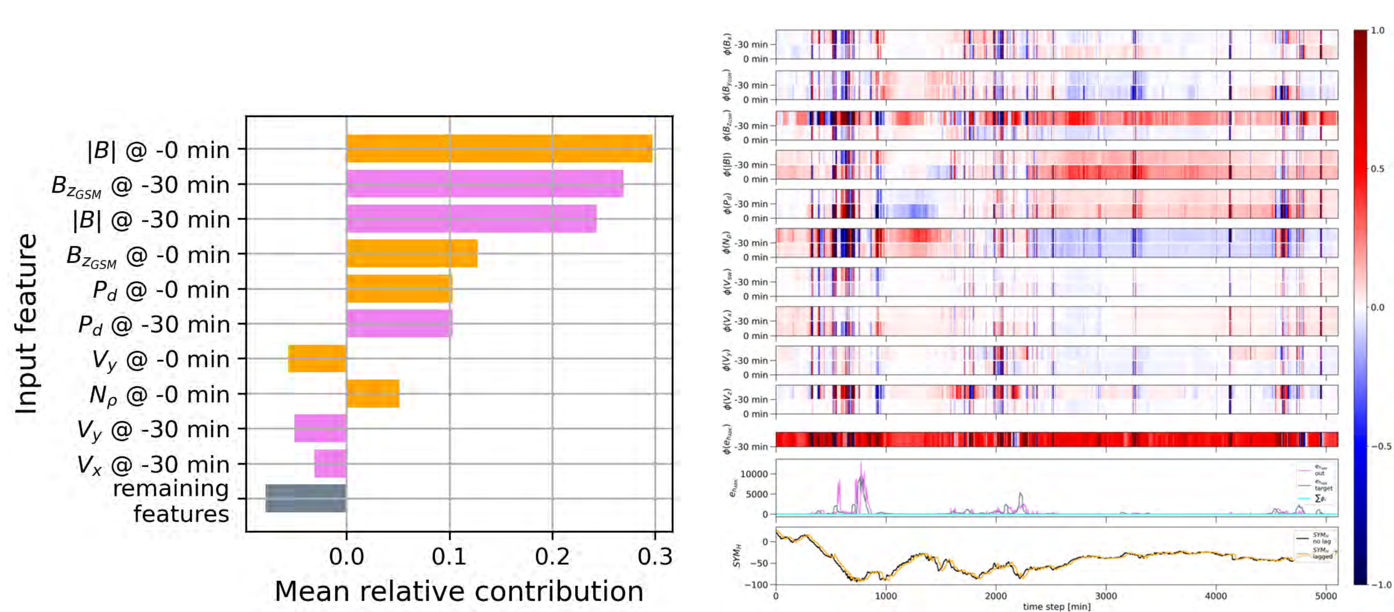
Knowledge Discovery in Time Series Data

Marelie H Davel, Stefan Lotz, Marthinus W Theunissen, Almaro de Villiers, Chara Grant, Randle Rabe, Stefan Schoombie and Cleo Conacher
National Institute for Theoretical and Computational Sciences, North-West University, South African National Space Agency, Centre for Artificial Intelligence Research,University of Cape Town, University of Stellenbosch, South African Grape and Wine Research Institute.

Introduction

- Complex time series data often encountered in scientific and engineering domains.
- Deep learning (DL) is particularly successful here:
  - large data sets, multivariate input and/or ouput,
  - highly complex sequences of interactions.
- Model interpretability:
  - Ability to understand a model's decisions in a given context [1].
  - Techniques typically not originally developed for time series data.
  - Time series interpretations themselves become uninterpretable.
- Knowledge Discovery:
  - DL has potential to reveal interesting patterns in large data sets.
  - Potential to produce novel insights about the task itself [2, 3].
- '*know-it*': Collaborative project that studies knowledge discovery in time series data.



Goal

- Develop a platform that simplifies:
  - the development of time series models,
  - interpreting these models,
  - interpreting the explanations.
- Probe the limitations of current interpretability techniques when applied to time series data, specifically.
- Apply to selected applications.

Platform

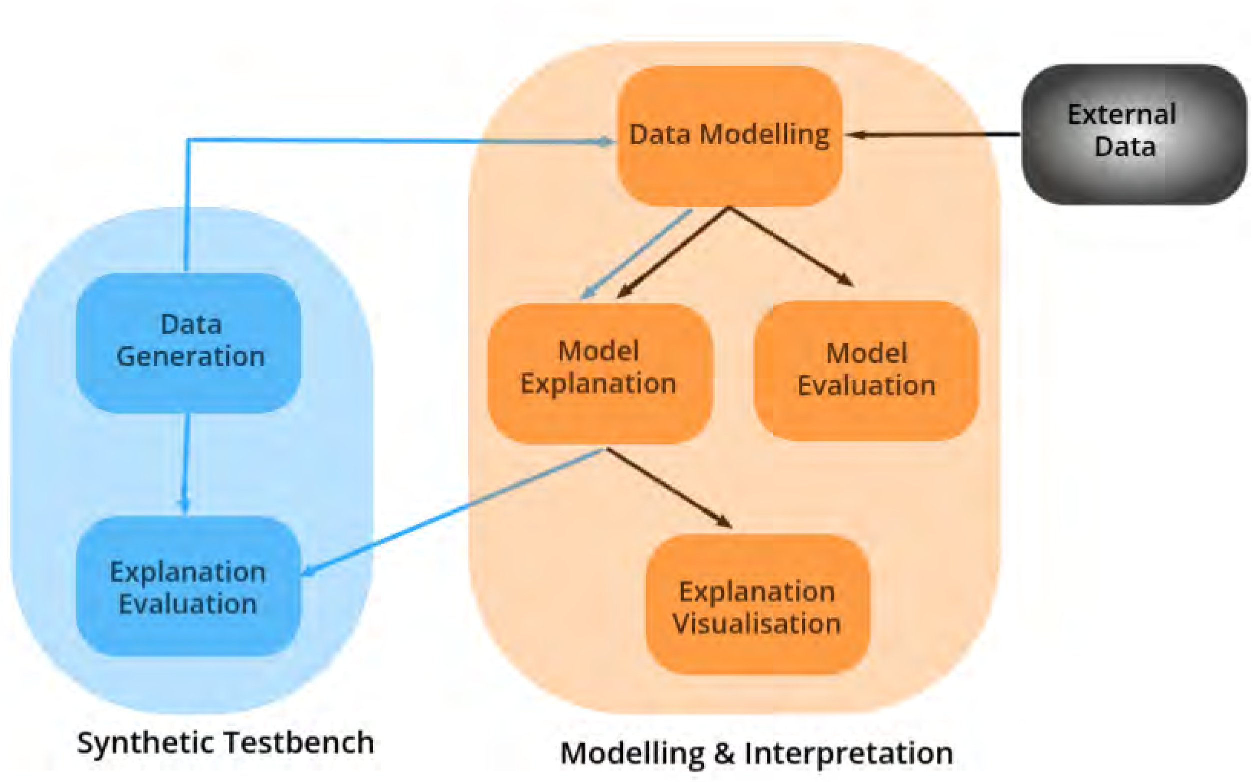


Figure 1: Main modules

- Both existing time series architectures (TCNs, RNNs, LSTMs, temporal transformers) and ability to import own architecture.
- Feature attribution focus currently on SHAP variations [4].
- Not as forecasting-focused (autoregressive) as related projects.
- Summarisation and visualisation still limited.

Synthetic Testbench

- Create time series data with known relationships.
- Testbench generates the data set.
- User uses *know-it* or external system to construct accurate model and extract explanations.
- Testbench evaluates explanations against known ground truth.

Synthetic Data

- Synthetic data should reflect real-world time series data.
- User configurable:
  - *generating function:* specifies underlying feature distributions,
  - underlying *co-variances* among input parameters,
  - *transfer function*: transforms input to output,
  - *time dependencies* between parameters.

Applications

- Environmental monitoring: Penguin prey capture events from bird-borne data loggers.

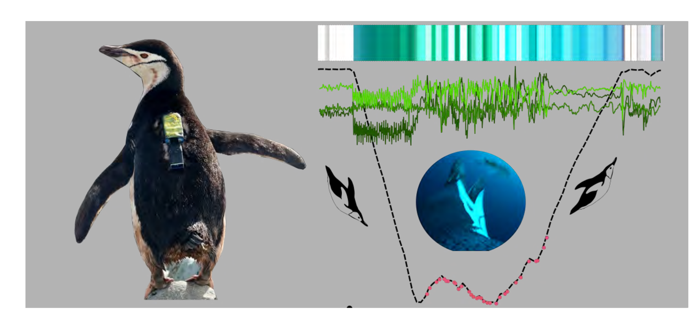


Figure 2: Accelerometer data logged by chinstrap penguins. Video confirmation used to train models in a supervised fashion.

- Microbial ecology: Modeling population dynamics of the wine yeast community through time, where yeast-yeast interactions affect fermentation outcomes.

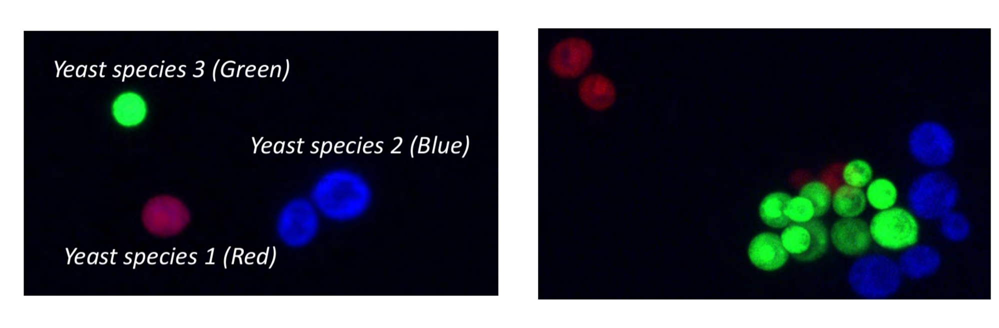


Figure 3: Microscopic images of distinctively labelled yeast species interacting in a constrained growth medium.

- Space weather tracking: Geomagnetic index prediction from solar wind data.

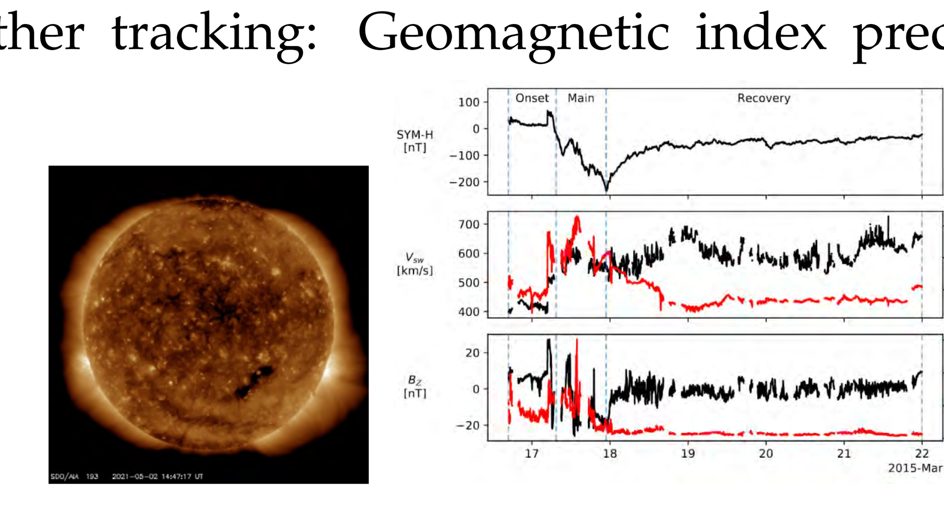


Figure 4: Predicting the effect of a solar storm on Earth's geomagnetic field.

References

[1] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 2019.

[2] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[3] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[4] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 4766–4775, 2017.