Topic #	P01
Domain	ML on Tabular Data
Title	**P01 - Impact of covid 19 on energy consumption in Baltic countries**
Description	We would like to find the impact of covid 19 on energy consumption for Baltic countries. The idea is to find how the covid 19 measurements such as lock down, shut down of schools, universities, commercial places etc has impacted the energy consumption. There is data available for all 3 countries from 2018 to 2022. Few methods such as data analysis, visualisation, stat models and time series analysis could be done to get proper results.
Is data available	Data is already available
Contact person	Neha Sharma (**is ready** to mentor teams)
Organization	University of Tartu
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Low (**Dima**: this project is probably a bit more about data exploration, so one would have to figure out ways to apply ML here)

Topic #	P02
Domain	Medical Imaging
Title	**P02 - Tumor Growth Monitoring from Xenograft Cancer Models using MRI and Deep Learning**
Description	The proposed project aims to develop an automated deep learning solution for monitoring tumor growth in patient-derived xenograft cancer models through serial non-contrast non-gated T2w MRI analysis. The student(s) working on this project will apply advanced deep learning techniques, such as 3D Convolutional Neural Networks (CNNs), for tumor segmentation and quantitative analysis.The specific goals of the project include achieving a high level of accuracy in tumor segmentation, benchmarking against existing methods, and processing a substantial volume of MRI data for robust results. Furthermore, our solution will be designed with scalability in mind to handle larger datasets or different types of MRI scans effectively. The automation introduced by this project promises to save significant time and reduce human error, ensuring more reliable tumor growth pattern analysis. The dataset for this project can be downloaded from here (https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=145752799). Successful completion holds the potential for impactful contributions to oncology research and precision medicine, with findings worthy of publication in relevant scientific journals.
Is data available	Data is already available
Contact person	Vijayachitra Modhukur (**is ready** to mentor teams)
Organization	University of Tartu
How many teams they are ready to supervise?	**There is a room for 2-3 teams**
Approx. Complexity	**Hard**

Topic #	P03
Domain	Satellite Imaging
Title	**P03 - Intersection finder**
Description	"We are building an offline military navigation app, that has the whole map of Estonia stored inside. We use 1000 x 1000 px map images, where 1 px = 1 m, we have around 50000 images. The problem: for military communication reasons, we need to map the coordinates of all the major road intersections in all the maps. The goal of the project: to develop a machine learning model, that takes in all the 50000 images an returns an array of intersection coordinates for each map image. It has to be adjustable to capture only main road intersections. An accuracy over 75% will already be of practical value for the app. Misclassified intersection points will not be a critical error in the app, just less useful points. The methods to use: Computer vision algorithms, like Yolo. P.S. I am also a student in the Machine Learning course and would like to lead the team."
Is data available	Data is already available
Contact person	Erkki Tikk (**student in the course**)
Organization	Asymmetric Systems OÜ
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium/High

Topic #	P04
Domain	ML on Tabular Data
Title	**P04 - Towards automating data quality specification**
Description	Data quality management, although is not new, but still very relevant topic that becomes even more relevant with the increase of the amount and variety of the data. However, data quality is a very multi-faceted topic, where a proper management of the above is a complicated task, including but not limited due to the need for domain knowledge with the reference to both the data and topic of data quality. For the latter, this is due to the concepts of data quality dimensions, rules and metrics, which make sit complicated to the end-user without respective DQ knowledge to conduct a DQ analysis. Hence, attempts towards automating data quality specification become popular. This is all the more needed in the light of wide popularity of third-party data such as open data - data that were generated / collected and processed by entity other than data user. Current approaches can be divided into rule-based, metadata-driven and ML-based, where the latter is the most promising but the least represented in both academia and practice. Hence, the objective of this project would be to propose such a ad-hoc approach towards automating DQ specification by means of extracting data quality requirements from data features, which would be expected to be done employing ML (composite/combined approach with the use of predefined rules or using metadata is welcome and can appear to be the most promising). This would mean that DQ rules would be extracted from the analysing the data (attributes / columns for tabular data) and determining the patterns such as email (if @domain is determined), date, post codes or any others, as well as consistency rule extraction (e.g., attribute names or metadata states that the dataset contains start and end date, so the comparison of the respective values, so that end date is after the start date would make sense). The requirements for the above can be partly retrieved from several existing DQ analysis tools supporting ML-based rule/check definition.
Is data available	Data is already available
Contact person	Anastsija Nikiforova (**is ready** to mentor teams)
Organization	University of Tartu
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium

Topic #	P05
Domain	ML on Tabular Data
Title	**P05 - Automated classification of open datasets to improve data findability on open government data portals**
Description	While many open government data (OGD) portals provide a large number of open datasets that are free to use and transform into value, not all of these data are actually used. In some cases, this is because these data are difficult to find due to the low level of detail presented in them, including, but not limited to the absence or inaccuracy of the category(-ies) and tags assigned to a particular dataset, which is a part of the data publisher task. In the case of some OGD portals, 1/3 of the datasets are not categorized, although the portal provides a rich list of data categories that are in line with best practices and allow to classify these datasets. This leads to cases where the dataset cannot be found if the user searches for data using catalog or tags (only using the search bar will return the dataset, if the search query matches the title or description of the provided dataset). This project is intended to propose an automated data classification mechanism, which, based on a dataset and the data provided on it (title, description of the dataset (! please, take into account that you will be asked to carry out at least a simplified text analytics), parameters of the dataset (if sufficiently expressive)), will suggest a categories and tags to be assigned to it. First, you will be expected to explore OGD portals and how datasets look like, and what can be scenarios for OGD user to search for a particular dataset. Then, a list of indicators will be defined, which should constitute the input for data classification (mostly in line with the above but can be enriched, if possible), and an appropriate solution will be developed. This would contribute to the FAIRness of the open data, although mainly referring to F – findability, but indirectly affecting other features that the OGD should meet in order to provide social, economic and technological benefits from individual users, SMEs and governments.
Is data available	Data is available online, but it needs to be fetched or scrapped
Contact person	Anastasija Nikiforova (**is ready** to mentor teams)
Organization	University of Tartu
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium/Hard

Topic #	P06
Domain	Time Series Analysis
Title	**P06 - Exploring use cases of different forecasting methods**
Description	We'd like to investigate which forecasting methods work well across the revenue of some of our products. The goal is to find which methods work and in what scenarios and if not, why (might be a data problem). We're not after a perfect forecasting model, though that would be great, but a series of recommendations. There's a lot of discovery and with this so we're open to any forecasting methods the students wish to use. But it would be good to start with the basics e.g. linear regression, ARIMA, Holt-Winter, Kalman filters etc.
Is data available	Data is available and will be provided.
Contact person	Chak Leung (ready to mentor students)
Organization	Adaptavist
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Hard -> this project is about time series which we do not talk too much in the course

Topic #	P07
Domain	ML on Tabular Data
Title	**P07 - Analyzing student activity in the Computer Programming course**
Description	The dataset consists of the numbers of attempts made by 467 students in 16 weekly quizzes and 13 homeworks in the introductory computer programming course. The goals are: 1) Predict the final score or grade, or at least identify the students who might be struggling, based on their early activity. 2) Classify the students based on their typical study patterns.
Is data available	Data is already available
Contact person	Reimo Palm (ready to mentor students)
Organization	University of Tartu
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium/Hard

Topic #	P08
Domain	ML on Tabular Data
Title	**P08 - Pattern recognition for quantification in chemical analysis**
Description	Problem: calibration graphs are used everywhere where chemical analysis is needed from clinical studies to metallurgy. Most analyses should follow linear models but signal saturation, high background noise, human errors, instrument malfunctioning, interferences, etc. cause deviations from linearity. In worse cases multiple of these happen at the same time causing patterns that are automatically hard to detect and if overlooked can cause severe mistakes in quantification, possibly leading to wrong dosage, damaged items in manufacturing, or inconclusive forensics analysis. To remove the points that do not follow linearity today expert knowledge is required, which is time-consuming, not scalable, and often not reproducible. Goal: train an ML model for pattern recognition in calibration data with 6 to 12 calibration points. So that this pattern could be used to automatically exclude points that are not in the linear range. Methods: pattern recognition or potentially advanced optimization.
Is data available	Data is available (n=1000) but additional labelling might be needed to some extent
Contact person	Anneli Kruve (ready to provide substantial mentorship)
Organization	Stockholm University
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Hard

Topic #	P09
Domain	Sentiment Analysis (NLP)
Title	**P09 - Generating Feature-Level Sentiment Summaries from App Reviews**
Description	Given a set of user reviews of a mobile application, the goal is to generate a sentiment summary at the level of app features. It would be good if students investigate the performance of pre-trained models like BERT and RoBERTa for the tasks of extracting app features and sentiment polarity detection.
Is data available	Data is already available
Contact person	Faiz Ali Shah (ready to provide some mentorship)
Organization	University of Tartu
How many teams they are ready to supervise?	**Only 1 team can participate (first come first serve basis)**
Approx. Complexity	Hard -> we are not really talking about NLP in the course

Topic #	P10
Domain	ML for Tabular Data
Title	**P10 - Using Machine learning to estimate river peak flows in Estonia**
Description	We want to use machine learning to improve the current engineering formulas in estimating peak flows in smaller river catchments in Estonia. These formulas use various catchment characteristics (land use, slope etc,) which also requires some working with GIS libraries and functions. These peak flows are used for hydraulically sizing culverts as well as other water conveyance sytems like ditches and even smaller bridges. The idea is to revise the current formulas to provide a better designing basis, but still easy to employ, for engineers in the relevant field.
Is data available	I will provide all the required data, online and offline
Contact person	Ottar Tamm (ready to provide considerable mentorship)
Organization	Estonian University of Life Sciences
How many teams they are ready to supervise?	**Only 1 team can participate**
Approx. Complexity	Low/Medium