**Research topic**: Proactive microservice workload predictions using deep learning

**Coordinator**: Dr. Virginia Niculescu

**Abstract:** Microservice architectures are becoming increasingly popular due to their flexibility and scalability which makes them a good fit for deployment in the cloud. In order to fully take advantage of on-demand resources and reduce manual tuning an auto-scaling approach is required, which means increasing or decreasing the number of deployed services according to incoming workloads.

This research project aims to provide a performant model for workload prediction using deep learning algorithms and offer a comparison with other classic and machine learning approaches from recent research. Furthermore the selection of the predicted variable and parameter values are optimised for the specific problem of scaling microservices in a production environment.

Some microservice deployments which are already used in the industry (e.g. Netflix) can take minutes to start-up. This means that an efficient auto-scaler should be able to make predictions of when the next spike is likely to appear. Therefore an efficient prediction window should be >=5 minutes to account for service startup time. The predicted variable should not be influenced by the prediction like for example: response time, which will be shorter when we scale up. Thus, the predicted value for this experiment is : user requests per time interval. This can be later used by an auto-scaler to calculate necessary instances of a service.