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, …

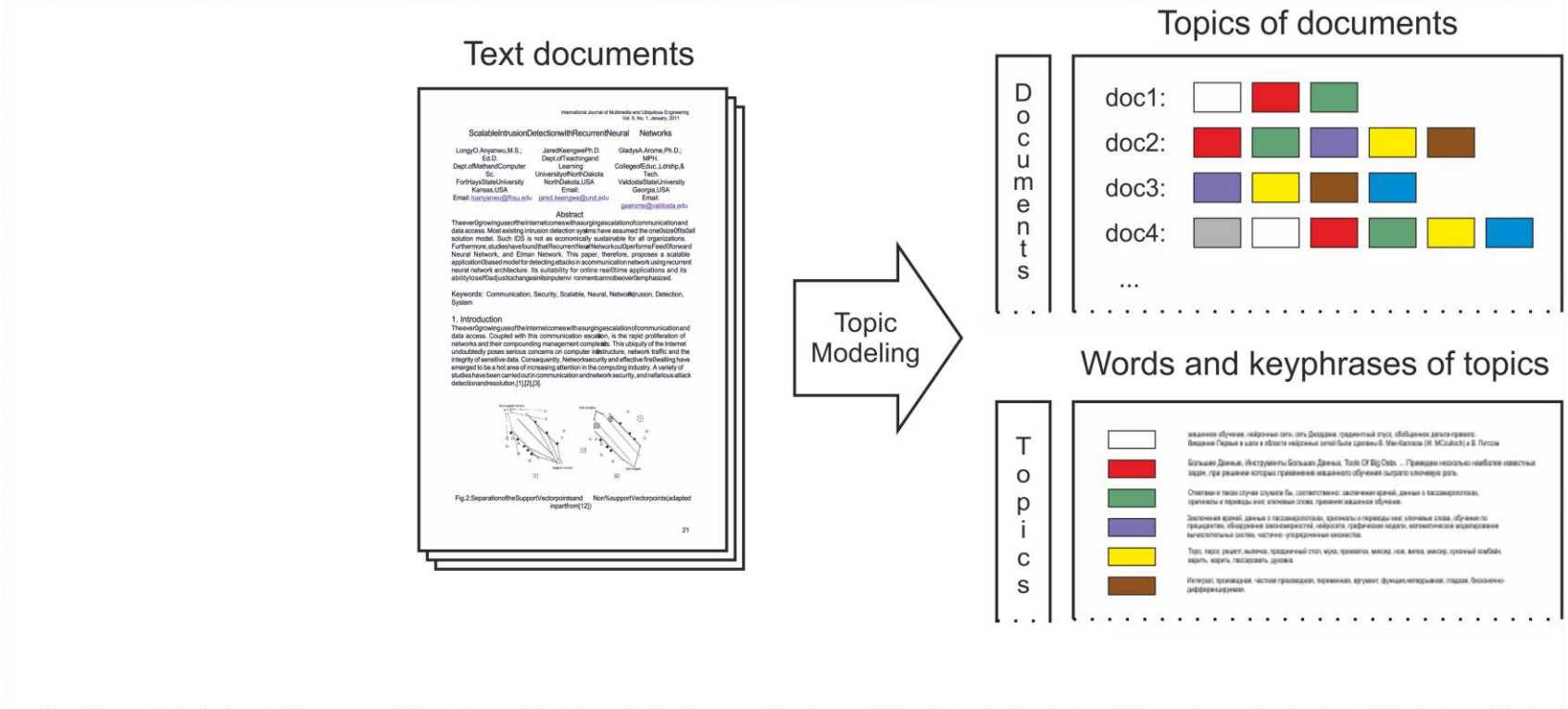
Activities

- Research
- Development
- Education
- Innovation

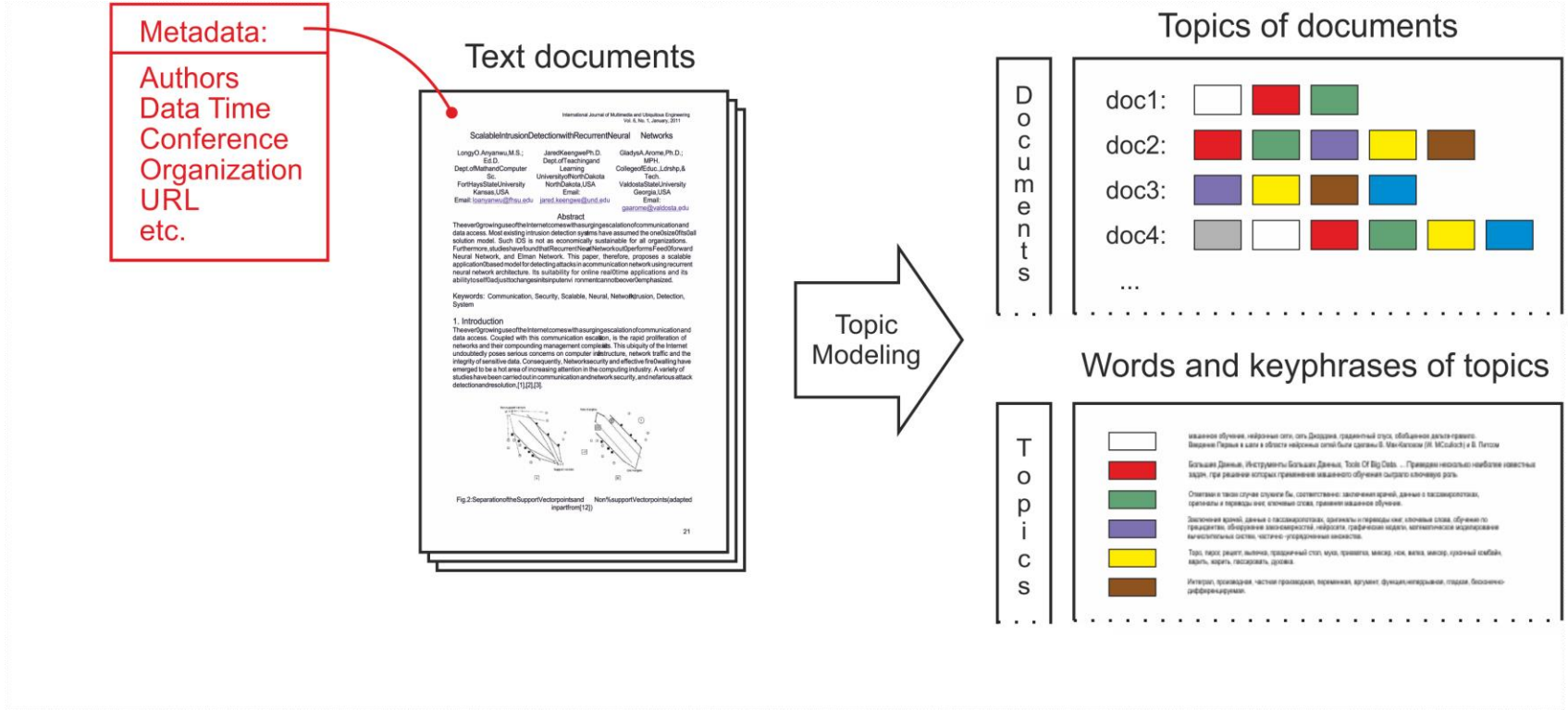
Projects of MIPT group:

- **Texts:**
  - 1. BigARTM for Topic Modeling
  - 2. Conference Hierarchy Tool
- **Signals:**
  - 3. ECG Multi-Disease Diagnostics
  - 4. Human Behavior Recognition

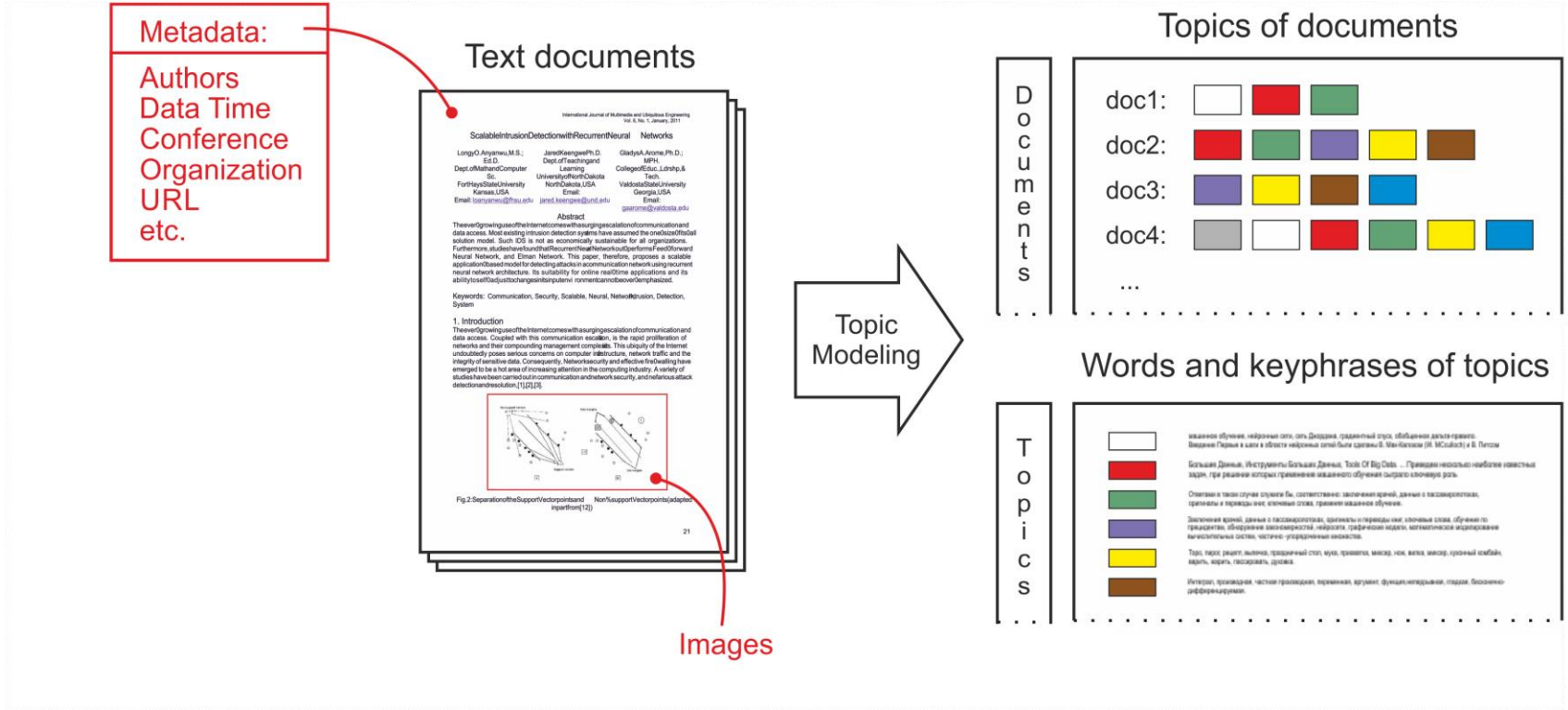
1. Topic Modeling for big text collections



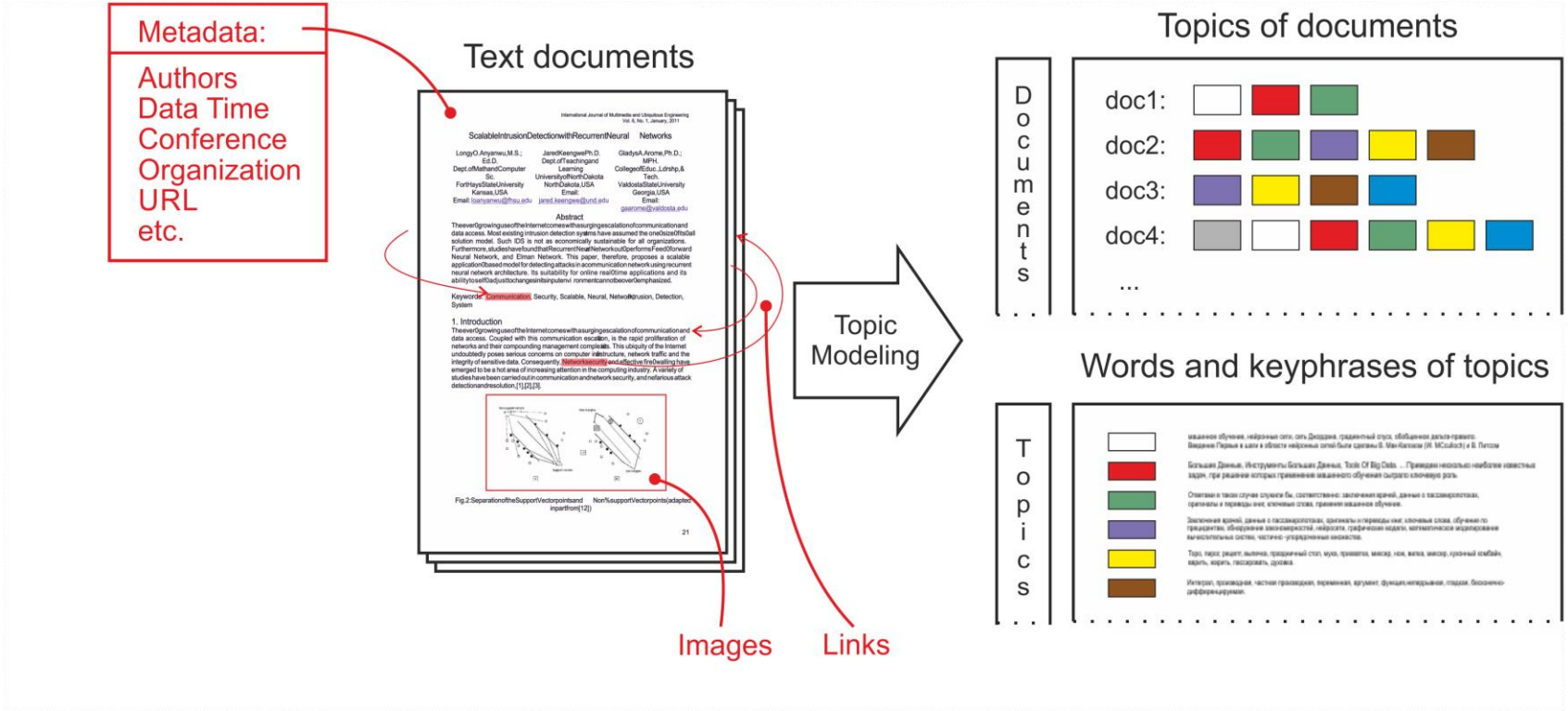
1. Topic Modeling for big text collections



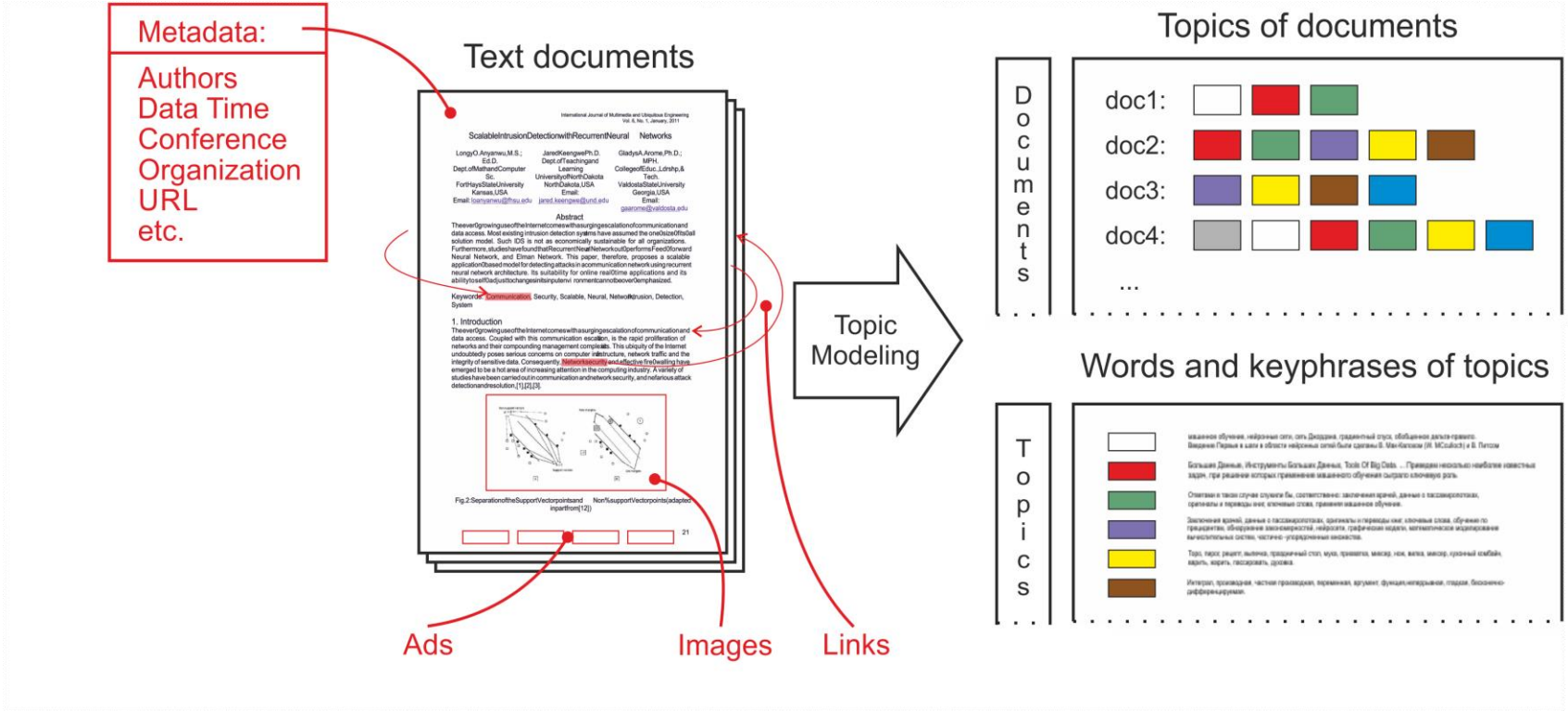
1. Topic Modeling for big text collections



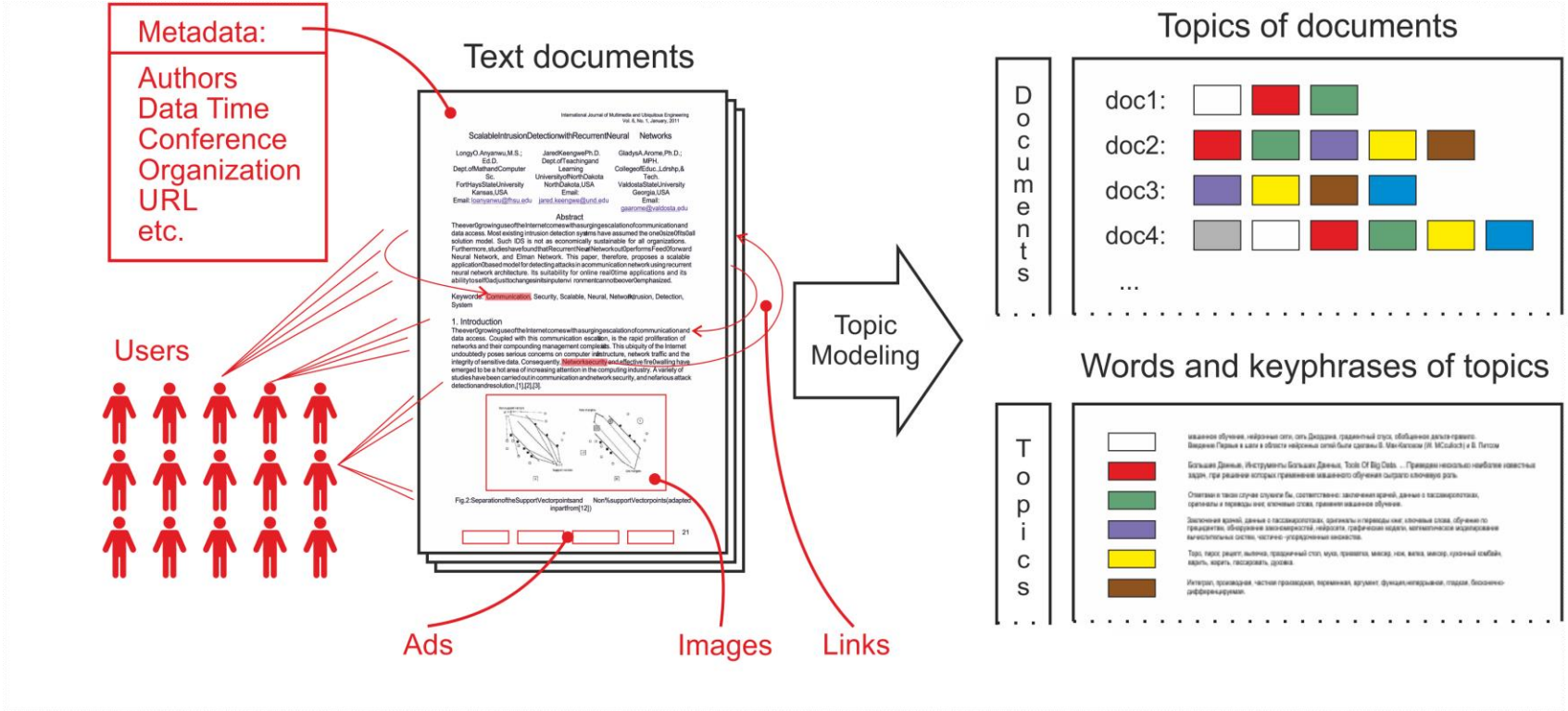
1. Topic Modeling for big text collections



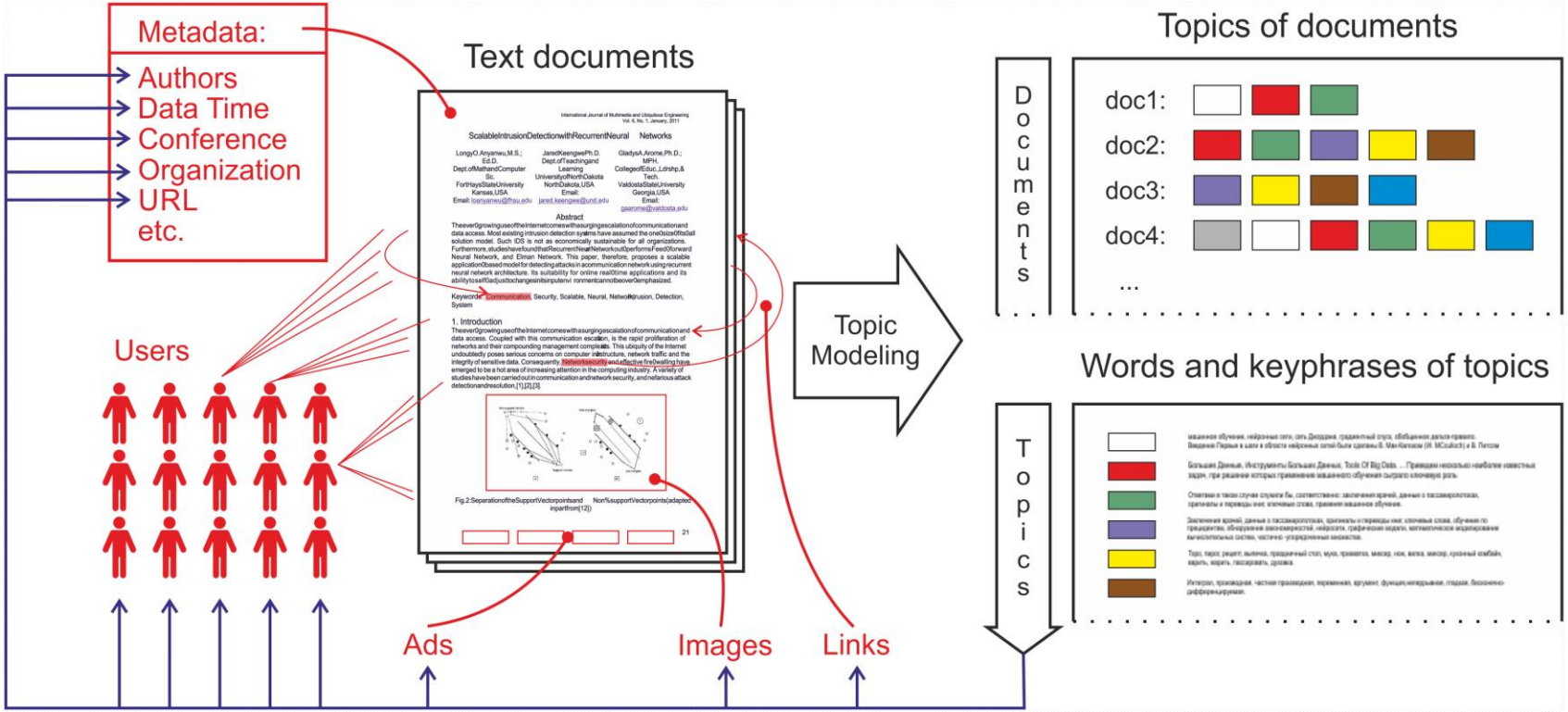
1. Topic Modeling for big text collections



1. Topic Modeling for big text collections

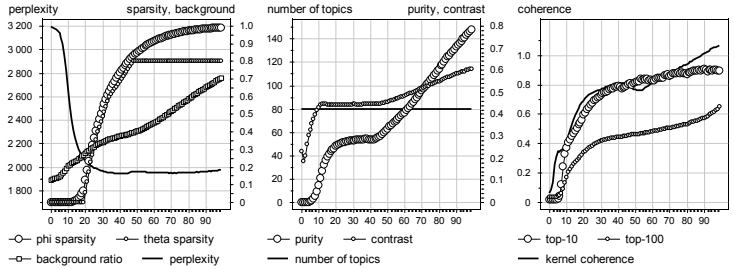


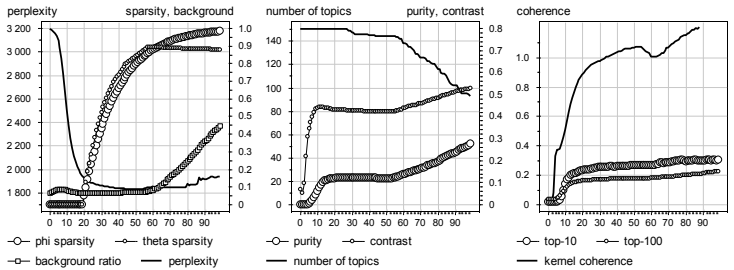
1. Topic Modeling for big text collections



1. Topic Modeling: BigARTM project

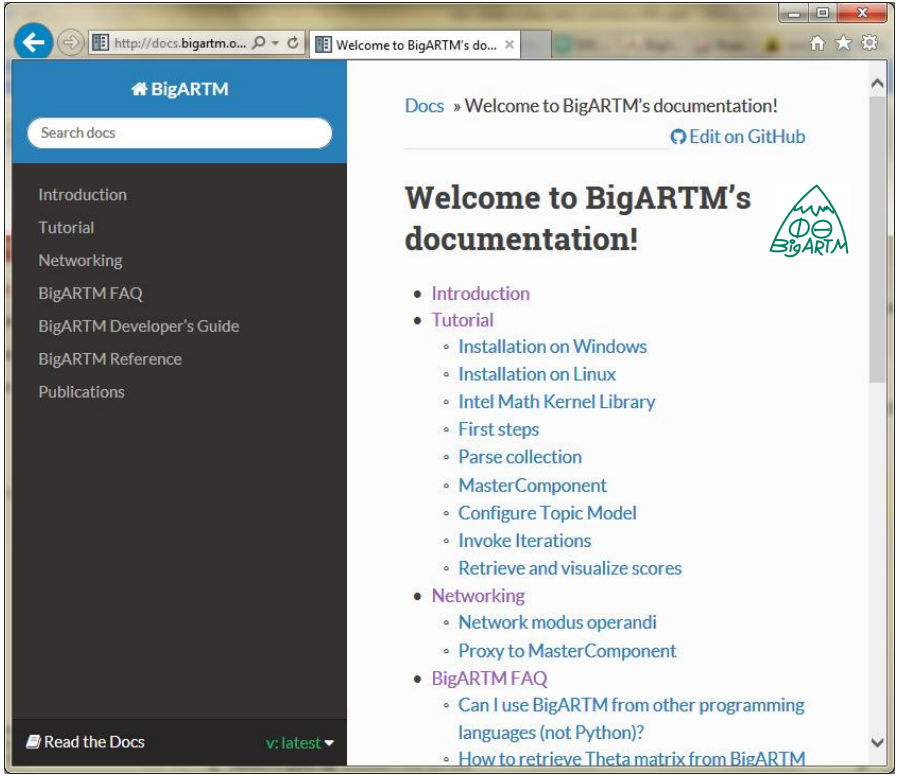
- **Challenge:** how to combine many functionalities in a single Topic Model
- **Theory:** ARTM – Additive Regularization for Topic Modeling
  (easy to understand – easy to design – easy to infer – easy to combine)
- **Implementation:** BigARTM – open-source with permissive license
- **Experiments:**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

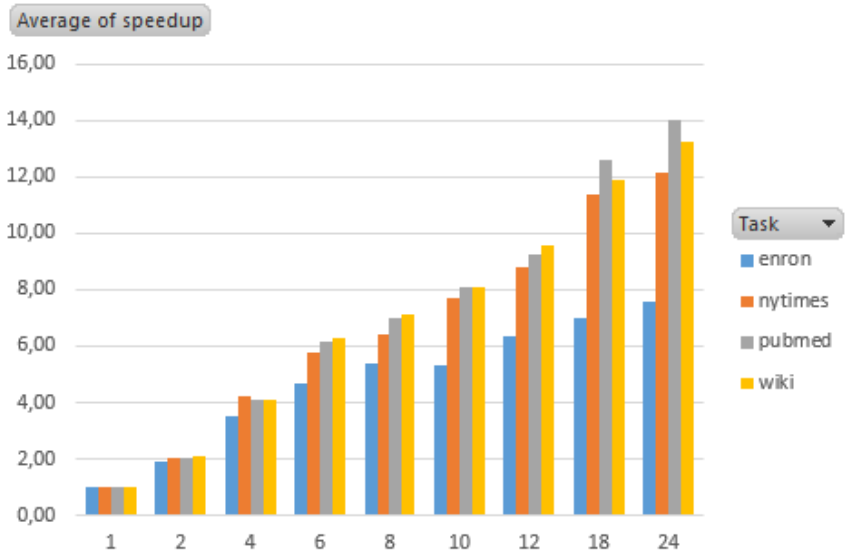


1. Topic Modeling: BigARTM project

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.

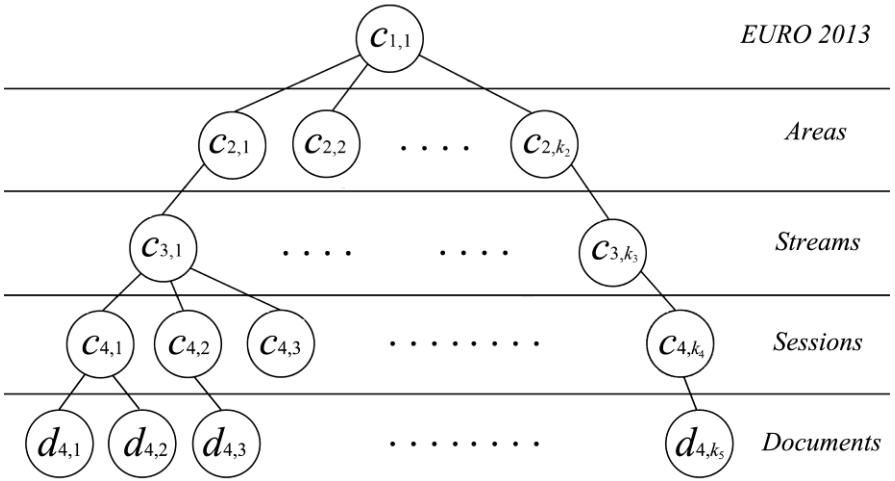
2. Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'2014, Analysis of Images, Social networks and Texts. ***Springer, Communications in Computer and Information Science***, 2014. Vol. 436. pp. 29–46.

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

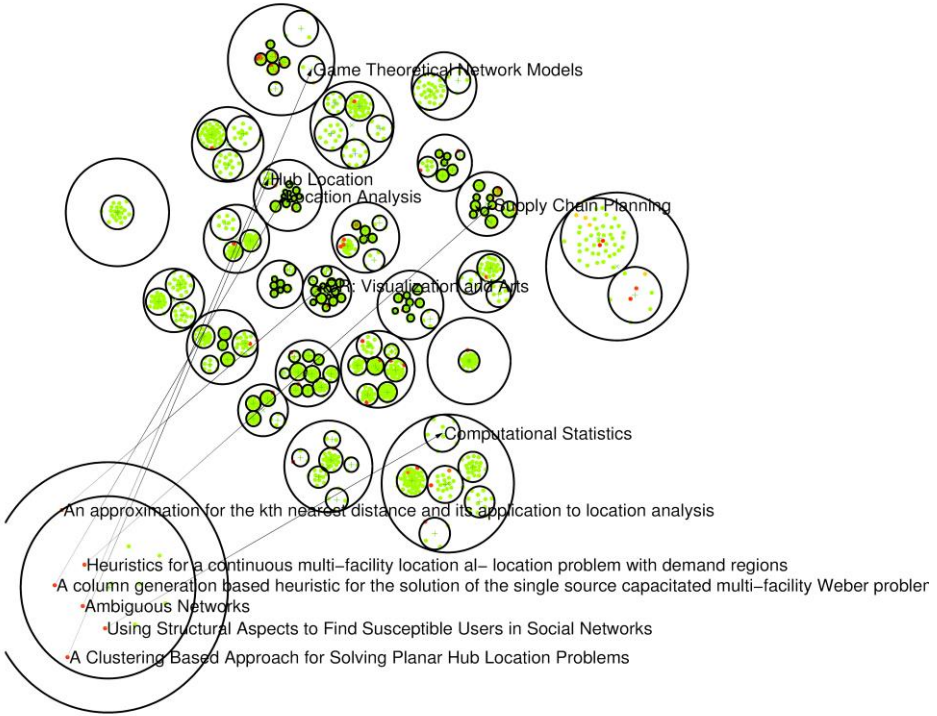
- **The goal:**
  automatic construction of
  a scientific conference program

- **EURO –** *European Conference on Operational Research:*
  >3500 participants,
  >200 experts,
  24 areas, 137 streams



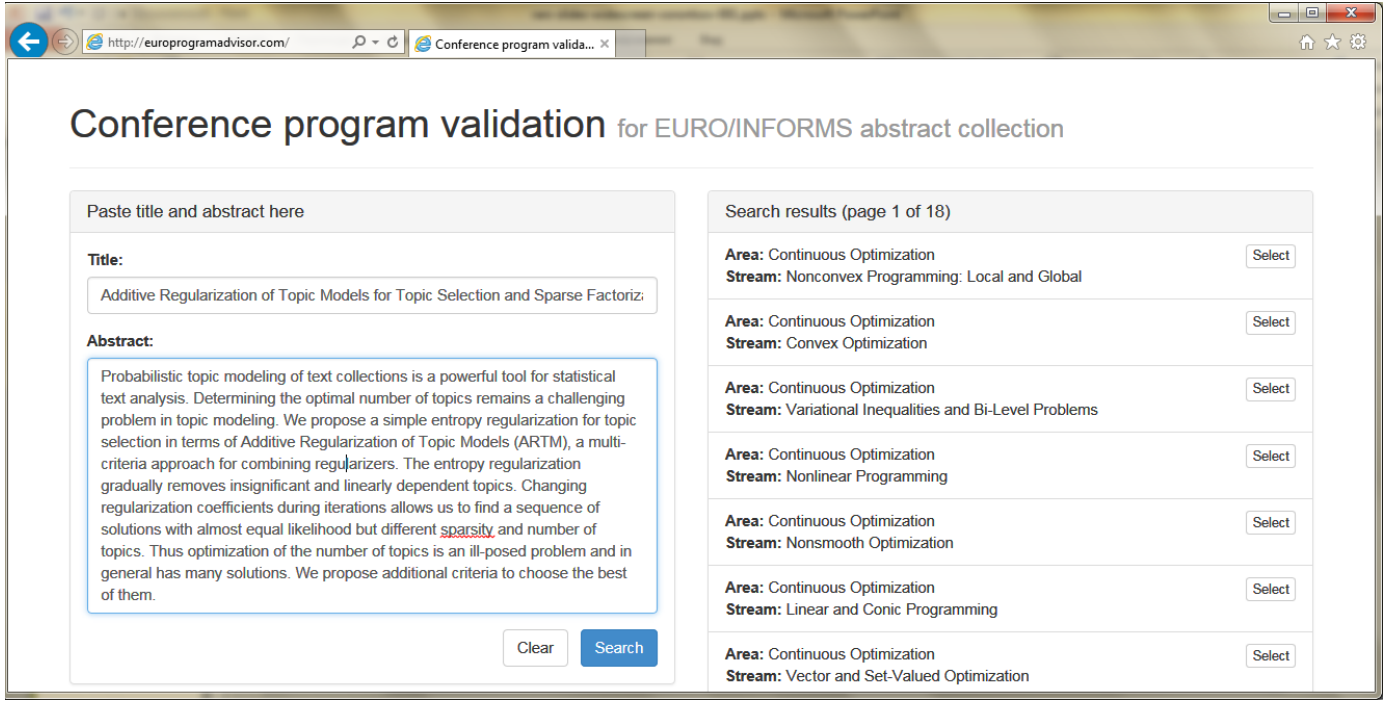
2. Hierarchical Conference Topic Model

- Visualization of inconsistencies:
- Green points
  – consistent submissions
- Red points
  – inconsistent submissions



2. Hierarchical Conference Topic Model

➜ http://EUROprogramAdvisor.com

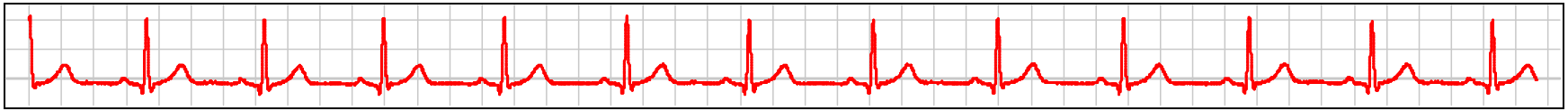


2. Hierarchical Conference Topic Model

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

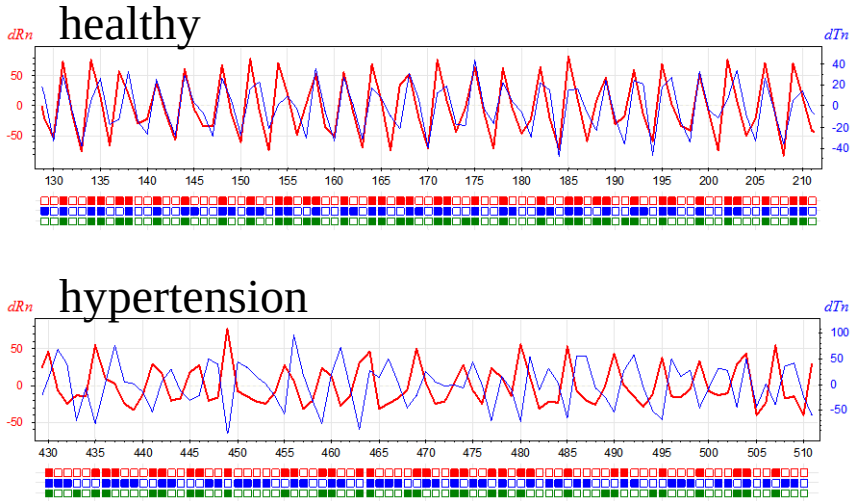


1. Measuring RR-interval and amplitude of each R-peak
2. Discretization

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

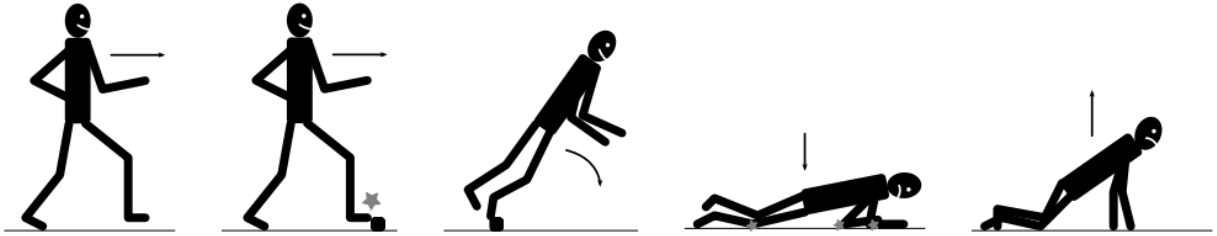
disease	cases	AUC, %	spec, % (sens=95%)
femoral head necrosis	327	99.19 ± 0.10	96.6 ± 1.76
cholelithiasis	277	98.98 ± 0.23	94.4 ± 1.54
coronary heart disease	1262	97.98 ± 0.14	91.1 ± 1.86
gastritis	321	97.76 ± 0.11	88.3 ± 2.64
hypertensive disease	1891	96.76 ± 0.09	84.7 ± 1.99
diabetes	868	96.75 ± 0.19	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

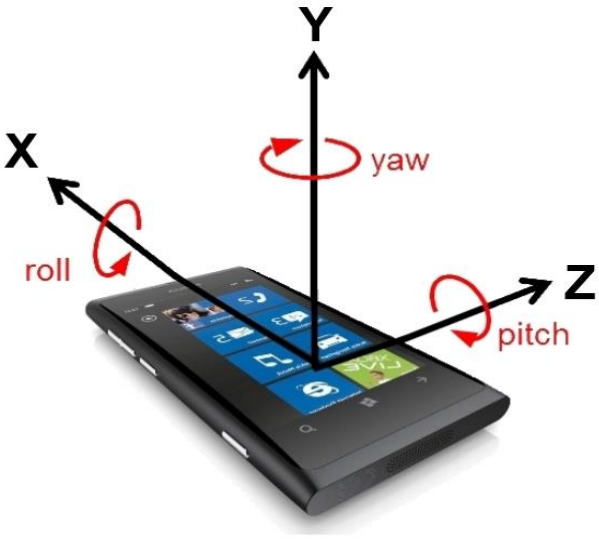
- **Goal:** to reveal sudden changes in user behavior



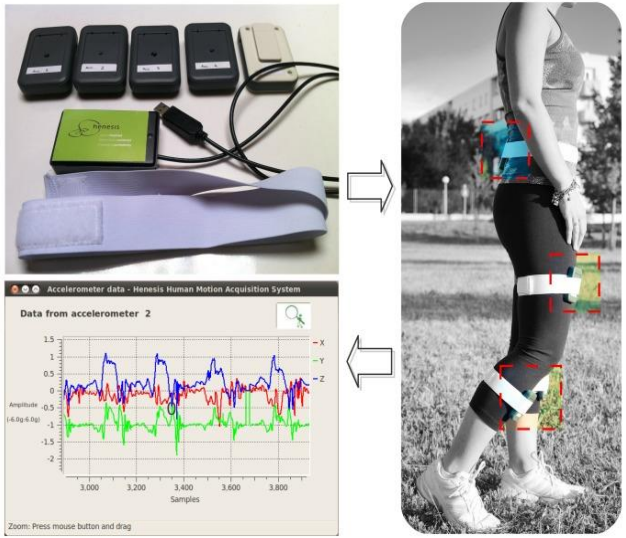
- **Data:** multidimensional time series captured from a wearable device.
- **Solution:** automatic generation of a Deep Learning network

4. Physical Activity and Behavior Recognition

- Human motion tracker on mobile phone

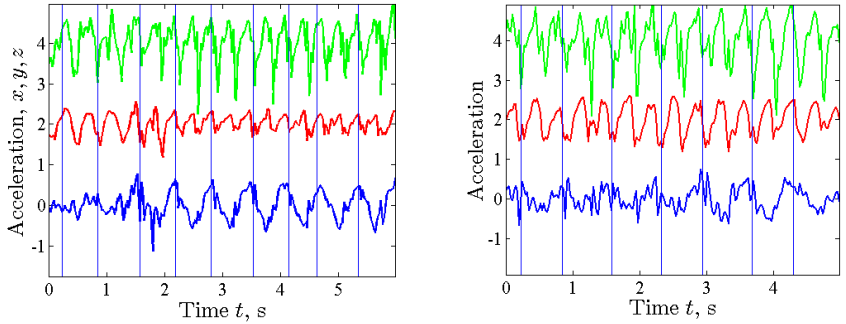


- Wearable sensing system

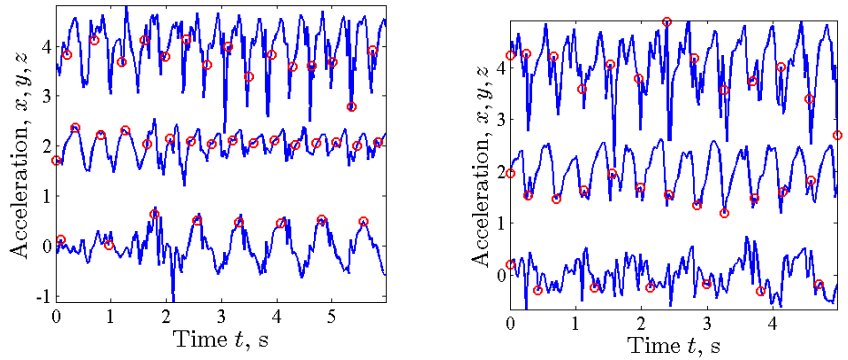


4. Physical Activity and Behavior Recognition

- 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	**0.99**
Walk	0	622	1	4	0	0	**0.99**
Up	1	2	154	5	0	0	**0.95**
Down	0	2	4	124	0	0	**0.95**
Sit	0	1	2	0	79	1	**0.95**
Stand	0	0	1	1	1	65	**0.96**

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

- Vadim Strijov
  e-mail: strijov@ccas.ru
  URL: http://www.ccas.ru/strijov