Московский Физико-Технический Институт


# Лаборатория Машинного Интеллекта

## http://mipt.ai

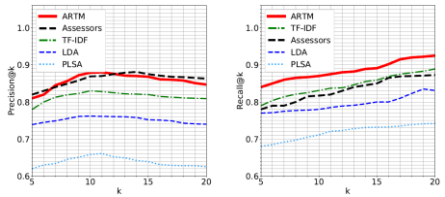
*заведующий лабораторией, д.ф.-м.н., проф.РАН*

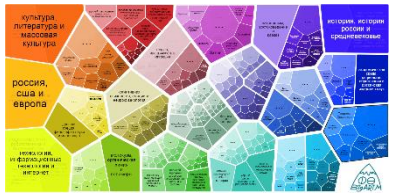
*Воронцов Константин Вячеславович*

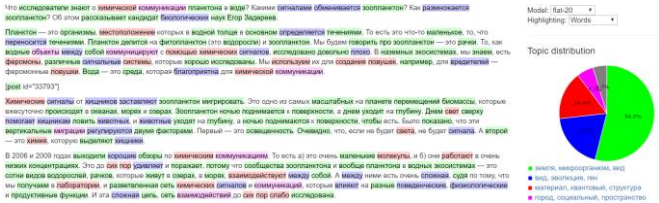
k.v.vorontsov@phystech.edu

# Направления исследований лаборатории:

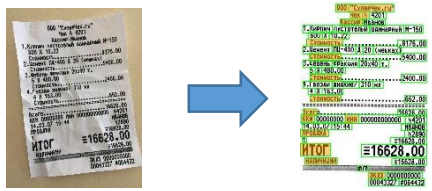
- **Анализ текстов и транзакционных данных**

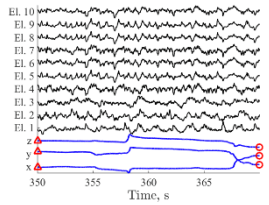



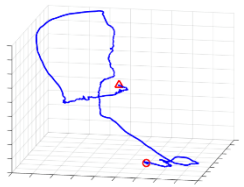

- **Анализ изображений и видео**

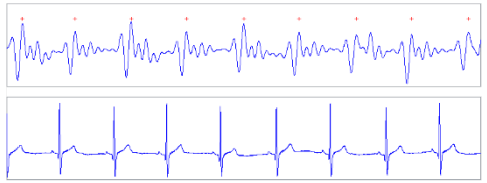


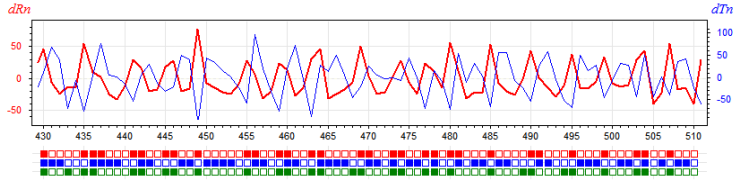


- **Анализ временных рядов и интернет вещей**

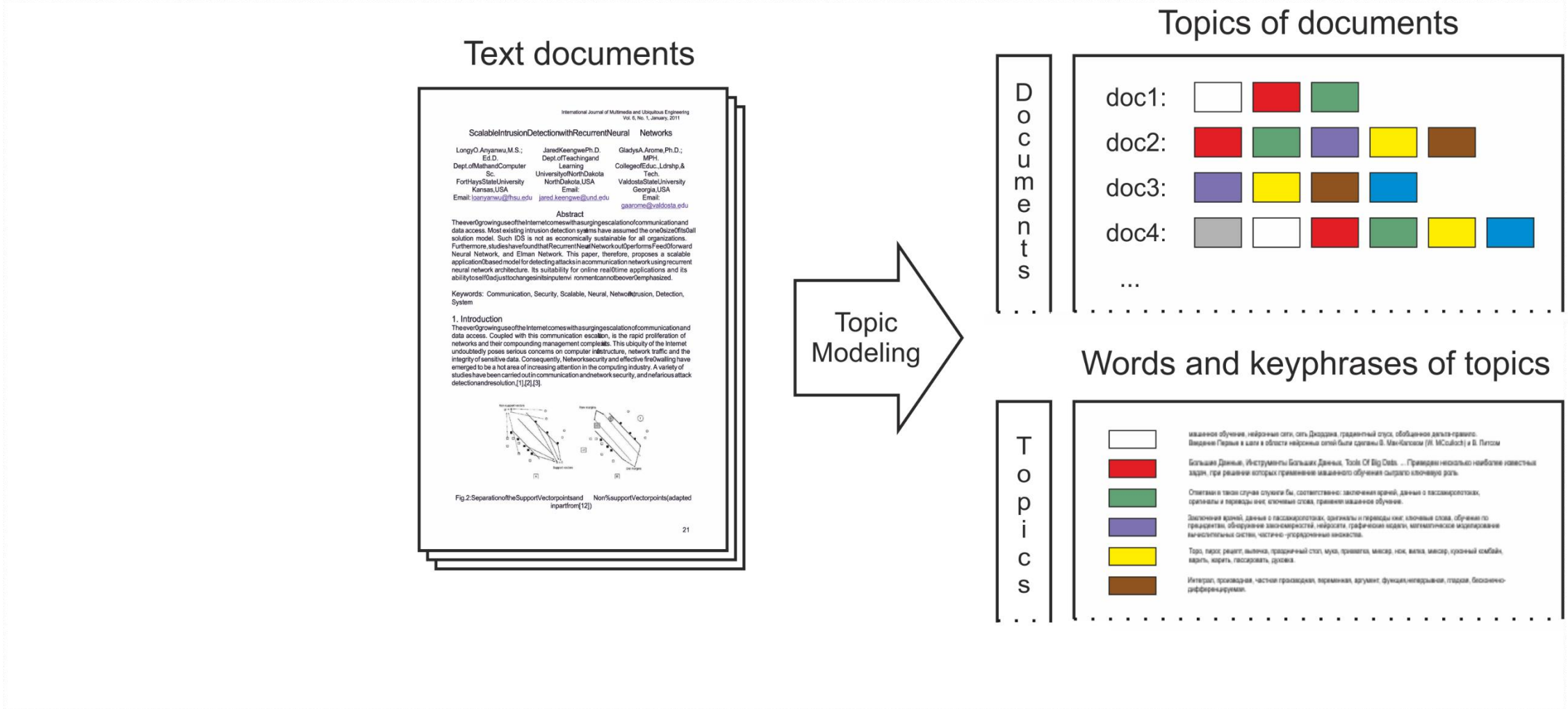


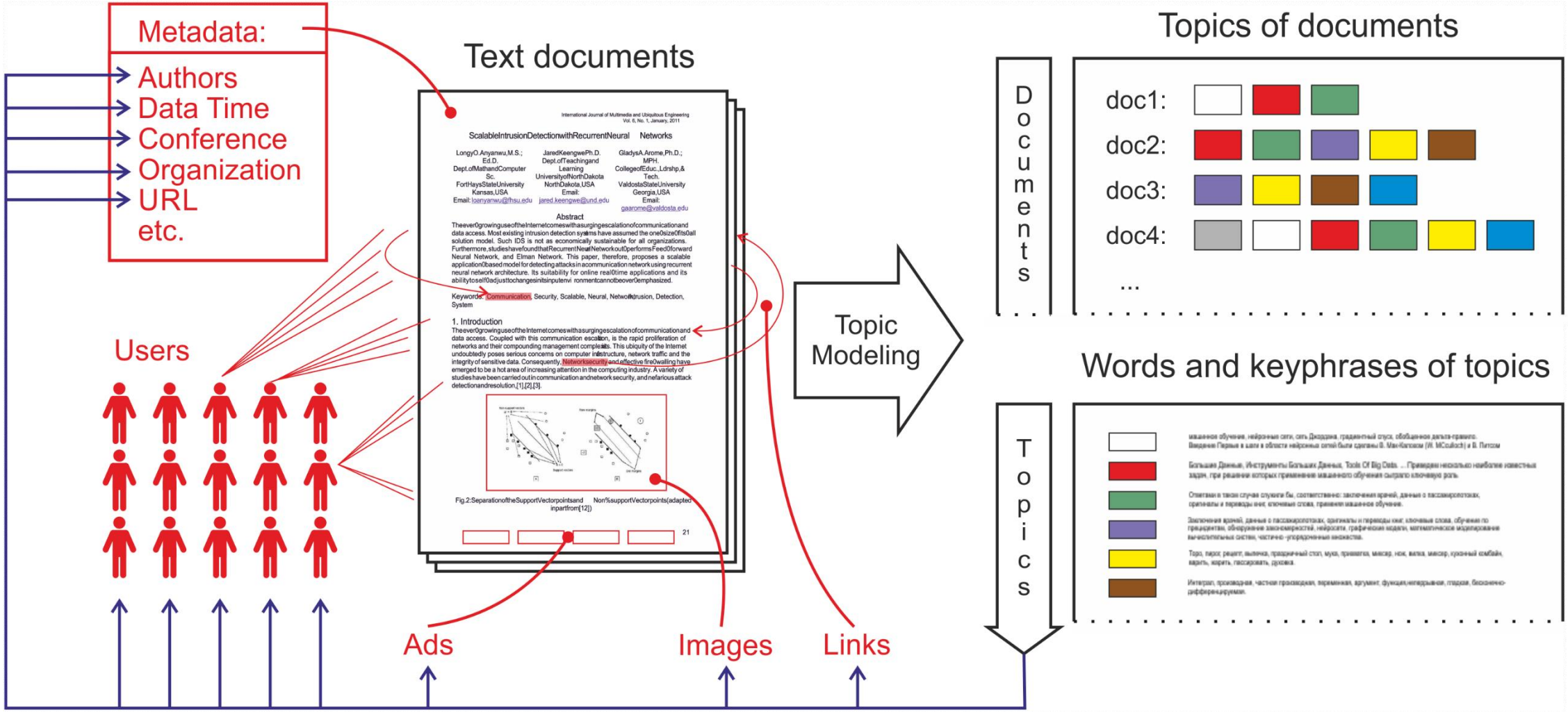




# Тематическое моделирование текстов



# 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.

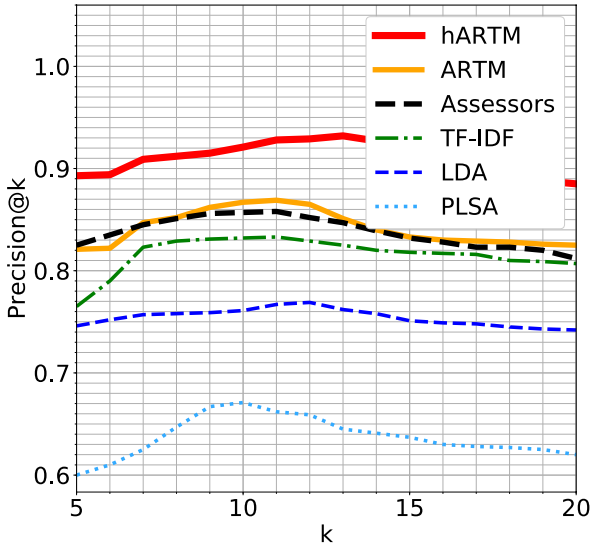
# Тематический разведочный поиск

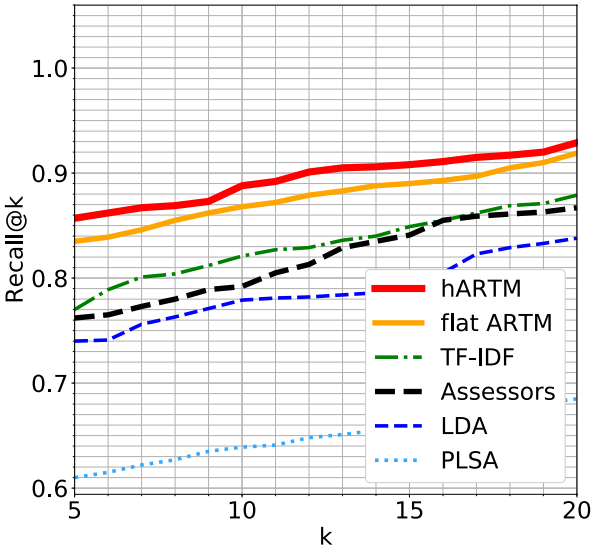
- Длинные запросы (1 стр. А4)
- 100 запросов
- 3 асессора на каждый запрос
- 30 минут в среднем на запрос
- Разметка на Яндекс.Толока
- Коллекции техно-новостей

(170K Russian docs)

(750K English docs)

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




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

# 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

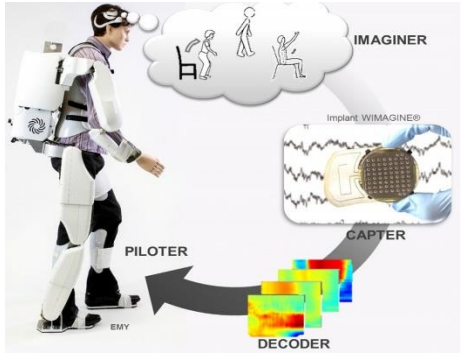
Автоматическое распознавание сканированных документов

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

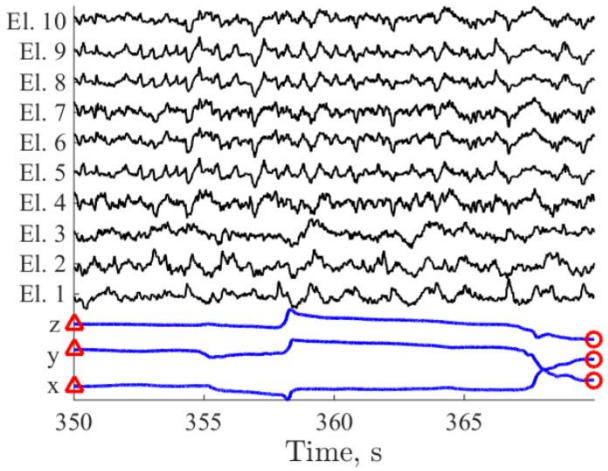


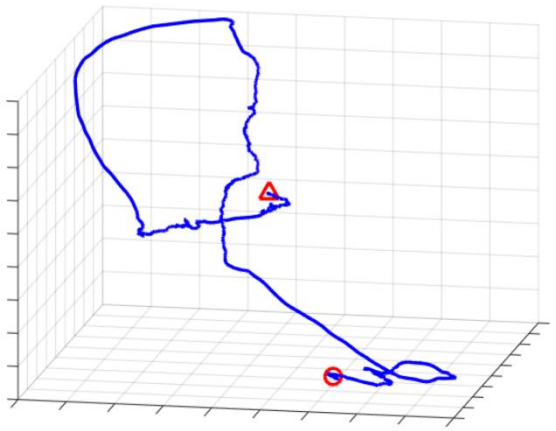
```
<elements>
    <element>
        <name>head</name>
        <type>Complex</type>
        <value>
            <element>
                <name>number</name>
                <type>Integer</type>
                <value>4201</value>
            </element>
            <element>
                <name>operator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

# Анализ временных рядов:
## распознавание движений по электрокортикограмме

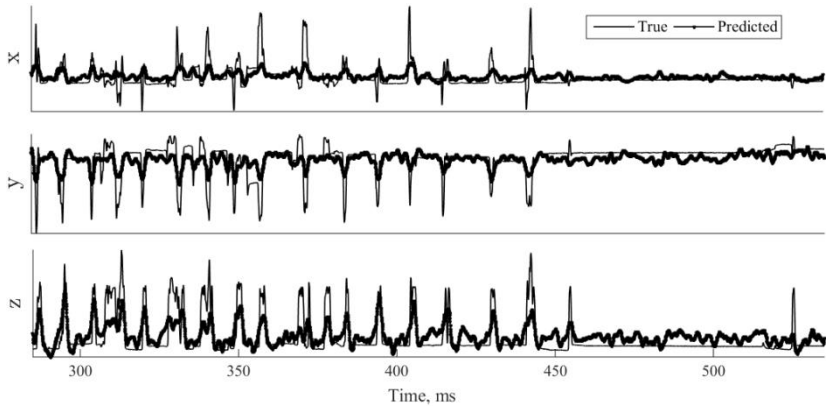



BCI-проект WIMAGINE (совместно с clinatec.fr).
**Цель** – создание системы компенсации нарушений двигательного аппарата человека



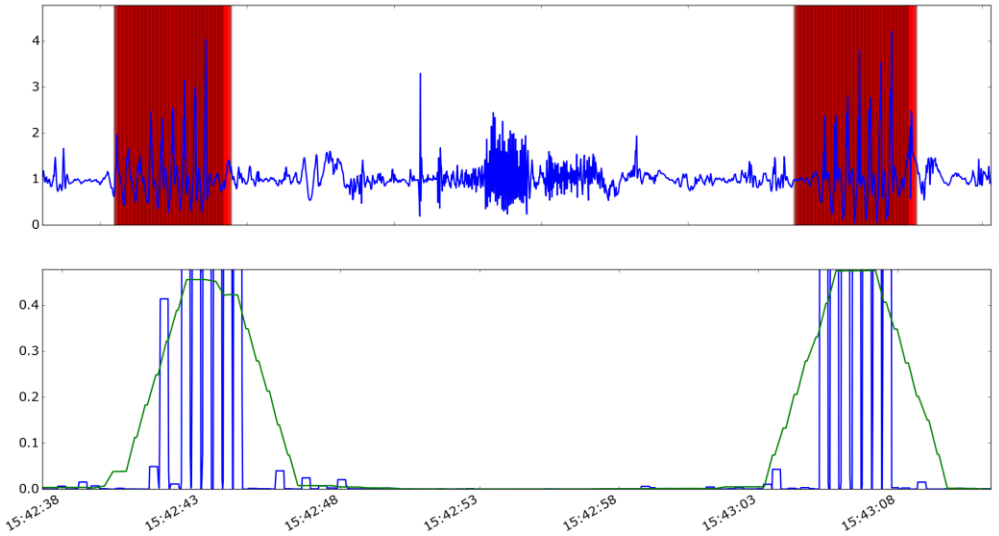
The wrist motion trajectory prediction



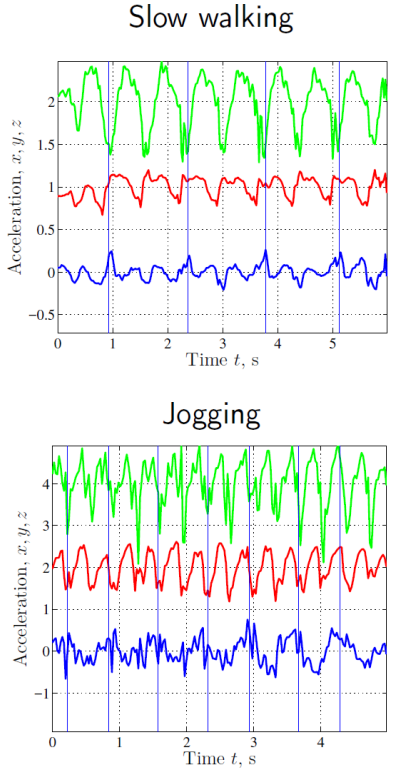

*A.Motrenko, V.Strijov.* Multi-way feature selection for ECoG-based BCI. Expert Systems with Applications, 2018.

# Анализ временных рядов:
## распознавание движений по данным акселерометров

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

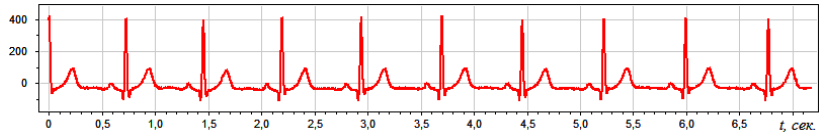



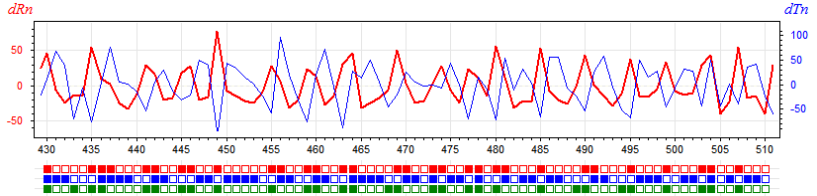

Пример: забивание гвоздя – активность, состоящая из элементарных движений



# Символьная динамика:
## оценивание рисков заболеваний по электрокардиограмме

1. Вычисление признаков кардиоциклов




2. Дискретизация и векторизация

```
DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE
```

3. Машинное обучение
4. Оценивание на тестовых данных

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

# 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