

Applicant: **Kiraly, Franz**

Funding Sought: **£308,575.06**

TPS2019\100078

PRIMARY APPLICANT DETAILS

Name	Franz
Surname	Kiraly
Email (Work)	fkiraly@turing.ac.uk

Section 3 - Submission Form

What is your work package title?

Turing time series modelling toolbox – sktime phase II

Lead Investigator:

Franz Kiraly

Co-Investigators:

Anthony Bagnall, Jason Lines, Sebastian Vollmer

Name of PI/Co-I present at scoping workshop (or Turing Director):

Franz Kiraly

Earliest possible start date

01 June 2019

Latest possible end date

31 December 2020

Duration (in Months):

18

Please select which one of the following project types your work package fits under.

☒ Tools

Please can you check which theme areas your work package addresses or aligns to. Please check up to 3 theme areas.

☒ Health

☒ AI for Science

Should you wish to include these, please attach any letters of support for your work package here. (PDF is the best format)

📄 **Sktime Application to SPF Tools, Practices and Systems call - CHolmes V2**
📅 29/04/2019
🕒 15:09:22
📄 pdf 111.27 KB

📄 **LoS4**
📅 29/04/2019
🕒 15:07:37
📄 pdf 436.62 KB

📄 **LoS1**
📅 29/04/2019
🕒 15:07:30
📄 pdf 436.62 KB

📄 **LoS5**
📅 29/04/2019
🕒 15:07:15
📄 pdf 79.36 KB

📄 **LoS2**
📅 29/04/2019
🕒 15:07:42
📄 pdf 886.55 KB

📄 **LoS3**
📅 29/04/2019
🕒 15:07:36
📄 pdf 482.79 KB

📄 **Contributors**
📅 29/04/2019
🕒 15:07:22
📄 pdf 182.54 KB

Abstract:

Learning with time series and temporal data is crucial to many applications, across wide areas of research in engineering, finance, health, the natural science, and many others. Open source capabilities in dealing with such data is limited, leading to unnecessary replication of coding work, or technically inappropriate (and therefore error-prone) reduction to cases off-shelf toolboxes can deal with (e.g., tabular data). One of the reasons for this situation is that there is a broad variety of learning tasks that arise in the context of temporal data. Unlike for “classical” supervised learning, there is not a “one-interface-fits-all” approach, and it necessitates development of a meta-language for model building and checking. To be more precise, some key tasks with stylized application areas are as follows (in rough order of technical complexity):

- (i) Supervised learning with time series features, including time series classification and regression – event classification in neuroscience, physics, object/motion recognition
- (ii) Time series annotation (supervised and unsupervised), including anomaly detection, segmentation – intensive care and medical monitoring, equipment health monitoring
- (iii) Forecasting (supervised and unsupervised) – trajectory prediction, weather/climate forecasting, ecosystem modelling, supply/demand (e.g., energy) forecasting
- (iv) Event modelling including survival modelling – electronic health records, clinical studies with survival outcome, behaviour modelling, predictive maintenance

The “optimal” solution to the issue would provide an interoperable system for modelling strategies and pipelines for each of the above, especially since the tasks are algorithmically related: for example, a forecasting strategy may be constructed by tabulating sliding windows and applying a time series regression method; or, an event modelling strategy may be obtained from binning, tabulation, and application of a supervised learning method.

Primary requirements for such a set of interoperable “time series” modelling toolbox modules are:

- (a) Availability of modelling atoms under a task-specific unified interface that exposes hyper-parameters and inference results. E.g., ARIMA models for forecasting, GLM panel regression for supervised forecasting,

Cox PH models for survival

(b) Abstract composition methodology for tuning, pipeline building. E.g., grid search or Bayesian optimization of hyperparameters.

(c) Abstract reduction methodology, i.e., meta-learning that mutates the task. E.g., leveraging supervised learning for forecasting, or time series classification for event modelling.

(d) Orchestration of model evaluation and success control workflows, of modelling strategies against data.

This proposal suggests to build on outcomes of the sktime project (Dec 2018 – May 2019) in order to complete the above vision, leveraging an expanding community of contributors, network of application case studies, co-development with SPF themes, and a consolidated code base that covers the basics of use case (i), supervised learning with time series features, in the form of a python toolbox compatible with the sklearn and pydata ecosystems.

An expanded scale, and a sufficient amount of manpower, would put use cases (ii) - (iv) in reach of development (see "research").

In addition, we anticipate close interlinkage and co-development with projects in the health programme (see "impact").

Route to Impact

Our goal is for sktime to facilitate the rapid development of good solutions to a range of problems using state-of-the-art algorithms assessed by a rigorous and unbiased framework of evaluation.

Our primary route to impact is to find domain project where the toolkit can be used to add value to existing research, and in turn inform development of the toolkit. We have a series of agreements with domain experts:

External project partner Modelling tasks

MRC Cognition & Brain Sciences Unit of the University of Cambridge (LoS1, LoS2.pdf) Early detection of dementia from MEG/EEG – multivariate time series classification

Great Ormond Street Hospital (LoS3.pdf) Predictive modelling using Intensive Care Unit data – panel data prediction

UEA physics (LoS4.pdf). Detecting radio-frequency interference (RFI) in astrophysics data – classification & annotation

French Agence Nationale de la Recherche (LoS5.pdf) Environmental monitoring (pollution, forests, agriculture) – classification, anomaly detection

Since the supervised learning module is, at the time being, the most developed, we will initially concentrate on problems of classification and regression. However, these four applications have potential problems related to all four tasks, and we will work with partners to create a range of use-cases.

In addition, the health programme has confirmed its support for co-development and deployment for tasks (ii) and (iv), and will provide additional time of a co-PI in-kind, dedicated to interfacing with the projects in the domain. Time-to-event modelling is a key task appearing pervasively in the domain's projects, and so is annotation of time series in the form of change-point/outlier detection and segmentation (see "connections" for a detail list).

Research challenges :

Besides sitting in the nexus of a wide range of application research challenges, there are a number of research challenges within the tooling and methodology development domains that the project will need to address. Some of the major open questions are stated below:

Tooling:

- (1) Design of unified object oriented interfaces for modelling strategies: within the key tasks (i)-(iv), and across the tasks. Design of a data-task-strategy interface.
 - (2) Design of appropriate data containers for linked and hierarchical data sets in which multiple samples of time series and multi-variate time series may be present. Typing formalism for “columns/variables” that are time series, sequence, or annotation valued.
 - (3) Interface formalism for model composition, tuning, pipelining, task reduction: how to expose hyper-parameters, deal with atoms of different task type, mutate interfaces.
- Can one define a (user-friendly) first-order type language for model building?

Methodology:

- (1) What are candidates for widely well-performing methods for tasks (i)-(iv)? Due to absence of off-shelf toolboxes, large-scale benchmark studies have so far not been conducted (timeseriesclassification.com and the Makridakis studies being notable exceptions)
- (2) What are robust and theoretically justified quantitative workflows for model assessment and model validation? What is the set of “simple baselines” for each task, especially considering “simple” reduction strategies, e.g., in (i) “tabulate and use sklearn”.
- (3) As a specific class of models, the toolbox should eventually allow construction of hybrid methods and composites using automated feature engineering and deep learning atoms. The performance and usefulness of such hybrids would be interesting to investigate.

Community benefit:

The sktime project has attracted a community of method contributors and application end users across six countries (see Contributors.pdf for details).

The toolkit will also play a central role in two EPSRC project proposals that will be submitted in 2019. JL will submit a new lecturer grant proposal to develop novel algorithms for on chip classification of streaming time series. AB will submit an EPSRC responsive mode grant on multivariate time series classification. Both will involve research using the sktime toolkit. The toolkit will be used on a range of existing projects that UEA are involved with, including an EPSRC CASE award with BT, a BBSRC case award with Scotch Whisky Research Institute and an ANR project at Rennes 2.

The sktime team are keen to develop better communication channels both internal and external to the Turing. We will instigate a formal mechanism for information sharing between development teams at the Turing to help foster a sense of community and spread best practice. We will set up regular surgery activities to allow ongoing projects to ask us what we could do for them, and to request guidance in dealing with time series data.

In addition, project development is open on GitHub, and we aim to further integrate with members of the pydata user and developer community, or Turing projects who consider project outputs useful.

Connections to other activities :

Natural connections with Turing internal projects are existing modelling tools efforts at the Turing (Shogun, Julia, R/mlr), and systems projects which involve an aspect of model building or model appraisal in the context of time series.

In addition to that, reproducible practices projects (such as the Turing Way) are natural partners as modelling toolboxes providing the basis for reproducible analyses through standardized code and workflow components.

In-principle, any project at the Turing in which time series or temporality is a core aspect has the potential to enter into a synergistic relationship.

Furthermore, as mentioned above, the health themes has confirmed its support, with co-development and deployment potential in the following SPF projects:

Project title Modelling tasks

Scottish Patients at Risk of Readmission - Part 2 time-to-event modelling, time series features
NHS Digital project time-to-event modelling, time series features
Biobank algorithm - Classify Fitbit data with parallel TensorFlow Time series annotation: segmentation, event detection
Development of a learning machine for supporting decision-making for clinicians time-to-event modelling, electronic health records
UCLC projects/partnership time-to-event modelling, scheduling

References and URLs:

No Response

Please can you fill in the costing sheet for your work package.

📎 **COSTINGBREAKDOWN timeseries**

📅 29/04/2019

🕒 21:13:47

📄 exe 21.22 KB

Section 4 - Equal Opportunities Monitoring and Reasonable Adjustments

How did you hear about this invitation to submit work packages ?

TPS workshop

Please indicate below if you would like to request or discuss any reasonable adjustments to the application process.

No Response

What is your date of birth?

12 January 1986

Do you have an impairment, health condition or learning difference that has a substantial or long term impact on your ability to carry out day to day activities? (tick all that apply)

☒ Prefer not to say

What is your gender?

☐ Male

What is your sexual orientation?

☐ Prefer not to say

What is your ethnic group?

☐ Prefer not to say

Do you have a religion or belief?

☐ No religion

Does your gender identity match your sex as registered at birth?

☒ Yes

Information about gender identity is considered sensitive personal data under the Data Protection Act. We want to make sure that we have permission to store this data for the purposes of monitoring and advancing equality and diversity in higher education. Please indicate if you give us permission to store this information and use it in this way.

☒ Yes

Do you have any caring responsibilities? (tick all that apply)

☒ None



UNIVERSITÉ RENNES 2
SKOL-VEUR ROAZHON 2

CAMPUS VILLEJEAN

Place du recteur
Henri Le Moal CS 24307
35043 Rennes cedex
France

T +33 (0)2 99 14 10 00
www.univ-rennes2.fr

**Letter of support for the
Turing sktime toolkit**

**Rennes,
26th April 2019**

Romain TAVENARD
Assistant Professor
romain.tavenard@univ-rennes2.fr

To whom it may concern,

I am primary investigator for the 48-month ANR project titled MACHine learning for environmental Time Series (MATS) which begins in April 2019 (Grant ANR-18-CE23-0006, <https://anr.fr/Project-ANR-18-CE23-0006>). This project focusses on developing novel machine learning algorithms for three key environmental issues: agricultural practices and their impact; forest preservation; and air quality monitoring based on dedicated datasets made available to the MATS research team through Kalideos project and Zone Atelier Armorique.

Jason Lines and Tony Bagnall are named collaborators on the MATS project and will contribute their expertise in time series classification. The MATS project was proposed and funded prior to the start of the sktime project and is based on a python toolkit named tslearn developed at University of Rennes 2. The functional overlap between tslearn and sktime is minimal, and the interfacing of these two packages will benefit the MATS project. Specifically, it will allow for more extensive benchmarking against state of the art (work package 1.2) and facilitate the more rapid development of a wide range of semi-supervised techniques (3.1) and metric learning (3.2) which will enhance the analysis of the datasets.

I fully support the continued development of the sktime toolkit and believe it will enhance the quality of the science on the MATS project.

**Romain Tavenard
Principal Investigator for ANR MATS project**

CAMPUS LA HARPE
Avenue Charles Tillon
CS 24414
35044 Rennes cedex

CAMPUS MAZIER
2, Avenue Antoine Mazier
22015 St-Brieuc cedex 1

Task		
(i)	Supervised learning with time series features	
(ii)	Time series annotation (supervised and unsupervised), including anomaly detection, segmentation	
(iii)	Forecasting (supervised and unsupervised)	
	Collaborations and Contributors	Role
	1 Professor Geoff Webb	Director of Monash University Centre for Data Science, Australia
	2 Dr Francois Petitjean	ARC DECRA Fellow, Senior Lecturer (Data Science)
	3 Ahmed Shifaz	PhD student, Monash, Australia
	4 Professor Jose Lozano	Scientific Director of the Basque Center for Applied Mathematics, Spain
	5 Amaia Abanda	PhD student, BCAM
	6 Professor Germain Forestier	Professor at Univ. of Haute-Alsace, IRIMAS, France
	7 Hassan Fawaz	PhD student, Haute-Alsace
	8 Dr Romain Tavenard	Assistant Professor, Rennes 2, France
	9 Professor Eamonn Keogh	University of California, Riverside, USA
	10 Anh Dau	PhD student, UCR, USA
	11 Yan Zhu	Google, USA
	12 Dr Jessica Lin	Associate Professor, George Mason, USA
	13 Yifeng Gao	PhD student, George Mason, USA
	14 Xiaosheng Li	PhD student, George Mason, USA
	15 Li Zhang	PhD student, George Mason, USA
	16 Dr Diego Silva	Assistant Professor, São Carlos, Brazil
	17 Dr Gavin Cawley	Senior Lecturer, UEA, UK
	18 George Oastler	PhD student, UEA, UK
	19 Michael Flynn	PhD student, UEA, UK
	20 James Large	PhD student, UEA, UK
	21 Matthew Middlehurst	PhD student, UEA, UK
	22 David Guijo	PhD student, Cordoba, Spain
		Area of Contribution
		(i) Proximity Forest for multivariate
		(i) Proximity Forest/Optimised DTW
		(i) Proximity Forest/TS-Chief
		(i) Development of kernel methods
		(i) distance based kernel methods
		(i) Full integration with keras
		(i) Full integration with keras
		(i) Integration of shapelets and CNN
		(ii) anomaly detection and segmentation
		(ii) matrix profile
		(ii) matrix profile
		(ii) Anomaly detection, motifs, segmentation
		(ii) Anomaly detection
		(ii) motifs
		(ii) segmentation
		(iii) forecasting
		(i) Kernel methods
		(i) Distance based classification
		(i) Spectral based classification
		(i) Dictionary based classification
		(ii) Anomaly detection
		(i) Shapelet based methods

17/04/19

To whom it may concern,

I am an MRC Programme Leader at the MRC Cognition & Brain Sciences Unit of the University of Cambridge, and been in discussion with Prof Tony Bagnall at the University of East Anglia. The purpose of this letter is to offer my support for the application of the Turing sktime toolkit to problems in the health domain, with specific focus on the MEG/EEG classification component of the project. We wish to determine whether MEG can be used as a diagnosis tool for the early detection of dementia, and the sktime toolkit offers us the opportunity to try a range of novel algorithms that have not been applied to this domain before. Over the last few years, we have accumulated a unique database of resting-state MEG data from approximately 100 patients with Mild Cognitive Impairment (MCI) – a potential prodromal stage of dementia – plus over a 100 age- and sex-matched controls. MEG offers more spatial degrees of freedom than previous datasets using EEG. Moreover, to study the effects of healthy ageing, we also have the same data on a normative sample of nearly 650 adults from the CamCAN project (www.cam-can.org). There is already a data-sharing agreement between the universities of Cambridge and East Anglia, so the data can be provided promptly on demand. I will also offer my expertise in which features of the MEG data are likely to be of most relevance to the project.

Yours faithfully,



Prof Richard Henson

24th April 2019

To whom it may concern,

Collaboration on Clinical Intensive Care Data Project

The Great Ormond Street Hospital Intensive Care Units and the UCL Great Ormond St Institute of Child Health Critical Care Group recently participated in a Data Study Group (DSG) at the Alan Turing Institute. The data was multivariate time series/panel data with measurements for vital body functions (heart rate, blood pressure, breathing rate, etc) for individual patients and additional time-constant information such as gender, age, hospital ward, hospital admission date, extubation date, reintubation date. The main challenge was to predict whether extubation was successful given the history of vitals previous to the extubation, where extubation was considered successful if the patient had not been reintubated within 48 hours of extubation. One of the DSG participants, Markus Loning, is part of the sktime team, and used sktime in the DSG. It became apparent that the sktime package would be useful for exploring this data.

We are interested in continued collaboration with the Alan Turing Institute and we see supporting the sktime toolkit as one mechanism for doing this. Specifically, this will involve providing continued access to the ICU data via the secure storage at the Turing Institute (subject to standard approval processes, see below). We will advise on formatting and processing the data as a multivariate classification problem to predict whether an extubation is successful or not.

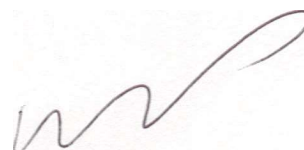
We also support the exploration of further potential applications of the toolkit to our data for classification, forecasting, anomaly detection and time to event modelling.

The data approval process of the Data Access Request Service has already been performed in order for the data to be used in the DSG. Repeating the process, if it is indeed required, should be a formality.

Yours faithfully,



Samiran Ray
Consultant
Paediatric Intensive Care Medicine



Mark Peters
Professor
Paediatric Intensive Care Medicine

17/04/19

To whom it may concern,

I am an MRC Programme Leader at the MRC Cognition & Brain Sciences Unit of the University of Cambridge, and been in discussion with Prof Tony Bagnall at the University of East Anglia. The purpose of this letter is to offer my support for the application of the Turing sktime toolkit to problems in the health domain, with specific focus on the MEG/EEG classification component of the project. We wish to determine whether MEG can be used as a diagnosis tool for the early detection of dementia, and the sktime toolkit offers us the opportunity to try a range of novel algorithms that have not been applied to this domain before. Over the last few years, we have accumulated a unique database of resting-state MEG data from approximately 100 patients with Mild Cognitive Impairment (MCI) – a potential prodromal stage of dementia – plus over a 100 age- and sex-matched controls. MEG offers more spatial degrees of freedom than previous datasets using EEG. Moreover, to study the effects of healthy ageing, we also have the same data on a normative sample of nearly 650 adults from the CamCAN project (www.cam-can.org). There is already a data-sharing agreement between the universities of Cambridge and East Anglia, so the data can be provided promptly on demand. I will also offer my expertise in which features of the MEG data are likely to be of most relevance to the project.

Yours faithfully,



Prof Richard Henson

Faculty of Medicine and Health Sciences
Norwich School of Medicine

Dr. Saber Sami (MED)
Senior Lecturer in Dementia Research
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom

S.samil@uea.ac.uk
Tel: +44 (0) 1603 597175
Fax: +44 (0) 1603 593752

26 April 2019

Dear Sir/Madam,

I'm writing to provide a letter of support for the proposed Turing time series modelling toolbox (sktime).

Based on nearly two decades of work in brain electrophysiology and imaging and as the current lead in this area in the department of Medicine at the University of East Anglia; I can clearly see the need for this toolbox. The conceptual approach envisioned by Bagnall et al. would dramatically help improve the efficacy, robustness and standardisation of sensor based time series analysis in a wide range of clinical settings we are currently testing. Recent advances in image based machine learning and standardisation efforts in computer vision have already led to improved and more robust diagnostic accuracy and has become the norm in today's evolving personalised medicines approach. However, clear standards and benchmarks are missing for widely adopted sensor based techniques like e.g. Electroencephalography (or EEG).

EEG is currently one of the most widely used non-invasive brain imaging tools in neuroscience and in the clinic, but surprisingly little clinical use has been made with the development of recent machine learning based algorithms for the detection and prediction of cognitive outcomes. My recent work in brain connectivity and machine learning for dementia patients shows that EEG could be explored as a biomarker of pathophysiologies and has the potential to predict treatment success likelihoods as well as providing an avenue for neuro-feedback/brain-computer based interventions. The possibility to integrate multiple sensor recording with advanced machine learning capabilities through this initiative will enhance clinical diagnostics and provide wider opportunities that can be applied to the rigorous standards for key endpoints in multi-site clinical trials. We aim to use the newly developed toolbox in our future dementia hackathon events co-hosted with the Dementia Platform UK (Oxford University) and the Alan Turing Institute to enable new clinical interventions.

In summary, more efficient and robust algorithms will mean less animals/human will need be tested in future trials. This initiative will provide funding for high-quality UK-based research projects focused on addressing grand challenges in neuroscience and neurodegeneration as part of larger international network initiatives.

Yours Faithfully,



Dr. Saber Sami

Selected Publications

Sami S, Hughes LE, Williams N, Cope T, Henson R, Rowe JB Neurophysiological signatures of Alzheimer's disease and Frontotemporal lobar degeneration: pathology versus phenotype *Brain*. 2018;141(8):2500-2510..

Passamonti L, Vázquez Rodríguez P, Hong YT, Allinson KS, Williamson D, Borchert RJ, **Sami S**, Cope TE, Bevan-Jones WR, Jones PS, Arnold R, Surendranathan A, Mak E, Su L, Fryer TD, Aigbirhio FI, O'Brien JT, Rowe JB. 18F-AV-1451 positron emission tomography in Alzheimer's disease and progressive supranuclear palsy. *Brain*. 2017 Mar 1;140(3):781-791

Genicot M, Absil PA, Gousenbourger PY, Lambiotte R, **Sami S**. Coupled Tensor Decomposition: a Step Towards Robust Components *IEEE proceedings of EUSIPCO* 2016

J. Borchert, R., Rittman, T., Passamont, L., Zheng, Y., **Sami, S.**, P. Jones, S., Nombela, C., Vázquez Rodríguez, P., Vatansever, D., Rae, C., E. Hughes, L., Robbins, T. W., B. Rowe, J. (2016) Atomoxetine Enhances Connectivity of Prefrontal Networks in Parkinson's Disease, *Neuropsychopharmacology* 41(8): 2171–2177

Sami, S., Robertson, E. M., & Miall, R. C. (2014). The Time Course of Task-Specific Memory Consolidation Effects in Resting State Networks. *The Journal of Neuroscience*, 34(11), 3982–3992.

The Alan Turing Institute

By email

Monday 29 April 2019

To whom it may concern,

This is to evidence the health programme's support for the time series toolbox proposal.

Much of the data that we consider in the health programme, such as electronic health records or clinical study data, involve a temporal sampling process. Examples include, longitudinal observations on patients, or a temporal outcome such as survival or time-to-event. This is almost a defining feature of health data.

Solid tooling for this situation is rare, especially if one wishes the model to ingest information or update predictions as time passes for the patient, or when occurrence of clinical events is to be modelled. The most critical issues with existing toolboxes are their assumption of a tabular presentation, and the inability to build probabilistic temporal models.

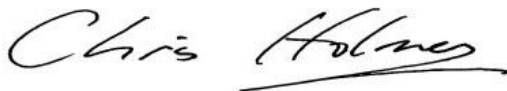
Insofar, we are quite enthusiastic about the toolbox proposal to extend Sktime, particularly the work packages on (not yet existing) probabilistic and event modelling functionality, which seems to be a crucial prerequisite for rapid prototyping and testing of clinically relevant models for electronic health records, such as fully Bayesian modelling pipelines for temporal event modelling.

As material support, the health programme will be happy to contribute 5% of Sebastian Vollmer's time (above the 5% for co-PI) towards co-development of critical toolbox functionality. The main purpose is identifying and leveraging use cases from within the health programme, contributing requirements, informing design and implementation as well as test-wise deployment.

Potential use cases with immediately spring to mind is Louis Aslett's SPARRA project with NHS Scotland, the collaboration with David Llewellyn and the dementia network, the collaborations with UCLC and NHS digital, or Mihaela van der Schaar's projects.

We are also open to recommend to project members to contribute modelling code to Sktime, though this would of course be conditional on a strong case of benefitting their home project.

Best wishes,



Professor Chris Holmes
Director, Health Programme
The Alan Turing Institute

The Alan Turing Institute

The British Library
96 Euston Road
London NW1 2DB

+44(0)30 0770 1912
info@turing.ac.uk
turing.ac.uk