Topic #	P01
Domain	Computer Vision
Title	**P01 - Rexplorer solar energy ML development**
Description	Rexplorer is helping to automate solar power planning on buildings to improve the energy efficiency of our homes, offices and other real-estate. The goal of the AI development project is to further develop our Gitlab library AI models used to detect installed PV & roof obstacles in solar project planning. The library currently has 2 image recognition models, set of jupyter notebook that takes as input a geojson Polygon or Multipolygon features of areas of interest. The pipeline is built so that it is run locally but all source files, running scripts and results are stored at separated protected server and could be accessed by SSH. The task for students is to train the models by annotation of orthomosaic images highlighting feature borders and detected objects, measuring the test dataset IoU (intersection over union), polishing the scripts and scaling the model to new areas.
Is data available	Data is already available
Contact person	Hendrik Hundt
Organization	Rexplorer OÜ (https://www.rexplorer.ee/)
How many teams they are ready to supervise?	**Only 2-3 teams can participate (1/3)** **Potentially a possibility to get a prize for the best solution.**
Instructors' comment	This project seems to involve some manual data labelling, which is often hard and time consuming.

Topic #	P02
Domain	ML on Tabular Data
Title	**P02 - Honey Fraud Detection through DNA-Based Taxonomical Composition Analysis**
Description	The project task is to develop a model that can distinguish between authentic and non-authentic honeys using Kraken output data files (https://ccb.jhu.edu/software/kraken/). Based on the data, a model that can reliably classify honey as either authentic or non-authentic should be developed. Background information: https://doi.org/10.1101/2024.07.31.605955
Is data available	Data is already available, but **NDA signing is necessary.**
Contact person	Priit Paluoja
Organization	Celvia CC AS (https://celvia.ee/en/)
How many teams they are ready to supervise?	**There is room for only 1 team in this project -> already taken.**
Instructors' comment	This project might require some basic understanding of biology and you should take it if understand a bit of this: https://doi.org/10.1101/2024.07.31.605955)

Topic #	P03
Domain	Autonomous Driving
Title	**P03 - Automatic labeling of traffic light detection dataset**
Description	Automatic labeling of traffic light detection dataset using Segment Anything 2 model. The created dataset should include annotations for sequences of images to allow detection of blinking lights, which is not possible with single image. Bonus is to train lightweight Yolo model on the created dataset.
Is data available	Data can be artificially generated
Contact person	Tambet Matiisen
Organization	Autonomous Driving Lab of University of Tartu
How many teams they are ready to supervise?	**Only 2-3 teams can participate (3/3) - this has been taken** **Potentially a possibility to get a prize for the best solution.**
Instructors' comment	This project will require you to run a model called SAM2, for labelling videos and then training computer vision models. This is going to be hard project, be aware.

Topic #	P04
Domain	Autonomous Driving
Title	**P04 - Automatic labeling of vehicle blinker detection dataset**
Description	Automatic labeling of vehicle blinker detection dataset using Segment Anything 2 model. Knowing the state of blinkers is important for the prediction module, to estimate if the vehicle is going to change lane or turn. Detecting blinkers from single video frame is not possible, as they are, you know, blinking. Bonus is to train Yolo detection model that outputs in addition to vehicle bounding boxes also the blinker state as class.
Is data available	Data can be artificially generated
Contact person	Tambet Matiisen
Organization	Autonomous Driving Lab of University of Tartu
How many teams they are ready to supervise?	**Only 2-3 teams can participate (1 / 3)** **Potentially a possibility to get a prize for the best solution.**
Instructors' comment	This project will require you to run a model called SAM2, for labelling videos and then training computer vision models. This is going to be hard project, be aware.

Topic #	P05
Domain	Autonomous Driving
Title	**P05 - Lidar-only self-supervised 3D object detection network**
Description	Train lidar-only self-supervised 3D object detection network. Replicate the LISO paper: https://baurst.github.io/liso/
Is data available	Data can be artificially generated
Contact person	Tambet Matiisen
Organization	Autonomous Driving Lab of University of Tartu
How many teams they are ready to supervise?	**Only 2-3 teams can participate (2 / 3)** **Potentially a possibility to get a prize for the best solution.**
Instructors' comment	Lidar point-clouds are hard and unusual data to work with.

Topic #	P06
Domain	Time Series Prediction
Title	**P06 - Time-series forecasting: hourly national electricity consumption**
Description	Problem: electricity production and consumption are matched through markets. The main one is the Day-Ahead which happens daily at 13:00, the day before delivery of electricity. Consumption forecast plays an important role in the formation of the price. This project would help having hourly forecasts of the consumption for the next day. The scope is the Baltic region (EE+LT+LV). Goal: develop a time-series forecasting algorithm based on weather forecasts + datetime features to forecast hourly national electricity consumption for EE, LT and LV. Method: inspiration can be gotten from last winter Kaggle competition (https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers). Gradient boosting should be the more efficient but some deep-learning approach especially Transformers could be tried. Additionally, time-series feature engineering is the key.
Is data available	Data is available online, but it needs to be fetched or scrapped
Contact person	Jean-Baptiste Scellier
Organization	Enefit / Eesti Energia
How many teams they are ready to supervise?	**There is room for only 1 team in this project -> already taken!**
Instructors' comment	Time series is not covered in our course, so you might need to do a lot of googling to figure out the right approaches.

Topic #	P07
Domain	Time Series Prediction
Title	**P07 - Time-series forecasting: renewable electricity production**
Description	Problem: electricity production and consumption are matched through markets. The main one is the Day-Ahead which happens daily at 13:00, the day before delivery of electricity. Renewable production forecast plays an important role in the formation of the price. This project would help having hourly forecasts of the renewable production for the next day. The scope is the Baltic region (EE+LT+LV), electricity production from solar and wind. Goal: develop a time-series forecasting algorithm based on weather forecasts to predict hourly national renewable electricity production for EE, LT and LV. Method: inspiration can be gotten from last winter Kaggle competition for PV production (https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers). Gradient boosting should be the more efficient but some deep-learning approach could be tried. Time-series feature engineering will be the key.
Is data available	Data is available online, but it needs to be fetched or scrapped
Contact person	Jean-Baptiste Scellier
Organization	Enefit / Eesti Energia
How many teams they are ready to supervise?	**There is room for only 1 team in this project -> already taken.**
Instructors' comment	Time series is not covered in our course, so you might need to do a lot of googling to figure out the right approaches.

Topic #	P08
Domain	Time Series Prediction
Title	**P08 - Time-series forecasting: electricity prices**
Description	Problem: electricity production and consumption are matched through markets. The main one is the Day-Ahead which happens daily at 13:00, the day before delivery of electricity. Given the market price, we are willing to produce or not electricity. This project would help having hourly forecasts of the price for the next day. The scope is the Baltic region (EE+LT+LV). Goal: develop a time-series forecasting algorithm based on consumption forecast, renewable production forecast, oil & gas prices, interconnectors capacity to forecast hourly day-ahead electricity prices in the Baltics. Method: Gradient boosting or some deep-learning approach especially Transformers could be tried. Good data preparation and time-series feature engineering will be key.
Is data available	Data is available online, but it needs to be fetched or scrapped
Contact person	Jean-Baptiste Scellier
Organization	Enefit / Eesti Energia
How many teams they are ready to supervise?	**There is room for only 1 team in this project - this has been taken.**
Instructors' comment	Time series is not covered in our course, so you might need to do a lot of googling to figure out the right approaches.

Topic #	P09
Domain	ML on Tabular Data
Title	**P09 - Predicting which businesses are unlikely to pay their fines**
Description	Using ML to predict which businesses are unlikely to pay their fines issued by the registry department of Tartu County Court (the Estonian Business Register) The back story All legal entities in Estonia are register in the Estonian Business Register. While a business is registered, they have the duty to keep their data up to date in the register and submit an annual fiscal report. Failing to do either, can result in the business being fined. There are thousands of fines issued by the Business Register every year. The levels of fines being payed by legal entities are not good. The goal Our goal is to find the legal entities that are unlikely to pay their fines and to do so even before the fine is issued. There is a real life use case and a successful project could help improve the Estonian business environment. The data There are 348 465 legal entities registered in the Estonian Business Register and there have been 34 528 fines issued in 2024 alone (as of 20.09.2024). Main reason for the fines being issued is the annual fiscal report not being submitted (on time). We have access to the data from the Business Register. This is open data, but will need to be anonymised before any work can be done on it. There is a possibility (dbc) to also add in open data from other government institutions. There is no one big dataset ready, but one could be put together based on data available and deemed interesting in relation to the project. There is a level of SQL work to be done before the ML work can start. The team I am taking part of the Machine Learning course and I’m looking for 1-3 people to join me, so we could form a team and tackle this challenge together.
Is data available	The data is available, but not in a formed and ready dataset. The dataset can be put together based on data available and deemed interesting in relation to the project. There is a level of SQL work to be done before the ML work can start.
Contact person	Kersti Mikkov (currently taking the course)

Organization	RIK - e-Äriregister (e-Business Register)
How many teams they are ready to supervise?	**Only 2-3 teams can participate (3/3) - please do not take this project anymore.**
Instructors' comment	Will likely take time to acquire (via SQL) and preprocess the data.

Topic #	P10
Domain	Multiclass Classification
Title	**P10 - Smart invoice classification with Trigon**
Description	One of the main tasks of an accountant is to categorize purchase invoices and expense receipts. This categorization depends on interpreting the country's tax laws. Trigon is a new accounting software that uses a standardized method for these ""classifiers."" Unlike many other systems, Trigon standardizes account and VAT rate entities, making the categorization process easier for accountants. We believe this approach will allow us to implement a machine learning (ML) solution to help accountants work faster and make fewer mistakes. Each invoice contains machine-readable data (either digitized or in e-invoice format), which can be used for categorization. The project's goal is to develop a recommendation model for these invoice classifiers (account and VAT rate entities). With this solution, Trigon can suggest appropriate accounts and VAT rates to accountants and alert them when they use an uncommon account and VAT rate combination.
Is data available	Data will be provided by Trigon's team
Contact person	Jaanus Karlson (can spend more than 5 hours monthly on mentoring)
Organization	Trigon OÜ
How many teams they are ready to supervise?	**There is room for only 1 team in this project -> has been taken.** **Potentially a possibility to get a prize for the best solution.**
Instructors' comment	Probably will take time to understand the task and the data.

Topic #	P11
Domain	Computer Vision: Segmentation
Title	**P11 - Segmentation Smackdown: Battle of the Models**
Description	The segmentation of fibrous structures from scanning electron microscopy images is often difficult due to the varying nature of the fiber morphology. The project's main goal is to improve the segmentation approach implemented thus far in FiBar (https://fibar.elixir.ut.ee/). The idea is to test state-of-the-art segmentation approaches (Segment Anything Model by Meta, U-Net models, and potentially other transformer models) and fine-tune these models to improve segmentation.
Is data available	Data is already available
Contact person	Marilin Moor
Organization	University of Tartu
How many teams they are ready to supervise?	**There is room for 2-3 teams (2 / 3)**
Instructors' comment	Requires some background in computer vision, preferably experience with segmentation task.

Topic #	P12
Domain	ML on Tabular Data
Title	**P12 - Factors of success in abstract theoretical courses**
Description	Courses containing a lot of mathematics are traditionally hard for students. However, there are still some students who excel in them. The objective is to find out what these students do differently and whether there are activities or behavioral patterns that could be recommended to the other students to make such courses easier for them. The dataset consists of coded activities of 119 students in the Theoretical Computer Science course each day throughout the whole semester. The activities are, for example, viewing study materials, submitting homework, participating in a lecture online, etc. The goals are: 1) predict the final grades of students based on their activity pattern; 2) discover activity patterns that separate high-performing students from low-performing students.
Is data available	Data is already available
Contact person	Reimo Palm
Organization	University of Tartu
How many teams they are ready to supervise?	**Has already been taken by enough teams.**
Instructors' comment	The data is small, and thus, it might be hard to train ML models.

Topic #	P13
Domain	NLP
Title	**P13 - Coupled seq2seq training**
Description	We have a new method of training sequence-to-sequence models that lets an existing pre-trained model (like Whisper / NLLB / mT5 / etc) to serve as a guide in training a new model. That way the new model can serve as an extension module to the guide model: their vector spaces will be compatible and their encoders/decoders can be recombined and used. We have tested the method only on machine translation, the project goal is to expand the testing to new modalities: speech, images, other NLP tasks besides translation, music generation, etc. You will have to (1) find more seq2seq models on HuggingFace and update the implementation of the method to support these models and (2) find datasets on HuggingFace / elsewhere that can be used to train such extension modules in order to evaluate the approach.
Is data available	Data is available online, but it needs to be fetched or scrapped
Contact person	Mark Fishel
Organization	University of Tartu, TartuNLP
How many teams they are ready to supervise?	As many teams can participate as would wish
Instructors' comment	I do not completely understand all the things that this project requires, so be mindful 🙂

Topic #	P14
Domain	Defect detection
Title	**P14 - Industrial defect detection pipeline**
Description	Synthetic data generation to train defect detection model. The solution can in theory automate many boring jobs at some factories. The problem is that data is hard to collect, and even harder to precisely annotate. My idea is to combine classic computer vision techniques with custom optimization algorithm and with the power of Blender to generate realistic annotated images.
Is data available	Data is hard to obtain
Contact person	Anton Vykhovanets
Organization	University of Tartu
How many teams they are ready to supervise?	**There is room for only 1 team in this project**
Instructors' comment	Getting the data will require NDA + probably this project will require some experience with computer vision