POSE: Phase I: Feature-wise Focused Open-Source Ecosystem for Time-Series Analysis.

Proposal submission on Research.gov or Grants.gov

.[DEADLINE: October, 2022.]

- ☐ Project Summary [One (1) page max].
- ☐ FastLane documentation
  - ☐ Collaborators and other affiliations
  - ☐ Current and pending support
  - ☐ Bio sketch
  - ☐ Budget
  - ☐ Data management
  - ☐ Equipment and facilities
  - ☐ Sub-award documentation

Project Description. -Seven (7) pages max-

- ☐ Context of OSE
- ☐ Broader Impacts
- ☐ Ecosystem Discovery
- ☐ Organization and Governance
- ☐ Community Building

Letters of collaboration. Minimum 3 and maximum 5.

- ☐ Letter of collaboration from Lilie Idea Lab.
- ☐ Letter of collaboration from CMU.
- ☐ Letter of collaboration from AWS (email contact from Lilie).
- ☐ letter of collaboration from LinkedIn (QQ).
- ☐ Letter of collaboration from Google (Haifeng). Open collective.
  Rice, Trane, NI, GM (maybe).

Other

- ☐ Security Plan

Project summary

1 Overview

The proposed open-source ecosystem aims to increase the reach of three open-source packages, Time-series Outlier Detection (TODS), AutoKeras, and PyOD, which have gathered more than 15,000 stars on GitHub from more than 10k users. The Open Source Ecosystem proposed here is designed to tackle distinct business intelligence challenges while enabling human interpretation to help subject matter experts understand, trust, and leverage advanced machine learning (ML) technologies. In addition, the proposed ecosystem intends to grow the current open-source developer community and reach new domain experts through a series of events designed to enable new pathways for the development of collaborative open-source in the time-series domain that could lead to new technology products or services that have broad societal impacts. This POSE proposal brings together world-class US universities (Carnegie Mellon University and University of Texas - Austin), top corporations that are deploying machine learning at a large scale (LinkedIn), leading idea labs for Innovation and Entrepreneurship (Lilie lab at Rice University), and small businesses with strong research capabilities.

2 Intellectual Merit

While the Pathways to Enable Open-Source Ecosystems (POSE) program is not intended to fund the development of open-source research products, the open-source presented as a foundation of the Open-Source Ecosystem (OSE) involves novel and significant advances in ML Automation and Interpretation applied to anomaly detection. Each advance has dramatically extended current ML techniques toward dealing with deep learning challenges raised in real-world business intelligence data. A flexible and modular end-to-end anomaly detection system, different from current empirical approaches, is needed by the open-source community to automatically decide the optimal configuration of complex modeling pipelines to tackle their academic and industrial time-series data challenges.

3 Broader impacts

The proposed project's successful outcome will expand the impact of AI POW's open-source developments in dealing with an emerging and critical ML automation and interpretation problem with significant industrial applications. One example is that this project's results will have an immediate and substantial impact on improving the manufacturing industry's performance, enabling production line managers to simplify the anomaly identification and prediction task while understanding the flaws' origin to elevate product quality.

Similarly, as data science and engineering become more critical for businesses competing in a globalized digital economy, it is paramount to consolidate a 21st-century data-capable workforce. Many large enterprises can implement analysis detection systems because they can afford a highly specialized data scientist team. The proposed project will help create an open-source ecosystem for domain experts to benefit from affordable and easy-to-use business intelligence digital tools tailored to software engineers with more basic programming understanding. A phase 2 follow-up proposal of this open-source ecosystem can allow a larger group of small and medium businesses to adopt and benefit from advanced ML tools and embrace the data revolution by using their data assets in new innovative ways.

Keywords: CISE; Outlier Detection; Machine Learning; Explainable AI; AutoML.

POSE: Phase I: Feature-wise open-source Ecosystem for Time Series Applications.

1 Project Description

In this proposal we introduce the open-source projects, Time-series Outlier Detection (TODS), AutoKeras, and Python Outlier Detection (PyOD), which combined have gathered more that 15,000 stars on GitHub and over 500,000 downloads, from more than 10,000 users. The proposed Open Source Ecosystem (OSE) combines the three packages that combined offer a flexible and modular end-to-end anomaly detection suite of open-source packages. Particularly, the proposing team developed TODS and PyOD as an AutoML solution to reduce human efforts in developing anomaly detection models. Such AutoML tool involves the computational elements to provide a highly efficient anomaly detection system by systematically investigating three types of automation: pipeline searching, neural architecture search, and active anomaly detection with meta-learning. Besides, to help domain experts better understand the results generated by the anomaly detection system, the OSE introduces features from AutoKeras to enable interpretability and provide human-friendly explanations to end-users. This proposed Pathways to Enable Open-Source Ecosystems (POSE) will expand the reach of our open-source solutions and allow domain experts with basic programming knowledge to detect, predict, and understand anomalies for their business analysis challenges in ways that previously were extremely difficult, if not impossible.

1.1 Open-Source Ecosystem, Current State of Innovation

We plan to develop a sustainable OSE based on a promising open-source product that has the following innovation in the emergent machine learning (ML)-based time-series outlier detection.

Highly modular end-to-end anomaly detection system. Our ecosystem has implemented various carefully selected components to provide all the necessary functionalities to construct a machine learning pipeline. The system architecture's flexibility is crucial to meet emergent national needs across different industries and to build a community around our proposed OSE.

AutoML for anomaly detection. To further save the engineering efforts and allow domain experts to build an anomaly detection system with limited machine learning knowledge, we employ AutoML to minimize the technical barrier of entry for non-computer scientists.

Interpretations for anomaly detection. Apart from a well-performing solution, domain experts also demand interpretability. We offer interpretable machine learning solutions for users to understand the mechanism behind the pipeline's decision; and the OSE also enables interpretations of AutoML for users to learn why the machine constructs such pipelines.

1.2 Context of the Open-Source Ecosystem

This POSE phase I proposal brings together two world-class US universities with strong state-of-the-art machine learning research (Carnegie Mellon University and University of Texas at Austin), leading corporations that are deploying machine learning at a large scale (Microsoft-LinkedIn), top idea labs for Innovation and Entrepreneurship (Lilie lab at Rice University), and small businesses with strong research capabilities (AI POW LLC).

The resulting OSE will lead to an open-source community that can easily help our current and new users to access features for their time-series analysis and applications. One immediate goal is to tailor this ecosystem to be a platform that can easily integrate multiple vertical applications depending on the interest of the domain experts. A few verticals where PyOD and TODS have been used are defect detection (energy), weather forecast (agriculture), quality assurance (manufacturing), fraud detection (finance) and remote asset monitoring (IoT), to name a few.

As part of our community building activities we have planned a five-day idea lab and a multidisciplinary hackathon that has commitment from our partnering companies and universities.

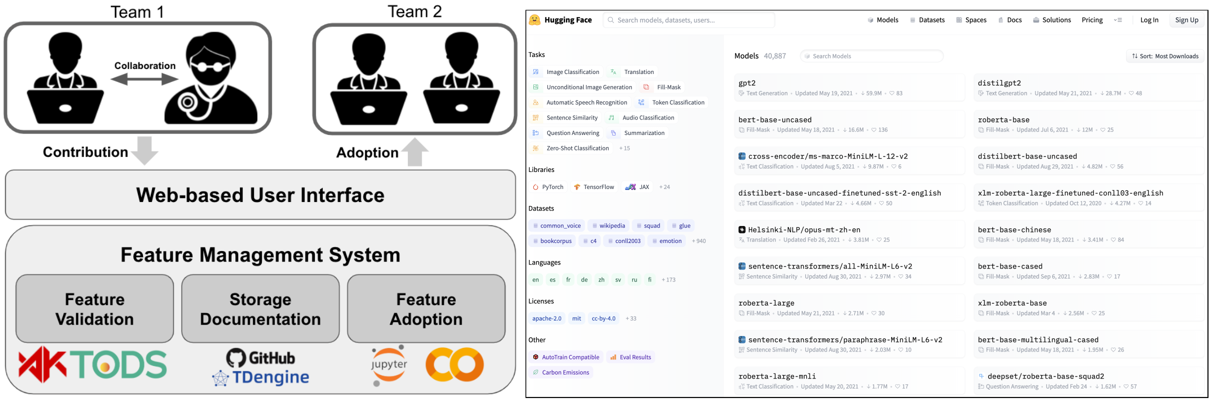

Figure 1: Illustration of open source ecosystem and an web-based interface example of HuggingFace.

2 Pain points Identification and Solutions

Pain Point Description. The applications of time series analysis range from future forecastings such as weather, traffic, or market forecasting and outlier detection such as fraudulent, money laundry, intrusion, and manufacturing fault detection. Analyzing time-series data is extremely difficult due to the complex temporal dependencies between individual time points and the correlations between individual variables (e.g., sensors, channels, features). Furthermore, given different tasks on the same time-series data, the corresponding analysis will be significantly different. Typically, similar tasks in time series data are usually equipped with similar data attributes. For example, sensors for weather monitoring are generally identical and, therefore, may share a mutual feature analysis process under certain circumstances. That is, features extracted for forecasting Seattle downtown may be applicable to forecasting Vancouver city's weather due to the spatial similarity. Another example could be the features extracted for analyzing a domain expert could adopt vehicle engine manufacturing line by analyzing HVAC system manufacturing line as they all use electricity, vibration, and torque sensors to monitor the conveyor belt of the manufacturing line. However, extracting new features from the data science team requires comprehensive studies on the data with domain experts, and many existed efforts cannot be well integrated due to the lack of a unified platform; therefore, identical works are repeatedly performed and lead to low efficiency and costly process. To this end, establishing an open-source ecosystem that allows developers and data scientists to share and collect valuable feature analysis methods is essential to an effective and efficient process for time series analysis.
Proposed Approaches Figure 1 illustrates the proposed open-source ecosystem. The entire ecosystem focuses on a feature management system, which focuses on feature validation, feature storage and documentation, and feature adoption. A web-based application serves as an interface to allow users to upload, document, and test the publicly available features for their tasks to enable users to contribute and adopt existing features in the ecosystem.

3 Organization and Governance

We propose the following methods for continuous development, integration and deployment processes. The AWS infrastructure will be provided by AI POW LLC and will enable further open, asynchronous, and distributed development of our OSE. The following measures have been established to ensure quality control, security and privacy of new content, and support for users:

- **Feature Validation.** As we aim to enable users to provide feature extraction methods within a few lines of code fraction, this module aims to ensure the quality of the contributed code fraction for feature extraction. Specifically, we will define a unified function structure (e.g., input/output of the function, data structure) to ensure the consistency and readability of the contributions. Then, to validate the quality and effectiveness, we will employ an automated machine learning framework including AutoKeras [1] and TODS [2] to plugin the contributed code fraction for extracting features from existing publicly available datasets. Then, the two AutoML engines will leverage the feature extracted from the contributed code fraction to validate the performance improvements on a wide range of datasets and tasks to ensure the quality of contributed features.

- ☐ **Storage and Documentation.** To store a wide range of publicly available datasets and the contributed code fractions with corresponding documentation for feature validation and adoption, this module aims to provide a unified structure to allow users to contribute code fractions with comprehensive documentation in an efficient way. To achieve this goal, we will develop a storage and documentation system upon GitHub API and TDEngine. Specifically, the contributed code will be stored in a publicly available GitHub repository and licensed by the GNU GPL [3] to ensure the copyright of the contributor. Additionally, we will collect and categorize existing publicly available datasets and leverage the TDEngine to store the datasets to enable efficient data retrieval for feature validation and adoption.
- ☐ **Feature Adoption.** To further allow users to testify the existing features on their application scenarios, we built a testing platform upon Google Colab and Jupyer notebook. In this way, users will be able to develop their solutions in our proposed open feature community. Additionally, AutoKeras and TODS will be available to users to provide efficient AutoML solutions to accelerate the development process.

The entire ecosystem will equip with a web-based user interface to allow users to exploit and contribute features without a complex environment and communicate with each other toward a better quality of code fractions. Having Huggingface as a reference point, the web-based user interface will be formed as a forum, where users can browse existing features based on their categories, target tasks, or tags. Furthermore, on the page of each feature, the documentation of the code fractions and corresponding discussions will be presented, and adopting the code fractions into Google Colab will function on the same page as well. In addition, an email registration will be required to access advanced functionalities such as feature adoption. Each member will be encouraged to contribute features to gain additional access to browse more features or highly effective features (e.g., features that lead to a certain level of performance improvement). In this way, a we aim to prompt users to contribute more quality features and discussions.

3.1 Project Team

The organizing lead team will coordinate the community building activities (Alfredo), and OSE development and deployment guidelines (Xia).

PI Alfredo Costilla-Reyes is Chief Research Officer at AI POW LLC, a Hispanic-led company established from research conducted at Rice University in the Department of Computer Science. Dr. Costilla is leading the product R&D in time-series anomaly detection and explainable AI for IoT devices. He is also the **first-generation graduate** from both the Entrepreneurship and Technology Commercialization program at Mays Business School and the doctorate program in Electrical Engineering, both from Texas A&M University. **Dr. Alfredo has been a recipient of the NSF Small Business Innovation Research grant, the Rice Innovation Fellowships, and the prestigious Mexico's National Award presented by the president of Mexico for his contributions to**

science, technology, and entrepreneurship. Specifically, Dr. Alfredo has led projects regarding embedded software and systems for future agriculture, battery-less wearable consumer electronics, application-specific integrated circuits, and wireless systems for IoT applications. His research and entrepreneurial endeavors have been previously funded by Venture Capital.

Senior Personnel Dr. Xia "Ben" Hu is an Associate Professor at Rice University in the Department of Computer Science. He has published more than 100 papers in major data mining venues. His articles have received seven Best Paper Awards (candidate), and he is the recipient of the JP Morgan AI Faculty Award, the Adobe Data Science Award, and the NSF CAREER Award. An open-source package developed by his group, namely AutoKeras, has become the most used automated deep learning system on Github (with over 8,500 stars and 1,400 forks). Dr. Hu's work on deep collaborative filtering, anomaly detection, and knowledge graphs is part of the TensorFlow package, Apple production system, and Bing production system. Hu's work has been cited more than 10,000 times. He was the conference General Co-Chair for WSDM 2020.

The team has co-authored publications of fundamental research in developing interpretable, automated, and interactive machine learning systems. Note that, while the preliminary work produced by the DARPA and NSF projects complements and strengthens our open-source code, *the proposed POSE project will investigate along with a significantly different direction* which focuses on developing a new OSE managing organization, responsible for the creation and maintenance of infrastructure needed for efficient and secure operation of an OSE based around TODS, PyOD, and AutoKeras. Our past collaborations have provided a solid ground to potentiate the synergies emerging from continuous multidisciplinary interactions dedicated to the proposed OSE.

4 Pointer to the existing publicly-available open-source product that is being transitioned

Several publicly-available open-source projects have incorporated feature extraction modules. **Ts-learn**[1] [4] is a general-purpose Python toolkit for time-series analysis. It provides various tools for pre-processing and feature extraction. Some basic machine learning models are also provided for standard tasks, such as classification and clustering. **Ts-fresh**[2] is another package that provides systematic time-series feature extraction functionaries. It includes statistics, signal processing, and nonlinear dynamics feature extraction modules. In addition, it provides feature section algorithms to forget irrelevant features. **sklearn**[3] is a general machine learning package, which also contains many functions for feature extraction. In addition to feature extraction, it also provides various machine learning modules for tasks such as classification and outlier detection. While these packages have identified the importance of feature sections in time-series analysis, they only provided limited choices of feature extraction modules. Moreover, they have not built an interactive community to allow users to contribute new feature extraction methods.

Similar to our objective, Hugging Face[4] [5] has built a community for natural language models. Users in Hugging Face can easily download the models of interest or contribute new models to the community. The interactive nature of Hugging Face makes a significant impact and attracts lots of users. Our efforts follow the core idea of Hugging Face but focus on the time-series analysis community. Specifically, instead of allowing users to contribute datasets or pre-trained, we encourage users to share good time-series feature extraction codes. Then other users can reuse these feature extraction codes in similar projects.

[1]https://github.com/tslearn-team/tslearn

[2]https://github.com/blue-yonder/tsfresh

[3]https://github.com/scikit-learn/scikit-learn

[4]https://github.com/huggingface

5 Details on the current status of the research product, development model, methods of dissemination, and user base

We have published several academic papers regarding time-series analysis, which will serve as the basis of our open-source community. Our research paper titled "Revisiting Time Series Outlier Detection: Definitions and Benchmarks" [6] has defined and studied different types of outliers in time-series data. Specifically, we proposed a behavior-driven taxonomy to clearly define the contexts of the outliers. Our taxonomy first categorizes time-series outliers as point-wise and pattern-wise outliers. The point-wise outlier is further categorized as global, contextual, and collective outliers, and the pattern-wise outliers are further divided into shapelet, seasonal, and trend outliers. Each type of outliers has clear context definitions. Based on the proposed behavior-driven taxonomy, we benchmark different types of algorithms. Our paper titled "Towards Similarity-Aware Time-Series Classification" [7] studied the time-series classification problem. This work examines the performance of traditional similarity-based methods and modern deep learning models for time-series classification under different amounts of supervision. Motivated by the various pros and cons of these two different lines of research, we proposed a new framework that leverages graph neural networks to incorporate similarity information into the automated feature representation learning of deep neural networks. We demonstrated that our proposed algorithm outperformed other state-of-the-art time-series classification algorithms on the benchmark datasets. Our research and codes will serve as the code basis of the open-source ecosystem. In addition, we have several publications of AutoML [1, 8–10], which will contribute to the development of the proposed AutoML system.

Several of our developed open-source projects have attracted the open-source community's attention. Their users will potentially become our first batch of contributors to our open-source ecosystem. First, the open-sourced system developed by the team, named Auto-Keras[5] [1], has made significant impacts in both industry and academia. It has become the most used (>8,500 stars, 1,400 forks, and 500,000 downloads) automated deep learning system on Github. This system has been reported by various news media such as Towards Data Science, Analytics Vidhya, Analytics India Magazine, and the most influential technology news media in China, such as Jiqizhixing. Another system, named Automated Time Series Outlier Detection System (TODS)[6] [11], is one of the most popular time-series outlier detection systems. It provides an exhaustive list of primitives for time-series processing, feature analysis, detection algorithms, and reinforcement, with more than 70 primitives in total. Users can flexibly create different types of pipelines based on the primitives. Our implementations of various preprocessing primitives will be our initial preprocessing modules in the ecosystem. We have the ability to redirect the users of TODS to our ecosystem. Our package AutoVideo[7] [12] further extends the TODS to video analysis. Like time series, videos are also streaming data, and thus many modules in AutoVideo could also be used to build the ecosystem. All our previous experiences and codes will be leveraged to build the proposed ecosystem quickly. The users of our existing packages are our early users.

We have been closely working with our collaborators in CMU, and Linkedin, who will help us promote the ecosystem. **Yue Zhao** from CMU is the author of PyOD[8] [13], the most popular outlier detection system on Github with >5,000 stars, 1,000 forks, and 6,000,000 downloads. PyOD has a large user base, and outlier detection is a key application of PyOD. We will collaborate with the PyOD team to build our ecosystem. We will ensure that the feature extraction modules in our

[5]https://github.com/keras-team/autokeras

[6]https://github.com/datamllab/tods

[7]https://github.com/datamllab/autovideo

[8]https://github.com/yzhao062/pyod

system are compatible with PyOD so that the users of PyOD can become part of the ecosystem. **Professor Xia (Ben) Hu** from Rice University and **Dr. Qingquan Song** from Linkedin are the main authors of AutoKeras. We will work closely with them to make the interface of our feature extraction module compatible with AutoKeras. And the algorithms in AutoKeras will be the core searching algorithms in our system.

6 Ecosystem Discovery

Given the early traction of our three open source packages reflected in starts, forks and downloads in GitHub, the TODS-POSE team is planning to further develop a high quality ecosystem using a automated machine learning approach. Our team has seen the need to implement a proof-of-concept automated machine to help evaluate the quality of the contributed features. Specifically, when current and new users submit a new feature, our system will automatically launch an automated search with the submitted features on all the benchmark datasets. The TODS-POSE partners will generate a report to reflect the performance of the contributed feature on datasets from the different domains currently supported by our open-source packages. We will only accept the feature that reaches a bar for quality control. The generated report will also give insights of the type of users that uses our tools, and the types of features needed in their domain. We believe that controlling and interpreting the quality of the contributed features early-on in the ecosystem discovery is particularly important because it can help attract users, which will be enabled by our open-source technology of automated machine learning.

This quality-assured feature-sharing ecosystem for time series analysis is a rigid demand and crucial **to properly identify researchers and engineers working in this domain**. First, although all kinds of time series data are equipped with timestamps, different applications require different feature extraction methods. For example, extracting features from the timestamp of the transaction log requires seasonality analysis, while extracting features from the timestamp of an IoT device requires periodicity. Therefore, our OSE web interface will be equipped with a simple feature recommender system based on our AutoML-based feature evaluation system for the target tasks. Second, similar analysis methods may be shared across datasets even though the attributes are different. For instance, The domain expert may directly adapt analysis methods of electricity sensors for fault detection in manufacturing lines to analyze electricity sensors for remote asset monitoring for HVAC systems due to the similar behavior of sensors. This way, by incorporating the feature sets with our proposed feature recommender system, the time cost of the development process will be significantly reduced for developers since they no longer need to start from scratch.

While we already have the core open-source technology to deal with the time-series data and automated machine learning tools to boost the model performance, OSE is necessary to further develop the audience we have attracted. On the other hand, the many users that visit our open-source products can potentially become code contributors in the later stages to help improve the quality of the codebase. OSE will significantly help us discover how we can more easily onboard new users that can benefit from our system and have a jumpstart on their time-series applications.

The TODS-OSE team knows that open-source users are more inclined to adopt a passive role of only using the available open-source code and not necessarily actively contributing, which is detrimental to the sustainability of our long-term vision of the OSE we propose. Therefore, the most challenging part of discovering the potential impact of our OSE is to prompt users to make a contribution. To identify such users and the favorite ways to facilitate code contributions, we will start with the current users of TODS, AutoKeras, and PyOD (the packages have 15k stars from more than 10k users on GitHub). In our community building activities we will provide various demonstrative use-cases with a limited number of features that can be directly adapted to the three packages without development efforts. Specifically, we will experiment with incorporating a web interface that can consolidate the three packages with limited numbers of features and multiple features developing templates. Then, we will ask users to do minimum modifications on different templates to create features to gain access to more effective features. This experiment will allow us to develop an ideal mechanism for the user contribution process and identify the type of users that may help build the virtuous cycle. The same type of users will be invited to experience a phase II POSE proposal and gain extra credits to access more functionalities and features.

7 Community Building

The specific activities that will aid in developing a strategy to engage potential content contributors who will help build and maintain the open-source product. We have planned the following mechanisms to engage with our communities:

Hackathon/competition: Our partners at CMU, UT-Austin, and Rice University will provide the venue to host a hackathon that will be fully focused on the proposed OSE-TODS. A competition organized in three major US Universities will allow the proposing team to demonstrate the interest of the academic and industrial community in the OSE. It is worth mentioning that the hackathon winners will be provided with mentoring at Rice University to participate in technology commercialization activities such as participation at the OwlSpark accelerator.

Workshops: Our collaborator from LinkedIn will provide valuable industry insights to help with workshops programmed for the general public. The sessions we have planned will cover introductory topics on time-series analysis for new users, as well as a more specific program that covers particular use cases in industries such as manufactury, energy, industrial Internet of Things (IoT), and remote asset monitoring for medium and advanced users that require time-series analysis in their day-to-day data science work.

Ideas Labs: The Liu Idea Lab for Innovation and Entrepreneurship (Lilie), and The Ken Kennedy Institute, both at Rice University, will help organize an Idea Lab that intends to attract domain experts from the US Southwest. This 'TODS Idea Lab' is programmed to be a five-day intensive and interactive work with a diverse group of participants from various disciplines and backgrounds to immerse in collaborative thinking processes to construct innovative approaches to use TODS in their domains. We expect to invite 35 participants that will identify and define the scope of the challenges relating to forming and sustaining an OSE based on TODS. Following the Ideas Lab, the teams with the best ideas may be invited to submit a phase 2 POSE proposal.

8 Broader Impacts

his POSE project will enhance partnerships between academia and small businesses with strong research capabilities in the US. We firmly believe that, by partnering with the Hispanic-led startup AI POW LLC, we will be driving diversity, excellence and innovation among our research members and third-party collaborations. Particularly, Mexico's presidential-award recipient, PI Alfredo Costilla-Reyes, brings a unique entrepreneurial approach to challenges that has helped him and his team recognize different opportunities and embrace them through AI POW's SBIR-funded engineering projects. We are confident that Dr. Costilla's strong technical and entrepreneurial acumen will further contribute to having an exceptional inclusive, equitable, and diverse POSE research partnership for many other students and collaborators at Rice University.

Harnessing the Data Revolution. As data science and engineering become more critical for businesses, it is paramount to consolidate a 21st-century data-capable workforce. This project aims to democratize easy-to-use tools tailored to software engineers with basic ML knowledge, **and accelerate the technology adoption for a large group of small and medium sized businesses that can now embrace the data revolution** to start using their data assets to innovate in new ways.

References

[1] Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1946–1956, 2019.

[2] Kwei-Herng Lai, Daochen Zha, Guanchu Wang, Junjie Xu, Yue Zhao, Devesh Kumar, Yile Chen, Purav Zumkhawaka, Minyang Wan, Diego Martinez, et al. Tods: An automated time series outlier detection system. *arXiv preprint arXiv:2009.09822*, 2020.

[3] Gnu general public license, version 3, June 2007. Last retrieved 2020-01-01.

[4] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *J. Mach. Learn. Res.*, 21(118):1–6, 2020.

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[6] Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[7] Daochen Zha, Kwei-Herng Lai, Kaixiong Zhou, and Xia Hu. Towards similarity-aware time-series classification. In *Proceedings of the 2017 SIAM international conference on data mining*, 2022.

[8] Qingquan Song, Dehua Cheng, Hanning Zhou, Jiyan Yang, Yuandong Tian, and Xia Hu. Towards automated neural interaction discovery for click-through rate prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 945–955, 2020.

[9] Yi-Wei Chen, Qingquan Song, and Xia Hu. Techniques for automated machine learning. *ACM SIGKDD Explorations Newsletter*, 22(2):35–50, 2021.

[10] Yuening Li, Zhengzhang Chen, Daochen Zha, Kaixiong Zhou, Haifeng Jin, Haifeng Chen, and Xia Hu. Automated anomaly detection via curiosity-guided search and self-imitation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[11] Kwei-Herng Lai, Daochen Zha, Guanchu Wang, Junjie Xu, Yue Zhao, Devesh Kumar, Yile Chen, Purav Zumkhawaka, Minyang Wan, Diego Martinez, et al. Tods: An automated time series outlier detection system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16060–16062, 2021.

[12] Daochen Zha, Zaid Pervaiz Bhat, Yi-Wei Chen, Yicheng Wang, Sirui Ding, Jiaben Chen, Kwei-Herng Lai, Anmoll Kumar Jain, Mohammad Qazim Bhat, Na Zou, et al. Autovideo: An automated video action recognition system. *arXiv preprint arXiv:2108.04212*, 2021.

[13] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*, 2019.