LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY
INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

9th International Workshop on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Druskininkai, Lithuania, Hotel "Europa Royale"
http://www.mii.lt/DAMSS

November 30 –December 2, 2017

Co-Chairmen:

Dr. Saulius Maskeliūnas (Lithuanian Computer Society)
Prof. Gintautas Dzemyda (Vilnius University, Lithuanian Academy of Sciences)

Programme Committee:

Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Hele-Maja Haav (Estonia)
Prof. Albertas Čaplinskas (Lithuania)
Prof. Ignacy Kaliszewski (Poland)
Prof. Yuriy Kharin (Belarus)
Prof. Tomas Krilavičius (Lithuania)
Prof. Julius Žilinskas (Lithuania)

Organizing Committee:

Dr. Jolita Bernatavičienė
Dr. Olga Kurasova
Aidas Žandaris

Contacts:

Dr. Jolita Bernatavičienė
`jolita.bernataviciene@mii.vu.lt`
Dr. Olga Kurasova
`olga.kurasova@mii.vu.lt`
Tel. (8~5) 2109 312

DOI: http://dx.doi.org/10.15388/DAMSS.2017

ISBN 978-9986-680-64-2

Preface

DAMSS-2017 is the 9th international workshop on data analysis methods for software systems, organized in Druskininkai, Lithuania, at the end of the year.

History of the workshop starts from 2009 with 16 presentations. The number of this year presentations is 68. The idea of such workshop came up at the Institute of Mathematics and Informatics. Lithuanian Academy of Sciences and the Lithuanian Computer Society supported this idea. This idea got approval both in the Lithuanian research community and abroad.

This year is special for the initiator of DAMSS conferences. On 2010, the Institute of Mathematics and Informatics became a member of Vilnius University, the largest university of Lithuania. Now, the institute changes its name into the Institute of Data Science and Digital Technologies. This name reflects activities of the institute much better. The renewed institute has seven research groups: Cognitive Computing, Image and Signal Analysis, Cyber-Social Systems Engineering, Statistics and Probability, Global Optimization, Operations Research, Education Systems.

The workshop has several main goals: to facilitate scientific networking that will lead to joint research projects and to initiate connections with foreign research institutions and scientists. Seeking to facilitate relations between science and business, computer science and IT business community will be introduced with research undertaken at Lithuanian and foreign universities in the fields of data science and software engineering.

Topics of the conference cover big data, bioinformatics, data science, deep learning, digital technologies, high-performance computing, visualization methods for multidimensional data, machine learning, medical informatics, ontological engineering, optimization in data science, business rules, and software engineering.

This book gives an overview of all presentations of DAMSS-2017.

Supported by:

General sponsors

Research Council of Lithuania
www.lmt.lt

Algoritmų sistemos
www.algoritmusistemos.lt

Main sponsor

BAIP
www.baip.lt

Sponsors

Baltic Amadeus
www.baltic-amadeus.lt

NetCode
www.netcode.lt

NRD CS
www.nrdcs.lt

Automatic Adjustment of Disparity in Stereo View by Eye Activity Based Markers

V. Abromavičius, A. Serackis, D. Navakauskas

Vilnius Gediminas Technical University
dalius.navakauskas@vgtu.lt

The presented investigation focus on eye fatigue problem. The problem arises during sustained visual perception of stereoscopic views. The high importance of this problems is supported by currently increased attention to the head mounted displays (virtual reality and augmented reality glasses) and immersive visual content (e.g. 360 degree video records). A stereoscopy effect is achieved by presenting individual views for each eye. The disparity between two views gives a possibility to support better depth perception. The higher disparity between image objects helps to distinguish between similar objects situated at different distances from the viewer. However, high disparities are related to the higher visual discomfort levels and may cause the eye fatigue during long lasting stereoscopic perception. The aim of the investigation is to extract multimodal data features, related to discomfort during stereoscopic perception of images with objects with different disparity level on the focus point. An experimental investigation with a control group was performed using special 120 stereo image dataset. The features were extracted from eye tracking, pupil size monitoring, EOG and EEG real-time signal measures. As a result of the investigation, the dependences of object disparity to dynamic changes of pupil size were estimated. A new technique was proposed for the identification of the moment from which the correct stereo perception is achieved on the image object with different level of disparity.

N Short Overview of Machine Learning Methods Used in Network Intrusion Detection Systems

L. Mišauskas, V. Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University
liudas.alisauskas@mii.vu.lt

Operation of typical network intrusion detection systems relies on signature and/or anomaly-based intrusion detection methods. Signature-based methods are strong at detecting already known intrusions, while anomaly-based methods are used to detect unknown network intrusions.

Results of signature-based methods directly depend on given intrusion recognition data features (e.g., patterns, rules, indicators of compromise), that basically are discovered from already detected and known network intrusion data. Usually, signature-based methods are computation resource effective and achieve high intrusion detection accuracy, but this type of methods cannot detect not known intrusions.

To solve not known intrusion detection problem, specific anomaly-based intrusion detection methods are used. The principle of anomaly-based methods is based on a comparison of normal network behavior model (provided or built from network flow observations) and online network behavior model (built from online network flow). Even though this type of methods uses more computational resources and results are probability based, this type of methods has proved themselves as able detect yet unknown network intrusions.

In this work, we present the results we have obtained from the research with the objective to overview and analyse machine learning methods that are used for anomaly detection in network intrusion detection systems in order to describe environments and tasks under which appropriate state-of-art methods perform the best.

IoT Mobile Network Node's Velocity Estimation via Curve Fitting

K. Bagdonas, A. Venčkauskas

Faculty of Informatics
Kaunas University of Technology
kazimieras.bagdonas@ktu.lt

We present a novel way to estimate the velocity of a linearly moving IoT network node using a single ranging source. Our proposed method is compatible with any ranging technique, that is capable of providing sufficient accuracy and sampling rate. Proposed method is able to mitigate bias in ranging estimation. We provide analytical proof for our method and a simulated study of Signal-to-Noise ratio, sample rate, and networks node's velocity influence on the accuracy of velocity estimation, based on curve fitting.

Computational Modelling of Active Suspensions of Growing Bacteria

R. Baronas[1], Ž. Ledas[1], R. Šimkus[2]

[1] Faculty of Mathematics and Informatics
Vilnius University
[2] Institute of Biochemistry
Vilnius University
romas.baronas@mif.vu.lt

Active colloids are small scale particles which are able to move autonomously through viscous fluids by converting energy extracted from their environment into directed motion.

Chemical-reaction-driven Janus particles and living bacteria are among the most prominent examples of artificial and natural active colloids. The direct movement of bacteria toward the highest concentration of food molecules and to flee from poisons is called chemotaxis and plays an important role in a wide range of biological processes.

In this work, the self-organization of Escherichia coli cells, as detected by bioluminescence imaging, was computationally modeled by the non-linear Keller-Segel equations of chemotaxis with logistic cell growth.

Software to extract pseudo-one-dimensional spatiotemporal plots from the two-dimensional images and to isolate and count high concentration areas-aggregates was developed, and the quantification of the self-organization process by counting the average number of unstable aggregates was performed. Signal processing techniques were used to determine the number of aggregates. Initially, the bioluminescence images were pre-processed for noise reduction by the one-dimensional Gaussian filter. Then, the Otsu's method was used to determine a threshold for each line. With the applied threshold, the number of highlighted areas at each time step was calculated. Lazarus integrated development environment and Free Pascal were used to develop this software. The computational results support the idea that E. coli can be viewed as a model of active colloids.

This research was funded by a grant (No. S-MIP-17-98) from the Research Council of Lithuania.

Heuristic Extension for Multi-Criteria Decision Making in Autonomous Ground Robot Navigation

R. Baušys, R. Semėnas

Vilnius Gediminas Technical University
rokas.semenas@vgtu.lt

Autonomous ground robot navigation is an important research field that has many real-life applications. Over the past decade human-uncontrolled search and rescue, exploration and area mapping problems grew in numbers, but there is still a lack of robust and optimal algorithms to tackle them.

Many different approaches were suggested, and among them Multi-criteria Decision Making methods show feasible results and huge potential, considering the amount of different parameters that can be presented to the robot. In this paper, the main algorithms and problems of so-called, state of the art methods are presented and discussed, several possible heuristics to extend them are proposed.

Combined Permutation Tests and Finite Sample Consistency

Stefano Bonnini

Department of Economics and Management
University of Ferrara
stefano.bonnini@unife.it

In several applications, the dataset presents a large number of variables and small sample sizes. A typical solution for a location problem where two multivariate distributions are compared, is the Hotelling's T-square test. It requires, among the other assumptions independent samples and multivariate normality. In the presence of small sample sizes, Hotelling's T-square test is not very powerful, and it cannot provide a suitable solution for one-sided (directional) alternative hypotheses.

Combined Permutation Tests represent a nonparametric proposal in order to deal with a large number of variables, especially in the presence of small sample sizes. This testing method mainly requires exchangeability under H0, and it is distribution free. Thus it is much more flexible and robust than other testing methods, especially parametric solutions (e.g., Hotelling's T-square test). This method is suitable for both two-sided and one-sided alternatives and with numeric, categorical or ixed variables.

Finite sample consistency is one important property of this methodology. According to this property, under mild conditions, the power of the test increases with the number of variables under study. The result holds even when the number of variables is much larger than the sample sizes.

It has been proved that the, for the multivariate two-sided problem, the combined permutation test is more powerful than the Hotelling T2 test. Some Monte Carlo simulation studies prove that similar results hold under Normal, Cauchy, Student's t and Pareto distributions. Thus it is robust with respect to the underlying multivariate distribution. The test is well approximated under H0, unbiased and consistent (in the classic sense) when sample sizes diverge.

We present some advances in the study of the mentioned methodology and property, through the results of a simulation study and an application example.

An Investigation of Early Cyber Threat Detection Using Ensembles of Machine Learning Methods

V. Bulavas, G.Dzemyda, V. Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University
viktoras.bulavas@ittc.vu.lt

According to PwC's Global economic crime survey, cybercrime has overall evolved into the second place after an asset misappropriation. According to Lithuanian National Cyber Security Centre Annual report for 2016, scanning of surveilled network devices since 2015 has increased fivefold. Lithuanian academic network LITNET is no different, observing persistent multiple step intrusion activities.

As nowadays it is impossible to detect and mitigate all threats manually, automatic tools are used on a 24/7 basis. The techniques utilized by current network intrusion detection appliances in use fall into three main categories: anomaly detection, misuse detection, and hybrid. Misuse detection systems use signatures that describe already known attacks and require regular ruleset update.

Machine learning based anomaly detection requires supervision and specialist review due to currently still high false positive rate of detecting previously unseen system behaviors. With a n increasing frequency of cyber-attack, reviews take more and more time of cyber security specialists, which is a challenge. This indicates the highly demanded area for research aiming to increase threat detection accuracy and training

speed. Until very recently there was little published research about successful early threat detection models.
Other authors proposed an ensemble of Machine Learning models as a probable way of solving abovementioned early detection problems. Therefore, in this work, authors perform an investigation of selected method ensembles and present results of the comparison.

Application of Machine Learning for MWE Identification

I. Bumbulienė[1], J. Mandravickaitė[1, 2], T. Krilavičius[1, 2]

[1] Baltic Institute of Advanced Technology
[2] Vytautas Magnus university
`t.krilavicius@bpti.lt`

Identification of Multiword Expressions is an important problem in Natural Language Processing, especially for machine translation and other semantic analysis tasks. Often, lexical association measures (LAM), such as pointwise mutual information (PMI), log likelihood ratio (LLR), Dice are used to identify MWE's. However, just LAMs are insufficient for MWE detection, especially for Lithuanian language, but could be very useful as additional features for Machine Learning (ML) algorithms. Early experiments with Lithuanian and Latvian languages show that using Random Forest with Resample filter, we can achieve almost 99% precision, 58% recall and 73% F-score.
We discuss experiments with delfi.lt based corpora, different features, including LAMs, as well as experiments with different ML methods, i.e., Naive Bayes, Random Forests, Support Vector Machines, Artificial Neural Networks and others.

High-Performance Management and Analysis of Omics Data: Experiences at University Magna Graecia of Catanzaro

Mario Cannataro

Data Analytics Research Center & Bioinformatics Laboratory
Department of Medical and Surgical Sciences
Università Magna Graecia di Catanzaro, Italy
cannataro@unicz.it

Genomics, proteomics, and interactomics are gaining an increasing interest in the scientific community due to the availability of novel platforms for the investigation of the cell machinery, such as mass spectrometry, microarray, next-generation sequencing, that are producing an overwhelming amount of experimental omics data.

This large volume of omics data poses new challenges both for the efficient storage and integration of the data and for their efficient pre-processing and analysis. Moreover, both raw experimental data and processed data are more and more stored in various databases spread all over the Internet, not fully integrated.

Thus, managing omics data requires both support and spaces for data storing as well as novel algorithms and tools for data pre-processing, analysis, and sharing. The resulting scenario comprises a set of methodologies and bioinformatics tools, often implemented as services, for the management and analysis of omics data stored locally or in geographically distributed biological databases.

The talk describes some parallel and distributed bioinformatics tools for the pre-processing and analysis of genomics, proteomics and interactomics data, developed at the Bioinformatics Laboratory of the University Magna Graecia of Catanzaro. Tools for gene expression and genotyping (SNP) data analysis (e.g., micro-CS, DMET-Analyzer, DMET-Miner, OSAnalyzer, coreSNP) as well as for proteomics data analysis (e.g., MS-Analyzer, EIPeptiDi) will be briefly underlined.

Investigation of Self-managed Mircroservice Architecture

A. Cholomskis, D. Mažeika

Vilnius Gediminas Technical University
aurimas.cholomskis@vgtu.lt

Cloud computing made a big impact on software architecture evolution. The demand to serve multiple tenants, to include continuous delivery practice into the development process as well as increased system load influenced the style of cloud based software architecture. Microservice architecture is preferred architecture despite its complexity when scalability is an essential attribute of quality of service. Microservices should be managed, i.e., hardware resources should be adjusted based on application load, as well as resiliency should be ensured. Popular IaaS and PaaS providers such as Amazon, Azure or OpenStack ensure auto-scaling and elasticity at the infrastructure level. This approach has the following limitations: (1) Scaling and resiliency is a part of the infrastructure and not emerging from application nature; (2) The software is locked in with a specific vendor; (3) It might be difficult to run and ensure smooth scalability by running software on different vendors at the same time.

The self-managed microservices concept was introduced to deal with these limitations. The concept of self-managed microservices is extended and evaluated in more details in this manuscript. Microservice architecture is established which enables to build microservices in a self-managed way. Frameworks, techniques, and tools are proposed to build such architecture that allows monitoring performance indicators and health status, reacting and executing auto-scaling instructions or ensuring recovery from failures.

Formalising the Concept of Taboo as a Prohibition to Speak

V. Čyras[1], F. Lachmayer[2]

[1] Faculty of Mathematics and Informatics
Vilnius University
[2] Faculty of Law, University of Innsbruck, Viena
vytautas.cyras@mif.vu.lt

A taboo is a prohibition of an action based on the belief that such behaviour is either too sacred or too accursed for ordinary individuals to

undertake. In the talk, we formalise taboo as a prohibition to speak (in general, to inform). An example is in Hans Christian Andersen's tale "The Emperor's New Clothes." When the Emperor parades before his subjects in his new clothes, a taboo is to say that they do not see any suit of clothes on him for fear that they will be seen as "unfit for their positions, stupid, or incompetent". Finally, a child cries out, "But he is not wearing anything at all!" We distinguish three levels of normative prohibitions. First, come basic prohibitions, Forbidden X, i.e., norms which prohibit basic actions, Norm(¬X), e.g., "No smoking". Second level norms consist of primary taboos which prohibit to inform about a fact or fake but permit it to happen, Norm(¬Inf(X)). An example is a soiree, where it is prohibited to speak about money. However, it is not prohibited to have money. Third level norms consist of meta-taboos which prohibit to inform that a primary taboo exists, Norm(¬Inf(Norm(¬Inf(X)))).

Taboos may be sensitive morally, religiously, culturally, socially, politically, legally. Taboo can have different social reasons such as top-down institutional repression, odd morality, etc. Taboo norms may be evaluated negatively. However, various positions are discussed, e.g., informing a bandit, informing an abusive government, and informing a just government. The view "No prohibition to inform when the government is just" has a place. Officials can camouflage a primary taboo on the essential causes A of an effect E with a fake relationship between certain facts B and E. We focus on modelling statements about taboos and present a conceptualisation. Exploring the function of taboos and social reasons are out of scope.

LitWordNet Ontology Learning from Corpus Texts

J. Dainauskas, D. Vitkutė-Adžgauskienė

Faculty of Informatics
Vytautas Magnus University
daiva.vitkute-adzgauskiene@vdu.lt

In order to design LitWordNet ontology (a Lithuanian version of WordNet lexical ontology) both selective integration of relevant already existing dictionaries and semantic networks, as well automatic and semi-automatic information extraction from the unstructured texts from the Corpus of Contemporary Lithuanian Language was used.

The LitWordNet is built using RDF/OWL format, and its structure is based on synsets (synonym rings), linked by corresponding semantic relations, such as hypernyms, hyponyms, and meronyms. A combination of different methods was used for ontology learning from texts – Clustering by Committee (CBC) and Near-Synonym System(NeSS) for sense recognition when assigning word senses to synsets, pattern-based methods for recognizing semantic relations and building the hierarchical structure of the ontology.

The research showed that close to exact semantic relations can be extracted from copora using mentioned algorithms. As a result, it was also possible to enhance the ontology structure by adding contextual words for synsets, and thus facilitating the use of ontology in further text analysis tasks. The collection of methods and supporting computerized tools for Lithuanian language text mining, that is presented as a result of this project, enables periodic update of the LitWordNet ontology contents, whenever new corpus resources become available.

The use of corpus resources of contemporary Lithuanian language also creates the advantage of constant updating of ontology with words that are used in everyday language. Currently, LitWordNet ontology contains more than 45000 unique synsets. It can be browsed via an online interface at http://mackus.vdu.lt/LitWordNet.

A Marine Traffic Prediction using Recurrent Neural Networks

A. Daranda, G. Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University
andrius.daranda@gmail.com

Nowadays various sensors embedded in vessels and the analysis of the marine traffic becomes more common and straightforward. The navigational data comes from the vessel's sensors like Global Positioning System, Automated Identification systems and etc. The navigational data which was used in the experiments was obtained from all vessels based in the North Sea and these data was collected for two months period. In this paper, we present a deep learning-based approach to predict the route of the vessel. This is extremely important because even modern navigation devices could not provide full navigation picture in the certain situations. Moreover, the huge problem is to predict the vessel's route and to find certain patterns movements if do not know the ship's particular parameters or port of destination. To solve these problems, we used the Recurrent Neural Networks (RNN). RNN are a type of artificial neural network designed to recognize patterns in sequences of data. Their application is quite wide for cases, where a primary data set can be transformed into a set of sequences. RNN use various types of hidden units, e.g. long short-term memory unit, gated recurrent unit. We apply the gated recurrent unit. By exploiting this capability the prediction model was implemented which was trained and tested on more than 150.000 points which includes the various navigational data. The proposed system was created by High-Level TensorFlow Machine Learning Framework. This allows enhancing the vessel route prediction and ensuring correct predictions in the dynamic and volatile marine environment.

Speaker Identification Using Deep Neural Networks

L. Dovydaitis, V. Rudžionis

Institute of Applied Informatics
Kaunas Faculty, Vilnius University
laurynas.dovydaitis@gmail.com

We present speaker identification test results on Lithuanian speakers database LIEPA.
In this speaker identification task, we extract MFCC features from speakers voice sample. Then speaker acoustic model is created using deep neural networks. Identification accuracy comparison is made between deep neural network model versus hidden Markov model.
We conclude that deep neural network helps to achieve from 3% to 6% increase in identification accuracy on LIEPA dataset.

Bayesian Approaches to Multiobjective Optimization and its Integration into Industry 4.0 Frameworks

Michael Emmerich

LIACS, Faculty of Science
Leiden University, Netherlands
m.t.m.emmerich@liacs.leidenuniv.nl

The talk deals with the utilization of big and small data sets on evaluations of systems to find improvements and making better decisions. Decision making is typically based on multiple objectives, and the goal is to find solutions that are efficient (avoiding lose-lose scenarios) and that represent interesting alternatives.

In systems optimization, control parameters and configurations are searched for that yield improvement with respect to the current state. As an example, we will look at industrial processes from steel production, chemical production processes, and biogas plant control, and discuss how to find optimal process setups and control strategies concerning multiple objectives (robustness, average performance, environmental impact, etc.). One of the challenges within the so-called "Industry 4.0" frameworks is to integrate advanced measurement data and simulation data in decision support systems. Systems optimization forms an essential ingredient of this. Techniques from Bayesian global optimization that originated in Lithuania in the 70ties appear to offer mathematically rigorous, yet flexible, methodologies.
The talk will focus on new developments of such techniques and their generalization to big data sets and multiobjective optimization, which makes them viable components in "Industry 4.0" solutions..

Tests of Independence Based on Multivariate Distance Correlation Matrix

Konstantinos Fokianos

Department of Mathematics & Statistics
University of Cyprus
fokianos@ucy.ac.cy

We introduce the notion of multivariate auto-distance covariance and correlation functions for time series analysis. These concepts have been recently discussed in the context of independent and time series data, but we extend them in a different direction by putting forward their matrix version.We discuss their interpretation, and we give consistent estimators for practical implementation. Additionally, we develop a test for testing the iid hypothesis for multivariate time series data. The proposed test statistic performs better than the standard multivariate version of Ljung-Box test statistic. Several computational aspects are discussed, and some data examples are provided for illustration of the methodology.

Classification of Motor Imagery Using PCA Features for Brain-Computer Interfaces

Sam Gilvine Samuvel
Kaunas University of Technology
gilvine21@gmail.com

A Brain Computer Interface (BCI) is an alternative pathway for the human brain to communicate with the outside world. Here, BCI can act as a decoder. The decoder process the Electro Encephalon Gram (EEG) signal obtain from the brain activities and then send a command or intent to an external interactive device, such as a wheelchair. In the work, the proposed model to implement an algorithm that would be able to classify four different motor imagery tasks, using electroencephalogram (EEG) data from the BCI Competition can test three feature extraction techniques: band power features, time domain parameters and PCA features with two classifiers namely linear discriminant analysis (LDA) and support vector machines (SVM). The effect is compare to LDA classifier the best results for most subjects were achieved when using the SVM classifier with band power features, Time domain features with PCA.

Image Quality and Radiation Dose of Low Dose Cardiac Cta in Obese Patients

D. Golubickas, A. Jankauskas, R. Slapikas, A. Basevičius,
K. Morkunaite, V. Veikutis

Lithuanian University of Health Sciences
domas.golubickas@lsmuni.lt

Cardiac Computed Tomography Angiography (CCTA) advances in better protocols permit lower radiation dose during cardiac imaging procedure. In individuals with the suitable body size, a lower tube voltage setting may allow for a decreased radiation dose while maintaining diagnostic image quality.

Investigate the effect of low kilovoltage CCTA on qualitative and quantitate image quality parameters and radiation dose.

We sought to investigate feasibility of 100 kV CCTA in obese adult patients group (BMI 30-35 kg/m2) by comparing radiation doses and image quality versus standardized 120 kV protocols. Seventeen patients were examined using a tube potential of 100 kV (mean age 54.5±8.3

years, mean BMI 32.15±1.6). 28 patients (mean age 58.3±9.5 years, mean BMI 31.9±1.8) matched for body-mass-index, heart rate was scanned with a tube potential of 120 kV and served as the control group. Qualitative and quantitate image quality parameters were determined in proximal and distal segments of the coronary arteries. Image quality was determined by two blinded readers using Likert scale. Quantitative assessment was determined by the contrast-to-noise ratio (CNR) and signal-to-noise ratio (SNR). The differences between the groups were compared using two-tailed Student ‘test. To determine inter-observer agreement for the qualitative image quality assessment, intra-class-correlation (ICC) and Spearman correlation coefficient were calculated. A p-value less than 0.05 was considered statistically significant.
A total of 360 segments (136 segments [100 kV] versus 224 segments [120 kV]) were assessed qualifiedly. At 100 kV, 132/136 segments and at 120 kV, 220/224 segments were deemed diagnostic. The comparison of qualitative image quality by two observers showed very good agreement with ICC of 0.9. The noise level in patients scanned with a tube 100 kV was not significantly higher compared to the patients scanned at 120 kV (34±5.9 versus 33.9±6.5 HU). The mean contrast-to-noise ratio (CNR) and signal-to-noise ratio (SNR) were not significantly higher at 100 kV versus 120 kV (CNR 10.8±3.8 [100kV] vs. 10.4 ±3.9 [120 kV], $p<0.05$ and SNR 13.2±3.8 [100 kV] vs.12.6±4.1 [120 kV], $p<0.05$). In this study, we tested feasibility of 100 kV protocols and demonstrated a significant reduction of radiation dose of 32% versus 120 kV protocols (1.7 mSV [1.2-2.1] versus 2.5 [1.5-3.6]). Our data demonstrates that in obese patients’ low dose protocol is feasible and results in radiation dose reduction of 32%. Image quality was found to be diagnostically acceptable in all cases.

Digital Evidence Object Model for Cybercrime Investigation

Š. Grigaliūnas, E. Toldinas

Faculty of Informatics
Kaunas University of Technology
sarunas.grigaliunas@ktu.lt

In cybercrime investigation, theoretical methodology and practical tools have become two essential technologies. Theoretical methodologies define the models to investigate the cybercrime, and practical tools provide abilities for investigators to extract the evidence to prove the

crime. Known models integrate knowledge of experts from the fields of digital forensics and software development, use events reconstruction, automatic knowledge extraction, and preservation of data integrity. The aim of our research is to propose digital evidence object model that combines the crime investigation process with the object oriented programming model informatively. We propose a novel digital evidence object (DEO) model that is defined as DEO=(Why, When, Where, What, Who). In the DEO model Why=(cd,ie,fi,cpv,ca) is formally defined by a set of five variables: cd - Criminal Damage; ie - Industrial Espionage; fi - Financial Investigations; cpv - Corporate Policy Violation; ca - Child Abuse. The second element When=(t, Δt, T, S) is an event of DEO characterization, where T - how long and S - how soon to investigate DEO. The third element Where=(s, p) - s is source and p is a place of DEO. The fourth element What=(FIevent), FIevent describe something that occurs in a certain place during a particular interval of time. The fifth element Who=(Person, Process) identifies an object of the crime.
The proposed model provides a methodology for digital investigation minimizing investigation cost and time. The DEO model can be directly mapped into a tool, applicable to investigate the various type of cyber threat.

Knowledge Granularity in the Enterprise Information Systems

S. Gudas

Institute of Data Science and Digital Technologies
Vilnius University
saulius.gudas@mii.vu.lt

A new and rapidly growing paradigm of information processing – granular computing and modelling is an umbrella term to cover any theories, methods and tools that make use of information / knowledge granules in complex problem solving. The concepts of information granules and knowledge granules are discussed in the context of enterprise software systems engineering. The principles of extracting granular structures in enterprise domain are presented and illustrated with examples. The granularity levels of knowledge and information from the perspective of enterprise management are being examined.

Satellite Imagery Application to Financial Markets via Machine Learning

P. Gudžius, O. Kurasova, E. Filatovas

Institute of Data Science and Digital Technologies
Vilnius University
gudzius@gmail.com

Due to enhancements in nano-satellites hardware technology satellite imagery data is growing exponentially and follows Moore's law. There are more Cube-satellites launched to Low Earth Orbit in 2017 than in the previous ten years combined. Improvements in Geo-spacial Data (GSD) frequency, quality, coverage, and wave-range combined with Machine Learning (ML) technology enables us to calculate global oil reserves, track tanker ships, forecast wheat yields or estimate retail revenue based on car traffic – all spectacularly valuable financial information.

This research will focus on ML techniques used to process GSD pixel datasets with trillions of data points. Such techniques as Support Vector Machines, Convolutional Neural Networks, Neuro-fuzzy and Genetic Algorithms will be explored in order to discover the most effective ones with largest future potential for generic application. In-depth research will then be conducted in order to apply these techniques to image processing and object recognition.

The core objective is to achieve an optimal and computationally efficient image recognition output and apply it to financial markets.

Challenges in Metric-based Identification of Critical Software Components

Marjan Heričko

Institute of Informatics
Faculty of Electrical Engineering and Computer Science
University of Maribor
marjan.hericko@um.si

Knowing the threshold of reliable software metrics can contribute to product quality evaluation and, consequently, increase the usefulness of software metrics in practice. In order to deliver a software product with the required quality, it is crucial to address and manage the quality assurance domain properly. Software metrics can be used as a reflection of the qualitative characteristics of software components in a quantitative way, presenting a control instrument in the software development and maintenance process. Software metrics assess software from different views, but overall, reflect the internal quality of software systems. The usefulness of metrics without knowing their reference values is very limited, due mainly to interpretation difficulties. To overcome the above-mentioned difficulties, it is important that reliable reference/threshold values of software metrics are available. Thresholds are heuristic values that are used to set ranges for desirable and undesirable metric values for measured software and, furthermore, used to identify anomalies, which may be an actual problem. Threshold values explain if a metric value is in the normal range and, consequently, provide a scale for paying attention to components that exceed the threshold. There are many approaches available to compute threshold values. In this presentation we are going to address some challenges related to threshold derivation approaches and their application on practical projects, i.e. finding a representative benchmark data; addressing different statistical properties of software metrics; using a suitable statistical approach for threshold derivation; applying the metric threshold values in practice to different code context and maximizing the intersection between the sets of detected code-smells within different tools.

BOINC Based Enterprise Desktop GRID

A. Jurgelevičius, L. Sakalauskas

Institute of Data Science and Digital Technologies
Vilnius University
albertas.jurgelevicius@mii.vu.lt

Volunteer computing is a type of distributed computing in which so-called volunteers provide computing resources to projects. Berkeley Open Infrastructure for Network Computing (BOINC) is the standard software for most volunteer computing projects and provides the means for organizations to adopt IT services with little cost. Research show that public distributed computing has the required potential and capabilities to handle big data mining and other computational resource demanding tasks, however, businesses and organizations are not considering adopting this model until security, reliability concerns and challenges associated with it are solved. While these issues are researched and solved in cloud computing applications, these problems remain open in distributed public computing models. We identified the fundamental security and reliability issues in distributed public computing model, preventing businesses and organizations from adopting this model and making them use less affordable and legal issues involving cloud computing model. BOINC based enterprise desktop GRID has the potential to be widely used for solving today's computational challenges from business stand point perspective if adoption issues are solved. We propose a BOINC based platform that can become a foundation for future research required to bring back attention to this low cost public distributed computing model and make it a suitable platform for business companies and organizations for solving computational resource demanding tasks.

Political Protest Groups and its Rhetoric's on Facebook

Rasa Kasperienė

Vytautas Magnus University
kasperiene@gmail.com

In the recent years, the European Union is witnessing the growth of radical political communities in Europe and Lithuania as well. Current events have shown that in Europe movements of the extreme left and right political groups are expanding into a larger epicentre. For example, organisation Soldiers of Odin in Finland. This organisation is anti-immigrant street patrol group was established as a response to thousands of asylum seekers arriving in Finland. Now the group also has a presence in Sweden, Norway, Estonia, Canada, and Denmark. In Lithuania, as in other EU countries, the growth of right and left wings groups is also noticeable.

Voices of leaders of Lithuanian political alternative groups are getting into public space that transfers their ideology. Most of them are against the EU, but in favour for "strong Lithuania." The others (for example, leaders of so-called pro-Russian groups) are against the EU, NATO and current Lithuanian authority and government, but encourage Russian politics and is for strong Lithuania too. Sometimes leaders' rhetoric progresses from radical into more revolutionary. For example, Žilvinas Razminas in real and virtual space is inviting people to "start a social revolution against occupation regime in Lithuania" (Razminas: 2015).

Political processes that are developing in real space have a reaction in virtual space too. Members and followers of political protest groups are acting mostly on the Facebook platform via virtual groups. The aim of this study ideological beliefs in terms of network and information of Lithuanian political protests groups on Facebook as well as what factors are important for words Lithuania, land, ES, NATO, JAV, Russia, war, Revolution to appear. Also, another aim of this study is clarify the members belonging to other Facebook political protests groups (social network mapping).

In this research is used interdisciplinary approach at this study. Data of 10 political protests groups on Facebook was downloaded using data scraping technology. We analysed more than 87868 posts dating from 2010 till now. Computer modulated text analysis (concordances, keywords, group behaviour dynamics) showed that members of pro-Russian groups support rhetoric, an ideology that is similar to the ones of members of nationalists Facebook groups.

Finally, social network analysis of selected Lithuanian political protest groups on Facebook showed that many members of pro-Russian groups belong to nationalists’ Facebook groups and vice versa.

Pareto Suboptimal Solutions to Large-Scale Multiobjective Multidimensional Knapsack Problems with Assessments of Pareto Suboptimality Gaps

Ignacy Kaliszewski[1, 2]

[1] Systems Research Institute
Polish Academy of Sciences
[2] Warsaw School of Information Technology
ignacy.kaliszewski@ibspan.waw.pl

In this work, we report on our pursuit for providing a methodological support for approximate multiobjective optimization methods. We present our recent theoretical results which identify a class of mulitobjective optimization problems for which so called upper shells can be easily generated. With upper shells, two-sided bounds on component values of approximate solutions can be calculated with a negligible effort. We discuss the practical applicability of the results to solving large-scale mulitobjective optimization problems. We point to enormous potential savings of computations if we are ready to accept approximate solutions of reasonable accuracy instead of exact ones, where accuracy is represented by the Pareto suboptimality gap, i.e., the difference between the corresponding upper and lower bounds of the two-sided bound pair.

The development has been tested on a set of large-scale multiobjective multidimensional knapsack problems solved by a commercial mixed-integer linear package.

Identification and Control of Human Response to 3D Face Stimuli Using Virtual Reality

V. Kaminskas, E. Ščiglinskas

Department of System Analysis
Vytautas Magnus University
vytautas.kaminskas@vdu.lt

This paper introduces how predictor-based control principles are applied to the control of human response – excitement signal. We use changing distance-between-eyes in a 3D face as a stimulus – control signal. The human 3D face is observed in virtual reality. Human responses to the stimuli are observed using EEG based signal that characterize excitement. Predictive and stable model building method with the smallest output prediction error is proposed. A predictor-based control law is synthesized by generalized minimizing minimum variance control criterion in an admissible domain. The admissible domain is composed of control signal boundaries. Relatively high prediction and control quality of excitement signals are demonstrated by modelling results. Control signal boundaries allow decreasing variation of changes in a virtual 3D face.

Fractal Dimensions of Speech Signals

R. Karbauskaitė, G. Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University
rasa.karbauskaite@mii.vu.lt

The linearized analysis in a processing of speech signals had a great success and found a wide application. Various coding, analysis, and synthesis techniques were proposed assuming a linear nature of the speech signal. However, the speech production process is inert and non-linear by its nature and determines such phenomena as co-articulation, a variation of the speaking rate and the fundamental frequency. In a phonetic level, we can observe the reduction, elision, assimilation, and accommodation of sounds. The description of all these properties using linear techniques becomes complicated. Therefore, different nonlinear

techniques have been proposed for the analysis and modelling of the speech signal: nonlinear autoregressive and autoregressive-moving average predictors, polynomial approximation techniques, energy operators, various modulation types, chaotic models, and fractal-based methods also.

A fractal can be defined as an abstract mathematical object that describes a particular set or sequence of values. The fractal reveals pattern-based fragmentation, self-similarity, and self-affinity of the analysed set. These fractal properties can be characterized using a fractal dimension (FD) value. In the case of a speech signal, the fractal dimension is capable to estimate the irregularity level or turbulence degree of the signal.

In this study, we have explored various fractal dimension techniques for a description of the speech signal. Katz, Higuchi, Castiglioni, and Hurst exponent-based dimensions were analysed. The two-level speech signal analysis was proposed using the fractal dimension value. The first level treated frame-level fractal properties of the speech; the second level analysis was dedicated for utterance-level or supra-segmental fractal properties. We have employed these features for the speech emotion recognition. The results obtained are encouraging: the average recognition rate for 7 Lithuanian emotions was 98,2 %.

Analysis of Big Data Based on Conditionally Nonlinear Autoregressive Model

Yuriy Kharin

Research Institute for Applied Problems of Mathematics and Informatics
Belarusian State University
kharin@bsu.by

Applications in genetics, finance, medicine, information protection and other fields need to develop algorithms for computer modelling and analysis of big data presented in the form of long discrete-valued time series. An universal long-memory model for such data is the homogeneous Markov chain of sufficiently large order s on some finite state space A, |A|=N, 2<=N<+∞. Unfortunately, the payment for this universality is exponential w.r.t. the order s number of parameters D=O(N^(s+1)). To identify such a model, we need to have big data sets and the computational complexity of order O(N^(s+1)). To avoid this “curse of dimensionality,” we propose to use the so-called parsimonious (“small-parametric”) models of high-order Markov chains that are

determined by a small number of parameters [.Kharin Yu.S. Robustness in Statistical Forecasting. N.Y.: Springer (2013)].
We present a short review of our results on algorithms for computer analysis of discrete time series based on known models: Jacobs – Lewis model, Raftery MTD-model, MC(s, r)-model and its modification.
We propose a new parsimonious model – binary conditionally nonlinear autoregressive model - and construct algorithms for the fitting of this model to observed data. Results of computer analysis of real data are given.

Retinal Imaging and Image Processing for Glaucoma Diagnosis

Radim Kolář

Department of Biomedical Engineering
Brno University of Technology
kolarr@feec.vutbr.cz

Retinal imaging is still a developing field of ophthalmology. Mainly, optical coherence tomography had changed the quality of retinal imaging. Nevertheless, other new modalities are also important, because they are able to image the functional properties of the retinal tissue. The progress in this research area helps to understand many retinal diseases, including glaucoma, which is still not well understood.
This talk will briefly discuss imaging techniques for glaucoma diagnosis and image processing techniques used in this area. The main part will describe our current research activities with Department of Ophthalmology, Friedrich-Alexander-University Erlangen–Nürnberg, the possibilities of video-ophthalmoscopy and parallel retinal imaging with the focus on functional aspects of this modality.

An Initial Investigation of Keyword Spotting using Convolutional Neural Network

G. Korvel, P. Treigys, G. Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University
grazina.korvel@mii.vu.lt

The goal of keyword spotting is the identification of particular word in a spoken stream. An effective keyword spotting can be applied for data mining, audio document indexing, voice command detection, real-time conversation monitoring, and surveillance purposes. The task of keyword spotting is a sub-field of automatic speech recognition and therefore faces the same challenges of non-stationariness of the speech signal, language-specific and speaker variability, noise and distortions. Because of this, most of the advanced keyword spotting approaches are quite limited in the accuracy so far.

In this study, convolutional neural network based keyword spotting approach is introduced. The neural network is trained to classify the words and the keyword detection is implemented via analysis of word sequences. For classification purposes, the utterances are represented by spectrograms, calculated as time-frequency function.

To assess the introduced keyword spotting approach, the initial experimental investigation was conducted. 11 words of the total 111 were randomly chosen as the keywords. Each of the keywords was represented by 2990 spectrogram samples obtained from pronunciations by various speakers with the added different noise level. The obtained average classification accuracy was more than 90 %. Analysis results of confusion matrix for all 11 keywords also show the potentials of the convolutional network for keyword spotting task.

Rough Sets Applied to Music Informatics

B. Kostek, G. Korvel

Faculty of Electronics
Gdansk University of Technology, Poland
bozena.kostek@pg.edu.pl

In this presentation music data processing and mining in large databases is investigated based on soft computing methods. First, principles of rule-based classifiers and particularly rough sets are presented, showing their

usability in music informatics. Several examples of music processing are shown, including music genre/mood classification, automatic music collection tagging, personal recommendation, composing a playlist, etc. For the purpose of this research study a large number of 30000 audio files divided into different music genres/music mood were gathered to form a database. All files contained in this database were parametrized and resulted in feature vectors of 173 parameters. To reduce the dimensionality of data the correlation analysis was performed. This was then compared to the rough set-based processing of the same feature vectors, as such an algorithm produced reducts containing the most promising descriptors in the context of music genre/mood recognition. Classification tests were conducted using the Rough Set Exploration System (RSES), a toolset for analysing data with the use of methods based on the rough set theory as well as in the WEKA environment with the use of k-Nearest Neighbors (kNN), Bayesian Network (Net) and Sequential Minimal Optimization (SMO) algorithms. All results were analysed in terms of the recognition rate and computation time efficiency. In conclusion, a potential of the rough set-based approach when applied to music informatics was underlined as it offers the possibility to deal with imprecise, vague and indiscernible data objects.

The New Parallel Multilevel Tool for Implementation of Applied Optimization Algorithms

R. Kriauzienė[1, 2], A. Bugajev[2], R. Čiegis[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Vilnius Gediminas Technical University
kriauziene@gmail.com

In this work, we consider a general class of optimization problems for which the computation of objective function requires the solution of M different subproblems, with a priori estimated sizes of subproblems. Such optimization problems are often very computationally intensive. Thus, the parallel computations are necessary.

In general case, we define three levels of parallelization. Our main goal is to investigate a workload distribution between processes for a three-level parallel algorithm. On the first level, we implement some local optimization algorithm. For a given an example we use the simplex downhill method. A modification the basic algorithm is proposed, which increases the number of independent tasks.

However, the parallelization of this method is very limited, so for the big number of processes one level parallelization is not sufficient. On the second level, for each task, we calculate solutions for independent M subproblems in parallel. Note, that computational sizes of these subproblems can be very different. In order to perform load balancing, we introduce the third level of parallelization when subproblems are solved using some efficient parallel algorithms. In our case subproblems are non-stationary 1D differential equations. At each time step, they are approximated by systems of linear equations with the tridiagonal matrix. We use Wang's algorithm to solve the obtained systems.

The third level can be used alone. However, the efficiency of Wang's algorithm is limited due to a well known speed-up saturation effect, i.e., computations are slowed down due to large communication costs when the number of processes is increased. Thus a well-balanced distribution of tasks at all three levels should be defined. The results of computational experiments are presented and analysed.

Mathematical Morphology Based Method for Cultured Cell Motility Imaging

A. Krisciukaitis, R. Ramonaite, R. Petrolis, J. Skieceviciene,
L. Kupcinskas

Lithuanian University of Health Sciences
algimantas.krisciukaitis@lsmuni.lt

Cell culture is a method for growing or maintaining cells in vitro under controlled conditions. Cell culture models are widely used as an alternative to animal models to study various physiological phenomena or pathologies related to cell functionality and/or viability. Dispersed cells that are cultured directly from a tumor or healthy tissue usually have a limited lifespan. However, there is a big variety of commercially available cell lines, which refer to immortalized cells that can be cultured indefinitely. Majority of them are derived from tumors and can be used to study cancer pathogenesis and treatment. Evaluation of viability or functionality of cultured cells must be done in, as much as possible, the noninvasive way by visual inspection or applying advanced imaging technics. Invasiveness of cancer cells could be evaluated by the feature of their functionality – motility. It could be observed only by means of time-lapse microscopy and special imaging technics. The aim of this work was to elaborate a method for quantitative evaluation of cultured cells motility. We applied it to characterize invasiveness of commercially available cell lines derived from colon (cell line HCT116) and gastric tumors (cell lines AGS and MKN28). Mathematical morphology methods were applied to determine individual cell movements. Generalized cell movement maps were calculated from time-lapse recordings allowing visual inspection and detail morphological investigation. The investigation revealed quantitatively distinct motility in the colon and gastric cell lines. Gastric cell lines AGS and MKN28 showed the different spatial distribution of registered movements in regard to cell location. Although cancer cell lines are considered as immortalized, one can expect certain difference in their functionality in regard to passages (times cells were sub cultured into new vessels). We observed certain decay in cell motility in later passages. Quantitative characterization of cultured cells functionality by using elaborated method could be used to reveal differences in cell types or to monitor/control the cell line during use in several passages.

Evaluation of Online Banking Acceptance

K. Lapin, T. Šturo

Faculty of Mathematics and Informatics
Vilnius University
kristina.lapin@mif.vu.lt

Online banking systems offer costumers cost savings, reduced wait times and convenient access to services. However, the adoption of these services is particularly affected by trust concerns. Based on literature research, this paper proposes a heuristic evaluation methodology for evaluating the acceptance of online banking systems. The methodology provides a set of questions, which should be answered to evaluate acceptance and check the level of trust and reliability of online banking systems. The methodology has been verified on a widely used online banking systems.

Statistical Analysis of Word Frequency Distribution in Texts of Different Genres: Comparison of Lithuanian and English

J. Mandravickaitė[1, 2], T. Krilavičius[2, 3]

[1] Vilnius University
[2] Baltic Institute of Advanced Technology
[3] Vytautas Magnus University
justina@bpti.lt

We report an ongoing study on statistical characteristics of texts written in different genres. It has been suggested that genres resonate with people because they provide familiarity and the shorthand of communication. Also, genres tend to shift hand-in-hand with public opinion and reflect widespread culture of certain period(s). From NLP perspective, genres come in use in text classification and categorization, natural language generation, etc.

At this stage, we present a statistical analysis of Lithuanian and English texts of genres. For our explorations, we use Corpus of the Contemporary Lithuanian Language (for Lithuanian part) and Freiburg-LOB Corpus of British English (F-LOB). The main points of interest are number of words, number of different words and word frequencies. Structural type distribution and Zipf's law were applied in order to describe the frequency distribution of words in different textual genres.

Zipf's law is one of the universal laws proposed to describe statistical regularities in language. Thus word frequencies and their derivative indicators could be used to characterize textual genres. Application of word rank-frequency distribution, type-token ratio, the percentage of hapax legomena, i.e., words that occur only once, and entropy for different genre groups (fiction, scientific articles, documents, news articles) supported the latter assumption.

Differences between languages (Lithuanian and English) were observed as well. As genres are rather complex phenomena that depend on various linguistic, cultural, societal, etc. factors, our future study includes research of additional frequency structure indicators as well as their combinations.

Model-Based Systems Engineering Approach for Creating Secure Complex Systems

D. Mažeika[1,2], R. Butleris[1]

[1] Faculty of Informatics
Kaunas University of Technology
[2] No Magic Europe
donatas.mazeika@ktu.lt

Model-Based System Engineering (MBSE) is an emerging approach that applies modelling instead of documents to support complex system requirements, design, analysis, and verification and validation activities beginning in the conceptual design phase and continuing throughout development and later life cycle phases. The base of MBSE is Systems Modelling Language (SysML) which is a lightweight extension of Unified Modelling Language (UML). SysML enables systems engineers to capture and trace various types of models that come from different engineering disciplines. MBSE deals with complex systems (e.g., automotive, aerospace, medical devices) and security concerns for these systems are very sensitive.

Also, security risks (e.g., confidentiality, integrity, availability) could be considered very early together with the business logic and system requirements, however, the SysML language does not address how security aspects of the systems should be specified.

This presentation reviews existing methods for identifying security issues and introduces how they could integrated into the MBSE processes.

Research of Context Influence on Business Process Modelling and Simulation

I. Mičiulytė, O. Vasilecas

Institute of Data Science and Digital Technologies
Vilnius University
ieva.miciulyte@mii.stud.vu.lt

Nowadays, organizations aim to adapt quickly to changing external and internal contexts, and therefore optimize their business processes (BP). BP simulation allows analysing of various business scenarios, helps to optimize the activities of organizations, quickly adapt to various changes, discover bottlenecks of business and eliminate them. Therefore, BP modelling and simulation becomes more and more valuable for business.

Equally, it is important to have a possibility to define process contexts at information systems analysis phase and to assess the effect of context changing using BP simulation. The analysing of recent publication in the field also shows that researchers try to improve existing context modelling methods in order better to adapt for context-aware process. However, BP simulation when processes react to the changing process context is not sufficiently investigated and not fully realized in simulation tools. Therefore, in the paper we investigate context-aware BP modelling and simulation and existing BP simulations tools. This paper review existing BP context modelling approaches and requirements that they meet, and presents extended approach for BP context modelling. Proposed context model describes internal and external context elements using Business Process Model and Notation (BPMN) and Semantics of Business Vocabulary and Business Rules. Moreover, we create constraints for context model based on requirements for elements used in proposed context model. The presented model and set of constraints are formalized using the UML and OCL. Proposed constraints and model was validated to show the feasibility of proposed method. A method allows to extend context modelling and simulation capabilities and allows to increase adaptability to the changing context.

Data-driven Decision Support with Multiobjective Optimization

Kaisa Miettinen

Industrial Optimization Group
Faculty of Information Technology
University of Jyvaskyla, Finland
kaisa.miettinen@jyu.fi

Thanks to digitalization, we can collect and have access to various types of data and the question of how to make the most of the data arises. We can use descriptive or predictive analytics but to make recommendations based on the data, we need prescriptive or decision analytics. If the decision problems derived contain multiple conflicting objectives, we must employ methods of multiobjective optimization. Lot sizing is an example of such a data-driven optimization problem. Lot sizing is important in production planning and inventory management where a decision maker needs support in particular when the demand is stochastic. We propose a problem formulation with three objectives and solve it with

interactive multiobjective optimization methods. In interactive methods, a decision maker directs the search for the best balance between the conflicting objectives by providing preference information. In this way, (s)he can learn about what kind of solutions are available for the problem and also learn about the feasibility of one's preferences.

We consider the lot sizing problem of a Finnish production company. The results of this data-driven interactive multiobjective optimization approach are encouraging.

Evolution of the Notions of Algorithm and Computation: a Systematic Mapping Study

J. Miliauskaitė, L. Paliulionienė

Institute of Data Science and Digital Technologies
Vilnius University
jolanta.miliauskaite@mii.vu.lt

The notion of algorithm and other related notions (e.g., computing models and computation) were being developed during a long period of time, starting from the intuitive understanding by ancient Greek and medieval Arab philosophers and scientists; to the elaborating of formal theories in the 19th to mid-20th century by Gödel, Hilbert, Church, Post and Turing; and lastly to further evolving and applying of them in the age of computers. Moreover, these notions are not limited to domains that are created by the human minds, like mathematical logic, mind philosophy, artificial intelligence, computer sciences, and information system engineering. Indeed, investigations of algorithmic properties of biological, physical, and other natural processes are being increasingly performed.

In our investigation, we have conducted a systematic mapping study in order to classify and summarise contributions produced over time in the area of algorithms, computing models, and computation. This work is a part of a more extensive research that aims to perform a high-level conceptual overview of the field and to present a great variety of opinions, attitudes, and approaches in this area.

On the Law of the Iterated Logarithm for Extremal Queue Length in an Open Queueing Network

S. Minkevičius, E. Greičius

Faculty of Mathematics and Informatics
Vilnius University
saulius.minkevicius@mii.vu.lt

Queueing networks models have been extensively used for the performance analysis of manufacturing systems, transportation systems, and computer and communication networks. The paper is designated to the analysis of queueing systems, arising in the network and communications theory (called an open queueing network). We deal with approximations of an open queueing network and present a theorem about the law of the iterated logarithm for the extreme value of the queue length of customers in an open queueing network. The proof method used in the paper (name it as a recurrence method) can be widely applied in the network theory, e. g. in the mixed-component open Jackson network (that is a generalization of the classical open Jackson network).

Impact of Colour on Colorectal Cancer Tissue Classification in Hematoxylin and Eosin Stained Histological Images

M. Morkūnas[1, 2], P. Treigys[1], J. Bernatavičienė[1],
A. Laurinavičius[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] National Center of Pathology
Affiliate of Vilnius University Hospital Santaros Klinikos
mindaugas.morkunas@mii.vu.lt

Because of the constant discovery of new tumour tissue biomarkers, there is substantial interest in advanced computational pathology algorithms that would accomplish highly specific tasks of cancer research. Digitized pathology slides are the object for computational pathology pipelines to precisely detect, classify, quantify, and segment multiple types of histological objects.

Tumour tissue classification is often fine-tuned to adapt to highly specific end-goals of comprehensive pathology research whether a new,

unexplored cancer type is concerned, or an emerging biological tumour property needs additional evaluation.

Typically, raw pathology image data is produced in RGB colour space. However, RGB space may appear not the optimal choice. More perceptually uniform colour spaces have been shown to outperform RGB space in texture classification. In this work we explore the influence of colour information on classification of H&E stained colorectal cancer tumours into tissue compartments using the convolutional neural network. We will explore how the network model responds to separate colour channels of the RGB and CIELab colour spaces.

CUDA and OpenMP Implementations for Solving SMACOF Problems

G. Ortega[1], F. Orts[1], E. M. Garzón[1], E. Filatovas[3], O. Kurasova[2]

[1] TIC-146 Supercomputación-Algoritmos, Dpt. of Informatics
University of Almería, Spain.
[2] Institute of Data Science and Digital Technologies, Vilnius University
[3] Vilnius Gediminas Technical University
gloriaortega@ual.es

Real-world data, such as speech signals, images, biomedical, financial, telecommunication and other data usually have a high dimensionality. Each data instance (point) is characterized by some features. The dimensionality of such data, as well as the amount of processed data, is constantly increasing but the requirement of processing these data within a reasonable time frame remains an open problem. Dimensionality reduction methods which aim to map high dimensional data into a lower dimensional space play extremely important role when exploring large datasets. Among such methods, Multidimensional Scaling (MDS) remains one of the most popular ones.

Several approaches have been developed to reduce the computational complexity of the MDS techniques. However, the MDS techniques remain in high time complexity order. Therefore, parallel strategies should be considered to accelerate the computation of the MDS procedure. A well-known algorithm for MDS is called SMACOF (Scaling by Majorizing a COmplicated Function).

The experimental investigation has demonstrated that SMACOF is most accurate algorithm comparing to others. It should be noted that the SMACOF algorithm is expensive, as its complexity is O(m^2), where m is the number of observations.

In this work, two parallel versions of the SMACOF algorithm have been developed and evaluated on multicore and GPU platforms. To help the user of SMACOF, we provide these parallel versions and a complementary Python code based on a heuristic approach to explore the optimal configuration of the parallel SMACOF algorithm on the available platforms in terms of energy efficiency (GFLOPs/watt). Three platforms, 64 and 12CPU-cores and a GPU device, have been considered for the experimental evaluation.

Investigation of Interpolation Methods for Virtual Rowing Simulator

A. Paulauskas, T. Valatkevičius, C. Canbulut

Faculty of Informatics
Kaunas University of Technology
cenker.canbulut@ktu.edu

The aim of this introduction is to provide an example of virtual reality solution for rowing training machine. In this introduction, we developed a “Virtual Reality” supporting application that takes data of a rowing machine where the user performs rowing exercise and translates it to a visual representation of a rowing experience. This lets users or performers to practice their daily training using the rowing simulator rather than performing it outside in bad weather conditions or busy life time periods. When it comes to exchanging the data of a rowing machine into virtual reality environment, there is a major issue comes with it. This issue occurs because of the different behaviour of virtual reality environment and real-life environment. It is necessary to translate the data which will behave and take action according to the performers behaviour in real life within the virtual reality. That is why we look at the various issues arising when integrating virtual reality with mechanical devices. Refresh rate between machine behaviour and the actual data behaviour depend on the data exchange between those two devices (rowing machine and Virtual Reality device). It is a necessity to provide high refresh rates to let users experience real rowing experience to get precise data on board and present the result reliable to the fact of the user effort. This way, we get close result if the user would provide the same exercise in real life.

We purpose and investigate a solution to this issue by using several algorithms to predict the behaviour of the rowing machine at the time we are lacking some data. Using some of these algorithms provide life-like rowing immersion while being quite accurate at the same time.

General-purpose Bilevel Solver BASBL: Implementation and Computational Study Using BASBLib Library

R. Paulavičius[1], Claire S. Adjiman[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2]Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, United Kingdom
remigijus.paulavicius@mii.vu.lt

Bilevel optimization problems are very challenging optimization problems arising in many important practical applications, including chemical and civil engineering, economics, transportation, pricing mechanisms, airline and telecommunication industry, machine learning and etc. The numerous applications of bilevel programming problems provide a strong incentive for developing efficient solvers for this large class of problems.

In this talk, we introduce the BASBL solver, an implementation of the deterministic global optimization algorithm Branch-and-Sandwich for mixed-integer nonconvex/nonlinear bilevel problems, within the open-source MINOTAUR toolkit. The BASBL solver stems from the original Branch-and-Sandwich algorithm and modifications proposed in our recent works. We also introduce BASBLib, an extensive online open-source library of bi-level benchmark problems collected from the literature. The library is designed to enable contributions from the bi-level optimization community. We use the problems from BASBLib, including problems derived from practical applications, to study the performance of BASBL using different algorithmic options, including a variety of bounding schemes, branching, and node selection strategies.

Optimization of Surface Wastewater Treatment Filter Filler Effectiveness Using Mathematical Modelling

N. Pozniak, L. Sakalauskas

Institute of Data Science and Digital Technologies
Vilnius University
natalija.pozniak@gmail.com

Inadequate treatment of surface wastewater may considerably impair the quality of water. With increasing urbanization, intensification of car traffic and increasing area of impervious surfaces, the pollution of surface water and negative impact on the aquatic environment are also rising. Increasing surface water pollution leads to intensive eutrophication. One of surface wastewater treatment technologies capable of reducing suspended solids, heavy metals and other pollutants is surface wastewater treatment filters.

Filters with different fillers are designed for treatment of main pollutants of surface wastewater: suspended solids, heavy metals (zinc, cadmium, copper, lead), BDS5, total carbon and nitrogen. The effectiveness of filters filled with construction waste and biocarbon was analysed using the kriging method with distance matrices. The developed method allows modelling filter characteristics with different filler ratios based on the previous experimental studies of filters.

Genetic Algorithm Optimization of Foundations Using BOINC Framework

M. Ramanauskas, D. Šešok

Vilnius Gediminas Technical University
mikalojus.ramanauskas@vgtu.lt

This work addresses the problem of global optimization of real-life civil construction: optimization of rostverk-type foundations. The problem is solved using a genetic algorithm. These algorithms depend on many parameters, and in order to find the best set of parameters, numerous experimental calculations are required. In order to accelerate the optimal detection of parameters, calculations were performed on the distributed calculations platform BOINC.

Extended Analysis of the Sapnai.net Dataset

A. Rapečka, J. L. Jedzinskas, G. Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University
aurimas.rapecka@mii.vu.lt

There are many types of recommendation systems and recommendation methods in today's world. Each method has its advantages and disadvantages. Most popular and widely distributed universal methods demonstrate good results in most data sets, but this method require large computing resources and often is to slow, especially for real time

calculations. With limited computation resources or large datasets, applications of most methods are not successful.

An example or large dataset is Sapnai.net Dreams Meanings Dataset. Dataset consists of 27 000 000 search records by 1 470 000 users in 4200 dreams meanings. This dataset is available for scientific researches and published online at *www.sapnai.net/db*.

The aim of this work is to present an extended analysis of this dataset and efficiency of several most popular recommendation methods in this dataset. We solved several challenges in this analysis: determination similar groups of users, relations between dreams by users, relations between week or month days and dreams popularity.

Efficient Human Motion Matching Algorithm for Depth Scanning Systems Based on Hausdorff Distance Metric

K. Ryselis

Faculty of Informatics
Kaunas University of Technology
karolis.ryselis@ktu.edu

Human motion tracking problem is solved in a variety of ways. Usually, it requires a depth scanning sensor that may or may not process its data and an algorithm to compare sensor's data to a predefined template. One of the most widespread depth scanning sensors is ``Microsoft Kinect". However, most algorithms for this sensor can only be used to track standing positions. One of the algorithms that can compare any shapes is Hausdorff distance. Unfortunately, it is suited to compare static poses, and there is no fast implementation available. An algorithm that is fast enough to use with “Microsoft Kinect” sensor and can track any positions was created. It is based on Hausdorff distance metric and ideas of shape transformations using Procrustes analysis. The algorithm's high performance comes with tradeoff in accuracy because the transformations used are best-guess only and not exact. Therefore, algorithm's performance and precision were both evaluated. The developed algorithm works in two stages: firstly, it transforms the two shapes so that they are in the same position and the same size in their frames using heuristic Procrustes analysis; secondly, it calculates the differences between the two shapes using Hausdorff difference matrix and a single dissimilarity measure as Hausdorff distance. These steps are repeated for all frames to track and live results are calculated.

It was found that the developed algorithm may work with frames sized up to 1.7 times larger than the body index frame size of “Kinect 2” in real time (at least 30 frames second) using modern hardware and ``Kinect 2" frames are handled even faster, suggesting that any hardware compatible with “Kinect 2”sensor may run motion tracking without dropping frames. It was determined that in order to get small error it is reasonable to calculate Procustes analysis coefficients at the beginning of the motion sequence with both tracked user and template in standing positions and reuse them throughout the motion sequence provided that motion sequence tracked is continuous and the same user is tracked throughout the sequence.

Using this method average error in Hausdorff distance metric exceeds 2%, and precise results are calculated for 91% of all frames. The results suggest that the algorithm works well in areas like yoga training and rehabilitation because they involve non-standing positions and require real time motion tracking.

Securing Fog Network Using Blockchain

D. Rudzika, A. Venčkauskas

Kaunas University of Technology
darius.rudzika@ktu.lt

Efficient IoT device communication with Cloud and security aspects of IoT devices and their network communication have been a burning issue for quite some time. In this talk we cover:

a) The concept of using Fog network for efficient IoT devices connectivity;

b) Inherent security issues of IoT devices and networks;

c) Blockchain technology based solution to improve IoT device and Fog network security.

In Fog network part we cover IoT communication shortcommings and how they are solved using Fog networks. Further, we cover most common security threats to IoT devices and networks, such as Impersonation, Man In the Middle, Collusion, Denial of Service, Eavesdropping, Jamming, Spam, Tampering, and Forgery. In the solution part, we shortly overview the Blockchain technology concepts (Distributed Ledger), classification (Public, Private, Federated) and most popular use cases (cryptocurrency, smart contracts).

Further, we propose the Blockchain technology based solution to improve the security of Fog network and connected IoT devices.

On the Local Geometric Approach to Global Optimization for Multidimensional Scaling

M. Sabaliauskas, G. Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University
martynas.sabaliauskas@mii.vu.lt

Multidimensional scaling (MDS) is the most popular method for a visual representation of high-dimensional data. Each low-dimensional visualization requires holding similarity and distances between multifaceted objects as possible. MDS ensures such objective. The proposed idea for MDS is to approach geometrically to the global extremal value of stress function for each multidimensional object separately and repeat this process until stress function value stops decreasing. The special geometric procedure, ensuring the local search of the position of projection of the separate object, is proposed. In the realization, the multistart descent is incorporated seeking for the global position of the separate object projection. According to experimental results, our method tends to solve the global optimization problem of MDS.

A Research on Safety Methods of Communication Protocols of Internet of Things

R. Savukynas, V. Marcinkevičius, D. Dzemydienė

Institute of Data Science and Digital Technologies
Vilnius University
raimundas.savukynas@mii.vu.lt

Currently, there are many standardized Internet communication protocols that belong to different levels of the interconnection model of open systems. These communication protocols are widespread, have different energy costs and support safety-critical techniques. Security methods allow you to identify objects of the Internet of Things, establish their operating rules, protect the sent confidential information, and verify the integrity of the messages received.

Security is one of the main components of the Internet of wisdom-related things, as these environments interact with people and other objects in the environment, and it is necessary to ensure secure communication between the objects and the users of intelligent environments. The agents in the Smart Objects Internet objects interact with each other with specific high-level protocols that must ensure minimum energy consumption, security of data transmission and reliability. Smart environments that are designed for manufacturing, military, health, and other security-critical applications, and it is necessary to protect these environments from various hazards and external malicious impacts that can be exploited by these environments.

An important safety aspect is the privacy of individuals and organizations because the smart environment introduced in people's living or working environments can be directly used to collect illegal and sensitive data about the environment. In this work, the methods of security of Internet communication protocols for smart objects, ensuring reliable communication and interoperability of objects of the Internet of things are investigated.

Predictive Novelty Discovery for Real-Time Stereoscopic Image Analysis and Decision Support

A. Serackis, V. Abromavičius, V. V. Borutinskaitė,
D. Navakauskas

Department of Electronic Systems
Vilnius Gediminas Technical University
arturas.serackis@vgtu.lt

The presented investigation focus on eye fatigue problem. The problem arises during sustained visual perception of stereoscopic views. The high importance of this problems is supported by currently increased attention to the head mounted displays (virtual reality and augmented reality glasses) and immersive visual content (e.g. 360-degree video records). A stereoscopy effect is achieved by presenting individual views for each eye. The disparity between two views gives a possibility to support better depth perception. The higher disparity between image objects helps to distinguish between similar objects situated at different distances from the viewer. However, high disparities are related to the higher visual discomfort levels and may cause the eye fatigue during long lasting stereoscopic perception. The aim of the investigation is to extract multimodal data features, related to discomfort during stereoscopic perception of images with objects with different disparity level on the focus point. An experimental investigation with a control group was performed using special 120 stereo image dataset. The features were extracted from eye tracking, pupil size monitoring, EOG and EEG real-time signal measures. As a result of the investigation, the dependences of object disparity to dynamic changes of pupil size were estimated. A new technique was proposed for the identification of the moment from which the correct stereo perception is achieved on the image object with different level of disparity.

Lipschitz Global Optimization

Yaroslav D. Sergeyev

Numerical Calculus Laboratory
Calabria University, Italy
N. I. Lobachevsky University of Nizhni Novgorod, Russia
yaro@si.dimes.unical.it

Global optimization is a thriving branch of applied mathematics, and an extensive literature is dedicated to it. In this lecture, the global optimization problem of a multidimensional function satisfying the Lipschitz condition over a hyperinterval with an unknown Lipschitz constant is considered. It is supposed that the objective function can be “black box,” multiextremal, and non-differentiable. It is also assumed that evaluation of the objective function at a point is a time-consuming operation. Many algorithms for solving this problem have been discussed in the literature. They can be distinguished, for example, by way of obtaining information about the Lipschitz constant and by the strategy of exploration of the search domain. Different exploration techniques based on various adaptive partition strategies are analysed. The main attention is dedicated to two types of algorithms. The first of them is based on using space-filling curves in global optimization. A family of derivative-free numerical algorithms applying space-filling curves to reduce the dimensionality of the global optimization problem is discussed. A number of unconventional ideas, such as adaptive strategies for estimating Lipschitz constant, balancing global and local information to accelerate the search, etc. are presented. Diagonal global optimization algorithms are the second type of methods under consideration. They have a number of attractive theoretical properties and have proved to be efficient in solving applied problems. In these algorithms, the search hyperinterval is adaptively partitioned into smaller hyperintervals, and the objective function is evaluated only at two vertices corresponding to the main diagonal of the generated hyperintervals. It is demonstrated that the traditional diagonal partition strategies do not fulfill the requirements of computational efficiency because of executing many redundant evaluations of the objective function.

A new adaptive diagonal partition strategy that allows one to avoid such computational redundancy is described. Some powerful multidimensional global optimization algorithms based on the new strategy are introduced. Extensive numerical experiments are performed on the GKLS-generator that is used nowadays in more than 40 countries in the world to test numerical methods.

Results of the tests demonstrate that proposed methods outperform their competitors in terms of both numbers of trials of the objective function and qualitative analysis of the search domain, which is characterized by the number of generated hyperintervals. A number of possible generalizations to problems with multiextremal partially generated constraints are mentioned. The usage of parallel computations and problems with multiextremal constraints are discussed briefly, and theoretical results on the possible speed-up are presented.

Automatic Artery Vein Ratio Measurements in Eye Fundus Images

G. Stabingis[1,2], J. Bernatavičienė[1], G. Dzemyda[1], A. Paunksnis[3], R. Vaičaitienė[4], L. Stabingienė[2]

[1] Institute of Mathematics and Informatics, Vilnius University.
[2] Klaipeda University.
[3] Telemedicine Research Center.
[4] Gen. J. Zemaitis Lithuanian Military Academy
giedrius.stabingis@mii.vu.lt

Retinal vascular imaging offers a noninvasive research tool for many diseases like diabetic retinopathy, arterial hypertension, ageing macula degeneration, glaucoma and others. Detecting diseases in their early stages are essential. One of the early symptoms is artery vein ratio in eye fundus images. It allows not only to detect but also to follow the course of the disease. Here, we report on the development of an automatic, operator-independent, computer-based method for measurements of the artery-vein ratio in eye fundus images.

The proposed method involves different image pre-processing methods used in different stages of the algorithm, mathematical morphology based preliminary blood vessel tree extraction, two-stage optic nerve disc detection, vessel measurement algorithm, vessel feature extraction, vessel classification and artery vein ratio calculations for upper and lower part of the fundus image.

The great importance of the proposed approach is its ability to process fundus images that are of different resolution and in lower quality. The low fundus image quality is a standard issue introduced by some diseases, in situations when image taking is complicated, and when specialist lacks camera usage experience. The proposed method uses vessel measurement algorithm which makes measurements independently of the extracted blood vessel tree. Obtained results are compared with Matlab based ARIA software on the publicly available database. For artery-vein ratio investigation, high-resolution images gathered with the Optomed OY digital mobile eye fundus camera Smartscope M5 PRO is used. Artery-vein ratio evaluation is tested on images obtained from healthy and ill people.

Improved DIRECT-type Algorithms for Generally Constrained Global Optimization Problems

L. Stripinis, R. Paulavičius

Institute of Data Science and Digital Technologies
Vilnius University
linas.stripinis@mii.vu.lt

In this talk, we consider a very general global optimization problem. The well-known derivative-free global-search DIRECT (DIvide a hyper-RECTangle) algorithm performs well on a subclass box-constrained problems. Unfortunately, quite often the efficiency of the DIRECT algorithm deteriorates on problems with many local optima or when the solution with a high accuracy is required. To overcome these difficulties, different regimes of the global and local search are introduced, or the algorithm is combined with a local optimization schemes. We investigate a different direction of addressing DIRECT inefficiencies and propose a new strategy for the selection of potentially optimal rectangles. Our new strategy does not require any additional parameters or local search subroutines.

Solving problems with linear or nonlinear constraints, the DIRECT algorithm does not naturally address them, and till now only very few DIRECT extensions were proposed, to the best of our knowledge. Thus, in this talk, several different general constraints handling methods within

the DIRECT framework are proposed and investigated. One of the existing approaches is based on exact L1 penalty functions. However, performed numerical investigation revealed potential inefficiency of this approach when inaccurate penalty parameters are used.
We propose several modifications to address such inefficiencies staying in a DIRECT algorithmic framework. An extensive experimental investigation reveals the effectiveness of the proposed enhancements.

Data Envelopment Analyses: from Education Systems Performance Assessment (Control) to Performance Management (Improvement)

D. Stumbrienė[1], A. S. Camanho[2], A.Jakaitienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] School of Engineering of the University of Porto
dovile.stumbriene@mii.vu.lt

Monitoring the performance of educational systems is at the top of the agenda of governments and educational authorities worldwide, but only a small number of studies have focused on a country or multi-country analysis. The comparison among educational systems based on primary, secondary or tertiary education individually does not represent all education system. Only the analyses of all the educational levels together can represent all education system in the country. While it is possible to evaluate performance using individual indicators and assess their evolution over time, it is not a trivial task to conduct multi-dimensional evaluations. Taking into account several indicators simultaneously and their progress over time require more advances techniques, such as the construction of composite indicators. This research focuses on the construction of composite indicators for the education monitoring, discussing the advantages and disadvantages of different modelling alternatives. The purposes of our research are to compare the traditional formulation of composite indicators (standard aggregation) with the DEA-based model (involving the optimization of weights) and to identify the impact of the different type of weight restrictions (fixed and flexible weighting systems). Cross-country analyses are important for policy making since they allow the benchmarking of educational policies. Our study uses annual data from 29 European countries over the 2014 year, collected from EUROSTAT and OECD databases. We went through the following five stages for the construction of composite indicators: data treatment, data normalisation, weighting, aggregation, and comparison of the results obtained, both in terms of the efficiency score as well as country rankings.

Evaluation and Analysis of the Power of Local Geomagnetic Field

V. Šiaučiūnaitė [1], M. Landauskas [1], A. Vainoras [2], M. Ragulskis [1]

[1] Kaunas University of Technology
[2] Lithuanian University of Health Sciences
vaiva91@gmail.com

A Recent collaboration between Department of Mathematical Modelling. in Kaunas university of technology (KTU) and cardiologists from Lithuanian university of health sciences (LUHS) are focused on the analysis of possible interconnections between the Earth's magnetic field and parameters of human's cardiovascular system. It is already shown that there exist connections between geomagnetic field and increased number of coronary diseases, heart attacks, blood pressure variations, HRV as well as some others.

It is important to first be able to measure the geomagnetic field precise enough and secondly evaluate its power in different frequency ranges in an extreme value-robust way. The first problem could be considered solved as LUHS operates the magnetometer which is located in central Lithuania (Baisogala). The device registers the intensity of local geomagnetic field in two directions in picotesla accuracy. It must be emphasized that such accuracy is sufficient to register a whole spectrum of magnetic signals: Schuman resonances, disturbances of the power grid, part of the Earth's natural magnetic fluctuations.

Our research focusses on several predefined frequency ranges of the local geomagnetic field. The most intuitive way to find the power of the given signal in particular frequency ranges is to crop then clip the spectrogram and sum the remaining values from the time interval of interest. The question is at what level we should perform the cropping operation. In this work, we present the approach which achieves that in an optimal way. For the criterion of the optimality, we introduce the aim function and solve the optimization problem.

Mathematical approaches presented will be employed in future research activities carried out by KTU and LUHS. Improved techniques for magnetic field power's computation will lead to assess the power in a certain frequency range more precisely. This is particularly important for frequency ranges spanning Schuman resonances and some of the lower frequency intervals as they are comparable in frequency to the brain activity, rhythms of the cardiovascular system and autonomic nervous system.

Construction of Active and Semi-Active Schedules for Job Shop with Fixed Post-Processing Periods

V. Tiešis

Institute of Data Science and Digital Technologies
Vilnius University
vytautas.tiesis@mii.vu.lt

Classical scheduling models (e.g., job shop) usually are too simple to represent all details of production. In the research, the extended job shop model with post-processing periods (cooling down, stiffening, etc.) necessary after extraction of a product from a machine is considered. It is assumed that duration of such periods is known in advance and buffers between machines are spacious enough to stow all products during post-processing periods.

The active or semi-active schedule generation scheme (SGS) is considered in which in each iteration some procedure extracts the set of eligible operations from the set of available operations. Then one operation from the set of eligible operations is selected for appending to schedule by some priority rules. The Giffler-Thompson procedure that is common for job shop problem and generates an active schedule was modified to be used for the case with post-processing periods.

It was proved that SGS with the modified Giffler-Thompson procedure and with any priority rules generates an active schedule for the extended job shop problem with post-processing periods.

The semi-active SGS common for job shop problem selects an available operation by some priority rule and schedules it after the previously scheduled operation as soon as possible. It was proved that such SGS also generates a semi-active schedule in the case of the extended job shop problem with post-processing periods.

In the research, the job shop model was also extended for the case when after the accomplishment of the setup, workers are free to do other operations but the prepared machine continues the previous operation. Appropriate SGSs were proposed for such problems.

Visualization of Relationships between ECG Parameters Using Optimal Lagrangian Difference Matrices

I. Timofejeva[1], A. Vainoras[2], M. Ragulskis[1]

[1] Department of Mathematical Modelling
Kaunas University of Technology
[2] Cardiology Institute, Lithuanian University of Health Sciences
inga.timofejeva@ktu.edu

A computational framework for the visualization of relationships between ECG parameters using optimal Lagrangian difference matrices is presented. Each participant's ECG (electrocardiogram) data was obtained during a bicycle ergometry exercise. A genetic algorithm optimization scheme is applied to the data in order to construct an optimal matrix describing the dynamics of the cardiovascular system during the load and recovery processes. This technique could provide valuable insight into the specific characteristics of each individual's cardiovascular system.

Analysis of Public Procurement Data Using Social Network Techniques

R. Užupytė[2,3], T. Krilavičius[1,3],

[1] Baltic Institute of Advanced Technology
[2] Vilnius University
[3] Vytautas Magnus University
ruta.uzupyte@bpti.lt

As more and more data become publicly available, the interest in extracting valuable information out of this data receives increasing attention. For each category of data, the advantages of open data analysis may be specific. In this research, we are interested in the open data of public procurement of Lithuania. Such kind of analysis may help reveal corruption and address the issues of transparency. However, the appropriate analysis technique should be carefully selected in order to obtain reliable and valuable results. In this part of the research, we focus on methods of knowledge discovery, such as social network analysis. This methodology was selected due to its ability to identify and visually represent relations among a large amount of data.

Research in High Frequency Statistical Arbitrage Strategies Applied to Microsecond and Nanosecond Information

M. Vaitonis, S. Masteika

Kaunas Faculty, Vilnius University
mantas.vaitonis@knf.vu.lt

Last decade of computer development changed the way the financial instruments are traded in the markets. Humans are no longer necessary to make trades; this task is performed by trading algorithms. The speed of trading is of the most importance. However, there are relatively few academic researches on the increased speed of trading from milliseconds to nanoseconds. In order to address the aforementioned shortcoming, this research measures the differences in the effectiveness of the pairs trading strategies, emerging when microsecond and nanosecond data are included. The effect of the increased speed of data on the pairs trading strategies is analysed. We present different pairs trading strategies and one pair selection algorithm, based on cointegration method. These trading strategies are implemented on five different commodity futures contracts using both microsecond and nanosecond historical data. The effectiveness is measured in accordance with the profit, generated at the end of the trading period.

In order to measure the effectiveness of all presented pairs trading strategies, the Sharpe Ratio method was introduced. The results revealed that all strategies are more effective when subject to higher frequency data of nanoseconds. Best results were obtained by the strategy presented by D. Herlemont, which resulted in the Sharpe Ratio of 2,6388. Higher frequency data provide the trader with the most recent information, allow to react faster and to notice every change in the market in the fastest manner. Thus, it is necessary to go deeper in speed, and nanosecond might be improved to even higher speeds.

Advanced Evaluation Methods of Multiple Application Software Interoperability

A. Valatavičius, S. Gudas

Institute of Data Science and Digital Technologies
Vilnius University
valatavicius.andrius@gmail.com

In informatics and mainly in application software interoperability there are many issues concerning data standardization and data quality, schema matching, orchestration, and choreography but practically no deterministic nor probabilistic evaluation methods have been established to evaluate whether two or more application software systems can be interoperable. This application software interoperability measurement should be the basis for improving interoperability methods. In this research, we explore existing evaluation methods for software interoperability and compare to the possibilities of using fuzzy matching and data mining techniques such as self-organizing maps, random forest, and artificial neural networks. This research is limited to enterprise applications developed using service-oriented architecture and mostly focus on software that uses web services and SOAP protocol for data transfer which meta-data is usually described using standardized WSDL documents. RESTful web service meta-data description is not standardized, and it is difficult to extract meta-data of the data structure. RESTful web services remain as a further improvement on the evaluation method established in this document.

The result of this research is an introduction to created evaluation model that should be a step towards the more standardized way of evaluation of the application software interoperability. Simply put, this research proposes a way towards determining whether two or more application software systems can be interoperable and to what extent. It is an effort to answering what components can be interoperable, to which extent the systems can be interoperable.

Knowledge-Based UML Models Generation Transformation Algorithms from Enterprise Model

I. Veitaitė, A. Lopata

Institute of Applied Informatics
Kaunas Faculty, Vilnius University
ilona.veitaite@knf.vu.lt

The business and information technologies (IT) alignment creates a lot of discussion on its terminology. Specialists and professionals of their field can offer that enterprises need to achieve strategic business and IT alignment to be best in competitiveness. Strategic business and IT alignment affect business efficiency and IT productivity.

Many enterprises generate business-facing roles that have a fundamental commitment to making and keeping connections among business and IT fields in order to support inducing perception and interaction. The capability to clarify business and IT alignment difficulties and problems with the equal clarity should be granted.

The today's situation and the suitability of the research were assessed by analysing the scientific literature, which is related to the information system engineering, information system development life cycle stages, enterprise modelling, UML, ISO standards, MOF architecture, model-driven development and other fields.

The main scope is to present knowledge-based Unified Modelling Language (UML) dynamic models generation from Enterprise model (EM) transformation algorithms. The transformation algorithms description is represented as some UML models with depiction an explanation of significant steps. The decisive result of the generation from enterprise model are UML dynamic models that refer composition of the particular problem domain and may be used by IS developers in further IS life cycle stages.

Research Action designated as COST Action CA15123: The European research network on types for programming and verification (EUTYPES).

Retraining Strategies of Modified SOM for Abnormal Marine Traffic Detection

J. Venskus, P. Treigys, J. Bernatavičienė, V. Medvedev

Institute of Data Science and Digital Technologies
Vilnius University
julius.venskus@gmail.com

The growth of marine traffic around the seaports raise the traffic control problems. These increase the workload for traffic service operators. The automated identification system (AIS) of vessel movement generates significant amounts of data that need to be analysed incrementally to train model as data become available gradually over time. A fast self-learning algorithms for the decision support system are requested to develop that could detect the abnormal vessel movement.

Investigation considers the modified self-organising map (SOM) with introduced pheromone algorithm. The work presents the experiments of different algorithm retraining strategies. The investigation will cover such retraining strategies aspects as random neuron winner initialisation; winner neurons initialised as those from the previously trained network. Different initialisation strategies and the use of pheromone evaporation function will be explored with the view to expose the data amount and classification accuracy relation as well as the time needed for the network retraining. The AIS collected data provided by the Klaipeda seaport will be used in the experiments.

Using Lempel-Ziv Jaccard Distance Measure in Data Stream Classification

K. Ząbkiewicz

Faculty of Economics and Informatics
University of Bialystok
kamil.zabkiewicz@gmail.com

Data streams are recently getting more and more attention. We can see their presence in marketing (e.g. analysis of customer‘s buying activity), healthcare (eg. measuring heartbeat by the wearable sensors to predict heart attack), behaviour analysis (eg. sentiment analysis of the messages in Twitter) or computer network security (eg. analysis of the network packets to detect the intrusion). Classification of this sort of data is a very challenging problem.

In our previous approach to deal with the classification of the streaming data sets, we took the Normalized Compression Distance (NCD). The main reason for choosing this method was that the NCD is parameter-free, feature-free, and alignment-free. There was no need to set various parameters, such as learning speed, the number of epochs, weights for features, etc.

New non-standard distance measure was proposed. It is called Lempel-Ziv Jaccard Distance (LZJD). Authors claim that is also parameter free and works with the raw binary data. What is more, this method is the real distance measure, satisfying triangle inequality and is computed faster than NCD. The basis of this approach is the computation of the minimum hash that leads us to Jaccard similarity measure. This similarity can be simply converted into the distance by subtracting it from 1. The main motivation of this work is to test how this measure works in data streams. We performed some experiments. Our testing methodology was based on the previously proposed one. We used nearest neighbor classifier with non-standard distance measure. Preliminary results will be shown in this work. We will also compare them with earlier ones done with Normalized Compression Distance.

How Can Statistical Physics Help in Data Mining Tasks?

Massimiliano Zanin

Centre for Biomedical Technology
Universidad Politécnica de Madrid, Spain
massimiliano.zanin@ctb.upm.es

The statistical physics concept of complex network shares many characteristics with data mining, more than what may prima facie appear. Not only do both share the same general goal, that of extracting information from data to ultimately create compact and quantifiable representations; but they also often address similar problems too. If these two scientific fields have mostly walked separated paths, these are now starting to converge. This talk will shortly review the concepts and hypotheses underlying both approaches, and how they have historically been used to perform different data-related tasks. We will additionally discuss how complex networks and data mining are expected to interact in the future, with a special focus on the emerging field of systems medicine.

Impact of Different Geometric and Functional Principles on Perception of Acceptability Areas: Differences in Computing and Humanities'Students

L. Zariņa, J. Šķilters, J. Borzovs

Faculty of Computing
University of Latvia
liga.zarina@lu.lv

Different spatial descriptions induce different spatial areas of acceptability in observers (Logan, & Sadler, 1996). In our study, we have tested different types of geometric and functional objects and the areas that users categorize as appropriate. Further, we also assumed that there might be differences between students of different fields (computing and humanities). Our study was complemented with the test of spatial abilities and skills that we explored both independently and dependently on the test of acceptability areas.

9th International Workshop

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Sudarytoja Jolita Bernatavičienė
Parengė spaudai ir išleido
Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų institutas
Akademijos g. 4, Vilnius
Spausdino UAB „BMK leidykla“
J.Jasinskio g. 16, Vilnius

General Sponsors

Research Council of Lithuania

algoritmų sistemos

Main Sponsor

Critical IT infrastructure

baip

Sponsors

NRD Cyber Security

NetCode

BALTIC AMADEUS