

Московский Физико-Технический Институт



Лаборатория Машинного Интеллекта

<http://mipt.ai>

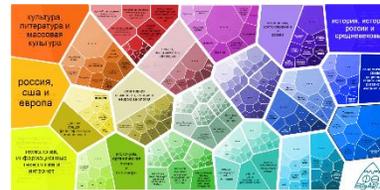
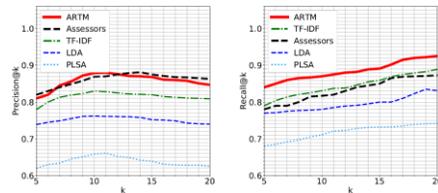
заведующий лабораторией, д.ф.-м.н., проф.РАН

Воронцов Константин Вячеславович

k.v.vorontsov@phystech.edu

Направления исследований лаборатории:

• Анализ текстов и транзакционных данных



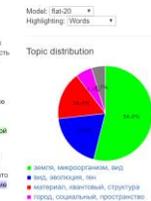
Что исследователи знают о химической коммуникации планктона в воде? Какими сигналами обмениваются зоопланктон? Как взаимодействуют зоопланктон? СД этим занимается кандидат биологических наук Егор Макаров.

Планктон — это организм, выходящий из воды только в состоянии перемещения. То есть это что-то маленькое, то, что перемещается течениями. Планктон делится на фитопланктон (это водоросли) и зоопланктон. Мы будем говорить про зоопланктон — это рыбы. То, как разные объекты между собой коммуницируют с помощью химических сигналов, исследовано довольно плохо. В химических экосистемах мы знаем, есть феромоны, различные химические системы, которые хорошо исследованы. Мы исследуем на дне океана планктон, например, для рыбачества — феромоны лобстера. Вода — это среда, которая благоприятна для химической коммуникации.

[DOI: 10.1371/journal.pone.0166228]

Химические сигналы от рыбок заставляет зоопланктон мигрировать. Это одно из самых масштабных на планете перемещений биомассы, которые электроно-приводятся в океане, море и озерах. Зоопланктон очень подвижен и поворачивается в одну сторону на секунду. Делая свой выбор, планктон вынужден поворачиваться, и вынужден уходить на глубину, и ночью поднимается в поверхность, чтобы есть. Было показано, что эти вертикальные миграции регулируются двумя факторами. Первый — это освещенность. Конечно, что если не будет света, не будет сигнала. А второй — это запах, который выделяет хищник.

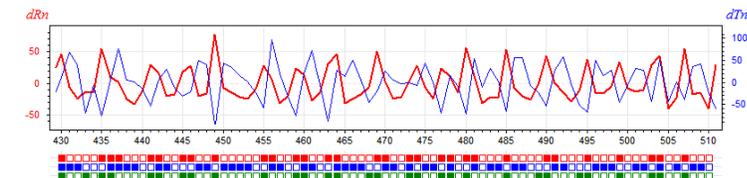
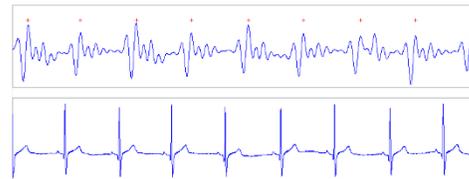
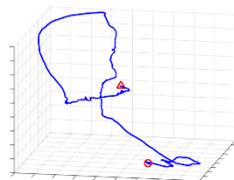
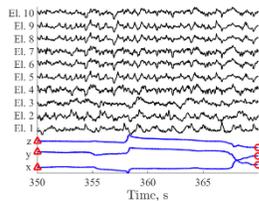
В 2008 и 2009 годах мы провели крупные выборы по химическим коммуникациям. То есть в это время, маленькие молекулы, и в) они работают в очень низкой концентрации. Это до сих пор остается и порождает, потому что сообщества зоопланктона и вообще планктон в разных экосистемах — это очень разные ассоциации рыбок, которые живут в озерах, морях, взаимодействуют между собой. А между ними есть очень много биомассы, сдвиг то, что мы получаем в лаборатории, и разветвленная сеть химических сигналов и коммуникаций, которые влияют на разные поведенческие, физиологические и продуктивные функции. И эти данные очень важны для понимания до сих пор слабо исследованной.



• Анализ изображений и видео



• Анализ временных рядов и интернет вещей



Тематическое моделирование текстов

Text documents

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

Scalable Intrusion Detection with Recurrent Neural Networks

Longyi O Anyanwu, M.S.; Jared Keengwe, Ph.D.; Gladys A. Arome, Ph.D.;
Ed.D.; Dept. of Teaching and Learning; MPH;
Dept. of Math and Computer Sc.; University of North Dakota; College of Educ., Ldrshp. & Tech.
Forthays State University; North Dakota, USA; Valdosta State University; Georgia, USA
Email: loanyanwu@fhsu.edu; jared.keengwe@und.edu; gaarome@valdosta.edu

Abstract
The ever growing use of the Internet comes with a surging escalation of communication and data access. Most existing intrusion detection systems have assumed the one-size-fits-all solution model. Such IDS is not as economically sustainable for all organizations. Furthermore, studies have found that Recurrent Neural Network outperforms 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-tune to changes in its environment cannot be overemphasized.

Keywords: Communication, Security, Scalable, Neural, Network, Intrusion, Detection, System

1. Introduction
The ever growing use of the Internet comes with a surging escalation of communication and data access. Coupled with this communication escalation, is the rapid proliferation of networks and their compounding management complexities. This ubiquity of the Internet undoubtedly poses serious concerns on computer infrastructure, 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 nefarious attack detection and resolution [1], [2], [3].

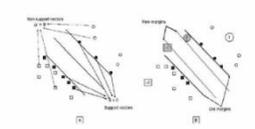
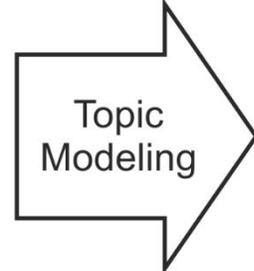


Fig. 2: Separation of the Support Vector points and Non-Support Vector points (adapted in part from [12])

21



Topics of documents

Documents

doc1:

doc2:

doc3:

doc4:

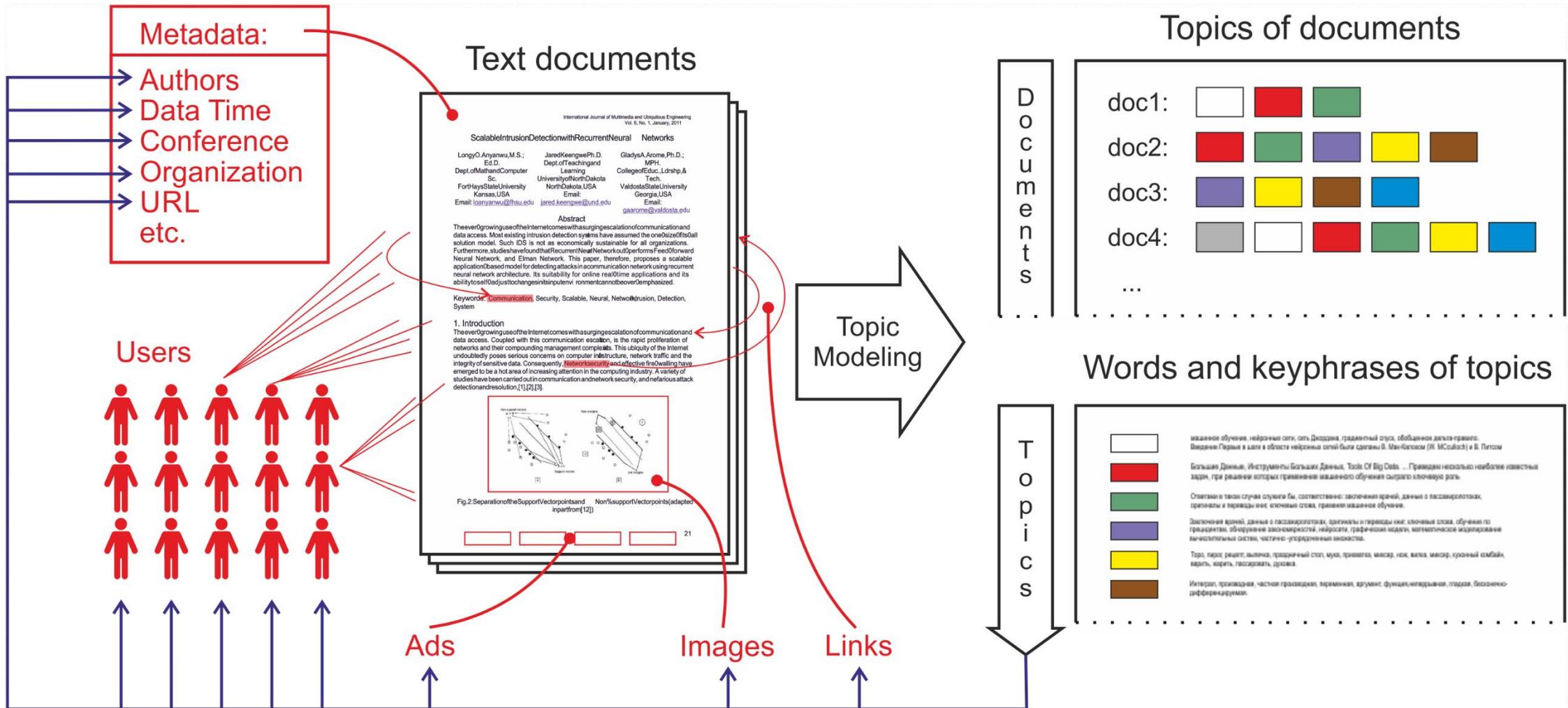
...

Words and keyphrases of topics

Topics

- машинное обучение, нейронные сети, сеть Джексона, градиентный спуск, обобщение данных, графы. Видение Перель в шаше в области нейронных сетей была сделана В. Мак-Калом (W. McCulloch) и В. Питсом
- Большая Данные, Инструменты Больших Данных, Tools Of Big Data. ... Примеры нескольких наиболее известных задач, при решении которых критически важно машинное обучение сыграло ключевую роль
- Отчеты в таком случае случились бы, соответственно: заключены арены, данные в паскариловых, критерии и парадигмы, ключевые слова, грамматики машинное обучение.
- Заключены арены, данные в паскариловых, критерии и парадигмы; ключевые слова, обучение по градиенту, обучение закономерностей, нейросети, графовые модели, компьютерное моделирование вычислительных систем, частично-упреждаемые вычисления.
- Таро, парок, реант, вычисления, градиентный спуск, мук, приемы, мексик, нек, вела, мексик, графический софт, верь, верь, паскарилов, дрова.
- Интернет, прохождение, частоты графовых, переменная, аргумент, функция, непрерывная, гитары, бесконечно-дифференцируемая.

Тематическое моделирование текстов



Тематическое моделирование текстов

- **Теория:** ARTM – Additive Regularization for Topic Modeling
- **Технология:** BigARTM – библиотека тематического моделирования с открытым кодом
- **Приложения:**
 - разведочный информационный поиск
 - обнаружение событий в новостных потоках
 - обработка записей разговоров в контакт-центрах
 - выявление типов потребления по банковским транзакциям
 - выявление видов экономической деятельности компаний



<http://bigartm.org>

	procs	T = 50		T = 200	
		time, m	perplexity	time, m	perplexity
BigARTM	1	42	5117	83	3347
BigARTM async	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM async	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM async	8	5	6220	10	4309
Gensim	8	88	6344	288	4263

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and Modular Regularized Topic Modelling. The 21st Conference of Open Innovations Association FRUCT: Intelligence, Social Media and Web, Finland, Helsinki, November 6-10, 2017.

Тематический разведочный поиск

- Длинные запросы (1 стр. А4)
- 100 запросов
- 3 ассессора на каждый запрос
- 30 минут в среднем на запрос
- Разметка на Яндекс.Толока
- Коллекции техно-новостей

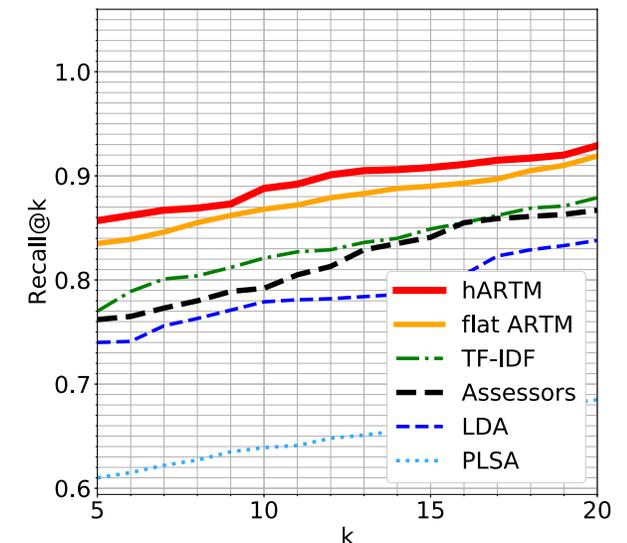
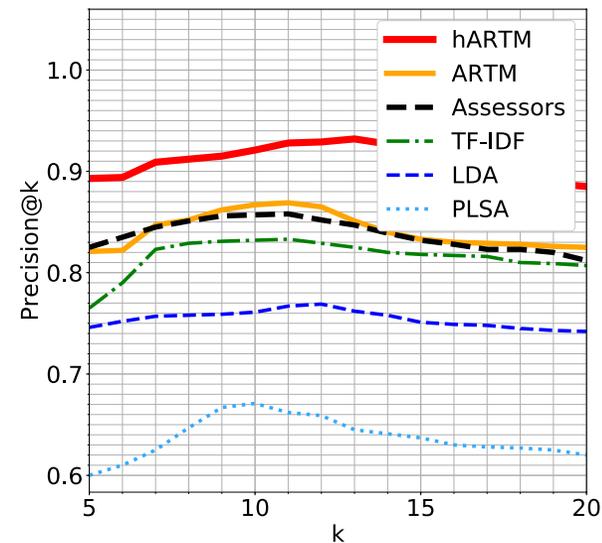


(170K Russian docs)



(750K English docs)

Результат:
точность (precision) и полнота (recall) поиска



Тематизация банковских транзакций

- **Транзакционные данные физических лиц:**

документ → клиент, слово → тип продавца, тема → тип потребления

Цель: формирование персональных предложений клиентам

- **Транзакционные данные юридических лиц:**

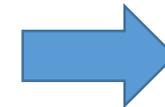
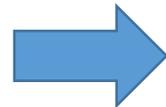
документ → компания, слово → контрагент, тема → вид деятельности

Цель: отраслевой консалтинг для компаний малого бизнеса

Анализ и распознавание изображений: Text detection & OCR

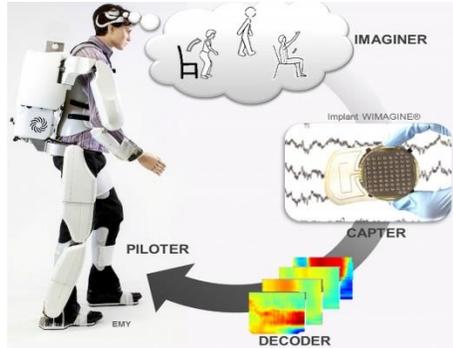
Автоматическое распознавание сканированных документов

- распознавание геометрической структуры документа
- обнаружение текстовых строк
- распознавание текста

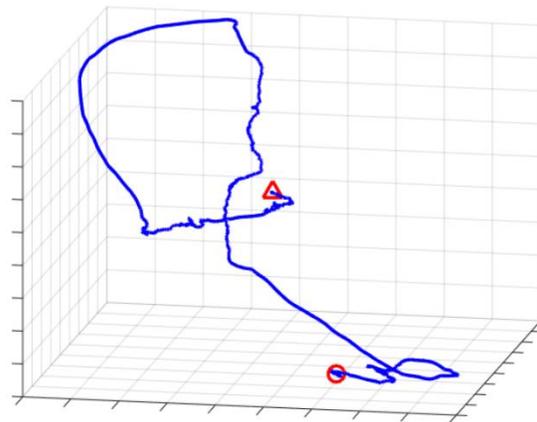
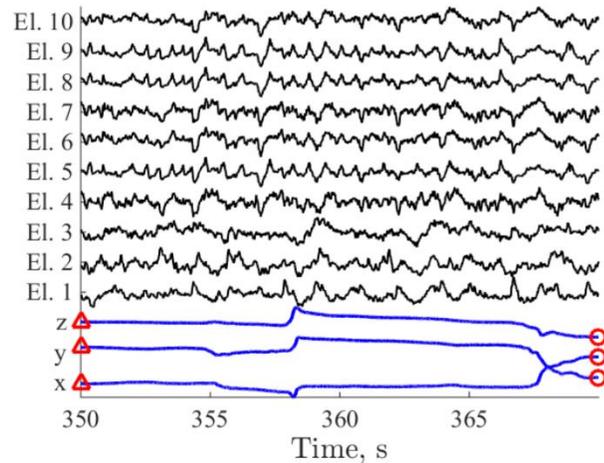


```
<elements>
  <element>
    <name>head</name>
    <type>Complex</type>
    <value>
      <element>
        <name>number</name>
        <type>Integer</type>
        <value>4201</value>
      </element>
      <element>
        <name>operator</name>
        <type>String</type>
        <value>Иванов</value>
      </element>
    </value>
  </element>
  <element>[ ]</element>
  <element>[ ]</element>
  <element>
    <name>control-block</name>
    <type>Complex</type>
    <value>
      <element>[ ]</element>
      <element>[ ]</element>
      <element>
        <name>control-id-2</name>
        <type>String</type>
        <value>084432</value>
      </element>
    </value>
  </element>
</elements>
```

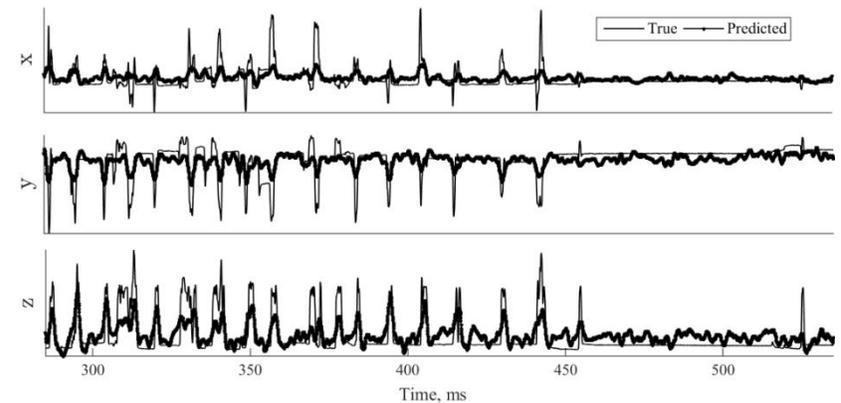
Анализ временных рядов: распознавание движений по электрокортикограмме



BCI-проект WIMAGINE (совместно с clinatec.fr).
Цель – создание системы компенсации нарушений двигательного аппарата человека

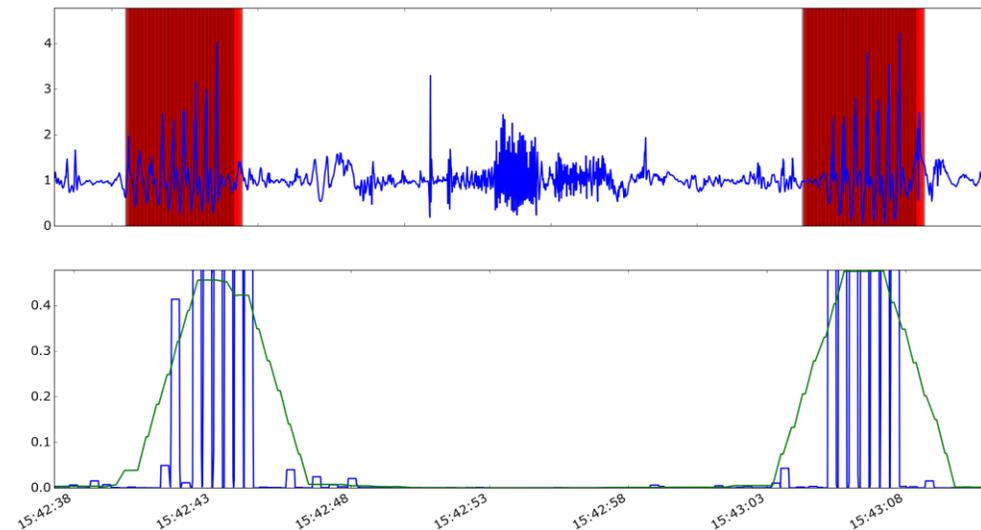


The wrist motion trajectory prediction

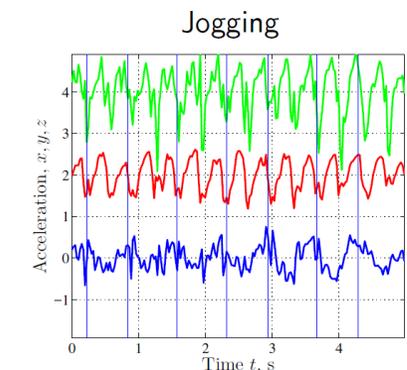
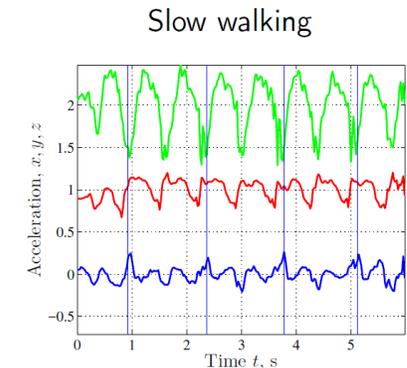


Анализ временных рядов: распознавание движений по данным акселерометров

- Бизнес-приложение: мониторинг физической активности рабочих на производстве или на стройке

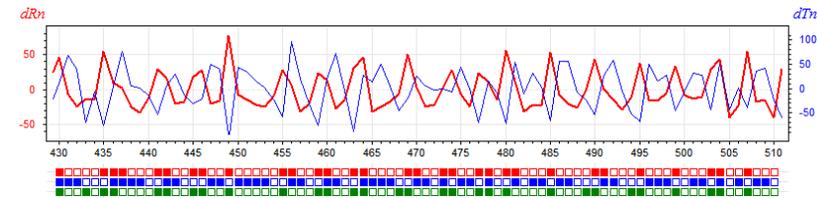
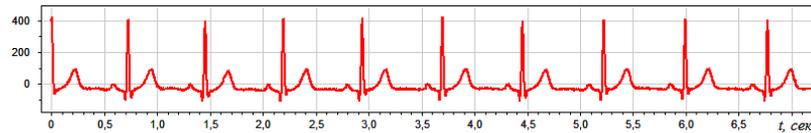


Пример: забивание гвоздя – активность,
состоящая из элементарных движений



Символьная динамика: оценивание рисков заболеваний по электрокардиограмме

1. Вычисление признаков кардиоциклов



2. Дискретизация и векторизация

```

DBEEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAFAFFAFAFFAFAFFAEBFAEBFAEFAAFAAFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFAEADFCAFBCADFFECFFAFAFFAFAFAEFFCACFCAEFFCAD
DAADBFAAFFAEFBABFACDFFAAFBAADFADFDAAFCFCDFCEEFCAEFBECBBBAADBAACFFAFAFFA
CFFCECFDAABDAEFFAFAFFCEDBFAAFFAEFFAEFBACFBADFEAFAFFCAFFDAFFAEBDAAADBBADDAFF
EABFCCAFDEEBDECFACFFAABFAADFBAFFACFFFAEFFACFFACFFCECFBAFFFAFFFAFFFAADFB
AABFACDFDAEFFAADBAEFFEAFBCECFDECCFBAFFAADFACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCEFFAFAFFABCFDAAFFADBFCAEFFAABFACBFABEFAEBFAEBCAFFBAFFFAAFFDACFDAAFB
CAFFAEFCFFACFFACDFDADFDAABFAEODDABBFACDVAFFFAFFCADFAADFACFFAEDFCACFCAEBC
    
```

3. Машинное обучение

4. Оценивание на тестовых данных

болезнь	выборка	AUC, %	C% при Ч=95%
некроз головки бедренной кости	327	99.19 ± 0.10	96.6 ± 1.76
желчнокаменная болезнь	277	98.98 ± 0.23	94.4 ± 1.54
ишемическая болезнь сердца	1262	97.98 ± 0.14	91.1 ± 1.86
гастрит	321	97.76 ± 0.11	88.3 ± 2.64
гипертоническая болезнь	1891	96.76 ± 0.09	84.7 ± 1.99
сахарный диабет	868	96.75 ± 0.19	85.3 ± 2.18
аденома простаты	257	96.49 ± 0.13	80.1 ± 3.19
рак	525	96.49 ± 0.28	82.2 ± 2.38
узловой зоб щитовидной железы	750	95.57 ± 0.16	73.5 ± 3.41
холецистит хронический	336	95.35 ± 0.12	74.8 ± 2.46
дискинезия ЖВП	714	94.99 ± 0.16	70.3 ± 4.67
мочекаменная болезнь	649	94.99 ± 0.11	69.3 ± 2.14
язвенная болезнь	779	94.62 ± 0.10	63.6 ± 2.55

Публикации 2017-2018

- *D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.* Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.
- *A.Ianina, L.Golitsyn, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.
- *Aduenko A.A., Motrenko A.P., Strijov V.V.* Object selection in credit scoring using covariance matrix of parameters estimations // Annals of Operations Research, 2017
- *Kulunchakov A.S., Strijov V.V.* Generation of simple structured Information Retrieval functions by genetic algorithm without stagnation // Expert Systems with Applications, 2017
- *Katrutsa A.M., Strijov V.V.* Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria // Expert Systems with Applications, 2017
- *Бочкарев А.М., Софронов И.Л., Стрижов В.В.* Порождение экспертно-интерпретируемых моделей для прогноза проницаемости горной породы // Системы и средства информатики, 2017
- *N.Skachkov, K.Vorontsov.* Improving topic models with segmental structure of texts. Dialog-2018. (accepted)
- *V.Bulatov, V.Alekseev, K.Vorontsov.* Intra-Text Coherence as a Measure of Topic Models Interpretability. Dialog-2018.
- *M.Selezneva, A.Sholokhov, A.Belyi.* Quality evaluation and improvement for hierarchical topic modeling. Dialog-2018.
- *A.Motrenko, V.Strijov.* Multi-way feature selection for ECoG-based BCI. Expert Systems with Applications, 2018