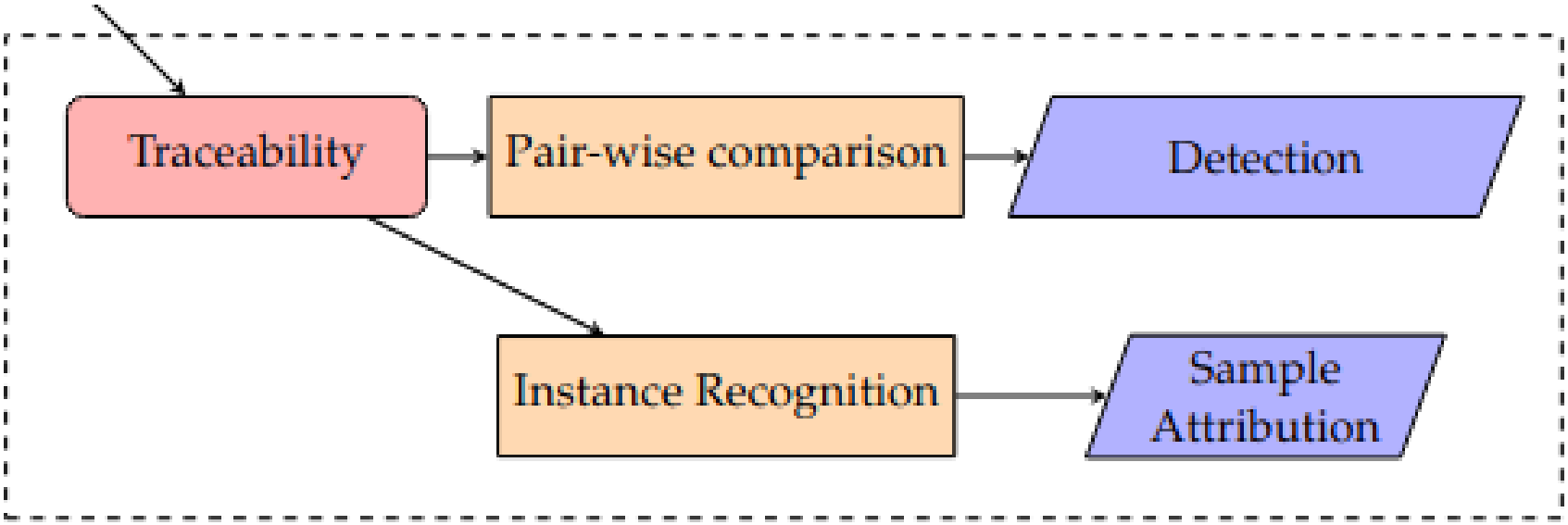
Research goal - contamination

Contamination detection

Contamination analysis

Contamination quantification

Traceability



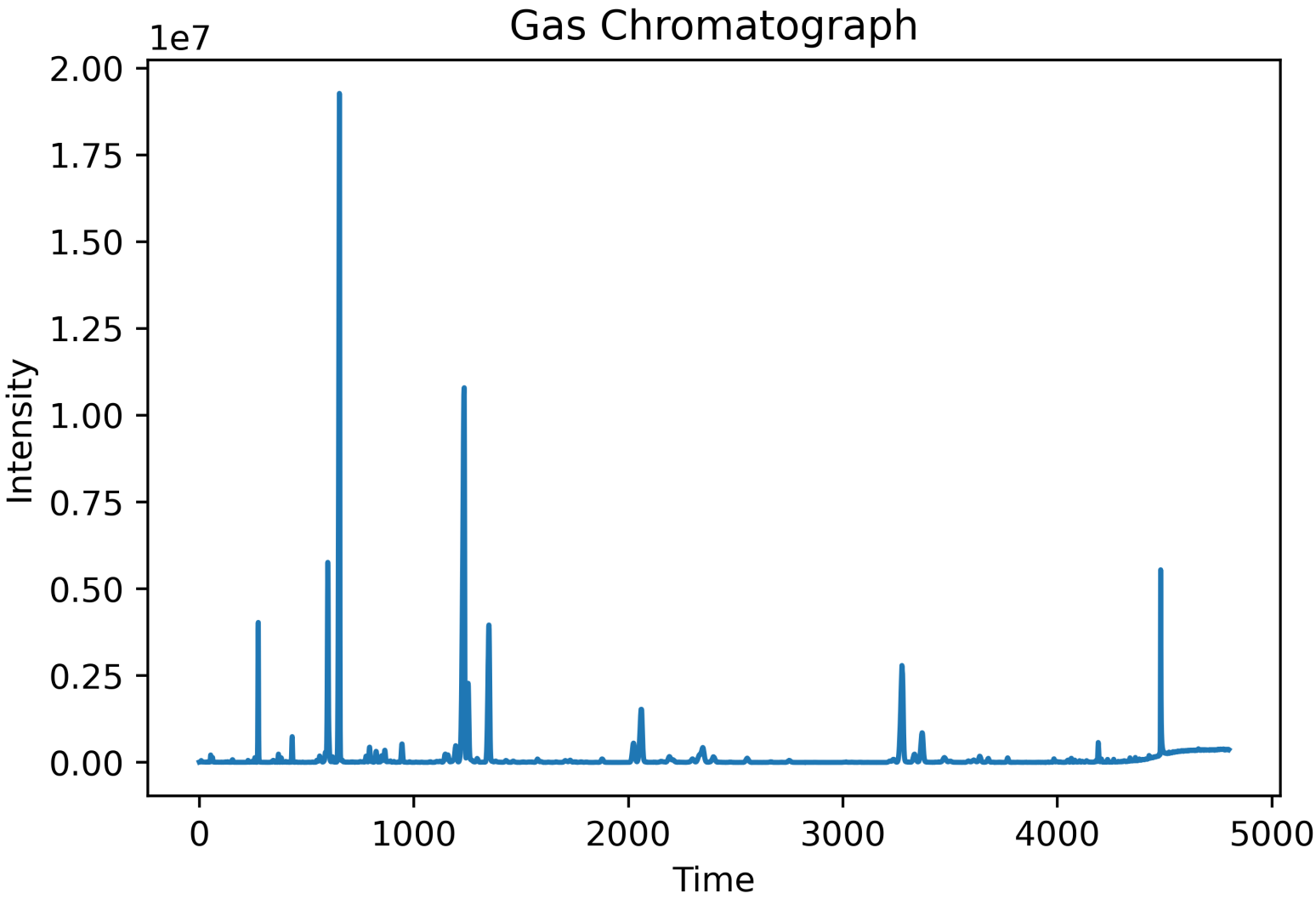
Research goal - traceability

Detection - pair-wise comparison

Sample attribution - instance recognition

Evolutionary Computation for gas chromatography

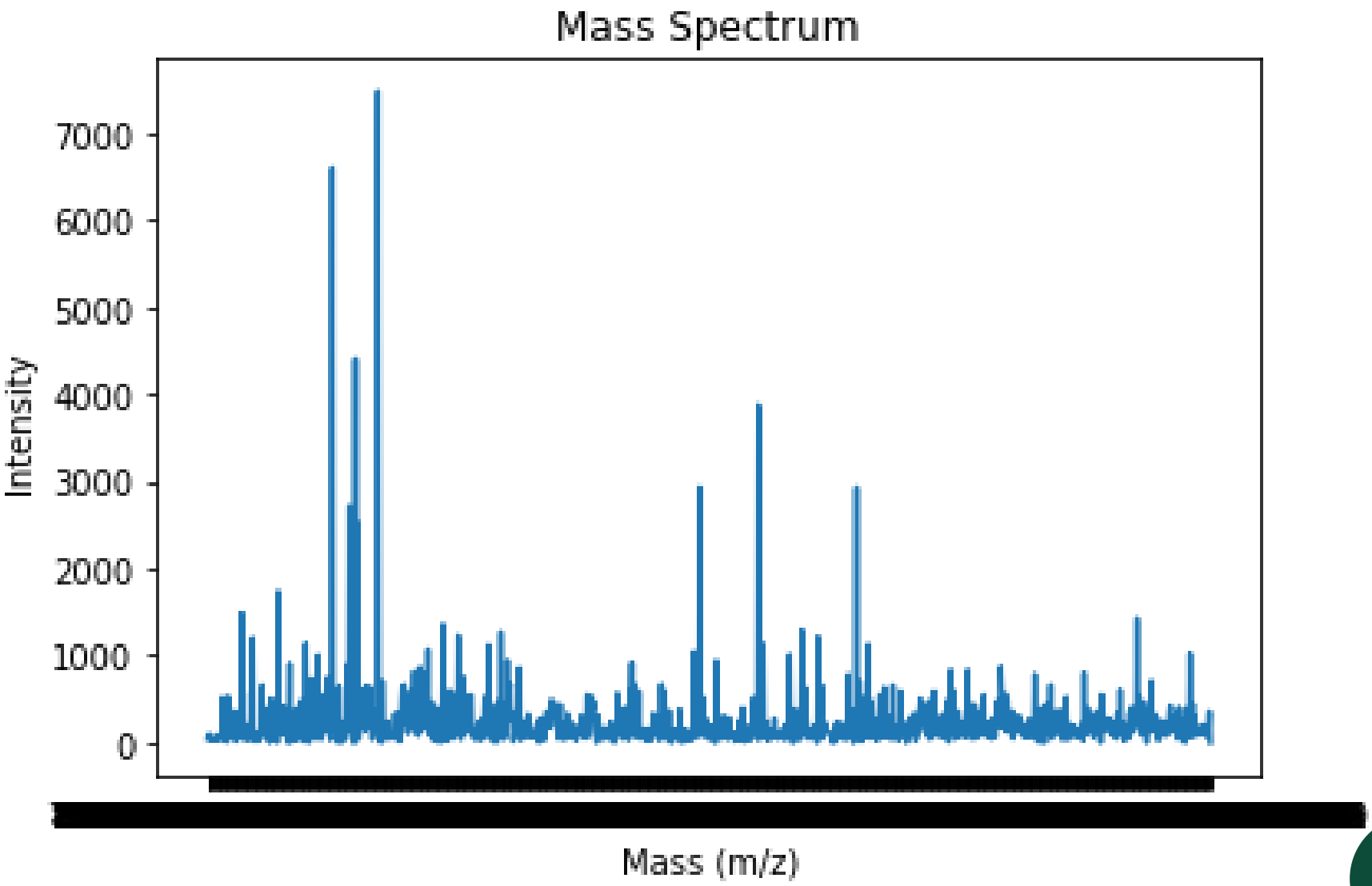
Gas chromatography



Gas chromatograph

Exploratory data analysis for rapid mass spectrometry

Mass spectrometry



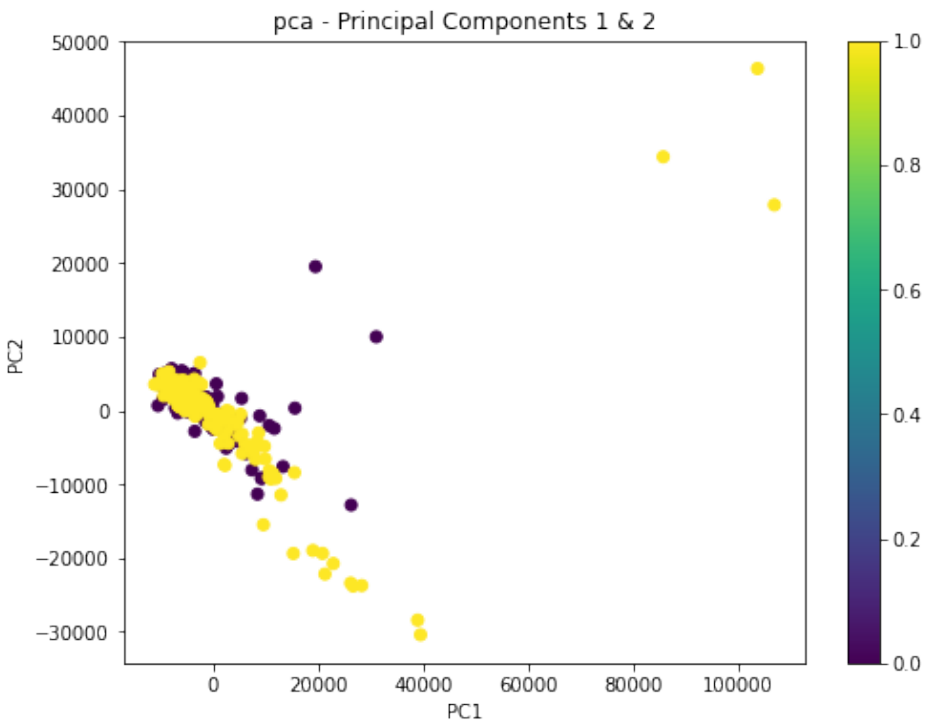
Mass spectrograph

Exploratory Data Analysis - PCA

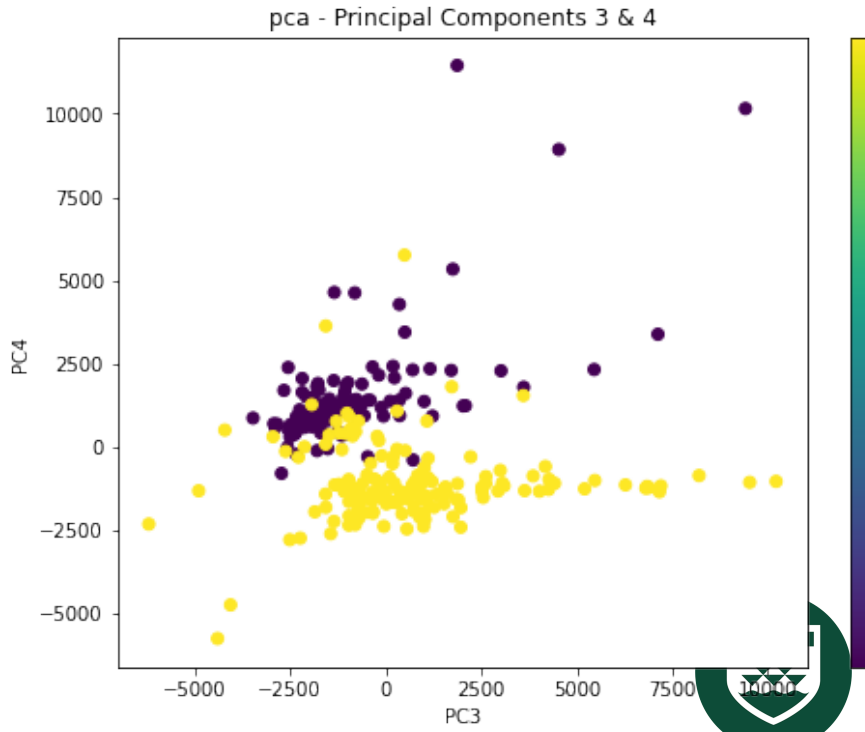
Visual intuition for dimensionality reduction techniques and their respective feature subsets

PCA [9]

Features 1 & 2



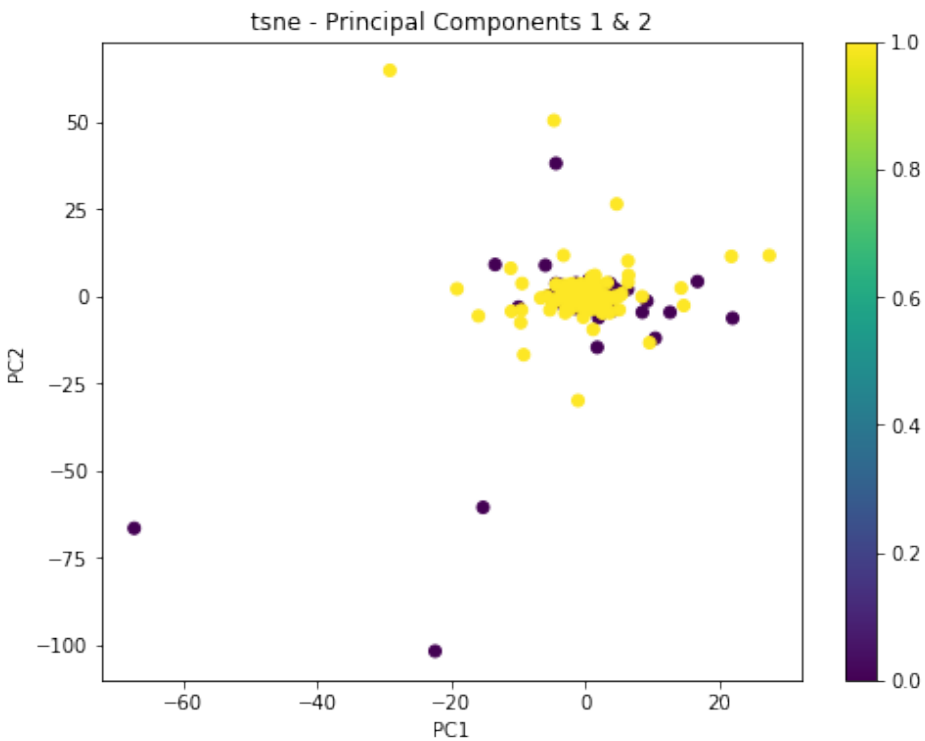
Features 3 & 4



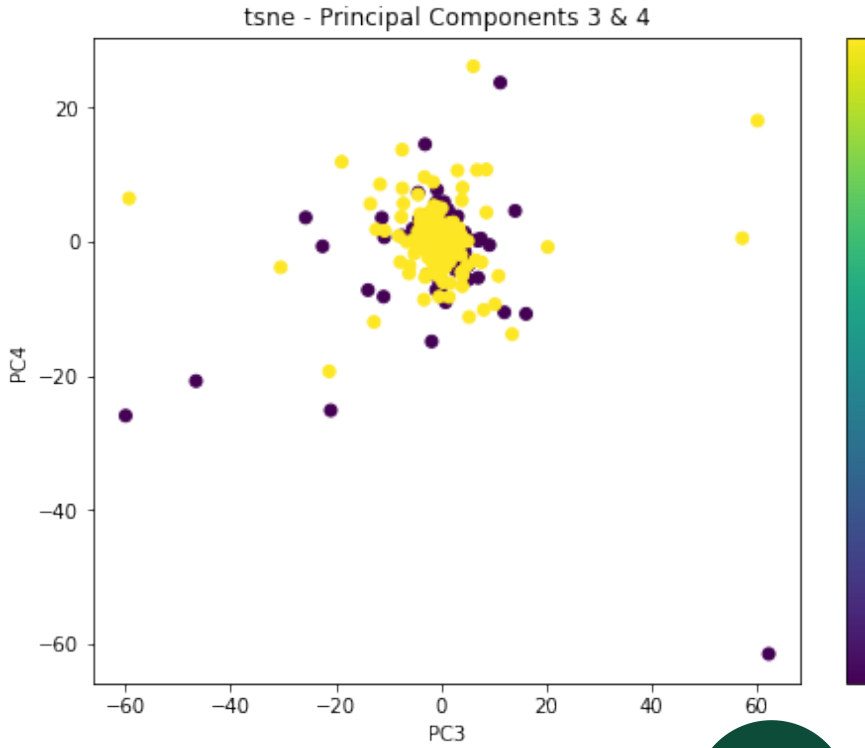
Exploratory Data Analysis - t-SNE

t-SNE [10]

Features 1 & 2



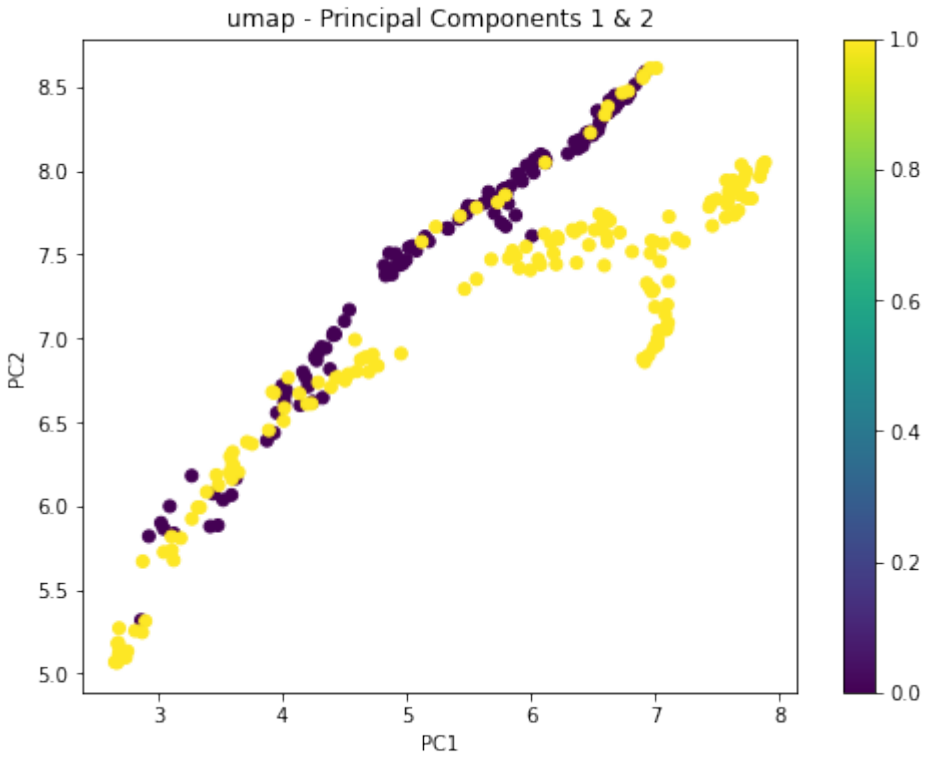
Features 3 & 4



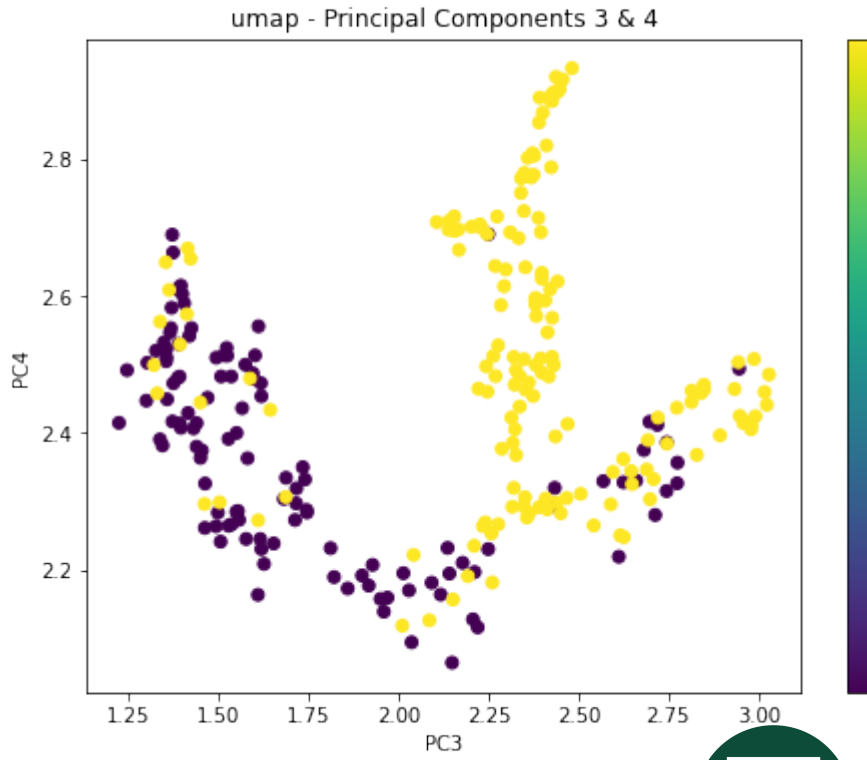
Exploratory Data Analysis - UMAP

UMAP [11]

Features 1 & 2



Features 3 & 4



Proposed contributions

The work will contribute novel methods for rapid determination of bulk composition and quality of marine biomass in Mass Spectrometry. The proposed contributions are as follows.

- **Fish identification** - a variable-width input multi-scale classification task.
- **Quantitative contaminant analysis** - an extension of existing outlier thresholding techniques that not only detect contamination but also provide quantitative and qualitative profiles of contaminants found.
- **Traceability** - a one/few-shot learning problem that employs novel pre-training strategies adapted from natural language processing.

Contributions

Fish identification - a variable-width input multi-scale classification task - Resolution-invariant (e.g. variable-width input) via data augmentation, synthetic datasets, and pretraining strategies like Next Spectra Prediction and Masked Spectra Modeling. The model learns multi-scale resolution-invariant representations that can handle low-resolution rapid mass spectrometry and high-resolution direct-infusion spectrometry.

This model determines the viability of rapid mass spectrometry for use in fish processing, compared and contrasted with the slower direct-infusion mass spectrometry, whose *rapid* nature will dramatically increase factory throughput.

1. Multi-scale resolution-invariant fish **species identification** with binary classification.
2. Multi-scale resolution-invariant fish **body part identification** with multi-class classification.

Contributions

Quantitative contaminant analysis - an extension of existing outlier thresholding techniques that not only detect contamination but also provide quantitative and qualitative profiles of contaminants found. The proposed model will have mechanistic interpretability for trusted use by domain experts in fish processing. Implemented with four methods:

1. Contamination **detection** of cross-species or mineral oil contaminants with binary classification.
2. Contamination **analysis** of cross-species or mineral oil contaminants with multi-label classification, which gives predicted contaminants present in the sample.
3. Contamination **quantification**, which gives relative percentages of each contaminant, with multi-label multi-output regression, where one or more contaminants are identified, and their relative percentage of the composition of the sample is estimated via regression.
4. **Black swans**, the unknown unknowns [12, 13]. An extension of outlier thresholding [14, 6] to detect out-of-distribution classes, classes not in the training data, as unknown contaminants, that can

Contributions

Traceability - a one/few-shot learning problem that employs novel pre-training strategies adapted from natural language processing, to decode and interpret a substantive repository of over 14,000 mass spectrometry datasets. Learning a universal feature embedding that is applied to the downstream tasks of:

1. **Detection** - detects if two samples come from the same individual fish via similarity learning for pair-wise comparison.
2. **Instance recognition** - assigns unique identifiers for each individual fish, to unseen instances that can be matched to existing fish that have been uniquely labelled via instance recognition. A novel few-shot similarity-based contrastive learning approach, for instance recognition.

Thesis outline

Abstract

Glossary

Chapter 1 - Introduction

Chapter 2 - Literature Survey

Chapter 3 - Datasets and Processing

Chapter 4 - Fish Species and Part Identification

Chapter 5 - Fish Quantitative Contaminant Analysis

Chapter 6 - Fish Traceability Analysis

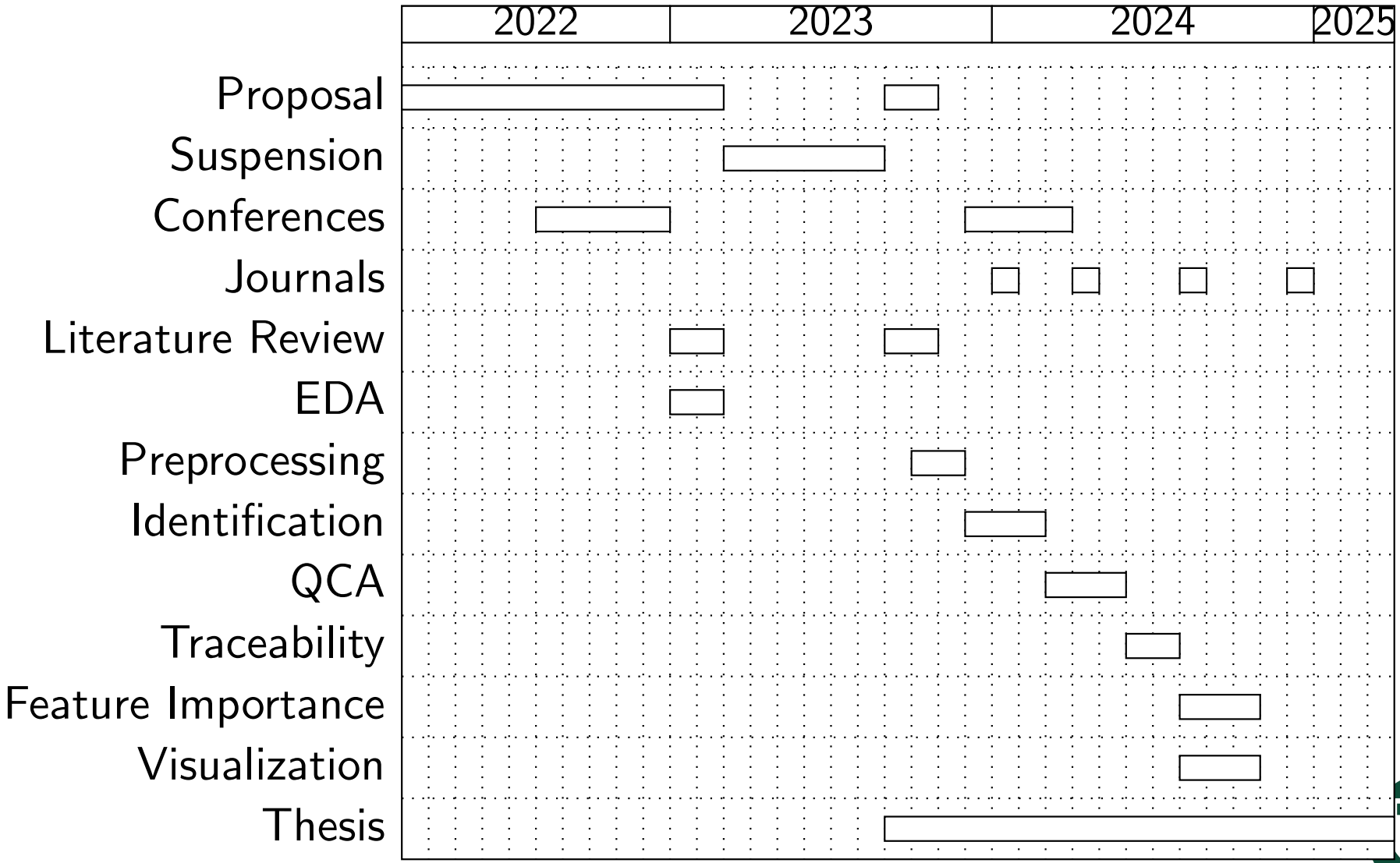
Chapter 7 - A Case Study, Demonstrations and Discussions

Chapter 8 - Conclusions

Bibliography

Index

Timeline


2022
2023
2024
2025
Proposal
Suspension
Conferences
Journals
Literature Review
EDA
Preprocessing
Identification
QCA
Traceability
Feature Importance
Visualization
Thesis

TLDR;

Novel AI variable-width input multi-scale resolution-invariant pre-trainied analysis algorithms for rapid mass spectrometry datasets of marine biomass.

1. fish **species** & body **part identification** $\rightarrow$ multi-scale resolution-invariant binary & multi-class classification, respectively;
2. **quantitative contaminant analysis** $\rightarrow$ multi-label classification & multi-output regression & out-of-distribution outlier thresholding; and
3. **traceability** $\rightarrow$ few-shot similarity-based contrastive learning for pair-wise comparison & instance recognition.

[1] H. P. F. D. M. Australia, "Melbourne restaurant hunky dory accused of serving catfish to customers instead of dory," May 2016. [Online]. Available: https://www.dailymail.co.uk/news/article-3611999/Melbourne-restaurant-Hunky-Dory-accused-serving-catfish-customers-ins html

[2] M. Á. Pardo, E. Jiménez, and B. Pérez-Villarreal, "Misdescription incidents in seafood sector," *Food Control*, vol. 62, pp. 277–283, 2016.

[3] K. Eder, "Gas chromatographic analysis of fatty acid methyl esters," *Journal of Chromatography B: Biomedical Sciences and Applications*, vol. 671, no. 1-2, pp. 113–131, 1995.

[4] A. P. Simopoulos, "Evolutionary aspects of diet: the omega-6/omega-3 ratio and the brain," *Molecular neurobiology*, vol. 44, no. 2, pp. 203–215, 2011.

[5] M. L. Panse and S. D. Phalke, "World market of omega-3 fatty acids," *Omega-3 Fatty Acids*, pp. 79–88, 2016.

[6] C. Black, O. P. Chevallier, K. M. Cooper, S. A. Haughey, J. Balog, Z. Takats, C. T. Elliott, and C. Cavin, "Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[7] K. Bi, D. Zhang, T. Qiu, and Y. Huang, "Gc-ms fingerprints profiling using machine learning models for food flavor prediction," *Processes*, vol. 8, no. 1, p. 23, 2019.

[8] D. D. Matyushin and A. K. Buryak, "Gas chromatographic retention index prediction using multimodal machine learning," *Ieee Access*, vol. 8, pp. 223 140–223 155, 2020.

[9] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[10] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[11] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[12] N. N. Taleb, *Fooled by randomness: The hidden role of chance in life and in the markets*. Random House Trade Paperbacks, 2005, vol. 1.

[13] ——, *The black swan: The impact of the highly improbable*. Random house, 2007, vol. 2.

[14] C. Black, O. P. Chevallier, S. A. Haughey, J. Balog, S. Stead, S. D. Pringle, M. V. Riina, F. Martucci, P. L. Acutis, M. Morris *et al.*, "A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry," *Metabolomics*, vol. 13, no. 12, pp. 1–13, 2017.