

Russian-American

# Research Symposium

*Sponsored by MIT in collaboration with Skoltech*

## Machine Learning for Big Data *Texts, Signals, Images and Video*

Professor Konstantin Vorontsov

Moscow Institute of Physics and Technology

December 15, 2014

# SDP: Machine Learning for Big Data (2014)

## Structure of research group:

- MIPT: K.Vorontsov, V.Strijov, ...
- MSU: D.Vetrov, A.Konushin, ...

## Activities

- Research
- Development
- Education
- Innovation

## Projects of MIPT group:

### → Texts:

- 1. BigARTM for Topic Modeling
- 2. Conference Hierarchy Tool

### → Signals:

- 3. ECG Multi-Disease Diagnostics
- 4. Human Behavior Recognition

# 1. Topic Modeling for big text collections

## Text documents

International Journal of Networks and Distributed Engineering  
Vol. 6, No. 1, January, 2011

### Scalable Intrusion Detection with Recurrent Neural Networks

Long O. Aiyem, M.S.; Jasef George Ph.D.; Gladys Aroca, Ph.D.;  
E.I.D.; Dept of Engineering Learning; MPh;  
Dept of Math and Computer Sci.; University of North Dakota; College of Educ., L. Inship & Tech.  
Fort Hays State University; North Dakota, USA; Valdosta State University  
Kansas, USA; Email: jasef.george@und.edu; Georgia, USA  
Email: loayem@fhsu.edu; jasef.george@und.edu; Email: garoca@valdosta.edu

**Abstract**  
The ever growing use of the Internet comes with a surge in escalation of communication and data access. Most existing intrusion detection systems have assumed the on-premise physical solution model. Such IDS is not as economically sustainable for all organizations. Furthermore, studies have found that recurrent neural networks outperform feedforward Neural Network, and Elman Network. This paper, therefore, proposes a scalable application based model for detecting attacks in a communication network using recurrent neural network architecture. Its suitability for online real-time applications and its ability to self-adjust to changes in network environment is emphasized.

**Keywords:** Communication, Security, Scalable, Neural, Network Intrusion, Detector, System

**1. Introduction**  
The ever growing use of the Internet comes with a surge in escalation of communication and data access. Coupled with the communication escalation, is the rapid proliferation of networks and their compounding management complexity. The ubiquity of the Internet undoubtedly poses serious security, network, and traffic and the integrity of sensitive data. Consequently, Network security and effective firewalls have emerged to be a hot area of increasing attention in the computing industry. A variety of studies have been carried out in communication and network security, and malware attack detection and resolution. [1][2][5]

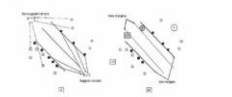
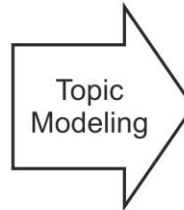


Fig. 2 Separation of the support vectors and non-support vectors (adapted from [12])

21



## Topics of documents

D  
o  
c  
u  
m  
e  
n  
t  
s

doc1:	<span style="width: 20px; height: 20px; background-color: white; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: red; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: green; border: 1px solid black;"></span>			
doc2:	<span style="width: 20px; height: 20px; background-color: red; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: green; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: purple; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: yellow; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: brown; border: 1px solid black;"></span>	
doc3:	<span style="width: 20px; height: 20px; background-color: purple; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: yellow; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: brown; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: blue; border: 1px solid black;"></span>		
doc4:	<span style="width: 20px; height: 20px; background-color: gray; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: white; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: red; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: green; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: yellow; border: 1px solid black;"></span>	<span style="width: 20px; height: 20px; background-color: blue; border: 1px solid black;"></span>
...						

## Words and keyphrases of topics

T  
o  
p  
i  
c  
s

<span style="width: 20px; height: 20px; background-color: white; border: 1px solid black;"></span>	машинное обучение, нейронные сети, сеть Девидов, градиентный спуск, обучение с подкреплением, Векторы Паркса в классе и области нейронные сети были созданы: В. Мейнхофф (ИТ МСБайт) и В. Петров
<span style="width: 20px; height: 20px; background-color: red; border: 1px solid black;"></span>	Большая Данные, Инструменты, Большая Данные, Tools Of Big Data ... Прямые наиболее наиболее известные задан, при решении которых применение машинного обучения сыграло ключевую роль
<span style="width: 20px; height: 20px; background-color: green; border: 1px solid black;"></span>	Смешанные и такие случаи обучения, соответственно: запонки вранья, данные в газодорожках, графовые и переходы чис, ключевые слова, графы и машинное обучение
<span style="width: 20px; height: 20px; background-color: purple; border: 1px solid black;"></span>	Закономерности, данные в газодорожках, сложность и трудности как ключевые слова, обучение по градиенту, обучение искусственных нейронных сетей, обучение нейронных сетей, искусственные нейронные сети, частные - упорядоченные векторы
<span style="width: 20px; height: 20px; background-color: yellow; border: 1px solid black;"></span>	Топы, папки, ресурсы, машинное обучение, градиентный спуск, муля, графовые, мексик, как, вектор, графовый обмен, мексик, мексик, паспорты, графы
<span style="width: 20px; height: 20px; background-color: brown; border: 1px solid black;"></span>	Интерес, градиентный спуск, частые градиентные, переменные, аргумент, функции градиента, графы, особенно дифференцируемые

# 1. Topic Modeling for big text collections

Metadata:

Authors  
Data Time  
Conference  
Organization  
URL  
etc.

Text documents

International Journal of Networks and Ubiquitous Engineering  
Vol. 6, No. 1, January, 2011

Scalable Intrusion Detection with Recurrent Neural Networks

Long O. Argyrakis, M.S.; Jose R. Gonzalez, Ph.D.; Gladys A. Alvarez, Ph.D.;  
E.I.D.; Dept. of Engineering Learning; MPh; College of Edu. L. L. Shapiro, &  
Dept. of Math and Computer Sci.; North Dakota State University; Valdosta State University;  
Fort Hays State University; Kansas, USA; Email: jrgonzalez@ndst.edu; george@valdosta.edu  
Email: loargy@fhsu.edu; jrgonzalez@ndst.edu; galarvez@valdosta.edu

**Abstract**  
The ever growing use of the Internet comes with a surge in escalation of communication and data access. Most existing intrusion detection systems have assumed the one-to-one (1:1) solution model. Such IDS is not as economically sustainable for all organizations. Furthermore, studies have been conducted on recurrent neural networks (RNN) for detecting network anomalies. This paper, therefore, proposes a scalable application based model for detecting attacks in communication network using recurrent neural network architecture. Its suitability for online real-time applications and its ability to self-adjust to changes in network environment is also emphasized.

**Keywords:** Communication, Security, Scalable, Neural, Network Intrusion, Detector, System

**1. Introduction**  
The ever growing use of the Internet comes with a surge in escalation of communication and data access. Coupled with the communication escalation, is the rapid proliferation of networks and their compounding management complexity. The ubiquity of the Internet undoubtedly poses serious security, reliability, and integrity of sensitive data. Consequently, network security and effective firewalls have emerged to be a hot area of research attention in the computing industry. A variety of studies have been carried out in communication and network security, and network attack detection and resolution. [1][2][3]

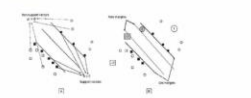
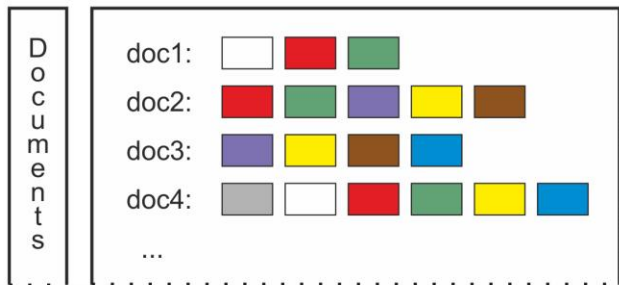


Fig. 2 Separation of the Support Vector Support and Non-Support Vector Support (adapted from [2])

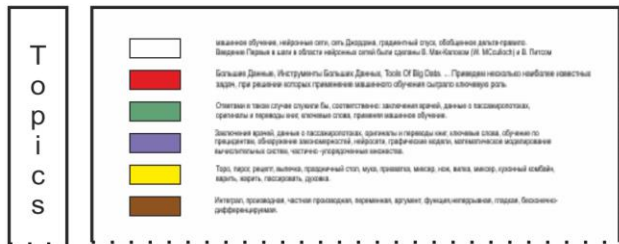
21

Topic Modeling

Topics of documents



Words and keyphrases of topics



# 1. Topic Modeling for big text collections

Metadata:

Authors  
Data Time  
Conference  
Organization  
URL  
etc.

Text documents

International Journal of Networks and Ubiquitous Engineering  
Vol. 6, No. 1, January, 2011

Scalable Intrusion Detection with Recurrent Neural Networks

Long O. Argyrakis, M.S.; Jeehoon George, Ph.D.; Gladys A. Alvarez, Ph.D.;  
E.D.; Dept. of Engineering Learning; MPh; College of Education, Linn and  
Dept. of Math and Computer Science; University of North Dakota; Tech; Valdosta State University  
Fort Hays State University; Kansas, USA; Email: jeehoon.george@und.edu; gladys.alvarez@valdosta.edu  
Email: loargrakis@fhsu.edu; jeehoon.george@und.edu

**Abstract**  
The ever growing use of the Internet comes with a surge in malicious communication and data access. Most existing intrusion detection systems have assumed the traditional signature solution model. Such IDS is not as economically sustainable for all organizations. Furthermore, studies have shown that recurrent neural networks outperform feedforward Neural Network, and Elman Network. This paper, therefore, proposes a scalable application based model for detecting attacks in communication network using recurrent neural network architecture. Its suitability for online real-time applications and its ability to self-adjust to changing network patterns is demonstrated over an empirical dataset.

**Keywords:** Communication, Security, Scalable, Neural, Network, Intrusion, Detector, System.

**1. Introduction**  
The ever growing use of the Internet comes with a surge in malicious communication and data access. Coupled with the communication explosion, is the rapid proliferation of networks and their compounding management complexity. The ubiquity of the Internet undoubtedly poses serious security concerns. Network traffic and the integrity of sensitive data. Consequently, Network security and effective firewalls have emerged to be a hot area of increasing attention in the computing industry. A variety of studies have been carried out in communication and network security, and network attack detection and resolution. [1][2][3]

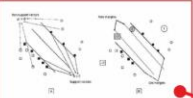


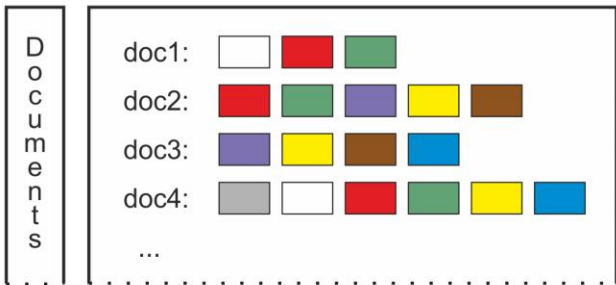
Fig.2 Separation of the support vectors and non-support vectors in a 2D space [12]

21

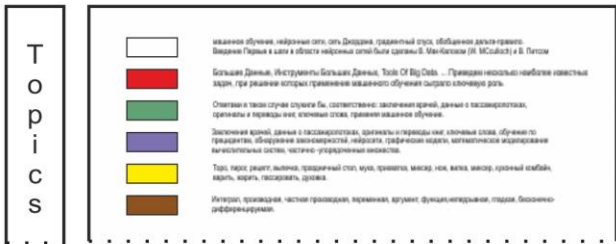
Images

Topic Modeling

Topics of documents



Words and keyphrases of topics

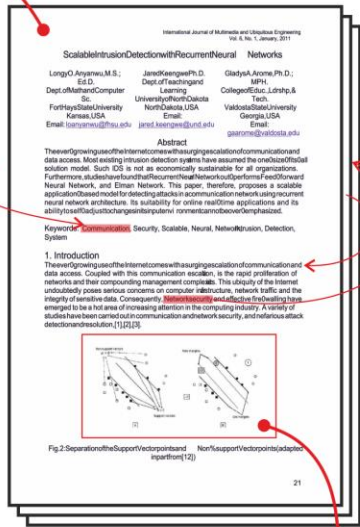


# 1. Topic Modeling for big text collections

**Metadata:**

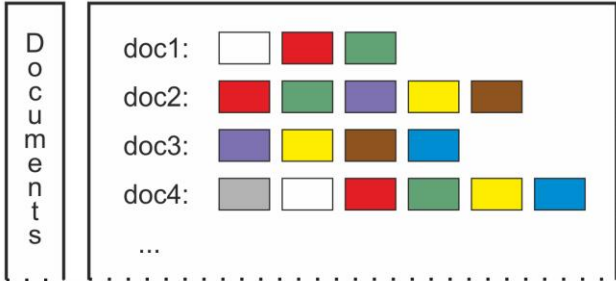
Authors  
Data  
Time  
Conference  
Organization  
URL  
etc.

## Text documents

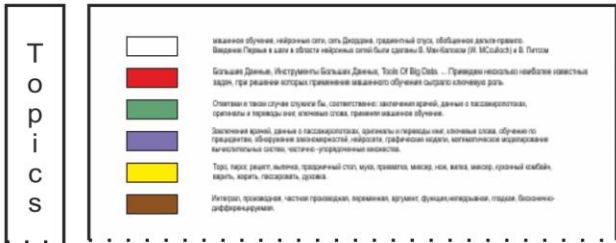


Images Links

## Topics of documents



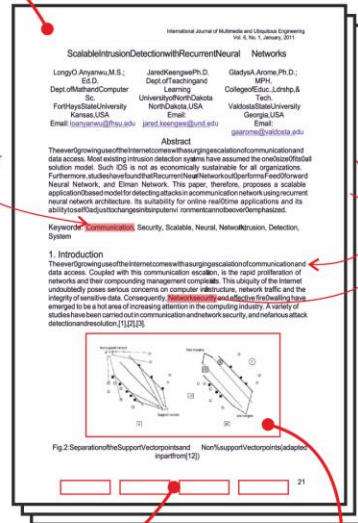
## Words and keyphrases of topics



# 1. Topic Modeling for big text collections

**Metadata:**  
Authors  
Data Time  
Conference  
Organization  
URL  
etc.

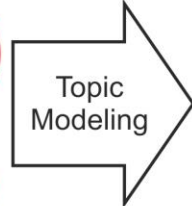
## Text documents



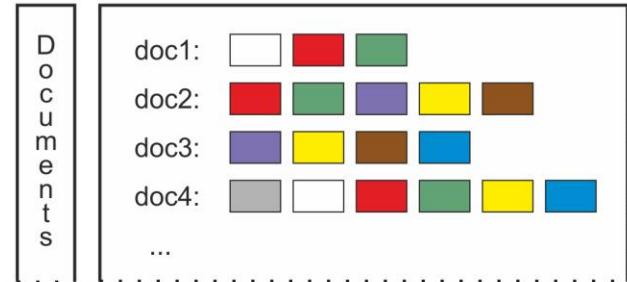
Ads

Images

Links



## Topics of documents



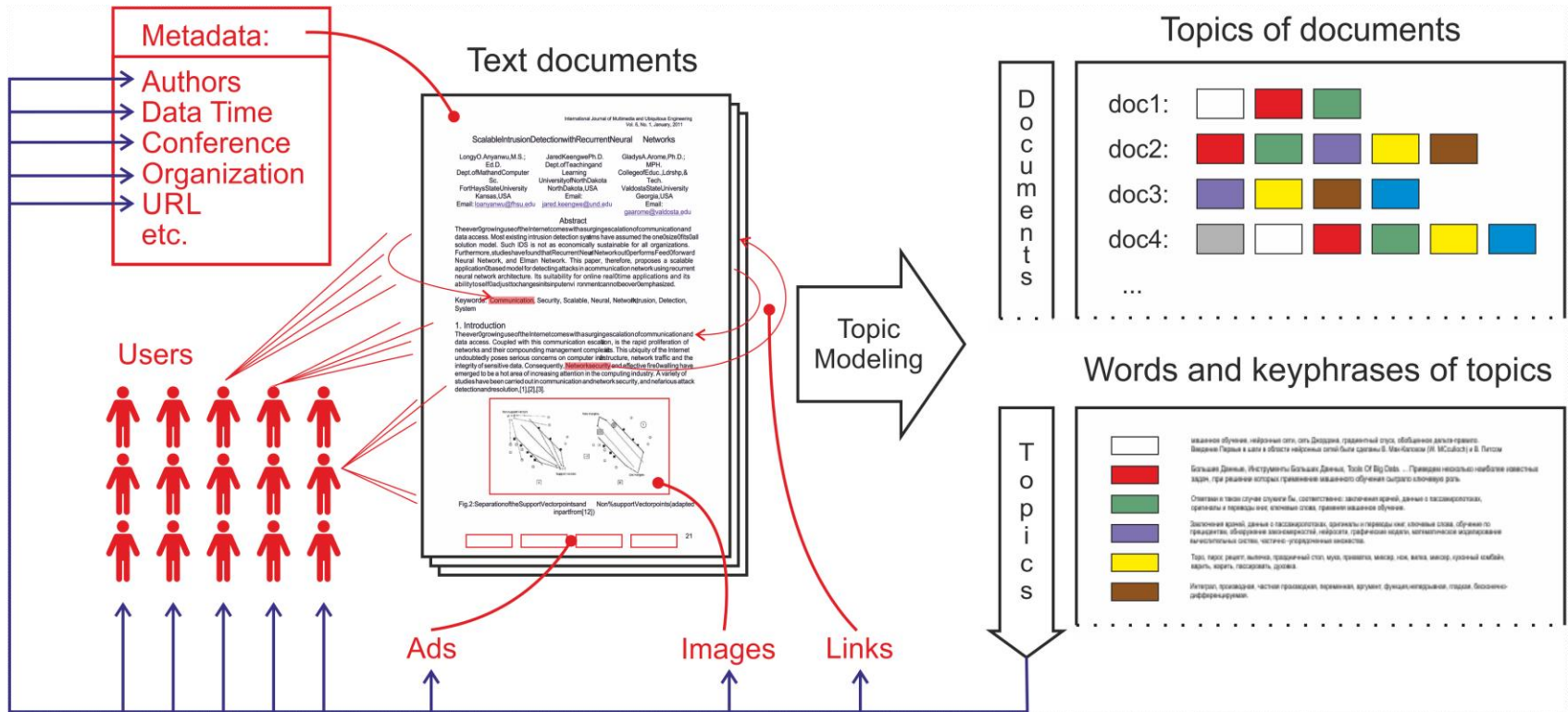
## Words and keyphrases of topics







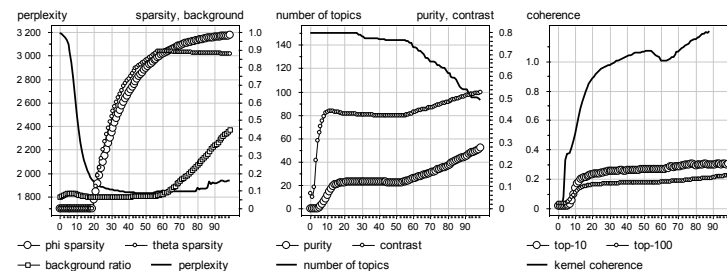
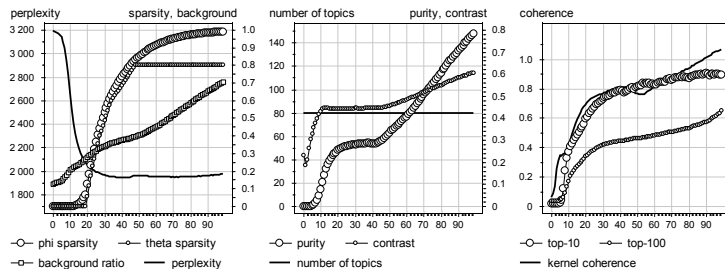
# 1. Topic Modeling for big text collections



# 1. Topic Modeling: BigARTM project



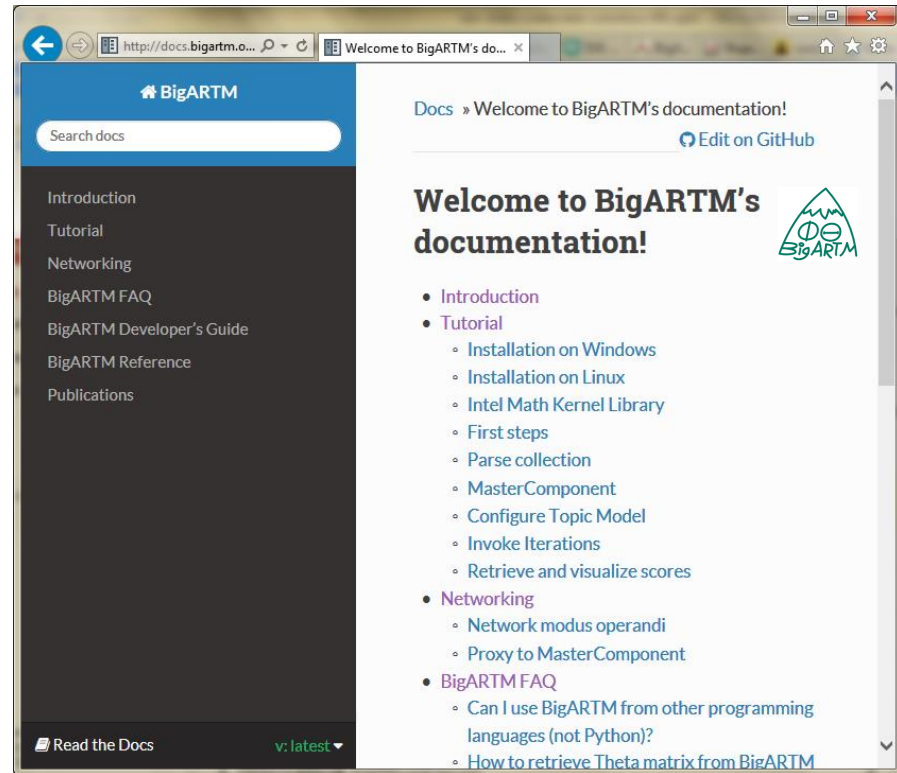
- **Challenge:** how to combine many functionalities in a single Topic Model
- **Theory:** ARTM – Additive Regularization for Topic Modeling  
(easy to understand – easy to design – easy to infer – easy to combine)
- **Implementation:** BigARTM – open-source with permissive license
- **Experiments:** multi-criteria optimization of Topic Models



# 1. Topic Modeling: BigARTM project



- Open-source (<http://bigartm.org>)
- Parallel, Distributed,
- Online, Fast, Sparse, Robust,
- Multi-modal,
- Multi-criteria,
- Multi-language,
- Semi-supervised,
- Temporal,
- Hierarchical, etc...



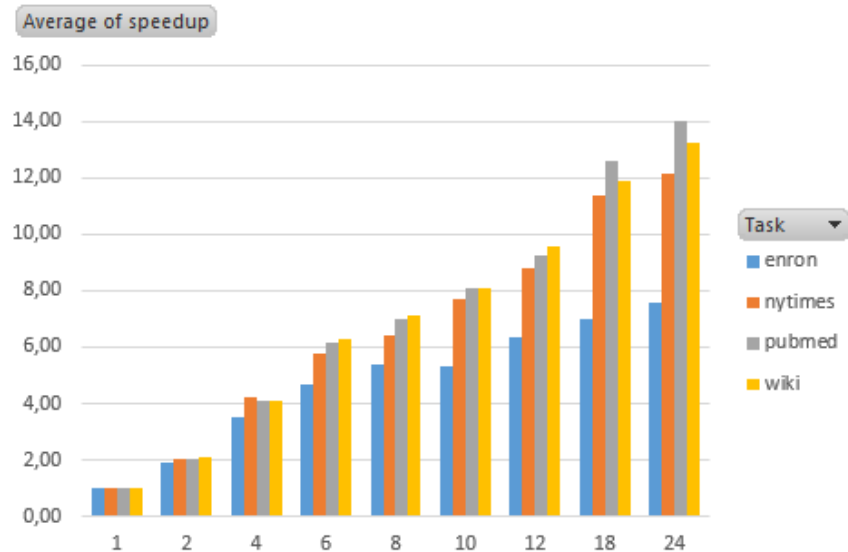
# 1. Topic Modeling: BigARTM project



Performance testing on datasets (W – size of vocabulary, D – number of documents):

- Enron (W=28 102, D=39 861)
- Nytimes (W=102 660, D = 300 000)
- Pubmed (W=141 043, D = 8 200 000)
- Wiki (W=100 000, D = 3 665 223)

#Proc	enron(sec)	nytimes(min)	pubmed(min)	wiki (min)
1	28,42	15,58	113,98	130,74
2	14,83	7,76	55,78	62,03
4	8,10	3,71	27,93	31,88
6	6,12	2,70	18,50	20,87
8	5,31	2,44	16,32	18,37
10	5,33	2,02	14,05	16,12
12	4,49	1,78	12,30	13,63
18	4,07	1,37	9,04	10,97
24	3,74	1,28	8,15	9,87



Intel® Xeon® CPU E5-2630 v2 @ 2.60 GHz  
(12 cores + hyper threading)

# 1. Topic Modeling: publications

---

1. Vorontsov K. V. Additive Regularization for Topic Models of Text Collections // ***Doklady Mathematics***. 2014, Pleiades Publishing, Ltd. — Vol. 89, No. 3, pp. 301–304.
2. Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'2014, Analysis of Images, Social networks and Texts. ***Springer, Communications in Computer and Information Science***, 2014. Vol. 436. pp. 29–46.
3. Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // ***Machine Learning***, Special Issue «Data Analysis and Intelligent Optimization», Springer, 2014. (to appear).

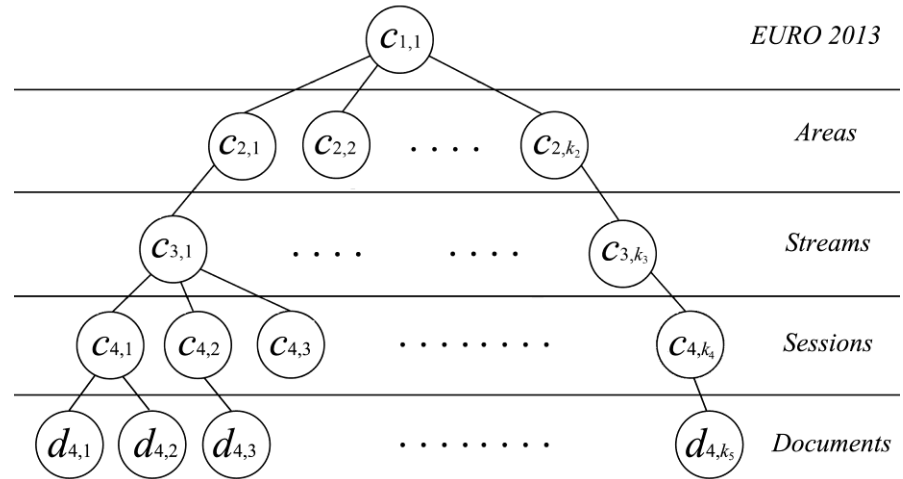
## 2. Hierarchical Conference Topic Model

→ **The goal:**

automatic construction of  
a scientific conference program

→ **EURO – European Conference  
on Operational Research:**

>3500 participants,  
>200 experts,  
24 areas, 137 streams



# 2. Hierarchical Conference Topic Model

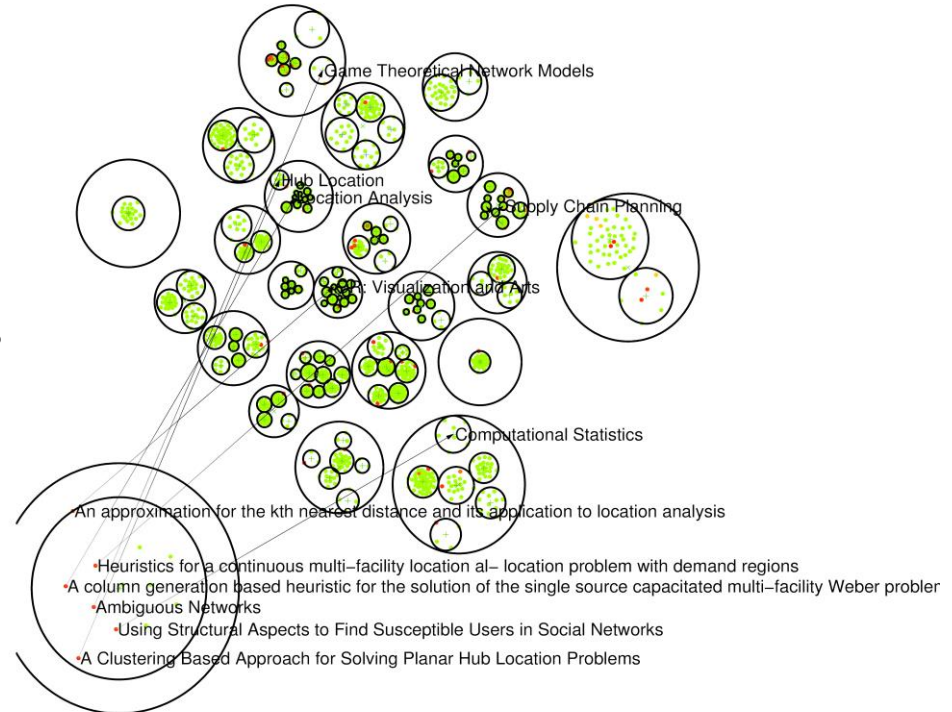
→ Visualization of inconsistencies:

→ Green points

– consistent submissions

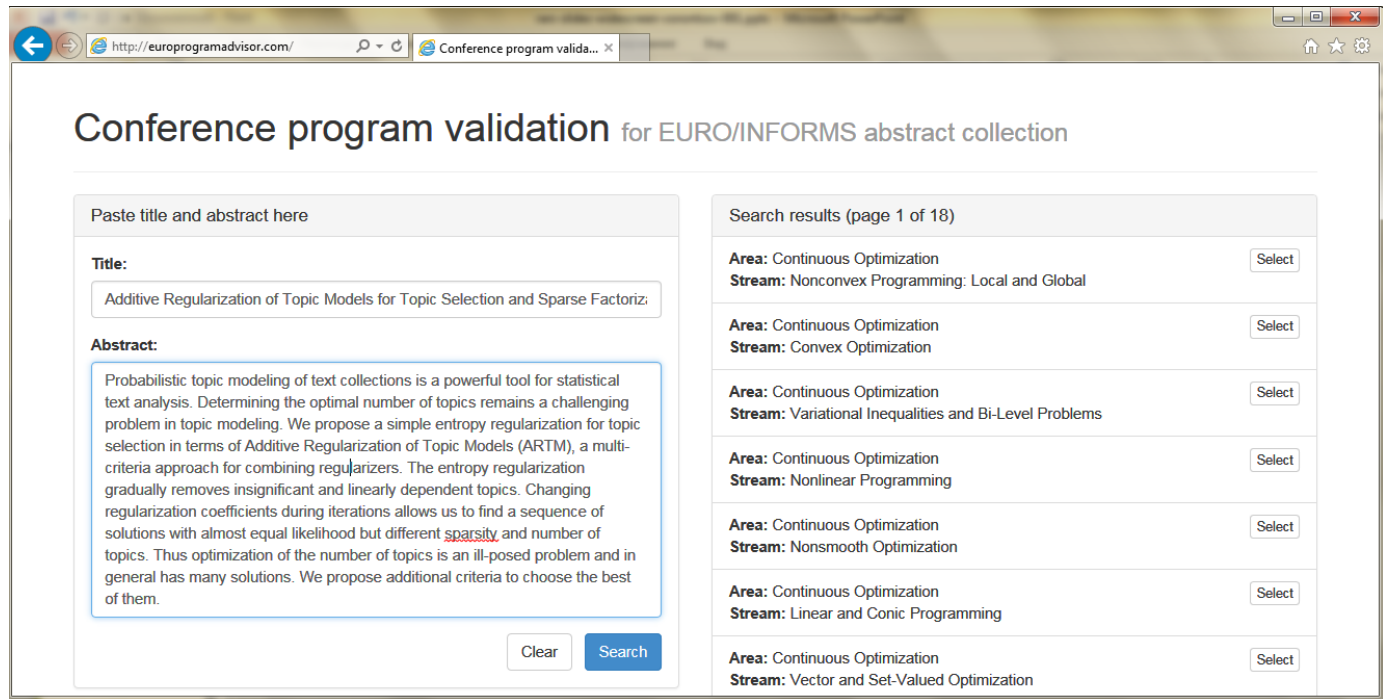
→ Red points

– inconsistent submissions



# 2. Hierarchical Conference Topic Model

→ <http://EUROprogramAdvisor.com>



The screenshot shows a web browser window with the URL <http://europrogramadvisor.com/>. The page title is "Conference program validation for EURO/INFORMS abstract collection". The main content area is divided into two columns. The left column contains a form for pasting a title and abstract. The right column displays search results for the entered text.

**Conference program validation for EURO/INFORMS abstract collection**

Paste title and abstract here

**Title:**  
Additive Regularization of Topic Models for Topic Selection and Sparse Factoriz.

**Abstract:**  
Probabilistic topic modeling of text collections is a powerful tool for statistical text analysis. Determining the optimal number of topics remains a challenging problem in topic modeling. We propose a simple entropy regularization for topic selection in terms of Additive Regularization of Topic Models (ARTM), a multi-criteria approach for combining regularizers. The entropy regularization gradually removes insignificant and linearly dependent topics. Changing regularization coefficients during iterations allows us to find a sequence of solutions with almost equal likelihood but different sparsity and number of topics. Thus optimization of the number of topics is an ill-posed problem and in general has many solutions. We propose additional criteria to choose the best of them.

Clear Search

Search results (page 1 of 18)

<b>Area:</b> Continuous Optimization <b>Stream:</b> Nonconvex Programming: Local and Global	Select
<b>Area:</b> Continuous Optimization <b>Stream:</b> Convex Optimization	Select
<b>Area:</b> Continuous Optimization <b>Stream:</b> Variational Inequalities and Bi-Level Problems	Select
<b>Area:</b> Continuous Optimization <b>Stream:</b> Nonlinear Programming	Select
<b>Area:</b> Continuous Optimization <b>Stream:</b> Nonsmooth Optimization	Select
<b>Area:</b> Continuous Optimization <b>Stream:</b> Linear and Conic Programming	Select
<b>Area:</b> Continuous Optimization <b>Stream:</b> Vector and Set-Valued Optimization	Select



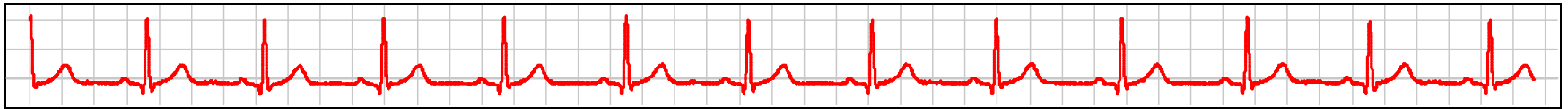
## 2. Hierarchical Conference Topic Model

---

1. Katrutsa A.M., Kuznetsov M.P., Strijov V.V., Rudakov K.V. Metric concentration search procedure using reduced matrix of pairwise distances // Intelligent Data Analysis, 2015, 19(5).
2. Kuzmin A.A., Aduenko A.A., Strijov V.V. Thematic classification using expert model for major conference abstracts // Information Technologies.
3. Kuzmin A.A., Aduenko A.A., Strijov V.V. Thematic Classification for EURO/IFORS Conference Using Expert Model // Conference of the International Federation of Operational Research Societies, 2014.

# 3. ECG processing for Multi-Disease Diagnostics

## The Technology of Informational Analysis of ECG-signal



1. Measuring RR-interval and amplitude of each R-peak

2. Discretization

DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAFAFFFAEAEFBAEFBAEFCAFFAAD  
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFAADFCAFBCADFFCEFFAFAFFFAEFFCAFCACFFCAD  
DAADBFAAFFAEFBFAACDFFAAFBAADFADDAFCECFCEDFCEEFCAEFBECBBBADBAACFFAFAFFA  
CFFCECFDABDAEFFFAFFCEDBFAFFAEFFAEFBACFBADFEAFAFFCAFFDAFFAEBDADBBADDAFF  
EABFCCAFDEEBDECFFACFFAABFAADFBAFFACFFFAEFFACFFACFFCECFBAFFFAFFFAFFAADF  
AABFACDFDAEFFAADBAEFFEAFBCECFDECCFBAFFAADFDACDFAFFAADFCAADFAEFBAFFCADFE  
AFFCECFCEFFAFAFFABCFDAAFFADBFCAEFFAABFACBFAEBFAEBFAFFBAFFFAFFDADFDAABFB  
CAFFAECCFFACFFACDFCADFDAABFAREDDABBFCAACDBAFAFFAFAFFCADFAADFACFFAEDEFCAFC  
CAEBCE

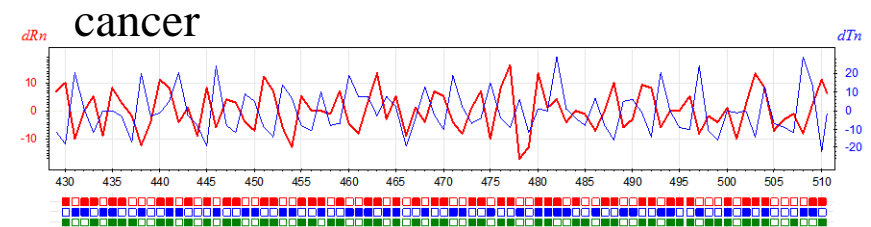
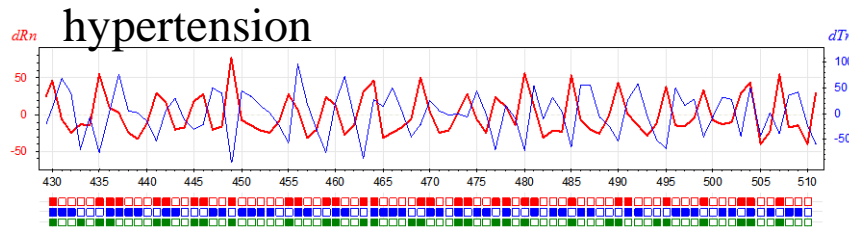
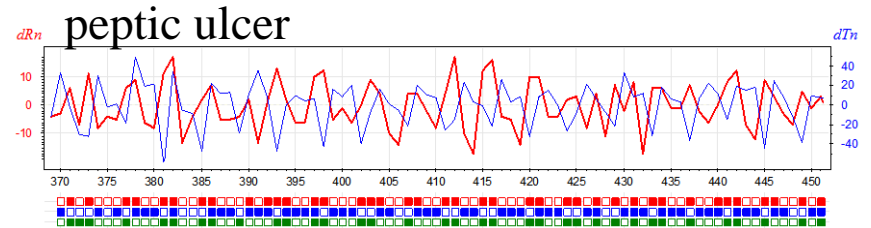
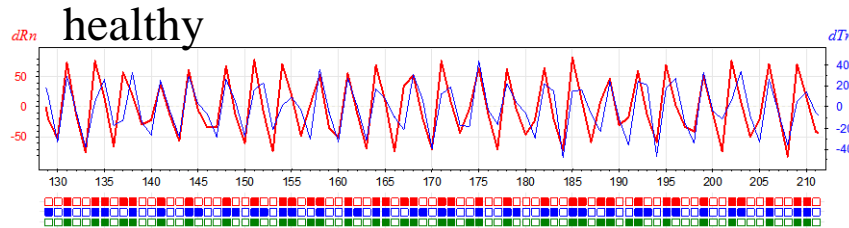
3. Vectorization

4. Machine Learning (Naïve Bayes, SVM, Topic Model, Deep Learning)

5. Estimation (Cross-Validation, Sensitivity-Specificity, AUC)

# 3. ECG processing for Multi-Disease Diagnostics

- Variations of RR-intervals and R-amplitudes carry information about the functioning of not only the heart, but all the systems of the body, and can be used for the diagnosis at any stage of the disease [V.M.Uspenskiy, 2008]



# 3. ECG processing for Multi-Disease Diagnostics

The results of our cross-validation experiments

Data set:

20 000 ECGs

5-7 minutes each

60 Gb total size

40 diseases

disease	cases	AUC, %	spec, % (sens=95%)
femoral head necrosis	327	99.19 ± 0.10	96.6 ± 1.76
cholelithiasis	277	98.98 ± 0.23	94.4 ± 1.54
coronary heart disease	1262	97.98 ± 0.14	91.1 ± 1.86
gastritis	321	97.76 ± 0.11	88.3 ± 2.64
hypertensive disease	1891	96.76 ± 0.09	84.7 ± 1.99
diabetes	868	96.75 ± 0.19	85.3 ± 2.18
benign prostatic hyperplasia	257	96.49 ± 0.13	80.1 ± 3.19
cancer	525	96.49 ± 0.28	82.2 ± 2.38
nodular goiter thyroid	750	95.57 ± 0.16	73.5 ± 3.41
chronic cholecystitis	336	95.35 ± 0.12	74.8 ± 2.46
biliary dyskinesia	714	94.99 ± 0.16	70.3 ± 4.67
urolithiasis	649	94.99 ± 0.11	69.3 ± 2.14
peptic ulcer	779	94.62 ± 0.10	63.6 ± 2.55

# 3. ECG processing for Multi-Disease Diagnostics

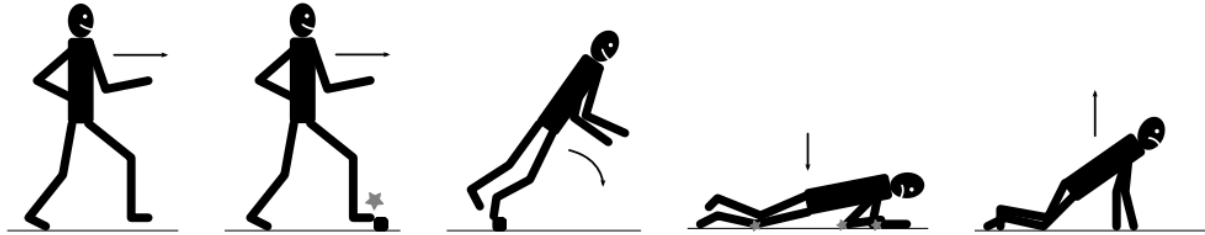
---

1. Uspenskiy V. M., Vorontsov K. V., Tselykh V. R., Bunakov V. A. Information Function of the Heart: Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic System // Advanced Mathematical and Computational Tools in Metrology — AMCTM 2014.
2. Uspenskiy V. M. Information Function of the Heart // Clinical Medicine, vol. 86, no. 5 (2008), pp. 4–13.
3. Uspenskiy V. M. Diagnostic System Based on the Information Analysis of Electrocardiogram. MECO 2012. Advances and Challenges in Embedded Computing (Bar, Montenegro, June 19-21, 2012), pp. 74–76.

# 4. Physical Activity and Behavior Recognition

---

→ **Goal:** to reveal sudden changes in user behavior



→ **Data:** multidimensional time series captured from a wearable device.

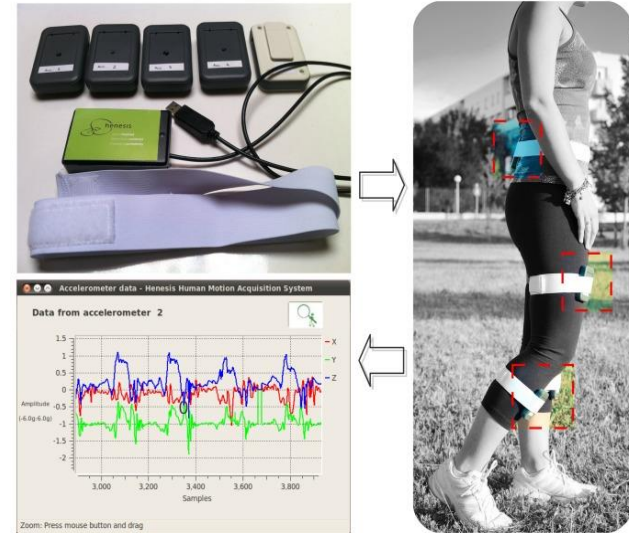
→ **Solution:** automatic generation of a Deep Learning network

# 4. Physical Activity and Behavior Recognition

→ Human motion tracker on mobile phone



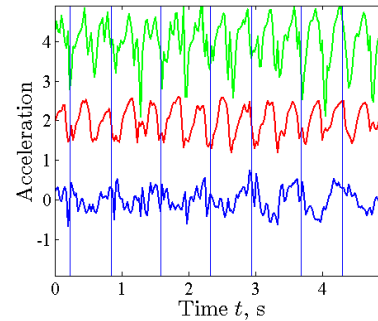
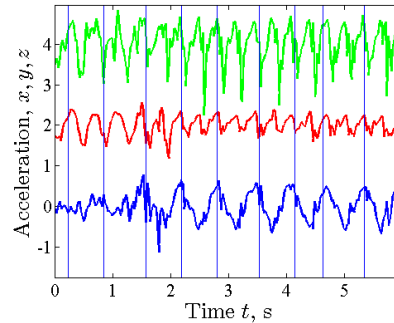
→ Wearable sensing system



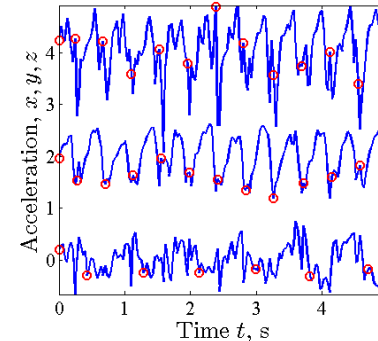
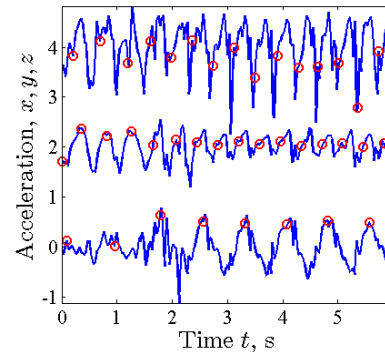
# 4. Physical Activity and Behavior Recognition

## → Time series segmentation for Jogging and Skipping

### → Manually



### → Automatically





# 4. Physical Activity and Behavior Recognition

## → Classification results

Data set:

1M examples

200 points each

	Predicted class						
	Jog	Walk	Up	Down	Sit	Stand	Accuracy
Jog	490	1	2	1	0	0	<b>0.99</b>
Walk	0	622	1	4	0	0	<b>0.99</b>
Up	1	2	154	5	0	0	<b>0.95</b>
Down	0	2	4	124	0	0	<b>0.95</b>
Sit	0	1	2	0	79	1	<b>0.95</b>
Stand	0	0	1	1	1	65	<b>0.96</b>

# 4. Physical Activity and Behavior Recognition

---

1. M.S. Popova, V.V. Strijov. *Selection of optimal physical activity classification model using measurements of accelerometer* // Informatics and applications, 2015.
2. A.D. Ignatov, V. V. Strijov. *Human activity types recognition using quasiperiodic sets of time series* // Multimedia Tools and Applications, 2015.
3. A. Motrenko, V. Strijov. *Extracting fundamental periods to segment human motion time series* // Journal of Biomedical and Health Informatics, 2015.
4. A. M. Katrutsa, M. P. Kuznetsov, V. V. Strijov, K. V. Rudakov. *Metric concentration search procedure using reduced matrix of pairwise distances* // Intelligent Data Analysis. Vol. 19(5). 2015.

# Skoltech students projects

---

1. *Fedor Chervinskii*. EEG Classification
2. *Alvis Logins*. TOUCH : In-Memory Spatial Join by Hierarchical Data-Oriented Partitioning
3. *Rustem Feyzkhanov*. Email filters generator.
4. *Sergei Kasatkin*. Determination of the type of human activity based on the data from the accelerometer
5. *Ekaterina Kotenko, Alexandra Kudryashova*. NDVI calculation for satellite images
6. *Mikhail Matrosov*. Short-term forecasting of musical compositions.
7. *Roman Prilepskiy*. Text detection.
8. *Oleg Urzhumtsev*. Dictionary builder.
9. *Irina Zhelavskaya*. Automatic Filters Generator for Gmail.
10. *Sergey Voronov*. Topic model for filtering scientific papers.

# Questions?

---

## Contacts:

Moscow Institute of Physics and Technology

→ Konstantin Vorontsov

e-mail: [voron@forecsys.ru](mailto:voron@forecsys.ru)

URL: <http://www.MachineLearning.ru/wiki> User:Vokov

→ Vadim Strijov

e-mail: [strijov@gmail.com](mailto:strijov@gmail.com)

URL: <http://www.strijov.com>