

Я

ПРОФИ

**СТУДЕНЧЕСКАЯ
ОЛИМПИАДА
Я – ПРОФЕССИОНАЛ**

Машинное обучение для анализа текстов и сложно структурированных данных

Воронцов Константин Вячеславович

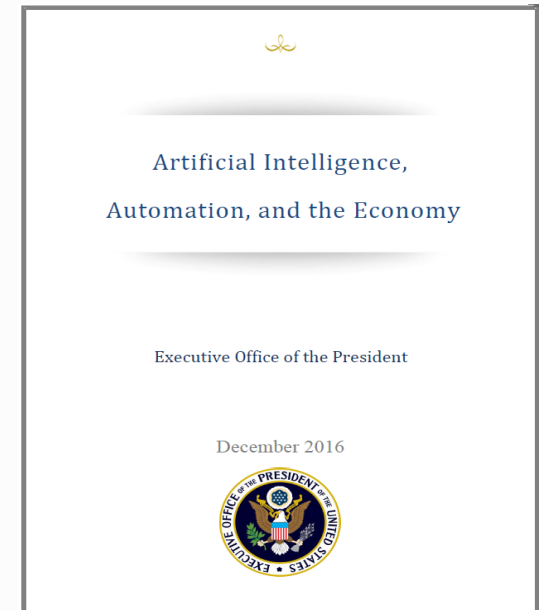
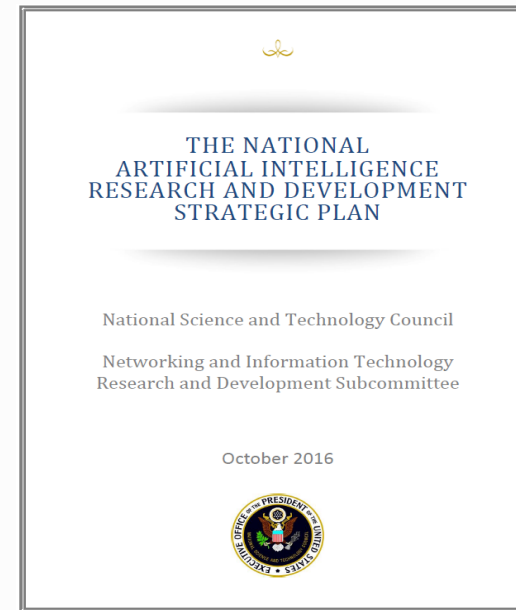
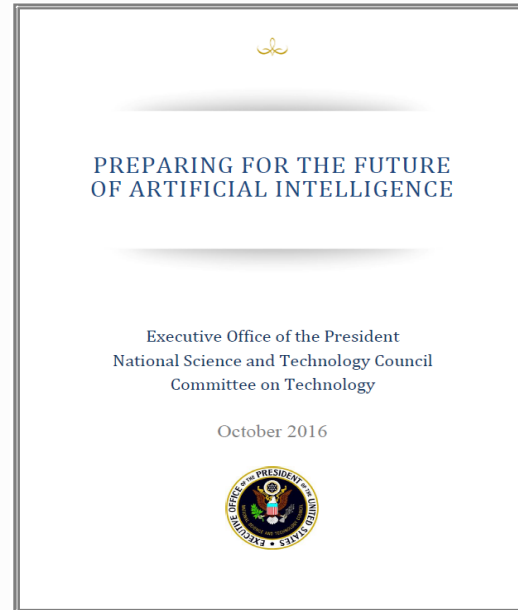
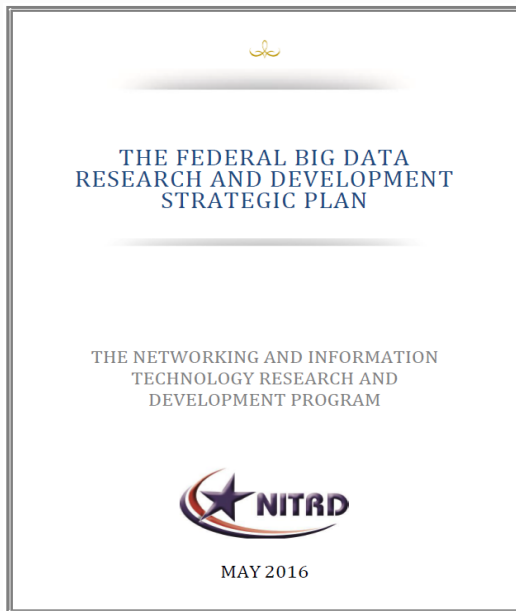
**д.ф.-м.н., профессор РАН,
руководитель лаборатории Машинного интеллекта МФТИ**

«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, *искусственном интеллекте и машинном обучении*» (2016)



Клаус Мартин Шваб,
президент Всемирного
экономического форума

Отчёты Белого дома США, май-октябрь 2016

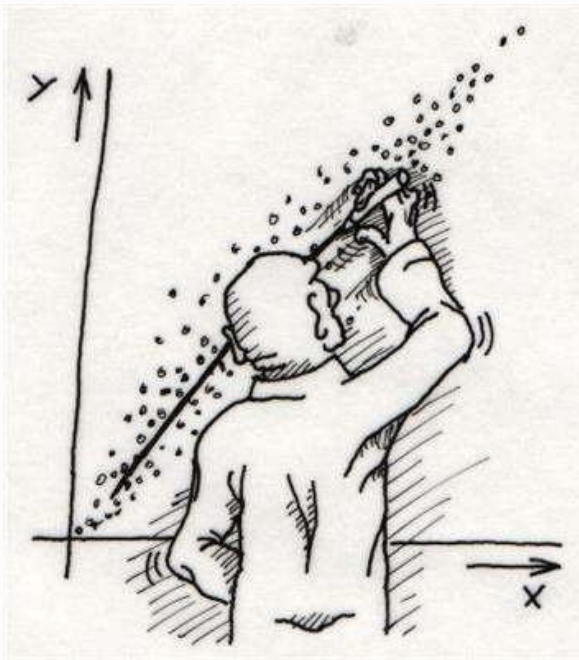


«Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future»

Машинное обучение

(Machine Learning, ML)

- одна из ключевых информационных технологий будущего
- наиболее успешное направление искусственного интеллекта, вытеснившее экспертные системы и инженерию знаний

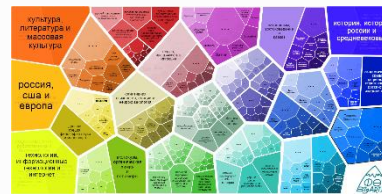
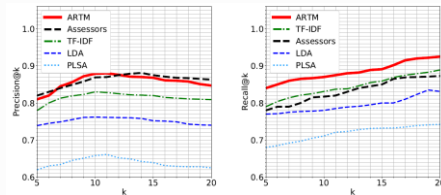


- *проведение функции через заданные точки в сложно устроенных пространствах*
- математическое моделирование в условиях, когда знаний мало, данных много
- тысячи различных методов и алгоритмов
- 100 тыс. научных публикаций в год

Машинный интеллект

тематика исследований лаборатории МФТИ

• Анализ текстов и транзакционных данных



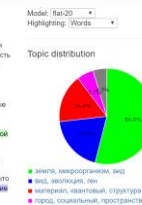
«Что исследователи знают о химической коммуникации планктона в воде? Какие сигналы обмениваются зоопланктоном? Как взаимодействуют зоопланктоны? СД этот рассказывает кандидат биологических наук Егор Шадрин.

Планктон — это организм, способный жить в водной толще в состоянии бесполого размножения. То есть это что-то маленькое, то, что переносится течениями. Планктон делится на фитопланктон (это водоросли) и зоопланктон. Мы будем говорить про зоопланктон — это рачки. То, как разные объекты между собой коммуницируют с помощью химических сигналов, исследовано довольно плохо. В основном зоопланктоны, мы знаем, есть феромоны, различные сигнальные системы, которые хорошо исследованы. Мы исследуем на дни создания планктона, например, для водорослей — феромоны лебедуш. Вода — это среда, которая благоприятна для химической коммуникации.

[DOI: 10.137097]

Химические сигналы от водорослей заставляет зоопланктон мигрировать. Это одно из самых масштабных на планете перемещений биомассы, которые естественным образом происходят в океанах, морях и озерах. Зоопланктон очень подвижен и поворачивается в разные стороны на планету. Делая свой путь, планктон способен менять высоту, и животные уходят на глубину и ночью поднимаются 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 фитопланктон.

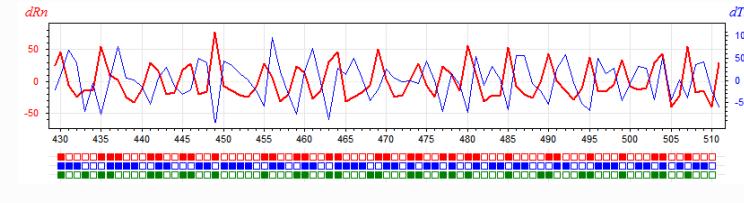
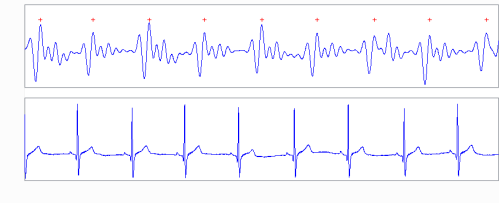
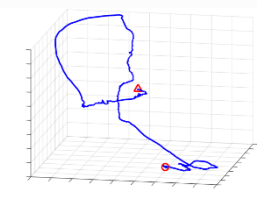
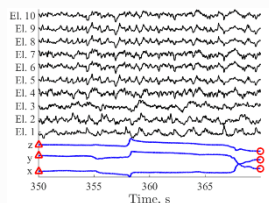
В 2008 и 2009 годах мы провели хорошие выборы по химической коммуникации. То есть а) это очень маленькие молекулы, и б) они работают в очень низкой концентрации. Это до сих пор остается и покажет, потому что сообщества зоопланктона и вообще планктона в разных экосистемах — это очень разные экосистемы рачков, которые живут в озерах и морях, взаимодействуют между собой. А между ними есть очень много биомассы, сдвиги по тому, что мы получаем в лаборатории, и разная сеть химических сигналов и коммуникаций, которые влияют на разных поведенческих, физиологических и продуктивных функций. И эти данные сеть, сеть взаимодействия до сих пор слабо исследована.



• Анализ изображений и видео



• Анализ временных рядов и интернет вещей



Topic Modeling

(тематическое моделирование текстов)

Text documents

International Journal of Networks 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. Acome, Ph.D.;
Ed.D.; Dept. of Teaching and Learning; MPH;
Dept. of Math and Computer Learning; College of Educ., Ldrshp, &
Sc.; University of North Dakota; Tech.
Forth State University; North Dakota, USA; Valdosta State University
Kansas, USA; Email: Valdosta, Georgia, USA
Email: loanyanwu@fhsu.edu; jared.keengwe@und.edu; gasrom@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 Networks 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-adjust to change in its input 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 firewalling 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].

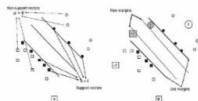
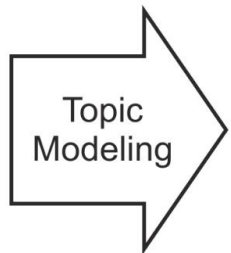
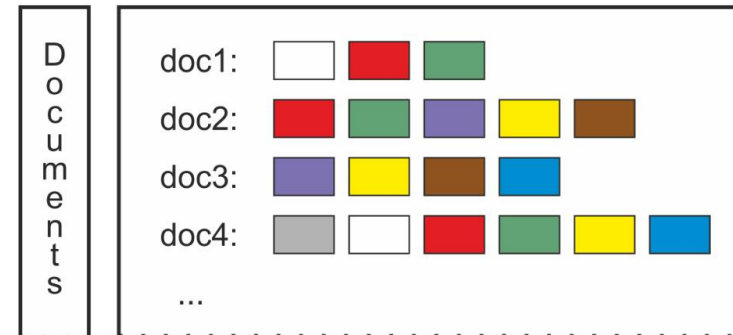


Fig. 2. Separation of the Support Vector points and Non-Support Vector points (adapted from [12])

21



Topics of documents

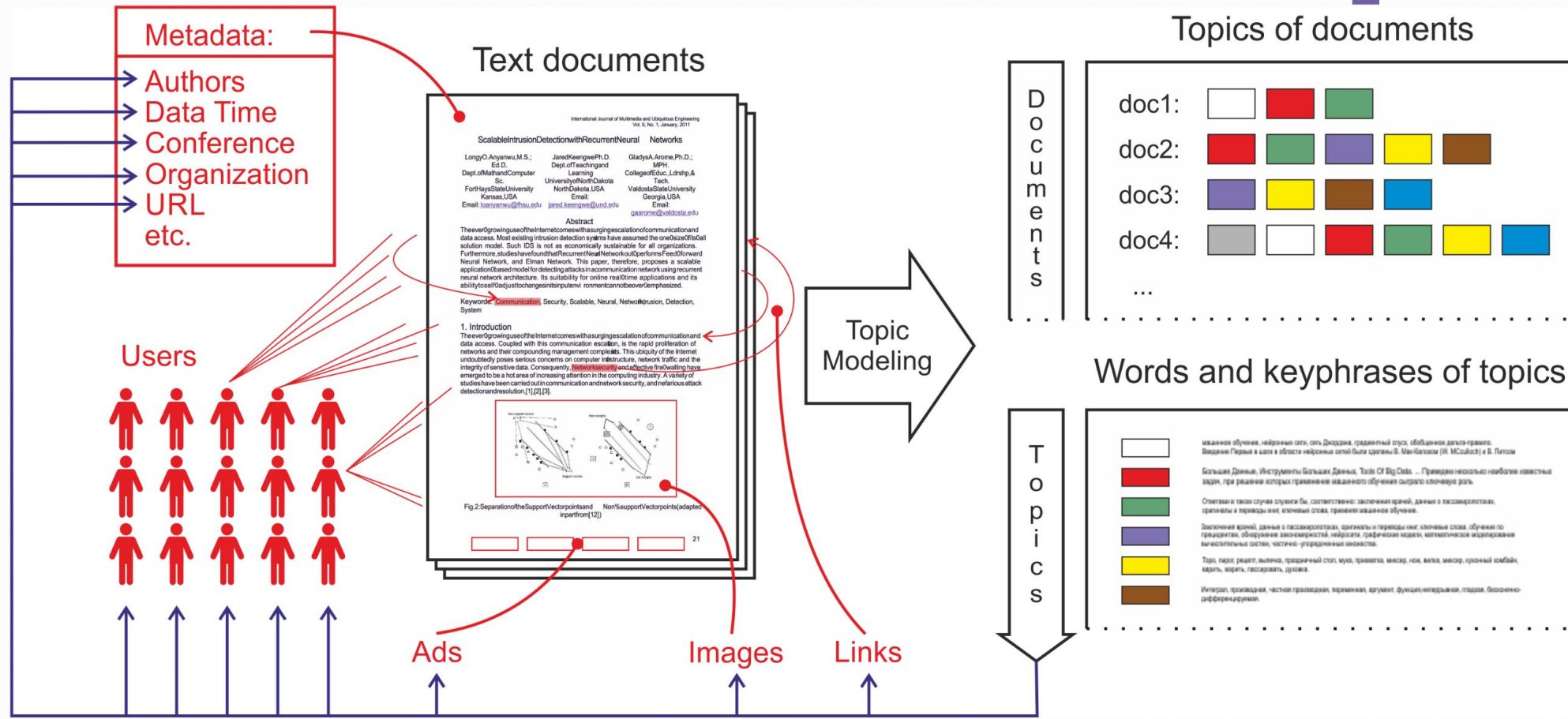


Words and keyphrases of topics



Multimodal Topic Modeling

(тематическое моделирование текстов)



Тематическое моделирование текстов и транзакционных данных

- **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>

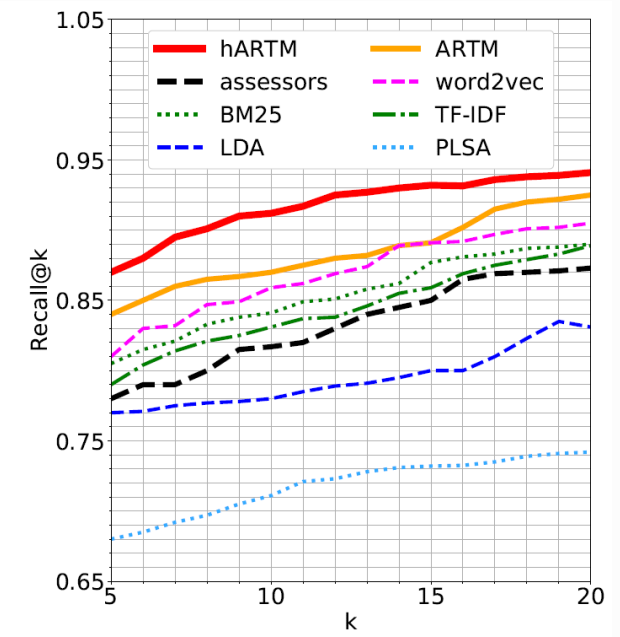
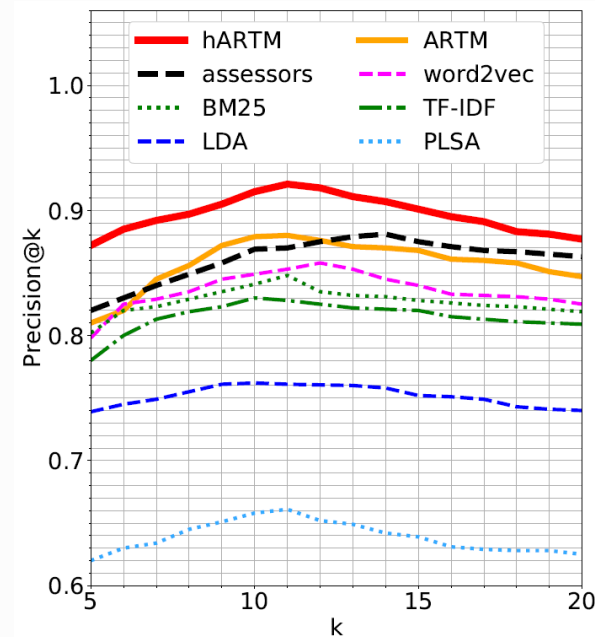
	procs	T = 50		T = 200	
		time, m	perplexity	time, m	perplexity
BigARTM	1	42	5117	83	3347
BigARTM async	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM async	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM async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

Exploratory search

Разведочный информационный 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

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



Тематизация банковских транзакционных данных



- **Транзакционные данные физических лиц:**
документ → клиент, слово → тип продавца, тема → тип потребления

Цель: формирование персональных предложений клиентам

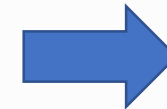
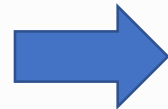
- **Транзакционные данные юридических лиц:**
документ → компания, слово → контрагент, тема → вид деятельности

Цель: отраслевой консалтинг для компаний малого бизнеса

Анализ изображений

Автоматическое распознавание сканированных документов

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

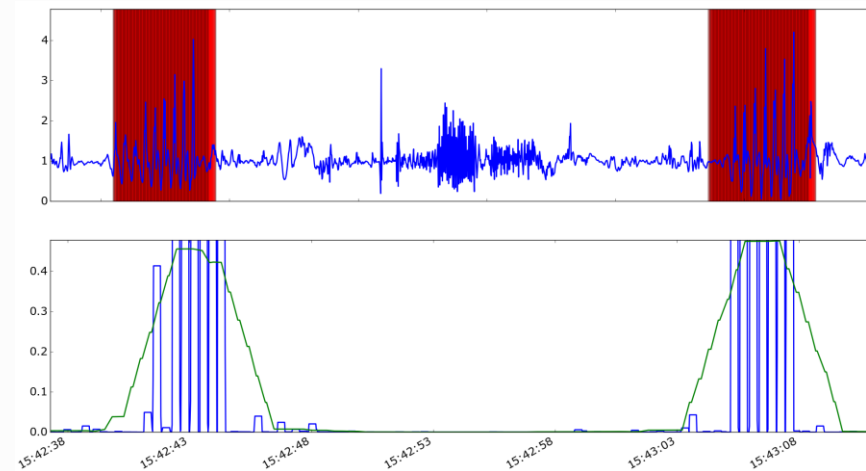


```
<elements>
  <element>
    <name>head</name>
    <type>Complex</type>
    <value>
      <element>
        <name>number</name>
        <type>Integer</type>
        <value>4201</value>
      </element>
      <element>
        <name>operator</name>
        <type>String</type>
        <value>Иванов</value>
      </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>
```

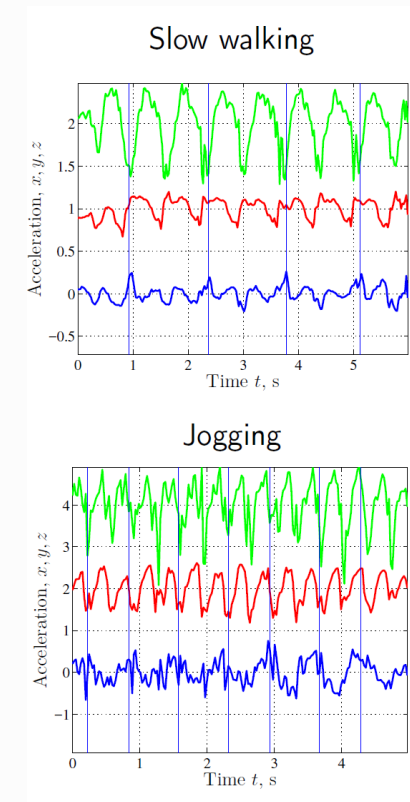


Анализ временных рядов

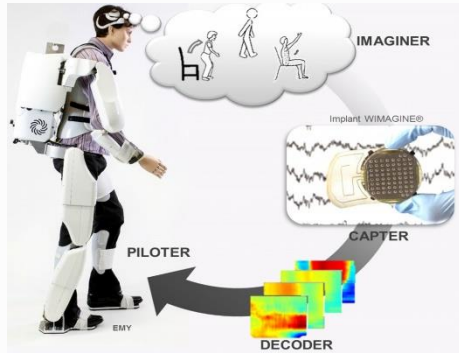
- Бизнес-приложение: мониторинг физической активности рабочих на производстве или в строительстве



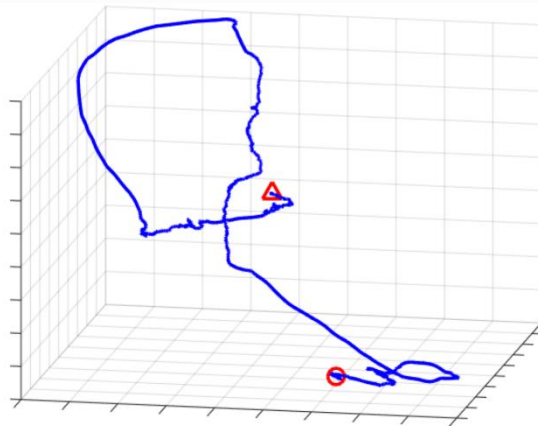
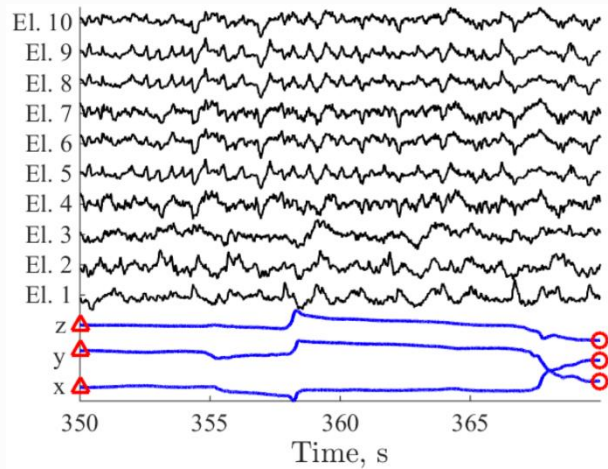
Пример: забивание гвоздя -- активность, состоящая из элементарных движений



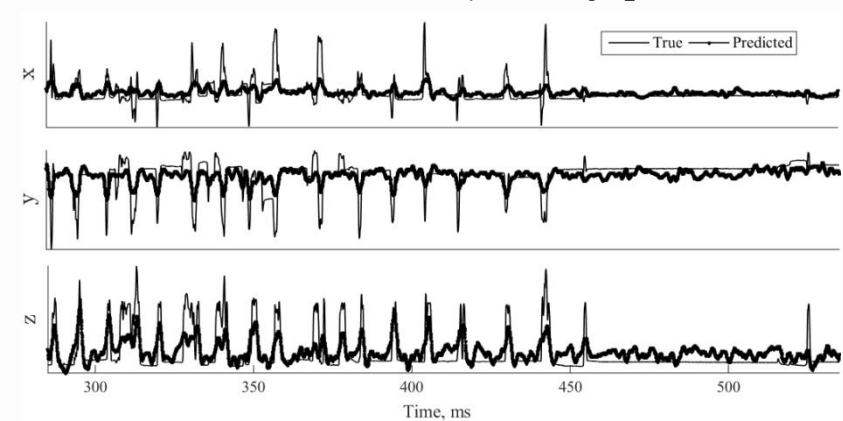
Анализ временных рядов



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



The wrist motion trajectory prediction

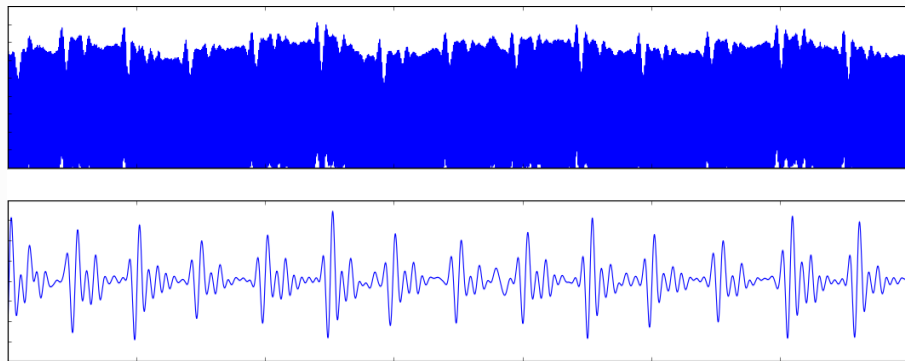


A.Motrenko, V.Strijov. Multi-way feature selection for ECoG-based BCI.
Expert Systems with Applications, 2018.

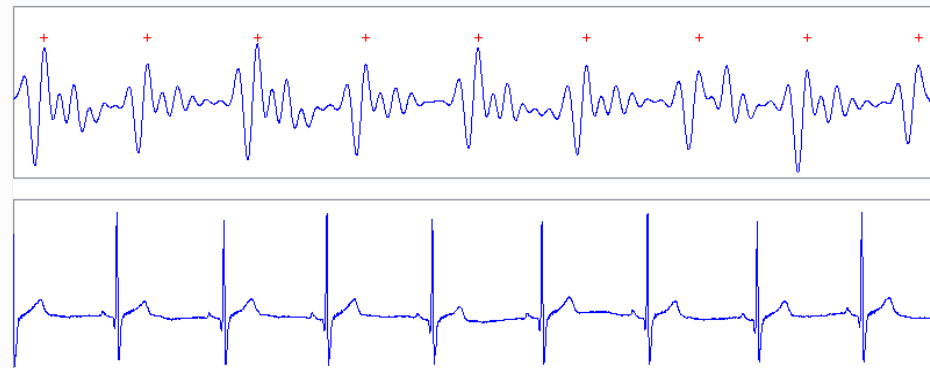
Анализ временных рядов электро- и баллистокардиография

- Мониторинг состояния больного
- Идентификация личности по 10-20-секундной БКГ
- Совместный анализ ЭКГ и БКГ, методы символьной динамики

Сырой и фильтрованный БКГ-сигнал



Сравнение БКГ и ЭКГ сигнала



Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,
руководитель лаборатории Машинного интеллекта МФТИ

k.v.vorontsov@phystech.edu

СПАСИБО
ЗА ВНИМАНИЕ

СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я - ПРОФЕССИОНАЛ

