Утверждено академическим руководителем

Образовательной Программы

«Системы больших данных» проф. Мальцева С.В.,

уровень образования: магистратура

одобрено Академическим Советом программы

14.10.2017

**1. Nonlinear dimension reduction methods**

Goals: describe and implement the non-linear dimension reduction methods for revealing the hidden patterns of data. The primary focus of the term paper should be on: (i) ISO-MAP method which enables to estimate the intrinsic geometry of data manifold based on a rough estimate of each data point’s neighbors (ii) t-NSE approach based on probabilistic recognition of the similar objects (ii) GP-LVM method based on the paradigm of space of latent variables. Other non-linear methods for dimension reduction such as non-negative matrix factorization or locally linear embedding may also be considered.

References:

1. ISO-MAP: http://isomap.stanford.edu
2. t-SNE <http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>
3. GP-LVM https://www.youtube.com/watch?v=l98Lw9KHzfc

**Supervisor: Petropavlovskiy**

**2. Markov Chain Monte-Carlo methods for Machine Learning**

Goals: describe and implement the Markov chains for probabilistic machine learning. The term paper should cover the main building blocks of modern Markov chain Monte Carlo such as the Metropolis-Hastings algorithm, Hamiltonian Monte Carlo (HMC) and Stochastic Gradient Decent - based MC. The applications of the MCMC methods such as dynamic Bayesian networks, optimal filtering, deep Gaussian networks should be considered.

References:

1. MCMC for machine learning <https://www.youtube.com/watch?v=qBf5EBdEw7Q>
2. http://www.cs.ubc.ca/
3. arnaud/andrieu\_defreitas\_doucet\_jordan\_intromontecarlomachinelearning.pdf
4. https://media.nips.cc/Conferences/2015/tutorialslidesnips-2015-monte-carlo-tutorial.pdf
5. <https://en.wikipedia.org/wiki/Inverse_transform_sampling>
6. http://www.mdpi.com/1996-1073/8/6/5538/htm

**Supervisor: Petropavlovskiy**

**3. Methods of Multi-Task Learning/**

## Goals: describe the aims, concept and methods of multi-task learning. The students should overview the basic methods of MT learning, hard and soft parameter sharing approaches in the first place. The connection between MTL and deep neural networks should be traced and clearly explained. The Cross-stitch Networks, low supervision paradigm and tensor factorization approaches also deserve a separate section of the term paper.

## References:

## Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing. Proceedings of the 25th International Conference on Machine Learning - ICML ’08, 20(1), 160–167.

## Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., & Pande, V. (2015). Massively Multitask Networks for Drug Discovery. [https://doi.org/https://arxiv.org/abs/1502.02072](https://doi.org/https:/arxiv.org/abs/1502.02072)

## Caruana, R. "Multitask learning: A knowledge-based source of inductive bias." Proceedings of the Tenth International Conference on Machine Learning. 1993

## Wikipedia

**Supervisor: Petropavlovskiy**

**4. Big data analysis in the industry N**

Abstract: Despite of all known characteristics of Big Data Systems (BDS) , they tend to be applied differently in various industries. Banking and Telecommunications, Public administration and Higher Education Institutions - all of those diverse industries have specific requirements for Big data Systems processing and application. The task for student here would be to make a systematic literature review of BDS application in existing projects (mainly using academia and business literature) and to create an integrated table with requirements from business and delivered results. Student may also use data from a company he/she works for, this however might require an additional permission from the corporate compliance group or signing an NDA.

This topic lays also foundation for a Master thesis and could be.

References:

1. Anna, Kuraeva, and Kazantsev Nikolay. "Survey on Big Data Analytics in Public Sector of Russian Federation." *Procedia Computer Science* 55 (2015): 905-911.
2. <http://ieeexplore.ieee.org/document/7067026/?arnumber=7067026>

# 3.     <http://www.sciencedirect.com/science/article/pii/S2213846313000114>

# 4.     <http://dl.acm.org/citation.cfm?id=2463712>

# Supervisor: Nikolay S. Kazantsev

**5. Virality of modern business models**

Abstract: Virality is one of the modern properties of successful business models. It describes the positive effects while social media marketing and enables users massively try and share positive references about the product or service. Recent success of popular apps such as MSQRD, PRISM and many others demonstrate how important is the integration between information systems and psychology. Despite of already existing competitors, some researchers claim that the virality was the decisive factor for their success, its spread looked more like a flash-mob than a planned marketing campaign.

In this course work student is expected to integrate case study material from various (mostly business) literature and create a set of recommendations for a start-up.

This hottest topic lays a good background for further master thesis and PhD dissertation worldwide.  It gives lots of tacit knowledge that one could apply starting own company or promoting products or services.

References:

1. <http://jonahberger.com/>
2. <https://books.google.ru/books?hl=ru&lr=&id=z_VBKjscPLkC&oi=fnd&pg=PP2&dq=Virality+of+modern+business+models&ots=IxqpCqXbk0&sig=0ri01YyXfU1h24flGkli22xi7Eo&redir_esc=y#v=onepage&q&f=false>
3. <https://www.amazon.com/Invisible-Influence-Hidden-Forces-Behavior/dp/1476759693/ref=as_li_ss_tl?ie=UTF8&linkCode=sl1&tag=jbsite-20&linkId=11406899823cffd33264b6d0ea568c6f>
4. <https://www.amazon.com/Contagious-Things-Catch-Jonah-Berger/dp/1451686579/>

# Supervisor: Nikolay S. Kazantsev

**6. Subject oriented approach to (big) Data-intensive workflows in aerospace industry (running project)**

Abstract: Subject-oriented business process management  (S-BPM) refers to a formal notation system for describing and executing business processes. In this system, the focus is on the subject or individual actor. S-BPM differs from other modelling languages in its low number of modelling symbols and close approximation of natural language—and therefore its similarity to the way people generally gather information, think and communicate. With this focus on the subject, actors can model their processes from a first-person perspective and experience them immediately. Because S-BPM processes can be executed immediately after modelling.

In this work a student is expected to find intersections between topics of S-BPM and Big Data. Would processing Big Data influence models built around employees and customers ? Student is expected to draw a set of S-BPM models for the cases of Big Data application in the aerospace industry.

References:

1.     Fleischmann, Albert, et al. *Subject-oriented business process management*. Springer Publishing Company, Incorporated, 2014.

2.     <https://www.metasonic.de/en/s-bpm>

3.     Schiefer, Josef, et al. "Process information factory: a data management approach for enhancing business process intelligence." *e-Commerce Technology, 2004. CEC 2004. Proceedings. IEEE International Conference on*. IEEE, 2004.

4.     Fleischmann, Albert, et al. "Subject-oriented modeling and execution of multi-agent business processes." Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02. IEEE Computer Society, 2013.

# Supervisor: Nikolay S. Kazantsev

**7.**      **Week signal revealing based on big data tools/Выявление слабых сигналов изменения ситуаций на основе инструментов работы с большими данными**

Abstract: For the problem solving there are hard and soft approaches to system modeling, monitoring. At that, ill-conditioned systems are characterized by problems that hardly can be extracted from analyzed control situation. This limits applicability of traditional methods for finding optimal (or even satisfactory) solution to control problems for such systems. In this process, we consider the intelligent activity of problem solver who should deep insight in problem situation of system under control and could offers a strategic decision based on analysis different scenarios. Problem solver is considered as a group expert thus we have to analysis and monitor a lot of information source: expert opinions and big data sets.

The characteristic feature of ill-structured system analysis consists in that the process of preparation and making decisions on control of ill-structured system is, as a rule, a group activity. Each participant of this process represents problem situation based on “his own” inherent representations and knowledge of situation (images, models of world). An image of world includes a set of convictions, perceptual features, cost and practical rules of an individual that guide his activity and influence the process of problem situation resolution. Thus, decision preparation and making in control problems for ill-structured systems should be considered as complex intellectual process of problem resolution that cannot be reduced to solely rational choice. With modern tools It seems to construct a supporting system that almost automatically clarify knowledge about problem, recommend significant variables. With processes automation and information development of society we got to big data set, produced by enterprise information system, internet. However the problem of lack information has not been solved. We got another problem with skewness in data, with unintegrated data and with quality of data.

* Information overload and information lack exist simultaneously.
* Basic principal of big data approach to statistical data analysis leads to loss possibility of weak signal analysis. In case of complex system development except for typical scenarios unexpected can be actualized. From the point of view supporting IT-system it’s needed only monitor of changes of putted factors of expert’s scenario. Our practical experience shows disadvantages of data-mining tools without supervising model, which can only reveal statistically significant influences, factors, or data clusters in general.
* The significance problem for analytical decision making is problem of data quality. For the data cleansing spend huge resources, but it is fact that more than 40% from total volume of data is “dirty”. There are variety of industrial tools to data cleansing from syntaxes mistakes and mistakes with missing data, but it remain time-consuming. There are two major reason of dirtiness of data: human induced (directly or indirectly embedded in IS especially in interface for data input); machine induced (technical interrupt leading to appear mistakes on a stage of data saving, updating.

It is important to note that the observed factors of the situation are heterogeneous, and accordingly, they have different arrays of accumulated data and information that cannot use similar monitoring tools.

In the present conditions of increasing instability and the speed of changes in situations, it is increasingly difficult to predict the nature of these changes and respond in a timely manner to them. Therefore, the development of approaches to monitoring as systematic monitoring of the system parameters with the purpose of revealing weak signals - early and odd signs of the onset of crisis situations, events that are essential for the development of the system - is becoming more important. (The concepts of weak and strong signals, as well as related levels of awareness of imminent changes, were introduced by I. Ansoff [5].)

In the framework of studies conducted to monitor the effect of weak effects of changes in observed factors.

References:

1. Avdeeva Z., Raikov A., Ermakov A. [Big Data Refining on the Base of Cognitive Modeling](https://publications.hse.ru/view/198195185) // IFAC-PapersOnLine. 2016. Vol. 49. No. 32. P. 147-152. [doi](http://doi.org/10.1016/j.ifacol.2016.12.205" \t "_blank)
2. Ansoff, I. H. Implanting strategic management. New Jersey: Prentice Hall. 1984.
3. Ansoff, I. H. Managing strategic surprise by response to weak signals. California Management Review, 18(2), 21-33. 1975.
4. Ansoff, Igor. Implanting Strategic Management. Prentice // Hall International Inc. 1984.
5. Cooper A. Weak Signals and Text Mining I – An Introduction to Weak Signals. URL: http://blogs.cetis.org.uk/adam/2011/05/12/weak-signals-andtext-mining-i-an-introduction-to-weak-signals/
6. Christos Ch., Awais R. and Paul J. Taylor. Weak Signals as Predictors of Real-World Phenomena in Social Media // 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Pp. 864-871. DOI: 10.1145/2808797.2809332
7. Ponomareva Julia V., Sokolova Anna V.. The identification of weak signals and wild cards in foresight methodology: stages and methods // Working papers. Series: Science, Technology and Innovation. National Research University Higher School of Economics. WP BRP 46/STI/2015. 2015. 26 p.
8. Tabatabaei N.. Detecting Weak Signals by Internet-Based Environmental Scanning // Waterloo, Ontario, Canada, 2011.
9. Yoon, J. Detecting weak signals for long-term business opportunities using text mining of Web news // Expert Systems with Applications, 39(16), 2012. 12543-12550.

# Supervisor: Zinaida K. Avdeeva

**8.**     **Tools for strategy correction in condition of changing business environment /Инструменты коррекции стратегии развития системы   
в режиме отслеживания изменений во внешней среде**

Abstract: The study raleted with problem of strategy monitoring and following correction of strategy because of exernal environment change and lead to problem of goals achivment. It need to develop approach to strategy control based on revealing the problem situation (considered as deviation between goals state and achievable state in existing condition) and solving it in whole cycle of external and internal changes monitoring.

In general, the control of system can be represented as construction of strategy for the system development, defining the main goals and general directions for their reaching, and its implementation. Revealing the system development problems influencing negatively achievement of strategic control goals is one of the key stages of construction of strategy.

Thus, the control problem consists in transfer of system into one of a nearest state corresponding to goal image. At that, the proposed approach allows determining the system state in both values of model factors and rates of factor changes. We presents the general scheme of method for strategy monitoring on the base of linear dynamic models on the base of cognitive map in regard to business systems .

Strategic monitoring plays one of the key roles in a cycle of supporting the formation, implementation and correction of the socio-economic system (SES) (industrial enterprises, corporations, cities, etc.) development. Its main purpose is to monitor the achievement of strategic goals of a system development in a changing external environment. Strategic monitoring is aimed at the timely detection of (1) favorable and unfavorable changes in the external environment, and (2) changes in the system properties (its strengths and weaknesses), which may affect the implementation of the adopted development strategy.

Considering complexity, weak formalizability of modern situations, on the one hand, strategic management experts note the necessity of expansion of the corresponding scientific and methodical support. On the other hand, methods of analysis and forecasting of system development on the basis of casual models, have proven to be pretty effective. Casual model (for example, cognitive map of a situation) is a model representing the expert knowledge of a situation in the form of structure of causal influences.

References:

1. Avdeeva Z., Kovriga S. The technology of the strategic goal-setting and monitoring of a system development on the basis of cognitive mapping In: Procedia Computer Science. 4rd International Conference on Information Technology and Quantitative Management, ITQM 2017  - IN PRINT.
2. Avdeeva Z., Kovriga S. (2008). Cognitive Approach in Simulation and Control. Proceedings of the 17th IFAC World Congress, Seoul, Korea, July 6-11, pp. 1613- 1620.
3. Avdeeva Z., Kovriga S. [Diagnosing of The Problematic Situation in Manufacturing System Development Based-On-Cognitive Map](https://publications.hse.ru/view/138532820) // Manufacturing Modelling, Management, and Control. 2013. Vol. 7, P.1. P. 964-968.
4. Gubanov, D., Korgin, N., Novikov, D., Raikov, A. (2014). E-Expertise: Modern Collective Intelligence, Springer. Series: Studies in Computational Intelligence, Vol. 558, XVIII, 112 p.
5. Kahraman, Cengiz et al. (2008) Fuzzy Multi-Criteria Decision Making -Springer. Pp.380
6. [Komarov M. M.](https://www.hse.ru/en/org/persons/14587151), Avdeeva Z. [Customer experience management for smart commerce based on cognitive maps](https://publications.hse.ru/view/161776450) // Annals of Data Science. 2016

# Supervisor: Zinaida K. Avdeeva

9. **The application of the theory of self-organized criticality to the detection of catastrophic events in microblogging social networks/Приложение теории самоорганизованной критичности к обнаружению катастрофических событий в микроблоггинговых социальных сетях**

During the past few decades, it has become increasingly apparent, that many chaotic and complicated systems do not yield to traditional analysis. Many composite systems naturally evolve to a critical state in which a minor event starts a chain reaction that can affect any number of elements in the system. Nowadays the interpretation of the occurrence of catastrophic events in composite systems is based on the theory of self-organized criticality (SOC) also known as the third paradigm of nonlinear dynamics.

In the course work it is planned to explain the appearance of disasters in microblogging social networks at the level of their topological structure and at the level of signals of networks, regardless of their structure.

The main resources used for the study are Mozdeh Big Data Text Analysis, NodeXL.

References:

1. Skvoretz, J. Complexity Theory and Models for Social Networks. Complexity. 2003(8). 47-55
2. Steyer, A., Zimmermann, J.B. Self Organised Criticality in Economic and Social Network. The Case of Innovation Diffusion. Lecture Notes in Economics and Mathematical Systems. 2001(503). 27-41
3. Bak P., Tang C., Wiesenfeld K. Self Organised Criticality. Phys. Rev. Lett. 1987(59). 381

# Supervisor: Andrey V. Dmitriev

10. **A nonlinear dynamical approach to the interpretation of microblogging network complexity /Приложение нелинейной динамики к интерпретации сложности микроблоггинговой социальной сети**

The aim of the study is to answer to the following question: are microblogging social networks complex systems? The theoretical and practical importance of receiving an answer to this question is due to the fact that complex systems have a tendency to generate catastrophic events. Catastrophic events mean unexpected events that cannot be predicted or extraordinary events that stand out from a series of related events. An important feature of the study is the fact that definition of the complexity of social networks is based on an analysis of the external manifestations of the system, regardless of its internal structure. This is a non-linear dynamic analysis of network time series obtained from Mozdeh Big Data Text Analysis and correlating the analysis results with the main features of complex systems.

References:

1. Skvoretz, J. Complexity Theory and Models for Social Networks. Complexity. 2003(8). 47-55
2. Steyer, A., Zimmermann, J.B. Self Organised Criticality in Economic and Social Network. The Case of Innovation Diffusion. Lecture Notes in Economics and Mathematical Systems. 2001(503). 27-41
3. Bak P., Tang C., Wiesenfeld K. Self Organised Criticality. Phys. Rev. Lett. 1987(59). 381

# Supervisor: Andrey V. Dmitriev

**11. Boilerplate detection for web pages/Исследование методов выделения шаблонов на веб страницах**

Abstract: Research and development of methods for analyzing web pages in order to extract the main content or build boilerplate for web page.

References:

* Cai, D., Yu, S., Wen, J.-R., Ma, W.-Y.: Extracting Content Structure for Web Pages Based on Visual Representation.
* C. Kohlschütter, P. Fankhauser, W. Nejdl. Boilerplate Detection using Shallow Text Features.

# Supervisor: Alexandr Bolkhoviteanov

**12. Duplicate detection in yellow pages services/Обнаружения дубликатов в справочниках организаций**

Abstract: Research and development of algorithms for searching similar organizations on maps in the yellow pages service database of contact information.

# Supervisor: Alexandr Bolkhoviteanov

**13. Machine learning in computational morphology/Исследование методов машинного обучения в задачах автоматического морфологического анализа**

Abstract: Research of the application of machine learning methods, including neural networks in the tasks of automatic morphological analysis and synthesis.

References:

* Durrett G., DeNero J. Supervised Learning of Complete Morphological Paradigms.
* Ruokolainen T., Kohonen O, etc. A Comparative Study of Minimally Supervised Morphological Segmentation.

# Supervisor: Alexandr Bolkhoviteanov

**14. New services based on open data/ Новые сервисы на основе открытых данных.**

Abstract: There should be presented new service based on open data concept, it should be described and analyzed with possible implementation (i.e. as mobile application). Literature review should show current existend solutions not just in Russia but all around the world. New services should not only utilize open data, but use open data as one part of the data source for the service.

**Supervisor: Mikhail M. Komarov**

**15. Context-based services/ Контекстно-ориентированные сервисы**

Abstract: Context-based services are services where real-world and digital data are utilized to provide new services and value to existing or potential customers. These services depend on "the ability for firms to aggregate and apply smart analytics to a widening array of new data — from customers' location information and social networks, to mobile apps, blogs, tweets, purchasing history, and more. The more that firms know about where their customers are, and what they're doing — analyzed in real-time — the more they will be able to deliver immersive and valuable services specifically tailored to them."

**Supervisor: Mikhail M. Komarov**

**16. New generation networks QoS for users/ Качество представления услуг для пользователей в сетях нового поколения.**

Abstract: Technologies for data transmission should be analyzed together with different approaches to measure quality of service for the end-users of the services which are provided through those networks. Positive and negative effects should be considered at the end for business-processe with the use of those technologies and dependancy of the proposed QoS metrics and business.

**Supervisor: Mikhail M. Komarov**

**17. Mobile technologies and Big Data: privacy issues / Мобильные технологии и большие данные: вопросы безопасности**

Abstract:  How to protect user’s privacy in Web-based applications (including websites, mobile apps etc). One way is to think of an efficient way to protect from web tracking in the light of the new GDPR regulation (and also ePrivacy regulation, that particularly covers web tracking). W3C Tracking Protection working group (<https://www.w3.org/2011/tracking-protection/>) has proposed a DNT policy that allows users to set up their privacy preferences, and describes how this policy should be enforced in Web-based applications. The same might apply to the mobile users. If there are any other proposals how to arrange privacy protection during big data time with the use of mobile technologies - students are welcome to propose and discuss it.

**Supervisor: Mikhail M. Komarov**

**18. Customer behavior management based on geolocation data / Управление поведением потребителей с использованием их данных геолокации.**

Abstract: Since companies have already begun to digitally transform their existing customer loyalty programs into mobile applications, new approaches and possibilities are resulting. By collecting and processing the mobility patterns of every single customer combined with the delivery of appropriate offers and incentives at the right time and the right place, may form the future’s customer approach. The processing of the customers’ location data may reveal new information and insights from the customer buying behavior

**Supervisor: Mikhail M. Komarov**

**19. The Creation Administrative and Automation Tools for Database of Big Data Project on Base Radioastronomy Data/Создание административных инструментов и средств автоматизации базы данных проекта больших радиоастрономических данных.**

Abstract: In 2012 the multi-beam feed array became operational on the BSA FIAN radio telescope [1,2,3]. Today it is capable of 24-hour observation using 96 beams in declination in the sky of -8 to +42 degrees (aboutly 40% of the sky) in the 109-111.5 MHz frequency band. The data receiving in 32 band mode (plus one common band) with a time constant of 12.5 ms (80 times per second), 33x96x80 four-byte data samples are produced per second, which equates to a daily data production of 87.5 gigabytes (up to 32 terabytes yearly, to this time 70 TB in archives). This data is an enormous opportunity for both short and long-term monitoring of various classes of radio sources (including radio transients), space weather and the Earth's ionosphere monitoring, search for different classes of radio sources, etc. For example, today more than 20 new pulsars from this data have been discovering (from 2,5 thousands in total), that one from best results for pulsar found commands in the word. At this moment tested different ways of data processing.

For these big data is constructed database on base of Postgresql [4]. But this database have not good administrative and automation tools for addition and pre-processing our big data science project. So we have in our data base only for 2012-2013, we not add last data and dada for big regime of observation. We need in systematic information from observational servers, also as preliminary preparation of data for stream processing (other course work). The science objesct table is need for constructing, it is need constructed some auto information for site of project (other course work) and so one.

Need skills from You: SQL (Postgresql preferable), Linux, html, good skills also with some scripts language (for example Perl or/and Python) and C++.

Scientific adviser: Dr. Samodurov V.A.1,2 (<http://www.hse.ru/org/persons/37253680>),

1 National research university Higher school of economics, Moscow, Russia

2Pushchino Radio Astronomy Observatory ASC LPI, Pushchino, Russia

References

1. Oreshko V.V. et al : 2012, Transactions of the Institute of Applied Astronomy (Russia), v. 24, p. 80
2. <http://astro.prao.ru/> - on-line database from daily monitoring of the radio sky in the 109-111.5 MHz (from 7 July 2012 to 20 October 2013).
3. Samodurov V. A., Rodin A.E., Kitaeva M. A., Isaev E., Dumsky D. V., Churakov D.D., Manzyuk M.O.: The daily 110 MHz sky survey (BSA FIAN): online database, science goals data processing by distributed computing. // Труды ХVII международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains” (DAMDID)). Обнинск: НИЯУ МИФИ, 2015. P. 127-128.
4. <https://www.postgresql.org/docs/9.5/static/docguide.html>

**Supervisor: Samodurov V.A**

**20. The processing of large radio astronomical data using a GPU (NVIDIA graphics cards)/Обработка больших радиоастрономических данных при помощи GPU (графических плат NVIDIA)**

Abstract: In 2012 the multi-beam feed array became operational on the BSA FIAN radio telescope [1,2,3]. Today it is capable of 24-hour observation using 96 beams in declination in the sky of -8 to +42 degrees (aboutly 40% of the sky) in the 109-111.5 MHz frequency band. The data receiving in 32 band mode (plus one common band) with a time constant of 12.5 ms (80 times per second), 33x96x80 four-byte data samples are produced per second, which equates to a daily data production of 87.5 gigabytes (up to 32 terabytes yearly, to this time 70 TB in archives). This data is an enormous opportunity for both short and long-term monitoring of various classes of radio sources (including radio transients), space weather and the Earth's ionosphere monitoring, search for different classes of radio sources, etc. For example, today more than 20 new pulsars from this data have been discovering (from 2,5 thousands in total), that one from best results for pulsar found commands in the word. At this moment tested different ways of data processing.

But at this moment we have not good instrument for stream computing of these big data. We need in launch the streaming data on various types of high-performance computing systems, including using a GPU (NVIDIA graphics cards) [4]. You must start-up, configuring and run NVIDIA graphics cards by CUDA, translate in CUDA language of some computing and astronomical algorithms (sorting, correlation, Fourier analytics etc. also as some astronomical algorithms). Need skills from You: C++ , CUDA technology, SQL (Postgresql preferable [5]), good skills also will Perl or/and Python.

Scientific adviser: Dr. Samodurov V.A.1,2 (<http://www.hse.ru/org/persons/37253680>),

1 National research university Higher school of economics, Moscow, Russia

2Pushchino Radio Astronomy Observatory ASC LPI, Pushchino, Russia

References

1. Oreshko V.V. et al : 2012, Transactions of the Institute of Applied Astronomy (Russia), v. 24, p. 80
2. <http://astro.prao.ru/> - on-line database from daily monitoring of the radio sky in the 109-111.5 MHz (from 7 July 2012 to 20 October 2013).
3. Samodurov V. A., Rodin A.E., Kitaeva M. A., Isaev E., Dumsky D. V., Churakov D.D., Manzyuk M.O.: The daily 110 MHz sky survey (BSA FIAN): online database, science goals data processing by distributed computing. // Труды ХVII международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains” (DAMDID)). Обнинск: НИЯУ МИФИ, 2015. P. 127-128.
4. <http://www.nvidia.ru/object/cuda-parallel-computing-ru.html>
5. <https://www.postgresql.org/docs/9.5/static/docguide.html>

**Supervisor: Samodurov V.A**

**21. Processing of large radio astronomical data using supercomputer computations/Обработка больших радиоастрономических данных при помощи суперкомпьютерных вычислений**

Abstract: In 2012 the multi-beam feed array became operational on the BSA FIAN radio telescope [1,2,3]. Today it is capable of 24-hour observation using 96 beams in declination in the sky of -8 to +42 degrees (aboutly 40% of the sky) in the 109-111.5 MHz frequency band. The data receiving in 32 band mode (plus one common band) with a time constant of 12.5 ms (80 times per second), 33x96x80 four-byte data samples are produced per second, which equates to a daily data production of 87.5 gigabytes (up to 32 terabytes yearly, to this time 70 TB in archives). This data is an enormous opportunity for both short and long-term monitoring of various classes of radio sources (including radio transients), space weather and the Earth's ionosphere monitoring, search for different classes of radio sources, etc. For example, today more than 20 new pulsars from this data have been discovering (from 2,5 thousands in total), that one from best results for pulsar found commands in the word. At this moment tested different ways of data processing.

But at this moment we have not good instrument for stream computing of these big data. We need in launch the streaming data on various types of high-performance computing systems, including using supercomputer computations. [4]. You need must translate on supercomputer of some computing and astronomical algorithms (sorting, correlation, Fourier analytics etc. also as some astronomical algorithms). Need skills from You: C++, high performance computing technology, SQL (Postgresql preferable [5]), good skills also will Perl or/and Python.

Scientific adviser: Dr. Samodurov V.A.1,2 (<http://www.hse.ru/org/persons/37253680>),

1 National research university Higher school of economics, Moscow, Russia

2Pushchino Radio Astronomy Observatory ASC LPI, Pushchino, Russia

References

1. Oreshko V.V. et al : 2012, Transactions of the Institute of Applied Astronomy (Russia), v. 24, p. 80
2. <http://astro.prao.ru/> - on-line database from daily monitoring of the radio sky in the 109-111.5 MHz (from 7 July 2012 to 20 October 2013).
3. Samodurov V. A., Rodin A.E., Kitaeva M. A., Isaev E., Dumsky D. V., Churakov D.D., Manzyuk M.O.: The daily 110 MHz sky survey (BSA FIAN): online database, science goals data processing by distributed computing. // Труды ХVII международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains” (DAMDID)). Обнинск: НИЯУ МИФИ, 2015. P. 127-128.
4. <https://parallel.ru/>
5. <https://www.postgresql.org/docs/9.5/static/docguide.html>

**Supervisor: Samodurov V.A**

**22. Analysis of the approaches to designing the systems of cross-lingual intelligent information access/Анализ подходов к разработке систем межъязыкового интеллектуального доступа к информации.**

Abstract: During last decade, one has been able to observe a quickly growing interest in the design of computer intelligent agents fulfilling cross-lingual information retrieval (CLIR) on the Web. It is a consequence of emerging a huge, permanently increasing number of Web-sources in languages being different from English. In September 2012, a seminar on Multilingual Semantic Web (MSW) was organized in Germany in the Dagstuhl Castle. The proceedings of this seminar contain the following data: in the year 2010, the number of non-English-speaking Internet users was three times higher than the number of English-speaking users (1430 million vs. 536 million users). That is why the problem of developing a MSW is very topical. It is broadly accepted that a promising approach to the realization of CLIR on the Web is employing a special semantic language-intermediary (SLI) in order to represent in the same format both semantic content of a user query and semantic content of the analysed fragment of a text in natural language (NL). The task of the term paper is to analyse the principoal existing approaches to designing the systems of cross-lingual intelligent information access, using a semantic language-intermediary.

References:

 1. Fomichov, V. A. [SK-languages as a Powerful and Flexible Semantic Formalism for the Systems of Cross-Lingual Intelligent Information Access](https://publications.hse.ru/view/208921070) // Informatica. An International Journal of Computing and Informatics. 2017. Vol. 41. No. 2. P. 221-232.

2. Lei Y., Uren V., Motta E. SemSearch: A Search Engine for the Semantic Web. S. Staab and V. Svatek (Eds.), EKAW 2006 - Proc. 15th Intern. Conf. on Knowledge Enguineering and Knowledge management, 2-6 Oct. 2006, Podebrady, Czech Republic. Lecture Notes in Artificial Intelligence. Berlin, Heidelberg: Springer, 2006, vol. 4248, pp. 238-245.

3. Fomichov V. A., Kirillov A. V. A Formal Model for Constructing Semantic Expansions of the Search Requests about the Achievements and Failures. Ramsay A., G. Agre G. (Eds). Artificial Intelligence: Methodology, Systems, and Applications. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, vol. 7557, pp. 296–304.

4. Aggarwal N., Polajnar T., Buitelaar P. Cross-Lingual Natural Language Querying over the Web of Data Metais E., Meziane  F., Saraee M., Sugumaran V., Vadera S. (eds.). Natural Language Processing and Information Systems. 18th Intern. Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 2013, Proceedings. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, vol. 7934, pp. 152-163.

5. Fomichov, V.A.: Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Springer, New York, Dordrecht, Heidelberg (2010a)

6. Fomichov, V.A.: Theory of K-representations as a Comprehensive Formal Framework for Developing a Multilingual Semantic Web. Informatica. An International Journal of Computing and Informatics  34 (3), 387-396 (2010b).

7. Fomichov, V.A.: A Broadly Applicable and Flexible Conceptual Metagrammar as a Basic Tool for Developing a Multilingual Semantic Web. In: Metais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds.) NLDB 2013. LNCS, Vol. 7934, 249-259. Springer, Heidelberg (2013)

8. Fomichov, V.A.: SK-Languages as a Comprehensive Formal Environment for Developing a Multilingual Semantic Web. In: Decker, H., Lhotska, L., Link, S., Spies, M, Wagner, R. R. (eds.) Database and Expert Systems Applications, 25th Intern. Conference, DEXA 2014, Munich, Germany, September 1-4, 2014, Proceedings, Part I, LNCS, Vol. 8644, 394-401. Springer, Cham, Heidelberg (2014)

# Supervisor: Vladimir A. Fomichov

**23.The analysis of knowledge-based approaches to natural language processing in business informatics/Анализ основанных на знаниях подходов к компьютерной обработке естественного языка в бизнес-информатике**

Abstract: Business Intelligence (BI) is a term used for denoting the models, methods, and applied computer systems that aim to provide businesses with support for decision-making  by means of collecting and processing information relevant to business leaders. The huge amount of textual information that is available in online sources and in corporative data bases makes it nearly impossible to follow and analyse all the relevant sources manually or to apply traditional storage and retrieval methods. Useful business information is hidden in  the mass of data that is permanently increasing and evolving.

Information Extraction (IE) deals with extracting relevant and high-quality information  from unstructured sources, , such as texts in natural language (NL) and document collections. The extraction process involves representing information contained in textual data  in a structured and normalized way. The Ontology-Based Information Extraction (OBIE) combines the processes of extracting and storing information and uses an ontology (a knowledge base) for extracting significant business events from NL-texts. The most broad prospects for representing the extracted information in a formal way are opened the Theory of K-representations (knowledge representations) developed by V. A. Fomichov. This theory includes, in particular, a new class of formal languages called SK-languages (standard knowledge languages).

References

1.      Arendarenko, E., Kakkonen, T. Ontology-based information and event extraction for business intelligence. In: A. Ramsay. G. Agre (Eds.). Artificial Intelligence: Methodology, Systems, and Applications. 15th Intern. Conference, AIMSA 2012, Varna, Bulgaria, September 2012, Proceedings. Springer, Lecture Notes in Artificial Intelligence, 2012, Vol. 7557, p.89-102.

2.      Wang C., Xiong M., Zhou Q., Yu Y. PANTO: a portable NL-interface to ontologies 4th European Semantic Web Conference Proceedings. Springer, 2007, p. 473-487.

3.      Fomichov, V.A.: Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Springer, New York, Dordrecht, Heidelberg, 2010.

4.      Razorenov, A. A., Fomichov, V. A. [A new formal approach to semantic parsing of instructions and to file manager design](https://publications.hse.ru/view/194073739" \t "_blank), In: Database and Expert Systems Applications. 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings Vol.. 9827. Part I. London, New York, Dordrecht, Heidlberg, Cham: Springer, 2016. P. 416-430.

# Supervisor: Vladimir A. Fomichov

**24. Большие данные в биоинформатике и здравоохранении / Big Data in Bioinformatics and Healthcare**

Abstract: Big data technologies are increasingly used for bioinformatics and health-care research. Large amounts of biological and clinical data have been generated and collected at an unprecedented speed and scale. Progress in bioinformatics and healthcare can be achieved only through effective, correct use of tools for big data processing. The role of big data techniques in bioinformatics applications is to provide data repositories, computing infrastructure, and efficient data manipulation tools for investigators to gather and analyze biological information. Biomedical scientists are facing new challenges of storing, managing, and analyzing massive amounts of datasets. For example, DNA sequencing is on the path to becoming an everyday tool in medicine; the new generation of sequencing technologies enables the processing of billions of DNA sequence data per day. However, there lack efficient big data architectures and tools for many important bioinformatics problems.

Term paper can be devoted to the following issues: Big data problems in bioinformatics; Methodologies and tools for big data analytics in bioinformatics; Computational biology perspective in big data era: challenges and future goals; Advanced datamining using RNAseq data; Analysis of genomic data in a cloud-computing environment; etc.

References

1. V. Marx, “Biology: The big challenges of big data,” Nature, vol. 498, no. 7453, pp. 255–260, 2013.
2. B. Wang, R. Li, W. Perrizo, “Big Data Analytics in Bioinformatics and Healthcare,” IGI Global (USA), 2014, 528 p.
3. M.D. Lytras, P. Papadopoulou, “Applying Big Data Analytics in Bioinformatics and Medicine,” IGI Global (USA), 2017, 465 p.
4. “Bioinformatics: Concepts, Methodologies, Tools, and Applications,” Ed.: Information Resources Management Association (USA), IGI Global (USA), 2013, (3 Volumes) 1783 p.
5. M. Reshma, M.A. Supaksha, N. Hemalatha “Application of Big Data in Bioinformatics,” Int. J. of Latest Trends in Engineering and Technology, Special Issue SACAIM 2016, pp. 206-212.
6. Jake Luo, Min Wu, Deepika Gopukumar, Yiqing Zhao “Big Data Application in Biomedical Research and Health Care: A Literature Review,” Biomed Inform Insights. 2016; 8: 1–10.

# Supervisor: Vasily V. Kornilov

**25. Большие данные в персонализированной медицине / Big Data in Personalized Medicine**

Abstract: Healthcare Informatics combines information and understanding from medical areas (pre-clinical, clinical and post-clinical), healthcare administration and management and information technology. Big data applications present new opportunities to discover new knowledge and create novel methods to improve the quality of health care. The application of big data in health care is a fast-growing field, with many new discoveries and methodologies to make people better. The large volumes of data from bioinformatics and healthcare informatics domains coupled with analytics can deliver preventive, predictive and personalized healthcare aids.

Term paper can be devoted to the following issues: The benefits of big data analytics in the healthcare sector; Methodologies and tools to manage massive and complex information of electronic medical records; Personalized disease phenotypes from massive OMICs data

References

1. C.E. Morr “Research Perspectives on the Role of Informatics in Health Policy and Management,” IGI Global (USA), 2013, 323 p.
2. R. Wachter “The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine’s Computer Age,” McGraw-Hill Education, 2017, 352 p.
3. L.L. Liang “Connected for Health: Using Electronic Health Records to Transform Care Delivery,” Pfeiffer, 2017, 272 p.
4. A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D.A. Beard, K. Najarian “Big Data Analytics in Healthcare,” BioMed Research International, Volume 2015, Article ID 370194, 16 pages
5. S. S.-L. Tan, G. Gao, S. Koch “Big Data and Analytics in Healthcare,” Methods Inf Med 2015; 54: 546–547

# Supervisor: Vasily V. Kornilov

**26. Цифровое преобразование бизнеса / Digital Business Transformation**

Abstract: “Digital business is the creation of new business designs that not only connect people and business, but also connect people and business with things to drive revenue and efficiency. Digital business helps to eliminate barriers that now exist among industry segments while creating new value chains and opportunities that traditional businesses cannot offer.” (Gartner, 2015).

Digital Transformation radically alters markets through the application of new digital technologies, and it challenges established business models. Never before has IT played a more pivotal role. In the age of IoE (the Internet of Everything) and Big Data, IT has become the business process. IT is helping companies develop new business models to adapt to market and competitive forces, fundamentally changing companies’ relationships with their customers. Digital business is built on new computing infrastructure – the pillars of mobile, cloud, Big Data, and analytics – accelerated by the Internet of Things (IoT), advances in machine learning, and innovations like blockchain. These disruptive technologies are giving companies the ability to radically change business models, and create new products and services.

The future of every industry is digital. Fully digital business processes will be intelligent systems that respond to any situation, connecting with both machines and people to enable critical, real-time responses and insight-driven decision-making.

What is driving digital transformation? What a digital strategy is? How do organizations embark upon digital business transformation? How the digital strategy can serve business objectives?

References

1. Martin Gill, Ted Schadler, “The Digital Business Transformation Playbook For 2017,” Forrester Research, 2017 // https://www.forrester.com/report/Digitize+Your+Business+Strategy/-/E-RES115755
2. Bernard Marr “IT Playing Strategic Role In Digital Business Transformation”, PARS International, 2017 // https://www.forbes.com/sites/bernardmarr/2017/08/21/it-playing-strategic-role-in-digital-business-transformation
3. David Rogers “The Digital Transformation Playbook: Rethink Your Business for the Digital Age”, Columbia University Press, 2016, 296 p.
4. Venkat Venkatraman “The Digital Matrix: New Rules for Business Transformation Through Technology”, LifeTree Media, 2017, 224 p.
5. Isaac Sacolick “Driving Digital: The Leader's Guide to Business Transformation Through Technology”, AMACOM, 2017, 224 p.

# Supervisor: Vasily V. Kornilov

**7. Domain adaptation using neural networks /Использование нейронных сетей для адаптации в обучении**

Abstract: Practical machine learning often involves handling small datasets. This restricts usage of powerful models such as neural networks. Although large datasets (that are somehow similar) are often available and can help solving the original problem. For example, one can use classification model trained on Wikipedia pages to classify small enterprise documents. It is known that neural networks are well fitted for such problems. This work includes learning essentials of neural networks; research on domain adaptation with neural networks; empirical comparison of these approaches on available datasets.

*Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. “How transferable are features in deep neural networks?.”* [*http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks*](http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks)

# Supervisor: Sergey Lisitsyn

**28. Multi-task learning using neural networks /Использование нейронных сетей для многозадачного обучения**

Abstract: One of promising subfields of machine learning is multitask learning. This way of learning is based on solving multiple similar problems at the same time. The approach is often helpful as different models may ‘help’ each other. For example, it is known that a person that speaks a few languages learns new languages faster. Architectural flexibility of neural networks fits multitask learning quite well. This work includes learning essentials of neural networks and modern approaches for multitask learning. Practical part of the work includes empirical research of multitask learning techniques using open datasets.

*Sebastian Ruder. “An Overview of Multi-Task Learning in Deep Neural Networks”,* [*https://arxiv.org/pdf/1706.05098.pdf*](https://arxiv.org/pdf/1706.05098.pdf)

# Supervisor: Sergey Lisitsyn

**29. Comparison of fastText and word2vec for text classification / Сравнение fastText и word2vec для задач классификации текстов**

Abstract: Lately, machine learning community has shifted to vector models for textual data. This shift started when word2vec was first introduced. The word2vec algorithm enabled us to build effective semantic vector models of natural languages. The fastText algorithm is a further improvement which also considers syntax similarities of words. This work includes learning basics of vector models for natural languages and empirical comparison of the methods using available datasets.

*FastText and Gensim word embeddings* [*https://rare-technologies.com/fasttext-and-gensim-word-embeddings/*](https://rare-technologies.com/fasttext-and-gensim-word-embeddings/)

# Supervisor: Sergey Lisitsyn

**30. Real-time cryptocurrency rate analyzing service development with Big Data methods, techniques and tools / Разработка сервиса анализа и предсказания курса крипто валюты в реальном времени с помощью методов и технологий Больших данных.**

Abstract: Cryptocurrency is an alternative medium of exchange consisting of numerous decentralized crypto coin types. The essence of each crypto coin is in its cryptographic foundation. Secure peer-to-peer transactions are enabled through cryptography in this secure and decentralized exchange network. Since its inception in 2009, the Bitcoin has become a digital commodity of interest, as some believe the crypto coins’ worth is comparable to that of traditional fiat currency. Past research has shown that real-time social media data can be used to predict market movement of securities and other financial instruments. In addition, cryptocurrency rate is very unstable and can change rapidly that is why decisions on buy or sell should be taken very fast.

References

1. Solution patterns for realtime streaming analytics - Srinath Perera, Sriskandarajah Suhothayan, 2015
2. DiAl: distributed streaming analytics anywhere, anytime – Ivo Santos, Marcel Tilly, Badrish Chandramouli, Jonathan Goldstein
3. Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies - Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, Chang Hun Kim, 17.08.2016, PLoS NE11(8): e0161197. <https://doi.org/10.1371/journal.pone.0161197>
4. Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules – Ramazan Gençay, 01.02.1999
5. Can We Predict the Winner in a Market with Network Effects? Competition in Cryptocurrency Market - Neil Gandal, Hanna Halaburda, 18.03.2016
6. Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis - Stuart Colianni, Stephanie Rosales, and Michael Signorotti(<http://cs229.stanford.edu/proj2015/029_report.pdf>)

# Supervisor: Ekaterina Filimonova

**31. Unified recommendation engine development using data science methods and Big Data technologies and tools/Разработка единого рекомендательного движка с применением методов data science и технологий Больших данных.**

Abstract: Recommender system is a specific type of intelligent systems, which users’ behavior in the Internet to make recommendations on items or services to them. It plays a critical role in a wide range of online shopping, e-commercial services and social networking applications. Data science provide methods and tools for building recommendation systems like neural networks, deep learning, logistic regression, decision trees etc. The problem is that many companies have different types of services and items they want to recommend, for which they need to track and catch different user behavior patterns. In this case, to avoid building recommendation system for each specific product is necessary to develop unified platform that will allow to set the parameters for targeting.

References

1. Rethinking Reddit: A Novel Hybrid Recommendation System - Spencer Kee, 2017,
2. Supporting Change Impact Analysis Using a Recommendation System: An Industrial Case Study in a Safety-Critical Context - Markus Borg, Krzysztof Wnuk, Björn Regnell, Per Runeson, 2017
3. An Insurance Recommendation System Using Bayesian Networks - Maleeha Qazi, Glenn M. Fung, Katie J. Meissner, Eduardo R. Fontes
4. Deep Learning based Large Scale Visual Recommendation and Search for E-Commerce - Devashish Shankar, Sujay Narumanchi, H A Ananya, Pramod Kompalli, Krishnendu Chaudhury, 07.03.2017
5. Ontology based web usage mining model - C. Ramesh, K.V. Chalapati Rao, A. Govardhan, 11.06.2017
6. Methods and systems for predictive engine evaluation and replay of engine performance - Ka Hou Chan, Simon Chan, Kit Pang Szeto, Yue Kwen Justin Yip, 04.05.2017

# Supervisor: Ekaterina Filimonova

**32. Real time signal (time series) processing/Обработка сигналов (временных рядов) в реальном времени.**

Abstract: The goal of this project is to implement and analyze optimal time series processing with a sliding window.

Main phases:

1. Generation of synthetic data: random function with specified stochastic properties and its measurement.
2. Computation of the optimal influence function of the processing sliding window with the given width and delay.
3. Analysis of the relation between processing quality, delay and the width of the sliding window. Such analysis should allow choosing a good compromise between the reconstruction quality, delay, and computational complexity.
4. Illustrations of the processing performance under different settings.

References:

1. Golubtsov P.V., Starikova O.V. Reduction of parametrically determined invariant measuring systems // Moscow University Physics Bulletin, V 56, No 6, с. 1-4, 2001
2. Filatova S. A., Golubtsov P. V. Invariant Measurement Computer Systems // Pattern Recognition and Image Analysis, 1, N 2, P. 224–235, 1991.
3. Filatova S. A., Golubtsov P. V. Invariance Considerations in Design of Image Formation Measurement Computer Systems // Proceedings of SPIE, V. 1960, Automatic Object Recognition III, P. 483–494, 1993.
4. Голубцов П. В., Филатова С. А. Оптимальная локальная редукция для пространственно-инвариантных измерительных систем // Математическое моделирование, 2, N 11, С. 61–66, 1990.

Supervisor: Peter V. Golubtsov

**33. Real time image processing for potentially infinite field of view /Обработка изображения с потенциально бесконечным полем рения.**

Abstract: The goal of this project is to implement and analyze optimal image processing with a sliding window.

Main phases:

1. Generation of synthetic data: measurement of an image, which includes smoothing-like distortion and random noise.
2. Computation of the optimal point spread function of the processing sliding window with the given size.
3. Analysis of the relation between processing quality and the size of the sliding window. Such analysis should allow choosing a good compromise between the reconstruction quality and computational complexity.
4. Illustrations of the processing performance under different settings.

References:

1. Filatova S. A., Golubtsov P. V. Invariant Measurement Computer Systems // Pattern Recognition and Image Analysis, 1, N 2, P. 224–235, 1991.
2. Filatova S. A., Golubtsov P. V. Invariance and Synthesis of Optimum Image Formation Measurement Computer Systems // Artificial Intelligence, Expert Systems and Symbolic Computing, Elsevier Science Publishers B. V. (North-Holland), P. 243–252, 1992.
3. Filatova S. A., Golubtsov P. V. Invariance Considerations in Design of Image Formation Measurement Computer Systems // Proceedings of SPIE, V. 1960, Automatic Object Recognition III, P. 483–494, 1993.
4. Golubtsov P.V., Sizarev D.V., Starikova О.V. Synthesis of optimal invariant image generation systems on a plane // Moscow University Physics Bulletin, V 58, No 2, pp 1-5, 2003

**Supervisor: Peter V. Golubtsov**

**34. Calibration in time series processing problems/Калибровка в задачах обработки временных рядов.**

Abstract: The goal of this project is to implement and analyze optimal time series processing for the system with the unknown measurement system (which represents distortion of the time series). It is assumed that originally the transformation, distorting the signal (time series) is initially unknown. As a result the reconstruction of the signal is not possible. However, we are able to obtain the true input signal with a certain delay. Such data can be used for the calibration of the measurement system.

Main phases:

1. Generation of synthetic data: random function with specified stochastic properties and its measurement.
2. Extraction of calibration data and accumulating it in an appropriate canonical form.
3. Gradual recalibration of the measurement system, based on the accumulated calibration canonical information.
4. Computation of the optimal influence function of the processing sliding window for the calibrated measurement system for various stages of calibration.
5. Illustrations of the processing performance under different settings.

References:

1. Filatova S.A., Golubtsov P.V. Invariance and Calibration in Image Processing Problems // The Third International Conference on Expert Systems for Numerical Computing. Computer Science Technical Reports Purdue University, pp 97-100 1993
2. Golubtsov P.V., Starikova O.V. Reduction of parametrically determined invariant measuring systems // Moscow University Physics Bulletin, V 56, No 6, с. 1-4, 2001
3. Голубцов П.В., Старикова О.В. Калибровка инвариантных преобразователей информации // Information Processes (Информационные процессы), том 1, № 1, с. 78-88 2001 [www.jip.ru/2001/golub-1.pdf](http://www.jip.ru/2001/golub-1.pdf)
4. Голубцов П.В., Старикова О.В. Проблема калибровки для инвариантных измерительно-вычислительных систем // Труды II международной конференции «Проблемы управления и моделирования в сложных системах», Издательство РАН Самара, с. 143-148. 2000
5. Голубцов П.В., Старикова О.В. Инвариантность и калибровка измерительно-вычислительных систем в непрерывно-дискретном случае // Труды III международной конференции «Проблемы управления и моделирования в сложных системах», Издательство РАН Самара, с. 635-641. 2001

**Supervisor: Peter V. Golubtsov**

**35. Crowding behavior in a supermarket and sales/Поведение толпы в супермаркете и продажи.**

Abstract: The goal of this project is to develop tools and visualization methods for analysis of customers’ flows in supermarkets and its relations to sales. The study will be based on real data, which includes RFID tracking of shoppers’ trajectories.

Main phases:

1. Visualization and comparison of customers’ flows during peak hours and off-peak hours.
2. Revealing relations between crowding behavior and sales.
3. Analysis of main shopping patterns.
4. Crowd dynamics analysis. Detection of zones with chaotic and regular flow.
5. Experimentation with various visualization techniques, such as: heat maps, vector fields, etc.
6. Possibility of trajectory prediction at different spots for different shopping patterns.

References:

1. Yada, K.: String Analysis Technique for Shopping Path in a Supermarket. Journal of Intelligent Information Systems 36, 385–402 (2011)
2. Kholod M., Nakahara T., Azuma H., and Yada K. The Influence of Shopping Path Length on Purchase Behavior in Grocery Store. Proceedings of KES 2010, Lecture Notes in Computer Science, 6278, pp.273-280, 2010.
3. Larson, J.S., Bradlow, E.T., Fader, P.S.: An Exploratory Look at Supermarket Shopping Paths. International Journal of Research in Marketing 22(4), 395–414 (2005)
4. Hui, S.K., Fader, P.S., Bradlow, E.T.: Path Data in Marketing: An Integrative Framework and Prospectus for Model Building. Marketing Science 28(2), 320–335 (2009)
5. Hui, S.K., Fader, P.S., Bradlow, E.T.: The Travelling Salesman Goes Shopping: The Systematic Deviations of Grocery Paths from TSP Optimality. Marketing Science 28(3), 566–572 (2009)
6. Hui, S.K., Bradlow, E.T., Fader, P.S.: Testing Behavioral Hypotheses using An Integrated Model of Grocery Store Shopping Path and Purchase Behavior. Journal of Consumer Research 36(3), 478–493 (2009)

**Supervisor: Peter V. Golubtsov**

**36. Fuzzy information and fuzzy experiments in big data context/Нечеткая информация и нечеткий эксперимент в контексте больших данных**

Abstract: Sometimes it is more convenient to represent uncertainty in terms of fuzzy sets as opposed to probability distributions. Fuzzy sets representations may better reflect such information as expert opinions.

The goals of this project are to

1. Design formal representation of an “expert assessment” process as a fuzzy measurement.
2. Develop methods for:
   1. updating fuzzy information, based on such measurement;
   2. extracting information from many independent fuzzy measurements and presenting it in a convenient intermediate form;
   3. combining pieces of such information;
   4. and deriving a decision from the accumulated information.
3. Visualize all the transformations of fuzzy information.

References:

1. Zadeh L. A. Fuzzy Sets // Inform. and Control, 8, P. 338–353, 1965.
2. Golubtsov P.V. Theory of Fuzzy Sets as a Theory of Uncertainty and Decision-Making Problems in Fuzzy Experiments // Problems of Information Transmission, V 30, No 3, pp 232-250, 1994
3. Dubois D., Prade H. Fuzzy Sets and Systems: Theory and Applications. New York: Academic Press, 1979.
4. Golubtsov P. V. Fuzzy logical semantics of Bayesian decision making // SPIE Proceedings Vol. 2493. Applications of Fuzzy Logic Technology II, pp 228–239, 1995.

**Supervisor: Peter V. Golubtsov**