Thesis Topics

**Robotic Automation of University Admission Processes**  
**Supervisor: Marlon Dumas**

The admission process for international students at University of Tartu involves several information systems that are not designed to talk to each other (e.g. DreamApply, SAIS, ÕIS, Urkund, and sometimes also Google Sheets). In order to move data across these systems, secretaries in different departments of the universities have to carry out manual, repetitive, and error-prone tasks (especially copy/pasting and searching).

Such manual tasks can be automated using an emerging type of technology called Robotic Process Automation (RPA). These tools allow business users to capture repetitive routines (e.g. moving data from one system into another via copy/paste actions). In this Masters project, you will analyze the existing process for admissions in at least two institutes of the university, and you will assess the possibility of automating parts of these processes (especially routine tasks) using RPA technology. As part of the project, you will review the capabilities of existing RPA tools, and you will implement at least two repetitive tasks using one of these tools in order to demonstrate the feasibility and potential benefits of using RPA in the university admission processes. This initial feasibility study and benefit assessment would be used as a basis for preparing a business case for the use of RPA technology to (partially) automate the admission process.

Given that RPA tools are meant to be used by business people, this project does not require any software development skills. However, knowledge of business process modeling and process automation (using a BPMS) would be very useful as a starting point. The topic is suitable for non-IT students.

**Title: CV-Keskus: resumes' analysis for creating prediction models.**  
**Supervisor(s): Rajesh Sharma (rajesh.sharma@ut.ee) and Jaan Masso**

Description: Job seekers often use online social platforms for job search. In this thesis, we will use CV-Keskus dataset either for descriptive or/and for predictive analysis with an aim to find various questions such as 1) what is the wage gap between Estonian and foreigners while comparing gender wages or 2) which fields attracts foreigners. Another possible problem could be building a rating score while applying to job by matching the experience, nationality, age, skills and using historical data of similar job descriptions. The thesis can be done either by investigating various small questions or by looking at one main research questions.  
Dataset: CV Keskus dataset.

**Title: Predicting box office revenues using wisdom of the crowd and movie critics**  
**Supervisor(s): Rajesh Sharma (rajesh.sharma@ut.ee) and Risko Russ**

Description: The term “wisdom of the crowd” refers to the collective opinion of a community or group. In comparison, expert views refer to the views expressed by the experts of a particular domain. In this thesis, you will investigate if it is the experts or if it’s the wisdom of the crowd, that can predict the box office outcome of the movies. This is continuation work of last year thesis, for which we would like to improve our approach to get better results. In particular, you will analyse tweets with respect to movies around the period of release date of movies. A dataset of tweets of around 600 movies will be provided. However, we also expect to
expand our analysis by collecting tweets about more movies during the period of thesis. The thesis involves, sentiment analysis of the tweets and subsequently proposal of the prediction model about predicting box office result of the movies. The topic lies at the intersection of big data analytics, text analytics, machine learning, business data analytics.

Dataset: Dataset will be provided.

**Title: Predictive analytics of companies**

**Supervisors:** Peep Kungas (peep.kungas@ir.ee) and Rajesh Sharma (rajesh.sharma@ut.ee)

**Description:** In this thesis, the student will be performing data science activities, in particular about predictive analysis about various companies with respect to financial status. In particular we are interested in tasks such as credit limit recommendation and late payment prediction. An example of credit limit recommendation could be if company A sells to company B, then A sends B an invoice. If the invoice amount exceeds credit limit of company B then the invoice must be paid by B before goods/services are received by B from A. An example of late payment prediction includes if company A sells to company B, then A sends B an invoice. If the invoice amount exceeds credit limit of company B then the invoice must be paid by B before goods/services are received by B from A. Otherwise payment date needs to be set. Typically the number of days between invoice date and due date is fixed in company policies, but when company is optimizing its sales, it is interested in increasing the delay if it helps to close deals or keep customers or reduce customer financial stress level.

Dataset will be provided for the thesis which include company credit score history, default company background data features - board membership network metrics, financials and financial indicators, (tax) debts, market sector, paid taxes etc.

**Title: Media monitoring**

**Supervisors:** Rajesh Sharma (rajesh.sharma@ut.ee) and Peter Ormosi, UEA, UK

**Title:** Measuring corporate reputation through online social media

**Supervisors:** Rajesh Sharma (rajesh dot sharma at ut dot ee) and Peter Ormosi

**Description:** When businesses are caught out engaging in illegal or immoral activities, their reputation might suffer. Corporate reputation is a reflection of how a business is regarded by its customers and the public in general. If corporate misbehaviour negatively affects a business’ reputation, customers might switch to rival businesses. For this reason, reputation has got a central role in free markets as it has the potential to deter businesses from misbehaving. The extent, to which corporate wrongdoings trigger a reputational loss is still debated and is subject to a large body of academic works. Most of these works are based on survey methods to measure reputation. This research relies on a more direct method to measure reputational changes, by conducting a sentiment analysis of how the public reacted on Twitter to some of the most high-profile corporate misconducts. In this particular work thesis, corporate reputation will be studied using the Volkswagen (VW) scandal as a case study and the public reaction it created on the Twitter. VW’s scandal has been chosen because it has been widely covered over time through both traditional and social media. Moreover we can measure how changes in media coverage and social media reaction affected VW’s financial performance. The dataset and related literature will be provided for speeding up the work.

Datset: Dataset for one use case will be provided however, we expect to collect some additional datasets for comparative study.

References:
[1] Corné Dijkmans, Peter Kerkhof, Camiel J. Beukeboom, A stage to engage: Social media use and
Title: Analysing Server Logs for predicting Job Failures.
Supervisor: Rajesh Sharma (rajesh.sharma@ut.ee) and Alina Sirbu

Server logs generally refer to files which are created for monitoring the activities being performed on servers. In recent years a lot of research has been performed in analysing server logs for analysing the status of the jobs or tasks that arrive on servers.

In this thesis, you will be analysing logs from Google cluster, which is a set of machines responsible for running real Google jobs for example, search queries. The research encompasses the domain of large scale data analytics and machine learning. The main contribution of the thesis includes proposing of model to predict the job failures on servers.

Dataset: Real dataset of Google traces will be provided along with related literature to ramp up the learning process.

Title: Gender-based segregation in company boards and well-being
Supervisors: Peep Kungas (peep.kungas@ir.ee) and Rajesh Sharma (rajesh.sharma@ut.ee)

Description: Segregation is an unjustified separation or distance in social environments (physical, working, or on-line) of individuals on the basis of any physical or cultural trait. A segregation index measures the segregation degree of a minority group within each of units (e.g. schools) weighing each unit by some relevance. The core hypothesis of this topic is that distribution of gender and age segregation reflect a change in the labor market. There are some initial findings (e.g. isolation index negatively correlates with an unemployment rate in case of young men in Lääne-Virumaa (2008-2015)), which seem to confirm this hypothesis, but further studies are required to confirm the hypothesis. Furthermore, researchers at University of Pisa have developed models for measuring segregation in boards of companies for which input data is available on daily basis. Hence, usage of segregation metrics to now-cast (un)employment means in practice that quarterly delays in measuring the effect of policy changes to (un)employment can be reduced to virtually zero. The latter allows raising the quality of decision-making wrt (un)employment. Typically higher values of the index mean higher segregation. In this thesis, the student will analyse the even distribution with respect to gender and age in boards of companies leads to improved credit risk management low credit risks of companies in a region and if they have strong positive correlation to well-being of the society.

Dataset: Dataset will be provided.

Title: Discovery of public-private corruption cases
Supervisors: Peep Kungas (peep.kungas@ir.ee) and Rajesh Sharma (rajesh.sharma@ut.ee)

Description: Corruption is a major obstacle to sustainable economic, political and social development. Overall, corruption reduces efficiency and increases inequality. CleanGovBiz estimates that the cost of corruption equals more than 5% of global GDP. Hence effective means to prevent corruption have significant effect to development. Corruption is also an issue in Estonia. In fact, the
number of revealed corruption cases is on the rise – registered number of corruption offences were 161 and 450 in 2012 and 2015 correspondingly. This project aims at reducing public-private corruption in Estonia by providing a mechanism for automatically revealing corruption patterns and using the mechanism to discover corruption cases already in their early stages. The approach is to apply a combination of social network analysis and machine learning techniques to analyze temporal networks of organizations, persons and assets (tenders, financial aid, real estate objects, etc) in order to find temporal network patterns, which describe existing corruption cases.

Datasets: Following data for the purpose of the project:
- Board members and owners of businesses
- Some features of businesses
- Real estate ownership data
- In addition the project will benefit from the following datasets:
  - Public tender data (sums, descriptions and winners of tenders)
  - Grants, financial aid and subsidies
  - Public sector officials/employees
  - Corruption cases (specific organizations and persons) for learning


Title. Media monitoring for business analysis
Supervisors: Peep Kungas (peep.kungas@ir.ee) and Rajesh Sharma (rajesh.sharma@ut.ee)

Description: Social media reach and engagement have turned out to be key metrics, which allow measuring the performance of posts. Reach tells the size of the audience reached by a post/mention, while engagement indicates the number of individuals, which reached to the post/mention. However, there is no model to measure the same for online mentions. In this thesis, a student can investigate the problem related to what is the measurable impact of company’s marketing / communication activities? [What are we measuring ? and how ?]. This problem can be analysed either by 1) using predictive modelling or 2) by alternative approach such as by define engagement in Web through user’s Web search action. In particular, media mentions of Estonian businesses is of interest to us.

Dataset: The dataset in form of web server logs of companies page visits and web visitor logs will be provided. In addition, we expect to crawl google trend logs would be crawled.

Title: Tonality of company media mentions
Supervisors: Peep Kungas (peep.kungas@ir.ee) and Rajesh Sharma (rajesh.sharma@ut.ee)

Description: Social media is often used by companies to reach out fast and to a broader set of audience. However, it can be used for companies to infer what is the public perception about the company itself. For example, a company might be interested in an online activity, which requires immediate action (e.g. for disaster prevention) with respect to the company. In a different case a company might be interested in determining what it should change in order to increase market share with respect to the competitors in terms of price, service/product, delivery etc. In particular we are interested in questions like what the measurable impact of company’s marketing / communication activities?
Title: Model for estimating dynamics of market size
Supervisors: Peep Kungas (peep.kungas@ir.ee) and Rajesh Sharma (rajesh.sharma@ut.ee)

Description: Companies offer struggles with the question whether to focus a) when to develop new products, or b) just stick to existing products. The basic questions which revolves around this topic are 1) What is the size of a market and 2) at which pace the market is growing/shrinking? In this thesis, a student will work on marketing activity data to explore some of the case studies.

Title: Social capital at work places
Supervisor: Rajesh Sharma (rajesh.sharma@ut.ee)

Description: Social capital is defined as contribution of a group in terms of resources as a whole. This is group social capital, compared to this, individual social capital is defined as collection of resources of the neighbours of a node. Using resource theory and social network analysis, a student will explore in particular in work place settings how social capital can play an important role in career advancement.
Dataset: Enron dataset with nodes and their description will be provided. However, we expect student to collect additional resources.

Title: Predictive Analysis on Twitter: Techniques and Applications
Supervisors: Rajesh Sharma (rajesh.sharma@ut.ee) and Anurag Singh

Predictive analysis of social media data has attracted considerable attention from the research community as well as the business world because of the essential and actionable information it can provide. Over the years, extensive experimentation and analysis for insights have been carried out using Twitter data in various domains such as healthcare, public health, politics, social sciences, and demographics. Some fine-grained analysis may be done, involving aspects such as sentiment, emotion, and the use of domain knowledge in the coarse-grained analysis of Twitter data for making decisions and taking actions, and relate a few success stories. Social media data has Classal ready enabled researchers to predict the trends and outcomes of several critical real-world events, and its reliability and coverage can further be improved by incorporating background knowledge.

Modeling Business Processes on an Blockchain Eco-System (BPMN) - Booked

Supervisor: Fredrik Milani (milani [ät] ut [dot] ee) and Luciano Garcia Banuelos

Blockchain technology is gaining traction and is speculated to be the next “internet”. Working with blockchain to improve processes, require working with conceptual process models. However, there are no standards for how to model processes running on blockchain enabled inter-organizational eco-systems. This thesis aim at investigating how such processes can be modelled using BPMN. The work for this thesis will involve modeling the current processes of a case and then modeling it as if it was running on a blockchain enabled eco-system. This would mean modeling BPMN processes, collaborations, and choreographies. The work also includes discussion and conclusions about BPMN for modeling such processes.
Modeling Business Processes on an Blockchain Eco-System (CMMN) Booked

Supervisor: Fredrik Milani (milani [ät] ut [dot] ee) and Luciano Garcia Banuelos

Blockchain technology is gaining traction and is speculated to be the next “internet”. Working with blockchain to improve processes, require working with conceptual process models. However, there are no standards for how to model processes running on blockchain enabled inter-organizational eco-systems. This thesis aim at investigating how such processes can be modelled using CMMN (Case Model Management and Notation). The work for this thesis will involve modeling the current processes of a case and then modeling it as if it was running on a blockchain enabled eco-system. The work also includes discussion and conclusions about CMMN for modeling such processes.

Framework for Process Mining in Industry – Systematic Literature Review

Supervisors: Fredrik Milani (milani åt ut dot ee) and Fabrizio Maggi

Process mining is a technique increasingly used in the industry. However, there is still some work left before it becomes as common as other data mining methods. One reason might be that industry is not well acquainted with what can be done with process mining and its value. This thesis is about conducting a systematic literature review, examine the list of relevant papers, and develop a framework that help industry finding the most suitable process mining tool depending on their needs. Some of the work for this thesis has already been conducted and as such, more emphasis will be given on framework development.

Business Process Redesign for Blockchain Solutions - Booked

Supervisor: Fredrik Milani (milani [ät] ut [dot] ee) and Luciano Garcia Banuelos

Process innovation and re-design is an important part of any business process initiative. Over the years, several methods and approaches have been developed for changing processes. Oftentimes, re-design is enabled by new technology. We are seeing blockchain technology emerging as a potential force of innovation and disruption. Process changes play an important role for enabling innovations to deliver value. This thesis topic is to about examining how existing methods and approaches for process redesign can be applied in the context of processes on blockchain technology. The thesis will require to get acquainted with existing methods, take an existing as-is process, and apply redesign on it to create the to-be processes. Following this, analysis is conducted and conclusions made.

GDPR and Blockchain Solutions

Supervisor: Fredrik Milani (milani [ät] ut [dot] ee) and Luciano Garcia Banuelos

Recently, GDPR went into effect. GDPR (General Data Protection Regulation) is a regulatory document affecting what kind of data companies can store and use. The implications are still being
investigated but all commercial enterprises operating within the European market must comply to GDPR. At the same time, we see blockchain technology based solutions grow. However, it is not known what parts of GDPR can be complied with when using blockchain technology. This thesis will examine what GDPR means and how it related to solutions based on blockchain. The work will include review of blockchain technology, mapping of what parts of GDPR can be fulfilled and which parts cannot. For instance, the right to be forgotten is very difficult to manage with blockchain. Following this, the work also includes investigating how the parts not fulfilled can be solved. The contribution of the thesis will be valuable for enterprises wishing to build compliant blockchain solutions.