# Project
# PORTFOLIO

Emilia Marchese

Personalized Digital Health: Depression and Suicide Risk Detection form Internet Browsing Traces

Multi-modal Question Answering on Electronic Health Records with Chest X-ray Images

Project in Reinforcement Learning: MountainCar & LunarLander

An efficient Google Maps Scraper – AWS Cloud

Explainable AI for Natural Language Processing: Feedback Prize – English Language Learning Kaggle Competition

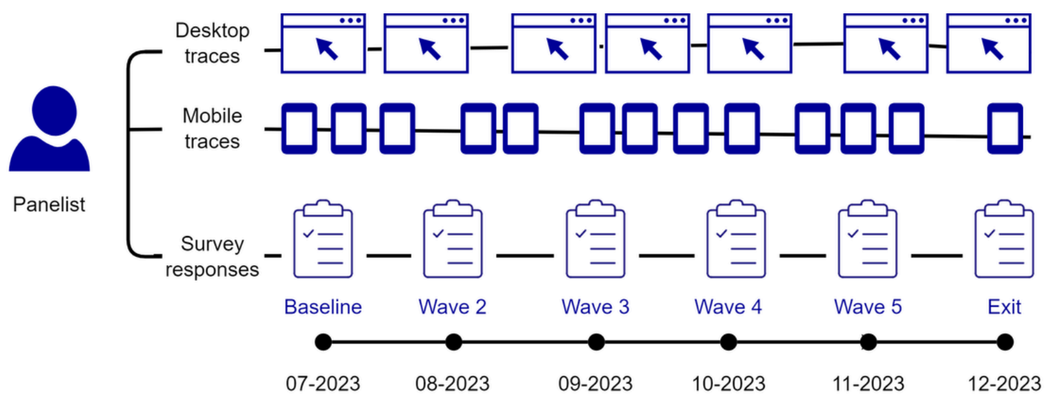
Bayesian Data Analysis: Modelling the Ebola Epidemic

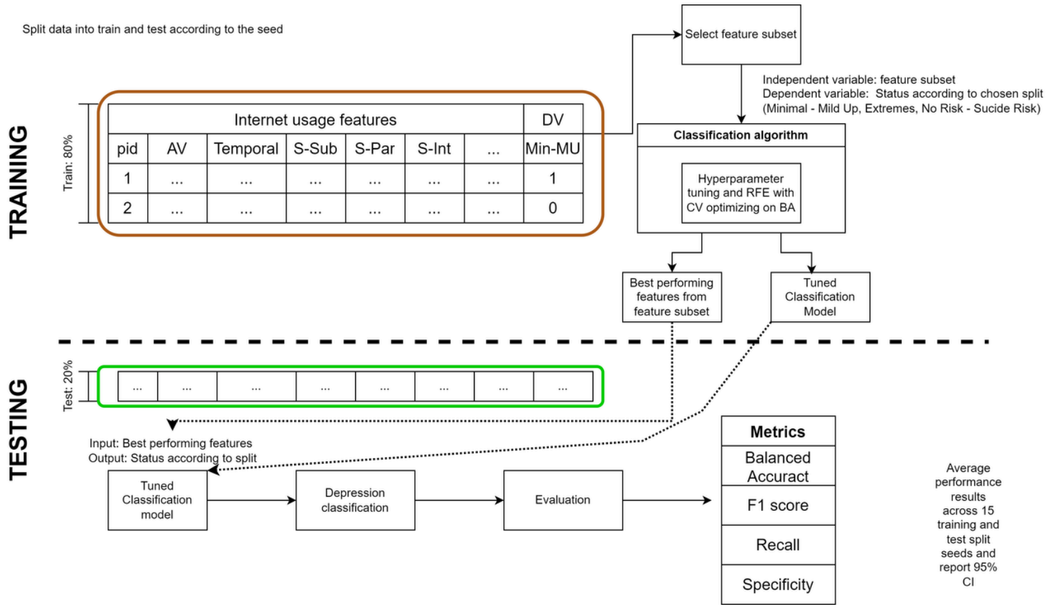
OSM Changes Visualizer - Observing Infrastructure Changes Over Time

An analysis on the effect of the COVID-19 pandemic on the housing demand in Finland

# Personalized Digital Health: Depression and Suicide Risk Detection form Internet Browsing Traces

The aim of the project is to gain **actionable insights** and identify internet usage behaviours that can be useful in the creation of **digitial health technologies** for depression prevention. For this project, I **pre-processed** 200GB of mobile and desktop traces from 900 German individuals, and performed **feature selection, model optimization and model selection** for depression classification.




To get a comprehensive overview on the potential of internet usage traces for derpession classification, I trained five ML models (XGBoost, RF, RL, SVM-RBF) across several splitting seeds. **The training was parallelized across seeds and models on the high performance computing (HPC) Aalto Triton cluster**, and the best performing internet usage features were found using recursive features elimination.
To observe longitudinal effects, I used **hierarchical mixed models** to find associations between internet usage and depression.

## SKILLS

Pre-processing
Fine-tuning
Model selection
XAI
HPC (Slurm)
Hierarchical models

## PRESENTATION

https://drive.google.com/file/d/1ASnznt2-R07_Ocv7cTnSx8MkKqAToPpY/view?usp=sharing

## THESIS

https://urn.fi/URN:NBN:fi:aalto-202403172759

## Department

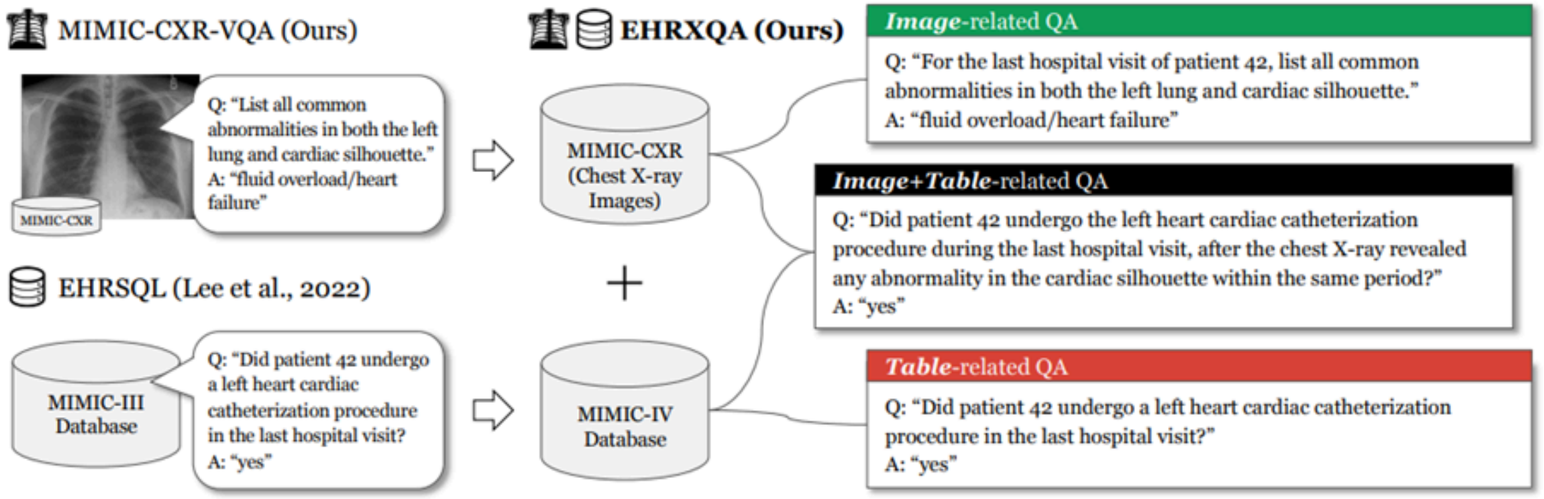
Department of CS, Aalto University, CSS group

# Multi-modal Question Answering on Electronic Health Records with Chest X-ray Images

## SKILLS

PyTorch
LLM fine-tuning
HPC
CV
HuggingFace API

Electronic Health Records (EHRs) contain a wealth of patient information across multiple modalities, including structured records, images, and clinical text. However, current EHR question answering (QA) systems focus on a single modality, overlooking the multi-modal nature of EHRs. To fully utilize EHR data and support clinical decision-making, it is crucial to develop versatile QA systems that can navigate across multiple modalities. This project aimed to create a **multi-modal QA system** that incorporates both **structured records** (MIMIC-IV) and **chest X-ray** images (MIMIC-CXR-JPG). This system needed to be able to answer questions spanning three modalities: *Image-related*, *Table-related*, and *Image+Table*-related. To query the electronic health records, **my teammate and I fine-tuned a Codes-1b LLM model on text-to-SQL tasks** on the MIMIC-IV database. We added  a Computer Vision function that calls upon **BiomedCLIP-PubMedBERT** to perform the **VQA** (visual question answering) task on the chest x-ray images. Our implementation achieves very good performance across all three modalities.




## REPOSITORY

https://github.com/emimarch/ml-for-healthcare-2

## COURSE

AI612 Machine Learning for Healthcare, Korean Advanced Institute of Science and Technology

# An efficient Google Maps Scraper – AWS Cloud

As part of my research assistant role in mobility at the Department of Economics, I migrated a Google Maps web-scraping infrastructure on AWS from **EC2 computing to EC2 spot-instances**, with the aim to reduce costs. Spot-instances are EC2 spare resources whose price depends on demand, and are often significantly cheaper. The caveat is that the resource might be withdrawn at any point if the demand increases. To mitigate the risk, **I automated the storage of the scrapers and the launch of new scrapers** whenever a scraper was shut down, either because of detection from Google Maps or due to the spot-instance being reserved. I did so using **EventBridge** events and **Lambda** functions.

## SKILLS

AWS
IAMs
EC2
S3
Lambda
EventBridge
Web-scraping

## DEPARTMENT

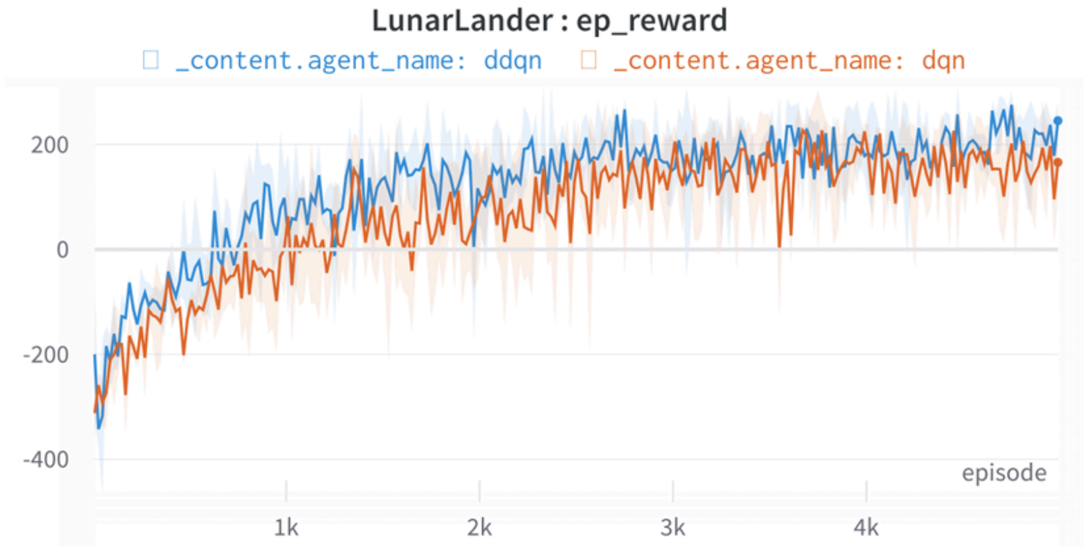
**Department of Economics,**
**Aalto University,**
**Urban Mobility**

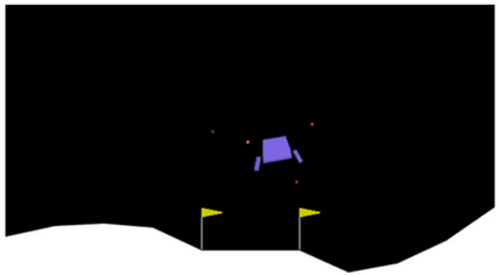
## SUPERVISOR

**Prof. Prottoy Akbar**

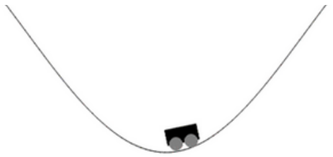
# Project in Reinforcement Learning: MountainCar & LunarLander

The aim of the project is to **tune hyperparameters of reinforcement learning algorithms** and present a performance comparison of algorithms for various **OpenAI Gym environments** with different degrees of complexity. The course provides an overview of mathematical models and algorithms behind optimal decision making i**n time-series systems,** with a focus on optimal decision making and control, reinforcement learning, and **decision making under uncertainty.** In this project, my teammate and I implemented four different RL algorithms on two environments, trained our agents across several seeds and compared the performance of the chosen algorithms to select the best performing agent for each environment. Experiments were monitored using wandb.ai.



(b) LunarLander



(a) MountainCarContinuous

## SKILLS

PyTorch
RL algorithms
Experiment tracking
wandb.ai

## REPOSITORY

https://github.com/emimarch/Reinforcement-Learning-Final-Project

## REPORT

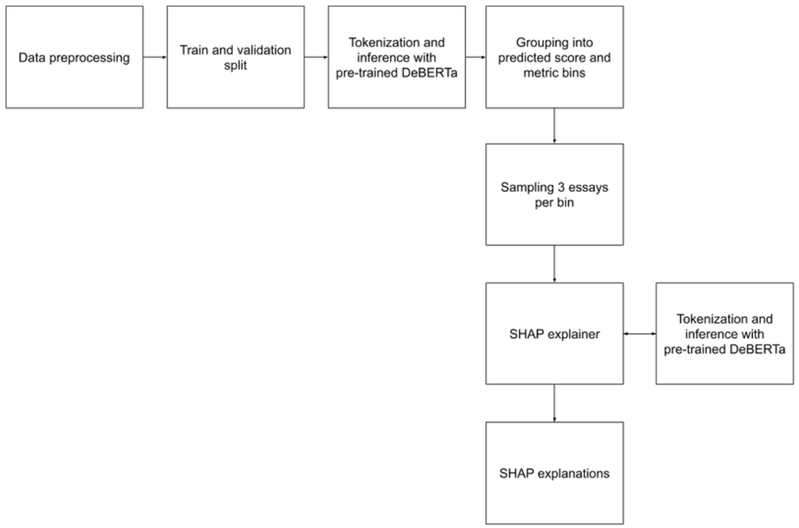
**https://github.com/emimarch/Reinforcement-Learning-Final-Project/blob/main/Report_anon_redacted.pdf**

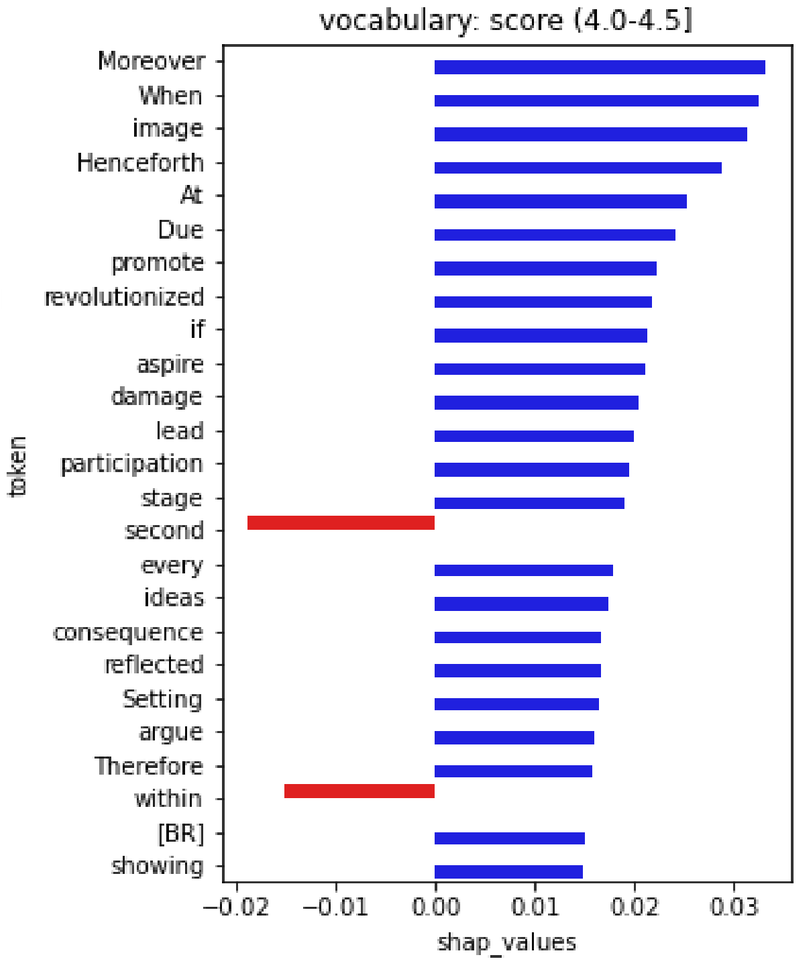
## COURSE

ELEC-E8125- Reinforcement Learning, Aalto University

# Explainable AI for Natural Language Processing: Feedback Prize- English Language Learning Kaggle Competition

The project's objective is to use **Explainable Artificial Intelligence** frameworks to interpret the output of an **NLP prediction task** based on the *Feedback Prize-English Language Learning* Kaggle competition. The competition texts "assesses the language proficiency of 8th-12th grade English Language Learners (ELLs)" in six different categories: *cohesion*, *syntax*, *vocabulary*, *phraseology*, *grammar*, *and conventions*. The aim is to analyze the motivations behind the **black-box NLP model** predictions for some of these categories to assess if they are sensible and if the explanations follow a logic that is akin to a human's. We use a fine-tuned *Deberta-v3-small* LLM model from the third winner of the Kaggle competition as our black-box model, and use the SHAP explainer for the explanations. We find that the results form the explainer are reasonable.




## SKILLS

SNLP
PyTorch
Tokenizers
Transformers
XAI

## REPORT

**https://github.com/emimarch/SNLP-XAI-Project/blob/main/NLP-project-report.pdf**

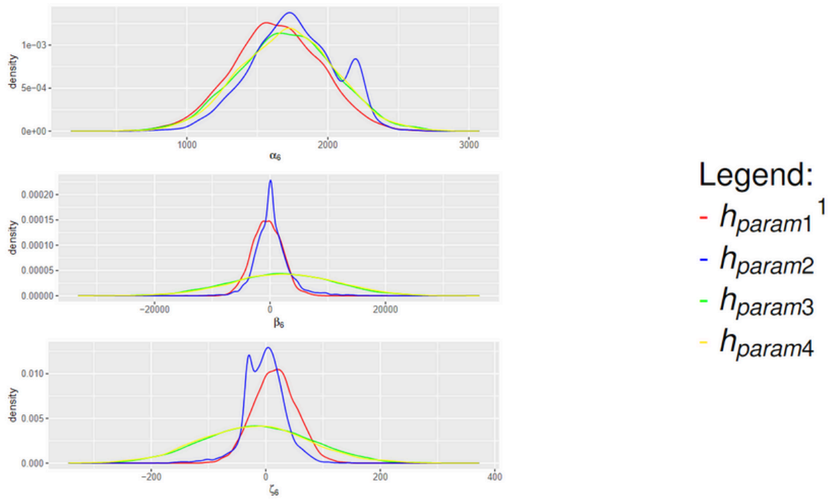
## COURSE

ELEC-E5550 Statistical Natural Language Processing, Aalto University

# Bayesian Data Analysis: Modelling the Ebola Epidemic

## Sensitivity Analyses: Hierarchical model

Case Study: Sierra Leone's model parameters




The project aims to apply the Bayesian data analysis workflow on a real-world scenario. To do so, my teammates and I **decided to model the number of ebola cases** in various countries in terms of the population density and the number of months passed from the start of the pandemic. The Bayesian **pooled priors** and **hierarchical priors** approaches are compared. A **convergence analysis for the MCMC** is conducted to ensure the models are satisfactory, and **model comparisons**, **posterior predictive checks** and **prior sensitivity analyses** are performed to select the best model.

## SKILLS

Stan
Bayesian Workflow
Priors
Hyper-priors
Posteriors

## REPOSITORY

https://github.com/emimarch/BDA-Project

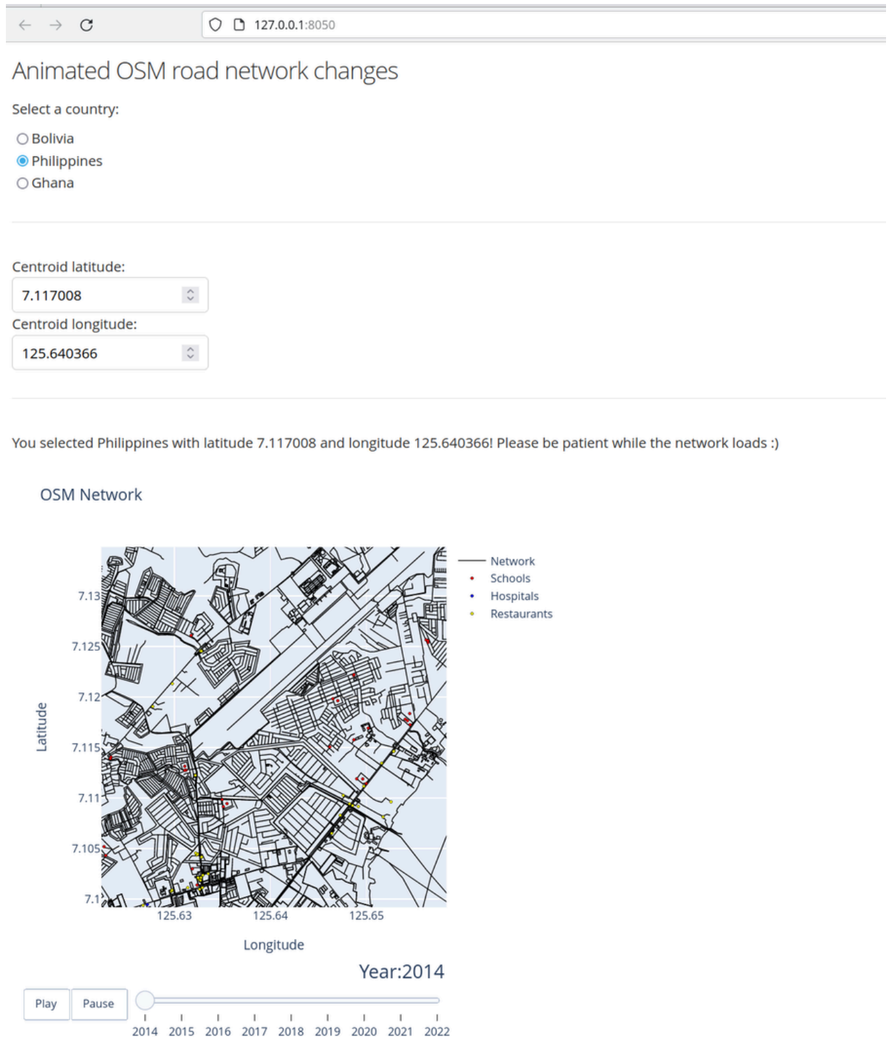
## REPORT

https://github.com/emimarch/BDA-Project/blob/master/Project.pdf

## COURSE

CS-E5710 Bayesian Data Analysis, Aalto University

# OSM-Change Visualizer: Observing Infrastructure Changes Over Time




An **OpenStreetMap** network changes visualizer as a **Dash** webapp. The webapp allows you to select a country and a centroid. After selecting your point of interest, you can observe how the infrastructure and amenities have changed in the area through time. The map is interactive.

## SKILLS

Dash
Plotly
OpenStreetMap
GIS
GeoPandas

## REPOSITORY

https://github.com/emimarch/OSM-Changes-Visualizer

## REPORT

https://github.com/emimarch/OSM-Changes-Visualizer/blob/master/EIV___report.pdf

## COURSE

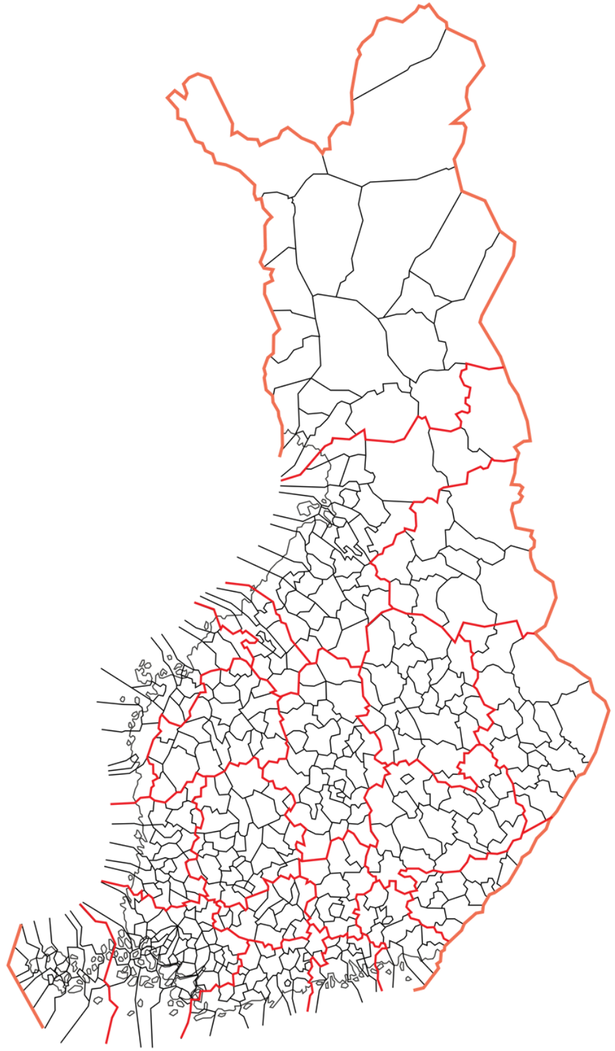
CS-E4450 - Exploratory Information Visualization, Aalto University

# An analysis on the effect of the COVID-19 pandemic on the housing demand in Finland

**SKILLS**

**Econometrics**
**GLM models**

This empirical work studies the effect of municipality characteristics on the demand for housing in Finland during the COVID-19 pandemic. It analyses how home prices, rents, the number of sales and the number of tenancy agreements have been affected by the pre-pandemic density, house and rent prices, households' average earned income, degree of urbanization and percentage of commuters. The purpose is to **elaborate on whether some municipality characteristics have had a negative or positive effect on housing demand** in order to infer whether this could possibly result in the settlement of a new **spatial equilibrium.**



## Presentation

https://drive.google.com/file/d/18X9dUY4JfHRBepjoEwFNB7kPq8k7qD0s/view?usp=sharing

## Thesis

https://urn.fi/URN:NBN:fi:aalto-202209115427