

# Classification of Sensor Time Series

## Pilot Study Milestone

Luciano Melodia, Christian Sauerhammer, Richard Lenz

Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg

Summer semester 2021



## Our Agenda for Today

This is our agenda for this meeting:

- Project overview.

- Current state of the project.

- Requirements that have arisen.

- Proposal for further course.

- Presentation of the classification tool.

- Experiments to be conducted.

- Other.

## Project Overview

Year	1. Year				2. Year				3. Year			
Quartal	1	2	3	4	1	2	3	4	1	2	3	4
Process analysis												
Literature analysis												
Market analysis												
Pilot study												
Evaluation												
Conception of the tool												
Prototype implementation												
Documentation												
Milestones	M1		M2/M3		M4	M5	M6		M7		M8	

# Current State of the Project



## Current State of the Project (I)

### Findings:

A method for interpolation of arbitrary multivariate time series.

A corresponding stop criterion when the interpolation should be stopped.

Hypothesis test for deciding the quality of the interpolation.

 <https://github.com/karhunenloeve/SIML>

 <https://arxiv.org/abs/1911.02922>

**Abstract:** In this study the Voronoi interpolation is used to interpolate a set of points drawn from a topological space with higher homology groups on its filtration. The technique is based on Voronoi tessellation, which induces a natural dual map to the Delaunay triangulation. Advantage is taken from this fact calculating the persistent homology on it after each iteration to capture the changing topology of the data. The boundary points are identified as critical. The Bottleneck and Wasserstein distance serve as a measure of quality between the original point set and the interpolation. If the norm of two distances exceeds a heuristically determined threshold, the algorithm terminates. We give the theoretical basis for this approach and justify its validity with numerical experiments.

## Current State of the Project (II)

### Findings:

A method to determine the width of neural networks.

Theoretical bridge to an exact solution for a special case.

Special case coincides with periodic or quasiperiodic time series.

 <https://github.com/karhunenloeve/NTOPL>

 <https://arxiv.org/abs/2004.02881>

**Abstract:** In this paper we present an approach to determine the smallest possible number of neurons in a layer of a neural network in such a way that the topology of the input space can be learned sufficiently well. We introduce a general procedure based on persistent homology to investigate topological invariants of the manifold on which we suspect the data set. We specify the required dimensions precisely, assuming that there is a smooth manifold on or near which the data are located. Furthermore, we require that this space is connected and has a commutative group structure in the mathematical sense. We use the representatives of the  $k$ -dimensional homology groups from the persistence landscape to determine an integer dimension for this decomposition.

## Current State of the Project (III)

### Summary:

We learned how to augment existing data for the problem of not having enough for some sensor readings. The natural neighbors method has been shown to work for arbitrary time series. We also found a solution for when to stop interpolation so as not to significantly change the original distribution of the data, or to do so, if one wishes.

We have built a theoretical bridge to topological data analytics and have shown that we can estimate the embedding dimension of neural networks for time series very well. This helps us to achieve a lower number of parameters in training without sacrificing the quality of the classifier.

# Presentation of the Classification Tool

by Christian Sauerhammer





# Experiments to be conducted



## Experiments to be conducted

Long Short-Term Memory (LSTM) networks still state of the art for time series.

Combination of theory of  $\mathcal{C}^k$ -differentiable neural architectures with LSTM, this reduces drastically parameter size.

How low can we go in parameters?

What is an appropriate  $k$ ?

How can we determine this by our previous work?

**Reward:** Time, energy and thus cost savings for training.

 <https://github.com/karhunenloeve/TwirlFlake>

# Requirements



## Hardware

**Hardware:** Two NVIDIA Quadro RTX 4000 with 8 GB of GDDR6, 36 computation units, 288 NVIDIA-Tensor units and 2.304 parallel CUDA computation units. **Price:**  $\approx$  900€.

A better option would be: Single NVIDIA Quadro RTX 8000 with 48 GB of GDDR6, 72 computation units, 576 NVIDIA-Tensor units and 4.608 parallel CUDA computation units. Less latency than many separated RTX 4000 graphic cards. **Price:**  $\approx$  5.600€.

**Time constraints:** After augmentation we have a dataset of about 7.2GB with more then 300.000 single files. We optimized the current graphic cards together with **Ubuntu 20.04**, **CUDA 11.0**, **Cudnn 8.4** and **Tensorflow 2.4**. This software is mandatory, no alternatives. Software is not in newest version, but in this configuration compatible. Neural networks have been reduced to about  $10^7$ -Parameters. Currently one neural network run takes about 7 **days**.

🔗 Installation guide: <https://gist.github.com/karhunenloeve/223dcc4a193b7fd33669f9e4326d289b>

## Requirements

**Problem:** We have almost completely exhausted our current hardware. In total, we compare 23 neural architectures with each other. Assuming an average of 5 days per run and ten runs per architecture to give an expected value for the results, we need about 575 days of pure computation time.

**Hardware:** We need about 18 more NVIDIA Quadro RTX 4000 for a total pure computation time of 29 days to conduct the experiments.

Currently Cuda is the most powerful language for parallel computations on graphic cards and requires **NVIDIA** graphic cards.

### Questions:

Is computational power available at Siemens AG?

Could you set up an Amazon AWS Account for this task?

Should we set up a small computation cluster ourselves?

## Costs for Amazons Service

Name	vCPU	ECU	Memory	Inst. Storage	Costs
p3.2xlarge	8	31	61 GiB	EBS Only	\$3.823 per h
p3.8xlarge	32	97	244 GiB	EBS Only	\$15.292 per h
p3.16xlarge	64	201	488 GiB	EBS Only	\$30.584 per h
p2.xlarge	4	16	61 GiB	EBS Only	\$1.326 per h
p2.8xlarge	32	97	488 GiB	EBS Only	\$10.608 per h
p2.16xlarge	64	201	732 GiB	EBS Only	\$21.216 per h
g3.4xlarge	16	58	122 GiB	EBS Only	\$1.425 per h
g3.8xlarge	32	97	244 GiB	EBS Only	\$2.85 per h
g3.16xlarge	64	201	488 GiB	EBS Only	\$5.70 per h

**Graphic cards:** in p2: NVIDIA Tesla K80, in p3: NVIDIA Tesla V100.

**CPU:** Intel Xeon E5.

**Source:** <https://aws.amazon.com/ec2/pricing/on-demand/>

## Other Options

**Google Cloud:** <https://cloud.google.com/products/ai/>

Hard to configure and the total bill is tough to estimate.

Easy to use interface for setup.

\$300 initial credit.

**AWS EC2:** <https://aws.amazon.com/ec2/>

Hard to configure and the total bill is tough to estimate.

Hard to use interface for setup.

Scalable, secure and reliable.

**Paperspace:** <https://www.paperspace.com/>

A bit more pricy than Google Cloud or AWS.

Takes security seriously.

Uses own servers and also third party servers.

# Proposal for further Course





## Changes in the project

### **I will leave the project and the chair as of 07/31.**

My contract with the Department of Computer Science 6 expires on this date. I have declined a contract extension. I will go to Augsburg to study theoretical mathematics. However, I will continue to supervise graduation theses and will be happy to provide support in their context.

### **Christian Sauerhammer is vacant as successor.**

Mr. Sauerhammer is currently writing his master thesis under my supervision and is comparing several neural network architectures for the classification of sensor signals. Furthermore, he has also designed a graphical user interface which is a prototype that can be used for the classification of sensor time series.

## Hurdles of the Past

### We should consider these hurdles for the project:

Constraints over one year of available resources due to CORONA pandemic.

Delay of the evaluation of the study due to necessary hardware procurement.

Delay due to change of project staff.

**Summary:** We are currently on schedule, the design of the study has been completed. However, at this point I recommend the expansion of this phase of the project. Due to higher influences and unforeseen circumstances, we had to work with the funds available to us. Nevertheless, the results so far are quite promising. The classification of time series data is by no means treated research terrain. We have already expanded the state of the art and have the opportunity to achieve unprecedented accuracy.

**I therefore recommend that the *M4/M5* project phase be extended to 6 quarters.**

## Problem: Schema Inference

### **Schema inference seems to be infeasible by the state of the art. Why is that?**

No theory if there are invariants characterizing a valid schema.

Invariants are numbers, that do not change if we exchange a valid schema by another one.

Simplicial databases are most promising approach.

We need to establish suitable invariants.

We need to establish algorithms to compute them efficiently.

We need to set up benchmark datasets.

### **We identify two strictly different tasks:**

Find invariants for samples from one sensor as a fingerprint.

This will guarantee an unsupervised way to find other related sources.

Investigating dependencies of invariants one can establish dependencies of signals.

Connect this information to create a suitable database schema. How do invariants from data relate to invariants from database schemas? Are they the same?

Other



## Other

**Student:** Noah Becker, fifth semester Computer Science Bachelor.

**Employment:** Currently for 5 hours a week at our chair and 15 hours at Siemens Energy AG.

**Mr. Becker's duties to date have been:**

- Set up and design of the front end for the classification application.

- Implementation of the interfaces for file upload.


- Implementation of a small library for file format conversion.

Can we employ Mr. Becker at Siemens and make him responsible for our project for half of his working time?

Mr. Becker's department has already given its consent to the project.

Your agreement was also obtained at the penultimate milestone meeting.

Thank you for your attention.  
**Any questions?**

Drop me a line:  
 [luciano.melodia@fau.de](mailto:luciano.melodia@fau.de).