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

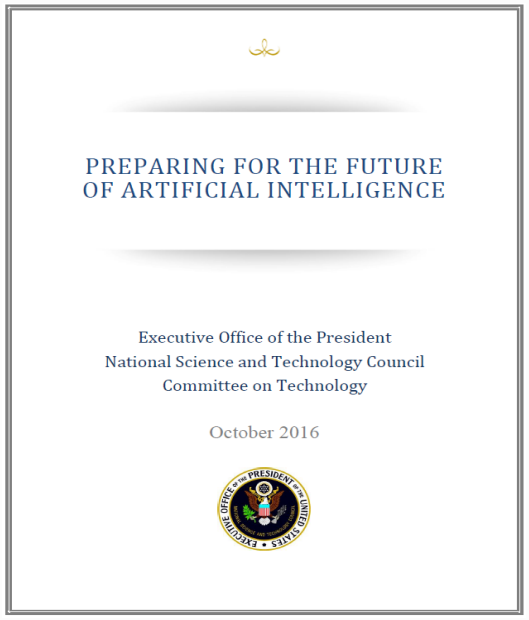
«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, *искусственном интеллекте* и *машинном обучении*» (2016)

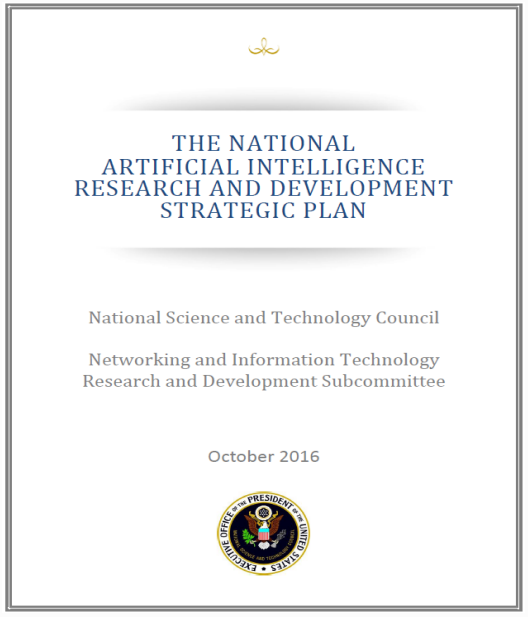
Клаус Мартин Шваб, президент Всемирного экономического форума

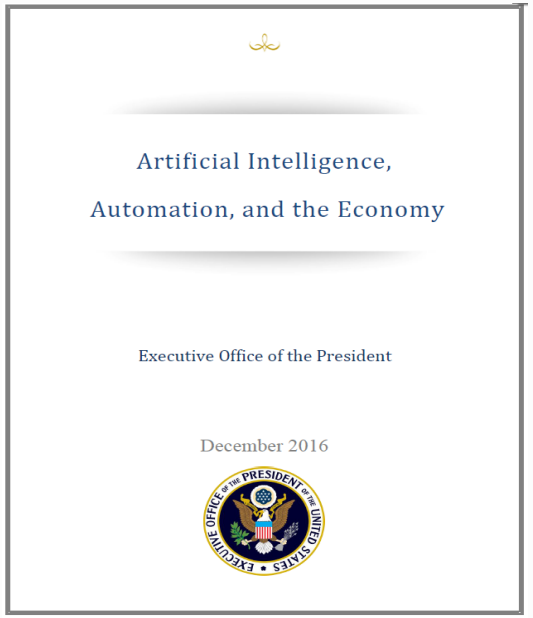
Я ПРОФИ
СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

МФТИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

Отчёты Белого дома США,
май-октябрь 2016







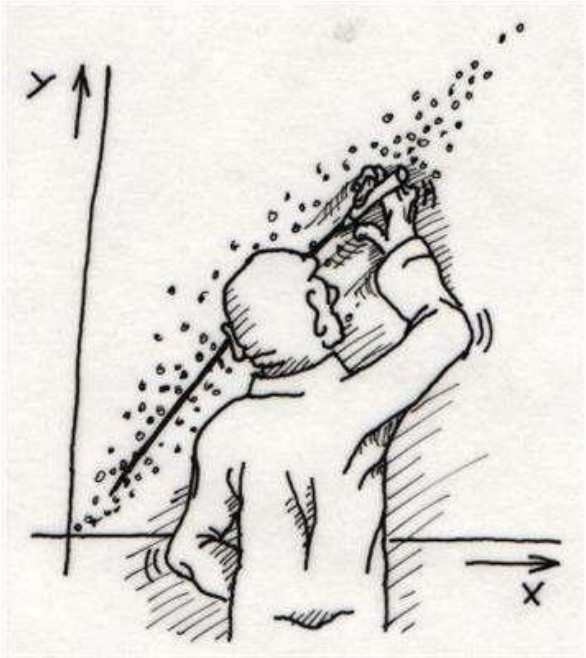
«Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future»

Я ПРОФИ
СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

МФТИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

Машинное обучение
(Machine Learning, ML)

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

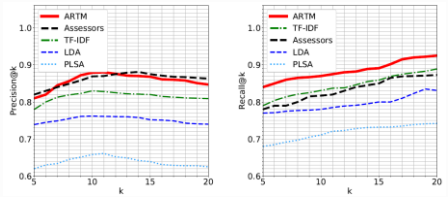


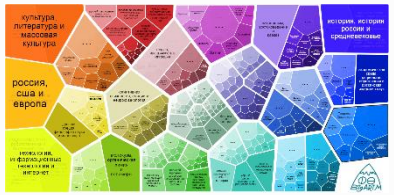
- *проведение функции через заданные точки в сложно устроенных пространствах*
- математическое моделирование в условиях, когда знаний мало, данных много
- тысячи различных методов и алгоритмов
- 100 тыс. научных публикаций в год

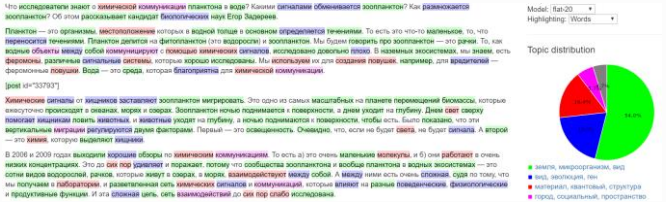
Машинный интеллект

тематика исследований лаборатории МФТИ

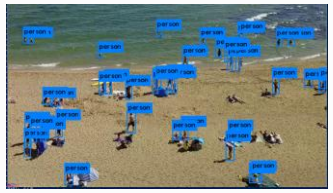
- Анализ текстов и транзакционных данных



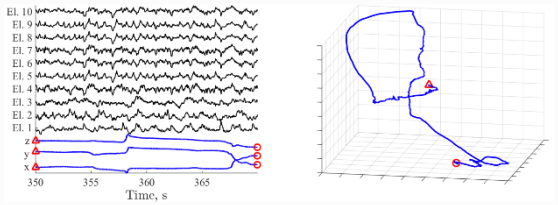


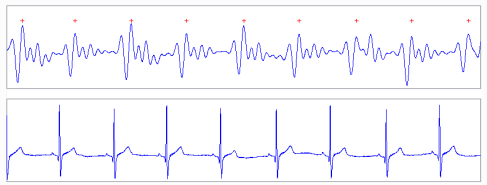


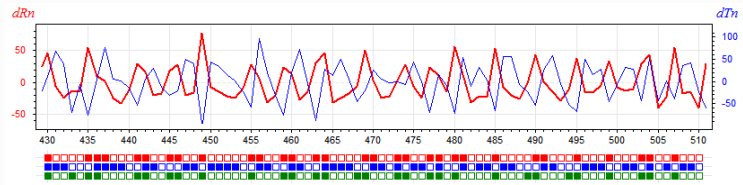
- Анализ изображений и видео



- Анализ временных рядов и интернет вещей



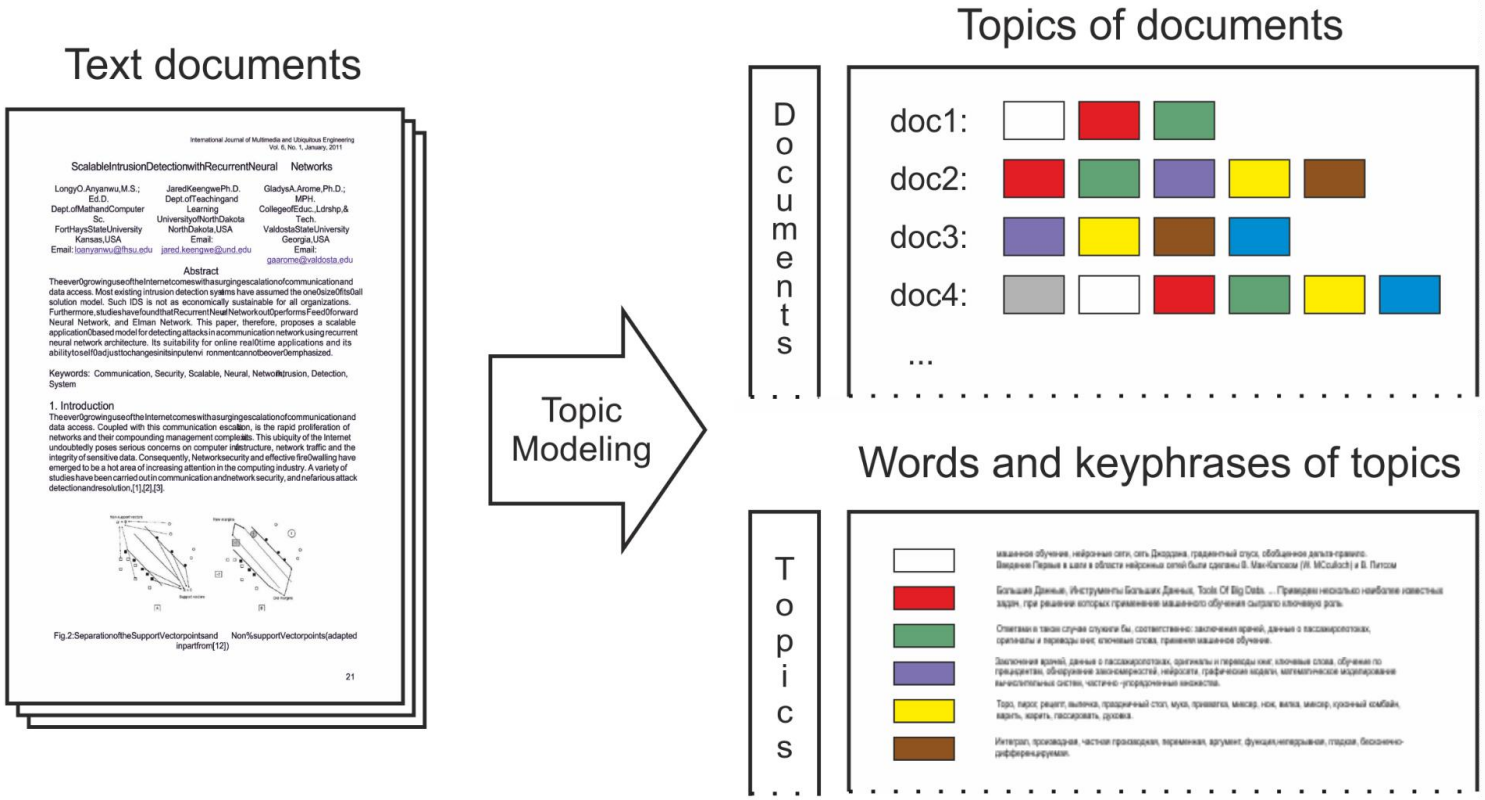




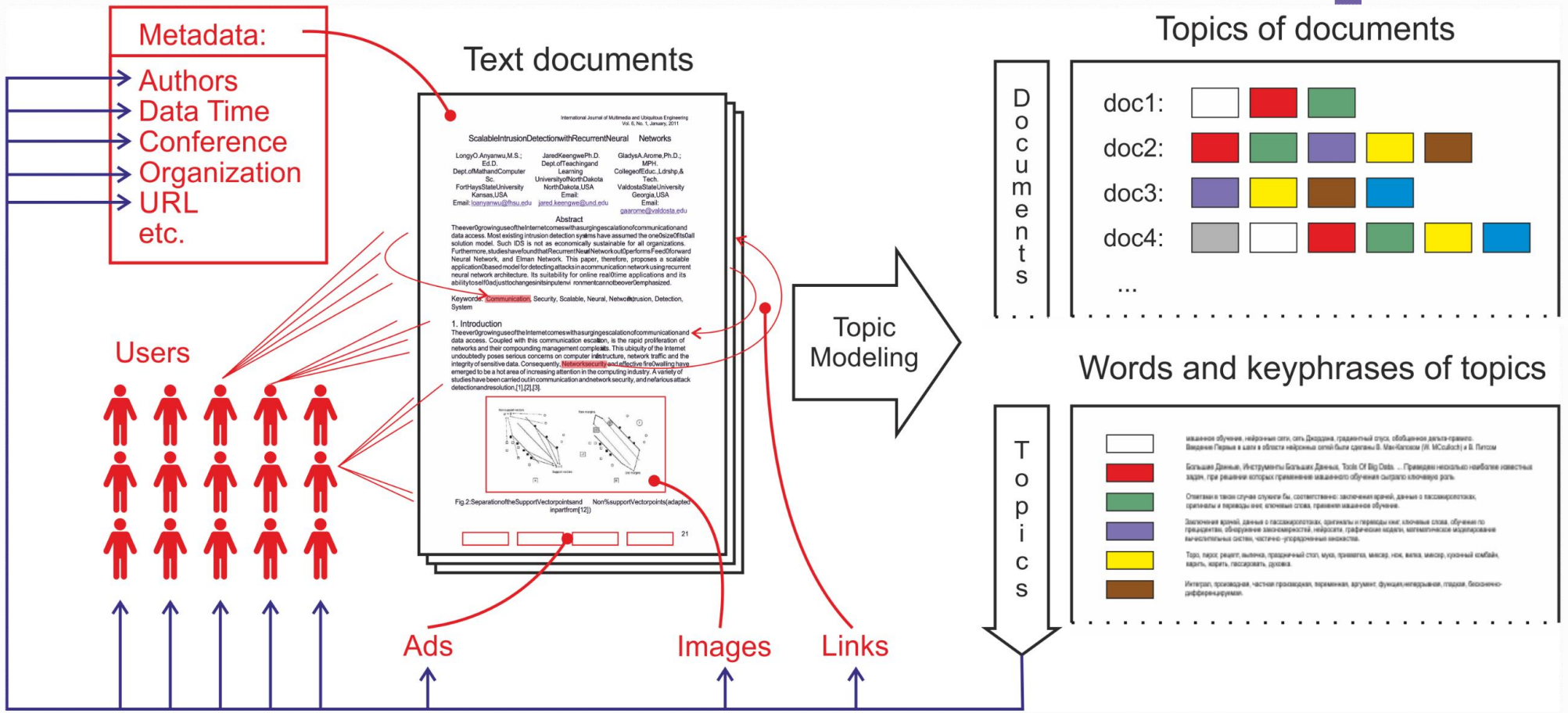
ПРОФИ
СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

МФТИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

Topic Modeling
(тематическое моделирование текстов)



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

Я ПРОФИ
СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

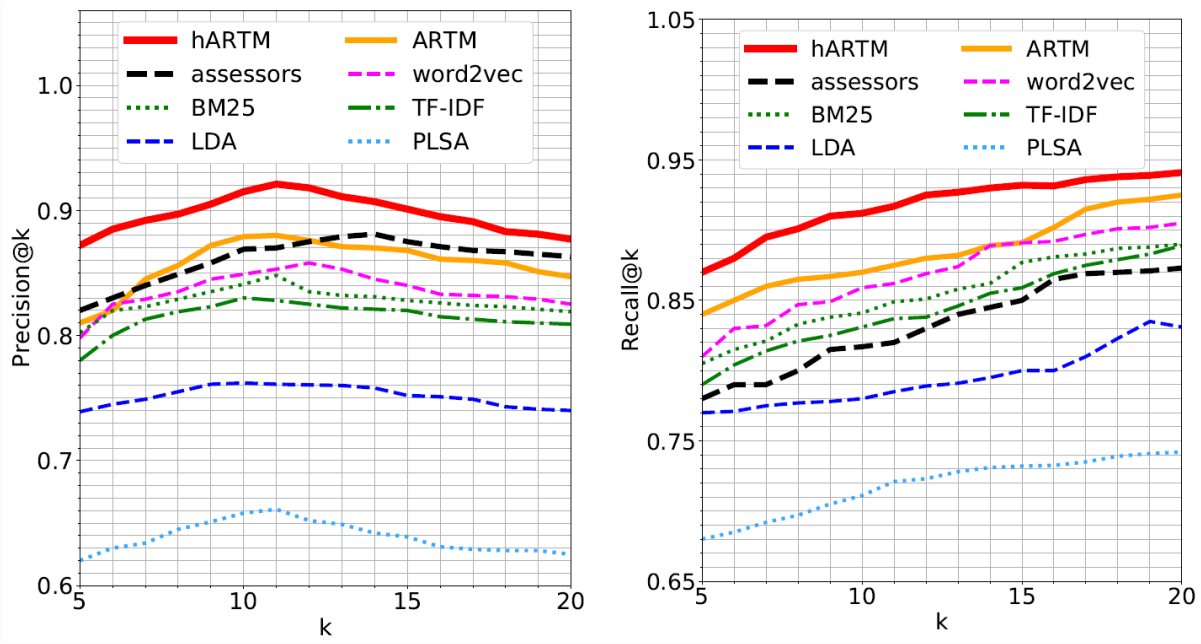
D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and Modular Regularized Topic Modelling. The 21st Conference of Open Innovations Association FRUCT: Intelligence, Social Media and Web, Finland, Helsinki, November 6-10, 2017.

МФТИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

Exploratory search
Разведочный информационный поиск

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

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



A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Тематизация банковских транзакционных данных

SBERBANK

- **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

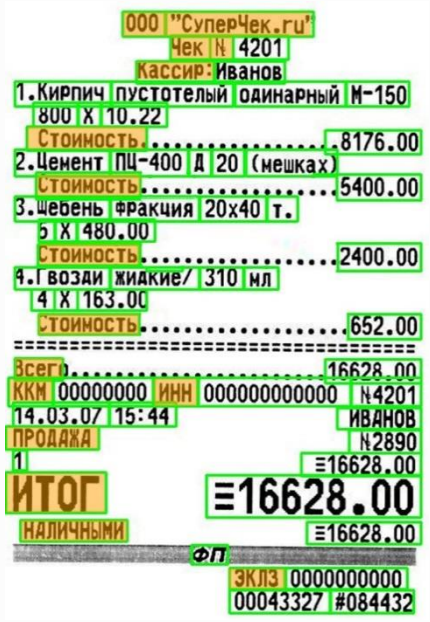
Я ПРОФИ
СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

МФТИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

Анализ изображений

Автоматическое распознавание сканированных документов

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



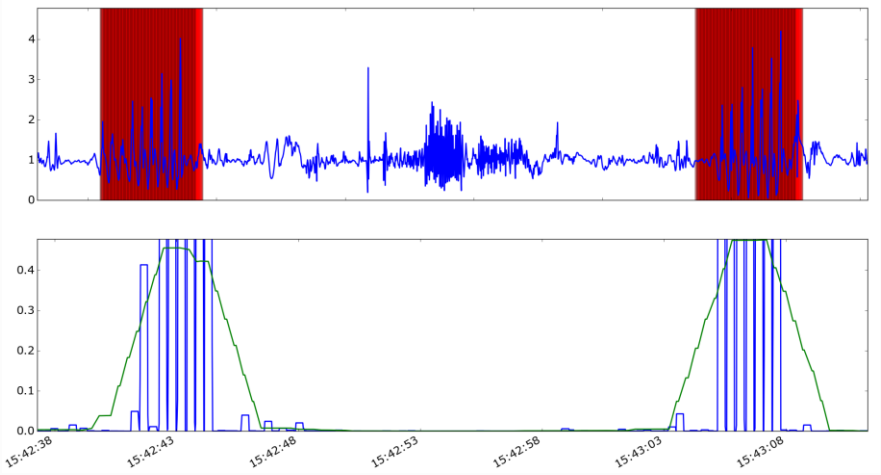

```
<elements>
    <element>
        <name>head</name>
        <type>Complex</type>
        <value>
            <element>
                <name>number</name>
                <type>Integer</type>
                <value>4201</value>
            </element>
            <element>
                <name>operator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

Я ПРОФИ
СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

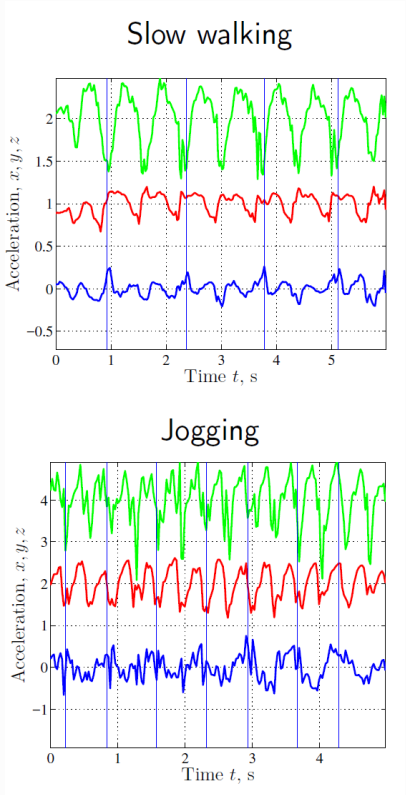
МФТИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

Анализ временных рядов

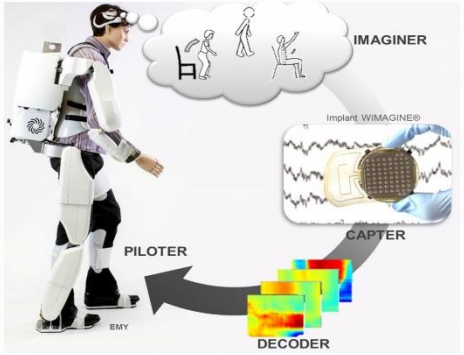
- Бизнес-приложение: мониторинг физической активности рабочих на производстве или в строительстве



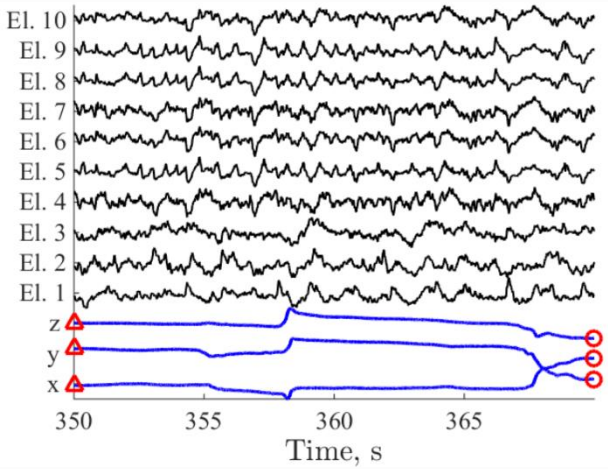
Пример: забивание гвоздя -- активность, состоящая из элементарных движений

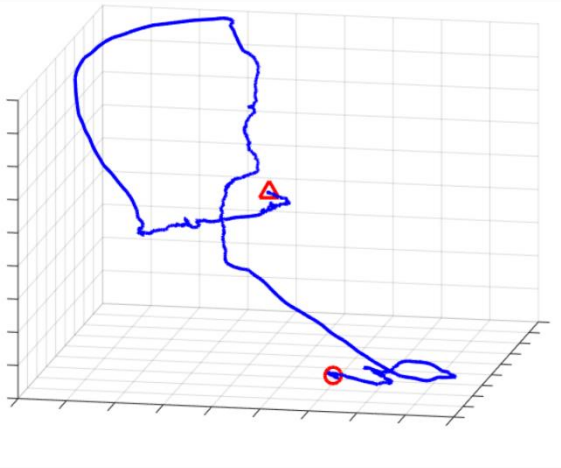


Анализ временных рядов

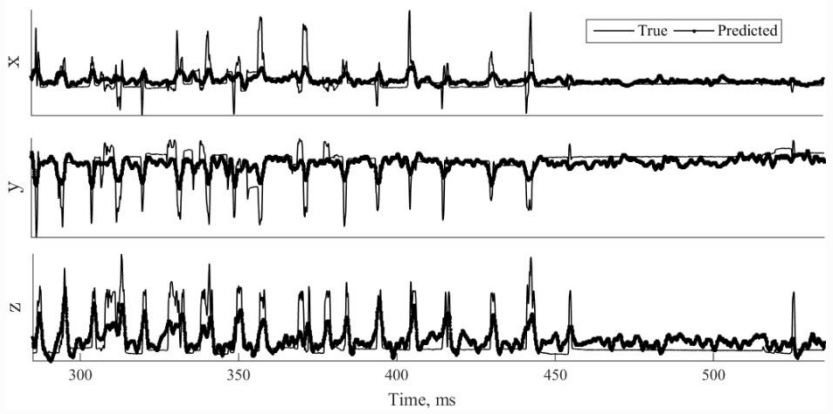


BCI-проект WIMAGINE (совместно с clinatec.fr).
Цель – создание системы компенсации нарушений двигательного аппарата человека





The wrist motion trajectory prediction

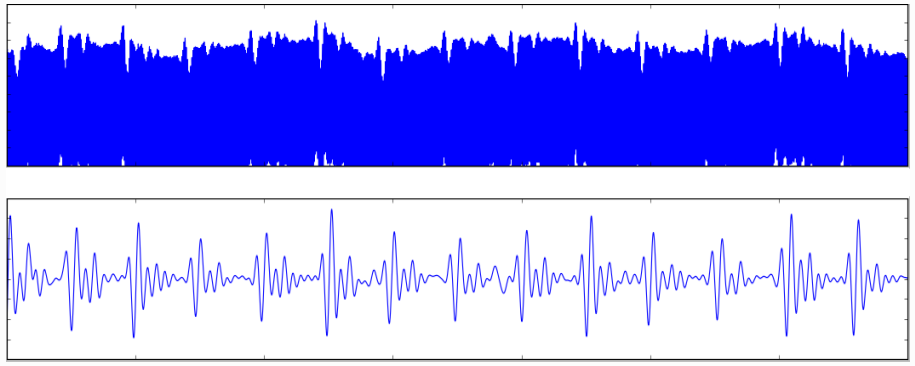


A.Motrenko, V.Strijov. Multi-way feature selection for ECoG-based BCI. Expert Systems with Applications, 2018.

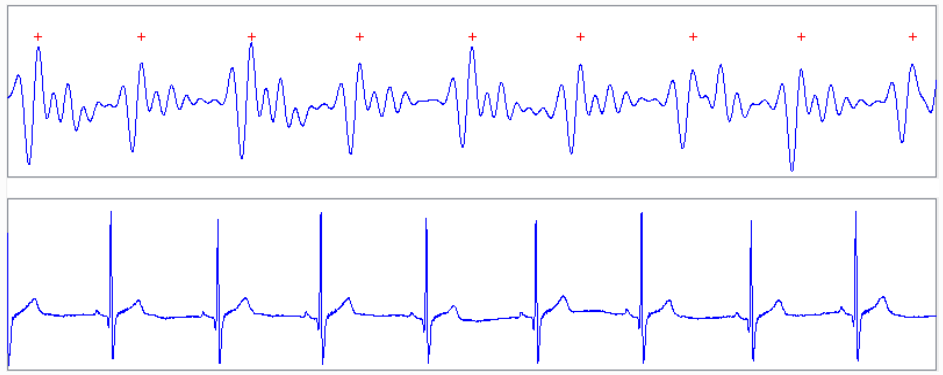
Анализ временных рядов
электро- и баллистокардиография

- 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



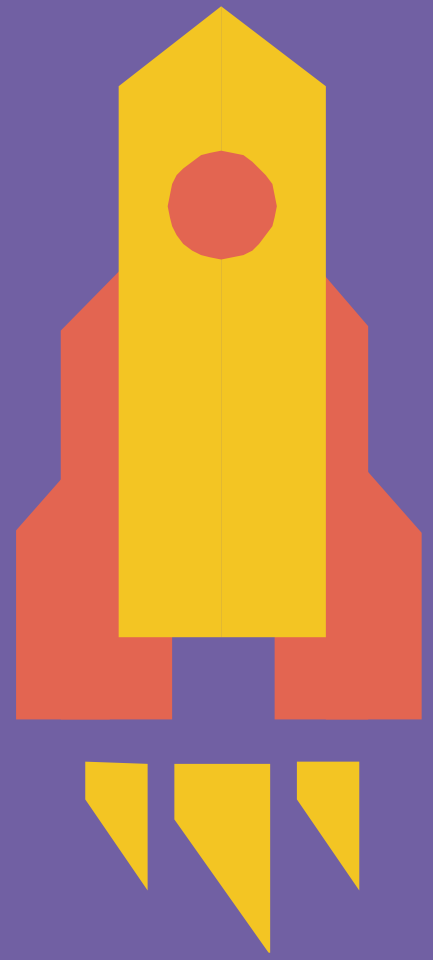
Я ПРОФИ
СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

МФТИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

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