Eventful: timely models for individual & societal health

Andrea Sipka[1], Andreas Dengel[2], David Selby[1], Ece Calikus[1], Muhammad Nabeel Asim[2], Rahul Sharma[1] and Sebastian Vollmer[1]

[1]Data Science and its Applications
[2]Smart Data & Knowledge Services

January 2023

eventful Framework
Resources for effective event modelling in practice

eventful Methods
Developing models for event-related tasks

eventful Impact
Enabling and demonstrating impact

Executive summary

An 'event' can describe any state with a timestamp. Some events are directly observable (death, crime reports, flooding) while others represent latent changes in an underlying process (flares in disease, social unrest, a cyber-attack). Predicting events—when, whether and where they occur—is an increasingly important task in science, industry and public policy.

This project focusses on the interconnected problems of measuring and predicting health and well-being. Health and well-being exist at two levels: within the individual, where we may observe or measure a patient's genetics, medical history and life course; and at the societal level, where the effect of the physical and socio-political environment might be apparent. Many public health epidemiologists hold that societal welfare and individual well-being are intrinsically linked. It may follow, therefore, that learning the dynamics of socio-political events can aid understanding of individual well-being, and *vice versa*.

We consider the world of events from two perspectives. 'Events as output' encompasses time-to-event analysis, where the task is to estimate the risk of an (adverse) event or the time until it occurs. 'Events as input' encompasses intensive longitudinal data analysis or the analysis of irregularly-spaced time series. In both areas, a key challenge is extracting useful features from high-dimensional, complex or dependent data structures to predict an outcome, be it dynamic or static. Existing machine learning tools and methods may need to be extended and adapted to handle such tasks.

The project explores methods of selecting and extracting features from whole genome data for survival problems (individual health), the effect of different levels of data aggregation on spatio-temporal models (societal health) and the challenges behind building a transparent, explainable and useful data science pipeline for such complex temporal data. A particular focus is paid to automation of the data cleaning process, as well as to effective combination of scientific and machine learning knowledge on problems where data are messy, sparse or only available in aggregate form.

Utility of the developed models and pipelines will be demonstrated through the collation of appropriate benchmarks and open data sets and application to several real-world problems, ultimately culminating in implementation as tools designed for the use of stakeholders and policymakers.

Contents

1 Motivation and goals

1.1 Motivation and vision

Health arises from interactions of individual, societal and environmental factors, which develop dynamically over a person's life course[1]. An individual's wellbeing can be dramatically affected not only by biological processes but also the occurrence and proximity of socio-political events. Epidemiology for public health, therefore, depends on robust understanding of both *individual-level* and *societal* trajectories to be able to explain the causes of adverse events and the effects of different interventions.

However, individual-level data, such as that captured in electronic health records, typically comprise an series of measurements sampled irregularly through time, often characterised by missing or erroneously recorded values. Meanwhile, at the societal level, many models do not yet take full advantage of the granularity of spatio-temporal data that has (recently) become available. In both cases, data may be extremely high-dimensional and it is often not apparent to analysts which features are likely to be relevant or which data processing pipeline may be most appropriate to the task at hand. Moreover, whilst machine learning frameworks for classification and regression problems are fairly well-developed, relatively little attention has been paid to facilitating transparent analysis of irregular, longitudinal time series or time-to-event data.

In this project, we work towards bridging these gaps. This involves a thorough review of the complete data analysis pipeline for individual and societal health modelling. Firstly, we evaluate the availability of data sets and tools for the tasks of modelling socio-political events and of survival analysis with multi-omic features. Automated data preprocessing pipelines will be compared, highlighting particular shortfalls in algorithmic cleaning of longitudinal, survival and spatio-temporal data sets. Additional attention is paid to feature extraction, both for 'events as input' (irregular time series) and 'events as output' (survival analysis for full genome data), and in the case of spatio-temporal tasks, the effect of binning (i.e. discretizing time and geolocation). Contrasting with so-called 'model-free' approaches to feature processing, we posit that generative models may perform better at handling missing, biased and unaggregated data and extracting knowledge from dynamic time series (e.g. patient health trajectories), and will test this hypothesis. Neither paradigm is necessarily useful or trustworthy, however, without the resulting models being explainable and transparent; to this end, we develop causal models to explain the effects (and their direction) of different social, genetic and environmental factors on individual and societal health.

Methods and results will be made reproducible and accessible through the release of an extensible, open source scripting library (Python, Julia and/or R) for event feature extraction and to cap off the project we propose a *useful* and *usable* tool for exploring the relative effects of individual and social events on health, in the form of an interactive Web interface, on simulated or open data.

1.2 Application domains

The Eventful project is motivated by employing AI to improve public health, characterized by two cross-cutting themes: **individual health** (within-subject), and **societal health** (or well-being). These two domains were chosen for their complexity, societal importance and potential benefits both to society at large and many possible industrial applications. We present a view on how two areas contrast in Figure 1. We propose two concrete test beds for these domains, and have identified and agreed in principle partnerships with relevant high-impact researchers, institutions and organisations. While a major part of this proposal focuses on methodological work, we use these test beds to demonstrate and test developed methodology and demonstrate impact.

Societal health: Impact of compound crises on youth in Lesotho Globally, we face a future of compound crises: extreme climate hazards, new pandemics and political polarisation leading to increased warfare and civil unrest. Understanding and modelling these crises is essential for mitigating the impacts on all aspects of life, including health, education and risk from violence. The impact of compound crises is even greater for populations that are already disadvantaged or particularly vulnerable (youth, women, global south). Informing data-driven evidence based policy making and interventions will require new approaches to multilayer modelling, taking into account economic, political, geographical, climate, and myriad of other event types. Rarely are these events recorded to same level of geographical or temporal granularity, adding an additional level of complexity. As we tackle problems from a methodological point of view, we intend to demonstrate their impact on a societally important test case of vulnerable youth in Lesotho. We have agreed in principle a collaboration with academics at University of Oxford (Lucie

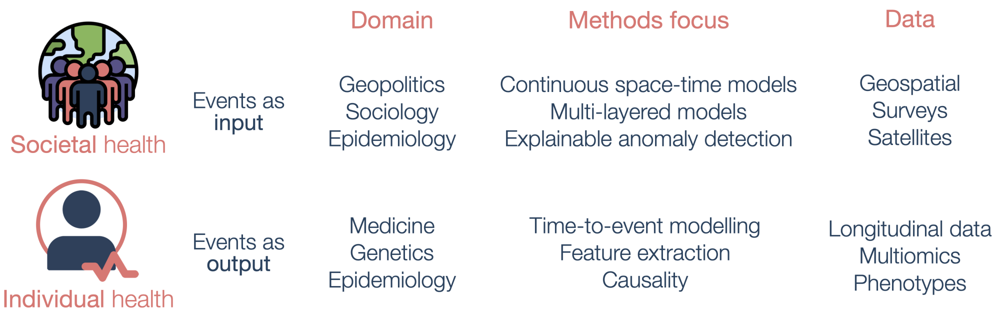

Figure 1: Application domains for Eventful

Cluver and Seth Flaxman, see Appendix A). They are in the process of consultations with a broad group of policy makers and relevant NGOs, including Lesotho Government, Unicef, World Health Organization, and UN Development Programme.

Individual health: Prognosis for insights from temporal patterns Time-to-event models have many uses in medicine and well-being research, including: describing the trajectory of health conditions; investigating associations and/or causal relationships among variables and outcomes of interest; estimating an individual's probability of experiencing different outcomes. To develop useful models that have a real domain impact requires high-quality data. To this end, we will avail ourselves of the data provided in the UK Biobank, an extensive clinical database containing more than 500,000 samples, including electronic health records (EHR) data, biomedical assay data, and high-throughput assay data (e.g. imaging data, whole genome sequencing data, and whole exome sequencing data). Deep learning approaches have already been implemented to analyze EHR data to predict the risk of the heart [2] and cardiovascular [3] diseases. However, these developments have mainly focussed on SNP analysis [4,5], looking at specific locations in DNA. Full genome sequencing, meanwhile, derives data for the three billion base pairs of DNA found in humans, potentially a richer source of predictive information. Integrating EHR and multi-omics data remains a challenge, which, if surmounted, promises more accurate time-to-event models. Further, taking into account the timings of patient measurements may lead to better insights into disease mechanisms. We will focus on cardiovascular diseases with our collaborators (Appendix A) Spiros Denaxas [6] and Ana Torralbo, and on diabetes with Andrew Hattersley and John Dennis.

Common challenges across both application domains include missing data, integrating multiple data sources and modalities, and building models that are simultaneously performant, explainable and trustworthy.

1.3 Scientific goals, challenges, contributions

In this section we outline the scientific goals of our proposal. For each goal, we state corresponding challenges, and our expected contributions towards meeting the challenges and thus fulfilling the goals.

Goal I: Rationalise event knowledge and data

Review the state of the art in modelling 'events as input' and 'events as outcomes', how and where these methods are employed to challenges in health of individuals and society. Create a knowledge base that will catalogue event-related open data sources that can be utilised for answering technical, scientific and policy-making questions in our application domains.

Contributions: We will perform a systematic review of available and currently-used methods, tools and open-source data sets for event-type data problems, collecting this information into a centralized resource. We will identify gaps in existing provision, such as automated machine learning pipelines' ability (or lack thereof) to handle temporal data, and the accessibility of tools for extracting useful features from sequence-type or irregular time series data. This 'knowledge base' will inform later developments in the project, such as the development and comparison of feature extraction methods and evaluation of automated data processing techniques.

Goal II: Advance data preprocessing for event modelling

Measure and understand the effects of different choices that can be made in the process of preprocessing event-type data. Improve the input side of events modelling by utilising, combining, improving or developing novel methods for data cleaning, preprocessing and variable selection or extraction.

Challenges: How do we extract more knowledge from aggregated data or data measured at different scales? How can we maximize use of available information, especially in the case of complex, interdependent structures such as genomic data, sequences of events in a patient trajectory or co-occurrence of factors in space and time? How can we improve the preprocessing of data that has sequential or hierarchical dependence structures, especially for aggregated and binned data?

Contributions: We study the effect of different choices in the event data analysis pipeline. Specifically, the choice of feature engineering, dimension reduction, data cleaning and binning and aggregation method or approach—and of course the type of modelling framework employed—may have profound influences on output results and performance according to machine learning benchmarks. This is an oft-studied topic in machine learning generally (e.g. Learn2Clean, AutoML) but spatio-temporal data and time-to-event modelling are two areas frequently overlooked, and thus we address them by adapting and extending methods accordingly.

Goal III: Advance event modelling applicability, explainability and trustworthiness

Extend, combine or improve state-of-the-art methods used to model events to increase their performance, explainability, trustworthiness and applicability to richer data sets.

Challenges: Achieving this goal requires overcoming **multiple challenges**:

- Can we design Emergent Hybrid Machine Learning and generative models competent in delivering state-of-the-art performance in the application areas?
- Which feature engineering techniques and deep learning models can extract most knowledge of multi-layered and multi-view (-omics and EHR) event data?
- How do we evaluate and improve explanation methods for event modelling?
- How do we improve survey-based collection of data about events and their impact on people by using active learning?
- How do we develop and evaluate deep learning based models for multi-layered events data?

Contributions: We propose applying novel techniques from the multidisciplinary area of 'scientific machine learning' to combine the benefits of expert-derived mechanistic models with more data-driven approaches. This will enable fitting continuous models to socio-political conflict data, where previously only discretisations were feasible.

We plan to investigate the extent to which time-to-event models can be improved using modern feature extraction techniques, in particular benchmarking survival models using whole-genome-derived features (from a biobank data set) against those using SNPs and standard feature selection techniques.

We assess the ability of state-of-the-art deep learning models to capture patterns in multi-layer event data in simulated and real-world use cases. We will develop and evaluate explainable methods in both areas and test different attention layers for that purpose. Borrowing methods from 'ecological inference' (also known as aggregate inference or distributional regression) we also endeavour to gain insights about individual-level phenomena from marginal, group-level or multi-layered data, including modelling individual trajectories for people or groups described by aggregated survey data. The capabilities of machine learning models to retrieve parameters from multi-layer systems will be tested on both simulated and real-world socio-economic survey data.

Goal IV: Increase utility of event modelling for scientific and technical community

Deliver useful and usable tools that implement methods and techniques that make event modelling more applicable, explainable and trustworthy.

Challenges: When delivering technical tools, can we find an effective trade-off between usability, usefulness, computational efficiency and extendability and maintainability?

Contributions: Modular packages that are useful beyond the applications at hand and that foster scientific and technological progress. We make these models available in model zoos (e.g. OpenML, Huggingface) where they are not present.

Goal V: Demonstrate utility of event modelling for society

Apply methods, models, techniques and tools to challenges important to individuals (in health context) and society at large (in the context of crisis modelling).

Challenges: Can we scope, design, model, and deliver a solution for the following applications:

- Forecasting cardiovascular events with input from clinicians using whole genome data
- Modelling socioeconomic events (e.g. periods of unemployment) from household survey data
- Predicting time to develop diabetes complication or modelling glycaemic progression using UK Biobank data
- Demonstration of the usefulness of active learning surveys through subsampling an existing survey and potentially in the real-world using a public polling firm
- Data-driven interventions to protect from (or reduce the effects of) compound crises on youth in Sub-Saharan Africa

Contributions: We scope, develop and deliver models or data products addressing above needs in collaboration with our network. We will disseminate our findings, open-source our solutions where possible, and submit research to reputable venues.

2 State of the art

Events data and where to find them The Armed Conflict Location & Event Data project ACLED;[7] provides an event-based data set describing 'actors' (people and groups), locations, dates and other characteristics of political violence, protests and civil unrest around the world. Similar data sets include the Uppsala Conflict Data Program Georeferenced Event Dataset, which focusses on violent events[8], the Integrated Crisis Early Warning System 'Dataverse' and the TERRIER project. Much of data in these projects is manually curated and extracted by domain experts; the process is not fully automated.

The Global Health Data Exchange disseminates information from national public health information management systems along with survey data from UNICEF and Demographic and Health Services that tracks events such as public health centre stockouts, immunization rates and cases of rare cancers and other rare diseases. Unfortunately, data quality is often inconsistent and offers only low levels of spatio-temporal resolution.

Previous attempts for central repositories for events data, such as `SurvSet`[9], collate only time-to-event data, mostly those bundled with R packages for survival modelling, which means the data are necessarily small in size. There remains no comprehensive 'meta-analysis' comparing or collating these different data contributions, nor a centralised repository. The Web site Papers With Code boasts just one data set associated with the task of event detection.

Event data preprocessing and cleaning Relatively little attention has been paid to longitudinal or time-to-event models in this context. For relational databases, HOLOCLEAN[10] provides a probabilistic framework for cleaning time series data while CURRENTCLEAN[11] allows identifying and cleaning stale data in a non-automated setting.

In other contexts, events data cleaning through iterative anomaly repairing[12] shows improvement in classification and parameter estimation in anomalous GPS trajectories and sensor readings. GeoSClean[13] demonstrates cleaning of anomalous GPS data securely. Reinforecement learning based methods like Learn2Clean[14] have also been used for cleaning data for classification, regression, and clustering tasks. The CleanML benchmark[15] considers classification tasks only. Automation of data analysis pipelines more generally may be framed as a human-in-the-loop task[16].

Time-to-event models Recently, benchmark experiments[17] have reviewed the state of the art of time-to-event models such as DeepHit[18], DeepSurv[19], DNNSurv[20], and CoxTime[21] in the context of non-whole genome data (SNPs) as well as feature selection in that application[22]. For high-dimensional data, random survival forests have been demonstrated to consistently perform well. These models are available in open source software including mlr3proba[23], survivalmodels[24], pysurvival[25], and scikit-survival[26]. Random survival forests are very sensitive to the choice of features and how they are preprocessed. In order to apply it full sequence genome we will build on existing work of SDS[27,28]. The same applies for the (temporal) clinical features where the state of the art is reviewed in[29].

Spatio-temporal modelling Recent advances in scientific machine learning[30] have demonstrated how systems of dynamic equations inspired by physical processes—represented via ordinary or partial differential equations (ODEs or PDEs)—may be combined with the latest deep learning models designed to model complex latent spatio-temporal dynamics. These 'neural ODEs/PDEs' permit flexible modelling of the latent structure via deep neural networks while learning complex dependencies from the observed event history[31], without necessarily having vast amounts of (complete) data. Example applications include predicting earthquakes and spread of infectious diseases.

In the societal context, statistical models of criminal behaviour[32] are able to capture relatively well the broad dynamics of hotspot sociological observations, but are not as flexible as deep learning. Gun violence, in particular, has been shown to 'diffuse' or spread between geographical locations[33]; understanding the spatio-temporal patterns may help inform efforts to prevent and reduce it.

Deng et. al.[34] present a systematic review of deep learning based methods for offline civil unrest (protests, strikes, etc.) and crime. DynamicGCN[35] forecasts short-term protests using unstructured text that demonstrated effectiveness in the prediction of protest events in multiple countries. Glean[36] uses knowledge graphs to predict concurrent political events of multiple types using heterogeneous data fusion. For crime sequence prediction on a weekly basis, Wang. et. al[37] proposed a sequence generative neural network. Jin et al.[38] developed a context-based sequence generative model in a generative adversarial network, Crime-GAN. Deep Temporal Multi-Graph Convolutional Network[39] and CrimeSTC[40] learn disentangled representation of spatio-temporal features and heterogeneous data for predicting crime.

Explainability Limited progress has been made towards posthoc explanation methods for time-to-event models. These include counterfactual explanations for survival models[41], using neural additive models[42], as well as derivatives of Lime-based[43] algorithm SurvLime[44,45]. None of these methods include temporal dimension in the final explanation, which is in turn crucial for predicting the survival conditional probability distribution. In explainability for spatio-temporal data, Deng et. al. introduced an interpretable deep learning framework for forecasting civil unrest events and providing multilevel explanation for predictions using knowledge graphs[46].

Additionally, there have been research directions exploring the selection of anomalous/outlying features or contexts for anomaly detection and explanation. Outlying aspect mining[47,48,49,50] and contextual anomaly detection[51,52,53] search for anomalous feature subspaces and contexts that explain the anomalous aspect of data detected by off-the-shelf detectors. However, the explainability of deep learning methods for spatio-temporal contextual anomaly detection has not been studied, and therefore discovering complex contextual relations in data in this setting remains black-box.

Survey methods Survey-based studies of human populations are commonly used by governments to track vital socioeconomic metrics (e.g. unemployment rate, or childhood stunting). Survey methodologies have scarcely been touched by statistical learning, except in geology, which has adapted to Bayesian techniques to optimize sampling[54] and recently ML models to fuse surveys with additional data sources[55]. Surveys involving human subjects are more complex, for example, needing significant weight adjustments to capture demographic proportions and response patterns. Recently, ML-based approaches have demonstrated significant weight-computation improvements using a model-averaging approach across multiple measures[56].

ML is also used frequently by researchers to perform spatio-temporal small-area estimation with additional data sets. WorldPop[57] pioneered such methods, producing gridded population data at the highest resolution to date across dozens of countries. World Food Programme's HungerMap combines survey data with satellite imagery and price-data in real-time to produce food insecurity updates[58].

There are limited examples of ML being used in survey design, with notable exceptions including reducing bias in data collection through ML monitoring of enumerators[59], and aiding in questionnaire design by identifying questions that align with qualitative interview answers about the concept[60].

3 DFKI Long-Term Research Framework

Goals of Eventful include improving the explainability and causal inferability of event-based models, thus contributing to **Credible and Trustworthy AI** stream of the DFKI research framework. We hypothesise that comprehensive approaches incorporating human knowledge and providing interpretable results can lead to more effective and practical event analysis and more successful application in the real-world. Additionally, the integration of *a priori* knowledge into neural networks is a promising approach to improve the models' predictive capabilities without increasing the parameter count and therefore memory and computational footprint.

The DFKI has made it its goal to tackle challenges regarding the combination of neural models with e.g. explicit or symbolic models, methods from information theory or Bayesian approaches (**Emergent Hybrid Machine Learning**). We plan to apply recent developments in the field of neural ordinary and partial differential equations—combinations of differential equations with neural networks model. The consequent reduction of parameters ties this work into **Resource-Conservative Systems**.

4 Industrial relevance

The scope of Eventful is time-to-event and spatio-temporal models, with extensions to multi-layer event modelling and a focus on explainability, usability and impact. Throughout the project, we will test our methodological advances and tools on the application domains of individual and societal health, to demonstrate relevance to **medicine**, **public health**, **geopolitics** and **geography**, **economic policy-making** and **security**.

However, time-to-event and spatio-temporal modelling inhabit a broader range of domains, albeit with differing terminology. For example, time-to-event analysis is also known as survival analysis (medicine), reliability analysis (engineering) and duration analysis (economics). We list a few specific examples of potential usage of this project in domains beyond those mentioned above.

Clinical prognosis Handling and performing predicting analyses on large, high-dimensional data sets—such as those increasingly seen in multi-omics and m-health research—may enable clinical researchers to better predict patient outcomes and measure treatment effects.

Early warning systems A proliferation of atmospheric sensors and other technology yields large quantities of climatological data, which is naturally timestamped and often geo-located, though sometimes difficult to wrangle owing to wide or irregularly-spaced sampling intervals.

Energy, agriculture and biodiversity Spatio-temporal modelling is a natural fit for analysis of land use, monitoring species distributions and measuring, for example, solar irradiation over different areas.

Finance Several recent advances combining deep learning with physics such as deep trajectory learning[61] and recursive multidimensional symbolic regression[62] have the potential to vastly transform the field and elucidate some of the most elusive questions for policy-makers—dynamics of asset-price bubbles, rapid onset monetary inflation, the whip-tail effect in supply-chains, the repercussions of economic shocks, and more.

Insurance The ability to predict the risk of events, especially catastrophic ones, is of especial interest to businesses and their insurers.

Manufacturing The recent pandemic has shown that individual events can have outsized effects on entire supply chains and even on multiple interconnected sectors. But not all relevant events must be of the scope of a global pandemic. Event analysis can help in identifying the impact of various events on manufacturing, pricing, sales and markets.

Marketing Data protection and other constraints often mean that customer data may only be collected or stored in aggregated form, however ecological inference may allow limited insights about individual consumer behaviour to be deduced.

5 Connections to existing DFKI projects

Racket One aspect of fault analysis in factory environments that is underdeveloped in Racket is the time-to-failure. Eventful aims to tackle models taking in high-dimensional data (like the many variables involved in factory environments) and returning possible events and their timeframe in WP 2.1.2 and WP 2.2.2. Additionally, we consider the explainability of these models in WP 2.2.1. Further, the underlying methodology could be useful for root cause analysis of the faults. Thus, there exists a natural potential for synergy.

FedWell FedWell's goal is to study and address the challenges of creating, maintaining and adapting user models to increase the benefit of adaptive systems for personal well-being and skill building. Monitoring and predicting users' well-being through models is an obvious use-case for improvement trough the type of models to be researched in Eventful: (i) Using a combination of deep learning and continuous statistical models allows for incorporating disruptive but irregular data like events (WP 2.1.1) into the models input, (ii) the prediction of future health events through the multitude of measurements going into the model (WP 2.1.2) as well as (iii) being a highly multi-layered problem, since health is a multi-layered problem with multi-layered events (WP 2.1.3).

DefuseNN DefuseNN aims to address three main challenges in deep learning: building a knowledge repository to improve understanding of the deep learning landscape, investigating and developing novel multi-modal fusion approaches for deep learning, and making use of external context within deep neural networks. Eventful has thematic overlaps with DefuseNN with a shifted focus to events: We want to add onto the existing deep learning knowledge base by incorporating databases and methods related to event data in WP 1.1. The use of multi-layered event data with deep learning in WP 2.1.3 ties into the ability to incorporate multi-modal signals in DNN architectures.

AScore The goal of this project is to create a simulation-based management system that supports municipalities in dealing with complex crisis using the example of the coronavirus pandemic. Predictions by the system can be improved by incorporating data that is irregular but disruptive. In WP 2.1.2 we intend to research and develop time-to-event models that use high-dimensional data as predictors to forecast events and their timing, tying well into and augmenting AScores' use of multi-sourced data to model crisis situations. AScore is based on agent based models that can be used as a ground truth for WP 2.1.3.

6 Work plan and packages

In achieving the goals of this project, we divide it into the following three work packages: (1) Building a general framework that enables better event analysis, (2) Integrating, developing and extending methods used for event modelling (3) Demonstrating and enabling impact by developing tools and executing a societally impactful project using developed framework, models and tools.

Figure 2: Overview of work packages (left) and sub-workpackages (right) to be delivered.

The overview of work packages and their deliverables is shown in Figure 2. Our work plan includes 147 senior researcher person-months (PM). The researchers will be supported by research assistants, and this has been included in the resource request. The work sizing in Table 1 shows senior researcher person-months to the sub-WP level and lower, where applicable. A more granular list of tasks and their dependencies is available in Appendix D. Potential risks for this project including their description, estimates for likelihood and impact, and our mitigation strategies are available in the Appendix B.

We outline the project schedule in Figure 3. Milestones are scheduled not to coincide with the end of each work package, but rather the earliest feasible delivery time for a minimal viable product (MVP). Hence, each milestone corresponds to a deliverable that may be a dependency of another, later (sub)-work package. Naturally, by the actual end of each work package the MVP will, through an iterative process, develop into more substantial outputs. Project milestones are as follows:

M1 Technical report of knowledge base accessible
M2 Time-to-event results for full genome using UK Biobank data
M3 Multilayered event modelling with at least 3 layers working in real-world data
M4 Explanations evaluated applications in both application domains
M5 Registering first tool in package manager (i.e. Cran, Pip)
M6 Proof of concept for the Capstone project delivered to end-users

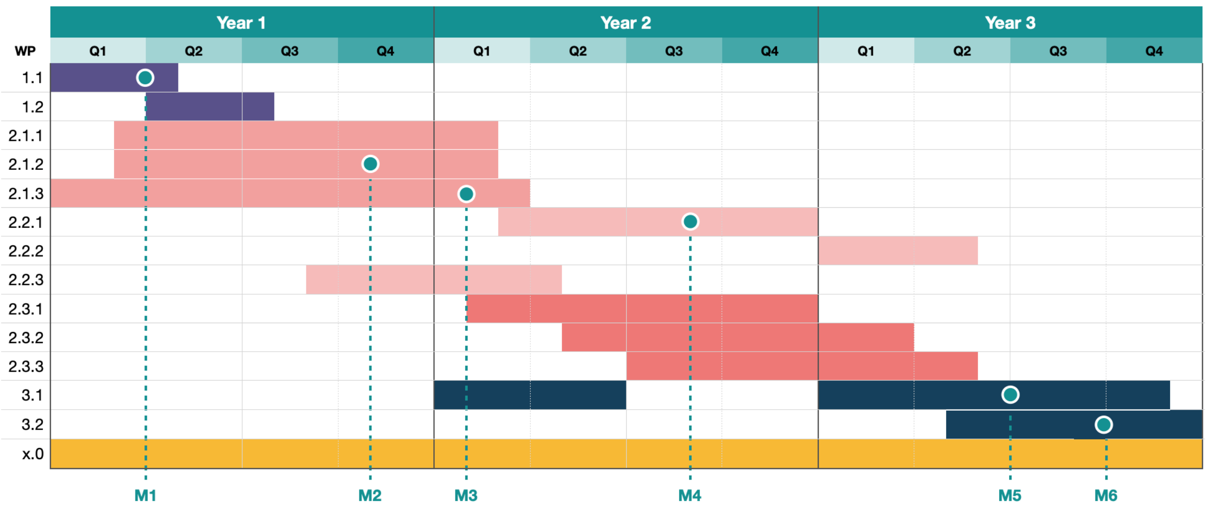

Figure 3: Work plan at sub-WP level, including milestones (M1-6).

6.1 Framework (WP 1.x)

The first work package will comprise a thorough review and compilation of the latest methods and resources for building high-dimensional temporal models and applying them in our two primary topic areas. The delivery will be supported by 3 person-months of management resource.

6.1.1 Events knowledge base (WP 1.1)

To create the knowledge base, we will do a comprehensive literature review into event-related tasks and data, focusing on (1) Collecting and categorising open data sets for events in our application domains; (2) Identifying state of the art methods and off the shelf models for methodological tasks in WP2; (3) Identifying opportunities for porting knowledge from other domains that might be tackling similar problems but using different terminology.

Our output will include a comprehensive guide to the various data repositories already available (e.g. conflict data sets such as ACLED and the Uppsala Conflict Data Program, and demographic health surveys) detailing their contents, methods by which they are generated, use history (e.g. which publications already mention them) as well as the level of accessibility (if not open access) and other relative merits. This guide will be accompanied by sample scripts and step-by-step tutorials demonstrating how to access

Paclage	Sub-package	Size
WP 1 Framework (14 PM)	1.0 Project management	3 PM
	1.1 Events knowledgebase	7 PM
	1.2 Automated preprocessing and cleaning	4 PM
WP 2 Methods (95 PM)	2.0 Project management	3 PM
2.1 Modelling	2.1.1 Continuous spatio-temporal modelling	11 PM
	2.1.2 Time-to-event models for high dimensional data	10 PM
	2.1.3 Multi-layered event data	18 PM
2.2 Explainability and causality	2.2.1 Explainability for time-to-event	8 PM
	2.2.2 Explainable anomaly detection	8 PM
	2.2.3 Explainable multi-layered models	8 PM
2.3 Variable resolution	2.3.1 Better binning and beyond	9 PM
	2.3.2 Better surveys	10 PM
	2.3.3 Beyond aggregation	10 PM
WP 3 Impact (38 PM)	3.0 Project management	10 PM
	3.1 Eventful tools	20 PM
	3.2 Capstone project	8 PM

Table 1: Sizing (in senior researcher person-months, PM) for work packages and subpackages

the data. We aim to publish this as a technical report and, as a stretch goal, as a web portal where this information will be accessible to researchers and practitioners.

6.1.2 Automated preprocessing and cleaning (WP 1.2)

Data preprocessing is the first stage of any machine learning pipeline and decisions made here have the potential for profound effects downstream. This work package will explore existing practices for cleaning, selecting and extracting features from data when either the features or the outcomes are temporal in nature. A systematic review will compare the various commonly-used data cleaning methods and tools available, cataloguing their support (if any) for irregularly-spaced time series and time-to-event problems. A limited benchmark study will evaluate the sensitivity of temporal models to choices of data preprocessing strategies.

Additionally, this work package will document recent developments in automated data science pipelines, and highlight any gaps in current provision for tools that can be applied to (spatio-)temporal data. A key output of this work package will be to fill any such lacuna by developing initial scripts that will extend existing automated data cleaning tools to add support for time-to-event analysis or temporally dependent data.

6.2 Methods for event modelling (WP 2.x)

The second work package concerns development of methods for the analysis of complex event-type data: in particular, spatio-temporal data of varying levels of resolution, and time-to-event models with high-dimensional predictors. To achieve this, we will build on our previous expertise in time-to-event [63,64,65]; explainable predictors [28,66]; survey methodology, results and ecological inference [67,68,69]; and continuous time models [70,71,72]. The delivery will be supported by 3 person-months of management resource.

6.2.1 Models (WP 2.1)

Continuous spatio-temporal modelling (2.1.1) To improve continuous spatio-temporal modelling we exploit the best characteristics of deep neural networks (flexible, but with limited support for sparse or incomplete data) with physics-inspired models that employ differential equations (offering the ability to impute missing data, at the expense of flexibility). This combined 'SciML' approach [30] using neural differential equations models—typically used for engineering applications—will be adapted for modelling the 'diffusion' of gun violence [32,33] and forecasting tumour growth [73] in continuous time, with a view to making such models more interpretable and flexible. We compare such neural differential equation models with more conventional methods, such as infectious disease models and discrete-time neural networks.

Time-to-event models for high-dimensional data (2.1.2) We explore the potential to improve the performance of these models, for instance the use of genome-wide sequence data—rather than more limited information offered by genotyped or imputed SNPs—to help predict the risk of adverse health events in patients. This would be achieved using a 'meta-modelling' approach to compare filter, wrapper, representation learning and expert knowledge-driven approaches—and combinations thereof—to extract and select features, borrowing techniques used in text mining [74,27] and meteorology [75]. Building on WP 1, we would compare our developed methods to public benchmarks [17] and then apply them to a real-world application using UK Biobank (which includes whole-genome data) to forecast (e.g.) cardiovascular events with input from clinicians. Ultimately, this work would assess the value of information contained within whole genome sequence data, relative to coarser representations such as genotypes and automatic or manually derived phenotypes.

Multi-layered event data (2.1.3) can be exploited to model complex spatio-temporal patterns, for instance predicting future rates or counts of crimes reported in different locations (or, analogously, incidence of natural disasters). By simulating a multi-layered system of discrete events using agent-based models or otherwise, we assess the ability of state-of-the-art deep learning models to retrieve known patterns of different levels of complexity. The developed models will then be applied to real-world data, including a case study of risk modelling for youth in Lesotho.

Typically, joint longitudinal data from the social, medical and financial health of an individual would not be available, however the Consumer Pyramids Household Survey, conducted by the Centre for the Monitoring of the Indian Economy (CMIE), offers a unique opportunity to study several event-type variables at once, including employment and unemployment spells, household illness, hospitalization and death, which can be modelled against societal background processes such as consumer purchase patterns and other financial indices.

6.2.2 Explainability and causality (WP 2.2)

Explainability for time-to-event (2.2.1) We will study the interplay between feature selection techniques (WP 2.1.2), imputation techniques for dealing with missing variables (WP 1.2) and the explanation methods including SurvSHAP [76] and others mentioned in Section 2. We will also build on our previous work on preservation of feature importance [77], which is a fundamental challenge in explainability for problems with many features. Next, we will consider causal inference and explanation, focusing on an important assumption of lack of unmeasured confounders. Only recent works consider missingness in variables that influence treatment selection [78]. We will extend these methodological advances in causality to time-to-event, conducting initial studies based on simulator and models from literature [79]. Our stretch goal is to apply this in non-simulated environment, building on our previous work on treatment selection and prognosis for diabetes [80,65,81]. We have agreed in principle on a project on studying glycaemic progression or time to developing diabetes complications using UK Biobank data with Prof. Hattersley and Dr. Dennis (see Appendix A).

Explainable anomaly detection (2.2.2) We evaluate and extend methods for explainable anomaly detection for spatio-temportal models. Our goal is to develop algorithms that effectively capture compherhensive spatio-temporal dependencies in the latent space and explain how anomalies occur under different contexts. This would result in efficient discovery of contextual anomalies, and we intend to build on recent work on anomaly detection on complex data [82,83] and extend it for multivariate spatio-temporal data sets, initially based on features generated by SPATE-GAN [84]. Our stretch goal is to build on our previous work on time series attribution techniques [85] to explain temporal precursors to anomalies.

Explainable multi-layered models (2.2.3) We will study statistical and deep learning models with and without interpretability to effectively evaluate the trade-off between feature preparation efforts and performance and their ability to generalize on future data under temporal dependencies. Then, to increase performance and explainability, we will study different types of attention layers [86] and perform neural network architecture search. Finally, we will build extend on existing models (e.g. Hossain [87]) and explore their neighbourhood in model space.

6.2.3 Variable resolution (WP 2.3)

Better binning and beyond (2.3.1) Only last year in-depth/detailed/quantitative investigations of the effect of binning have been studied [88]—showing glaring gaps in existing literature. This is particularly important for multi-layered event modelling since resolution of raw data is often on different scales [89]. We will study choice of binning and ways of imputing finer resolutions [90] in the context of pipelines, models, simulators and data sets investigated in complex spatio-temporal modelling (WP 2.1.3). Additionally, we will contrast this with methods that can directly model the data without binning such as a potential extension (WP 2.1.1) to multi-layered event data.

Better surveys (2.3.2) We will develop and test active learning for surveys. This can be seen as multi-output active learning for classification, and is a natural place to start (insert reference) as most surveys primarily measure discrete variables. We will demonstrate our methods by applying them to a problem where same results can be longitudinally achieved through an intelligent subsample from a relevant existing survey. We have identified Consumer Pyramids Household Survey (joint longitudinal data from the social, medical and financial health of 112k households—3 times a year) as a potential use case for this purpose, and have agreed in principle a collaboration for this with Prof. Shah (Appendix A).

Upon obtaining a baseline by the multi-output active learning, we will study whether a competitive method using Bayesian quadrature can be built. As surveys are used to allocate resources and intervene on the livelihoods of people, biases and non-representiveness can have dire effects [91]. The next step is to evaluate our Bayesian quadrature-based method and others across multiple performance metrics, fundamentally optimizing total survey cost, ensuring demographic representation, while maximizing the accuracy of the estimates.

Finally, we plan a live active learning survey to test, demonstrate and evaluate the methodology. We intend to run a field experiment using a public polling firm in the context of the World Food Programme's mobile Vulnerability Analysis and Mapping survey; the Consumer Pyramids Household Survey economic monitoring initiative in India; or in the context of the youth affected by the compound crises in Sub-Saharan Africa. The choice of use case will depend on the usefulness for the application, the ability to develop methodology and the availability of 'ground truth' for effective comparison.

Beyond aggregation (2.3.3) We will investigate methods that can recover joint distributions or even individual trajectories (down to the groups or sub-group level) where only aggregated counts/rates or sparse (marginal) data are available. This offers the opportunity to tie together both of our cross-cutting themes of individual health and societal well-being. Traditional examples of this problem formulation include voter transition models [92], which were extended with the application of Gaussian processes [93]. We will consider reconstructing a potential event sequence experienced by clusters of individuals based on count observation by extending the work of Singh et al. [94] to continuous-time Markov chains.

As a stretch goal, we will consider the following two applications: for the Lesotho project, we would like to breakdown the knowledge about people affected by multiple crises (e.g. health crisis and danger of child marriage) by multiple demographic attributes from aggregate separately collected data. Alternatively, we might consider breaking down internal migration by different groups building on [95] working on data of global internal displacement data base. Throughout, we endeavour to fuse the existing data with weakly informative data such as micro-census or satellite data [96].

6.3 Ensuring and demonstrating impact (WP 3.x)

The aim of this work package is to enable the work undertaken in WP 1 and WP 2 to be impactful by making it available as software package, where possible (Eventful in research). Then, we aim to demonstrate the impact by delivering a societally beneficial capstone project that will include multiple aspects of our work (Eventful in practice). The delivery will be supported by 12 person-months of management resource.

We are well-placed to do this due to our track record in collaborative development and contributions to statistical modelling software in R, Python and Julia (MLJ, rOpenSci, `drugprepr`, `doseminer`, `rdflib`, hephaistos), and broad experience delivering socially beneficial data solutions to government and charity sectors through our Data Science for Social Good fellowship since 2019 [97].

6.3.1 Development of Eventful tools (WP 3.1)

As methods and techniques are developed in WP 1 and 2, they will be aggregated into our 'work-in-progress' code to a central GitHub organisation. In this WP, we will package and extend this work in a way that can be used in practice by as broad community as possible. Code will be developed collaboratively on GitHub with clear documentation for users and contributors. The core infrastructure will be organized into one or more small software packages designed to follow 'best practice' coding principles as closely as possible, paying particular attention to readability, usability and extensibility.

For preprocessing, we aim to deliver a package for automated data cleaning, implementing similar methodology as Learn2Clean in time-to-event context. For modelling, we envision modelling packages for full sequence genomics and patient dynamics for time-to-event and a pipelining tool implementing methodology developed for multi-layered event data (stretch goal). For explainability, we will provide a web application with a package of explainable methods that will allow researchers to upload their trained deep learning models and receive insights about the role of neurons and different layers in the model. At feature level, the tool will highlight most informative features on the basis of which model is making a decision. As a stretch goal, we will explore delivering tools corresponding to variable resolution.

6.3.2 Capstone project (WP 3.2)

Our goal is to create a functioning solution that addresses a significant and complex societal issue, while showcasing our methodological advancements and tools. To achieve this, we will collaborate with stakeholders across the policy-making spectrum and ensure that the solution provides actionable, easily understandable insights. We will leverage existing partnerships with the University of Oxford who are already engaging with organizations such as the UN and the World Health Organization to address the impact of compound crises on youth in Africa.

The exact scope of this project is still being established. While our preliminary scoping and analysis indicates that a societal health project would be the most impactful and appropriate demonstration of our methods, we will also consider opportunities in individual health. As this aspect of Eventful is planned to be delivered in 2 years, we will commit to the exact scope by Q4 of the second year of the project, taking into account world events, population needs, and availability of data.

We see an opportunity to demonstrate multiple aspects of Eventful in a project such as protection of youth from compound crises in Sub-Saharan Africa. We could use the knowledge base (WP 1.1) to find additional data to augment government and third sector data, leading to better predictors. Preprocessing and data cleaning (WP 1.2) is an omnipresent step in every data-related project. Multi-layer models (WP 2.1.3) could then be used as the nature of the problem is exactly appropriate for this method—compound crises are multi-layered by nature and thus have multi-layered events.

The crucial part of this will be the integration of explainability for multi-layered models (WP 2.2.3), and we will pay particular attention to feedback from end users on the explanations that model will give. Finally, integration of the demographic surveys is often an important feature subset in similar problems, and deliverables of WP 2.3.3 could be used if the availability of survey data is affected (e.g. for remote areas) or if variables are sparsely measured. Potentially, active learning for multi-output classification (WP 2.3.2) might need to be used to devise a better surveying strategy, improving data quality. The deliverable of this work package is going to be an end-to-end data science pipeline integrated with an interface that end users such as socioeconomic policymakers can use directly for data driven evidence based policymaking.

7 Project evaluation

To evaluate the methods developed in the Eventful project we will implement and apply them to a series of exemplary real-world data sets and compare them against existing benchmarks. Specific applications include the diffusion of gun violence, cardiovascular risk prediction in the Biobank, glycaemic progression for diabetes patients and the Consumer Pyramids Household Survey.

As each of our work packages are quite different, we require different methods of evaluation for each, but to ensure comparison and evaluation is fair we largely group our evaluation into three methods: *Task driven*—can an independent auditor reproduce an example from our project using our data and tools? *Data driven*—can we solve a particular problem using a given data set? *Feedback, engagement and publication*—have we received positive feedback from a community interested in our work?

Independent reproduction of results of experiments and analysis is the gold-standard method to evaluate the success of tools and methodology. Our software outputs will ideally adhere to best practice, for example including continuous integration and a high level of unit test coverage.

Throughout the project we will submit papers summarising our work to high impact journals, conferences, and workshops. In particular for WP 1 we will aim for publication in broad audience journals to ensure our work is found by researchers across different events domains. For WP 2 we will aim for high-impact journals and conferences to ensure our methodological developments are received by expert audiences. For WP 3 we will submit to software journals such as JOSS. As well as evaluation by formal peer review we will also self-evaluate our software and outputs by utilising community metrics such as

Chaoss. We will also pay attention to feedback from users and collaborators. We will use these to influence code design and quality, documentation and coding direction.

8 Follow-up work

As our proposal includes a review of the state of the art, this is likely to point towards additional research gaps, which could be tackled by our department or the broader scientific or industrial community. Of those presented in Section 4, from today's viewpoint, we identify the following three application areas as ones that would potentially benefit the most from the methods and tools developed: energy and biodiversity; insurance risk estimation and manufacturing.

Whereas much of this proposal focusses on sourcing and selecting features from existing events data sets, an follow-on project might involve the use of natural language processing to generate new 'event'-type data sets from text, for example generating a stream of timestamped socio-political events from news reports. Additionally, performance and interpretability may be improved through exploitation of knowledge graph structures, which are not considered in this project.

Further opportunities for follow-up work will be identified during the project.

9 Resources

We plan for this project to be delivered by 4 senior researchers over the period of 36 months, plus 3 person months of senior resource for applied project scoping. Each researcher will be supported by 18 hours of student assistant work per week. In addition, the project will require the usual administrative support (e.g. HR, legal, purchasing). In addition to the standard work equipment, aspects of this projects might require computing power, data storage, and virtualisation hardware and mobile workstations. We foresee costs related to acquisition of existing data (UK Biobank) and data collection for surveys. Finally, reasonable travel will need to be covered for the duration of the project, including to present progress at SAB meetings, or to attend relevant conferences when presenting project work.

Appendix

A Collaborators

Prof. Dr. Lucie Cluver, University of Oxford - Professor, Centre for Evidence-Based Social Intervention, Department of Social Policy and Intervention

Prof. Dr. Spiros Denaxas, University College London - Professor of Biomedical Informatics, Deputy Institute Director for Research at the Institute of Health Informatics

Dr. John Dennis, University of Exeter - Senior Independent Research Fellow, University of Exeter Medical School

Dr. Seth Flaxman, University of Oxford - Associate Professor, Department of Computer Science, researching scalable methods and flexible models for spatio-temporal statistics and Bayesian machine learning, applied to public policy and social science

Prof. Andrew Hattersley FRCP FMedSci FRS, University of Exeter - Professor of Molecular Medicine and Gillings Chair in Precision Medicine

Dr. Ajay Shah, CMIE and Jindal Global Business School - Director at the Centre for Monitoring Indian Economy, Professor of Social Sciences & Humanities at Jindal Global Business School, Co-Founder of XKDR FORUM

Dr Ana Torralbo, University College London - Research Fellow in Health Data Science

B Risks

There are several risks or threats to successfully completing the project smoothly and efficiently. We identify a few of these potential risks and ways to mitigate them. For each risk, we also estimate likelihood (as either: very unlikely, not likely, possible, probable, or very likely) and impact (negligible, low, moderate, significant, or catastrophic).

Human resource Areas of the project require specific scientific knowledge and/or engineering skills. Failure to recruit or retain researchers with relevant specialized skills is possible and represents a moderate risk to efficient project completion (likelihood: possible; impact: moderate).

To mitigate this, the DSA research group is partnering with the SDS team within the DFKI and capitalize on existing partnerships with external collaborators (e.g. Oxford, UCL) and our broader group at Rhineland-Palatinate Technical University. Additionally, we can draw on a wider recruitment network through the Data Science for Social Good programme and by offering internships to local and international students. In any case, for at least the first four quarters of the project, specific people are already assigned to the planned tasks.

Data availability Some projects rely on access to specific data types. In the very unlikely event of failing to gain access to such data, there is a low risk of thwarting comprehensive evaluation of developed methods (likelihood: very unlikely; impact: low).

However, scoping is intentionally broad, allowing multiple different options of data sets for most applications: for example, if data from Lesotho is not available, we can seek data from another locale. No 'societal well-being' applications are intrinsically tied to a particular use case: children at risk and climatology are global problems with a diversity of places to find relevant data. The risk is higher for 'individual health', because personal health data is by its nature sensitive and less readily available, however we overcome this restriction by choosing to purchase access to the UK Biobank. As a final back-up option we may be able to avail ourselves of our collaboration with Mainz on the Clusters4Future Curatime project, which would also have relevant data.

Impact of tools and packages It is possible that on developing open-source tools and software outputs, they are not found to be useful or widely adopted by end-users. This would have a low but appreciable effect on the overall impact of our project (likelihood: possible; impact: low).

To maximize the usefulness and appeal of our tools we would prioritize work on those seen to be most 'useful', allocating time and resources accordingly. By engaging with the research community through discussions with relevant stakeholders (especially domain experts and existing analysts) we would hope to maintain engagement throughout the project and ensure tools are up-to-date and relevant to users' needs. Moreover, any tools developed can form part of a greater ecosystem already developed (namely MLJ and mlr3, the former having 1500 stars on GitHub), which increases its visibility and connects it with existing popular tools.

Scoping and planning of applied projects, survey deployment It is possible that (external) partners may disengage from the project, end-users become unreachable for feedback or projects are otherwise unable to be made a priority. The impact of this would be negligible as we can simply mitigate it by switching focus to other projects, especially as most sub-projects have multiple alternative sources of data and options for applications (likelihood: possible; impact: negligible).

At the start of each work package, a 1-month scoping review will identify such risks and adjust scope, where needed, to correct for any such missing external dependencies. In most cases, the application or case-study can simply be switched to one of several already-identified alternatives. Some tasks may additionally be assigned as Data Science for Social Good summer fellowship projects, due to strategic alignment between Eventful and fellowship. Finally, our external collaborators at other institutions may also be able to assist in certain scenarios.

C Abbrivations

Acled Armed Conflict Location & Event Data project

CMIE Centre for the Monitoring of the Indian Economy

DNN deep neural network

DSA Data Science and its Applications [DFKI research group]

EHR electronic health record

ML machine learning

NGO non-governmental organisation

ODE/PDE ordinary differential equation / partial differential equation

PM person-month(s)

SciML scientific machine learning

SDS Smart Data and Knowledge Services [DFKI research group]

SNP single-nucleotide polymorphism, a way of coding genetic data by reading at specific locations in DNA where relevant variation is present in the population

UCL University College London

WP work package

D Delivery task list

WP	#	Subtask	Req.	Size
1.0	1	Scoping applied projects, managing delivery		3
1.1	2	Research and create knowledgebase		6
	3	Web portal development (Stretch)	2	1
1.2	4	Review existing practices for cleaning, sensitivity benchmark study	2	3
	5	Design a tool, develop initial scripts	4	1
2.0	6	Managing delivery		3
2.1.1	7	Neural diff. eq. modelling, comparison with state of the art, package development		8
	8	Application to diffusion of gun violence and tumour growth forecasting	7	3
2.1.2	9	Model development, benchmarking, package development		7
	10	Application to cardiovascular events (Biobank)	9	3
2.1.3	11	Building a simulator		4
	12	Model development		8
	13	Applications to Lesotho and with CMIE		6
2.2.1	14	Feature selection vs missingness vs explanations vs importance	4, 9	2
	15	Methods for causality for time-to-event		3
	16	Application to diabetes (Stretch)	14, 15	3
2.2.2	17	Methods for explainable anomaly detection		5
	18	Explanations for temporal precursors to anomalies (Stretch)		3
2.2.3	19	Evaluate trade-off between feature performance and generalizability	12	3
	20	Study different types of attention layers; NN architecture search; extend models and explore model space	12	5
2.3.1	21	Binning strategies, Super Resolution strategies, pipeline integration	12	4
	22	Comparison with Continuous Methods , novel method (Stretch)		5
2.3.2	23	Active learning for survey prediction, application with CMIE		4
	24	Bayesian active learning for survey prediction		4
	25	Live active learning surveys (Stretch)		2
2.3.3	26	Modern ecological inference and AI derivates	22	3
	27	Multiple continuous time HMM Inference from aggregate data		4
	28	Application in migration (Stretch)	22	3
3.0	29	Project scoping, stakeholder, product and delivery management		10
3.1	30	Develop Learn2Clean for time-to-event	5, 9	5
	31	Package for full sequence genomics and patient dynamics for time-to-event	9	5
	32	Pipelining tool implementing methodology developed in 2.1.3 (Stretch)	12	3
	33	Web application with a package for explainable methods	14, 15	4
	34	Tools corresponding to 2.3 (Stretch)	21–28	3
3.2	36	Execute (model), deliver (product), evaluation	12, 19, 20	8
			Total	147

Table 2: List of subtasks, their sizings (in person-months) and dependencies

References

[1] Johannes Bircher and Eckhart G. Hahn. Understanding the nature of health: New perspectives for medicine and public health. Improved wellbeing at lower costs. *F1000Research*, 5:167, February 2016. ISSN 2046-1402. doi: 10.12688/f1000research.7849.1. URL `https://f1000research.com/articles/5-167/v1`.

[2] Shaan Khurshid, Samuel Friedman, Christopher Reeder, Paolo Di Achille, Nathaniel Diamant, Pulkit Singh, Lia X. Harrington, Xin Wang, Mostafa A. Al-Alusi, Gopal Sarma, Andrea S. Foulkes, Patrick T. Ellinor, Christopher D. Anderson, Jennifer E. Ho, Anthony A. Philippakis, Puneet Batra, and Steven A. Lubitz. ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation. *Circulation*, 145(2):122–133, January 2022. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIRCULATIONAHA.121.057480. URL `https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.121.057480`.

[3] Iain S Forrest, Ben O Petrazzini, Áine Duffy, Joshua K Park, Carla Marquez-Luna, Daniel M Jordan, Ghislain Rocheleau, Judy H Cho, Robert S Rosenson, Jagat Narula, Girish N Nadkarni, and Ron Do. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. *The Lancet*, 401(10372): 215–225, January 2023. ISSN 01406736. doi: 10.1016/S0140-6736(22)02079-7. URL `https://linkinghub.elsevier.com/retrieve/pii/S0140673622020797`.

[4] Wenjian Bi, Lars G. Fritsche, Bhramar Mukherjee, Sehee Kim, and Seunggeun Lee. A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank. *The American Journal of Human Genetics*, 107(2):222–233, August 2020. ISSN 00029297. doi: 10.1016/j.ajhg.2020.06.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929720301919`.

[5] Peilin Meng, Jing Ye, Xiaomeng Chu, Bolun Cheng, Shiqiang Cheng, Li Liu, Xuena Yang, Chujun Liang, and Feng Zhang. Associations between genetic loci, environment factors and mental disorders: a genome-wide survival analysis using the UK Biobank data. *Translational Psychiatry*, 12(1):17, January 2022. ISSN 2158-3188. doi: 10.1038/s41398-022-01782-8. URL `https://www.nature.com/articles/s41398-022-01782-8`.

[6] Eleni Rapsomaniki, Adam Timmis, Julie George, Mar Pujades-Rodriguez, Anoop D Shah, Spiros Denaxas, Ian R White, Mark J Caulfield, John E Deanfield, Liam Smeeth, Bryan Williams, Aroon Hingorani, and Harry Hemingway. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. *The Lancet*, 383(9932):1899–1911, May 2014. ISSN 01406736. doi: 10.1016/S0140-6736(14)60685-1. URL `https://linkinghub.elsevier.com/retrieve/pii/S0140673614606851`.

[7] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. *Journal of Peace Research*, 47(5):651–660, September 2010. ISSN 0022-3433, 1460-3578. doi: 10.1177/0022343310378914. URL `http://journals.sagepub.com/doi/10.1177/0022343310378914`.

[8] Kristine Eck. In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Cooperation and Conflict*, 47(1):124–141, March 2012. ISSN 0010-8367, 1460-3691. doi: 10.1177/0010836711434463. URL `http://journals.sagepub.com/doi/10.1177/0010836711434463`.

[9] Erik Drysdale. SurvSet: An open-source time-to-event dataset repository. Technical Report arXiv:2203.03094, arXiv, March 2022. URL `http://arxiv.org/abs/2203.03094`. arXiv:2203.03094 [cs, stat] type: article.

[10] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. HoloClean: Holistic Data Repairs with Probabilistic Inference. 2017. doi: 10.48550/ARXIV.1702.00820. URL `https://arxiv.org/abs/1702.00820`. Publisher: arXiv Version Number: 1.

[11] Zheng Zheng, Tri Minh Quach, Ziyi Jin, Fei Chiang, and Mostafa Milani. CurrentClean: Interactive Change Exploration and Cleaning of Stale Data. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2917–2920, Beijing China, November 2019. ACM. ISBN 978-1-4503-6976-3. doi: 10.1145/3357384.3357839. URL `https://dl.acm.org/doi/10.1145/3357384.3357839`.

[12] Aoqian Zhang, Shaoxu Song, Jianmin Wang, and Philip S. Yu. Time series data cleaning: from anomaly detection to anomaly repairing. *Proceedings of the VLDB Endowment*, 10(10):1046–1057, June 2017. ISSN 2150-8097. doi: 10.14778/3115404.3115410. URL `https://dl.acm.org/doi/10.14778/3115404.3115410`.

[13] Vikram Patil, Priyanka Singh, Shivam Parikh, and Pradeep K. Atrey. GeoSClean: Secure Cleaning of GPS Trajectory Data Using Anomaly Detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 166–169, Miami, FL, April 2018. IEEE. ISBN 978-1-5386-1857-8. doi: 10.1109/MIPR.2018.00037. URL `https://ieeexplore.ieee.org/document/8396993/`.

[14] Laure Berti-Equille. Learn2Clean: Optimizing the Sequence of Tasks for Web Data Preparation. In *The World Wide Web Conference on - WWW '19*, pages 2580–2586, San Francisco, CA, USA, 2019. ACM Press. ISBN 978-1-4503-6674-8. doi: 10.1145/3308558.3313602. URL `http://dl.acm.org/citation.cfm?doid=3308558.3313602`.

[15] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks. 2019. doi: 10.48550/ARXIV.1904.09483. URL `https://arxiv.org/abs/1904.09483`. Publisher: arXiv Version Number: 3.

[16] Sergey Redyuk, Zoi Kaoudi, Sebastian Schelter, and Volker Markl. DORIAN in action: assisted design of data science pipelines. *Proceedings of the VLDB Endowment*, 15(12):3714–3717, August 2022. ISSN 2150-8097. doi: 10.14778/3554821.3554882. URL `https://dl.acm.org/doi/10.14778/3554821.3554882`.

[17] Moritz Herrmann, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3):bbaa167, May 2021. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbaa167. URL `https://academic.oup.com/bib/article/doi/10.1093/bib/bbaa167/5895463`.

[18] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[19] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1. URL `https://doi.org/10.1186/s12874-018-0482-1`.

[20] Lili Zhao and Dai Feng. DNNSurv: Deep Neural Networks for Survival Analysis Using Pseudo Values. 2020. URL `https://arxiv.org/abs/1908.02337`. arXiv: 1908.02337.

[21] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. ISSN 1533-7928.

[22] Andrea Bommert, Thomas Welchowski, Matthias Schmid, and Jörg Rahnenführer. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics*, 23(1):bbab354, January 2022. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbab354. URL `https://academic.oup.com/bib/article/doi/10.1093/bib/bbab354/6366322`.

[23] Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. mlr3proba: An R Package for Machine Learning in Survival Analysis. *Bioinformatics*, February 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab039. URL `https://cran.r-project.org/package=mlr3proba`.

[24] Raphael Sonabend. survivalmodels: Models for Survival Analysis, 2020. URL `https://raphaels1.r-universe.dev/ui#package:survivalmodels`.

[25] Stephane Fotso. PySurvival: Open source package for Survival Analysis modeling, 2019. URL `https://www.pysurvival.io/`.

[26] Sebastian P\"olsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL `http://jmlr.org/papers/v21/20-729.html`.

[27] Muhammad Nabeel Asim, Muhammad Usman Ghani, Muhammad Ali Ibrahim, Waqar Mahmood, Andreas Dengel, and Sheraz Ahmed. Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification. *Neural Computing and Applications*, 33(11):5437–5469, June 2021. ISSN 0941-0643, 1433-3058. doi: 10.1007/s00521-020-05321-8. URL `https://link.springer.com/10.1007/s00521-020-05321-8`.

[28] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Christoph Zehe, Olivier Cloarec, Johan Trygg, Andreas Dengel, and Sheraz Ahmed. EL-RMLocNet: An explainable LSTM network for RNA-associated multi-compartment localization prediction. *Computational and Structural Biotechnology Journal*, 20:3986–4002, 2022. ISSN 20010370. doi: 10.1016/j.csbj.2022.07.031. URL `https://linkinghub.elsevier.com/retrieve/pii/S2001037022003117`.

[29] Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W. Jim Zheng, and Kirk Roberts. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. *Journal of Biomedical Informatics*, 115:103671, March 2021. ISSN 15320464. doi: 10.1016/j.jbi.2020.103671. URL `https://linkinghub.elsevier.com/retrieve/pii/S1532046420302999`.

[30] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal Differential Equations for Scientific Machine Learning, November 2021. URL `http://arxiv.org/abs/2001.04385`. arXiv:2001.04385 [cs, math, q-bio, stat].

[31] Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Neural Spatio-Temporal Point Processes, March 2021. URL `http://arxiv.org/abs/2011.04583`. arXiv:2011.04583 [cs].

[32] M. B. Short, M. R. D'Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes. A STATISTICAL MODEL OF CRIMINAL BEHAVIOR. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267, August 2008. ISSN 0218-2025, 1793-6314. doi: 10.1142/S0218202508003029. URL `https://www.worldscientific.com/doi/abs/10.1142/S0218202508003029`.

[33] Charles Loeffler and Seth Flaxman. Is Gun Violence Contagious? A Spatiotemporal Test. *Journal of Quantitative Criminology*, 34(4):999–1017, December 2018. ISSN 0748-4518, 1573-7799. doi: 10.1007/s10940-017-9363-8. URL `http://link.springer.com/10.1007/s10940-017-9363-8`.

[34] Songgaojun Deng and Yue Ning. A Survey on Societal Event Forecasting with Deep Learning, December 2021. URL `http://arxiv.org/abs/2112.06345`. arXiv:2112.06345 [cs].

[35] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. Learning Dynamic Context Graphs for Predicting Social Events. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1007–1016, Anchorage AK USA, July 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330919. URL `https://dl.acm.org/doi/10.1145/3292500.3330919`.

[36] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. Dynamic Knowledge Graph based Multi-Event Forecasting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1585–1595, Virtual Event CA USA, August 2020. ACM. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403209. URL https://dl.acm.org/doi/10.1145/3394486.3403209.

[37] Qi Wang, Guangyin Jin, Xia Zhao, Yanghe Feng, and Jincai Huang. CSAN: A neural network benchmark model for crime forecasting in spatio-temporal scale. *Knowledge-Based Systems*, 189:105120, February 2020. ISSN 09507051. doi: 10.1016/j.knosys.2019.105120. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705119304873.

[38] Guangyin Jin, Qi Wang, Xia Zhao, Yanghe Feng, Qing Cheng, and Jincai Huang. Crime-GAN: A Context-based Sequence Generative Network for Crime Forecasting with Adversarial Loss. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1460–1469, Los Angeles, CA, USA, December 2019. IEEE. ISBN 978-1-72810-858-2. doi: 10.1109/BigData47090.2019.9006388. URL https://ieeexplore.ieee.org/document/9006388/.

[39] Yaqian Wang, Liang Ge, Siyu Li, and Feng Chang. Deep Temporal Multi-Graph Convolutional Network for Crime Prediction. In Gillian Dobbie, Ulrich Frank, Gerti Kappel, Stephen W. Liddle, and Heinrich C. Mayr, editors, *Conceptual Modeling*, volume 12400, pages 525–538. Springer International Publishing, Cham, 2020. ISBN 978-3-030-62521-4 978-3-030-62522-1. doi: 10.1007/978-3-030-62522-1_39. URL http://link.springer.com/10.1007/978-3-030-62522-1_39. Series Title: Lecture Notes in Computer Science.

[40] Yue Wei, Weichao Liang, Youquan Wang, and Jie Cao. CrimeSTC: A Deep Spatial-Temporal-Categorical Network for Citywide Crime Prediction. In *2020 The 3rd International Conference on Computational Intelligence and Intelligent Systems*, pages 75–79, Tokyo Japan, November 2020. ACM. ISBN 978-1-4503-8808-5. doi: 10.1145/3440840.3440850. URL https://dl.acm.org/doi/10.1145/3440840.3440850.

[41] Maxim Kovalev, Lev Utkin, Frank Coolen, and Andrei Konstantinov. Counterfactual Explanation of Machine Learning Survival Models. *Informatica*, 32(4):817–847, 2021. ISSN 0868-4952, 1822-8844. doi: 10.15388/21-INFOR468. URL https://informatica.vu.lt/doi/10.15388/21-INFOR468.

[42] Lev V. Utkin, Egor D. Satyukov, and Andrei V. Konstantinov. SurvNAM: The machine learning survival model explanation. *Neural Networks*, 147:81–102, March 2022. ISSN 08936080. doi: 10.1016/j.neunet.2021.12.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608021004949.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL https://dl.acm.org/doi/10.1145/2939672.2939778.

[44] Maxim S. Kovalev and Lev V. Utkin. A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov–Smirnov bounds. *Neural Networks*, 132:1–18, December 2020. ISSN 08936080. doi: 10.1016/j.neunet.2020.08.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608020302963.

[45] Maxim S. Kovalev, Lev V. Utkin, and Ernest M. Kasimov. SurvLIME: A method for explaining machine learning survival models. *Knowledge-Based Systems*, 203:106164, September 2020. ISSN 09507051. doi: 10.1016/j.knosys.2020.106164. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705120304044.

[46] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. Understanding Event Predictions via Contextualized Multilevel Feature Learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 342–351, Virtual Event Queensland Australia, October 2021. ACM. ISBN 978-1-4503-8446-9. doi: 10.1145/3459637.3482309. URL https://dl.acm.org/doi/10.1145/3459637.3482309. Deng36.

[47] Fabrizio Angiulli, Fabio Fassetti, Giuseppe Manco, and Luigi Palopoli. Outlying property detection with numerical attributes. *Data Mining and Knowledge Discovery*, 31(1):134–163, January 2017. ISSN 1573-756X. doi: 10.1007/s10618-016-0458-x. URL https://doi.org/10.1007/s10618-016-0458-x.

[48] Lei Duan, Guanting Tang, Jian Pei, James Bailey, Akiko Campbell, and Changjie Tang. Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery*, 29(5):1116–1151, September 2015. ISSN 1573-756X. doi: 10.1007/s10618-014-0398-2. URL https://doi.org/10.1007/s10618-014-0398-2.

[49] Nikhil Gupta, Dhivya Eswaran, Neil Shah, Leman Akoglu, and Christos Faloutsos. Beyond Outlier Detection: LookOut for Pictorial Explanation. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 122–138, Cham, 2019. Springer International Publishing. ISBN 9783030109257. doi: 10.1007/978-3-030-10925-7_8.

[50] Haoyu Liu, Fenglong Ma, Yaqing Wang, Shibo He, Jiming Chen, and Jing Gao. LP-Explain: Local Pictorial Explanation for Outliers. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 372–381, November 2020. doi: 10.1109/ICDM50108.2020.00046. ISSN: 2374-8486.

[51] Jiongqian Liang and Srinivasan Parthasarathy. Robust Contextual Outlier Detection: Where Context Meets Sparsity. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2167–2172, New York, NY, USA, October 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983660. URL https://doi.org/10.1145/2983323.2983660.

[52] M. Y. Meghanath, Deepak Pai, and Leman Akoglu. ConOut: Contextual Outlier Detection with Multiple Contexts: Application to Ad Fraud. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 139–156, Cham, 2019. Springer International Publishing. ISBN 9783030109257. doi: 10.1007/978-3-030-10925-7_9.

[53] Ece Calikus, Sławomir Nowaczyk, Mohamed-Rafik Bouguelia, and Onur Dikmen. Wisdom of the contexts: active ensemble learning for contextual anomaly detection. *Data Mining and Knowledge Discovery*, 36(6):2410–2458, November 2022. ISSN 1573-756X. doi: 10.1007/s10618-022-00868-7. URL https://doi.org/10.1007/s10618-022-00868-7.

[54] Darrell Coles and Andrew Curtis. Efficient nonlinear Bayesian survey design using D_N optimization. *GEOPHYSICS*, 76(2):Q1–Q8, March 2011. ISSN 0016-8033, 1942-2156. doi: 10.1190/1.3552645. URL https://library.seg.org/doi/10.1190/1.3552645.

[55] Hojat Shirmard, Ehsan Farahbakhsh, R. Dietmar Müller, and Rohitash Chandra. A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sensing of Environment*, 268:112750, January 2022. ISSN 00344257. doi: 10.1016/j.rse.2021.112750. URL https://linkinghub.elsevier.com/retrieve/pii/S0034425721004703.

[56] Yuying Xie, Yeying Zhu, Cecilia A Cotton, and Pan Wu. A model averaging approach for estimating propensity scores by optimizing balance. *Statistical Methods in Medical Research*, 28(1):84–101, January 2019. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280217715487. URL http://journals.sagepub.com/doi/10.1177/0962280217715487.

[57] Christopher T. Lloyd, Alessandro Sorichetta, and Andrew J. Tatem. High resolution global gridded data for use in population studies. *Scientific Data*, 4(1):170001, January 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.1. URL https://www.nature.com/articles/sdata20171.

[58] Giulia Martini, Alberto Bracci, Sejal Jaiswal, Matteo Corea, Lorenzo Riches, Jonathan Rivers, and Elisa Omodei. Nowcasting food insecurity on a global scale. preprint, Nutrition, July 2021. URL http://medrxiv.org/lookup/doi/10.1101/2021.06.23.21259419.

[59] Neha Shah, Diwakar Mohan, Jean Juste Harisson Bashingwa, Osama Ummer, Arpita Chakraborty, and Amnesty E LeFevre. Using Machine Learning to Optimize the Quality of Survey Data: Protocol for a Use Case in India. *JMIR Research Protocols*, 9(8):e17619, August 2020. ISSN 1929-0748. doi: 10.2196/17619. URL https://www.researchprotocols.org/2020/8/e17619.

[60] Seema Jayachandran, Monica Biradavolu, and Jan Cooper. Using machine learning and qualitative interviews to design a five-question survey module for women's agency. *World Development*, 161:106076, January 2023. ISSN 0305750X. doi: 10.1016/j.worlddev.2022.106076. URL https://linkinghub.elsevier.com/retrieve/pii/S0305750X22002662.

[61] Roger Guimerà, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel Miranda, Jordi Pallarès, and Marta Sales-Pardo. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 6(5):eaav6971, January 2020. ISSN 2375-2548. doi: 10.1126/sciadv.aav6971. URL https://www.science.org/doi/10.1126/sciadv.aav6971.

[62] Silviu-Marian Udrescu and Max Tegmark. AI Feynman: a Physics-Inspired Method for Symbolic Regression. 2019. doi: 10.48550/ARXIV.1905.11481. URL https://arxiv.org/abs/1905.11481. Publisher: arXiv Version Number: 2.

[63] Raphael Sonabend, Andreas Bender, and Sebastian Vollmer. Avoiding C-hacking when evaluating survival distribution predictions with discrimination measures. *Bioinformatics*, 38(17):4178–4184, September 2022. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btac451. URL https://academic.oup.com/bioinformatics/article/38/17/4178/6640155.

[64] Raphael Sonabend, Florian Pfisterer, Alan Mishler, Moritz Schauer, Lukas Burk, Sumantrak Mukherjee, and Sebastian Vollmer. Flexible Group Fairness Metrics for Survival Analysis, July 2022. URL http://arxiv.org/abs/2206.03256. arXiv:2206.03256 [cs, stat].

[65] John M. Dennis, Bilal A. Mateen, Raphael Sonabend, Nicholas J. Thomas, Kashyap A. Patel, Andrew T. Hattersley, Spiros Denaxas, Andrew P. McGovern, and Sebastian J. Vollmer. Type 2 Diabetes and COVID-19–Related Mortality in the Critical Care Setting: A National Cohort Study in England, March–July 2020. *Diabetes Care*, 44(1):50–57, January 2021. ISSN 0149-5992, 1935-5548. doi: 10.2337/dc20-1444. URL https://diabetesjournals.org/care/article/44/1/50/33095/Type-2-Diabetes-and-COVID-19-Related-Mortality-in.

[66] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. ADH-PPI: An attention-based deep hybrid model for protein-protein interaction prediction. *iScience*, 25(10):105169, October 2022. ISSN 25890042. doi: 10.1016/j.isci.2022.105169. URL https://linkinghub.elsevier.com/retrieve/pii/S2589004222014419.

[67] Bilal Akhter Mateen, Harrison Wilde, John M Dennis, Andrew Duncan, Nick Thomas, Andrew McGovern, Spiros Denaxas, Matt Keeling, and Sebastian Vollmer. Hospital bed capacity and usage across secondary healthcare providers in England during the first wave of the COVID-19 pandemic: a descriptive analysis. *BMJ Open*, 11(1):e042945, January 2021. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2020-042945. URL https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2020-042945.

[68] David Antony Selby. *Statistical modelling of citation networks, research influence and journal prestige*. PhD thesis, University of Warwick, Coventry, UK, June 2020. URL http://wrap.warwick.ac.uk/158562.

[69] Christine Lamanna, Kusum Hachhethu, Sabrina Chesterman, Gaurav Singhal, Beatrice Mwongela, Mary Ng'endo, Silvia Passeri, Arghanoon Farhikhtah, Suneetha Kadiyala, Jean-Martin Bauer, and Todd S. Rosenstock. Strengths and limitations of computer assisted telephone interviews (CATI) for nutrition data collection in rural Kenya. *PLOS ONE*, 14(1):e0210050, January 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0210050. URL https://dx.plos.org/10.1371/journal.pone.0210050.

[70] Sebastian J. Vollmer. Dimension-Independent MCMC Sampling for Inverse Problems with Non-Gaussian Priors. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):535–561, January 2015. ISSN 2166-2525. doi: 10.1137/130929904. URL http://epubs.siam.org/doi/10.1137/130929904.

[71] Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6), December 2014. ISSN 1050-5164. doi: 10.1214/13-AAP982. URL https://projecteuclid.org/journals/annals-of-applied-probability/volume-24/issue-6/Spectral-gaps-for-a-MetropolisHastings-algorithm-in-infinite-dimensions/10.1214/13-AAP982.full.

[72] Alexandre Bouchard-Côté, Sebastian J. Vollmer, and Arnaud Doucet. The Bouncy Particle Sampler: A Nonreversible Rejection-Free Markov Chain Monte Carlo Method. *Journal of the American Statistical Association*, 113(522):855–867, April 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1294075. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1294075.

[73] Mohammad El Wajeh, Falco Jung, Dominik Bongartz, Chrysoula Dimitra Kappatou, Narmin Ghaffari Laleh, Alexander Mitsos, and Jakob Nikolas Kather. Can the Kuznetsov Model Replicate and Predict Cancer Growth in Humans? *Bulletin of Mathematical Biology*, 84(11):130, November 2022. ISSN 0092-8240, 1522-9602. doi: 10.1007/s11538-022-01075-7. URL https://link.springer.com/10.1007/s11538-022-01075-7.

[74] Abdur Rehman, Kashif Javed, Haroon A. Babri, and Muhammad Nabeel Asim. Selection of the most relevant terms based on a max-min ratio metric for text classification. *Expert Systems with Applications*, 114:78–96, December 2018. ISSN 09574174. doi: 10.1016/j.eswa.2018.07.028. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417418304457.

[75] Faiza Mehmood, Muhammad Usman Ghani, Muhammad Nabeel Asim, Rehab Shahzadi, Aamir Mehmood, and Waqar Mahmood. MPF-Net: A computational multi-regional solar power forecasting framework. *Renewable and Sustainable Energy Reviews*, 151:111559, November 2021. ISSN 13640321. doi: 10.1016/j.rser.2021.111559. URL https://linkinghub.elsevier.com/retrieve/pii/S1364032121008376.

[76] Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. SurvSHAP(t): Time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262:110234, February 2023. ISSN 09507051. doi: 10.1016/j.knosys.2022.110234. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705122013302.

[77] Oscar Giles, Kasra Hosseini, Grigorios Mingas, Oliver Strickson, Louise Bowler, Camila Rangel Smith, Harrison Wilde, Jen Ning Lim, Bilal Mateen, Kasun Amarasinghe, Rayid Ghani, Alison Heppenstall, Nik Lomax, Nick Malleson, Martin O'Reilly, and Sebastian Vollmerteke. Faking feature importance: A cautionary tale on the use of differentially-private synthetic data, March 2022. URL http://arxiv.org/abs/2203.01363. arXiv:2203.01363 [cs, stat].

[78] Jeroen Berrevoets, Fergus Imrie, Trent Kyono, James Jordon, and Mihaela van der Schaar. To Impute or not to Impute? Missing Data in Treatment Effect Estimation. 2022. doi: 10.48550/ARXIV.2202.02096. URL https://arxiv.org/abs/2202.02096. Publisher: arXiv Version Number: 3.

[79] Alicia Curth, Changhee Lee, and Mihaela van der Schaar. SurvITE: Learning Heterogeneous Treatment Effects from Time-to-Event Data. In *Advances in Neural Information Processing Systems*, volume 34, pages 26740–26753. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc//paper/2021/hash/e0eacd983971634327ae1819ea8b6214-Abstract.html.

[80] John M Dennis, Katherine G Young, Andrew P McGovern, Bilal A Mateen, Sebastian J Vollmer, Michael D Simpson, William E Henley, Rury R Holman, Naveed Sattar, Ewan R Pearson, Andrew T Hattersley, Angus G Jones, and Beverley M Shields. Development of a treatment selection algorithm for SGLT2 and DPP-4 inhibitor therapies in people with type 2 diabetes: a retrospective cohort study. *The Lancet Digital Health*, 4(12):e873–e883, December 2022. ISSN 25897500. doi: 10.1016/S2589-7500(22)00174-1. URL https://linkinghub.elsevier.com/retrieve/pii/S2589750022001741.

[81] Andrew P. McGovern, Nick J. Thomas, Sebastian J. Vollmer, Andrew T. Hattersley, Bilal A. Mateen, and John M. Dennis. The disproportionate excess mortality risk of COVID-19 in younger people with diabetes warrants vaccination prioritisation. *Diabetologia*, 64(5):1184–1186, May 2021. ISSN 0012-186X, 1432-0428. doi: 10.1007/s00125-021-05404-8. URL http://link.springer.com/10.1007/s00125-021-05404-8.

[82] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep One-Class Classification.

[83] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. DEEP SEMI-SUPERVISED ANOMALY DETECTION. 2020.

[84] Konstantin Klemmer, Tianlin Xu, Beatrice Acciaio, and Daniel B. Neill. SPATE-GAN: Improved Generative Modeling of Dynamic Spatio-Temporal Patterns with an Autoregressive Embedding Loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4523–4531, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i4.20375. URL https://ojs.aaai.org/index.php/AAAI/article/view/20375.

[85] Dominique Mercier, Jwalin Bhatt, Andreas Dengel, and Sheraz Ahmed. Time to Focus: A Comprehensive Benchmark using Time Series Attribution Methods:. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, pages 562–573, Online Streaming, — Select a Country —, 2022. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-547-0. doi: 10.5220/0010904400003116. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010904400003116.

[86] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, September 2021. ISSN 09252312. doi: 10.1016/j.neucom.2021.03.091. URL `https://linkinghub.elsevier.com/retrieve/pii/S092523122100477X`.

[87] K. S. M. Tozammel Hossain, Hrayr Harutyunyan, Yue Ning, Brendan Kennedy, Naren Ramakrishnan, and Aram Galstyan. Identifying geopolitical event precursors using attention-based LSTMs. *Frontiers in Artificial Intelligence*, 5: 893875, October 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.893875. URL `https://www.frontiersin.org/articles/10.3389/frai.2022.893875/full`.

[88] Sebastian Schutte and Claire Kelling. A Monte Carlo analysis of false inference in spatial conflict event studies. *PLOS ONE*, 17(4):e0266010, April 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0266010. URL `https://dx.plos.org/10.1371/journal.pone.0266010`.

[89] Tim C. D. Lucas, Anita K. Nandi, Elisabeth G. Chestnutt, Katherine A. Twohig, Suzanne H. Keddie, Emma L. Collins, Rosalind E. Howes, Michele Nguyen, Susan F. Rumisha, Andre Python, Rohan Arambepola, Amelia Bertozzi-Villa, Penelope Hancock, Punam Amratia, Katherine E. Battle, Ewan Cameron, Peter W. Gething, and Daniel J. Weiss. Mapping malaria by sharing spatial information between incidence and prevalence data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(3):733–749, June 2021. ISSN 0035-9254, 1467-9876. doi: 10.1111/rssc.12484. URL `https://onlinelibrary.wiley.com/doi/10.1111/rssc.12484`.

[90] Yusuke Tanaka, Tomoharu Iwata, Toshiyuki Tanaka, Takeshi Kurashima, Maya Okawa, and Hiroyuki Toda. Refining Coarse-Grained Spatial Data Using Auxiliary Spatial Data Sets with Various Granularities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5091–5099, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33015091. URL `https://ojs.aaai.org/index.php/AAAI/article/view/4442`.

[91] Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890):695–700, December 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-04198-4. URL `https://www.nature.com/articles/s41586-021-04198-4`.

[92] Philip J. Brown and Clive D. Payne. Aggregate Data, Ecological Regression, and Voting Transitions. *Journal of the American Statistical Association*, 81(394):452–460, June 1986. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1986.10478290. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478290`.

[93] Seth R. Flaxman, Yu-Xiang Wang, and Alexander J. Smola. Who Supported Obama in 2012?: Ecological Inference through Distribution Regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298, Sydney NSW Australia, August 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783300. URL `https://dl.acm.org/doi/10.1145/2783258.2783300`.

[94] Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Learning Hidden Markov Models from Aggregate Observations, November 2021. URL `http://arxiv.org/abs/2011.11236`. arXiv:2011.11236 [cs, eess].

[95] Tomoharu Iwata and Hitoshi Shimizu. Neural Collective Graphical Models for Estimating Spatio-Temporal Population Flow from Aggregated Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3935–3942, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33013935. URL `https://ojs.aaai.org/index.php/AAAI/article/view/4283`.

[96] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, July 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24638-z. URL `https://www.nature.com/articles/s41467-021-24638-z`.

[97] Harrison Wilde, Lucia L. Chen, Austin Nguyen, Zoe Kimpel, Joshua Sidgwick, Adolfo De Unanue, Davide Veronese, Bilal Mateen, Rayid Ghani, and Sebastian Vollmer. A recommendation and risk classification system for connecting rough sleepers to essential outreach services. *Data & Policy*, 3:e2, 2021. ISSN 2632-3249. doi: 10.1017/dap.2020.23. URL `https://www.cambridge.org/core/product/identifier/S2632324920000231/type/journal_article`.