Master's thesis topics

1. Algorithms

	Title	Abstract	Supervisors	Link
1	Information Extraction for Database Population from Scientific Articles	Scientific data are scattered among huge amount of articles produced by different research groups. Keeping track of all studies is extremely challenging yet it is important for certain research questions. E.g. to study an influence of climate change on animal communities it is important to know the basic information, since different types of animals live in different kinds of environments that are bound to different climate. This information is spread among numerous papers but could be collected and extracted automatically—this task is called Information Extraction (IE). The aim of the thesis is to implement one of the state-of-the-art IE methods populate a database developed at the Finnish Museum of Natural History. The method will most probably utilize transformers, RNN, or other deep learning network. This thesis is a good opportunity to start learning NLP, since IE is one of the core NLP tasks.	Lidia Pivovarova lidia.pivovarova@helsinki.fi	https://www.proquest.com/docview/2307372176
2	Greedy algorithm for Shortest Cyclic Superstring	The Shortest Linear Superstring problem is one of the most known problems in stringology (string algorithms). The aim is, for a set of strings as input, to find a shortest linear string which contains all the strings of the input. Even if this problem has been strongly studied, a lot of questions stays open. In particular, a 30 old conjecture says that one of the easiest algorithm (the greedy algorithm) has an approximation ratio better than all the approximation ratio prooved for all the other algorithms. The topic of this thesis is to understand how this greedy algorithm works in the Shortest Circular Superstring problem where the superstring is not linear but circular. References: • Avrim Blum, Tao Jiang, Ming Li, John Tromp, Mihalis Yannakakis: Linear Approximation of Shortest Superstrings. STOC 1991: 328-336 • Alexander Golovnev, Alexander S. Kulikov, Ivan Mihajlin: Approximating Shortest Superstring Problem Using de Bruijn Graphs. CPM 2013: 120-129	Bastien Cazaux, Veli Mäkinen	https://arxiv.org/abs/1809.08669
3	Fast nearest neighbor search.	There are several approaches to speed up (approximate) nearest neighbor queries in large data sets. Generally, they involve an initial stage where an index data structure is constructed. The index can be used to perform queries when new points arrive. The thesis can review and compare various different approaches (tree-based, locality-sensitive hashing, random projections, ...) and/or experiment with new variantions of the theme. References: Zezula, P., Amato, G., Dohnal, V., Batko, M., Similarity Search: The Metric Space Approach, Springer, 2006. V. Hyvönen, T. Pitkänen, S. Tasoulis, E. Jääsaari, R. Tuomainen, L. Wang, J. Corander, and T. Roos (2016). Fast nearest neighbor search through sparse random projections and voting, in Proc. 2016 IEEE International Conference on Big Data (IEEE Big-Data 2016), Washington DC, Dec. 5–8.	Teemu Roos	
4	Data mining historical textual traditions	Historical textual traditions, including manuscripts, early printed materials and collections of oral traditions, offer a rich source of information for data mining. Various exploratory data analysis methods can be used to summarize and visualize the data. A thesis project could apply known methods to new data collections available through public repositories or ongoing collaborative projects. References: J. Tehrani, Q. Nguyen, and T. Roos, (2016). Oral fairy tale or literary fake? Investigating the origins of Little Red Riding Hood using phylogenetic network analysis, Digital Scholarship in the Humanities 31(3):611–636.	Teemu Roos	https://www.helsinki.fi/en/researchgroups/digital-humanities http://republicofletters.stanford.edu/
5	"White-box" machine learning	Deep learning methods based on massive neural network architectures tend to be very hard to understand and interpret. They can be described as "black-boxes" that are good for mapping inputs into ouputs but it is nearly impossible to see (and understand) what happens inbetween. There have been initiatives to construct "white-box" methods, i.e., corresponding techniques with the added benefit of being understandable.	Teemu Roos	http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731
6	Solving data analysis / machine learning problems cost-optimally using constraint solvers	Constraint optimization, maching learning, and data mining are today well-established and thriving research fields within computer science. Each of the fields have contributed fundamental techniques and algorithmic solutions that are today routinely applied for addressing hard computational problems in various real-world context. However, the possibilities of harnessing the highly efficient constraint solving technology available today in providing generic and efficient solutions to various machine learning and data mining problems, such as different kinds of classification, structure learning, probabilistic reasoning, and pattern mining tasks, have only recently been realized, and there is plenty of opportunities for developing novel algorithmic solutions with optimality-guarantees to machine learning tasks via employing constraint solving. This area offers topics for several MSc theses, and the specifics (including the problem focused on) can be agreed on together with the thesis supervisor.	Matti Järvisalo	http://www.hiit.fi/cosco/coreo/
7	Instance structure and empirical hardness of NP-hard problems	NP-complete problems, such as the Boolean satisfiability problem (SAT), are deemed "intractable" by classical complexity theory. Despite this, modern exact algorithms for SAT, i.e., SAT solvers, are today routinely used for solving extraordinarily large problem instances of NP-hard problems arising from various industrial (e.g., software and hardware verification) and AI (e.g., automated planning) problem domains. However, the question of why exactly SAT solvers are so powerful in solving instances of NP-hard problems in the real world is not well understood. This problem setting gives rise to various research questions suitable for MSc theses, ranging from empirical to theoretical studies aiming at understanding the relationship between the underlying structure of real-world problem instances and the algorithmic techniques implemented in modern SAT solvers.	Matti Järvisalo	http://www.hiit.fi/cosco/coreo/
8	Testing and debugging constraint optimization solvers	Constraint solvers, such as those for the integer programming and propositional satisfiability (SAT) paradigms, offer off-the-self generic tools for solving NP-hard problems at large via declaratively expressing problems using mathematical constraints and feeding them to a constraint solver for obtaining an exact solution to the original problem. While this approach has turned out to be a successful approach to solving various hard computational problems, obtaining correct answers (especially in cases when no solutions exist) relies on the correctness of the solver implementations. This MSc thesis topic combines testing and debugging techniques from software development and constraint solving. The aim is to develop practical fuzz testing and delta debugging tools for constraint optimization (especially, Boolean optimization) which help solver developers to test and debug their solver implementations. This topic is suited for both algorithms and software systems students.	Matti Järvisalo	http://www.hiit.fi/cosco/coreo/
9	Algorithmic Uses of the Gumbel-Max Trick	There are powerful paradigms for solving optimization problems, like (integer) linear programming and branch-and-bound, which however do not readily allow random sampling from the distribution proportional to the objective function. Recently, an old, simple reduction from sampling to optimization, called the Gumbel-Max trick has found non-trivial algorithmic uses in machine learning and artificial intelligence. The thesis will review that literature. References: 1. Chris J. Maddison, Daniel Tarlow, Tom Minka: A* Sampling. NIPS 2014: 3086-3094. 2. Carolyn Kim, Ashish Sabharwal, Stefano Ermon: Exact Sampling with Integer Linear Programs and Random Perturbations. AAAI 2016: 3248-3254	Mikko Koivisto	

10	Strong Exponential Time Hypothesis	The Strong Exponential Time Hypothesis (SETH) asserts, roughly, that the classical CNF Satisfiability problem with n variables cannot be solved in time O(c^n) for any constant c < 2. Recent research has proved SETH-based lower bounds for various important problems, often using ingenious reductions. The thesis will review the state of the art and discuss the key open problems in the area. References: 1. Mihai Patrascu, Ryan Williams: On the Possibility of Faster SAT Algorithms. SODA 2010: 1065-1075	Mikko Koivisto	
11	Compensating for small labeled data	*Project context*: a system for web-scale surveillance of news media. Massive volumes of news stories are analyzed, classified and clustered according to various criteria using deep learning neural networks. For example, we try to determine what type of event is reported in the story (from a large set of different event types), or the sentiment – whether a story is positive /negative for a given entity (person/company/etc.) mentioned in the story. A major bottleneck in deep learning is the shortage of labeled data (as in most supervised learning). The project will explore methods for *data augmentation* – leveraging small amounts of labeled data by transforming it to produce more novel data usable for learning. Data augmentation has been successful in some applications in image analysis. An alternative for a similar purpose is *transfer learning* – improving performance on a given task by using data that was labeled for a different task.	Roman Yangarber	http://puls.cs.helsinki.fi http://newsweb.cs.helsinki.fi
12	Models of language evolution	Over the last 15 years, several models have been proposed for language evolution. Methods have been developed for modeling historical relationships among languages in a language family. But little has been done in the way of formal comparison of the effectiveness of the models. The models can be *"applied"* in various ways: e.g., to build family trees, which can then be compared against trees proposed by linguists, based on manual analysis. Some models may be used to predict unseen data, or to reconstruct ancestor word forms. The models typically find regularities in lexical data – in lists of related words. Better models should find more regularity in the data. Alternatively, a probabilistic model assigns probability to observed data, which gives us a measure of model quality: better models assign higher probability to the data. *Question:* do intrinsic and extrinsic measures of model quality correlate as we expect? For example, does a theoretically better model produce trees that agree better with the linguists' results? The project may involve literature surveys, empirical studies using existing models and language data, or both.	Roman Yangarber	http://anthology.aclweb.org/K/K16/K16-1.pdf#page=154 http://etymon.cs.helsinki.fi
13	Modeling knowledge states in language learning	*Project context:* a system for learning languages. Learners use arbitrary texts, which they upload to the system, from which the system then creates a wide variety of exercises for practice in a game-like environment, while the system tracks the learners' progress. *Problem 1:* modeling the learners' competence. *Problem 2:* modeling the complexity of a text "in general" and the complexity of a text relative to a specific learner. While standardized tests treat learner's competence as a value on a linear scale, in sophisticated models of learning, competence is never a scalar. It can be a state in a large space of possible knowledge states, or a probability distribution over a set of possible states. The paths through the space are not arbitrary, they are constrained by the nature of the subject being learned. Can we infer the structure of the knowledge space from data – by observing correct vs. incorrect answers to many exercises from many learners ? We also seek multi-criterion estimates of the complexity of a text that a student chooses for learning. We consider both objective and subjective criteria. As objective criteria, people previously used lexical and syntactic complexity – frequent usage of rare words, complex syntactic constructions, etc. Among a population of students, the system can rank students' competence: we expect that learners with "lower" competence make more mistakes on any given text (on average) than learners with "higher" competence. Therefore, subjective measures may predict that text A is more complex than text B if students – at a fixed level of competence – make more mistakes on A than on B. How well do objective and subjective criteria that have been proposed in the literature correlate on actual user data? (One motivating goal is: when a learner uploads a text, the system should rate the text on a scale – too easy/just right/too difficult for you.) The project may involve a literature survey, an empirical study, or both.	Roman Yangarber	https://revita.cs.helsinki.fi Martin Schrepp. "Extracting knowledge structures from observed data." British journal of mathematical and statistical psychology, 52(2) Yudelson, Koedinger, Gordon. (2013) "Individualized bayesian knowledge tracing models". Artificial Intelligence in Education.
14	Human powered hierarchical clustering with relative distance comparisons	The input to a hierarchical clustering algorithm is typically a distance matrix that contains pairwise distances between the data items to be clustered. Commonly this matrix is defined e.g. in terms of the Euclidean distance between feature vectors. However, in some applications computing absolute (Euclidean or other) distances is not feasible. This happens for example if the distance information is collected using a crowd of human annotators. Humans are rather poor at consistently evaluating some notion of absolute semantic distance (on some arbitrary scale) between, say, the contents of two photographs. On the other hand, humans are fairly adept at comparing items relative to each other. The objective of this project is to design and analyse an agglomerative hierarchical clustering algorithm that uses (possibly noisy) relative distance comparisons to compute a clustering dendrogram.	Antti Ukkonen	
15	Machine learning in classic games	One of the early successes in applying machine learning to board games was Tesauro's NeuroGammon. In chess, on the other hand, champion level performance was obtained mainly without machine learning. Recently, reinforcement learning has been used in breakthrought in computer go and poker. This area offers various possible thesis topics, for example focussing on a particular game, or a particular learning technique.	Jyrki Kivinen	Libratus AlphaGo
16	Automating a music composition process	There are numerous methods for automated or algorithmic composition of music. This thesis project will approach this goal from a unique point of view: modelling and implementing the composition process of an actual human composer, who will participate in this project. The work will include conceptualisation and formalisation of the composition process together with the composer, in sufficient detail so that it can be implemented as a computational model, and experimented with. Some background in music is useful, so is interest in computational creativity.	Hannu Toivonen	

<table>
<tr><td>17</td><td>Variant projection in pan-genomics</td><td>**Added September 16, 2019**

Variant calling is the process of identifying how an individual differs from the consensus (reference) genome of the species. The standard approach is to align high-throughput sequencing reads on the reference, and then detect places where many alignments support the same variant. With such methods, several sequencing projects have gathered a huge catalogue of human variation and there is now an active field of research pondering how to exploit this *pan-genomic* information for the next projects. This research aims to identify ways to amend the linear reference genome with extra information to improve accuracy of variant calling. One such approach is to replace the reference genome with an *ad hoc* genome tailored for the sub-population under study [1]. Such *ad hoc* genome is closer to the sequenced individual, and hence the alignment and variant calling can be conducted more reliably. However, the variant calls need to be *projected* back to the reference genome in order to be compatible for downstream analysis tasks. We have implemented one projection strategy, but there is an alternative that might alleviate some of the challenges of the current approach.

In this project, we want to implement the alternative projection strategy and compare it experimentally with the current one. In addition, the current framework does not optimally support all heterozygous variants, and the goal is to incorporate handling of these in both the current and the new strategy.

The tool (PanVC), where these new features are to be added, is mainly implemented in C++.

In addition to the rather experimental implementation work, the thesis content can focus on more theoretical considerations around the theme.</td><td>Veli Mäkinen</td><td>[1] Valenzuela, Norri, Välimäki, Pitkänen, Mäkinen. Towards pan-genome read alignment to improve variant calling. *BMC Genomics*, 19:87, 2018. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4465-8</td></tr>
<tr><td>18</td><td>Parallelization of string matching algorithms extended to graphs</td><td>**Added September 18, 2019**

Current research in algorithmic bioinformatics focuses on extending string algorithms to work on a string and a labeled directed graph (instead of just two strings). This is motivated, for example, by the rise of pan-genomics, where a “pan-genome” is a labeled graph encoding all genomes observed in a population.

However, several such problems, for example exact string matching, can only be solved in quadratic time on graphs, as opposed to linear time on strings. This motivates the search for alternative strategies that work in practice, on inputs of billions of characters.

In this thesis you will develop methods to “decompose” a graph in several components such that these algorithms can be run in parallel on each such “component”, and then combine the results for each “component”. The focus is on aligning a string in a labeled directed acyclic graph (DAG) under the models of exact matching and edit distance. You will review (serial) algorithms for a string and a graph, review parallelization strategies for computing the edit distance between two strings, develop such decomposition strategies, prove their correctness, implement them and then perform experimental evaluations.

Ideally, you are already familiar with string algorithms (e.g. through the String Processing Algorithms course) and have experience with parallel programming. Moreover, ideally you also have some experience with GPU programming, in case the algorithms can also be implemented there.</td><td>Alexandru Tomescu (alexandru.tomescu@helsinki.fi)

Leena Salmela

(leena.salmela@cs.helsinki.fi)</td><td></td></tr>
<tr><td>19</td><td>Learning and utilizing document plans in end-to-end text generation</td><td>**Added December 4, 2019**

An overview (holds for this and the two following topics):

We offer several MSc thesis topics related to automated generation of textual reports describing given statistical data. This so-called data-to-text generation is an extensively studied topic in natural language generation. Most of the research is done using standardized data, such as weather forecasts or sport games outcomes. The thesis topics we offer take place in a more general context: automated production of a textual report from statistical data of various nature – from unemployment rate to greenhouse emissions. An additional difficulty is to make these methods work in a multilingual setting, e.g. Finnish and Swedish in addition to English.

One of the main drawbacks of end-to-end models, i.e. systems that learn to generate text from given structured data, is that their output lacks high-level structure: even though each sentence looks natural, the text as a whole fails to maintain a coherent narrative. The goal of this work is to develop a machine-learning approach for deciding the information content of the text from training examples. These plans would then be used in a hybrid NLG system that is partially rule-based. Plans will consist of a sequence of database records that should be one by one transformed into natural language sentences and/or phrases. Plan generation can be interpreted as classification and ranking machine learning problems. First, given a set of data tables and a set of texts, where each sentence is linked to one table record, it is necessary to train a classifier, which would decide for each record whether it should be included into the plan. Second, the selected records should be sorted (ranked) into a final sequence. Alternatively, this can be seen as an iterative contextual ranking problem; given a specific context (the present state of the document plan), the goal is to find the most suitable next record to include in the plan. The student would implement one or more methods to solve this problem and then test them on the dataset mentioned above.

Suggested reading:

1. Puduppully, Ratish, Li Dong, and Mirella Lapata. "Data-to-text generation with content selection and planning." *AAAI 2019: Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6908–6915.
2. Gehrmann, Sebastian, Falcon Z. Dai, Henry Elder, and Alexander M. Rush "End-to-End Content and Plan Selection for Data-to-Text Generation." *Proceedings of the 11th International Conference on Natural Language Generation*, 2018, pp. 46–56.</td><td>Lidia Pivovarova

lidia.pivovarova@helsinki.fi

Leo Leppänen

leo.leppänen@helsinki.fi

Hannu Toivonen</td><td>EMBEDDIA</td></tr>
<tr><td>20</td><td>Learning sentence templates for data-to-text generation</td><td>Using manually produced or at least verified templates to generate text makes the output controllable and interpretable. Learning templates is done by replacing certain word groups in the text with names of data slots. This could be done by a set of rules using mapping between sentences and data rows.

E.g. consider sentence “External operating revenue declined by 3.7 per cent.”

and the following database record:

<table><tr><th></th><th>2019</th><th>Change, %</th></tr><tr><td>Operating revenue total</td><td>5 045</td><td>-3,7</td></tr></table>
Given this database record the sentence could be converted into template “<PARAMETER> declined by <NUMBER> per cent”. Templatization could be done using manually compiled rules. Alternatively, a small set of seed rules could be extended using bootstrapping. The student will implement one or more template-learning methods and test them on the dataset mentioned above.

Suggested reading:

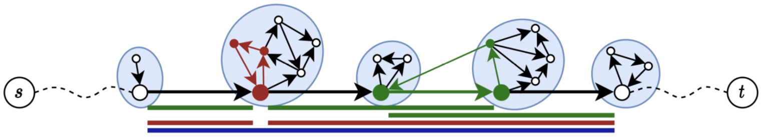
1. van der Lee, Chris, Emiel Krahmer, and Sander Wubben. "Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods." *Proceedings of the 11th International Conference on Natural Language Generation*, 2018, pp. 35–45.
2. Wiseman, Sam, Stuart Shieber, and Alexander Rush. "Learning Neural Templates for Text Generation." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 2018, pp. 3174-3187.</td><td>Lidia Pivovarova

lidia.pivovarova@helsinki.fi

Leo Leppänen

leo.leppänen@helsinki.fi

Hannu Toivonen</td><td>EMBEDDIA</td></tr>
</table>

21	Information Extraction for verification of generated texts	Information Extraction (IE) can be seen as a task opposite to data-to-text generation since it transforms unstructured text into a structured form. For instance, if a text generation task is to generate a news article from a table with sport match outcome, then the corresponding IE task would be to construct such table from a free text describing the event. This thesis topic will investigate the use of IE to verify the output of a natural language generation system. The student will implement an IE system, that could be trained using texts produced by humans from the dataset mentioned above. A possible approach to building an IE system is to start from manually constructed patterns and then automatically bootstrap more complex templates. Once the IE system is working with reasonable performance on human-generated text it can be used to assess performance of text generation by transforming text into a data table and then measuring the overlap between this reconstructed table and the original input data. In a more advanced setting the IE system could be used to improve a baseline data-to-text generation model by rejecting sentences that do not have support in the data. Suggested reading: 1. Nie, Feng, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin "A Simple Recipe towards Reducing Hallucination in Neural Surface Realisation." *ACL 2019: The 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2673–2679. 2. Puduppully, Ratish, Li Dong, and Mirella Lapata. "Data-to-text generation with entity modeling." *ACL 2019: The 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2023–2035.	Lidia Pivovarova lidia.pivovarova@helsinki.fi Leo Leppänen leo.leppänen@helsinki.fi Hannu Toivonen	EMBEDDIA
22	Event detection in historical news using permutation tests	**Added December 4, 2019** The main objective of this thesis is to find significant changes in word usages in historical newspapers, which potentially may correspond to events interesting for historical research. The main idea of permutation test is straight-forward: a time point corresponds to signal increase or decrease if the difference between its left and right area is significantly higher than an average difference that can be obtained in a large number of permutations. However, historical newspaper impose a number of problems related to incomplete and noisy data, multidimensionality and data imbalance. The student will have to develop a basic idea into a working algorithm, which would be applicable to digital humanities research.	Lidia Pivovarova lidia.pivovarova@helsinki.fi Hannu Toivonen	NewsEye - a Digital Investigator for Historical Newspapers
23	Experimental Evaluation of Dynamic Graph Algorithms	**Added February 26, 2020** Most of the graph applications in the real world deal with graphs that keep changing with time. These changes can be in the form of insertion or deletion of vertices or edges. An algorithmic graph problem is modeled in the dynamic environment as follows. There is an online sequence of insertion and deletion of edges and the aim is to maintain the solution of the given problem after every edge update. To achieve this aim, there is a need to maintain some clever data structure for the problem such that the time taken to update the solution after an update is much smaller than that of the best static algorithm. For the past three decades various dynamic graph algorithms have been developed but most of the work have focused on only theoretical evaluation of such algorithms. This often results in complicated algorithms that improves the theoretical bounds but are impractical for use in applications. Thus, an equally important aspect is the empirical performance of an algorithm on real world graphs. After all, the ideal goal is to design an algorithm having a theoretical guarantee of efficiency in the worst case as well as superior performance on real graphs. Often such an empirical analysis also leads to the design of simpler algorithms that are extremely efficient in real applications. Thus, such an analysis bridges the gap between theory and practice. The aim of the project would be to do experimental evaluation of some state of the art algorithms for a given graph problem, and to understand their behavior on the various kinds of graphs which may result in the development of simpler algorithms. Based on the progress and performance of the student, we can also consider a paid Research Assistant position as the project progresses.	Shahbaz Khan (shahbaz.khan@helsinki.fi) Alexandru Tomescu (alexandru.tomescu@helsinki.fi)	
24	Implementation, optimization, and experimental evaluation of graph algorithms for genome assembly	**Added on August 5th, 2020** In this project you will implement some very recent algorithms for graph problems motivated by the genome assembly problem in Bioinformatics, optimize these algorithms, and modify them to deal with some aspects of real data. The algorithm are presented in this preprint: https://arxiv.org/abs/2007.04726 In short, the algorithms look for paths (called *safe*) common to all s-t paths, trails and walks of a directed graph. There is freedom to the student to steer the project either in a direction focused more on algorithm engineering, or on practical bioinformatics aspects. If the project progresses well, and the work done is at "research level", then the student can also be hired as Research Assistant.  **Figure 1** Safe walks under different models for s-t-safety. The figure shows a sequence of s-t bridges as bold arrows and their bridge components as blue regions. Thick blue, red and green lines show the answers to the MaxSafe stPaths, stTrails and stWalks problems, respectively. Trail breakers and walk breakers have been highlighted in red and green, respectively.	Alexandru Tomescu (alexandru.tomescu@helsinki.fi)	https://arxiv.org/abs/2007.04726

2. Networking and services

Title	Abstract	Supervisors	Link
6G Systems and Cloud Continuum	Given the emerging high-density networks with the advent of versatile connectivity and new verticals, such as IoT and Industry 4.0, and their evolution, it is evident that data gathering and processing will be inherently distributed in 6G. The distributed processing is supported by a programmable network in terms of fine-grained network slicing and edge/fog computing for local processing capabilities. We have many M.Sc. thesis topics available pertaining to cloud continuum in 6G.	Sasu Tarkoma and Ashwin Rao	https://www.cs.helsinki.fi/u/starkoma/ https://www.cs.helsinki.fi/u/arao/

Cellular network meets AI	The developing 5G and future 6G, while improving networking capacities, also pose unprecedented challenges to the base stations, namely more complicated mobility management and protocol adaptation with lower latency tolerance. In this project, we investigate the potential of different machine learning techniques in tackling such challenges. The master thesis will focus on implementing state-of-the-art baseline and our learning-based algorithms in our customised NS3 simulator, and performance analysis of different learning algorithms in varied user mobility models and their impact on the performance of futuristic augmented reality applications. The thesis may also involve implementing the algorithms in our Software Defined Radio (SDR) stations to apply techniques to reality.	Pengyuan Zhou and Pan Hui	https://pengyuan-zhou.github.io/
Effectiveness of Opensource 5G core implementation	This work will include working with the Kumpula cellular test network and will be done in the context of the 5G Force project. The indoor cellular base stations in Exactum are currently offering networking connectivity using three different core networks: one from Nokia, one from OpenAirInterface , and one from a custom core network built by researchers at Aalto University. The proposed work will include expanding the network to include the 5G implementation of OpenAir, and also the free5Gc. This work will give the thesis work experience on building and maintaining networks along with the ability to work on tools and techniques to evaluate and improve the performance of communication networks. **Updated November 11, 2019**	Sasu Tarkoma and Ashwin Rao (name.surname@helsinki.fi)	https://www.cs.helsinki.fi/u/starkoma/, https://www.cs.helsinki.fi/u/arao/. https://5g-force.org/, https://www.openairinterface.org/, https://www.free5gc.org/.
Efficient monitoring in the far-edge.	This work will be done in collaboration with Nokia Bell Labs and will be based around a system called Infobus developed by them. Infobus uses a publish-subscribe mechanism built on top of Apache Kafka to disseminate the information gathered by nodes spread across the network. This thesis work is aimed at proposing, implementing, and evaluating a solution to extend the capabilities of the Infobus for edge networks including the far-edge. The project will also give the student the experience of collaborating with industrial research labs including the possibility to file for patents. **Updated on October 3, 2019.**	Sasu Tarkoma and Ashwin Rao (name.surname@helsinki.fi)	https://www.cs.helsinki.fi/u/starkoma/ https://www.cs.helsinki.fi/u/arao/ Previous thesis based on Infobus: https://helda.helsinki.fi/handle/10138/228837 Reference article summarizing edge computing: https://queue.acm.org/detail.cfm?id=3313377

Applications of blockchains in mobile systems	Blockchain technology provides, among other purposes such as cryptocurrencies, means to verify integrity of log information from potentially large systems. One example of an important log in mobile communication systems, is the log of phone calls and data usage. This is important because the log relates to the eventual bills sent to the mobile user. In case of roaming, fees may become surprisingly big, although inside European Union the high roaming fees are vanishing. The study would find out how blockchain technology could be used for such purpose in mobile systems. Alternatively, the study could focus on identifying other use cases of blockchains in mobile communications systems.	Valtteri Niemi	
Novel uses of human-computer interaction technologies	Humans are creative tool users and able to discover novel solutions for practical problems in their work and everyday life tasks. This creative potential remains however scientifically weakly understood. A master's thesis in in this area may be, for example, about development of interactive prototypes and services for creative practices; studies of artists and/or knowledge workers and their uses of IT in problem solving; or studies aiming to understand human creativity in interactive settings.	Antti Salovaara	https://www.cs.helsinki.fi/u/aksalova/
Studying the future of IT use with interactive prototypes	Computer science is probably the only discipline that is able to create futuristic prototypes and study their use in action, with end users. Master's theses in this area may focus on developing research methods for studying futuristic use contexts with end users or development of futuristic technologies and studying how they could be used.	Antti Salovaara	http://dl.acm.org/citation.cfm?doid=3025453.3025658
Understanding changes in human life through social media and messaging	Social media and messaging services provide a means to understand transformations and evolution of human life. Some domains of knowledge work, such as frontend development, are examples of contexts where knowledge is extremely short-lived: JavaScript frameworks and tools replace each other so quickly that they pose serious challenges for developers to stay up to date. Another context is consumer life: trends such as vegetarian food, healthy living, fashion trends etc. change consumption patterns, and companies have sometimes difficulties in being able to react to rapidly emerging phenomena. However, these challenges in both contexts can be addressed by studying discussions in social media (e.g., developer forums such as Reddit in the case of first example) and chat-based messaging (e.g., Slack), trying to identify features that indicate fluctuations in discussion. Master's theses in this area may involve uses of machine learning to understand messaging-based communication; qualitative and/or quantitative analyses of large text masses; or development of methods to visualize the ongoing large-scale changes in communications.	Antti Salovaara	

Security services in 5G testbed	Department of Computer Science runs a 5G technology testbed that contains 20+ base stations and a core network that is built using virtualization and cloud technologies. Among other things, the testbed allows experiments with new security and privacy services that could be built on top of future 5G networks, or alternatively, as native parts of such networks. An example of such service is extended protection against location and identity tracking of 5G users. The study would identify an interesting potential service, build a simple implementation and run an experiment over the 5G testbed. Finally, results of the experiment are analysed and reported.	Valtteri Niemi	https://www.cs.helsinki.fi/u/aksalova/publications/salovaara-tuunainen-ECIS2015-preprint.pdf https://www.cs.helsinki.fi/u/aksalova/publications/salovaara2013-software-developers-online-chat-author-copy.pdf
Overview of CINCO group thesis topics	This general overview of the CINCO research groups working area indicates three categories of topics: - designing and implementing business services (i.e. applications of interest in any business domain) using service-oriented engineering methods and/or model-driven production tools - renewing Pilarcos ecosystem infrastructure services such as populator, monitoring, binding management (communication channels in similar roles than ESBs), reputation systems, trust management evaluations (implementing decision-making algorithms and comparing to other already tested algorithms); and creating new services such as privacy-preservation facilities - composing new services using collaboration models and the pool of existing services in the ecosystem. **Specific MSc thesis topics available in connection with the CINCOLab resources** 1. Comparative feature and performance analysis on reputation/trust decision-making algorithms 2. Reflective construct of collaboratin contact 3. Open binding factory / Construction of adaptive communication channels 4. Controlling business processes 5. Role of software testing in open service ecosystem environment 6. Engineering tool for composing business processes to nets 7. Modeling language design for business policy (on a particular business area) 8. Business activity monitoring for open service ecosystems 9. Roles of cloud computing in open service ecosystems 10. Realtime rule engine techniques 11. Privacy requirements declaration languages 12. Detecting information leaks across ecosystems	Lea Kutvonen	introductory slideset.
Efficiency of computational service monitoring	In inter-enterprise computing, monitoring of services have a key role in intercepting messaging not confirming to eContracts that govern the collaboration. Monitoring can thus enable privacy-preservation, prevent unfortunate transactions when a collaborating partner proves unworthy of trust, and control quality of service agreements. In inter-enterprise collaboration environments, each party is resposible of protecting its own services and resources, and observing the fulfilment of eContract rules. Therefore, monitors can be bottlenecks in the system, unless carefully designed and placed. The research question of this thesis focuses on the analysis of monitoring cost, utilising both calculus and practical measurements.	Lea Kutvonen	
Tracing ownership of information	Digital rights management is an area where access to information - or rather, media - can be restricted to an identified group of users. The domain of research provides solutions focusing on access control while other solutions focus on marking the target object itself with copyright or ownership information. In the inter-enterprise collaboration environments, an essential aspect is controlling the legitimate and noticing illegitimate flows of information. This thesis should provide a survey of DRM techniques applicable in this kind of environment.	Lea Kutvonen	

Trust calculus / Privacy calculus / Reputation calculus	In service ecosystems, trust and privacy-preservation must be systematically supported by the ecosystem infrastructure - no isolated, partner-wide solution is sufficient. In recent litterature, steps have been taken to address the modeling principles in a formal manner. The purpose of these two topic settings is to analyse this litterature (survey), and critically select the appropriate ones for Pilarcos style ecosystems. The thesis should cover only trust or privacy - or reputation restricted to reputation-based trust systems.	Lea Kutvonen	
Investigating intra-blockchain transactions and network performance	Investigating intra-blockchain transactions and network performance: Understanding the blockchain solutions for decentralised exchanges which enables intra-block chain exchanges. Investigate or simulate the network performance and propose a new routing solution for them. This work requires investigate the source code analysis of different block chain implementations and implementing a new exchange network.	Mohammad Hoque	
Understanding the synchronisation techniques of blockchains	The fundamental backbone of any blockchain network is the distributed consensus mechanism by which a consensus on the order of the blocks in the chain is reached in a distributed fashion. Every participant in the network must have a synchronised view on this order of the blocks in the chain. With the absence of any centralised node participants leverage the consensus algorithm to achieve the required synchronisation. However, the network latency involved on broadcasting any newly created block introduces “forks” at different nodes creating different views, a symptom for inconsistencies among the participants in the network. Hence, it should be avoided whoever possible. Toward this aim, it is important to study different network properties which could minimise forks as much as possible. This work requires investigating the source code different block chain implementations and implementing new syncing mechanism. Knowledge about different synchronisation techniques, such as RSYNC, RDC are plus.	Mohammad Hoque	
Big data analysis on block chain transactions	Our particular interest is Ethereum based tokens (ERC). Analaysing Ethereum transactions and finding out the number of different tokens per block. Identifying incentivise transactions and other interesting facts such as congestion in the network by identifying transaction delay and correlating with real life incidents. And investigate the any kind of patterns in the transaction for some particular patters so the we can predict future fate for some tokens. The student must have taken the Big data framework course and motivated to the take the challenges.	Mohammad Hoque, Sasu Tarkoma	
Privacy-preserving indoor positioning	Indoor positioning is a challenging task because global satellite based systems, such as GPS, are not available. Various methods based on, e.g. WiFi signal strength, have been proposed but in these kind of systems a specific server is needed to calculate user's exact location. This violates user's location privacy. The study is about combining privacy-preserving techniques based on cryptography with state of the art indoor location algorithms.	Kimmo Järvinen	
Edge cloud server placement in wild	Edge cloud computing is an up and coming research field which proposes to place smaller compute cloud servers on network edge such as base station, wireless access points etc. along with a typical centralized cloud. The availability of computing and storage resources near users helps provide low latency for several time-constrained applications such as automated vehicles, drones, etc. One of the prominent research questions that still requires an answer is the placement of these servers in the real world. The master thesis in this area will focus on finding the ideal edge server locations driven by following criteria. 1. use-case latency 2. operation cost 3. availability The study may involve evaluating the findings by integrating real world network traces and available simulators	Jussi Kangasharju	
Implementing cross layer multipath TCP scheduler in Linux	Multipath TCP (MPTCP) is a variant of TCP which aims at allowing a TCP connection to use multiple paths on all available network interfaces such as WiFi, cellular etc. simultaneously. MPTCP aims to maximize available network usage and provide inherent redundancy. In our previous studies, we found that the default scheduler for MPTCP hinders its capability to achieve maximum performance and several locally available system parameters can be leverage to provide a more holistic network connectivity. The master thesis in this area will focus on implementing our QueueAware scheduler in Linux. The study will also involve behavior analysis of several Linux active queue management (AQM) algorithms and their impact on the performance of available MPTCP schedulers.	Jussi Kangasharju	
Critical data detection	Edge Computing is a new computing paradigm where server resources, ranging from a credit-card size computer to a small data center, are placed closer to data and information generation sources. With edge computing we can significantly decrease the volumes of data that must be moved. However, how to identify the critical data that we do not want to move. A thesis project could apply known methods of anomaly or critical data detection to a console log data repository of a distributed monitoring system.	Samu Varjonen, Francois Christophe	

3. Software systems

Title	Abstract	Supervisors	Link

MLOps	MLOps -- as a derivative of DevOps -- is about practice and tools for ML-based systems that technically enable iterative software engineering practice. We have several funded positions in the area of MLOps in our research projects (IMLE4 https://itea4.org/project/iml4e.html and AIGA https://ai-governance.eu/) that can be tailored to the interest of the applicant. For further details, contact Mikko (firs.last@helsinki.fi).	Mikko Raatikainen (Jukka K Nurminen) (Tommi Mikkonen)	
AI System Testing	As machine learning moves from laboratories to real use, good ways to deploy, test, and maintain AI solutions become increasingly important. There are multiple topics about this research theme: • experimentation (e.g. Coverage testing of neural networks**,** where the aim is to study how we can measure which parts of the neural network are activated and look for ideas how to use this information for improved testing and maintenance). • literature review (e.g. Survey of technical/process solutions for AI system testing) • other aspects agreed with the teacher	Jukka K Nurminen	E.g. https://arxiv.org/abs/1705.06640 https://arxiv.org/abs/1812.05389
Computational Moral	When computers are increasingly making decisions, such as who gets a loan or how does an autonomous vehicle behave in case of emergency, the ethical questions of such decision making become important. The thesis could look at technical issues and solutions to assist in ensuring fair decisions. Alternatives include: • measurement study (e.g. investigate tools which detect and correct bias in AI based decision making) • literature review (e.g. techniques to deal with ethical problems in decision making in autonomous systems)	Jukka K Nurminen	https://arxiv.org/pdf/1810.01943.pdf https://www.nature.com/articles/s41586-018-0637-6
Programming of Quantum Computers	Quantum computing is considered a promising direction for efficient solution of e.g. combinatorial optimization problems, which are common in machine learning and operations research. The aim is to look at the issue from practical perspective: what can be done today (e.g. with D-WAVE, IBM-Q), how to formulate the problems for quantum computing, understand what are the main bottlenecks, and what are the most promising future directions. The work could include experimentation with quantum computers and their simulators and software development toolkits.	Jukka K Nurminen	https://www.nature.com/articles/npjqi201523 https://www.research.ibm.com/ibm-q/technology/experience/ https://www.dwavesys.com/take-leap
Cre-at-ively Self-Ad-apt-ive Soft-ware Ar-chi-tec-tures	We have recently started exciting research in the intersection between the research fields of self-adaptive software and computational creativity, with the goal of developing novel software architectures that can creatively adapt themselves in unforeseen situations. This initiative is a new research collaboration between Discovery Group of Prof. Hannu Toivonen and ESE. There are different options for thesis work with either of the groups.	Tomi Männistö (Self-adaptive, architecture), Hannu Toivonen (Computational Creativity)	https://www.helsinki.fi/en/researchgroups/empirical-software-engineering/offered-msc-thesis-topics http://computationalcreativity.net/iccc2017/ICCC_17_accepted_submissions...

Robotics software and software architectures	We are building an interesting line of research in the area of software and software architectures for robotics. This area is an intersection of software engineering and artificial cognitive systems, and takes into account knowledge from different domain areas where robots perform tasks in the physical world. Thesis work in this area can range from more technical and practical to theoretical. The perspectives include both questions about traits of the robotics platform architecture that make development of robotics applications easier and questions about implementing software for robotics systems in different kinds of physical environments.	Niko Mäkitalo	https://www.helsinki.fi/en/researchgroups/empirical-software-engineering/offered-msc-thesis-topics
Open Source Software Development	Open Source Software development is characterised by openly available online collaboration and communication systems. There is a growing body of work examining the data accumulating in such systems. Descriptive studies have examined, e.g., how the development process unfolds and how the social communication structure corresponds to technical actions in the code. Other studies have tried to leverage the the repository data for improving software quality, easing communication, or automating development tasks. Theses in this area could focus on, e.g., analysis of communication patterns using Natural Language Processing techniques, collecting and using software metrics, automated development process support, or methods for analysing specific kinds of repository data. Keywords: Mining software repositories, Open Source Software	Tommi Mikkonen	https://www.helsinki.fi/en/researchgroups/empirical-software-engineering/offered-msc-thesis-topics Guzzi et al., Communication in open source software development mailing lists, in Mining Software Repositories (MSR), 2013. - https://ossmeter.com/
Pro-gram-mable World	The emergence of millions of remotely programmable devices in our surroundings will pose significant challenges for software developers. A roadmap from today's cloud-centric, data-centric Internet of Things systems to the Programmable World highlights those challenges that haven't received enough attention.	Tommi Mikkonen	https://www.helsinki.fi/en/researchgroups/empirical-software-engineering/offered-msc-thesis-topics http://blogs.helsinki.fi/ese-blog/2017/02/09/a-roadmap-to-the-programmable-world-software-challenges-in-the-iot-era/

Di-git-al-iz-a-tion and Di-gital Trans-form-a-tions: Im-pacts on Soft-ware En-gin-eer-ing And Sys-tems De-vel-op-ment	How should digitalization be taken into account in software development processes? What is the role of customer/user involvement in software-intensive systems development (e.g., digital services)? What are the key quality attributes? What new software engineering skills and competencies may be needed? What is the role of software (and IT) in general in different digital transformations (e.g., vs. business process development)? How is digitalization related to traditional software engineering and computer science disciplines in different contexts?	Petri Kettunen	https://www.helsinki.fi/en/researchgroups/empirical-software-engineering/offered-msc-thesis-topics
High Per-form-ing Soft-ware Teams	How is (high) performance defined and measured in software development (e.g., productivity)? Which factors affect it - either positively or negatively - and how strongly (e.g., development tools, team composition)? Can we "build" high-performing software teams in systematic ways, or do they merely emerge under certain favorable conditions? What are suitable organizational designs and environments for hosting and supporting such teams?	Petri Kettunen	https://www.helsinki.fi/en/researchgroups/empirical-software-engineering/offered-msc-thesis-topics https://www.cs.helsinki.fi/node/65141 http://www.cs.helsinki.fi/node/65143
Modeling human brain-signals	Our current interaction with information systems and digital information rely on explicit interaction. Could we mine the interest of the user directly from the human mind? Could unsupervised machine learning methods reveal interesting patterns of neural signals when we are engaged with digital information?	Tuukka Ruotsalo	
Conversational search	There is a gradual shift towards searching and presenting the information in a conversational form. Chatbots, personal assistants in our phones and eyes-free devices are being used increasingly more for different purposes, including information retrieval and exploration. With the recent success of deep learning in different areas of natural language processing, this appears to be the right foundation to power search conversationalization.	Tuukka Ruotsalo	http://augmentedresearch.hiit.fi/
Crowdsourced natural language training data for machine learning	Natural language user interfaces (e.g. chatbots) allow the user to simply talk to the computer, much like they would to another person. To use machine learning for building a natural language UI requires usually very large amounts of training data. This training data consists of pairs of utterances in natural language, and their "meaning", e.g. some user interface action. Generating such training data for specialised applications is challenging, because in the absence of a working system it is often difficult to predict in advance what kind of language the users will use. The objective of this project is to study the use of crowdsourcing techniques, for example "games with a purpose" for collecting such training data without implementing the, possibly costly, final user interface.	Antti Ukkonen	http://dx.doi.org/10.1145/1378704.1378719
Testing and debugging constraint optimization solvers	*See topic description above under "Algorithms" specialization.* This topic is suited for both algorithms and software systems students.	Matti Järvisalo	
Transforming business-level policies to monitoring rules	Inter-enterprise collaborations are governed by eContracts that define what are the required business processes between partners, what business services each partner provides, and especially, what are the nonfunctional properties required in the collaboration. The nonfunctional properties traditionally involve technical quality of service levels, but when enhanced to business area, interesting examples include nonfunctional properties capture trust, privacy, transactionality of interactions, and dependability of service. The collaborations can be controlled through eContracts and enterprises' local policies. Technically, the automated control is performed by low-level monitors. From administrative perspective, the policy rules must be declared using a high-level language - or rather, a family of languages working together. The goal of this thesis is to a) select a small set of languages fitting together and b) implementing transformation tools to translate these administrator designed rules to a low-level language run by the monitors.	Lea Kutvonen	
Performance optimization on big data platform	A cloud-hosted application is expected to support millions of end users with terabytes of data. To accommodate such large-scale workloads, it is common to deploy thousands of servers in one data center. Meanwhile, existing big data platforms (e.g., Hadoop or Spark) employ naive scheduling algorithms, which consider neither heterogeneity of resources nor differences of jobs. This motivates a more advanced scheduling scheme and performance optimization algorithms in big data environments. The goal of this thesis is to a) understand the main scheduling algorithms on Hadoop and Spark; and b) to implement performance optimization tools to improve the performance of the systems.	Jiaheng Lu	http://udbms.cs.helsinki.fi/?projects/hetePlatformAAwe

Artifact Recognition Based on Images Captured by Vision System in Autonomous ships or harbour	Cameras are used in state-of-the-art autonomous systems in order to get detailed information of the environment. However, when cameras are used outdoors they easily get dirty or scratches on the lens which leads to image artifacts that can deteriorate a system's performance. Work has been done on how to detect and cope with specific artifacts, such as raindrops. A thesis project could apply known methods of weather recognition or artifact recognition to a repository of maritime images and review and compare different approaches. Can be extended to include the thermal images in the repository.	Samu Varjonen, Francois Christophe	