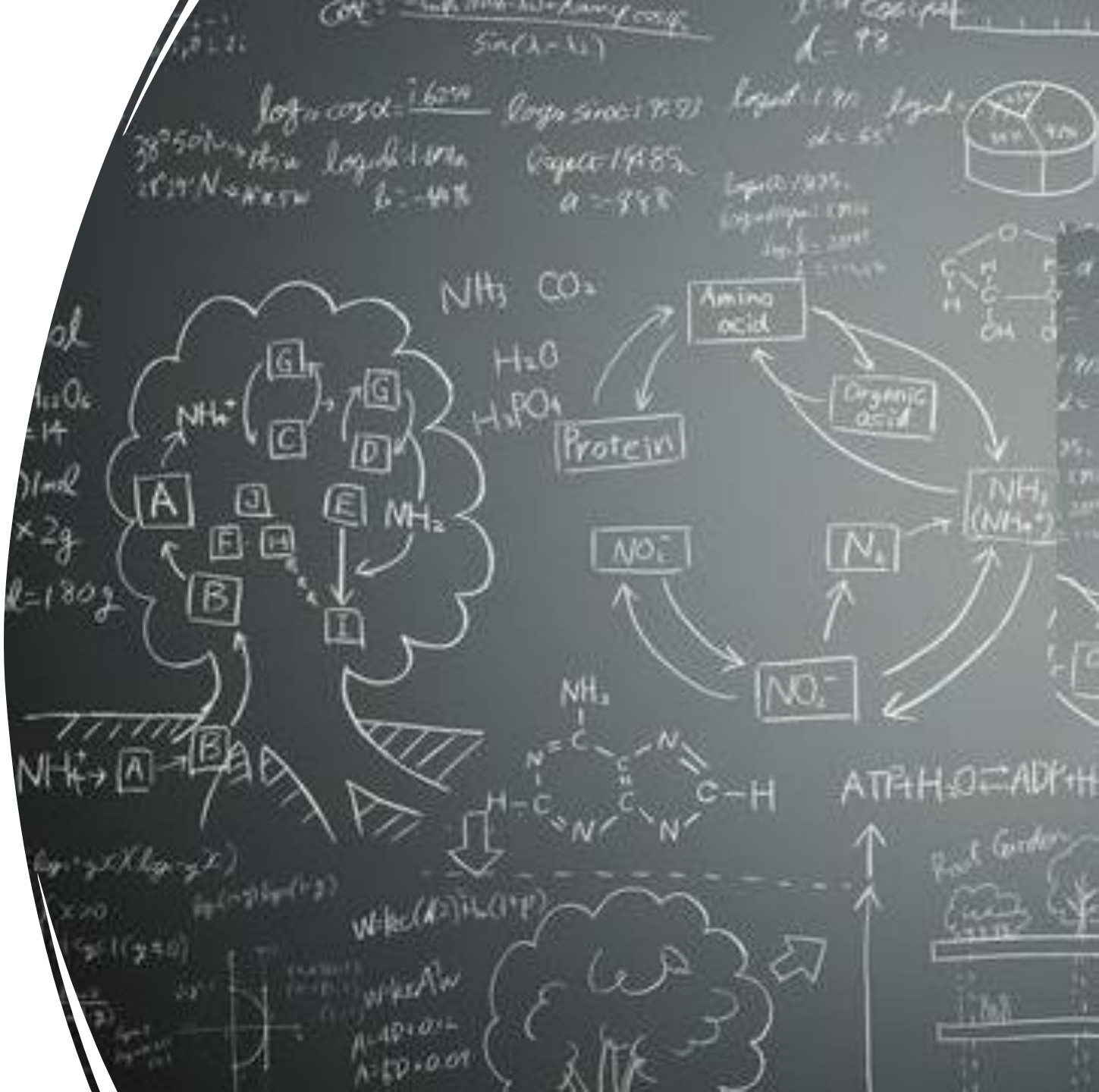
UCT Department of Statistical Sciences
Data Science Research and Supervision Network

What is Data Science?

- Data science is an interdisciplinary field:
  - Statistics,
  - Computer science,
  - Mathematics and
  - domain specific fields such as Astronomy, Physics, Bioinformatics, Economics, Finance etc.

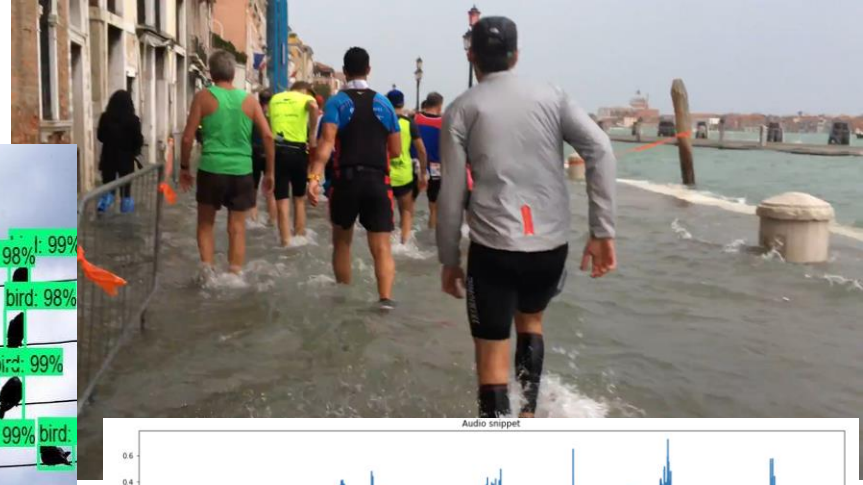
What is Data Science?

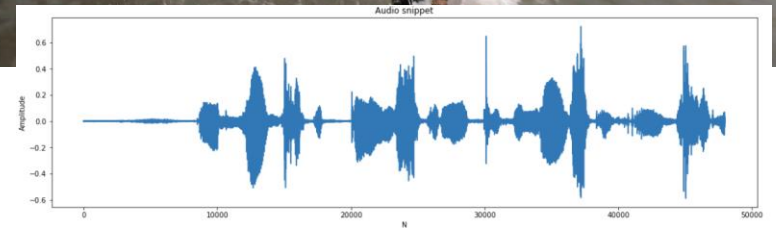
- Data science involves to collect, process, analyze, and interpret data in order to make data-driven decisions.

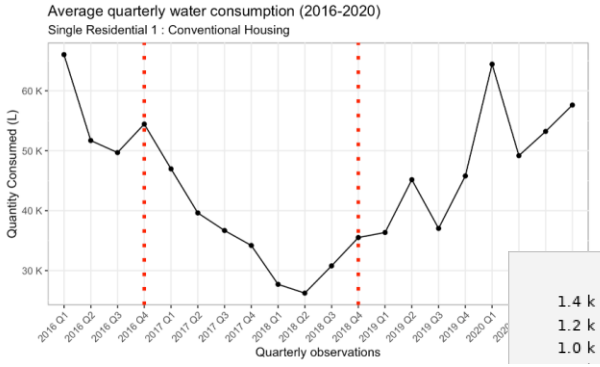


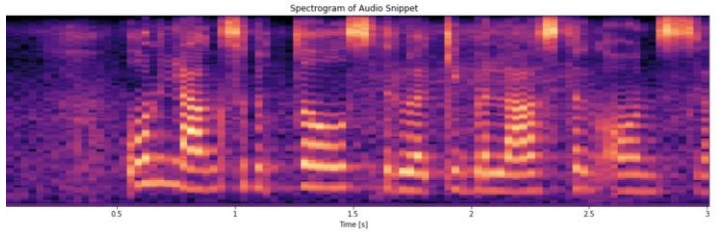
What type of Data?

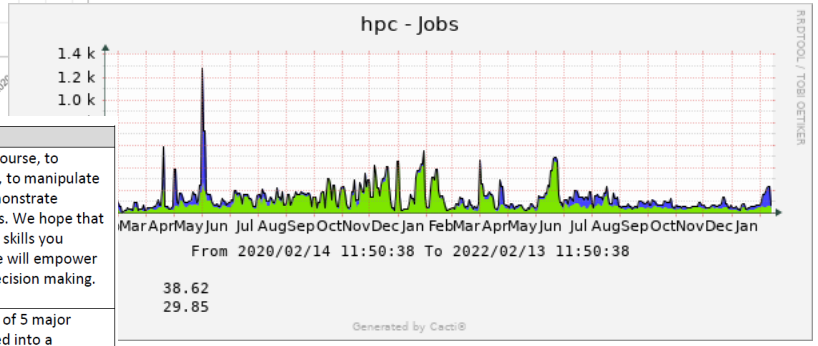
- Data scientists work with large and complex data sets, which may include structured and unstructured data from various sources such as databases, sensors, social media, and the internet.











Text from the Course Information Pack	Question	Answer
There is also a strong emphasis on using Excel in this course, to perform calculations, to manipulate data and also to demonstrate principles of statistics. We hope that you will find that the skills you acquire in this course will empower you in your future decision making.	Why is Excel used in this course?	Excel is used in this course, to perform calculations, to manipulate data and also to demonstrate principles of statistics. We hope that you will find that the skills you acquire in this course will empower you in your future decision making.
The syllabus consists of 5 major modules, each divided into a number of work units. These will be laid out for you on Vula under Course Modules.	What is the syllabus of this course?	The syllabus consists of 5 major modules, each divided into a number of work units. These are laid out for you on Vula under Course Modules.
Within each work unit, you will be given some material to engage with (read notes, watch videos, complete quizzes, etc) in a step-wise fashion. This material will become available sequentially (a new section is revealed every Friday	When is new course material made available?	New course material will become available every Friday afternoon. Within each work unit, you will be given some material to engage with (read notes, watch videos, complete quizzes, etc) in a step-wise fashion.

Data Science Domains

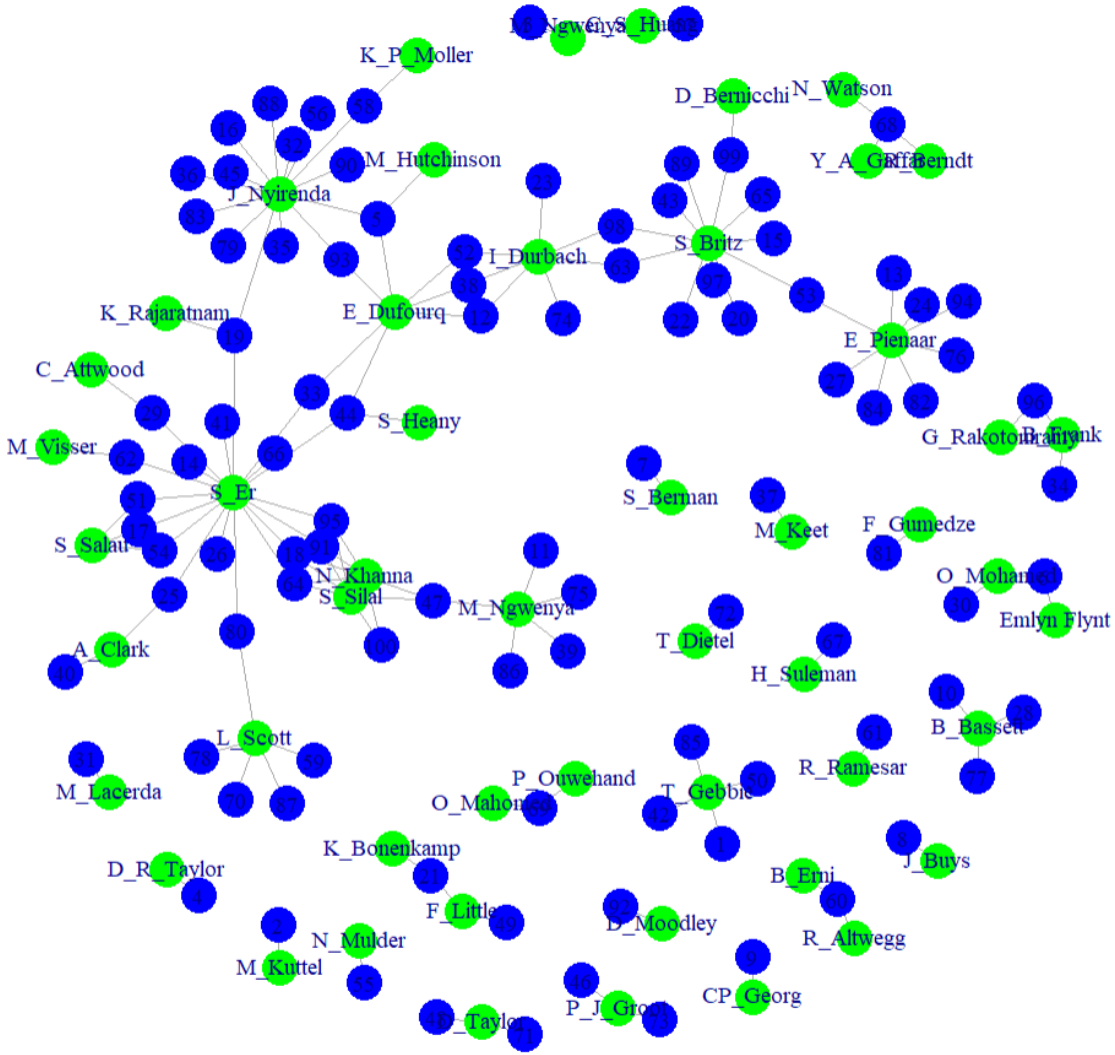
- As Prof. Rosling Says We Are Surrounded by Statistics. In which areas?
  - Student evaluation
  - Economics
  - Supermarkets
  - Air traffic control systems
  - Management
  - Sports – Football, Basketball, Swimming etc. Coin toss before a match.
  - Spam Filtering – Google!
  - Crime Statistics
  - Medicine
  - Food technologies
  - Engineering
  - Gambling
  - Customer Satisfaction
  - Detecting Credit Card Fraud

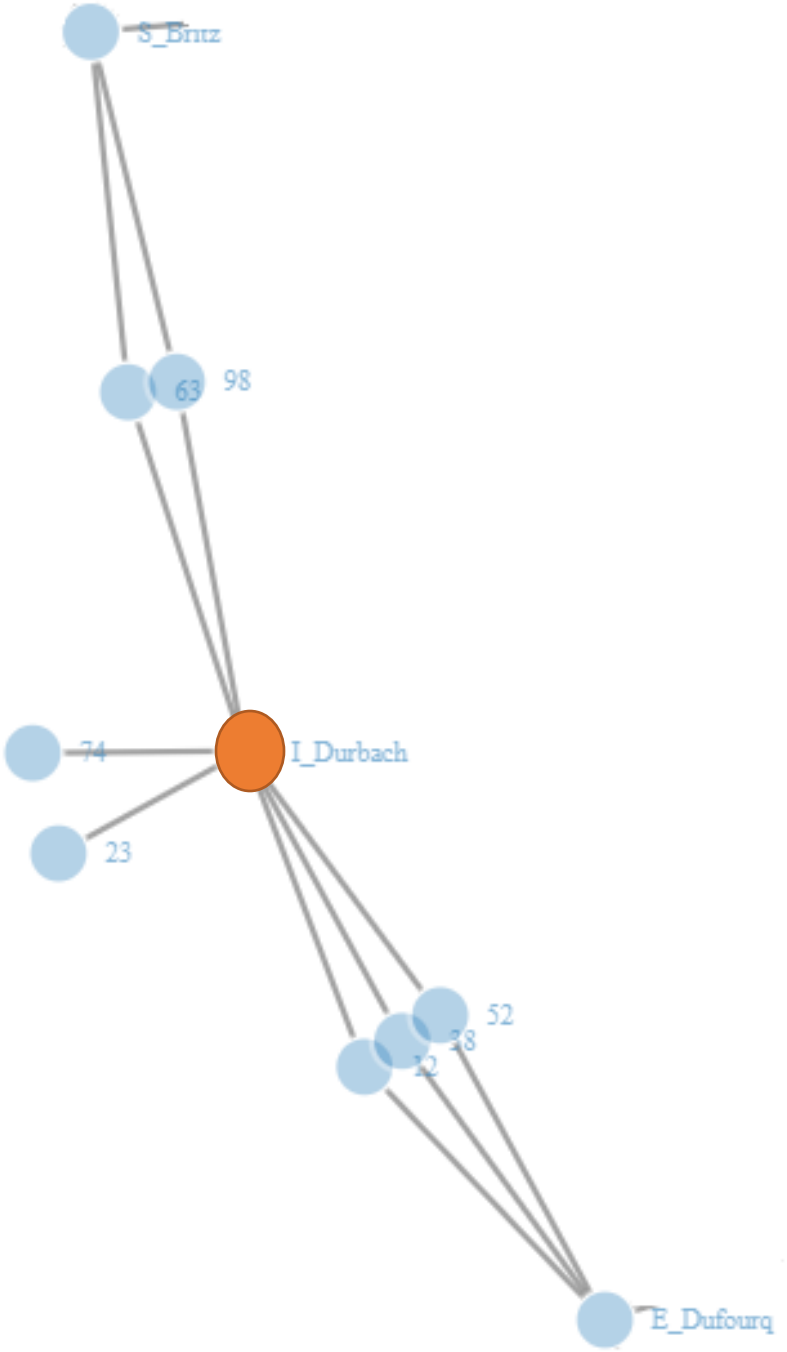
MSc specializing in DS – 2017 to date

MSc specializing in Data Science degree

- Collaborative program
- Students can choose 120/60 or 90/90 option. Both require a research component on a data science related real problem.

The network of data science projects at UCT (2018-2023)





Ian Durbach

- NLP
- Image analysis
- Sound analysis
- Ecology

Ian Durbach

- **Natural Language Processing** on Data Warehouses (2019)

Value	Quantity	Date	Product	Subcategory	Category	Shop	City	Province	Country
205.146	9	2006-07-01	Full-Finger Gloves, M	Gloves	Clothing	Top Sports Supply	Edmonton	Alberta	Canada
7873.146	9	2006-04-01	Road-450 Red, 58	Road Bikes	Bikes	Retail Mall	Richmond	British Columbia	Canada
71.988	2	2007-04-01	Men's Sports Shorts, M	Shorts	Clothing	Original Bicycle Supply Company	Toronto	Ontario	Canada
418.512	2	2006-09-01	ML Mountain Frame - Black, 38	Mountain Frames	Components	Rapid Bikes	Toronto	Ontario	Canada
2458.9178	2	2007-02-01	Mountain-200 Black, 38	Mountain Bikes	Bikes	Leisure Activities	Hull	Quebec	Canada
76.2	2	2008-05-01	Classic Vest, S	Vests	Clothing	Uttermost Bike Shop	Bracknell	England	United Kingdom
5102.97	5	2007-07-01	Road-350-W Yellow, 40	Road Bikes	Bikes	Prosperous Tours	London	England	United Kingdom
1336.23	3	2007-12-01	Touring-3000 Yellow, 44	Touring Bikes	Bikes	Fun Toys and Bikes	Tucson	Arizona	United States
84.7734	6	2006-12-01	Half-Finger Gloves, M	Gloves	Clothing	Eastside Department Store	Union City	California	United States
1070.694	3	2007-11-01	ML Road Frame-W - Yellow, 38	Road Frames	Components	General Associates	Hollywood	Florida	United States
744.2727	1	2006-07-01	HL Mountain Frame - Silver, 42	Mountain Frames	Components	Valuable Bike Parts Company	Orlando	Florida	United States
3887.964	6	2006-11-01	Mountain-300 Black, 38	Mountain Bikes	Bikes	Noiseless Gear Company	Columbus	Georgia	United States
838.9178	2	2006-02-01	Road-650 Black, 60	Road Bikes	Bikes	District Mall	Ferguson	Missouri	United States
400.104	2	2008-06-01	LL Touring Frame - Yellow, 50	Touring Frames	Components	Sleek Bikes	Denby	South Dakota	United States
35.994	1	2006-11-01	Men's Sports Shorts, S	Shorts	Clothing	Genial Bike Associates	Humble	Texas	United States
392.658	2	2007-02-01	HL Mountain Rear Wheel	Wheels	Components	Genial Bike Associates	Humble	Texas	United States
606.996	3	2006-10-01	LL Road Frame - Red, 48	Road Frames	Components	Friendly Bike Shop	Bellingham	Washington	United States
377.946	9	2008-06-01	Women's Mountain Shorts, L	Shorts	Clothing	Metro Cycle Shop	Tacoma	Washington	United States

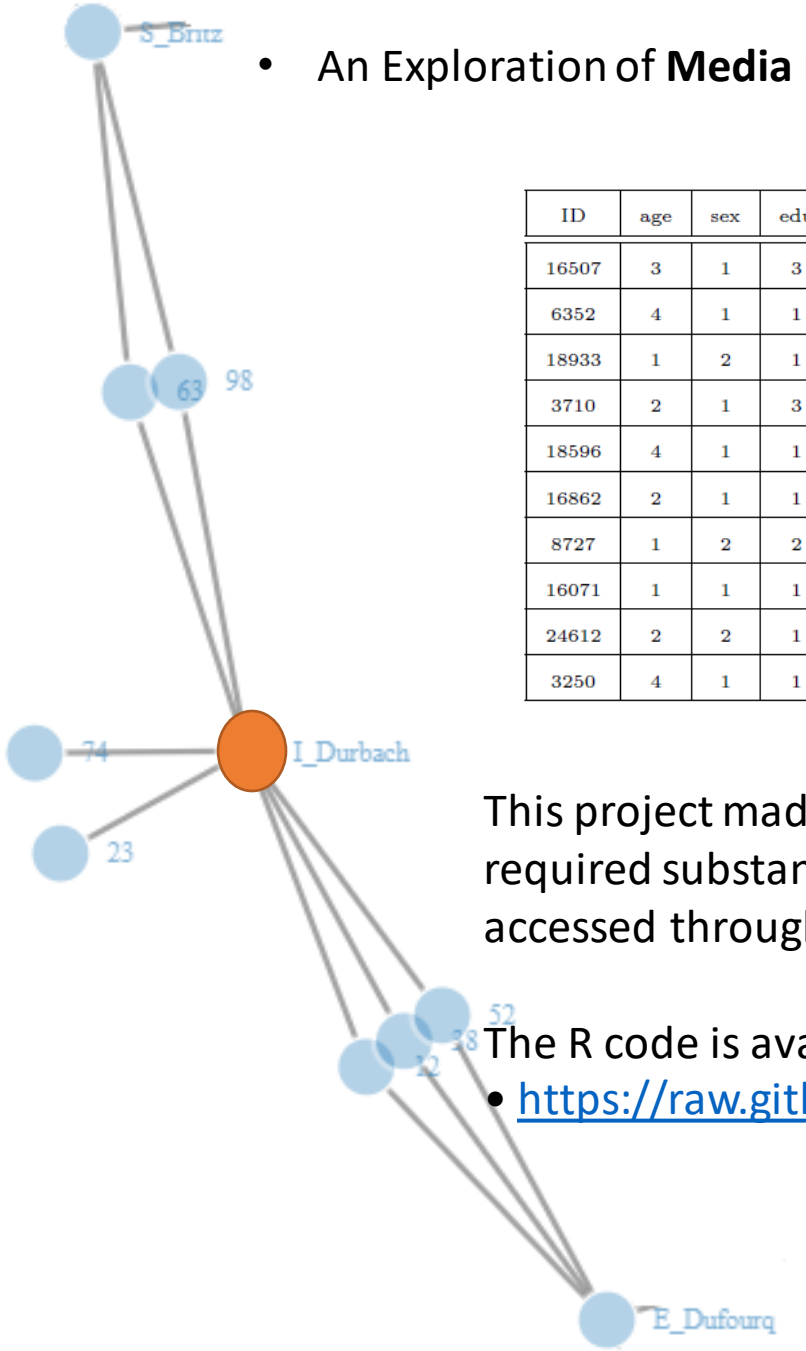
Figure 4.1: A sample of the Adventureworks data were given to respondents of the survey in order to complete the questions.

what
what is
what is the
what is the for
what is the for at
what is the for at united
what is the for at united states
what is the for at united states for
what is the for at united states for accessories

The main problem addressed in this research was to use natural language to query data in a data warehouse (**Adventureworks** open source dataset).

Two natural language processing models were developed and compared on a classic star-schema sales data warehouse with sales facts and date, location and item dimensions. Utterances are queries that people make with natural language, for example, **"What is the sales value for mountain bikes in Georgia for 1 July 2005?"**

The practical application of the research is that these models can be used as a **component in a chatbot on data warehouses**. Combined with a Structured Query Language query generation component, and building Application Programming Interfaces on top of it, this facilitates the quick and easy distribution of data; no knowledge of a programming language such as Structured Query Language is needed to query the data



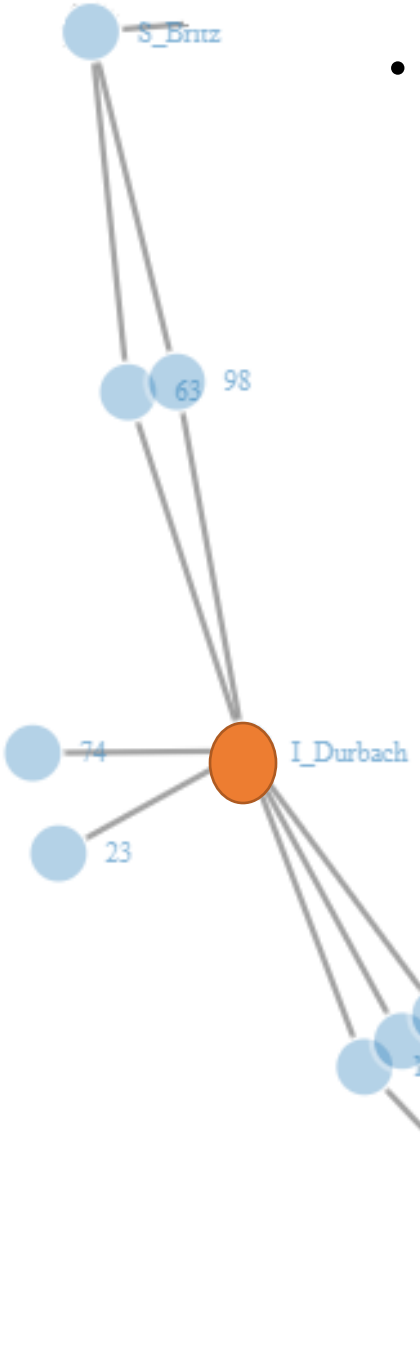
- An Exploration of **Media Repertoires** in South Africa: 2002-2014 (2019)

ID	age	sex	edu	race	lsm	magazines	radio	tv	internet	all	Ilanga	Topcar	Kaya.FM	int_tv
16507	3	1	3	4	5	2	6	3	2	0	0	0	0	0
6352	4	1	1	2	4	0	3	13	0	-1	0	0	0	0
18933	1	2	1	2	5	15	2	6	6	2	0	0	0	0
3710	2	1	3	4	5	8	0	12	1	1	0	0	0	0
18596	4	1	1	2	3	0	1	12	0	-1	0	0	0	0
16862	2	1	1	4	4	0	9	11	9	5	0	0	0	0
8727	1	2	2	1	4	11	8	15	6	5	0	1	2	0
16071	1	1	1	2	4	10	3	12	9	4	0	0	0	0
24612	2	2	1	1	2	0	0	10	0	-3	0	0	0	0
3250	4	1	1	1	5	8	8	9	0	5	0	0	0	0

This project made use exclusively of various All Media and Products Surveys (AMPS) that required substantial wrangling to prepare for analyses. The data used in this study were accessed through DataFirst

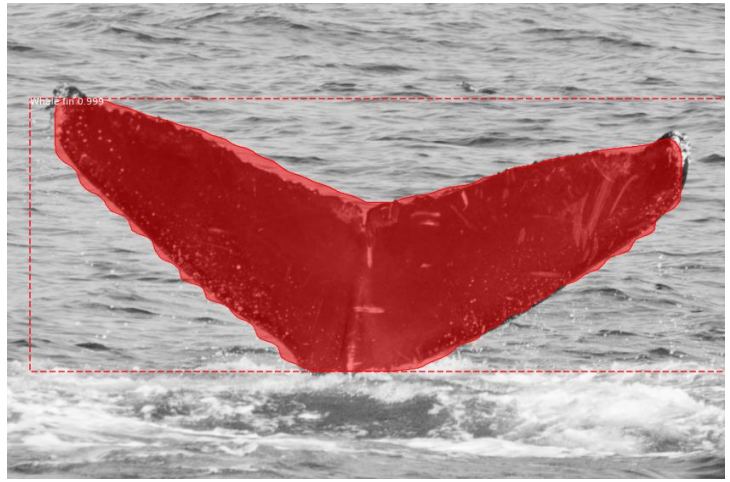
The R code is available in Github:

- https://raw.githubusercontent.com/hanspeter6



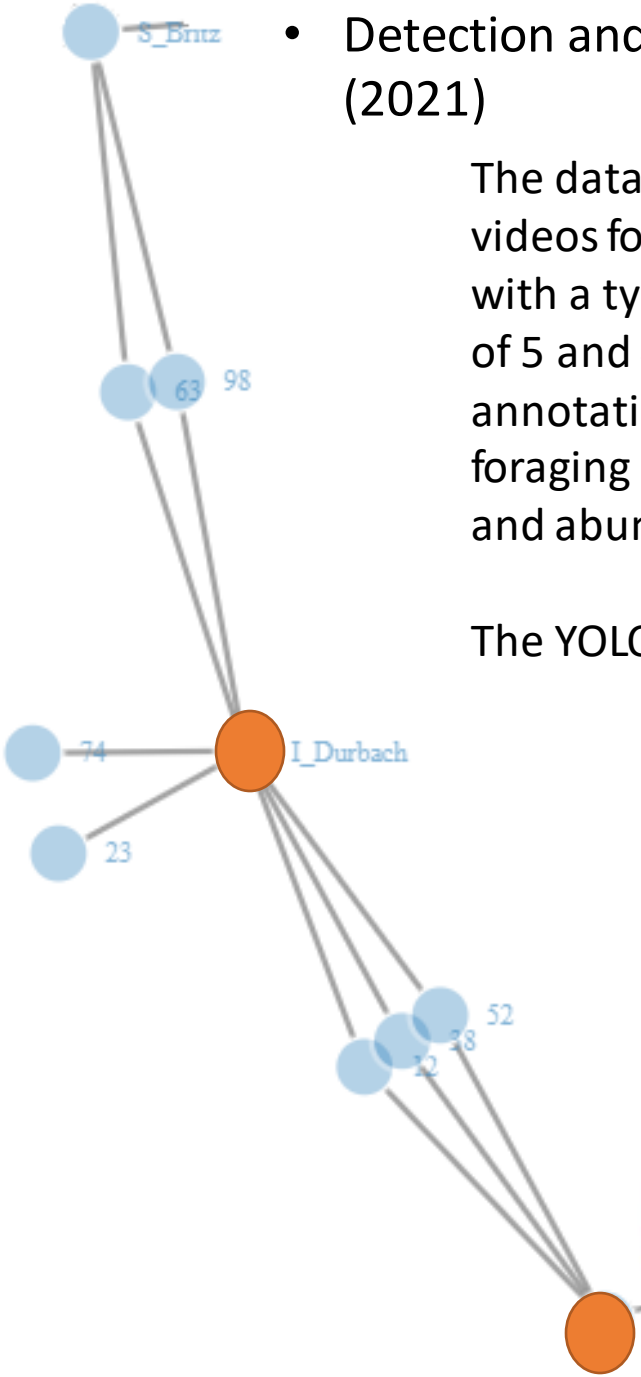
- Using Siamese neural networks to **identify individual animals** (2022)

This project uses two datasets, (1) humpback whales and (2) bottlenose dolphins. The humpback whale dataset contains 5,004 individuals in 25,362 images and comes from Happywhale's Database gathered from research institutions and public contributors [kaggle, 2019].



Siamese neural networks were used and the code used for the analyses in this chapter is available at

https://github.com/TinoMadzingira/Msc. Individual Matching SNN



- Detection and Isolation of Prey Capture Events in **Animal-Borne Images** (2021)

The dataset consists of video footage from animal-borne cameras, mounted on little penguins. The videos focused on the foraging behaviour of the penguins. These amounted to 95 hours of footage, with a typical video being 20 minutes long. The data came from 21 different penguins over a period of 5 and 4 days in October and November 2016 respectively. The data was also accompanied with an annotation file, that was manually complied by a subject expert. The annotation file contains foraging information such things as the type, abundance and capture of prey as well as the presence and abundance of conspecifics or heterospecifics (the presences of other penguins or birds).

The YOLO model was used for the object detection problem

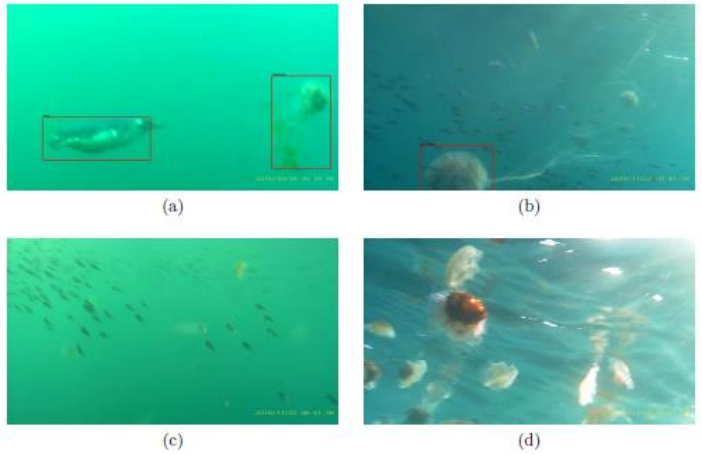
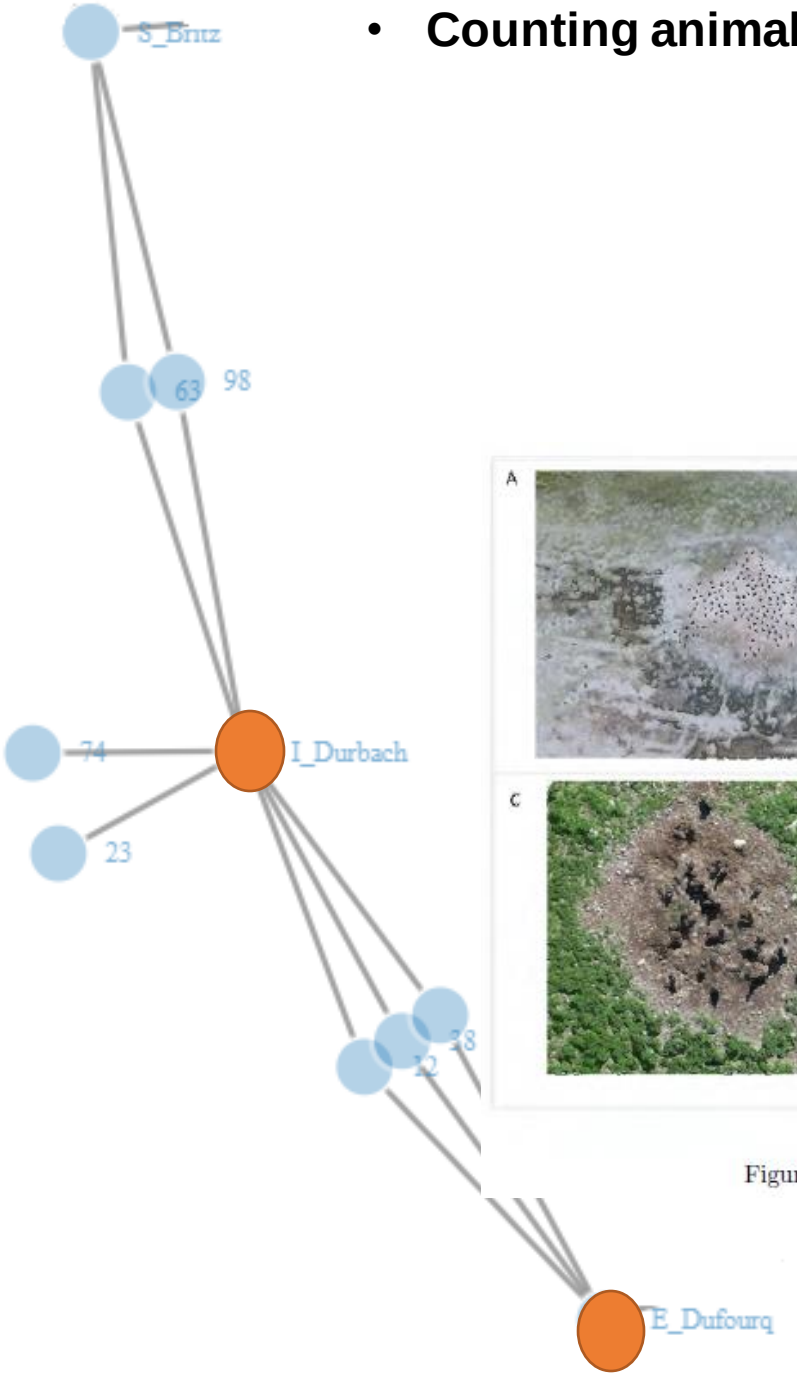


Figure 5.5: Images that the isolation model was unable to identify correctly. Frame (a) contains a jellyfish and penguin which the model labeled as a fish. Similarly, in frame (b) the model identified the jellyfish, however, missed the school of fish that was also present. While no detection were made in frames (c) and (d) though the images contains a school of fish and a group of jellyfish respectively.

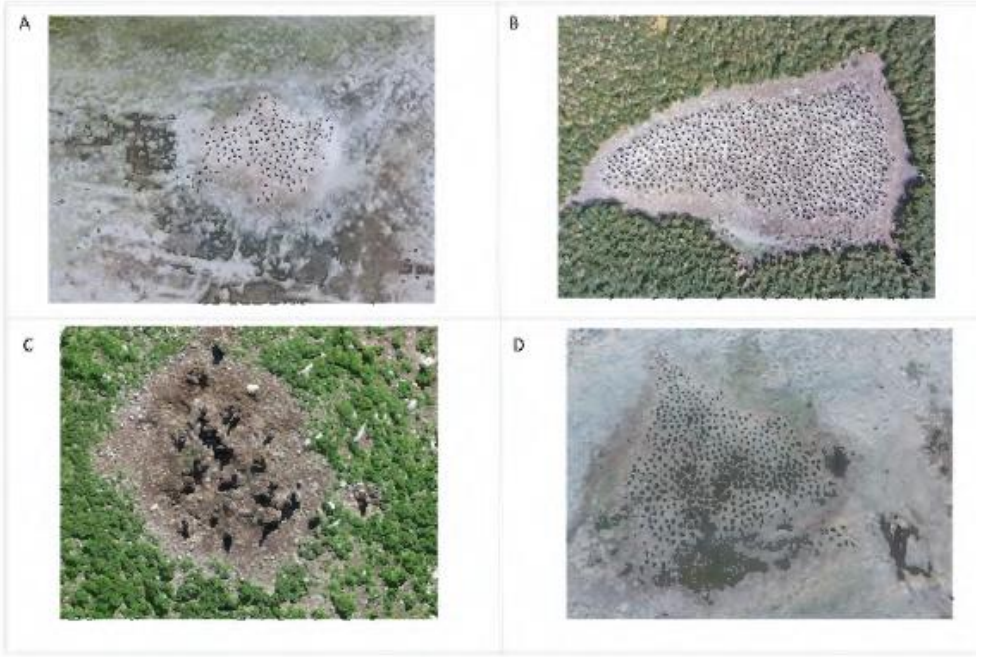
The code available in the GitHub repository link:
https://github.com/Temweka/Isolation_and_Detection_Prey

	Train	Validation	Test	Total
Jellyfish	1286	359	741	2386
Schools of Fish	983	90	317	1390



- **Counting animals** in ecological images (2022)

An opportunistic dataset of 4 aerial images of seabird colonies were obtained from drone surveys, specifically, Gentoo colonies across different breeding sites on the Falkland Islands monitored by the Falkland Islands Seabird Monitoring Programme. These images form the basis of the decisions made to develop the image blob counter. YOLOv3 was used.



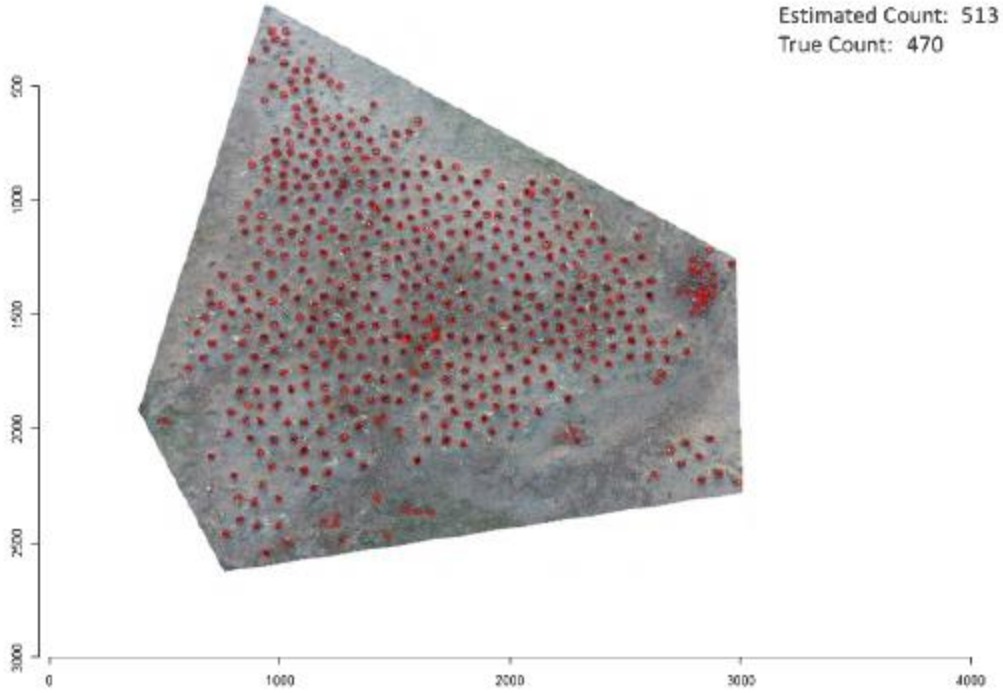


Figure 3.1: Aerial images of seabird colonies.

The code for the blob counter web application and object detection model can be found in the following Github repository: https://github.com/Nakkita/MSc_Data-Science

- Using convolutional neural networks to classify **Hainan gibbon calls** (ongoing)
- Adapting Large-Scale Speaker-Independent **Automatic Speech Recognition** to Dysarthric Speech (2022)

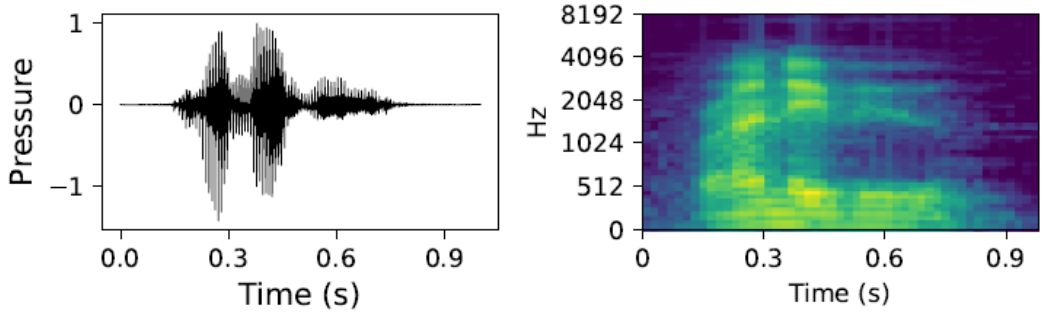


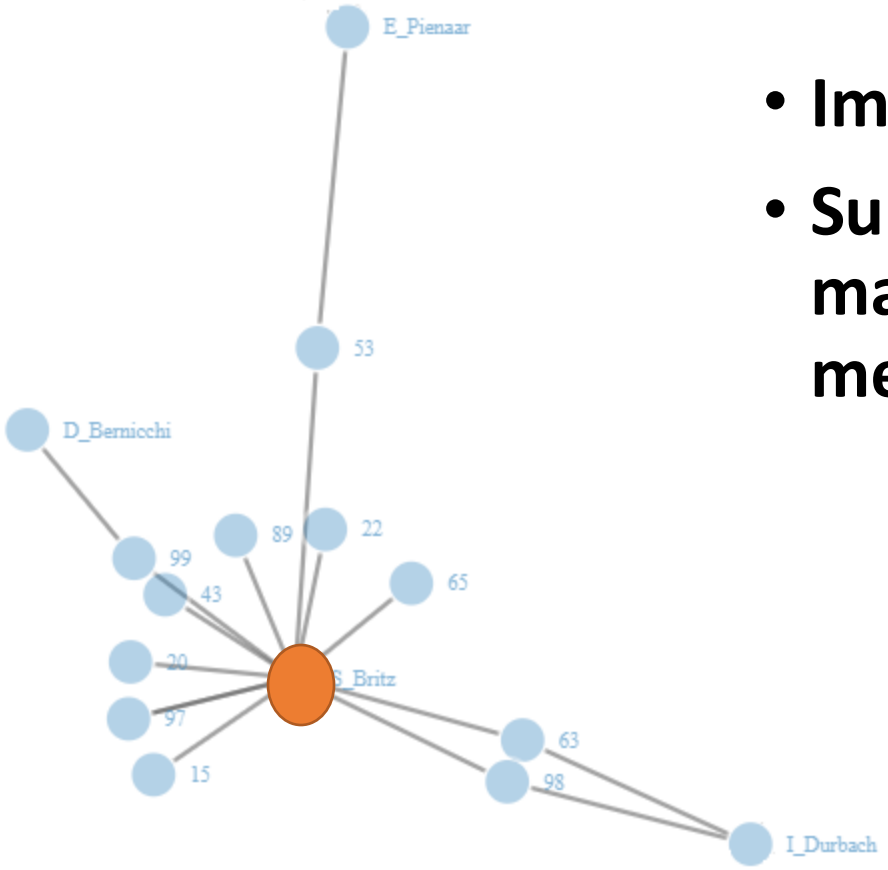
Figure 3.8: Raw waveform and resultant log Mel spectrogram of speaker M10 (high intelligibility) uttering "rendezvous".

Dysarthria is a collective name for such neurogenic speech disorders

The Universal Access Research Speech Database (Kim et al., 2008), or UASpeech for short, is comprised of dysarthric speech contributed by 15 speakers with cerebral palsy. 11 of the speakers have spastic cerebral palsy, two mixed and two unknown. Given that the majority of speakers have spastic dysarthria, the speech is characteristically strained, hoarse, hyper-nasal and slow, with imprecise articulation (Enderby, 2013).

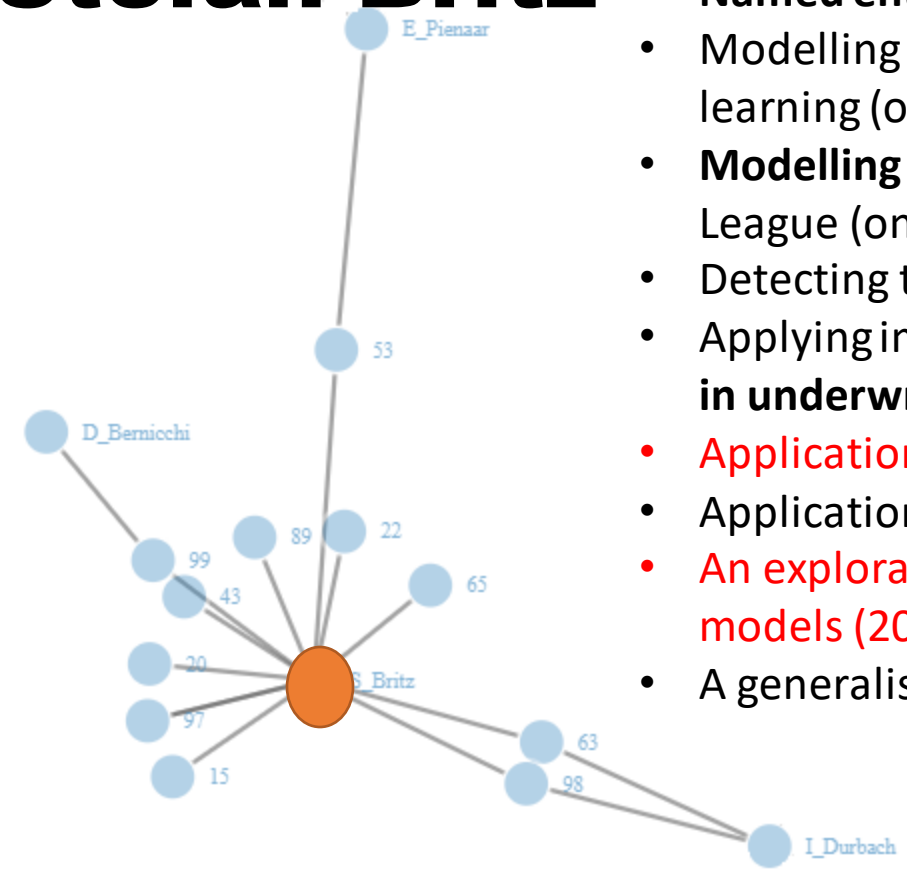
A seven-channel microphone array was used to capture the audio. DeepSpeech, RNNs were utilised.
The code for this thesis can be found at the following GitHub repository:
https://github.com/CharlesRHouston/dysarthria-project

Stefan Britz

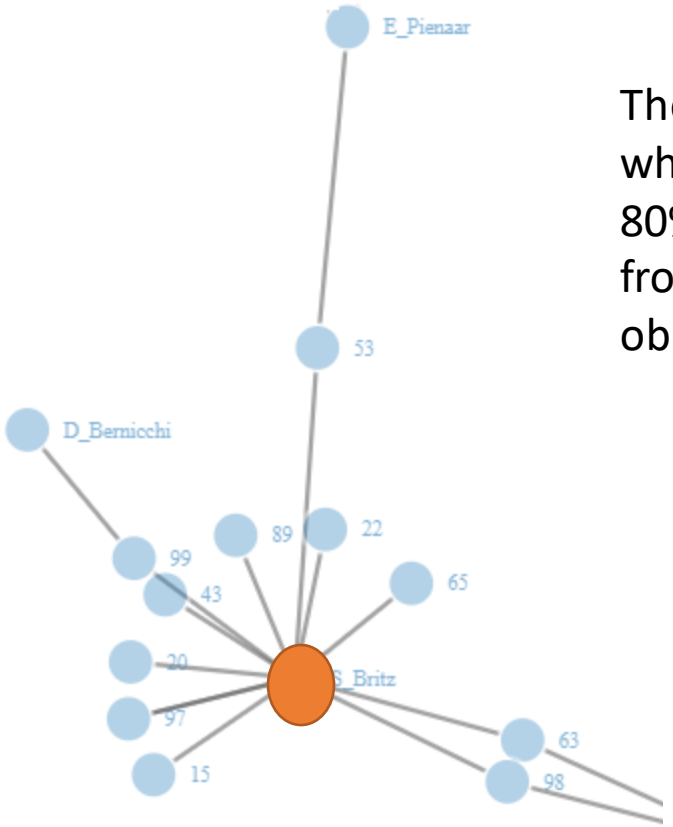


- Image analysis
- Supervised, machine learning methods
- Ecology
- Finance (fraud detection)
- Agriculture

Stefan Britz

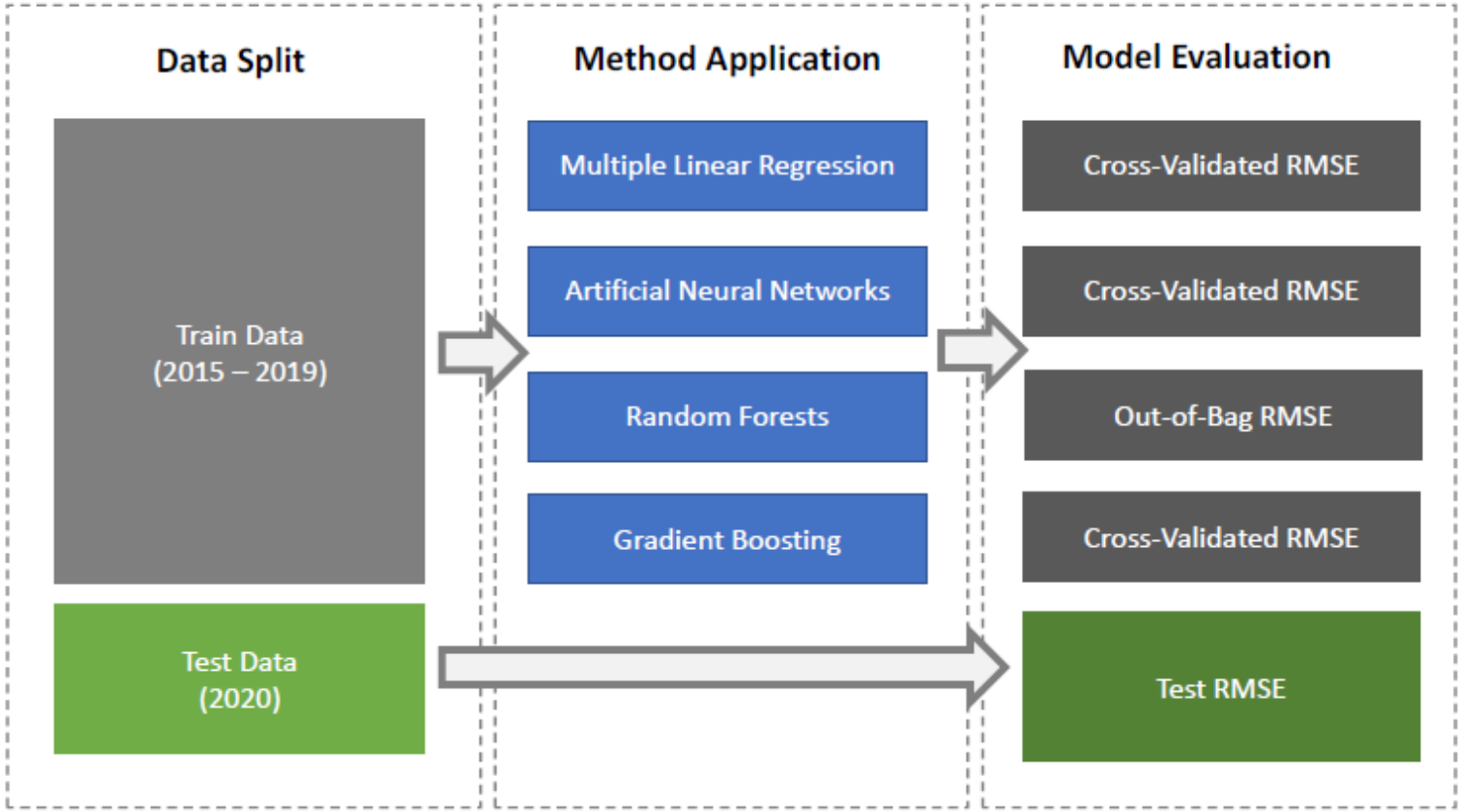


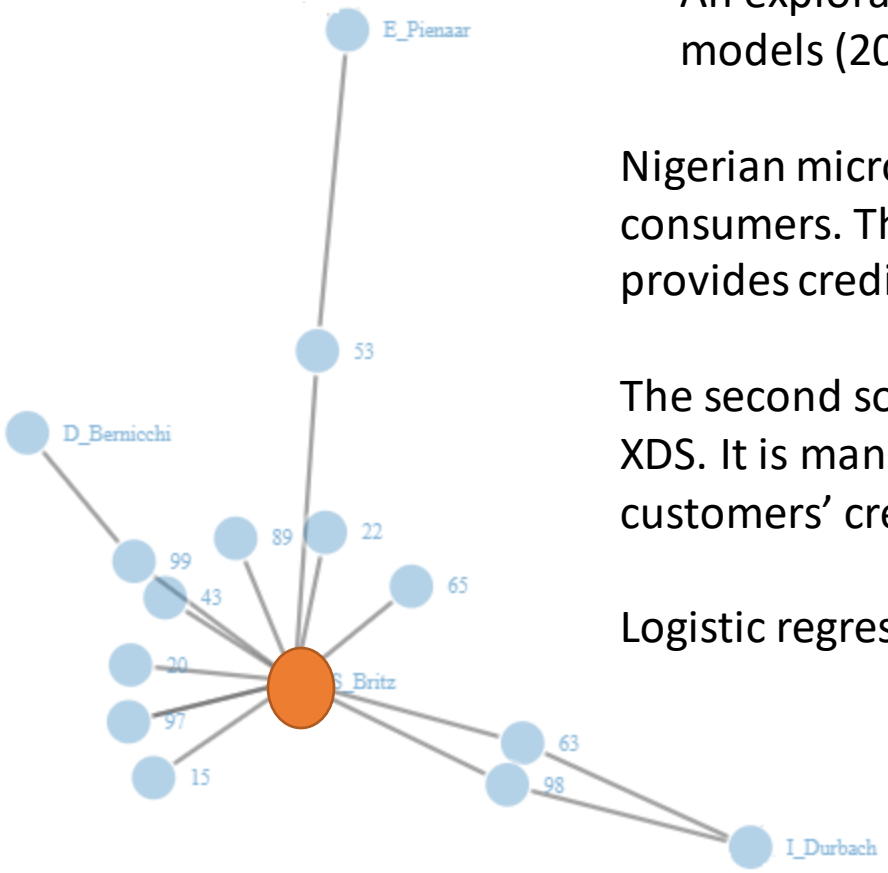
- **Named entity recognition** using neural networks (ongoing)
- Modelling highly imbalanced credit card **fraud detection** data using statistical learning (ongoing)
- **Modelling first innings** totals in T20 cricket: applications in the Indian Premier League (ongoing)
- Detecting the **possibility of fraud** based on claim **transactional** data (ongoing)
- Applying imputation and statistical learning to **predict gamma-glutamyl transferase in underwriting data** (ongoing)
- Applications of Machine Learning in **Apple Crop Yield Prediction** (2021)
- Application of gcForests in **cassava leaf image classification** (ongoing)
- An exploration of alternative features in **micro-finance loan default prediction** models (2020)
- A generalised approach to **body-aware virtual try-on** (ongoing)



- Applications of Machine Learning in **Apple Crop Yield Prediction** (2021)

The data were collected from farmers in theWestern Cape province of South Africa, which is the main apple producing region in the country contributing to approximately 80% of South Africa's total apple production. The research uses apple production data from 2016 until 2020. The dataset includes data from 745 orchards and there are 2800 observations in total.





- An exploration of alternative features in **micro-finance loan default prediction** models (2020)

Nigerian micro-finance institution that has disbursed loans to more than 250,000 consumers. The institution is an application (app) based lender and currently only provides credit to android users

The second source of data for this project was the Nigerian credit bureaus CRC, CRS and XDS. It is mandatory for credit providing institutions in Nigeria to submit their customers' credit performance data to these credit bureaus

Logistic regression, random forests, extreme gradient boosting, neural networks

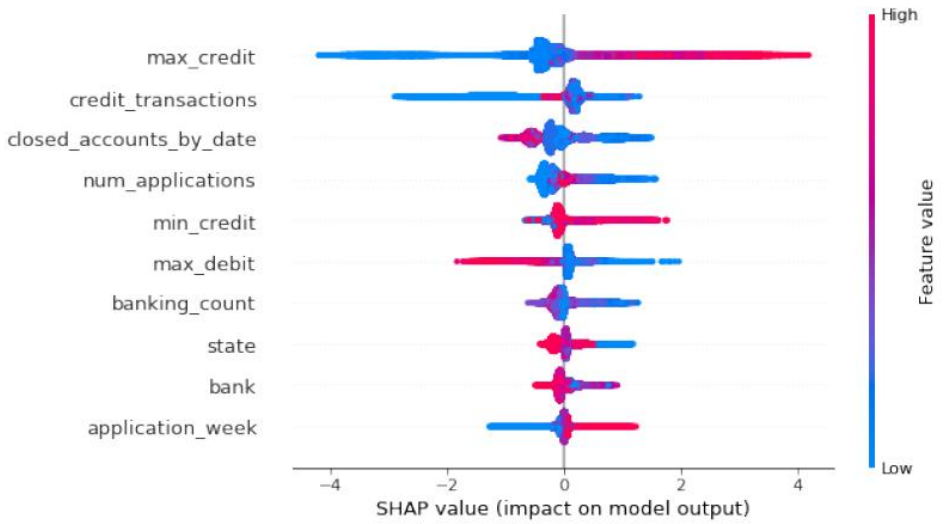
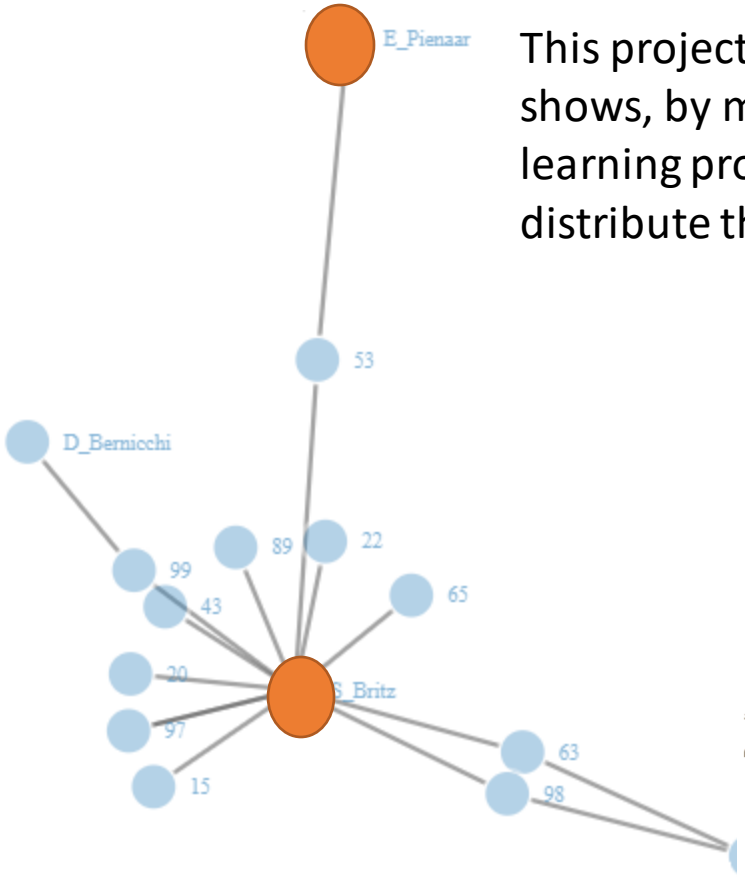


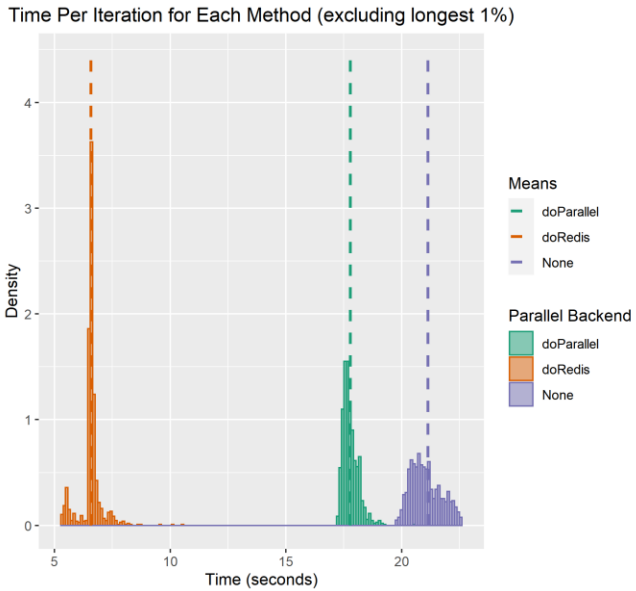
FIGURE 5.7: SHAP Values of Best Performing XGBoost Model

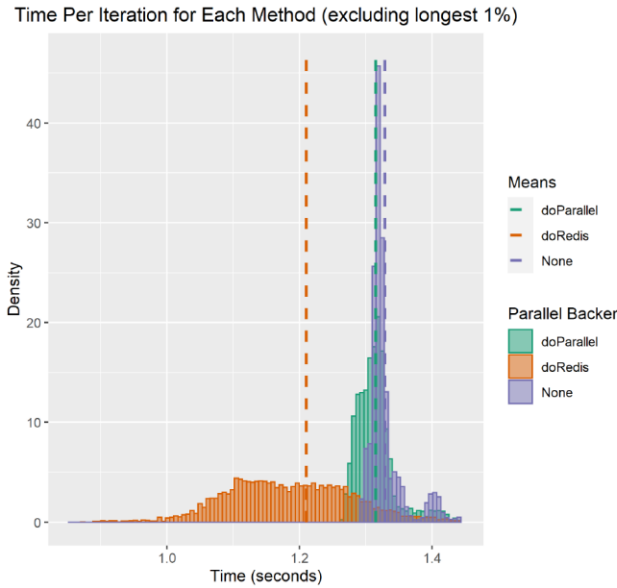
- **Small-scale distributed machine learning** in R



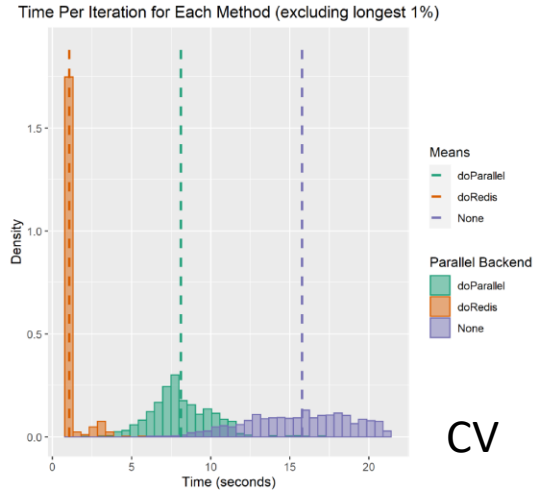
This project describes core features of the doRedis package and shows, by means of applying certain aspects of the machine learning process, that it is both viable and beneficial to distribute the machine learning aspects.

Random forests



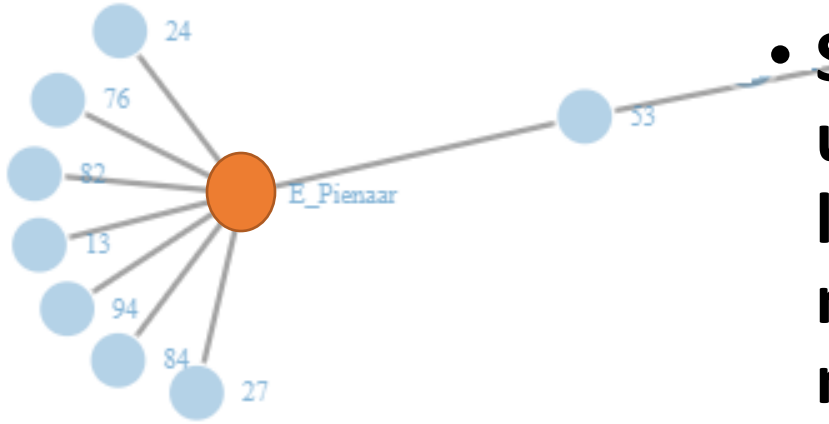


T-test

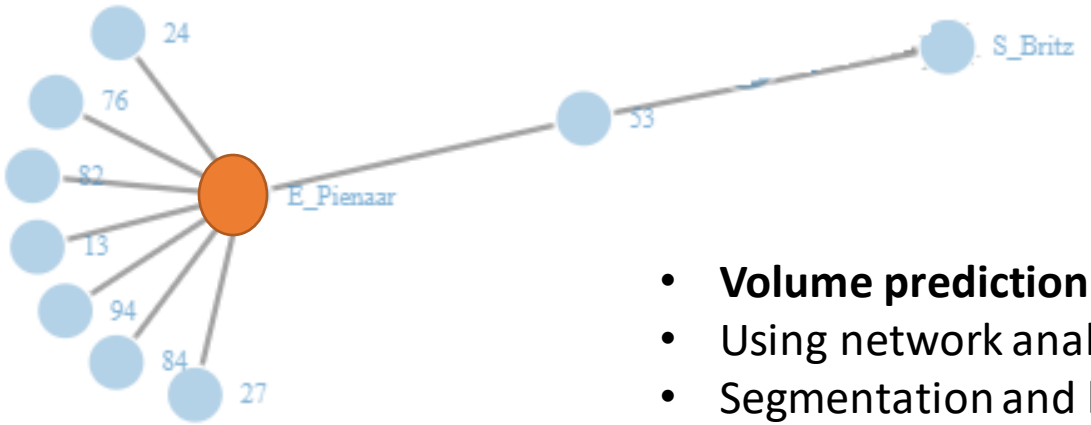


CV

Etienne Pienaar

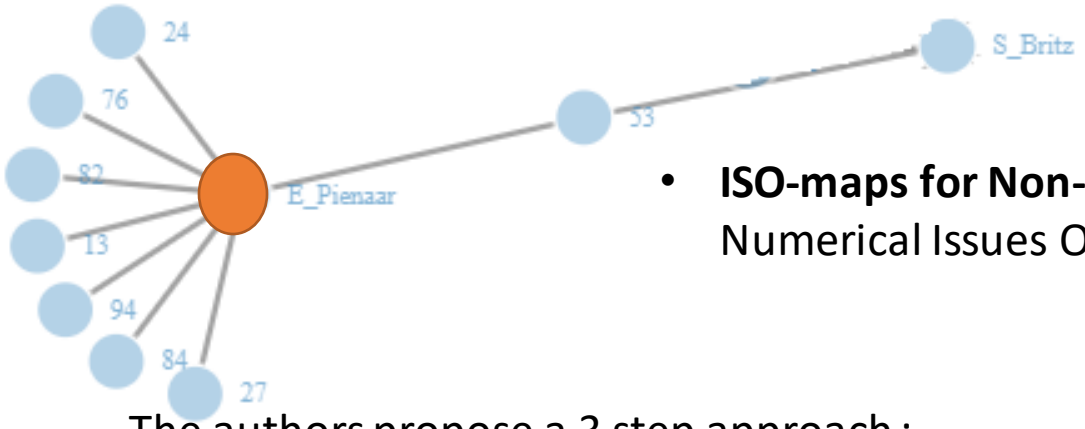


- Image analysis
- Supervised, unsupervised learning and machine learning methods
- Finance (fraud detection)
- Insurance
- Industry



Etienne Pienaar

- **Volume prediction** when a mobile data site is upgraded to 4G (ongoing)
- Using network analysis for **fraud prevention** (ongoing)
- Segmentation and localization of product attributes of **retail fashion images** using CNNs (ongoing)
- **Prediction of a priority ranking service level index** (SLI) on a mobile telephone network using machine learning and analysis of data from passive probing and network elements (ongoing)
- **ISO-maps for Non-linear Dimension Reduction** - Addressing Geometric and Numerical Issues Observed in Practice. (2022)
- Exploring applications of **tree-based ensemble methods** in managing **client persistency in the life insurance industry** (ongoing)
- Development of a test suite for **single object tracking algorithms in video** (2021)



- **ISO-maps for Non-linear Dimension Reduction** - Addressing Geometric and Numerical Issues Observed in Practice. (2022)

The authors propose a 3 step approach :

Step 1 - constructing a distance graph by defining the neighbours of each data point such that it is contained within one section of the manifold,

Step 2 - approximating the geodesic distances between all points using the graph constructed in Step 1,

Step 3 - applying Classical MDS to this distance graph to achieve dimensionality reduction.

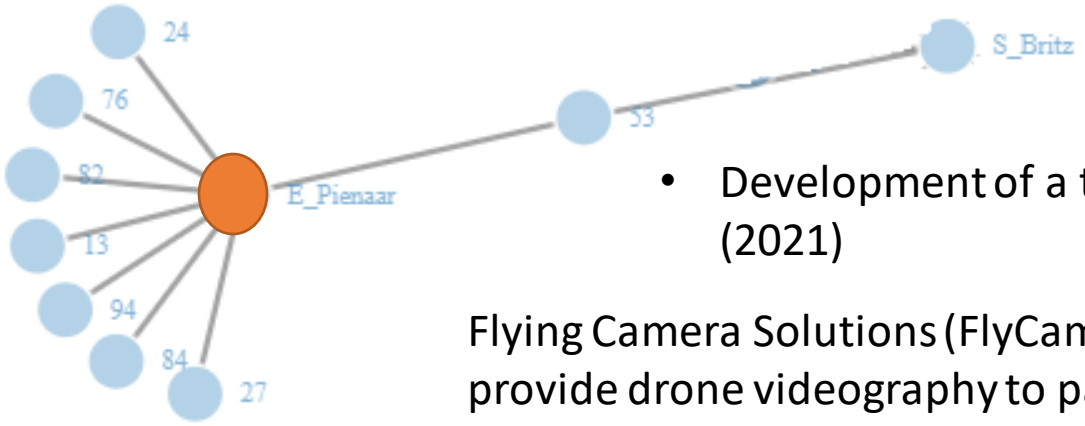
Applied to Swiss Roll, MNIST, 3 Kaggle datasets (Gesture, Churn, Cancer).

TABLE 7.3: Table showing the train and test accuracy when applying ISOmap, remediation and any-path algorithm to the **CHURN** dataset - SVM Linear Kernel

Dataset	Train Accuracy	Test Accuracy	Train Improvement	Test Improvement
Original dataset	87%	90%		
PCA on Original dataset	87%	90%		
PCA on Geodesic distances (baseline)	87%	90%	0 ppts	0 ppts
PCA with Geodesic distances of Remediated dataset	87%	90%	0 ppts	0 ppts
PCA with Geodesic distances of Remediated dataset and Any Path algorithm	87%	90%	0 ppts	0 ppts

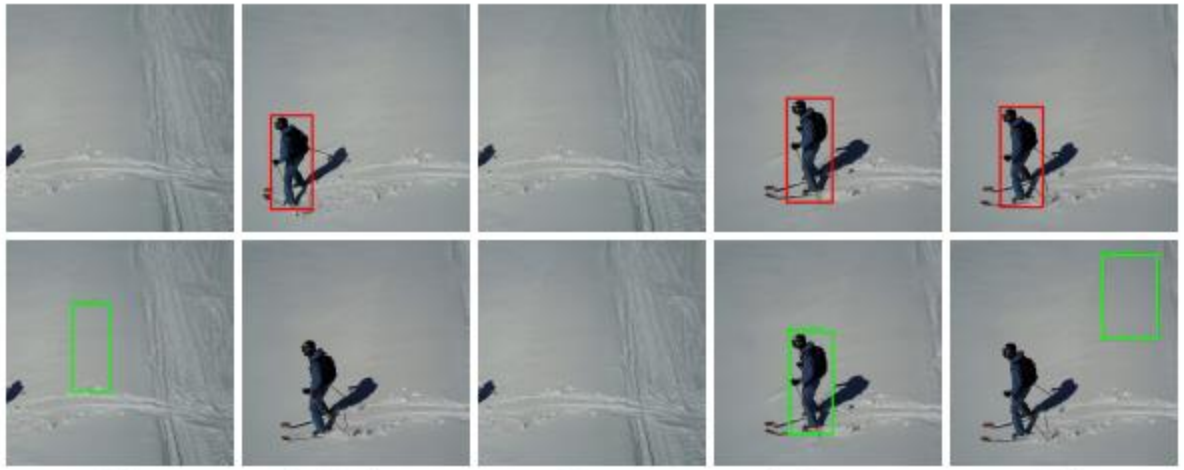
TABLE 7.4: Table showing the train and test accuracy when applying ISOmap, remediation and any-path algorithm to the **CANCER** dataset

Dataset	Train Accuracy	Test Accuracy	Train Improvement	Test Improvement
Original dataset	91%	87%		
PCA on Original dataset	91%	88%		
PCA on Geodesic distances (baseline)	76%	55%	-15 ppts	-33 ppts
PCA with Geodesic distances of Remediated dataset	71%	56%	-5 ppts	+1 ppt
PCA with Geodesic distances of Remediated dataset and Any Path algorithm	68%	56%	-8 ppts	+1 ppt



- Development of a test suite for **single object tracking algorithms in video** (2021)

Flying Camera Solutions (FlyCam), within Sony Lund's startup accelerator, intends to provide drone videography to paying customers in ski resorts: a customer should be able to go about their activity as usual while a drone films them. Visual object tracking, enabling the drone to track the customer throughout the activity, is a primary obstacle in creating a viable autonomous videography service



a. Failure (false positive) b. Failure (false negative: miss) c. Success (true negative) d. Success (true positive) e. Failure (false positive: exceeds threshold)

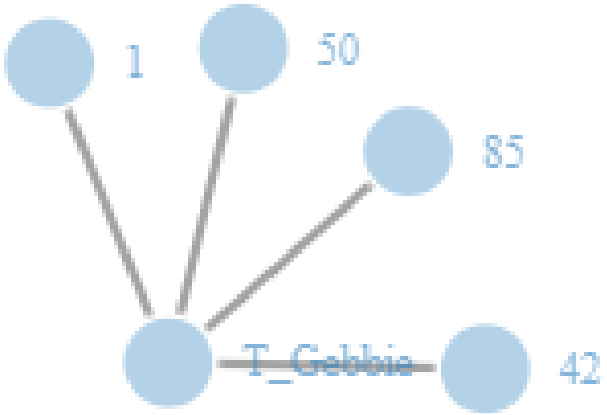
Ten performance metrics were adapted from multi-object to single-object tracking.

Nine tracking algorithms were then run on each of the 18 test video clips at varying resolutions to produce 375 tracking observations for analysis.

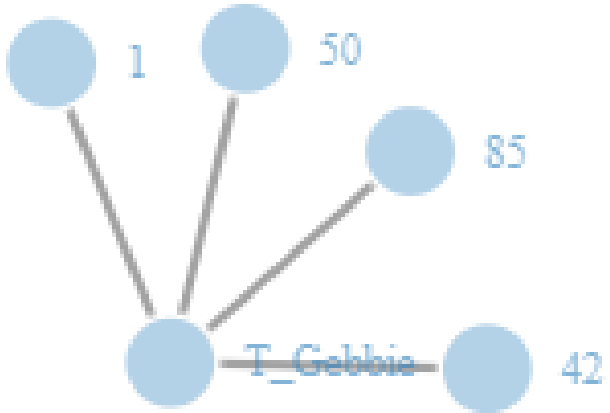
The evaluation results revealed the optimal tracking algorithm to be Re3: a recurrent-convolutional neural network tracker which runs at respectable speeds on a consumer laptop.

[Minor Dissertation Presentation | MSc. Data Science (2019) - YouTube]

Tim Gebbie



- Supervised, unsupervised learning and machine learning
- Portfolio theory
- Financial markets



Tim Gebbie

- **Sentiment Analysis** in Financial Markets: A South african perspective
- **Portfolio Construction** using Cleaned Correlation Matrices
- **Market State** Discovery (2022)
- A Reproducible Approach to **Equity Backtesting** (2019)

Research findings relating to anomalous equity returns should ideally be repeatable by others. Usually, only a small subset of the decisions made in a particular backtest workflow are released, which limits **reproducability**. **Data collection and cleaning, parameter setting, algorithm development and report generation** are often done with manual point-and-click tools which do not log user actions. This problem is compounded by the fact that the trial-and-error approach of researchers increases the probability of **backtest overfitting**.

The research introduces a set of scripts that **completely automate a portfolio-based, event-driven backtest**. Based on free, open source tools, these scripts can completely capture the decisions made by a researcher, resulting in a distributable code package that allows easy reproduction of results.

All of these tests were conducted on the lag-correction branch of https://github.com/riazarbi between the 1st and 5th February 2019.

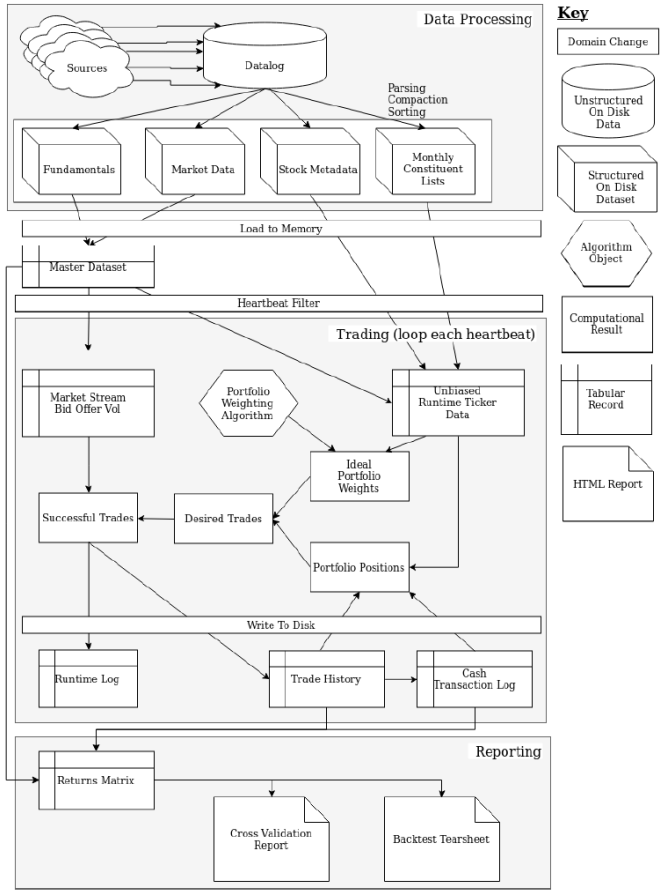
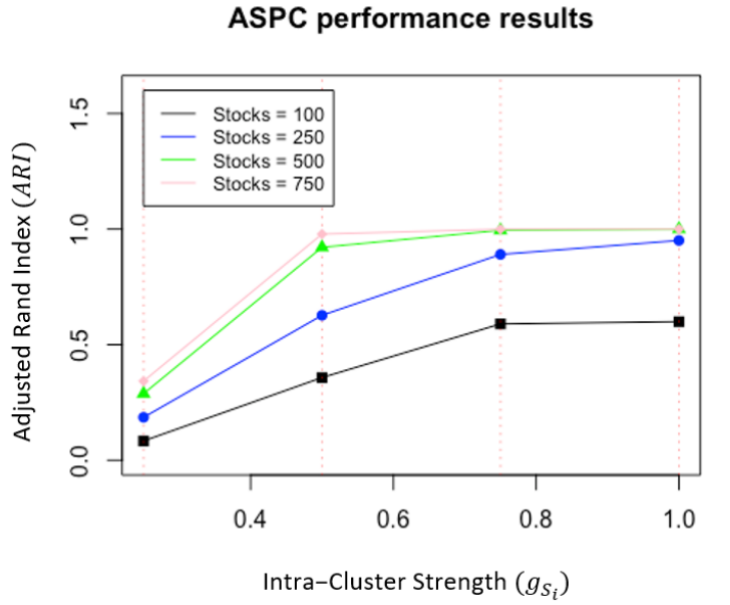
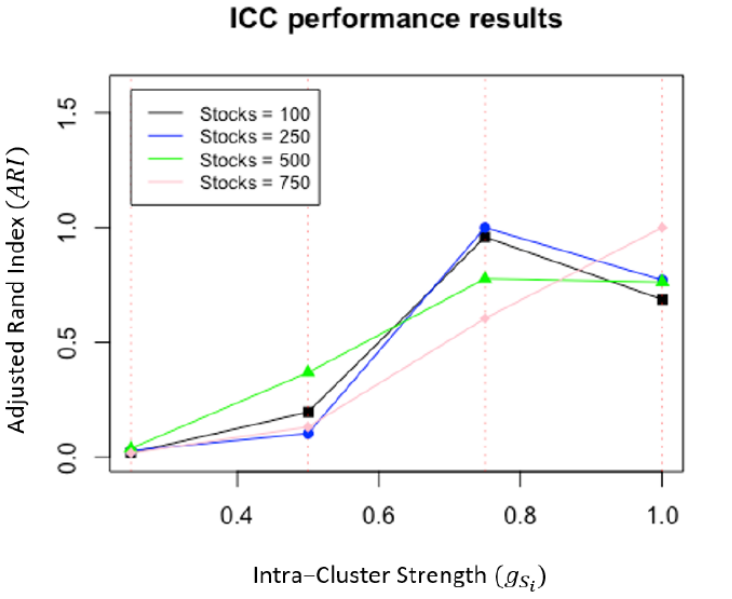


Figure 5.1: Data Flow Diagram



e 24: ASPC Stress Case (3-State Market) Results (Section 4.1.3.1):



23: ICC Stress Case (3-State) Market Results (Section 4.1.3.1)

Tim Gebbie

- **Market State** Discovery (2022)

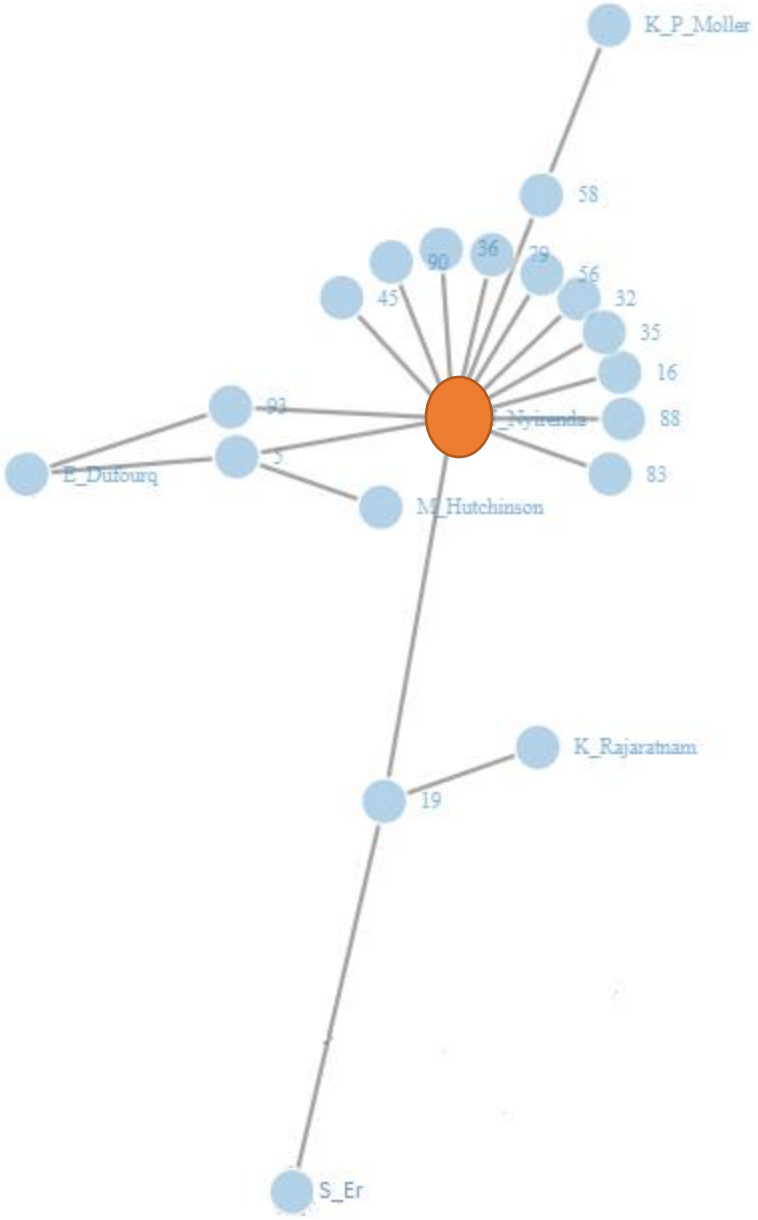
The concept of financial market state discovery is explored by assessing the robustness of **two unsupervised machine learning algorithms**: **Inverse Covariance Clustering (ICC)** and **Agglomerative Super Paramagnetic Clustering (ASPC)** trying to address three key issues.

First, popular algorithms require pre-selection of the number of states to be estimated.
Second, distributional assumptions are often imposed on the price dynamics.
Third, many algorithms lack the ability to handle high dimensional datasets without compromising on computationally efficiency and performance.

Inverse Covariance Clustering (ICC) and **Agglomerative Super-Paramagnetic Clustering (ASPC)** are two algorithms of interest that attempt to address these problems directly.

The assessment is carried out by: **simulating market datasets** varying in complexity; implementing ICC and ASPC to estimate the underlying states (using only **simulated log-returns as inputs**); and measuring the algorithms' ability to recover the underlying states, using the **Adjusted Rand Index (ARI)** as a performance metric.

A real life application is also carried out.

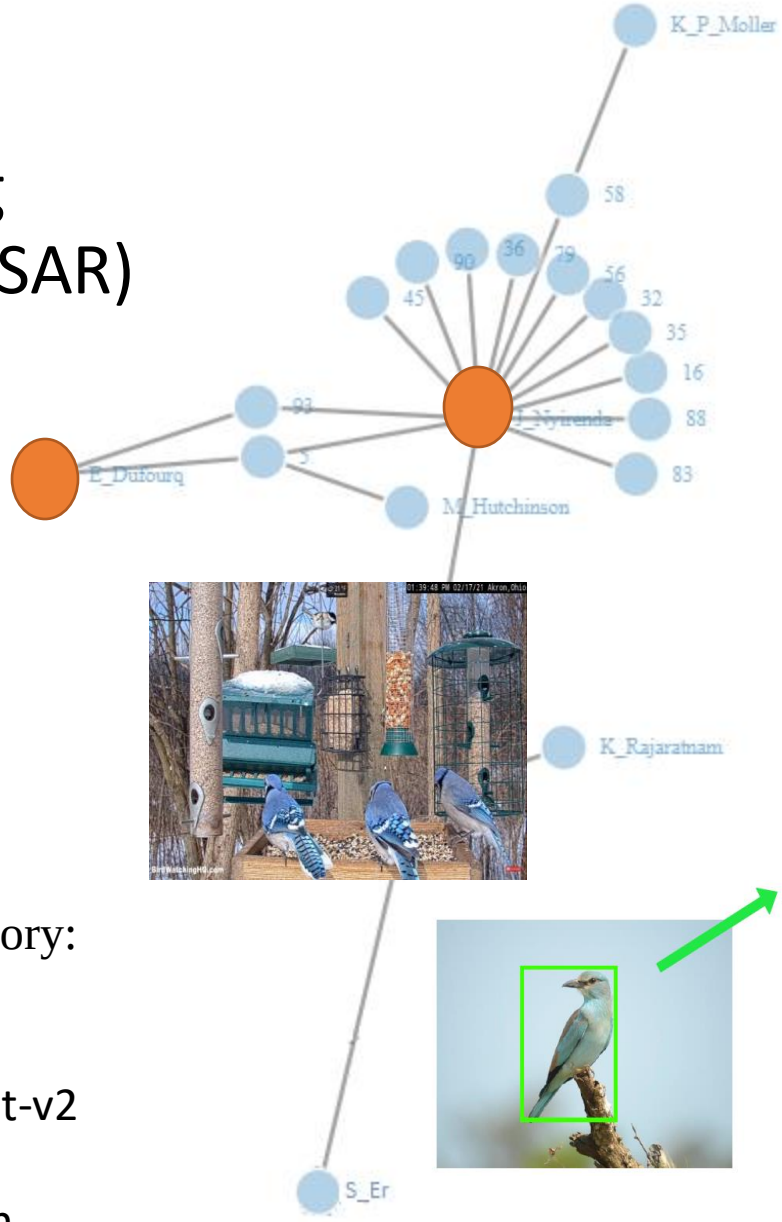


Juwa Nyirenda

- NLP
- Image analysis
- Time series analysis
- Ecology
- Sustainability
- Social sciences

Juwa Nyirenda

- Water Body Detection and Classification Using Satellite **Imagery** and Surface Apeture Radar (SAR) in Ohio (ongoing)
- Investigating **automated bird detection from webcams** using machine learning (2022)

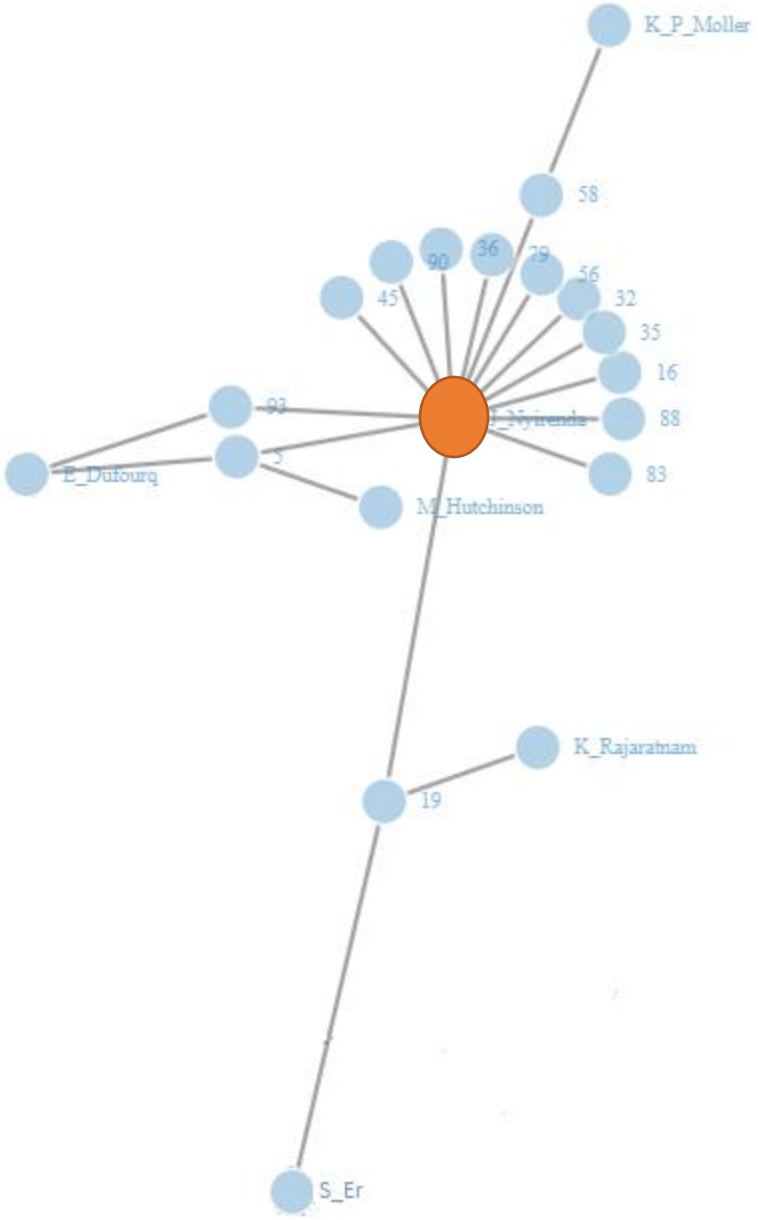


The dataset was made publicly available on *Zenodo* (https://zenodo.org/record/5172214#.YRMYdYgzZhF).

The student (Alex Mirugwe) has all the code available in a Github repository: available at https://github.com/mirugwe1/bird_detection.git

Faster R-CNN, SSD, ResNet-50, ResNet-101, ResNet-152, Inception ResNet-v2

Work has been published in Proceedings of 43rd Conference of the South African Institute of Computer Scientists and Information Technologists



- **Wineinformatics**: Exploring the Determinants of Wine Quality using Sensory Data Reviews and Machine Learning Techniques
- Video **Sentiment** Analysis
- Toward a sustainable energy future: Peak load shaving in commercial properties to reduce cost of energy (2022)
- Predicting social unrest events in South Africa using **LSTM** neural networks (2021)
- Exploring the Application of **Word2Vec to Basket Transaction Data** in the Grocery Retail Industry (2022)
- Exploring the application of **Natural Language Processing** to scientific medical cannabis publications (2022)
- Automated monitoring of **informal settlements** using neural networks
- A **machine learning model** for octane number prediction

- Toward a sustainable energy future: Peak load shaving in commercial properties to reduce cost of energy (2022)

In South Africa, commercial properties are billed per kWh and can incur an additional demand charge that often accounts for a substantial portion of the energy bill, depending on the load factor.

This project investigates **peak load shaving** as a solution for commercial properties to **reduce their cost of electricity while supporting the transition to a greener energy future**.

This project introduces **a novel approach** that employs **clustering** the energy demand profile shapes and training separate learning agents to target specific demand shapes, thereby reducing the complexity of the problem presented to the individual agents.

The reinforcement learning model was trained on historical data from a commercial shopping centre in Cape Town using a hypothetical battery.

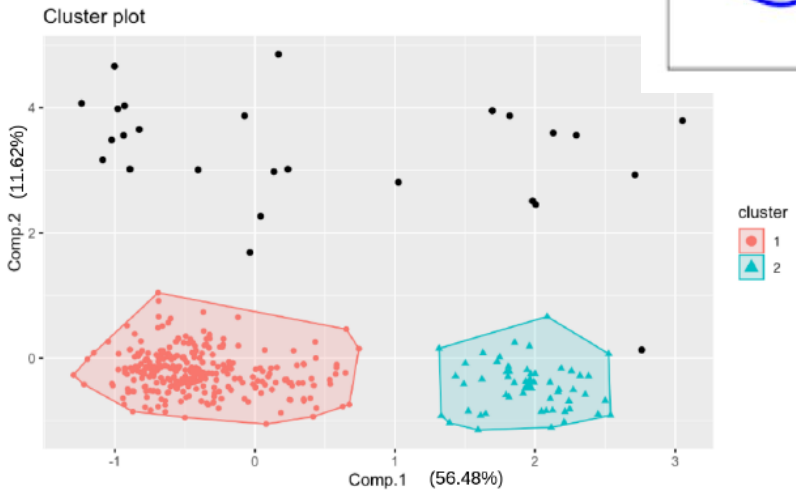


Figure 13: DBSCAN output: profiles plotted across the first two principal components. Two distinct clusters dots indicating the energy demand noise.

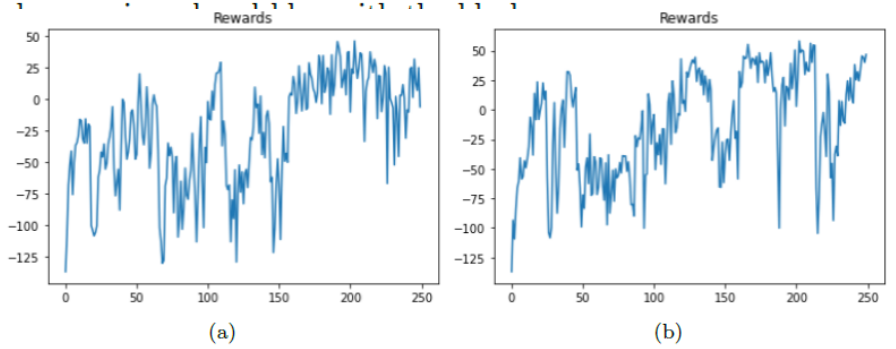


Figure 22: Cluster 1 reinforcement learning agent's undiscounted accumulated rewards averaged over 50 episode iterations: Subfigure (a) represents the agent trained without temperature and Subfigure (b) represents the agent trained with temperature.

- Predicting social unrest events in South Africa using **LSTM** neural networks (2021)

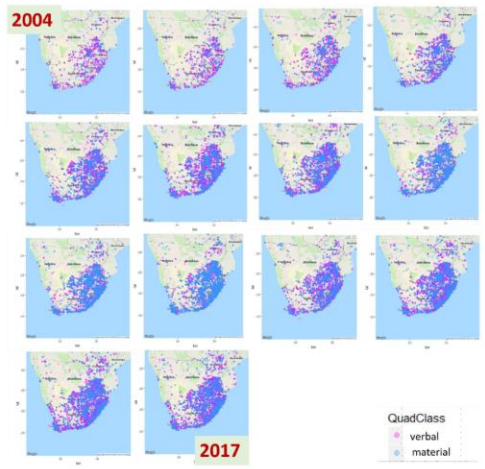


Figure 1.1: From left to right unrest events in South Africa 2004-2017.

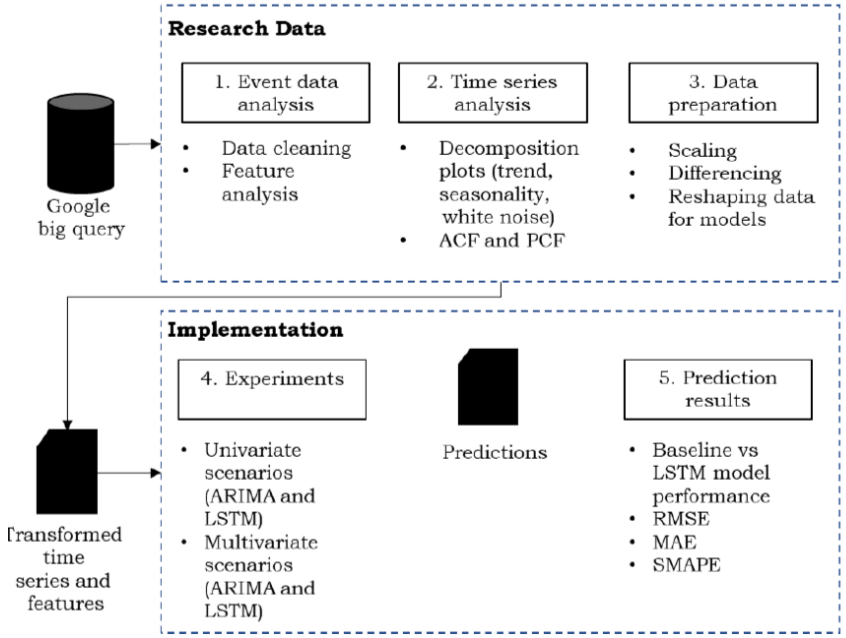


Figure 3.24: The proposed research design framework.

GDELT monitors news media across the world to identify events and currently contains a list of over 300 event categories.

These events include “protests”, “peaceful appeals”, “diplomatic agreements”, “diplomatic apologies”, “public demands”, and reports on threats. Each event record has 61 fields capturing details of the event, including its georeferenced location and actors mentioned. Actors are entities such as individuals, countries, identity groups, religious groups, political parties, and organizations.

This project aims to predict social unrest using the baseline variables in GDELT data used in Smith et al. (2017). These were SQLDATE, GoldsteinScale, NumMentions, NumSources, NumArticles, AvgTone, ActionGeo_Lat, ActionGeo_Long and ActionGeo_CountryCode.

- Exploring the Application of **Word2Vec to Basket Transaction Data** in the Grocery Retail Industry (2022)

Data from a transaction database of a large retailer in South Africa was extracted. The transaction database represents a complete history of information collected through EPOS. The information include: identity of the store, store size, location of store, product, brand, customer details and much more.

The application of Word2vec to basket transaction data was explored with the objective of the exploration is to establish whether the application of Word2vec to basket transaction data would generate product embeddings that represent a useful relationship between products

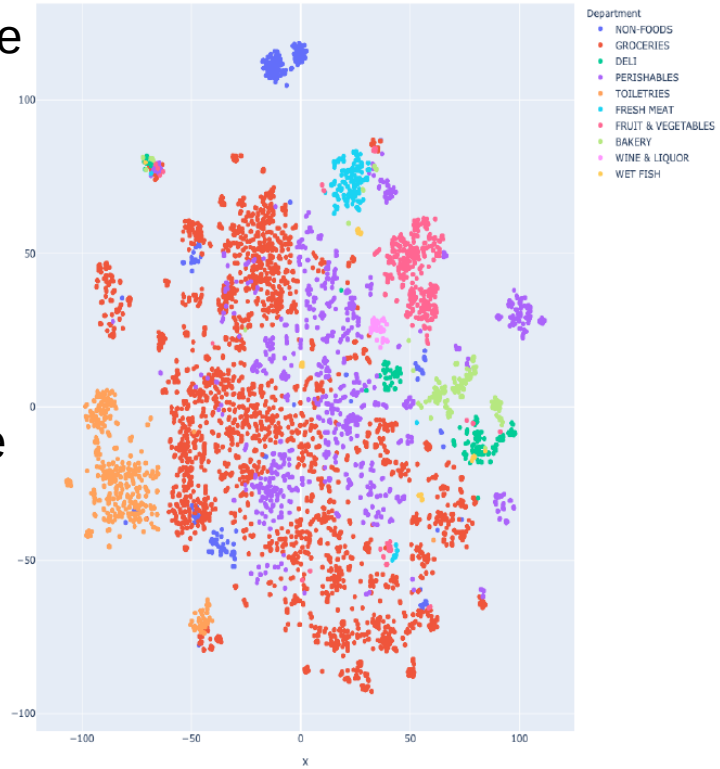


FIGURE 6.2: Scatter plot: Department representation

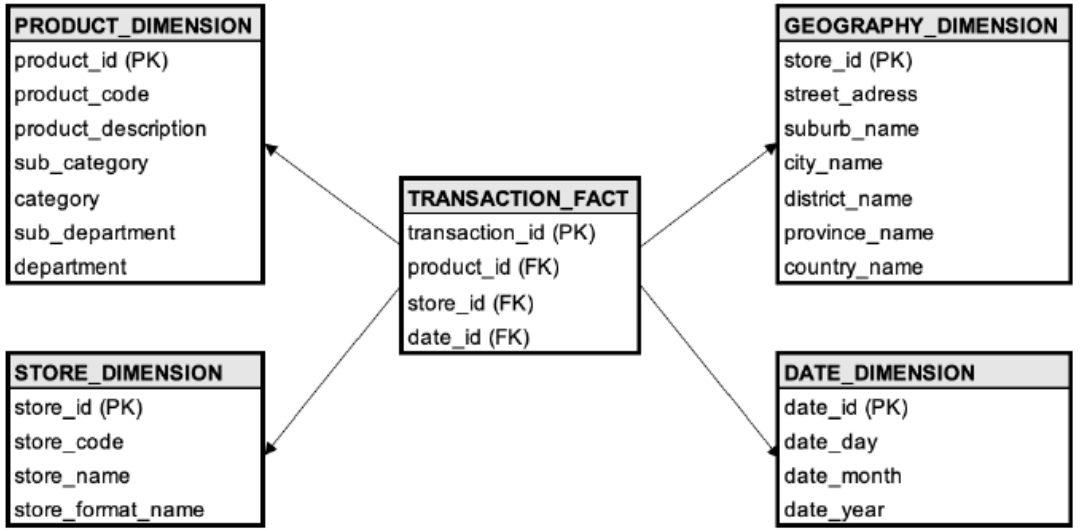


FIGURE 4.1: Database illustration (ER diagram)

- Exploring the application of **Natural Language Processing** to scientific medical cannabis publications (2022)

This project aims to develop an **appropriate method** to extract the key connections **between cannabis compounds, human physiology and disease from the existing medical literature**.

First, **natural language processing** techniques (such as document clustering and topic modelling, global vector word embeddings and supervised document classifiers) are used **to group 500 journal articles from the general literature on cannabis** according to broad research topics; analyse the interaction between cannabis compounds, human physiology and diseases; and train a classifier to classify unseen documents.

Second, the connections generated through this quantitative process are assessed qualitatively against those in a **manual dataset of research findings from more than 500 studies** collated over a number of years and provided by a medical company specialising in cannabis research.

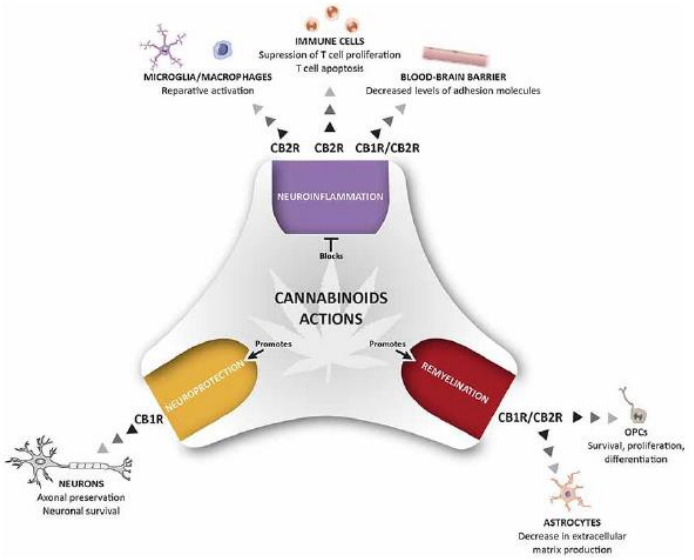
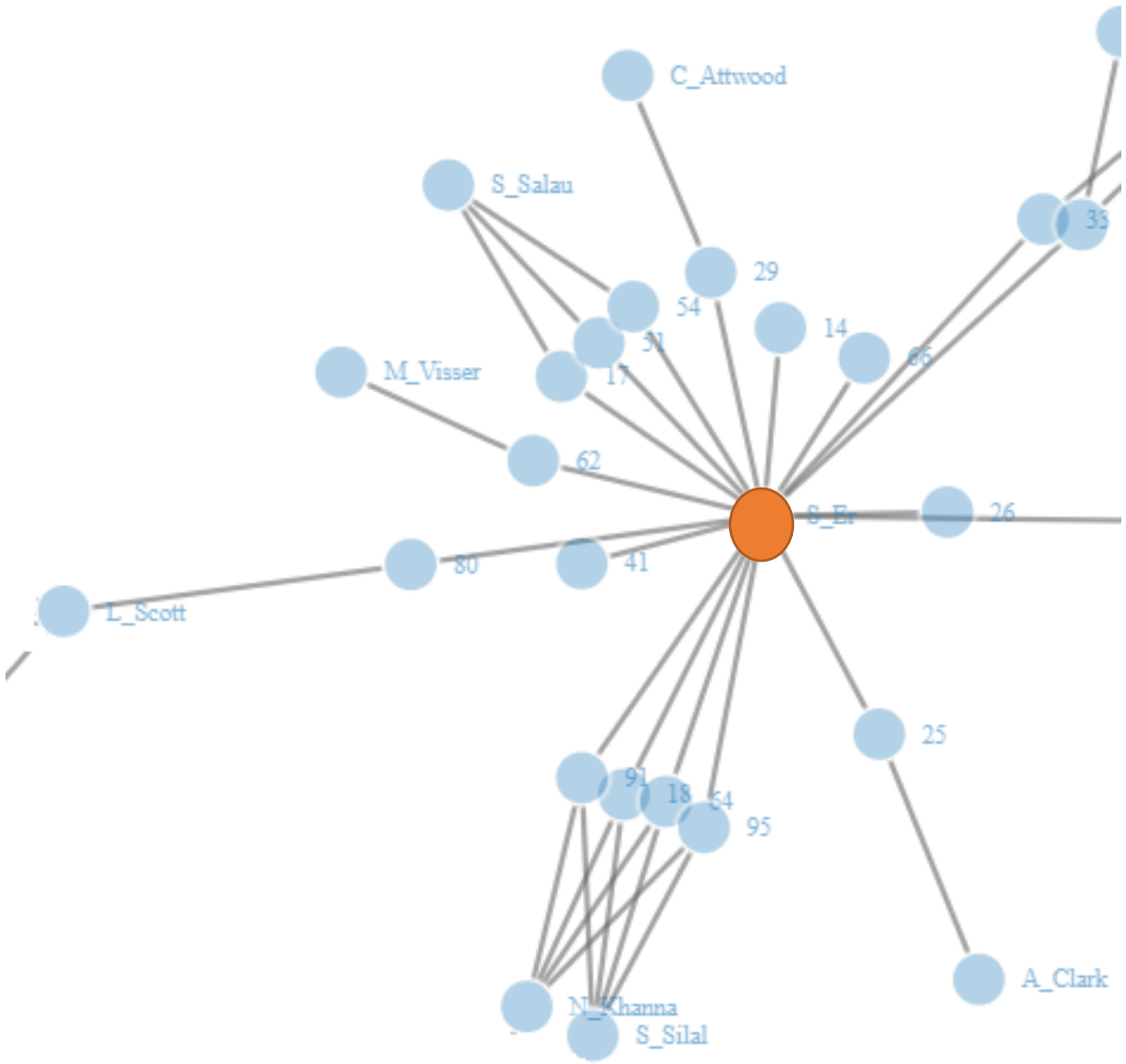


Figure 2: Cannabinoid Primary Actions (Mecha,2018)

Table 4: Sample of Dataset Paper Titles

Title	Author
Cannabidiol, a novel inverse agonist for GPR12	Kevin J. Brown
Beneficial effect of the non-psychotropic plant cannabinoid cannabigerol on experimental inflammatory bowel disease	Francesca Borrelli
Cannabinoids increase lung cancer cell lysis by lymphokine-activated killer cells via upregulation of ICAM-1	Maria Haustein
ACEA (a highly selective cannabinoid CB1 receptor agonist) stimulates hippocampal neurogenesis in mice treated with antiepileptic drugs	Marta Andres-Mach
Alzheimer's disease –mechansims-cause-factors-prevalence	Compaq
Cannabinoid 2 receptor is a novel anti-inflammatory target in experimental proliferative vitreoretinopathy	Anna-Maria Szczesniak
Cannabidiol increases survival and promotes rescue of cognitive function in a murine model of cerebral malaria	A.C. Campos

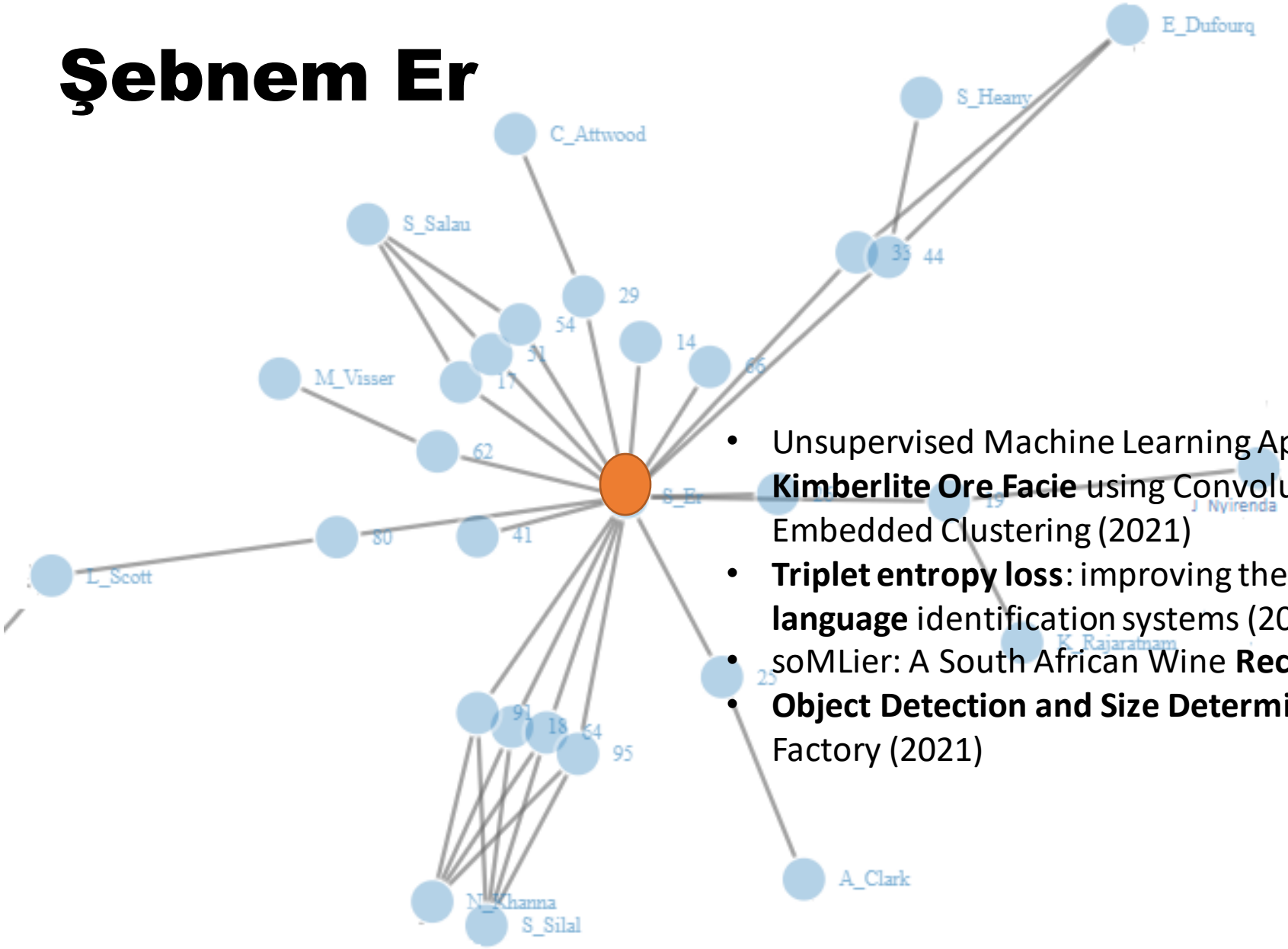


Şebnem Er

- NLP
- Image analysis
- Supervised and unsupervised learning methods
- Ecology/Biology
- Social sciences
- Education
- Health sciences

etc.

Şebnem Er



- Unsupervised Machine Learning Application for the Identification of **Kimberlite Ore Facie** using Convolutional Neural Networks and Deep Embedded Clustering (2021)
- **Triplet entropy loss**: improving the generalisation of short speech **language** identification systems (2021)
- soMLier: A South African Wine **Recommender System** (2022)
- **Object Detection and Size Determination** of **Pineapple** Fruit at a Juicing Factory (2021)

- Unsupervised Machine Learning Application for the Identification of **Kimberlite Ore Facie** using Convolutional Neural Networks and Deep Embedded Clustering (2021)

The aim of this research is to demonstrate the viability of **implementing** a **computer vision solution** to provide online information of the **composition** of material entering the plant, thus allowing the plant operators to adjust equipment settings and process parameters accordingly

Figure 4.3: *Camera and floodlight installation at Kao Diamond Mine, above the primary feed conveyor.*

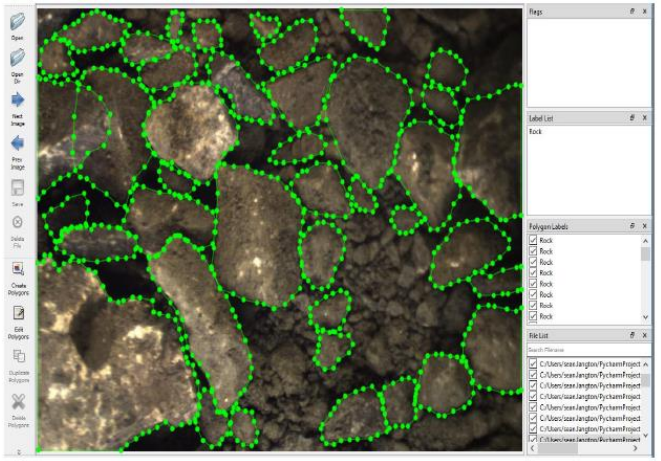


Figure 4.4: *992x662 images labelled using Labelme.py package (Wada 2016).*

The instance segmentation model consisted of a Mask R-CNN solution, making use of transfer learning on the COCO data set, as well as a ResNet 101 type architecture.

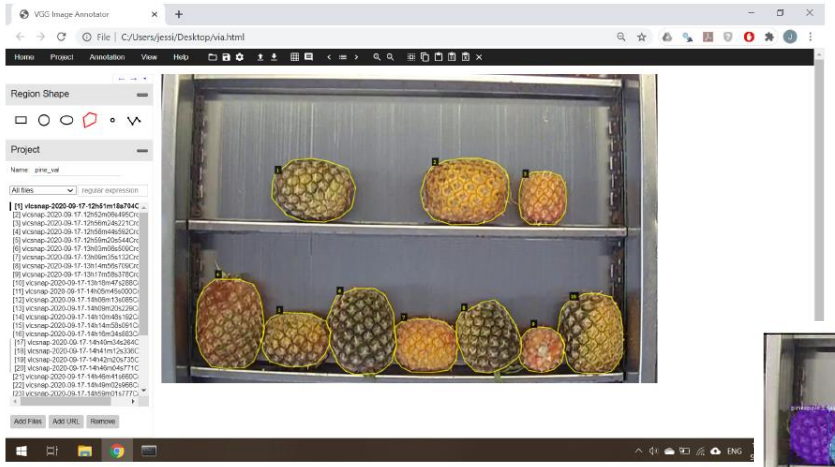
The second phase of modelling leveraged an unsupervised clustering method known as Convolutional Deep Embedded Clustering with Data Augmentation (ConvDEC-DA).

- **Object Detection and Size Determination** of **Pineapple** Fruit at a Juicing Factory (2021)

In the **Eastern Cape of South Africa**, the Smooth Cayenne pineapple cultivar is primarily grown for juicing purposes, with the juice being sold as a concentrated product with a prescribed sugar content.

Theoretically, larger pineapples should have a higher juice yield because a smaller proportion of the fruit is comprised of peel, which is removed in the first step of processing.

The image data used in this research comes from a pineapple juicing factory in the EasternCape of South Africa, where two conveyor belts carry the fresh pineapples into the factory for processing. Video footage was collected using two progressive scan CMOS cameras (Hikvision DS-2CD2145FWD-I(S)) located above the two conveyor belts (Camera A and Camera B) delivering pineapples to the factory



Model #	Model name	Augmentation	Min. Val. Loss	# epochs	Validation AP	
					IoU= 0.50	IoU= 0.50:0.05:0.95
3	COCO_NoAug_Res50	ResNet50	0.192	30	1.000	0.884
4	**COCO_Fliplr_Res50**	Horizontal flip	0.113	28	1.000	**0.914**
5	COCO_GaussNB_Res50	Gaussian noise and blur	0.110	24	1.000	0.905
6	COCO_Colour_Res50	Lightening & darkening	0.119	29	1.000	0.898
7	COCO_All_Res50	All of the above	0.107	29	1.000	0.898

- **Triplet entropy loss**: improving the generalisation of short speech **language** identification systems (2021)

Spoken language identification systems

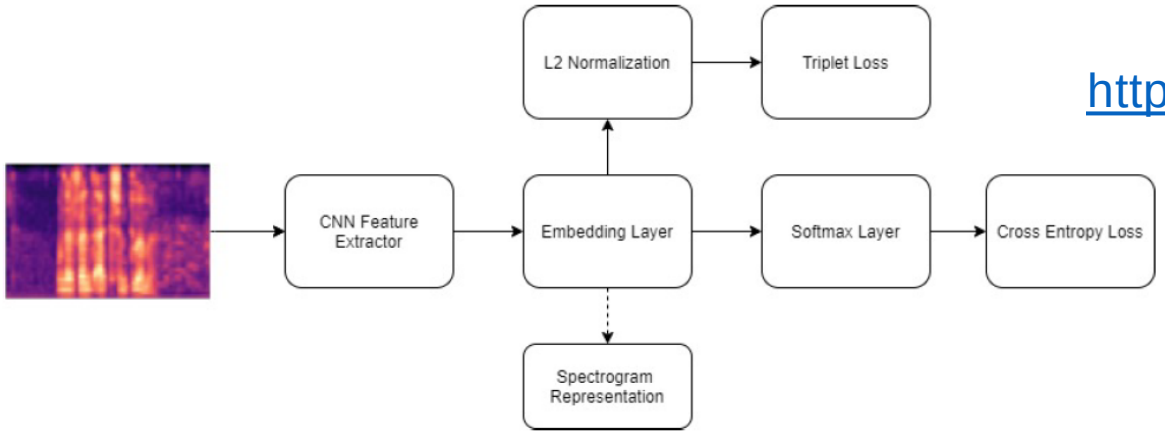
The research investigates several methods to improve the generalisation of language identification systems to new speakers and to new domains.

These methods involve Spectral augmentation

The research also introduces the novel Triplet Entropy Loss training method.

This training method involves training a network simultaneously using Cross Entropy and Triplet loss.

Several tests were performed in a South African context on six languages, namely Afrikaans, English, Sepedi, Setswanna, Xhosa and Zulu.



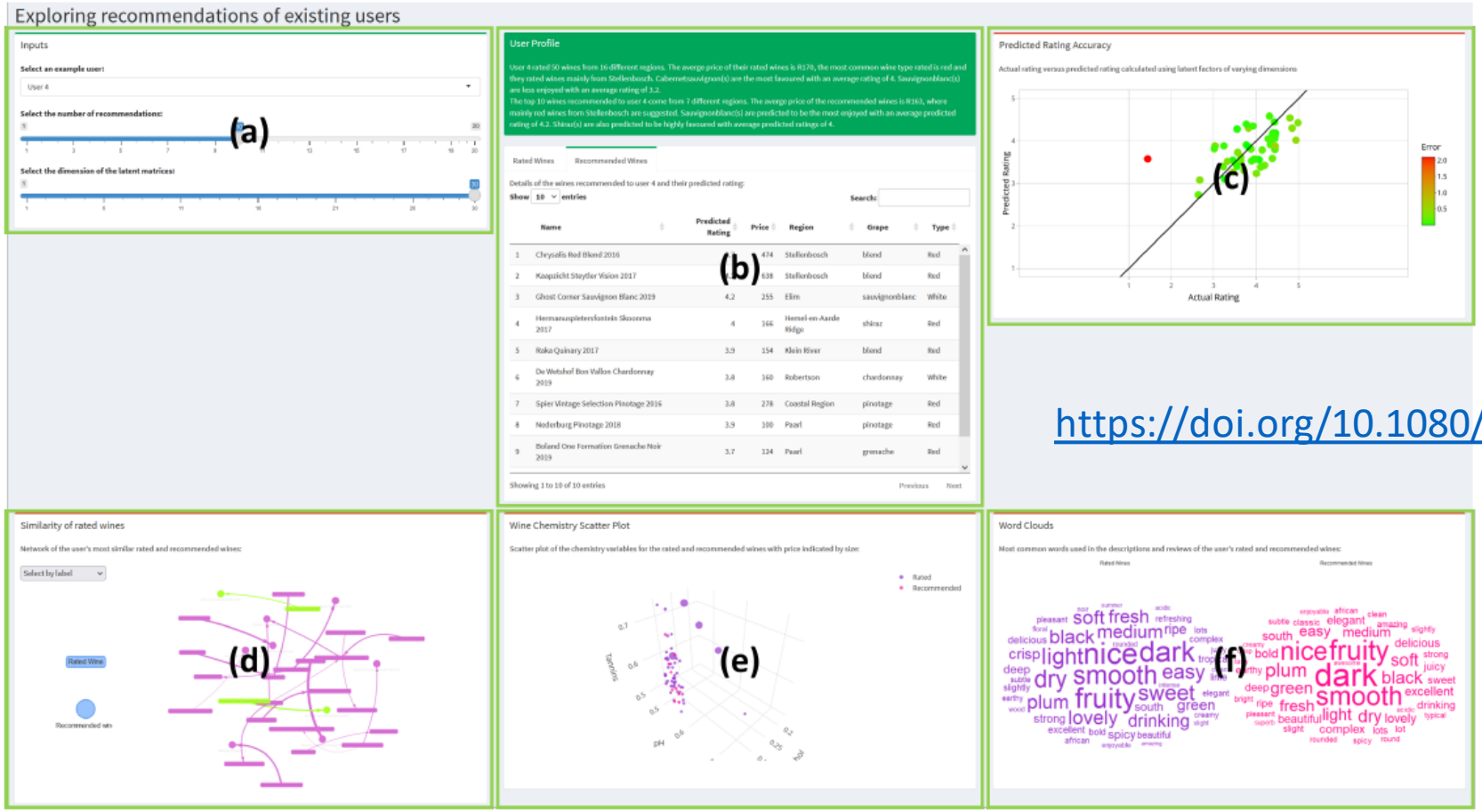
https://arxiv.org/pdf/2012.03775

Figure 4.8: Triplet Entropy Loss high level overview

- soMLier: A South African Wine **Recommender System** (2022)

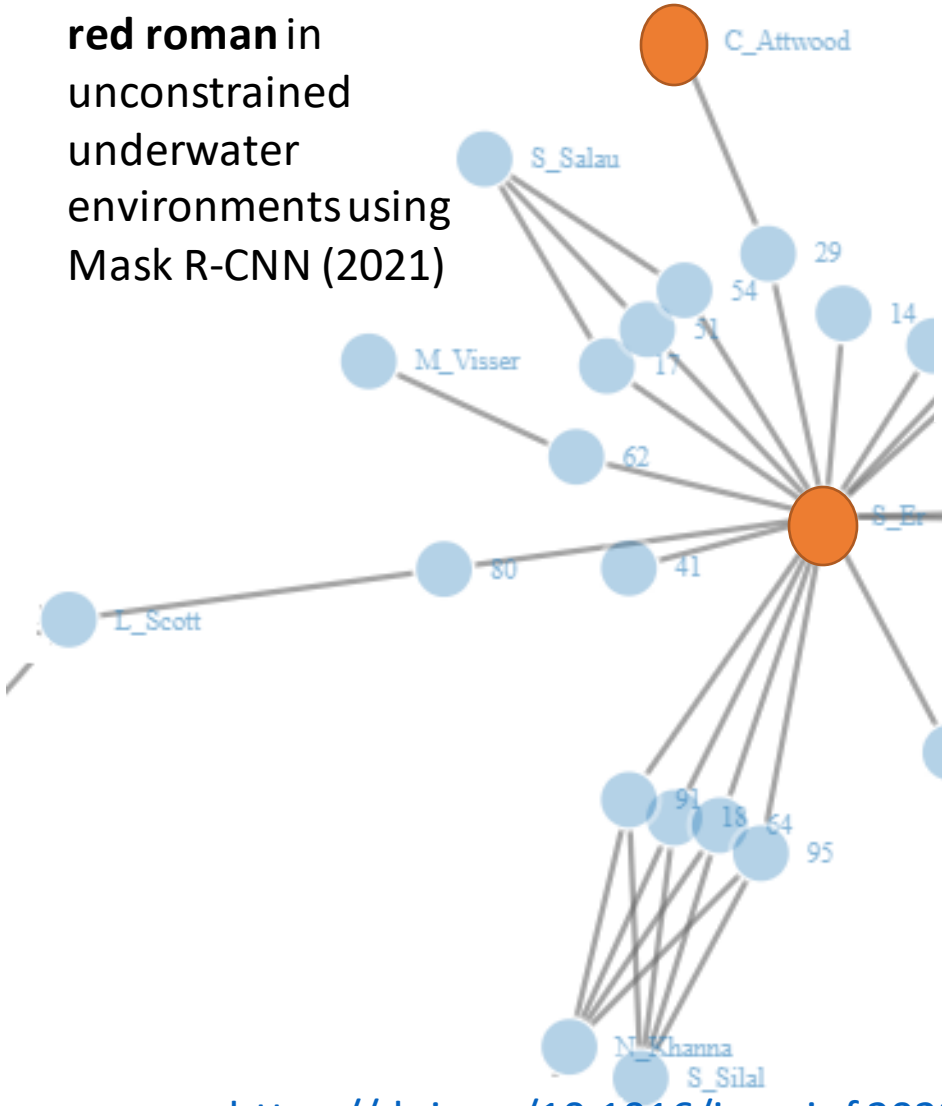
The datasets used to build the wine recommendation system consists of, firstly, a database of wines and secondly, a set of ratings and reviews for those wines. The first database contains attributes and descriptions for a set of 1,640 South African wines supplied by wine.co.za (WineNet, 2022).

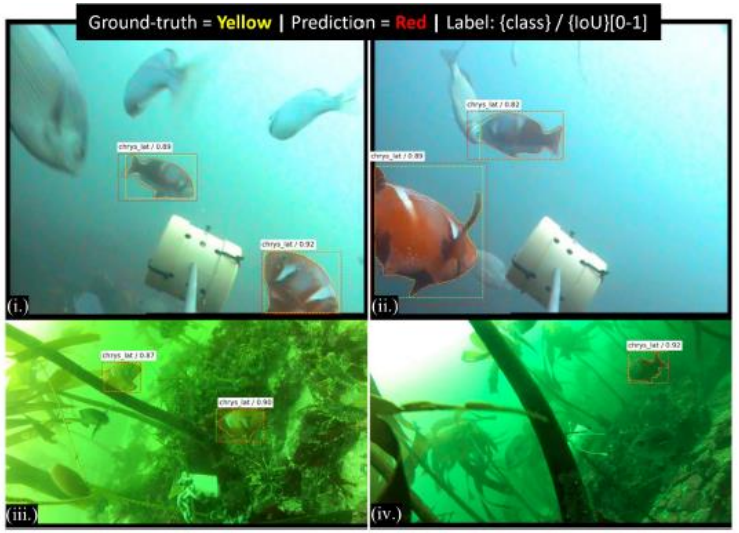
The second dataset consists of 210,605 ratings made by 92,514 users for the 1,640 wines; all ratings (made in the interval [1, 5] in increments of 0.5) of these wines were scraped from Vivino (2022) during August 2021.

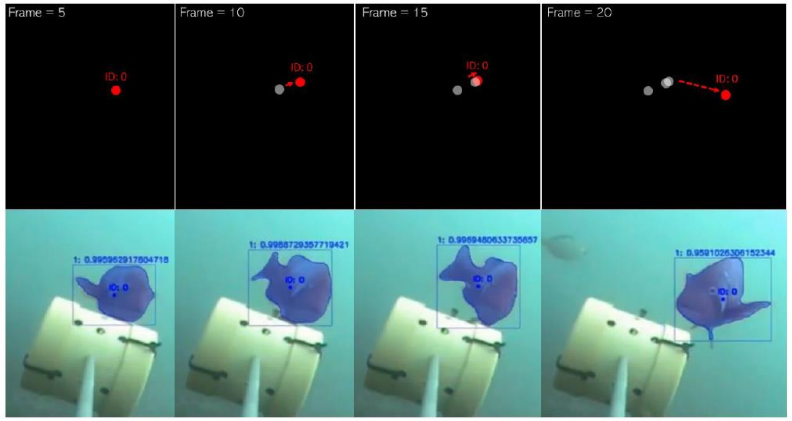


https://doi.org/10.1080/09571264.2023.2184333

- Automated **detection and classification of red roman** in unconstrained underwater environments using Mask R-CNN (2021)



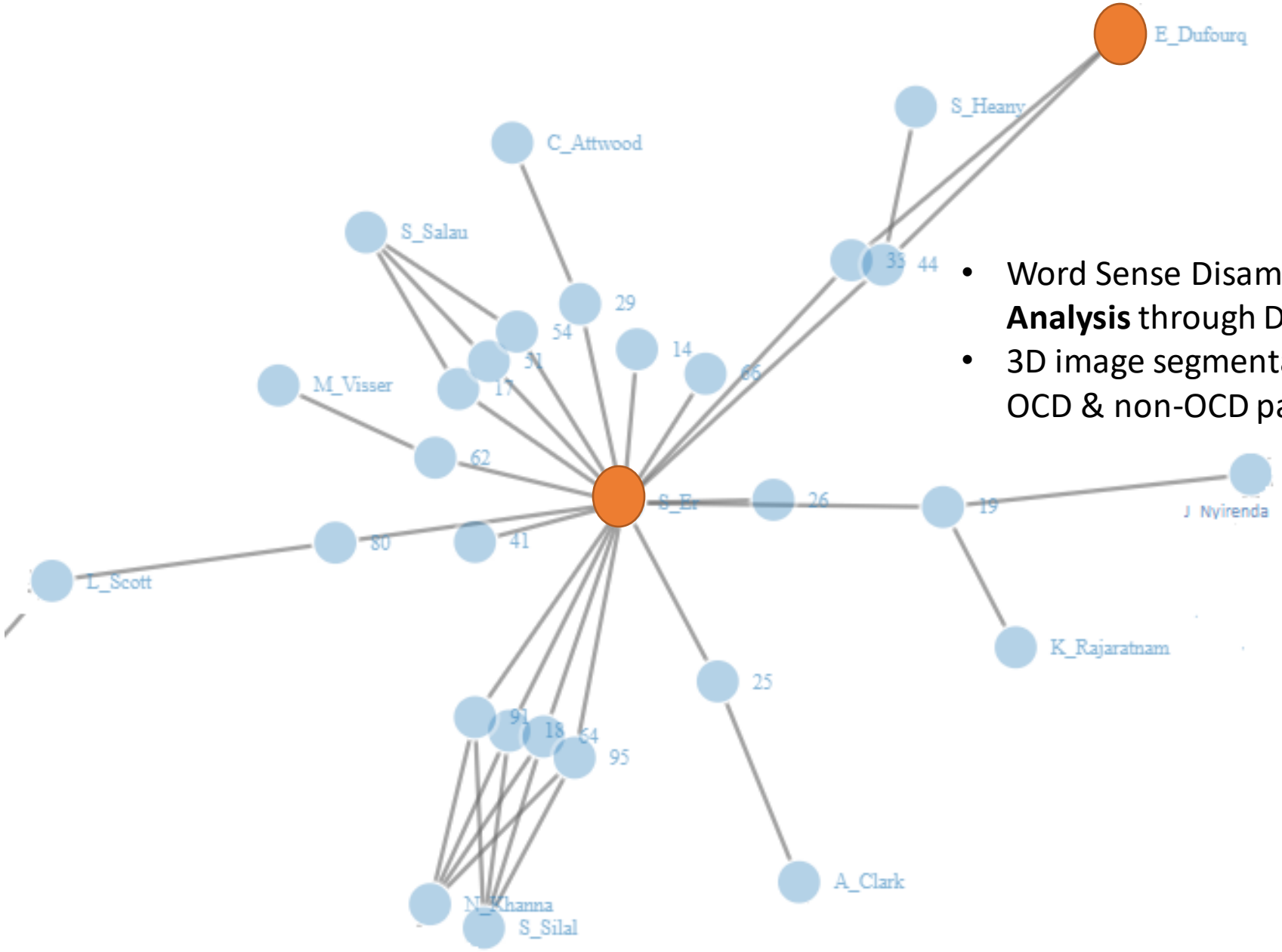




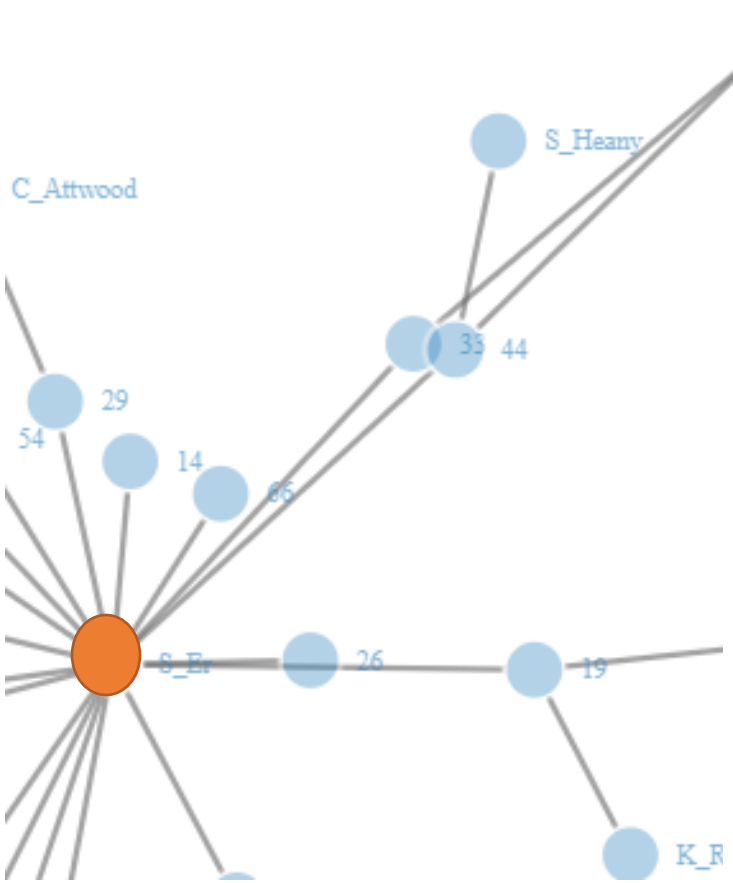
The red roman dataset comprises 2015 images and 2541 unique instance segmentation. The dataset is free to download and use for academic purposes

The red roman Mask R-CNN model developed in this research will assist marine research efforts where accurate, cheap and fast automated methods of video analysis of marine environments are necessary. The model will serve as a proof-of-concept that machine learning based methods of locating, identifying, counting and tracking fish in underwater video can replace or at least supplement human efforts

- https://doi.org/10.1016/j.ecoinf.2022.101593



- Word Sense Disambiguation in the domain of **Sentiment Analysis** through Deep Learning (2022)
- 3D image segmentation & classification of brain MRI for OCD & non-OCD patients (ongoing)

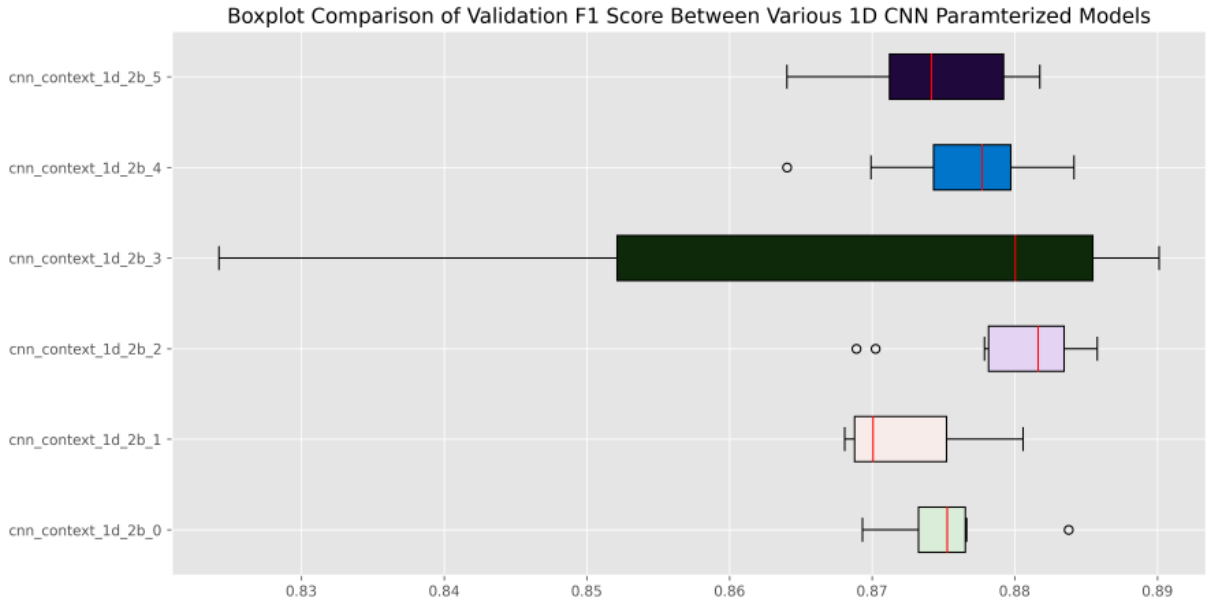


- Word Sense Disambiguation in the domain of **Sentiment Analysis** through Deep Learning (2022)

The aim of this research is to explore WSD in sentiment analysis. Amazon product review dataset was used. CNNs and LSTMs were used.

Table 4.2: Selected Amazon product categories & number of reviews

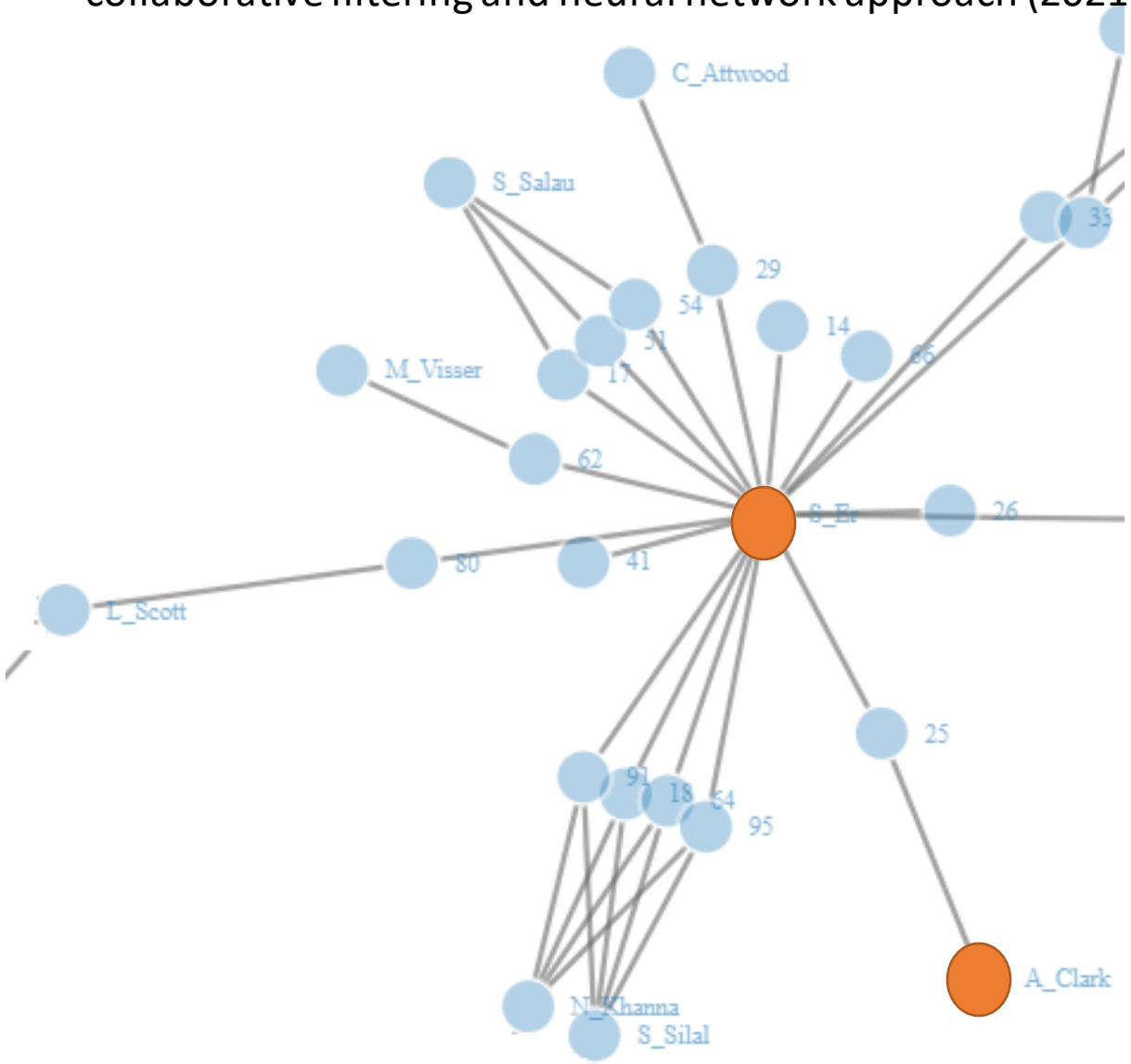
Rating	Sentiment	Sentiment Indicator
1,2,3	Negative	0
4,5	Positive	1



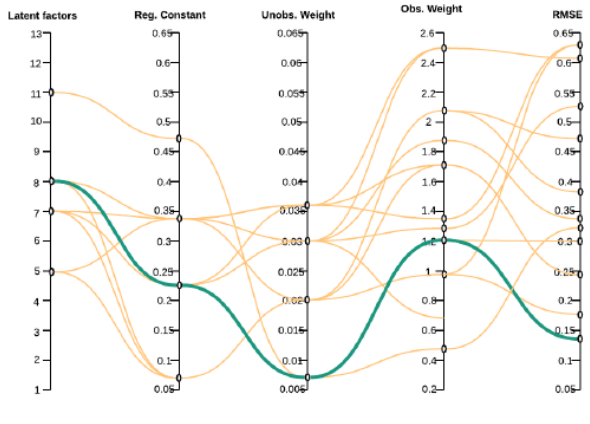
ICONDATA
INTERNATIONAL CONFERENCE ON DATA SCIENCE & APPLICATIONS

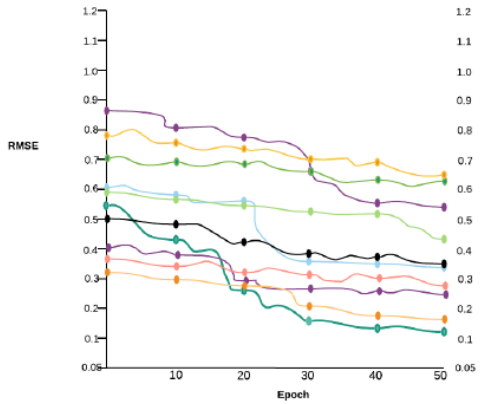
Figure 5.3: F1 score comparison between 1D CNN hyper-parameter tuned models

- **Insurance recommendation engine** using a combined collaborative filtering and neural network approach (2021)



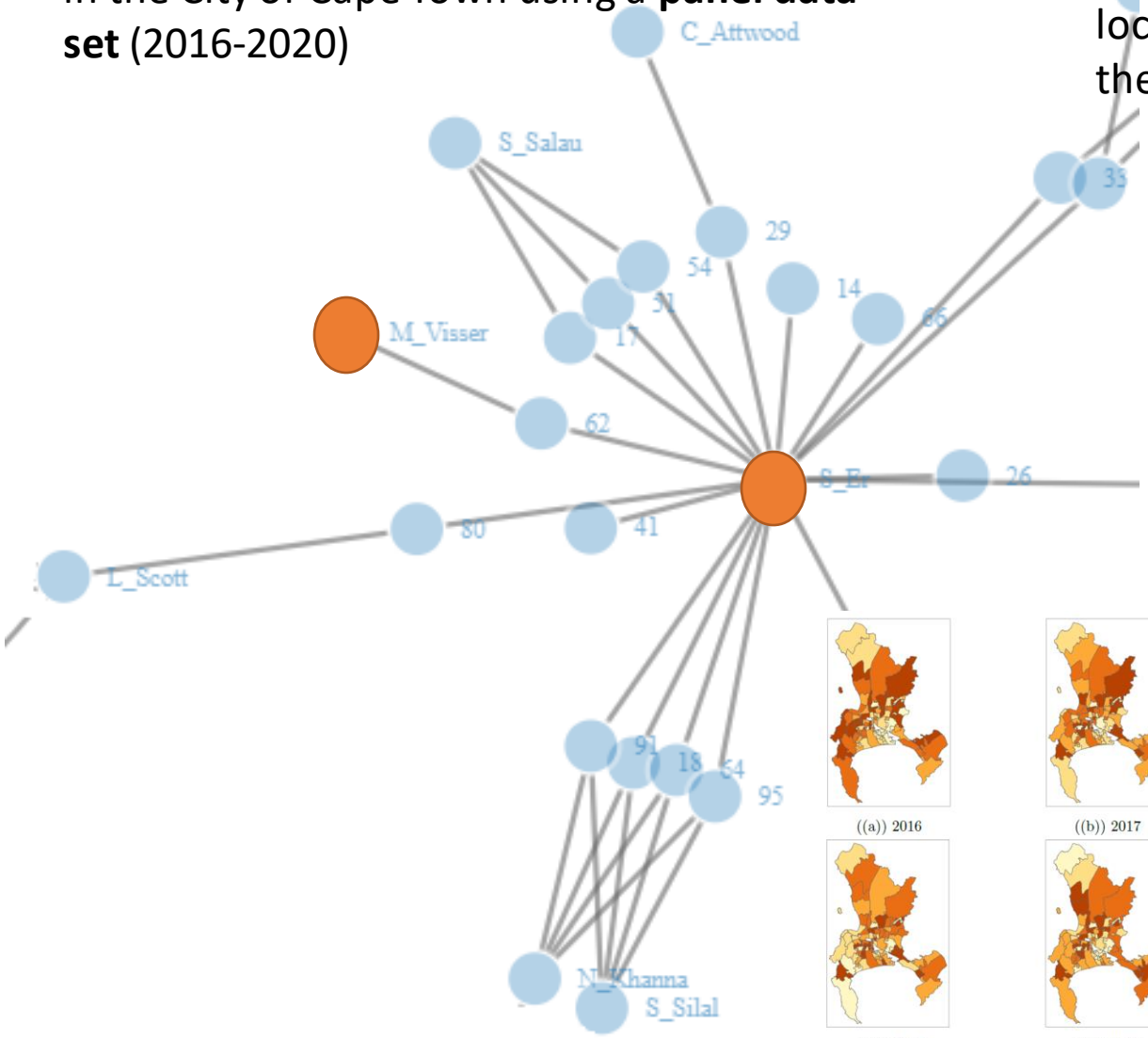
- The recommendation engine is built using both a collaborative filtering technique, as well as neural networks.
- The collaborative fitering technique is primarily used to provide existing users with recommendations, while the neural networks learning system is used to provide recommendations for new users.





- An analysis of household water consumption in the City of Cape Town using a **panel data set** (2016-2020)

The data in its raw form, is a collection of repeated household monthly observations. Each household has an associated point location and falls within the perimeters of one of the 116 wards the CoCT is divided into.



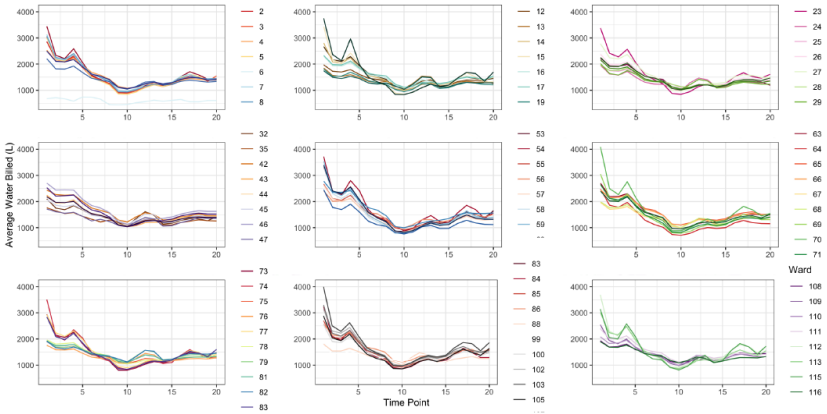


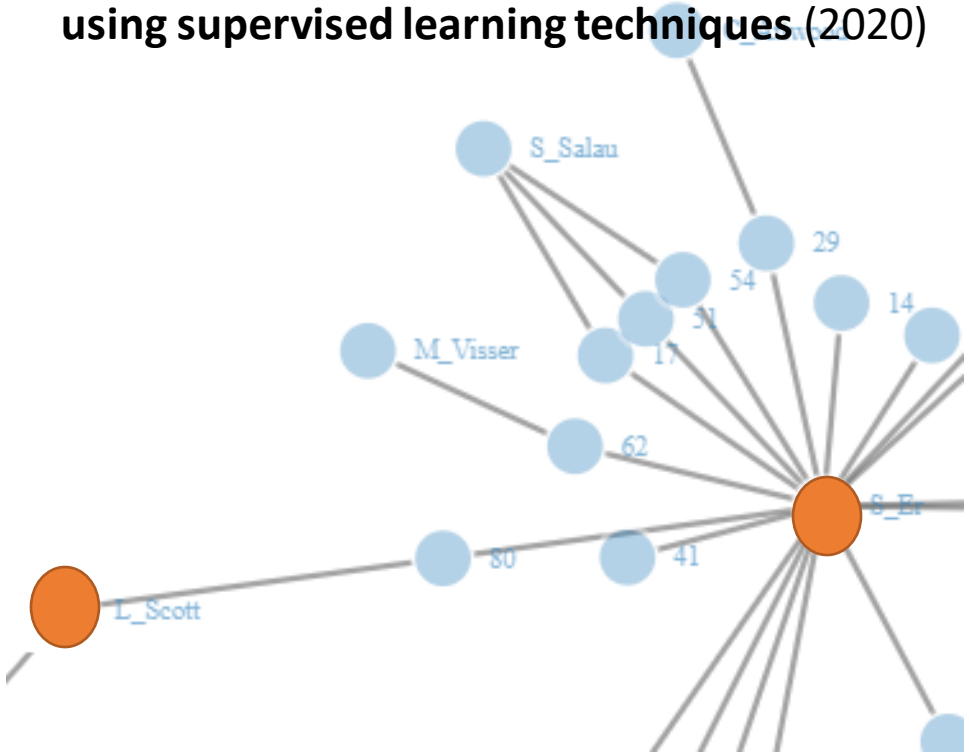
Figure A.1: Average quarterly water consumption (L), by ward

((a)) 2016 ((b)) 2017 ((c)) 2018

((d)) 2019 ((e)) 2020

Figure 4.5: Average ward water consumption, by year

- Building a question answering system for the introduction to statistics (STA1000) course **using supervised learning techniques** (2020)



The questions posed by this study are:

- Which features preserved the meaning of questions the best?
- Which classification model performed the best in predicting the question category?
- Was it a good idea to apply augmentation to the base dataset to increase the number of training samples?
- How can the best past answer be selected for a new question?
- Can sentiment analysis be used to help prioritise the answering of questions?

Table 5.5: Top 10 feature types with best performing models

Feature Type	# Samples	# Features	Best Model	F1-measure		Rank
				Train	Test	
TF-IDF unigrams + bigrams	587	5 688	MLR	0.978	0.848	1
LSA TF-IDF unigrams (probabilities)	587	55	MLR	0.844	0.844	2
BoW counts	587	1 101	MLR	0.934	0.841	3
TF-IDF unigrams	587	1 101	SGD	0.951	0.835	4
LSA TF-IDF unigrams (raw values)	587	55	MLR	0.886	0.831	5
LSA TF-IDF unigrams + extra RTT data (probabilities)	6 457	49	MLR	0.783	0.825	8
TF-IDF unigrams + bigrams + trigrams	587	10 972	SGD	0.998	0.821	9
TF-IDF unigrams + bigrams + trigrams + extra RTT data	6 457	53 279	SGD	0.977	0.817	12
BoW binary values	587	1 101	SGD	0.947	0.817	13
Engineered features	587	61	SGD	0.911	0.813	15

- BoW
- TFIDF
- LDiA

Table 3.4: Example questions with a mapping from question date to course week

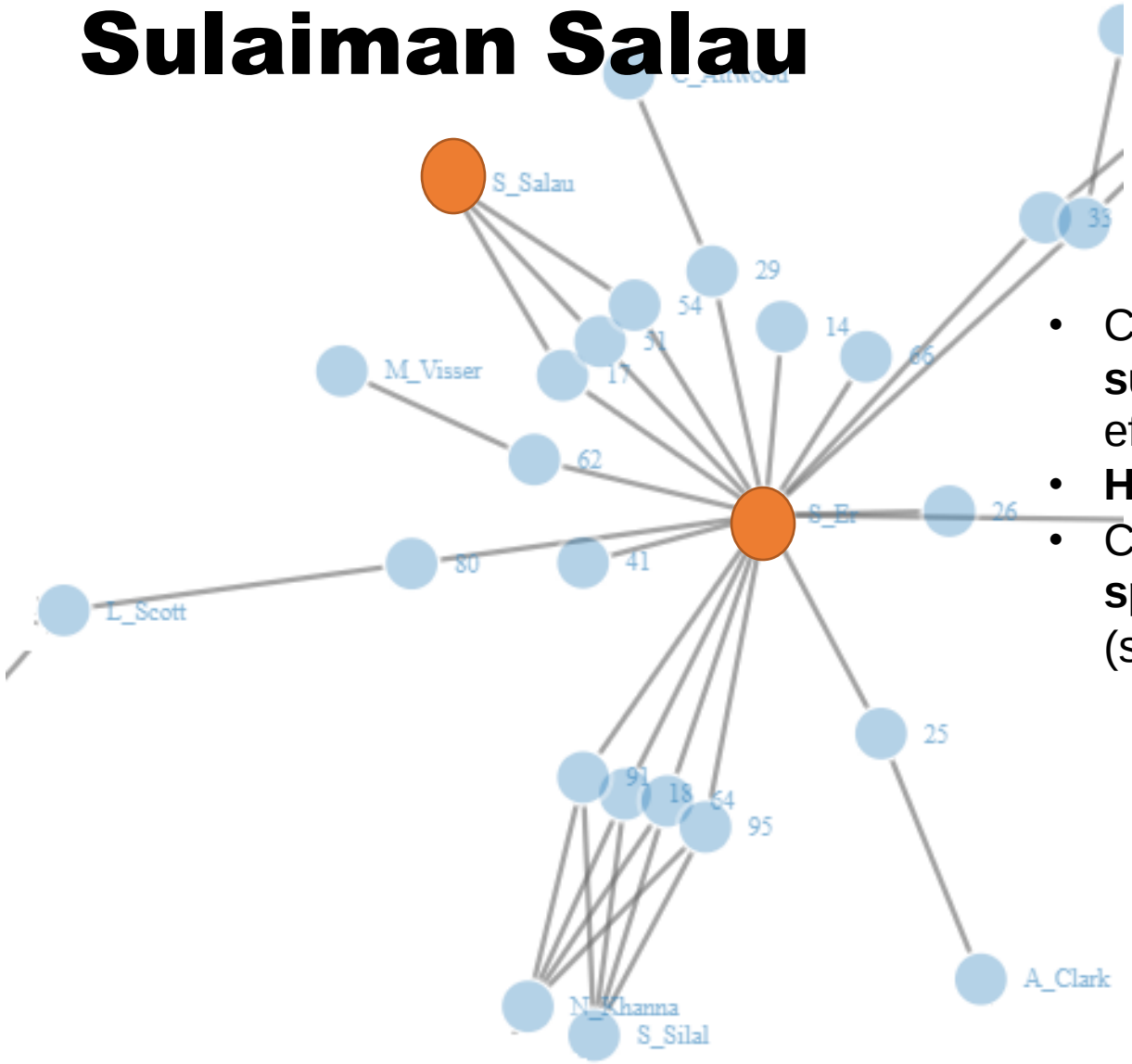
Question Text	Question Date	Course Week*
Hi, I also have a clash for both workshop slots. Is it possible for another slot to become available. Perhaps on a Monday ?	2015-02-16	1
I have a lecture clash ,can I still send my timetable in so I can be placed in a suitable workshop.	2015-02-27	2
I have a timetable clash with every workshop other than the Wednesday 9 to 10 slot. But it is already full. What should I do?	2015-07-21	1
HI [*sic*] I have a clash with all the listed lecture times. Is it okay if I only do the workshops?	2015-07-26	1

* The course week number is derived from the question date, taking into account that the course runs multiple times a year.

Table 5.12: Evaluation of a test question for the "Workshops" category

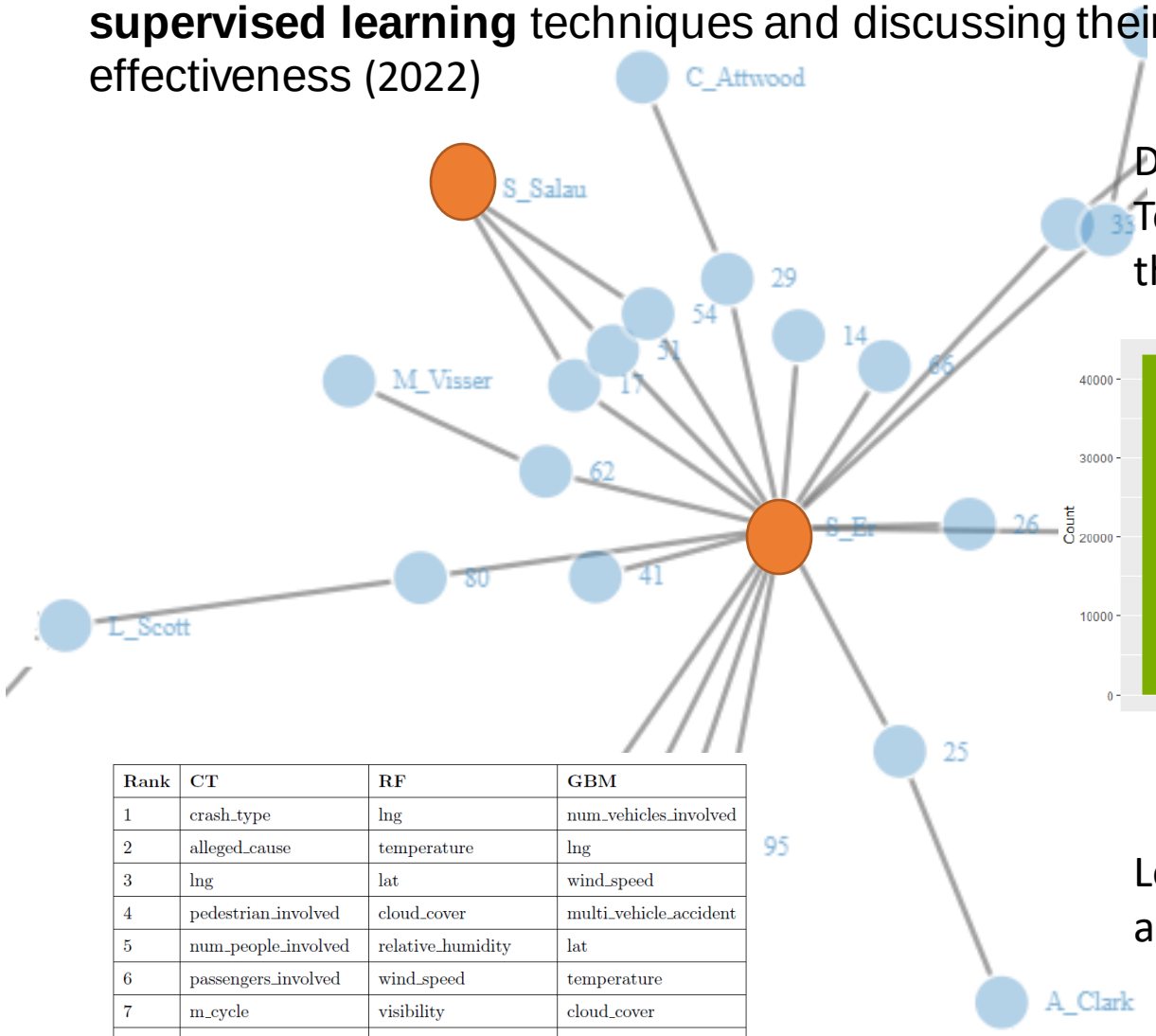
Type	Question Text	Question Tokens	Answer
Test question	If we failed test 1, will we need to attend a workshop this week?	{"fail", "test", "will", "need", "attend", "workshop", "week"}	If you failed Test 1 then you will be required to attend both workshops and tutorials (as a DP requirement).
Most similar past question	I failed test 1 but passed test 2 - am I still obliged to attend the weekly workshops?	{"fail", "test", "pass", "test", "still", "oblige", "attend", "weekly", "workshop"}	No, you are not, well done on the improvement! Although, if you reckon that attending the workshops helped you improve, then it might be wise to attend the last few as well.
Human evaluation	Are workshops compulsory?	{"workshop", "compulsory"}	No, but should you fail Test1 [*sic*], then it will be become compulsory for you to attend one workshop a week (in addition to tutorials).

Sulaiman Salau

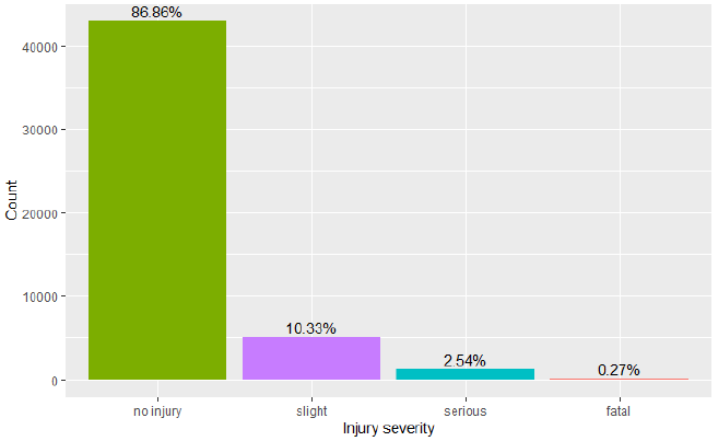


- Cape Town road traffic accident analysis: Utilising **supervised learning** techniques and discussing their effectiveness (2022)
- **Hospital Readmission** Risk (ongoing)
- Cape Town Airbnb price prediction: an exploration of **spatial statistic and machine learning** methods (submitted)

- Cape Town road traffic accident analysis: Utilising **supervised learning** techniques and discussing their effectiveness (2022)



Data on RTAs in Cape Town were sourced from the City of Cape Town. The dataset contains records of more than 82 000 RTAs that occurred during the 2015-2017 period.



Imbalanced nature of data was incorporated into the analysis with different sampling schemes.

Rank	CT	RF	GBM
1	crash_type	lng	num_vehicles_involved
2	alleged_cause	temperature	lng
3	lng	lat	wind_speed
4	pedestrian_involved	cloud_cover	multi_vehicle_accident
5	num_people_involved	relative_humidity	lat
6	passengers_involved	wind_speed	temperature
7	m_cycle	visibility	cloud_cover
8	multi_vehicle_accident	crash_type	visibility
9	num_vehicles_involved	alleged_cause	relative_humidity
10		num_vehicles_involved	crash_type

Logistic regression, random forests, GBMs, and ANNs were applied.

- Cape Town Airbnb price prediction: an exploration of **spatial statistic and machine learning** methods (submitted)

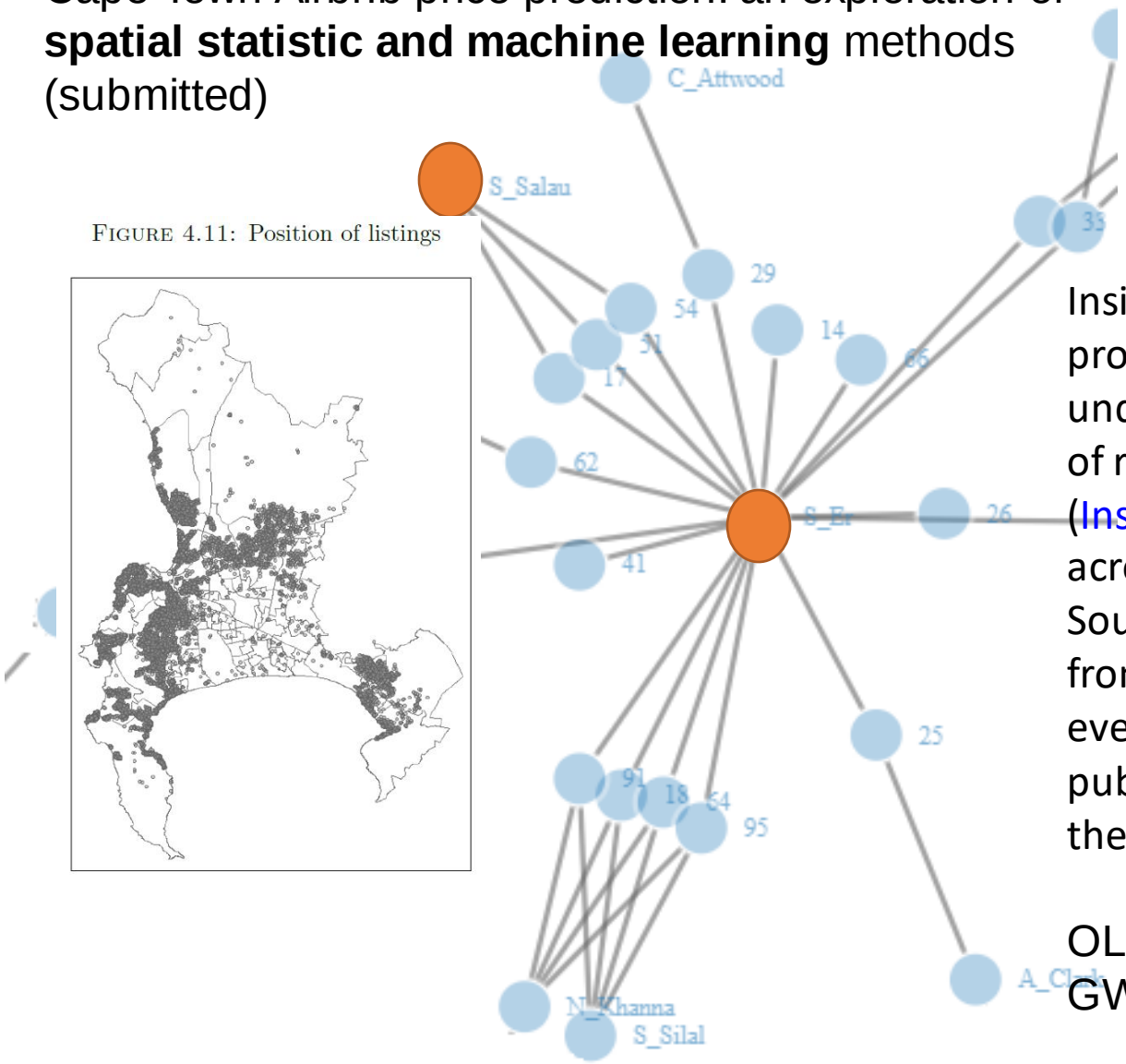


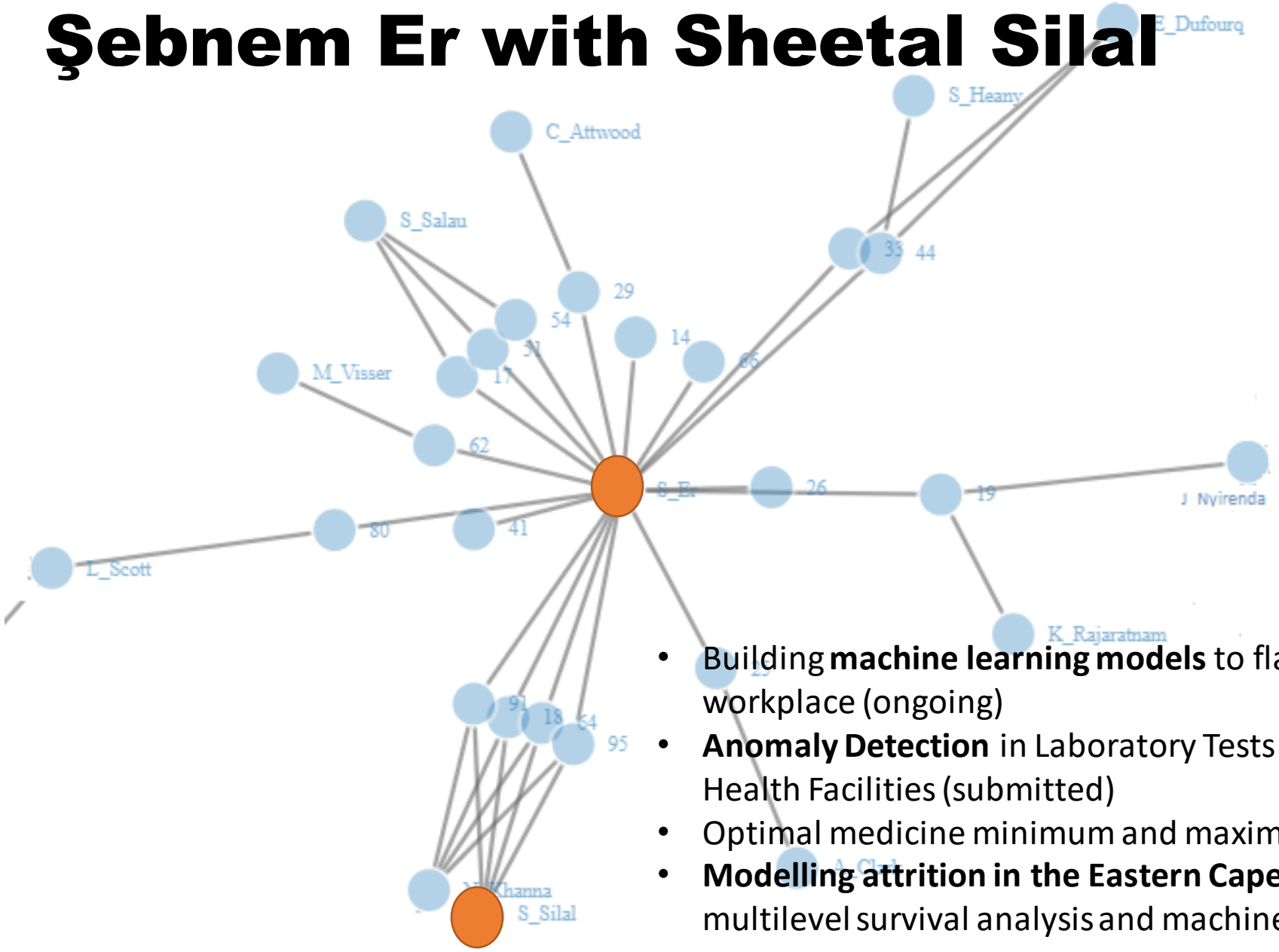
FIGURE 4.11: Position of listings

Inside Airbnb is a project that aims to provide data that can be used to understand, decide and control the role of renting residential homes to tourists" (InsideAirbnb, 2023). For cities across the world, including Cape Town, South Africa, Inside Airbnb scrapes data from the Airbnb website approximately every three months, and make the data publicly available on their website.

OLS, Spatial lag and error models, GWR

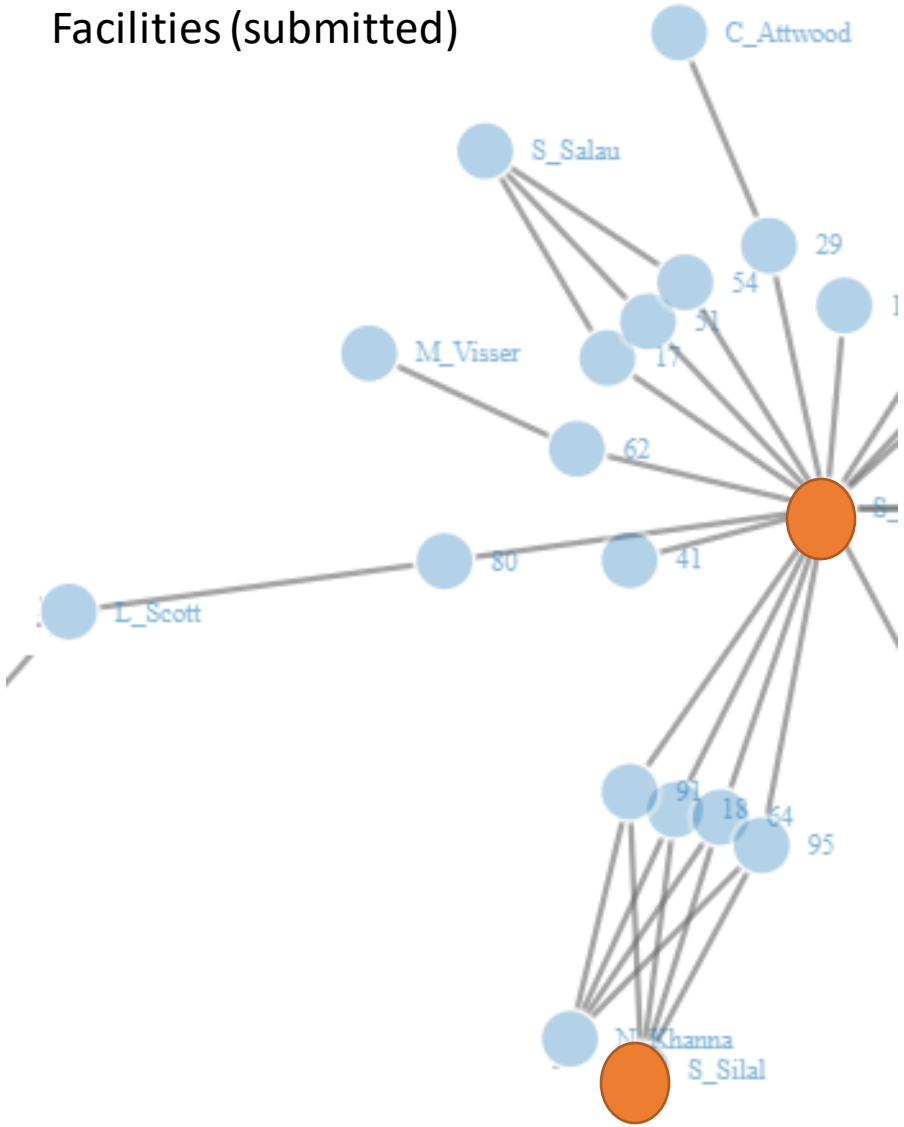
Category	Variable
Basic	Accommodates
	Room type: Hotel room
	Room type: Private/ shared room
Amenities	Wifi
	Free parking
	Pool
	Hot tub
	Kitchen
	Aircon
	Wash
	Beach / water front
Host	Verified host identity
	Host listings count
	House rules
Location	Exact location
	Suburb: City Bowl
	Suburb: Eastern
	Suburb: Northern
	Suburb: South East
	Suburb: South Peninsula
	Suburb: Southern
	Suburb: West Coast
	Distance to airport
	Distance to nearest attraction
Reviews	Number of reviews
	Reviews: 100
	Reviews: 93.99
	Reviews: null rating
Costs	Cleaning fee
	Security deposit
	All inclusive
	Additional charge
Availability	Short term only
	Availability 30 days
	Availability 365 days
	Instantly bookable
	Cancellation policy

Şebnem Er with Sheetal Silal



- Building **machine learning models** to flag **gender inequality** in the workplace (ongoing)
- **Anomaly Detection** in Laboratory Tests Subject to Gatekeeping in Selected Health Facilities (submitted)
- Optimal medicine minimum and maximum **stock level forecast** (ongoing)
- **Modelling attrition in the Eastern Cape public health system** using multilevel survival analysis and machine learning methods (submitted)

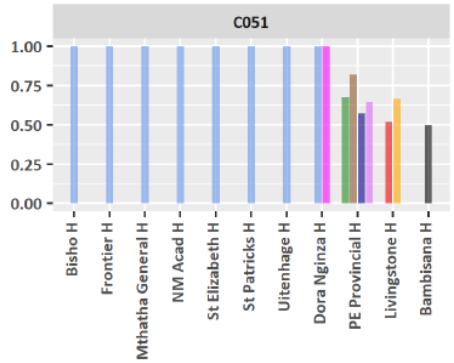
- **Anomaly Detection** in Laboratory Tests Subject to Gatekeeping in Selected Health Facilities (submitted)



The electronic gatekeeping (eGK) system in South Africa is a standardised set of rules that was developed by the National Department of Health (NDOH), NHLS pathologists and clinicians from the individual provinces of South Africa (NHLS 2017; Smit et al., 2015). The eGK system restricts test ordering by applying a given set of rules to tests ordered by a medical official for each patient

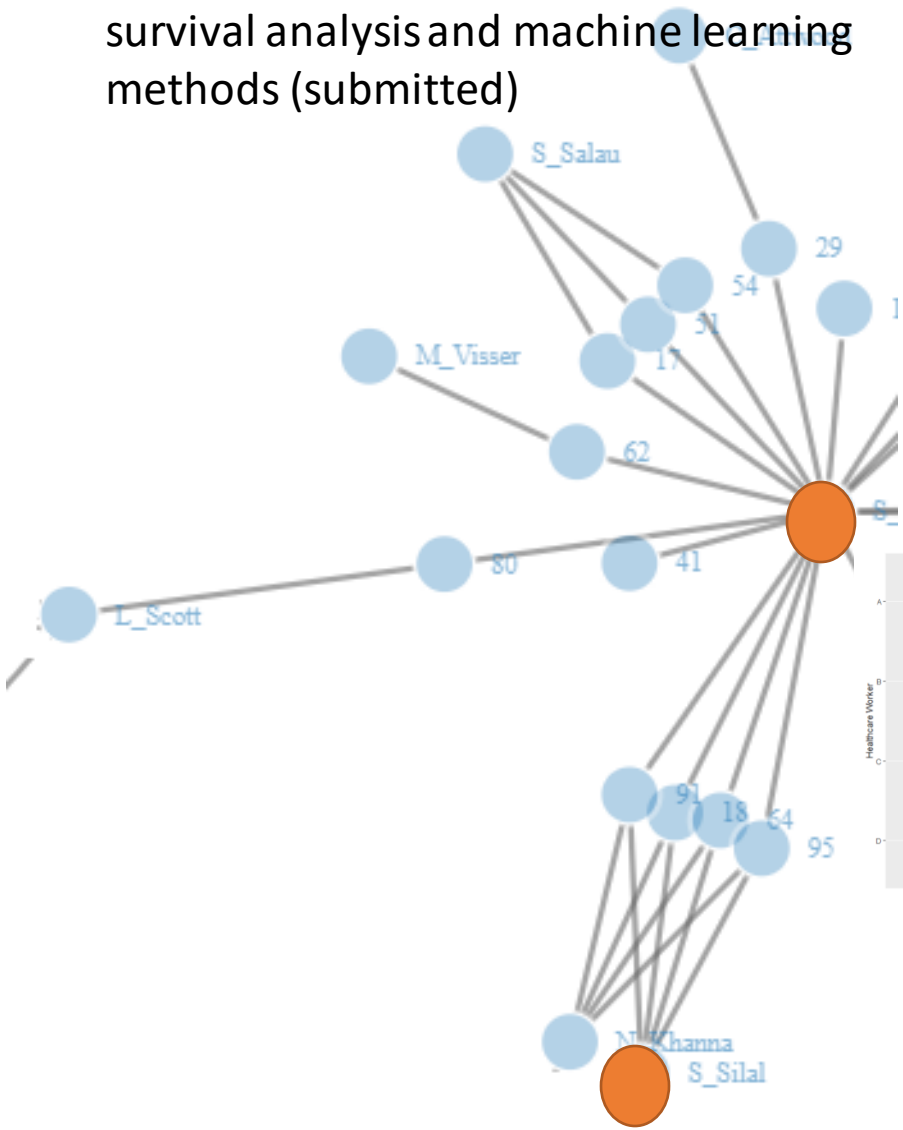
The aim of this project is to use anomaly detection methods to identify facilities in the Eastern Cape that are incurring more violations of eGK rules compared to other facilities.

Several methods were applied, K-means, K-medoids, Isolation forests, one-class SVMs



Test Code	Test Description
C002	Creatinine
C017	Urea
C051	Calcium
C053	Magnesium
C054	Inorganic phosphate
C056	Total protein
C057	Albumin
C058	Total bilirubin
C059	Conjugated bilirubin
C062	Alkaline phosphatase (ALP)
C063	Gamma-glutamyl transferase (GGT)
C093	C-reactive protein
C104	Procalcitonin
H001	Full blood count

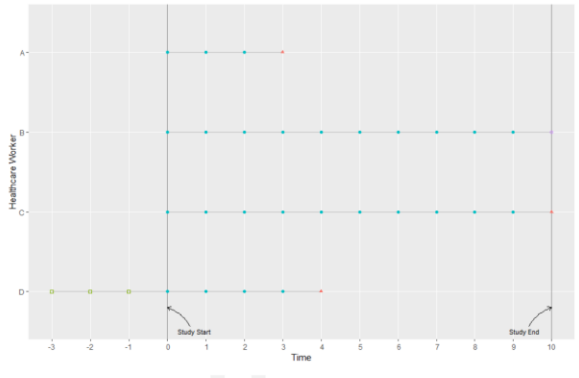
- **Modelling attrition in the Eastern Cape public health system** using multilevel survival analysis and machine learning methods (submitted)

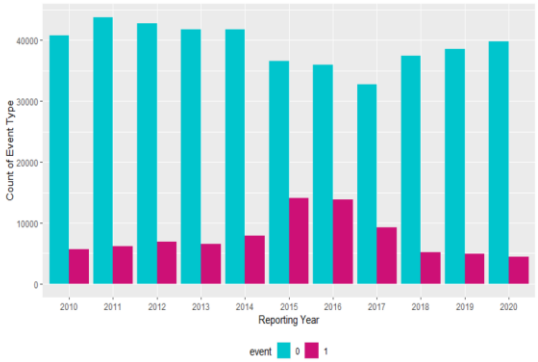


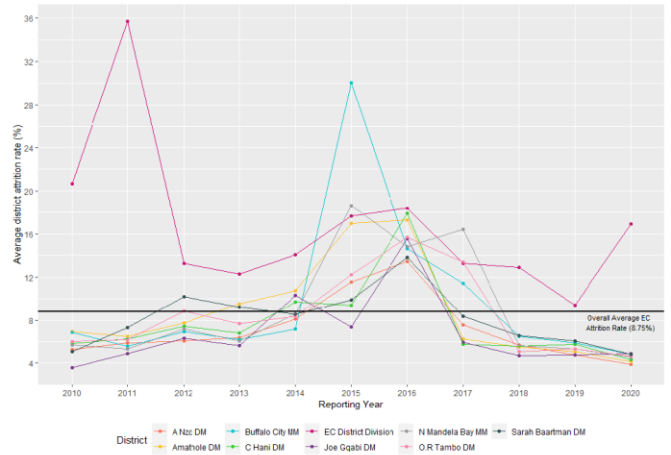
South Africa's public health system budget currently accounts for an annual 5% attrition rate for health facilities in general. This rate does not consider fluctuations in attrition rates between cadres, across facilities, or across districts. Presently, there are no guidelines or models for predicting attrition within the Eastern Cape (EC) public health care system from an individual, cadre, facility, or district level. As a result, staffing levels are determined entirely by the discretion of facility or departmental managers.

The study aims to develop and compare attrition prediction models in the EC public health sector.

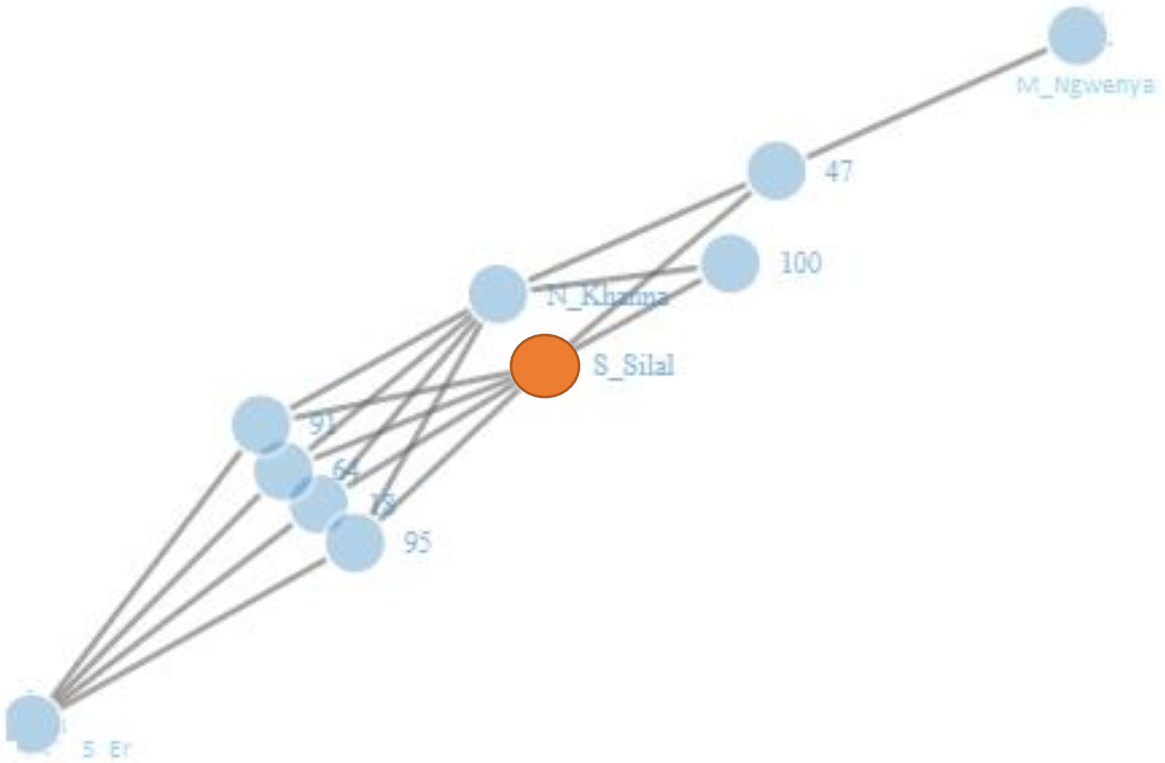
Multilevel Discrete-Time Event (MDTE) models and three ML modelling methods (MLP NNs, GLMM trees, and TBME models) have been applied to such data with the intent to predict discrete attrition events and determine factors influencing attrition.





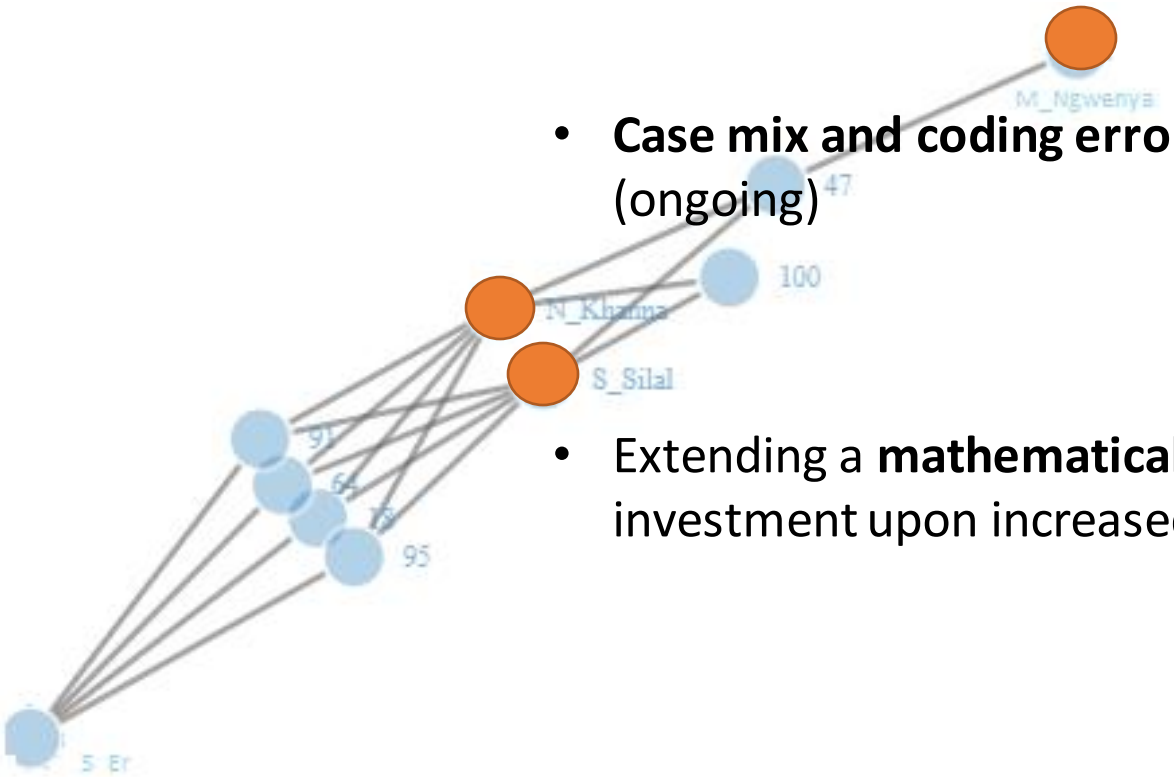


Sheetal Silal

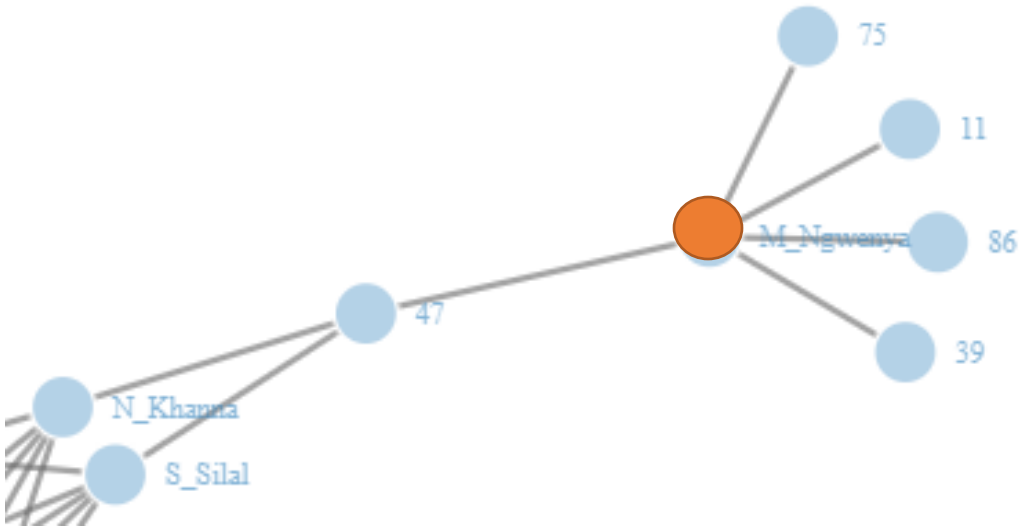


- Mathematical modelling
- Supervised and unsupervised learning methods
- Health sciences

Sheetal Silal



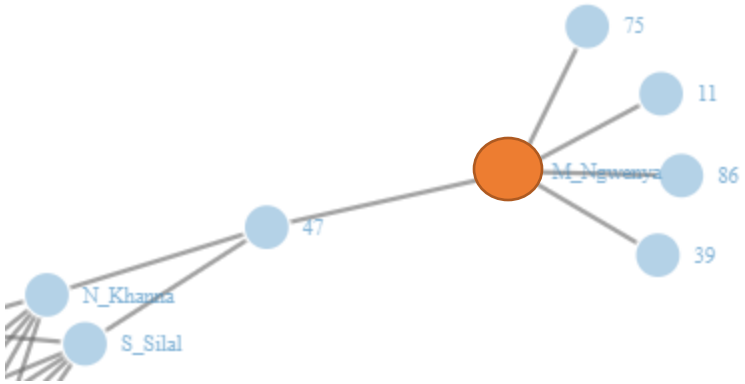
- **Case mix and coding error detection** in Western Cape healthcare facilities (ongoing)
- Extending a **mathematical model** of **syphilis** in South Africa, to develop a case for investment upon increased intervention of certain subpopulations. (ongoing)



Mzabalazo Ngwenya

- Mathematical modelling
- Supervised and unsupervised learning methods
- Image
- Health sciences
- Industry

Mzabalazo Ngwenya



- **COVID-19 fake news detection** using machine learning algorithms (ongoing)
- Buying **pattern analysis** of products for a global retail (ongoing)
- **Anomaly detection** in a Mobile Data Network (2019)
- A temporal prognostic model based on **dynamic Bayesian networks**: mining medical insurance data (2021)

- **Anomaly detection** in a Mobile Data Network (2019)

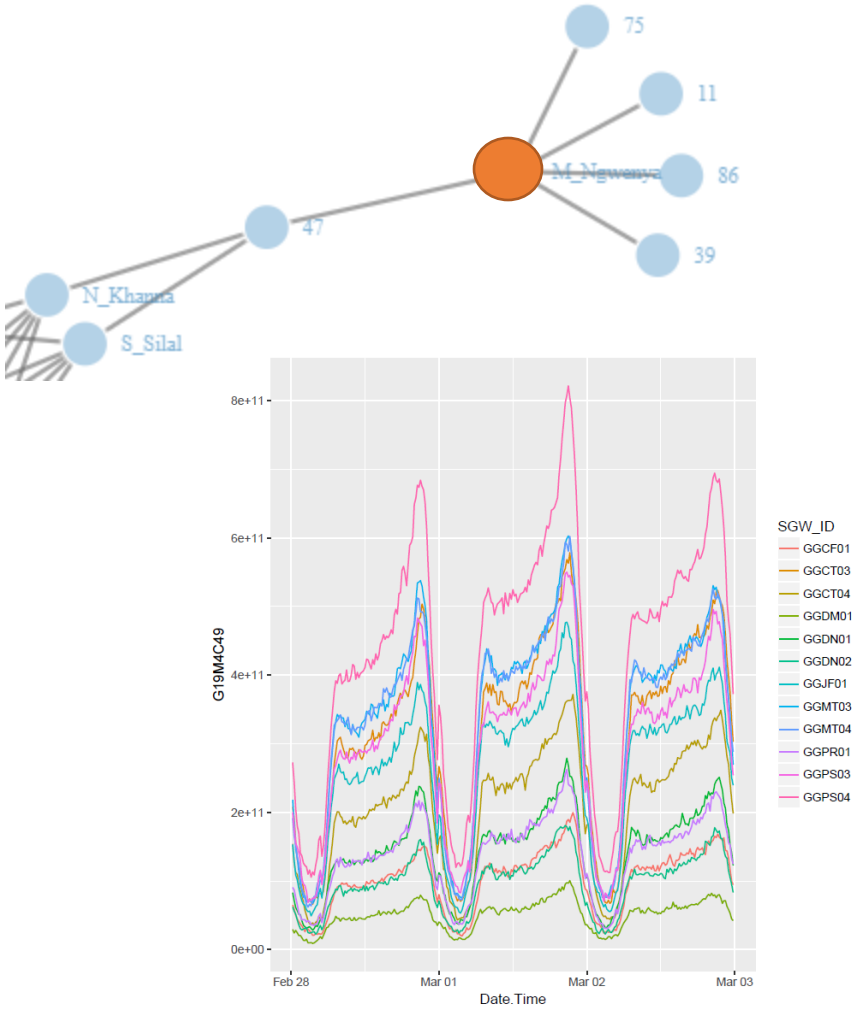


Figure 2.3: S1U Down link Bytes per SGW.

The source of the data are the GGSN/SGW/PGW nodes in the mobile network

Classic methods such as KNN, one class SVM etc and auto-encoders were applied.

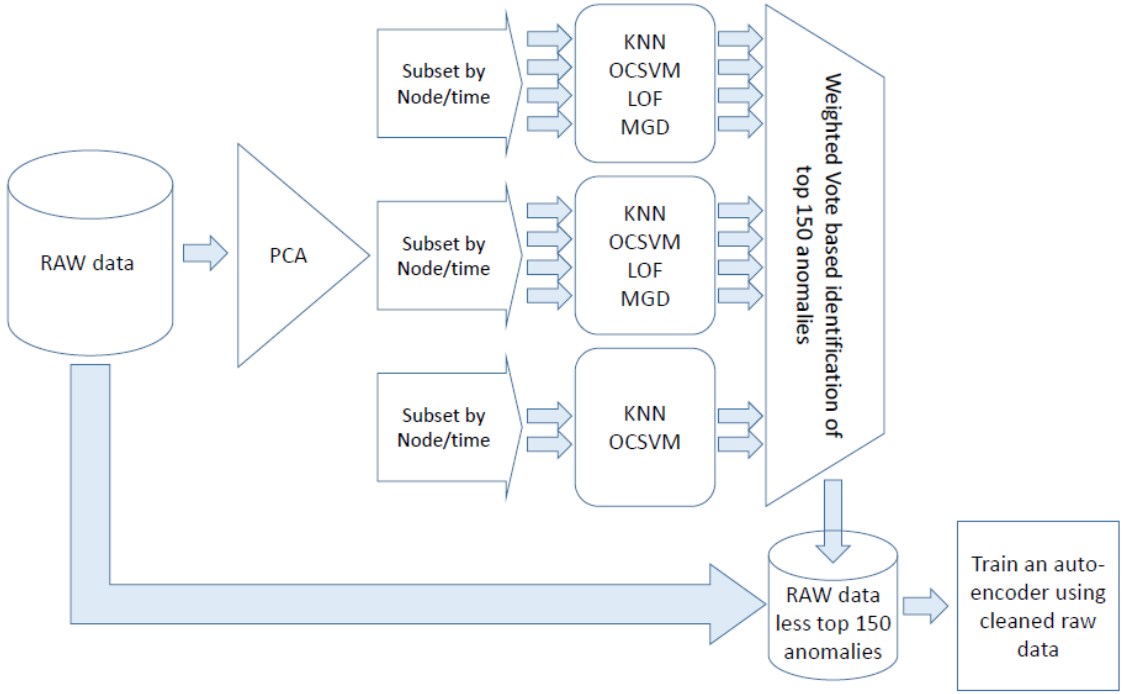
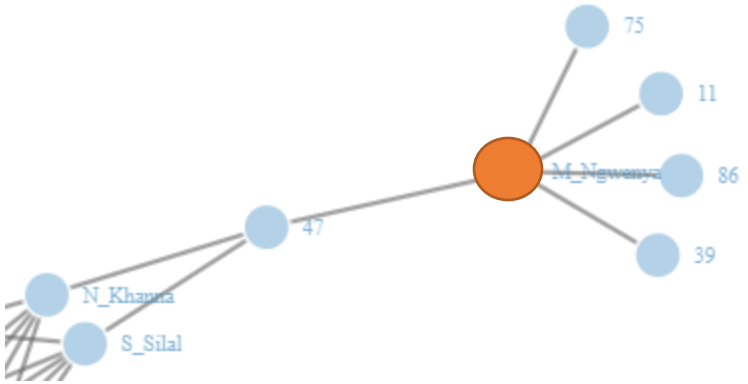


Figure 2.23: Data preparation process removing anomalies for autoencoder training.

- A temporal prognostic model based on **dynamic Bayesian networks**: mining medical insurance data (2021)



The structure of the model is presented in Figure 3.1.

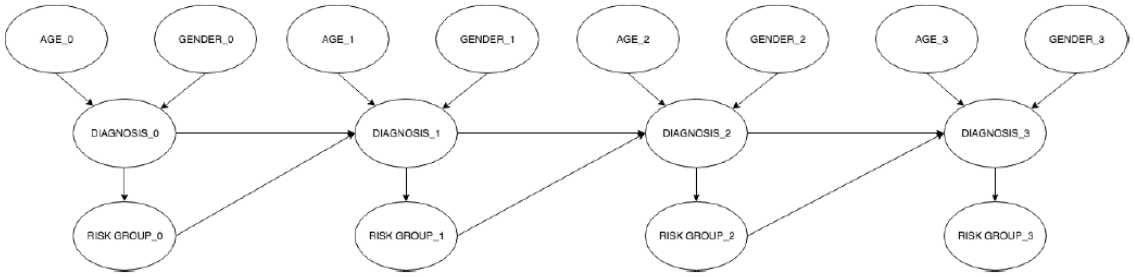


Figure 3. 1 Prognosis Dynamic Bayesian Network

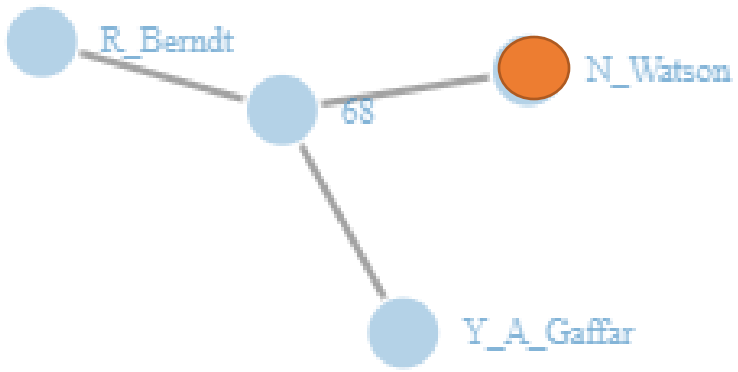
The models are used for prediction purposes of guiding doctors to make a smart diagnosis, patient-specific decisions or help in planning the utilization of resources for patient groups who have similar prognostic paths.

Dynamic Bayesian networks theoretically provide a very expressive and flexible model to solve temporal problems in medicine.

Challenges about the DBN methodology and implementation include the lack of tools that allow easy modelling of temporal processes. Overcoming this challenge will help to solve various clinical temporal reasoning problems.

In this project, these challenges were addressed while building **a temporal network** with explanations of the effects of predisposing factors, such as age and gender, and the progression information of all diagnoses using claims data from an insurance company in Kenya.

Neil Watson



- Operations research
- Supervised and unsupervised learning methods
- Sports sciences

Neil Watson

- **Radar-Based Multi-Target Classification** Using Deep Learning (2022)

Figure 4.1: Radar on top of the roof of building 44 at CSIR's main campus in Pretoria [34].

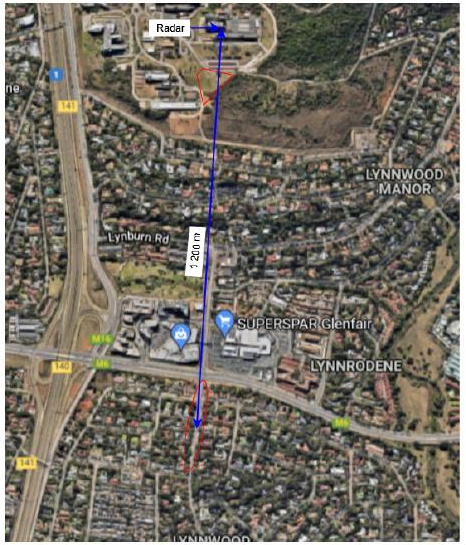


Figure 4.4: Field and road (marked in red) from which human activity measurements were taken.

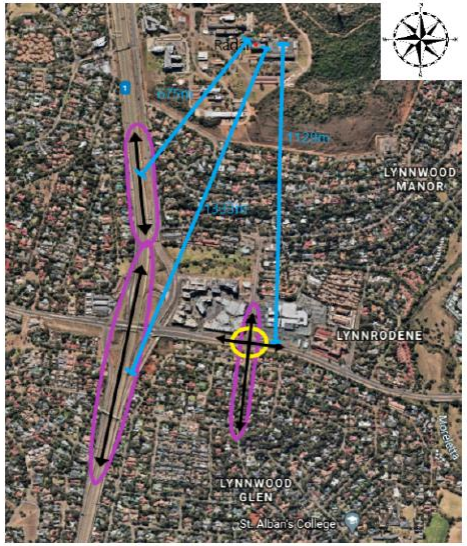


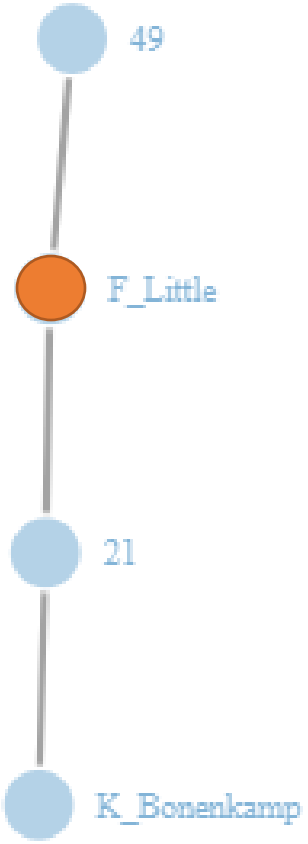
Figure 4.5: Roads (marked in purple) from which moving vehicles were measured [34].

Real-time, radar-based human activity and target recognition has several applications in various fields. Examples include hand gesture recognition, border and home surveillance, pedestrian recognition for automotive safety and fall detection for assisted living.

This project aims to improve the **speed** and **accuracy** of a previously developed model **classifying human activity** and **targets using radar data for outdoor surveillance purposes**.

A seven-layer residual network was proposed, which utilised batch normalisation (BN), global average pooling (GAP), and residual connections to achieve a classification accuracy of 92.90% and 87.72% on the validation and test data, respectively.

Francesca Little



- Statistical modelling
- Machine learning methods
- Health sciences

- **Modelling** gametocytes in the presence of **interval-censoring** (2022)

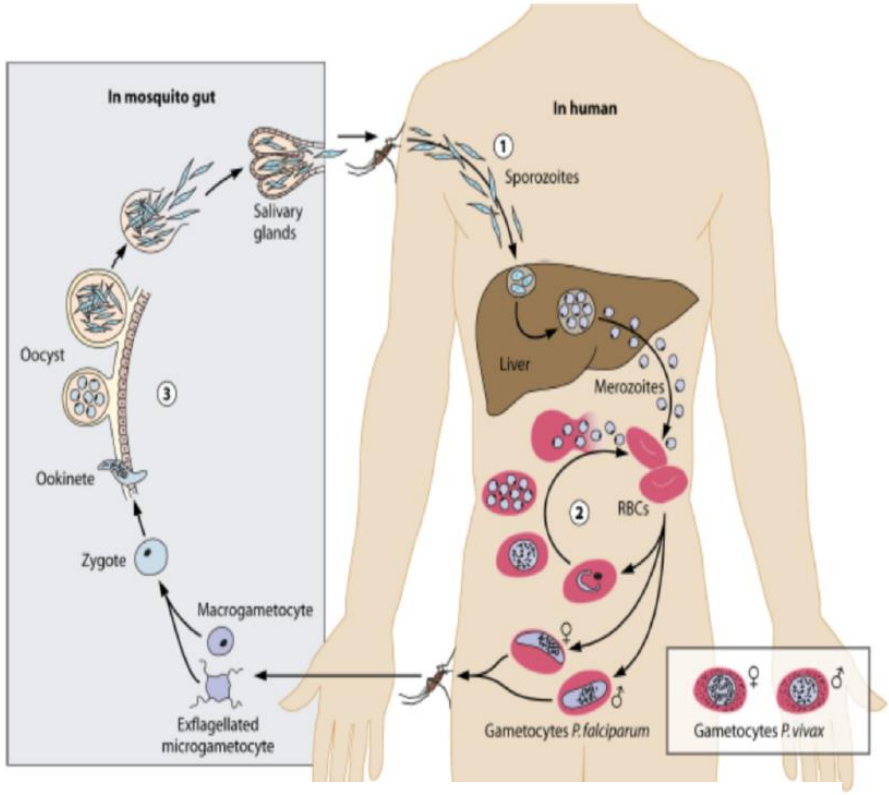
Malaria is a parasitic disease that has afflicted many over the years, with Plasmodium falciparum malaria accounting for many deaths.

The data analysed was obtained from a series of **clinical trials** conducted between **2002 and 2004 in Mpumalanga, South Africa and Mozambique**.

Patients were observed on days **0, 3, 7, 14, 21, 28 and 42**, where blood samples were collected and were analysed, providing gametocyte densities and other information.

This research **aims to directly model gametocytes while taking into account censoring**.

Interval-censored techniques were applied to the data. Several survival analysis models were applied to the gametocyte data. These models included the **Cox Proportional Hazards (PH)** model, **parametric PH** model and **accelerated failure time** models, which were used to illustrate how results may differ based on whether interval censoring was taken into account or not.



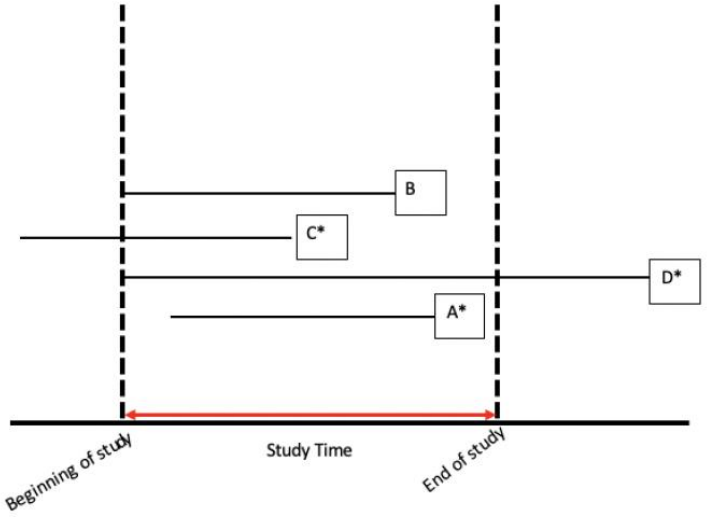
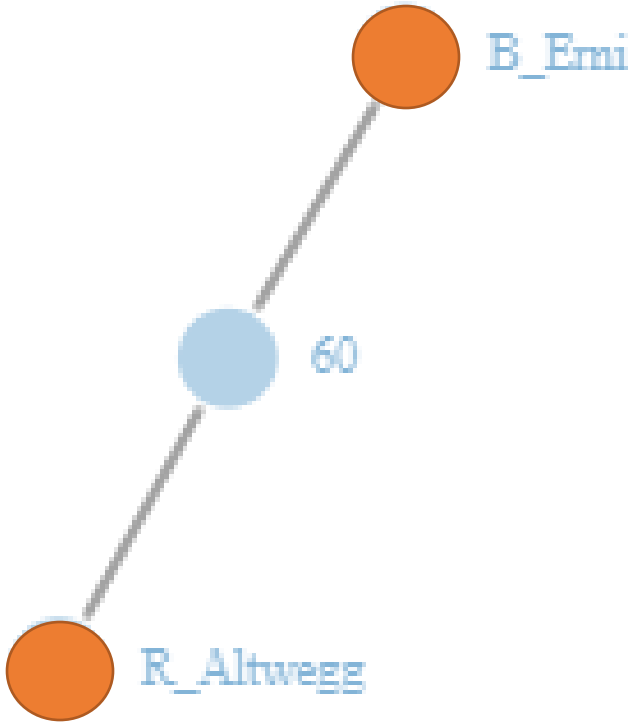


Figure 9: Examples of point censored observations

Res Altwegg and Birgit Erni

- Spatial methods
- Ecological sciences



- Using **state-space time series analysis** on wetland bird species to formulate effective bioindicators in the **Barberspan** wetland (2022

The **Coordinated Waterbird Count dataset (CWAC)** is a dataset containing waterbird counts from wetlands across South Africa, going as far back as 1970.

These data contain valuable information on population sizes and their trends over time.

The aim of this dissertation is **to bridge the gap between the CWAC dataset and the end users** (for both experts and non-experts).

A state-space time series model was applied to the waterbird counts in the CWAC dataset to determine waterbird population trends over the years.

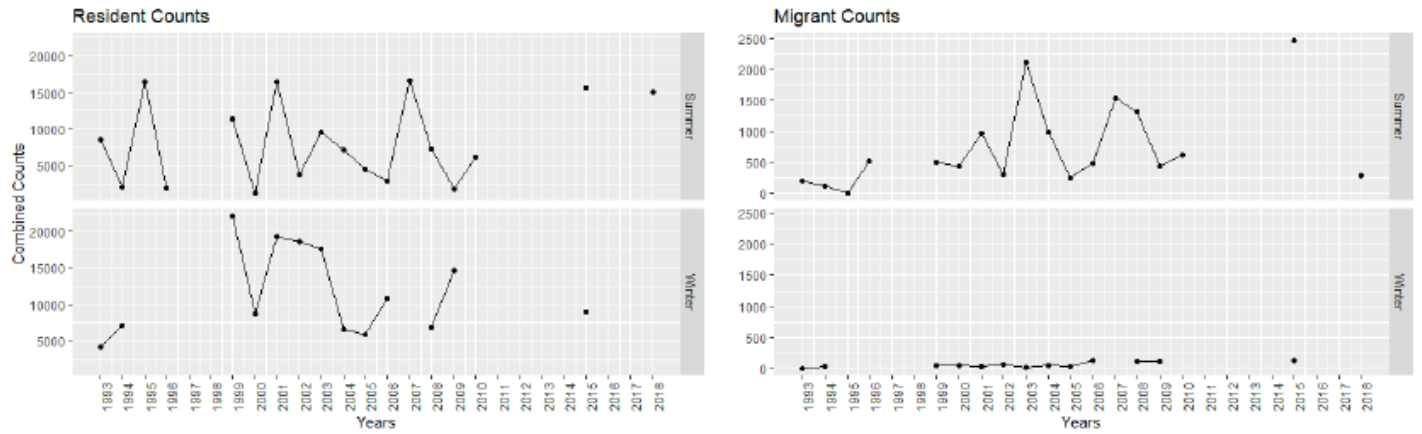
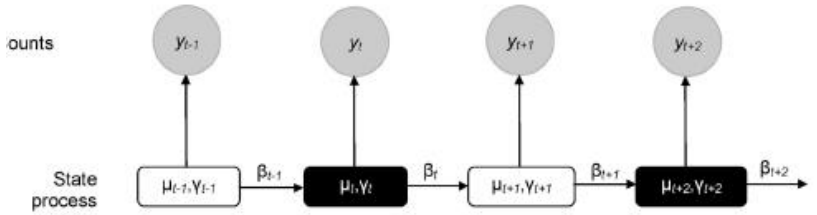
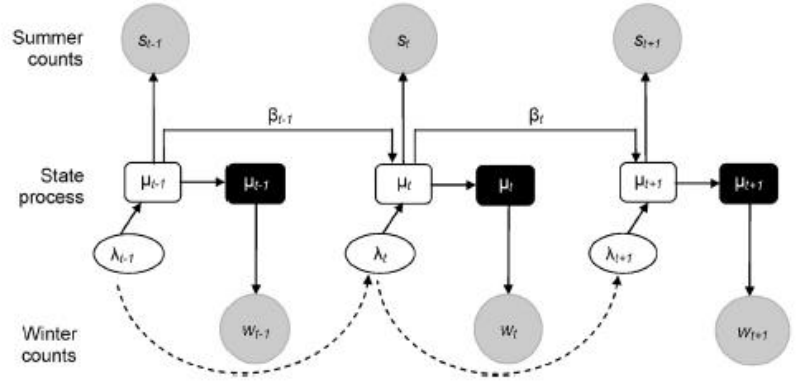


Figure 2: Combined counts of waterbirds per year at Barberspan wetland separated by migratory status with summer counts displayed on the top plot and winter counts displayed on the bottom plot.

(a) Model 1



(b) Model 2



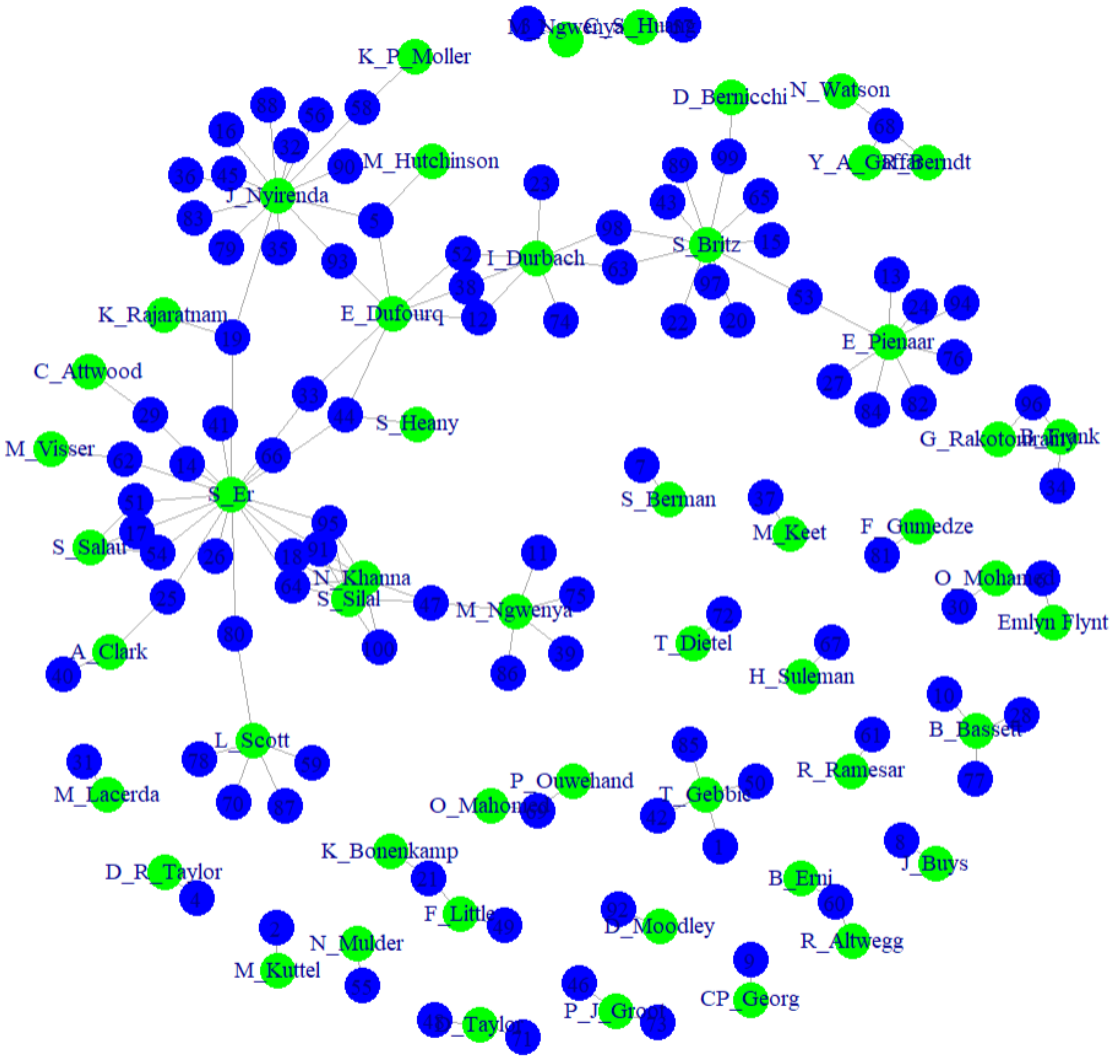
Other projects

- **Optimising** City of Cape Safety and Security resources emergency and policing service requests – supervised by **Leanne Scott**
- **Analytical performance management dashboard** for a South African city - a data science approach – supervised by **Leanne Scott**
- **Analysis** of the effect of course structure and pattern of usage on efficacy of **online/blended courses** – supervised by **Leanne Scott**
- **A Predictive Model** of Ilifu (Ilifu, a big data infrastructure for data-intensive research, enables South African researchers to be world leaders in the strategic science domains of astronomy and bioinformatics. Operated by a consortium of universities and research organisations, ilifu is a node in the national data infrastructure, partly funded by the Department of Science and Innovation to support the National Integrated Cyberinfrastructure System of South Africa) – supervised by **Georgina Rakotonirainy with Brad Frank**
- **Predicting hospital admissions** for members of a medical scheme using **machine learning techniques** – supervised by **Freedom Gumedze**

Conclusions

- We supervise very interesting projects with real life applications

- The supervision network diagram is very sparse
- Find the right expert in the department
- Collaborate
- Get involved
- Publish the good work!



- Thank you all…
- Questions?