Plan Overview

A Data Management Plan created using DMPonline

Title: AmiCI: Adaptive Microservice Configuration and Inference Using Neurosymbolic AI

Creator: Ruben Ruiz Torrubiano

Principal Investigator: Ruben Ruiz Torrubiano

Data Manager: Lukas Madl

Affiliation: Other

Template: DCC Template

ORCID iD: 0000-0001-9314-0739

Project abstract:

Cloud computing and the microservice architecture have become some of the key concepts that underlie the applications and services that surround us on a daily basis, from e-mail clients and social networks to critical applications in energy, infrastructure and healthcare, to name a few. On the one side, cloud computing has enabled the development of highly-scalable applications on the basis of a reliable and cost-effective virtualized infrastructure. On the other side, application developers that work on Software-as-a-Service (SaaS) solutions are faced with a considerable amount of design and deployment choices, which may also become more complicated depending on the constraints introduced by the application at hand. For instance, an application that processes critical and highly sensitive personal data might impose the constraint that the data can only be stored on certain data centers in the jurisdiction of a given country. Additionally, cost constraints might drive the application developers to deploy some parts of the application in other environments, like public clouds from major providers. There might also be configuration choices that need to be done based on customers requirements and previous experience. These factors greatly increase the development and deployment complexity, so that algorithms and tools are needed to assist application developers in this task.

The main goal of this project is to investigate how to solve this type of problems in an unified and adaptive way using machine learning and advanced statistical inference techniques. On the one side, rules-based or symbolic methods would be well suited for assisting in configuring a given project in a clear and reproducible way by following a certain logical path, usually determined by expert knowledge. On the other side, machine learning and data driven methods allow for learning from past projects and making suggestions for new ones. Since each type of method alone would result in suboptimal decisions being made, we aim at combining symbolic and machine learning methods to solve the problem. Our approach can be framed in the broader context of Neurosymbolic AI, a recent research trend that looks for solving the main limitations of machine learning methods by combining them with knowledge-based reasoning. In particular, our approach makes use of ontologies, knowledge graphs and structural causal models (SCM) on the symbolic side, and deep learning methods like transformers and recursive neural networks (RNN) on the data-driven side to learn from past projects. Additionally, combinatorial optimization methods combined with machine learning can be used to find the optimal configuration when an objective function is defined, like costs or energy consumption of the final application.

ID: 118773

Start date: 01-01-2024

End date: 31-12-2026

Last modified: 01-03-2023

AmiCI: Adaptive Microservice Configuration and Inference Using Neurosymbolic AI

Data Collection

What data will you collect or create?

In this project, the following data will be created and collected:

- Network and traffic data (inside of a virtual data center)
- Network and traffic data (between virtual data centers)
- Application usage data (log files)

These types of data might include the following data fields:

- Source IP address
- Destination IP adress
- Source MAC address
- Destination MAC address
- Timestamp

Payloads might include arbitrary application-dependent data.
Network data is typically of high frequency - it is therefore expected to generate large amounts of data.

How will the data be collected or created?

The data for this project will be always collected in a **controlled** environment. That means that no real-usage data from real users will be collected at any time. Data will be stored in virtualized storage services that will be read via REST API.

1. Project management data will be provided by the industry partner (BOC Group) where personal data will be anonymized accordingly.
2. Performance and network data will be collected in a cloud environment in a controlled way.

The data will be stored in a central shared storage using the following naming schema: Data\<project ID>\project management data for 1) and Data\<project ID>\performance data\<service ID>\<timestamp begin execution> for 2).
Using this schema, it will be possible to guarantee the consistency of the data and ensure that the conditions of the experiments can be repeated.

Documentation and Metadata

What documentation and metadata will accompany the data?

For each configuration project, it will be documented which services run, how to run the applications with a given user profile and how to collect the resulting data (e.g. which API calls are necessary to collect the data). Additionally, metadata about each project will be collected (name, description, type of project).

Ethics and Legal Compliance

How will you manage any ethical issues?

Ethical issues might arise from the following factors:

- Personal information will be present in the project management data. Identifiers and names will be anonymized before being used in the project. Once that data is effectively anonymized, it can be used by the project's participants. In case that stronger

methods are needed, it will be resorted to differential privacy for more guarantees.
- Any issue regarding bias/discrimination resulting from the model's decision will be documented in detail. One of the project's goals is to improve the explainability of the results, therefore we expect to detect such issues by design.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

All data will be generated by test applications to be also developed during the project. Personal data will be anonymized as detailed before. The data will be therefore owned by the project participants and there are no restrictions on third-party data. The data obtained in this way can be safely shared with other researchers to increase the reproducibility and robustness of the results.

Storage and Backup

How will the data be stored and backed up during the research?

Data will be stored in a separated shared cloud drive. Backups are automatically performed and data can be recovered in case of an incident. In case there is no sufficient storage, the amount of storage can be increased by the IT department on demand.
Regular data review iterations will ensure that only the data that is necessary to perform experiments is kept - All the other data will be deleted, so that the amount of storage used does not increase unnecessarily.

How will you manage access and security?

Data will be shared via cloud drives with the rest of the project team. Since all collaborators agree to use the same data storage platform, which is supported by the participant institutions, existing organizational credentials can be used to access the data in a secure way. If the data is deemed to be of particular strategic importance for the research project, it will be considered to add a two-factor authentication layer.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

IIn general, the data will be used for building realistic networking scenarios and training machine learning models with the generated data. Therefore, in order to reproduce the performance and the predictions of a given model, the same data need to be used for training. These data will be kept accesible for efficient access during the research project. After the research project has ended, data will be archived by compressing it and storing it securely in the institutional cloud storage environment.

What is the long-term preservation plan for the dataset?

After the research project, data will be archived and stored securely in the institutional cloud environment. Fragments of data used for reproducing published results will also be stored in an open data repository, so that the scientific community can reproduce and inspect the findings of the project.
Before archiving the data, metadata will be included in a documentation system to ensure that future uses of the data are done in an appropriate way.

Data Sharing

How will you share the data?

Data necessary to reproduce the findings of published results will be stored in a public repository that is well-known in the scientific community, assigns a persistent identifier (DOI) and can be searched in an easy way. Additionally, researchers that ask for access to

parts of the data that can not be kept in the open access repository (because they are owned by the industry partner) will be granted access on demand after reviewing the request with BOC Group.

Are any restrictions on data sharing required?

The only restrictions to be considered when sharing and using the data involve citing the authors of this research project and the relevant publication(s) in an appropriate way. In particular cases, like project management data, these data might be subject to formal agreement.

Responsibilities and Resources

Who will be responsible for data management?

Due to the nature of the research project, the exploitation partner Lukas Madl, will assume the responsibility for implementing the data management plan and monitoring its implementation. Data management activities will thus be kept central and the individual project contributors will have to adhere with the previously defined rules regarding naming and curation. All partners of the consortium agree on this matter.

What resources will you require to deliver your plan?

The needed resources for long-term storage are given by the IT infrastructure of the lead institution, IMC Krems. Short-term data storage will be performed in the corresponding virtual data centers that are needed as research infrastructure and provided by our cloud provider. Charges due to this short-term storage are already accounted for in the project budget. No additional expertise will be required, since all research members of the project team have a strong data management and analysis background.