



Научный семинар «Цифровая среда»  
Институт цифровых гуманитарных исследований  
Сибирского федерального университета

## Машинное обучение и семантический анализ

- тематическое моделирование в ДН
- автоматизация контент-анализа
- «мастерская знаний»: концепция, цели, задачи

*Воронцов Константин Вячеславович*

д.ф.-м.н., профессор РАН,  
зав. кафедрой машинного обучения и цифровой гуманитаристики МФТИ,  
руководитель лаборатории машинного обучения и семантического анализа  
Института искусственного интеллекта МГУ им. М.В. Ломоносова



# Эволюция подходов в обработке естественного языка

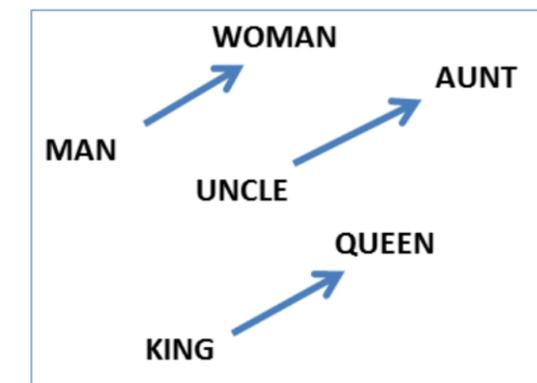
## Как решали задачи анализа текстов 10 лет назад

- морфологический анализ, лемматизация, опечатки, ...
- синтаксический анализ, выделение терминов, NER, ...
- семантический анализ, выделение фактов, тем, ...



## Модели контекстно-независимой векторизации слов

- модели дистрибутивной семантики: word2vec [Mikolov, 2013], FastText [Bojanowski, 2016], ...
- тематические модели LDA [Blei, 2003], ARTM [2014], ...



## Большие модели (LLM) контекстной векторизации слов

- рекуррентные нейронные сети: LSTM, GRU, ...
- «end-to-end» модели внимания и трансформеры: машинный перевод [2017], BERT [2018], GPT-4 [2023], ...

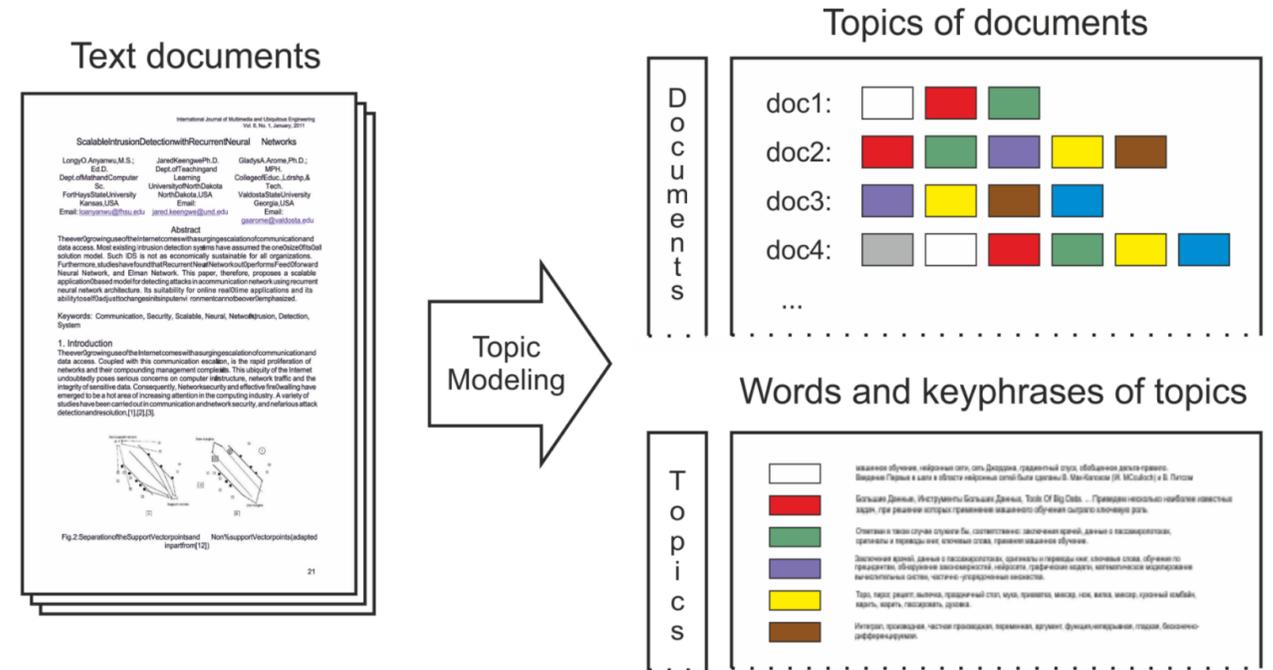
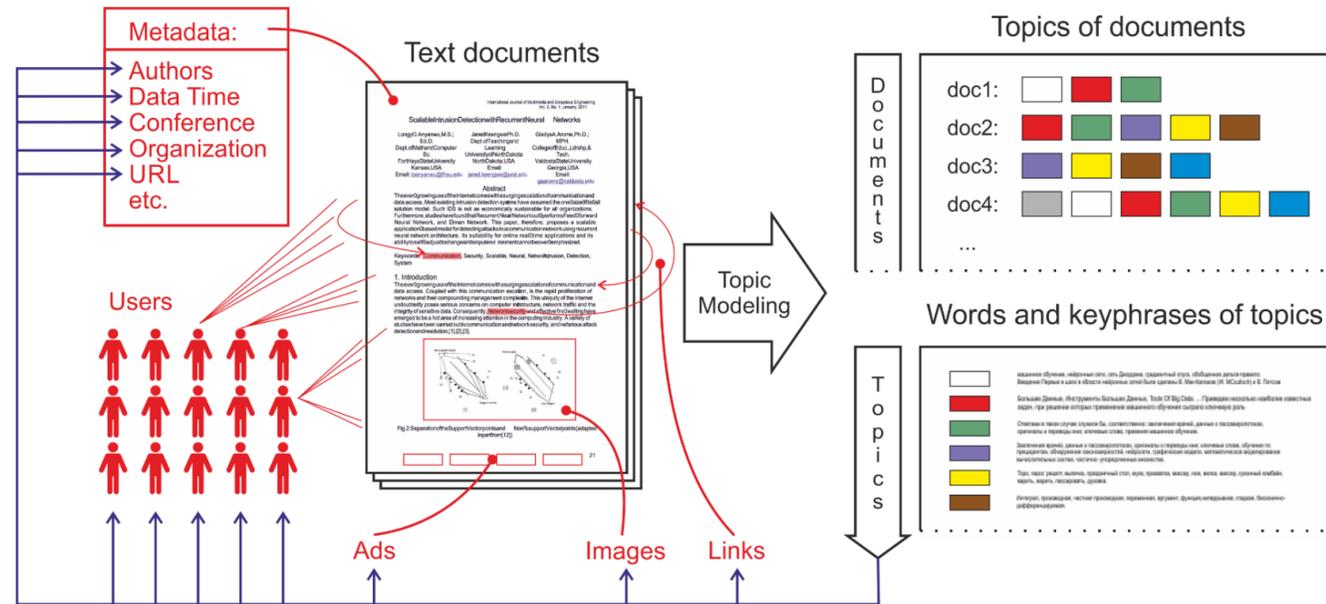
$$\text{softmax} \left( \frac{\begin{matrix} \mathbf{Q} & \mathbf{K}^T \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} & \times & \begin{matrix} \square & \square \\ \square & \square \end{matrix} \end{matrix}}{\sqrt{d}} \right) \mathbf{V}$$

The diagram shows a matrix multiplication of a 2x3 matrix (Q) and a 2x2 matrix (K<sup>T</sup>), followed by a division by the square root of d (√d). The result is a 2x2 matrix (V) passed through a softmax function.

# Тематическое моделирование

## Тематическая модель (ТМ) коллекции текстовых документов определяет

- какие темы есть в каждом документе
- из каких слов состоит каждая тема



## Мультимодальная ТМ определяет тематику токенов разных модальностей:

- авторы, время, n-граммы, теги, ссылки,...
- языки в мультязычной коллекции

Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URSS. 2025. 224с.

<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

# Что даёт тематическое моделирование

для каждой темы — *качественный анализ*:

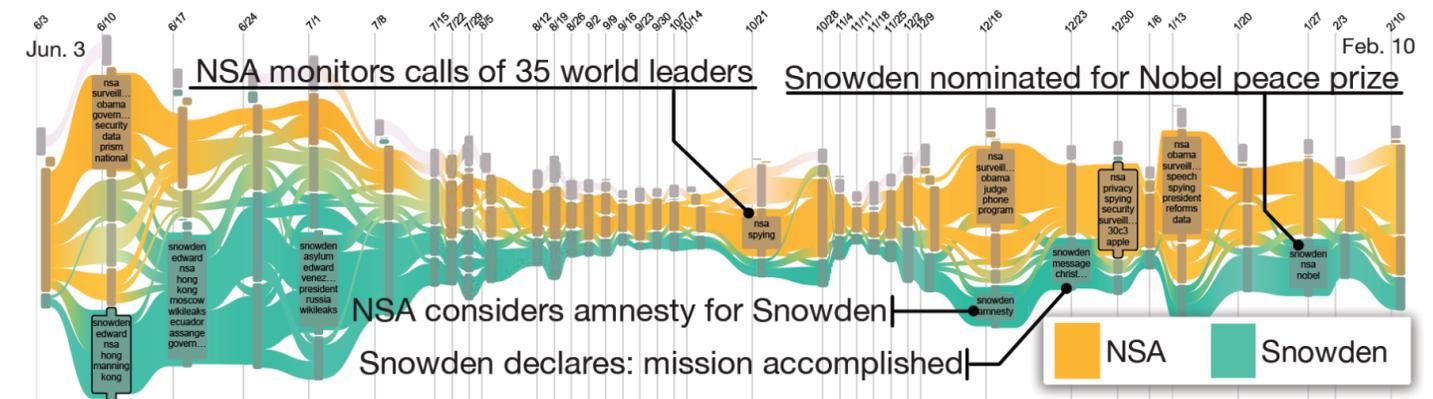
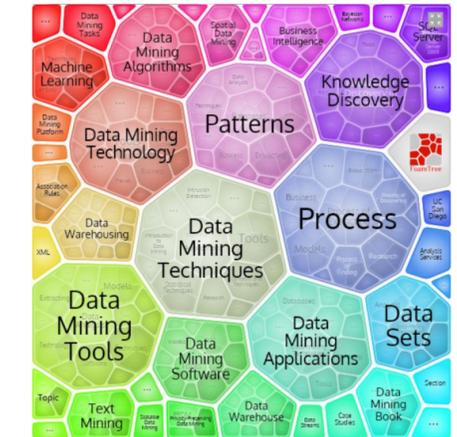
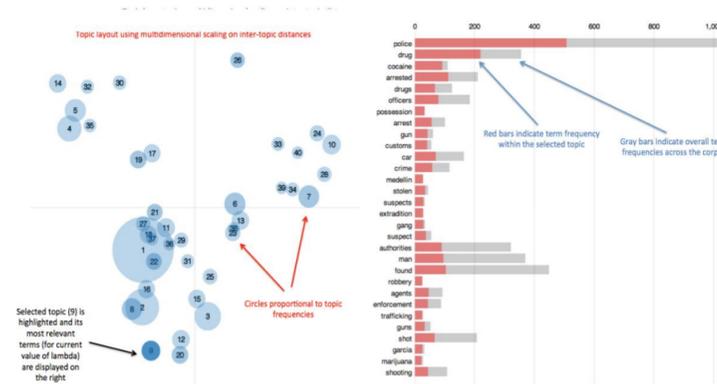
- слова, n-граммы, термины, фразы
- «рассказ о себе»: название, суммаризация

для каждого документа:

- состав тем
- тематическая сегментация

для коллекции в целом:

- *визуализации*: динамика, иерархия, спектр, карта
- *количественный анализ* в разрезе времени, источников, авторов, тегов, геолокаций, языков и т.д.



Воронцов К.В. Вероятностное тематическое моделирование: Теория регуляризации ARTM и библиотека с открытым кодом BigARTM. — Москва: URSS. — 2025. — 224 с..  
<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

# ARTM — Аддитивная Регуляризация ТМ

Математическая теория, позволяющая оптимизировать сумму критериев (**регуляризаторов**) для построения моделей с заданными свойствами.

Проще, гибче, технологичнее, чем байесовские модели на основе LDA.

**Дано:** коллекция текстовых документов как «мешков-слов»

- $n_{dw}$  — частота слова (терма)  $w \in W$  в документе  $d \in D$
- $|T|$  — сколько тем хотим определить в коллекции  $D$

**Найти:** тематическую языковую модель (в.п.  $D \times W \times T$ )

- $p(w|d) = \sum_{t \in T} p(w | \cancel{d}, t) p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$
- $p(w|t) = \phi_{wt}$  — из каких слов  $w$  состоит каждая тема  $t \in T$
- $p(t|d) = \theta_{td}$  — из каких тем  $t$  состоит каждый документ  $d$

**Критерий:** правдоподобие предсказания слов  $w$  в документах  $d$

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

Максимизация логарифма правдоподобия **с регуляризатором:**

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$$

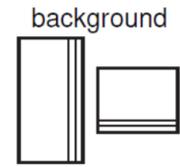
EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); & n_{td} = \sum_{w \in d} n_{dw} p_{tdw} \end{cases} \end{cases}$$

Воронцов К.В. Вероятностное тематическое моделирование: Теория регуляризации ARTM и библиотека с открытым кодом BigARTM. — Москва: URSS. — 2025. — 224 с..

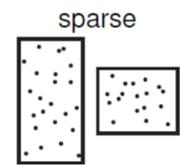
<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

# Примеры регуляризаторов



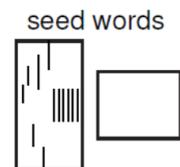
Сглаживание фоновых тем  $B \subset T$ :

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$

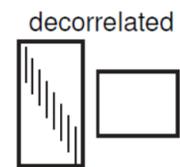


Разреживание предметных тем  $S = T \setminus B$ :

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \phi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$

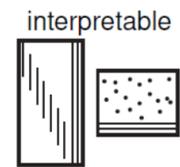


Сглаживание для выделения релевантных тем с помощью словаря «затравочных» ключевых слов

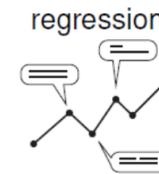


Декоррелирование для повышения различности тем:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем



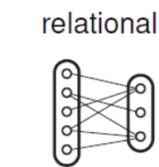
Линейная модель регрессии  $\hat{y}_d = \langle v, \theta_d \rangle$  документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left( y_d - \sum_{t \in T} v_t \theta_{td} \right)^2$$



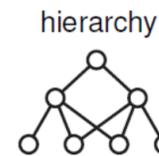
Связи сочетаемости слов ( $n_{uv}$  — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$



Связи или ссылки между документами:

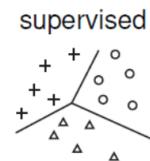
$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc}$$



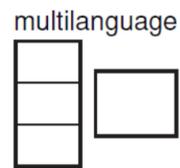
Связи родительских тем  $t$  с дочерними подтемами  $s$ :

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}$$

# Примеры регуляризаторов

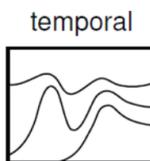


Модальности меток классов или категорий для задач классификации и категоризации текстов.



Модальность языков и регуляризация со словарём  $\pi_{uwt} = p(u|w, t)$  переводов с языка  $k$  на  $\ell$ :

$$R(\Phi, \Pi) = \tau \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in W^\ell} \pi_{uwt} \phi_{wt}$$



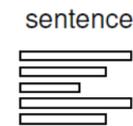
Темпоральные модели с модальностью времени  $i$ :

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} |\phi_{it} - \phi_{i-1,t}|.$$

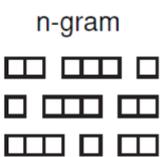


Модальность геолокаций  $g$  с близостью  $S_{gg'}$ :

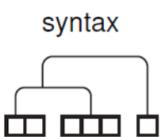
$$R(\Phi) = -\frac{\tau}{2} \sum_{g, g' \in G} S_{gg'} \sum_{t \in T} n_t^2 \left( \frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}} \right)^2$$



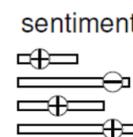
Тематические модели, учитывающие границы предложений, абзацев и секций документов



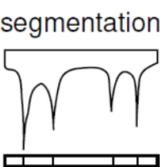
Модели с модальностями  $n$ -грамм, коллокаций, именованных сущностей (используем TopMine)



Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)



Модели выделения мнений на основе тональностей, фактов, семантических ролей именованных сущностей



Тематические модели сегментации с автоматическим определением границ сегментов

# BigARTM — технология тематического моделирования

## Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайн-параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и мер качества

## Сообщество:

- Открытый код <https://github.com/bigartm>  
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



## Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

3.7М статей Википедии, 100К слов: время min (перплексия)

проц.	T	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

Мурат Апишев. Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией. Диссертация к.т.н., ФИЦ ИУ РАН. 2020.

Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A. BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST'2015

# Разведочный поиск в технологических блогах

**Цель:** поиск документов

по длинным текстовым запросам

— Habr.ru (175K документов),

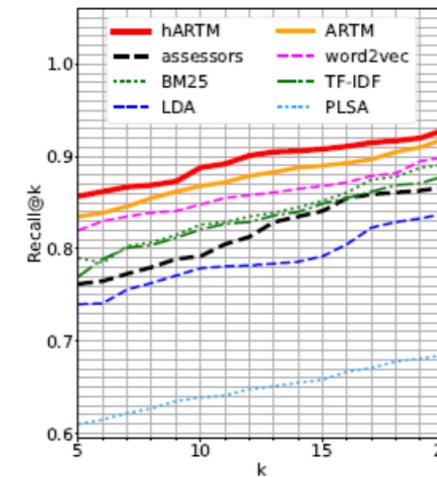
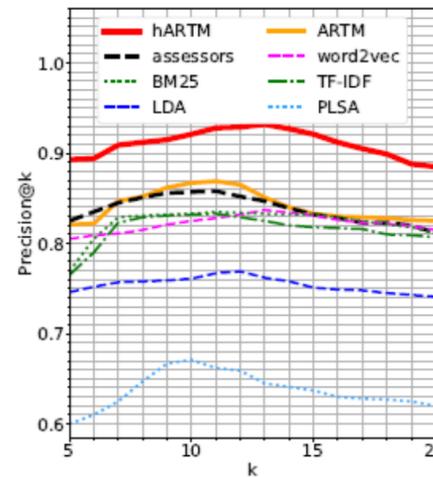
— TechCrunch.com (760K док.).

**Регуляризаторы:**

$$\mathcal{L} \left( \begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left( \begin{array}{c} \text{hierarchy} \\ \text{graph} \end{array} \right) + R \left( \begin{array}{c} \text{interpretable} \\ \text{bar chart} \quad \text{scatter plot} \end{array} \right) + R \left( \begin{array}{c} \text{multimodal} \\ \text{stacked bar} \quad \text{square} \end{array} \right) + R \left( \begin{array}{c} \text{n-gram} \\ \text{grid of boxes} \end{array} \right) \rightarrow \max$$

**Результаты:**

- Точность и полнота **93%**, превосходит ассессоров и другие методы (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- Увеличилась оптимальная размерность векторов:  
200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).



# «Поиск и классификация иголок в стоге сена»

**Цель:** выявление как можно большего числа тем о национальностях и межнациональных отношениях (затравка — словарь 300 этнонимов).

**Регуляризаторы:**

$$\mathcal{L} \left( \begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left( \begin{array}{c} \text{seed words} \\ \text{[bar chart]} \end{array} \right) + R \left( \begin{array}{c} \text{interpretable} \\ \text{[scatter plot]} \end{array} \right) + R \left( \begin{array}{c} \text{multimodal} \\ \text{[stacked bars]} \end{array} \right) \\ + R \left( \begin{array}{c} \text{temporal} \\ \text{[line graph]} \end{array} \right) + R \left( \begin{array}{c} \text{geospatial} \\ \text{[map]} \end{array} \right) + R \left( \begin{array}{c} \text{sentiment} \\ \text{[sentiment scale]} \end{array} \right) \rightarrow \max$$

**Результаты:** число релевантных тем: 45 (LDA) → 83 (ARTM).

*M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, K. Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.*

–, –, –, –, –. Mining ethnic content online with additively regularized topic models. 2016.

(японцы): японский, япония, корей, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,  
 (норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,  
 (венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадура, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,  
 (китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,  
 (азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,  
 (грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,  
 (осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,  
 (цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

османский  
 восточноевропейский  
 эвенк  
 швейцарская  
 аланский  
 саамский  
 латыш  
 литовец  
 цыганка  
 ханты-мансийский  
 карачаевский  
 кубинка  
 гагаузский  
 русич  
 сингапурец  
 перуанский  
 словенский  
 вепсский  
 ниггер  
 адыги  
 сомалиец  
 абхаз  
 темнокожий  
 нигериец  
 лягушатник  
 камбоджиец



# Поиск поляризации мнений в политических новостях

**Цель:** найти признаки, по которым событийная тема разделяется на кластеры-мнения

Modalities	<i>Pr</i>	<i>Rec</i>	<i>F1</i>
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
<b>All</b>	<b>0.77</b>	<b>0.97</b>	<b>0.86</b>

**Регуляризаторы:**

$$\mathcal{L} \left( \begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left( \begin{array}{c} \text{interpretable} \\ \text{[Bar Chart]} \quad \text{[Scatter Plot]} \end{array} \right) + R \left( \begin{array}{c} \text{multimodal} \\ \text{[Image]} \quad \text{[Text]} \end{array} \right) + R \left( \begin{array}{c} \text{n-gram} \\ \text{[Grid]} \end{array} \right) + R \left( \begin{array}{c} \text{syntax} \\ \text{[Tree]} \end{array} \right) \rightarrow \max$$

**Результаты:**

- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей:
  - SPO — факты как триплеты «субъект–предикат–объект»
  - FR — семантические роли слов по Филлмору
  - Sent — тональности именованных существностей

... Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР "национализировали" украинские предприятия ... При этом Кремль защитил конфискацию предприятий в ЛДНР ... Украина потребует расширить санкции ... За все эти действия обязательно наступит наказание. Украина потребует расширения санкций на тех, кто украл украинские предприятия ... (Kiev opinion)

... По словам Захарченко, Киев встретит свой "ужасный хвост" ... Киев возьмется за ум, и в целях спасения собственной промышленности снимет блокаду ... Обстановка, которую искусственно создала Украина с блокадой Донбасса, вынудила ... кошмарит свой народ ... если в Киеве были приняты какое-либо постановление ... положительные результаты, как в республиках, так и в России ... Если им удастся сместить Порошенко и при этом не развалить Украину, то все вернется на свои места ... (Moscow opinion)

Subject   Object   Agent   Locative   Negative lexicon   Dependent word

Слова «Порошенко», «Россия», «Украина» встречаются в тексте-1 и тексте-2 одинаково часто, однако:

- «Порошенко» — субъект в тексте-1 и объект в тексте-2;
- «Россия» — агент в тексте-1 и локация в тексте-2;
- негативная тональность: «Россия», «Кремль» в тексте-1, «Киев», «Украина» в тексте-2.

# Примеры тематического моделирования в исторических исследованиях (газетные архивы)

## [1] Корпус *Pennsylvania Gazette* 1728--1800, 25М слов:

- выделение последовательности событийных тем, изучение синхронности событий
- комбинирование автоматического анализа и ручного.

## [2] Газеты Техаса от гражданской войны до наших дней:

- выделение динамики всех тем, связанных с хлопком;

## [3] Газеты и периодика Финляндии (1854--1917):

- выделение тем о церкви, религии, образовании;
- тренды модернизации и секуляризации финского общества.

---

1. *D.Newman, S.Block*. Probabilistic topic decomposition of an eighteenth-century American newspaper. 2006.

2. *Tze-I Yang, A.J.Torget, R.Mihalcea*. Topic modeling on historical newspapers. 2011.

3. *J.Marjanen et al*. Topic modelling discourse dynamics in historical newspapers.

# Примеры тематического моделирования в исторических исследованиях (летописи и дневники)

## [4] Двужычный корпус книг на английском и немецком языке:

--- все темы, связанные с эпистемологией

## [5] Корпус текстов на китайском языке (1644--1912):

--- все темы, связанные с бандитизмом, преступлениями;

--- анализ контекста для установления типа преступления.

## [6] Дневник Martha Ballard (1735–1812), охватывает 27 лет:

--- выделение тем событийных и перманентных, персональных и исторических

--- специфичный английский XVIII века.

---

4. *M.Erlin*. Topic modeling, epistemology, and the English and German novel. 2017.

5. *Ian Matthew Miller*. Rebellion, crime and violence in Qing China, 2013.

6. *Cameron Blevins*. <http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary>.

# Примеры тематического моделирования в исторических исследованиях (журнальная периодика)

Статьи коллекции JSTOR доступны в виде «мешков слов».

## [7] Научные журналы XX века:

- различия тематики на английском и немецком языках;
- особенно исследовались различия, связанные со 2МВ;
- для объединения тем использовались интервики Википедии.

## [8] Более 100 лет литературно-художественной периодики:

- как менялись темы;
- как менялись значения слов внутри каждой темы;
- как менялась тема насилия (violence, power, fear, blood, death, murder, act, guilt).

---

7. *D.Mimno*. Computational historiography: Data mining in a century of classics journals. 2012.

8. *A.Goldstone, T.Underwood*. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. 2014.

# Примеры тематического моделирования в политологии (анализ публичных выступлений)

## [9] Выступления (210К) в Европарламенте, 1999--2014:

- выявление событийных тем и эволюции перманентных тем;
- как члены и комитеты ЕП влияют на формирование тем

## [10] Модель контрастных мнений (Contrastive Opinion Modeling):

- выступления в Сенате США ([www.votesmart.org](http://www.votesmart.org));
- СМИ: New York Times, Xinhua News, The Hindu, 2009—2010

## [11] Выступления в Совбезе ООН по Афганистану, 2001--2017:

- динамика отношения разных стран к проблемам Афганистана

---

9. *D.Greene, J.P.Cross*. Unveiling the political agenda of the European Parliament plenary: a topical analysis. 2015.

10. *Fang Y. et al.* Mining contrastive opinions on political texts using cross-perspective topic model. 2012.

11. *M.Schonfeld*. Discursive landscapes and unsupervised topic modeling in IR: a validation of text-as-data approaches through a new corpus of UN Security Council speeches on Afghanistan. 2018.

# Примеры тематического моделирования в политологии (анализ СМИ и социальных медиа)

## [12] Тематика изменения климата в СМИ Пакистана, 2010--2021

--- выявление, группирование и динамика тем

## [13] Выявление поляризации новостей (AULIEN COVID-19):

--- 1,5М новостей, 440 источников СМИ, 11.2019--07.2020

## [14] Выявление политических взглядов пользователей Twitter

## [15] Что пишет NYT о ядерных технологиях с 1945 по н/в

---

12. *W.Ejaz et al.* Politics triumphs: A topic modeling approach for analyzing news media coverage of climate change in Pakistan. 2023

13. *Zihao He.* Detecting polarized topics using partisanship-aware contextualized topic embeddings. 2021

14. *R.Cohen, D.Ruths.* Classifying Political Orientation on Twitter: It's Not Easy! 2013.

15. *C.Jacobi.* Quantitative analysis of large amounts of journalistic texts using topic modelling. 2015.

16. *H.Jelodar et al.* Latent Dirichlet allocation ({LDA}) and topic modeling: models, applications, a survey. 2019.

# Подытожим. Тематическое моделирование для цифровых гуманитарных исследований

## Почему ТМ не теряет актуальности в эпоху больших языковых моделей:

- полнота тематической кластеризации текстовой коллекции
- скорость и дешевизна вычислений
- интерпретируемость тем

## Открытые проблемы и создание нового «стандарта де-факто» в ТМ:

- переход от гипотезы «мешка слов» к модели тематического внимания
- как гарантировать 100% интерпретируемости тем?
- связаны ли «мусорные темы» и темы-дубликаты с дисбалансом тем?
- когда тема «рассказывает о себе», как измерить качество этого рассказа?
- как автоматизировать выбор числа тем и коэффициентов регуляризации?

## Машинное обучение и семантический анализ

- тематическое моделирование в ДН
- автоматизация контент-анализа
- «мастерская знаний»: концепция, цели, задачи

# Конкурс ПРО//ЧТЕНИЕ

(<http://ai.upgreat.one>)



**Задача:** автоматическая разметка смысловых ошибок в сочинениях ЕГЭ по русскому яз., литературе, истории, обществознанию, английскому яз.

**Период:** декабрь 2019 — июнь 2022, три цикла испытаний.

**Призовой фонд:** ₹100М русский язык + ₹100М английский язык

**Типов ошибок:** 152

(р:70 л:16 о:23 и:20 а:23)

**Подтипов ошибок:** 236

(р:112 л:19 о:29 и:26 а:50)

Помимо выделения ошибок, надо давать их объяснения.

## ФАКТИЧЕСКАЯ ОШИБКА

автор высказывания А.Франц

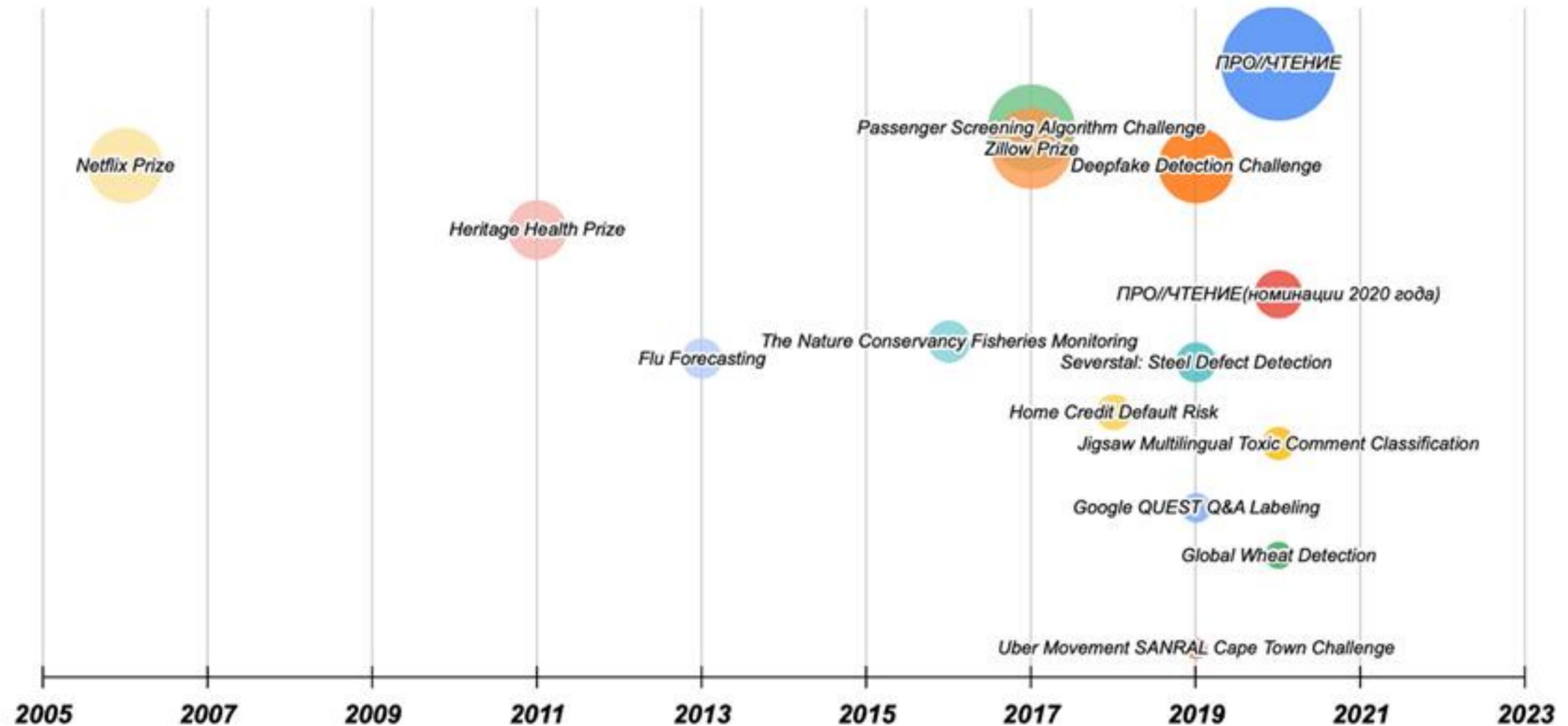
В своем высказывании «Если человек зависит от природы, то и она от него зависит» Д. Мережковский говорит о необходимости защиты природы.

## ЛОГИЧЕСКАЯ ОШИБКА

тезис не обоснован



# Конкурсы анализа данных — драйверы цифрового развития отраслей



# Конкурсы SemEval по детекции пропаганды

Базовая разметка: «фрагмент, метка класса»



Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum **omnium fortissimi** sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea **quae ad effeminandos animos pertinent important**, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa **Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt**, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language

Attack on Reputation: Smears

Manipulative Wording: Exaggeration

Justification: Appeal to Values



Commissio  
PopulusQue  
Europaea

Упрощённая разметка: «предложение, метка класса»

Обобщённая разметка: «фрагмент→мишень, метка класса»

- SemEval-2023 task 3. Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup. <https://propaganda.math.unipd.it/semEval2023task3>
- *G.Martino, P.Nakov et al.* A survey on computational propaganda detection. 2020.
- *F.Alam, P.Nakov et al.* Overview of the WANLP 2022 shared task on propaganda detection in Arabic. 2022.

# Контент-анализ: обобщение и автоматизация

**Обобщённый контент-анализ** — четыре базовые операции с текстом:

- 1) выделить фрагмент
- 2) классифицировать (тегировать) фрагмент по рубрикатору
- 3) связать несколько фрагментов
- 4) дать комментарий (затекст) к фрагменту или связи

**Цель** — автоматизировать контент-анализ больших текстовых массивов по небольшим размеченным корпусам, в любой предметной области

**Три задачи** построения обучаемой модели разметки:

- 1) разработка рубрикатора, инструкций разметчика, организация разметки
- 2) выбор большой языковой модели и её (до)обучение по разметке
- 3) оценивание качества разметки, сравнение и выбор моделей

# Разметка текста: обобщённый контент-анализ

Пик научной фантастики (и советской, и западной) пришелся на 1960–1970-е годы. Однако в 1970-х годах этот жанр начал постепенно затухать и сходить на нет, уже в 1980-х на Западе начинает набирать силу жанр фэнтези. Конечно же, это неслучайно. Именно 1960-е годы стали пиком научно-технического прогресса в XX веке. К тому времени закончилась первая половина XX столетия, за эти полсотни лет было изобретено столько, что все казалось возможным, верилось, что прогресс будет нарастать по экспоненте. **1960-е — это мир безудержного социального и культурно-технического оптимизма.** Человек полетел в космос, запустил искусственные спутники и задумался об освоении других планет.

Но этот порыв человечества в будущее создавал определенную угрозу для власти имущих как на Западе, так и в Советском Союзе. И уже в 1960-е годы перед сотрудниками Тавистокского института изучения человека в Великобритании (причем по иронии судьбы он располагается в графстве Девоншир, рядом с дартмурскими болотами, где разыгрывалась мрачная драма «Собаки Баскервильей» Конан Дойля) **была поставлена задача притормозить научно-технический прогресс путем внедрения определенных информационно-психологических и организационных моделей.** В частности, стартовала работа по созданию молодежных и женских субкультур и движений (именно в это время как по заказу появились The Beatles, The Rolling Stones, стал развиваться экологизм).

Одна из главных задач, поставленных перед Тавистокком, звучала так: to stamp out the cultural optimism of the 1960s (искоренить, вырубить, вытравить культурный оптимизм 1960-х годов). А **научная фантастика, особенно советская, безусловно, была оптимистической по своему настрою.**

Некоторые менее оптимистические ноты (не могу их назвать пессимистическими, но они выглядели более сложными, чем просто оптимизм) прослеживались у ряда писателей в соцлагере, в частности в книгах Станислава Лема (достаточно почитать его «Астронавтов» и «Магелланово облако»). Однако общий настрой советской фантастики до середины 1960-х годов был преимущественно оптимистичным — это видно и по творчеству братьев Стругацких, и по романам Ивана Ефремова.

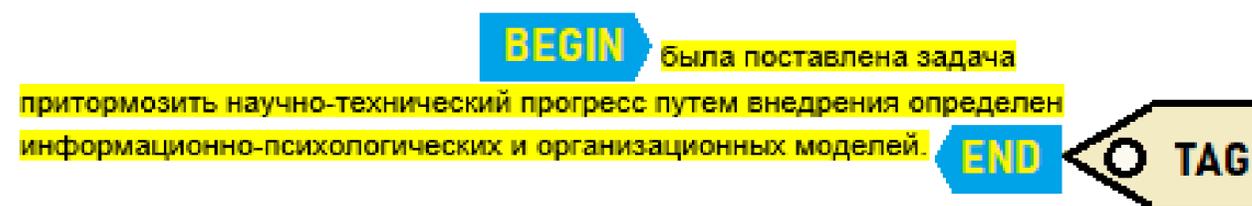
Первый доклад Римскому клубу (он создан в 1968 году) назывался «Пределы роста». В нем утверждалось, что человечество в своем индустриальном развитии достигло пределов, избыточно давит на природную среду, надо тормозить промышленно-экономическое развитие, перейдя к «нулевому росту». То есть 50 процентов всех средств должно идти на нейтрализацию негативных последствий, которые несет индустриальное развитие.

## Разметка состоит из элементов

**Элемент разметки** — несколько взаимосвязанных фрагментов, затекстов и тегов

**Теги (классы)** выбираются из рубрикатора

**Фрагмент** задаётся началом и концом, может иметь один или несколько тегов:



**Затекст** — комментарий, объяснение, дополнительная информация и т.п., может иметь один или несколько тегов

# Инструмент разметки

<https://markup.mlsa-iai.ru>

The screenshot displays the 'Markup' tool interface. At the top, there are navigation tabs: 'Задания', 'Теги', 'Документы', 'Постобработка', and 'Пользователь'. The 'Документы' tab is active, showing a document titled 'Документ №1: "23 февраля в войсковой части 5526..."'. The document text is shown with several blue highlights. Below the text is a 'Добавить фрагмент' button. To the right, the 'Постобработка' (Post-processing) panel shows a list of three extracted elements. Each element has a dropdown arrow, a checkbox, and a title. The first element is 'Элемент №2' with tags 'Здоровье' and 'Тональность: Положительная', and a selected snippet 'в стремлении к здоровому образу жизни'. The second element is also 'Элемент №2' with tags 'Уважение традиций' and 'Тональность: Положительная', and a selected snippet 'уважению к традициям'. The third element is 'Элемент №3' with tags 'Справедливость' and 'Тональность: Положительная', and a selected snippet 'служить можно по-разному, и не обязательно с оружием в руках. Служить можно в любом возрасте, служить можно и парню, и девушке. Служить можно и словом, и делом!'. At the bottom of the post-processing panel is a 'Сохранить и выйти' button.

← Задания Теги Документы Постобработка Пользователь

Список тегов

Документ №1: "23 февраля в войсковой части 5526..."

23 февраля в войсковой части 5526 прошло чествование воспитанников военно-патриотического клуба «Крепость». Заместитель командира части по идеологической работе майор Олег Ляшук отметил, что воспитание подрастающего поколения в патриотическом ключе, в стремлении к здоровому образу жизни, уважению к традициям, культурным ценностям и исторической памяти государства является главным профилактическим фактором безнравственности и аморальности. «Вместе мы вносим огромный вклад в будущее страны и нравственное здоровье нашего общества. Служба Родине во все времена была почетна. А служить можно по-разному, и не обязательно с оружием в руках. Служить можно в любом возрасте, служить можно и парню, и девушке. Служить можно и словом, и делом!» – подытожил Олег Ляшук.

Добавить фрагмент

Разметка от 09.02.2025: "23 февраля в войсковой части 5526..."

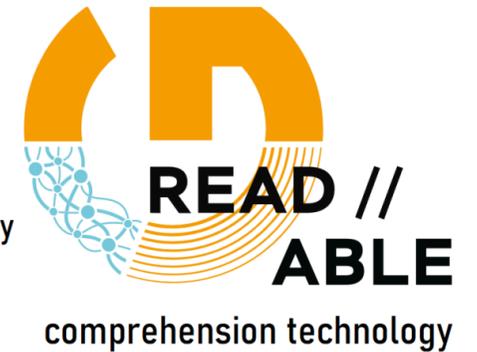
Выбрано объектов: 0 Выбрать все Убрать все

- Элемент №2 ×
  - Здоровье × Тональность: Положительная ×
  - "в стремлении к здоровому образу жизни" ×
- Элемент №2 ×
  - Уважение традиций × Тональность: Положительная ×
  - "уважению к традициям" ×
- Элемент №3 ×
  - Справедливость × Тональность: Положительная ×
  - "служить можно по-разному, и не обязательно с оружием в руках. Служить можно в любом возрасте, служить можно и парню, и девушке. Служить можно и словом, и делом!" ×

Сохранить и выйти

# Оценивание алгоритмической разметки

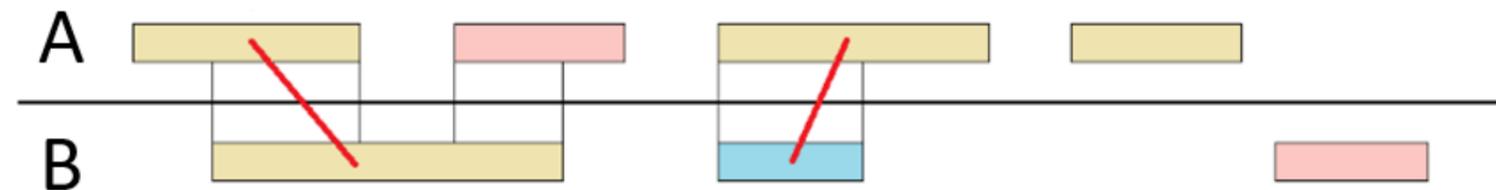
Up Great  
technology  
contest



- В основе методики — парное сравнение разметок текста:  
«алгоритм ↔ эксперт», «эксперт-1 ↔ эксперт-2»  
на основе оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок  $Con_{1,...,5}(A,B)$
- Вводится их средневзвешенная согласованность  $Con(A,B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по выборке  $Con(A,E)$  разметок алгоритма A и эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по выборке  $Con(E1,E2)$  разметок двух экспертов, E1 и E2
- ОТАР = СТАР / СТЭР, если больше 100%, то модель лучше экспертов

# Критерии согласованности разметок

## Оптимальное сопоставление элементов разметок A и B



Критерии (числовые величины от 0 до 1; чем выше, тем лучше):

**Con1** = доля фрагментов, для которых найдено сопоставление

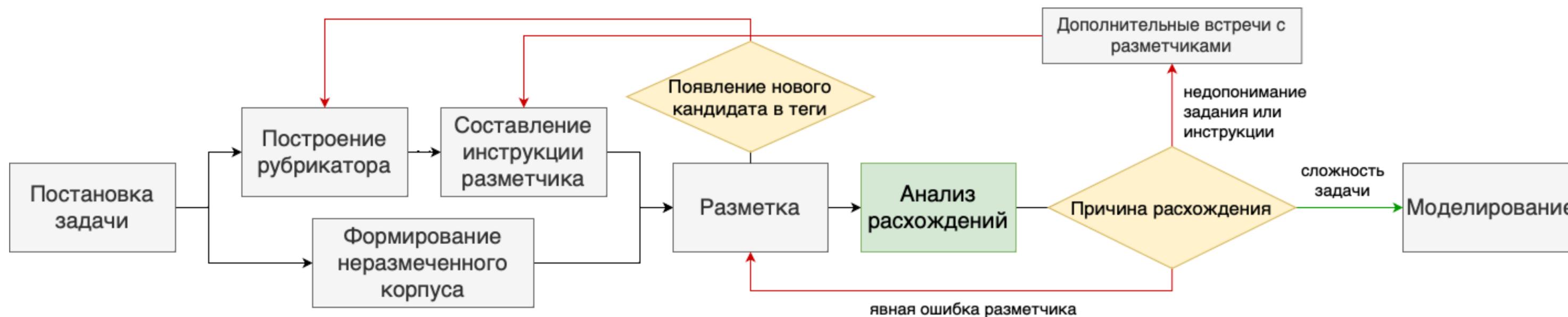
**Con2** = точность наложения сопоставленных фрагментов

**Con3** = точность совпадения тегов сопоставленных фрагментов

**Con4** = точность совпадения связей сопоставленных фрагментов

**Con5** = точность совпадения затекстов сопоставленных фрагментов

# Организация процесса разметки



- каждый документ размечается несколькими экспертами (2 или 3)
- документы ранжируются по согласованности экспертов  $Con(E, E')$
- наибольшие расхождения обсуждаются, вырабатывается консенсус
- происходит доработка инструкции и/или переразметка документов

# Детекция ценностей социокультурного кода

Создание рубрикатора на основе кросс-дисциплинарного подхода

Исследователи	Предмет исследований
Милтон Рокич	Ценностные ориентации людей - психология, социология
Герт Хоффстеде	Культурные характеристики народов - социология
Шолом Шварц	Теория базовых человеческих ценностей - социальная психология
Рональд Инглхарт	Исследование мировых ценностей - политология, социология
Сэмюэл Хантингтон	Этнокультурное описание цивилизаций - политология, социология
Юрий Сергеевич Степанов	Концепты русской культуры - лингвистика
Александр Александрович Аузан	Культурные коды экономики - экономика

Указ президента Российской Федерации № 400 от 2-07-2021 «О стратегии нацбезопасности»

Указ президента Российской Федерации № 809 от 9-11-2022 «Об утверждении основ госполитики по сохранению и укреплению традиционных российских духовно-нравственных ценностей»

# Какие ценности брать для рубрикатора: «аксиоматический» подход

- **Общественная значимость**  
ценность — это то, что оказывает влияние на социальную жизнь
- **Индивидуальная значимость**  
то, что влияет на принятие решений отдельными людьми
- **Субъективная измеримость**  
то, что человек может принимать или отвергать
- **Манипулятивность**  
то, на отношение к чему можно повлиять в процессе коммуникации
- **Текстуальность**  
то, что возможно описать, выразить текстом, фразой, историей
- **Атомарность**  
то, что не сводится к набору других ценностей

# Рубрикатор ценностей (1 из 2)

М. Рокич	Ш. Шварц	Г. Хофстеде
Ю.С. Степанов	Р. Инглхарт	Б.С. Ерасов

Группа социальных ценностей	
• Социальные ценности	• Личные границы
• Авторитет	• Материальные ценности
• Альтруизм	• Патриархальность
• Благородство происхождения / аристократизм	• Патриотизм
• Важность общественного мнения	• Пацифизм / мир во всём мире
• Воспитание	• Полезность (созидательный труд)
• Гендерное разнообразие	• Профессиональный успех
• Дети	• Репутация
• Долгосрочная ориентация	• Семья
• Дружба	• Социальное признание
• Избегание неопределённости	• Суеверия
• Индивидуализм	• Трудолюбие / продуктивность
• Интеллигентность	• Чувство принадлежности / единство народов
• Коллективизм	• Этничность
• Культура (нормы) поведения	• Язык

Группа витальных ценностей
• Витальные (необходимые) ценности
• Безопасность (личная)
• Время
• Еда
• Жизнь
• Жилище
• Здоровье
• Природа

Группа политических ценностей
• Политические ценности
• Власть
• Выборность власти (демократия)
• Институциональное доверие
• Историческая память и преемственность поколений
• Либерализм
• Национальная безопасность
• Права и свободы
• Правосознание (гражданская активность, гражданственность)
• Справедливость

# Рубрикатор ценностей (2 из 2)

М. Рокич	Ш. Шварц	Г. Хофстеде
Ю.С. Степанов	Р. Инглхарт	Б.С. Ерасов

## Группа религиозных ценностей

- Религиозные ценности
- Благочестивость
- Бог
- Религиозность
- Эзотерика

## Группа эстетических и гедонистических ценностей

- Эстетические и гедонистические ценности
- Жизнь, полная впечатлений
- Красота
- Культура и искусство
- Наслаждение жизнью
- Потворство желаниям
- Творчество
- Эстетика

## Группа экзистенциальных и познавательных ценностей

- Экзистенциальные и познавательные ценности
- Интеллект
- Критическое мышление
- Любовь
- Любознательность
- Мудрость
- Образование
- Перфекционизм
- Познание
- Принятие жизни
- Развитие
- Самостоятельность (выбор собственных целей)
- Смелость
- Смысл жизни
- Спокойствие (внутренняя гармония)
- Счастье
- Талант
- Твёрдая воля
- Целеустремлённость
- Широта взглядов

## Группа морально-нравственных ценностей

- Морально-нравственные ценности
- Аккуратность
- Беспечность
- Вежливость
- Верность
- Взаимопомощь и взаимоуважение
- Гордость
- Гостеприимство
- Гуманизм
- Дисциплина
- Достоинство (самоуважение, самооценочность)
- Духовность (приоритет духовного над материальным)
- Заботливость
- Искренность
- Милосердие
- Ответственность
- Скромность
- Смирение (послушание, кротость)
- Совесть (нравственный закон, мораль)
- Терпение
- Уважение традиций
- Целомудрие
- Честность
- Чувство юмора

# Ценностный ландшафт

Самые частотные теги  
(по количеству элементов)

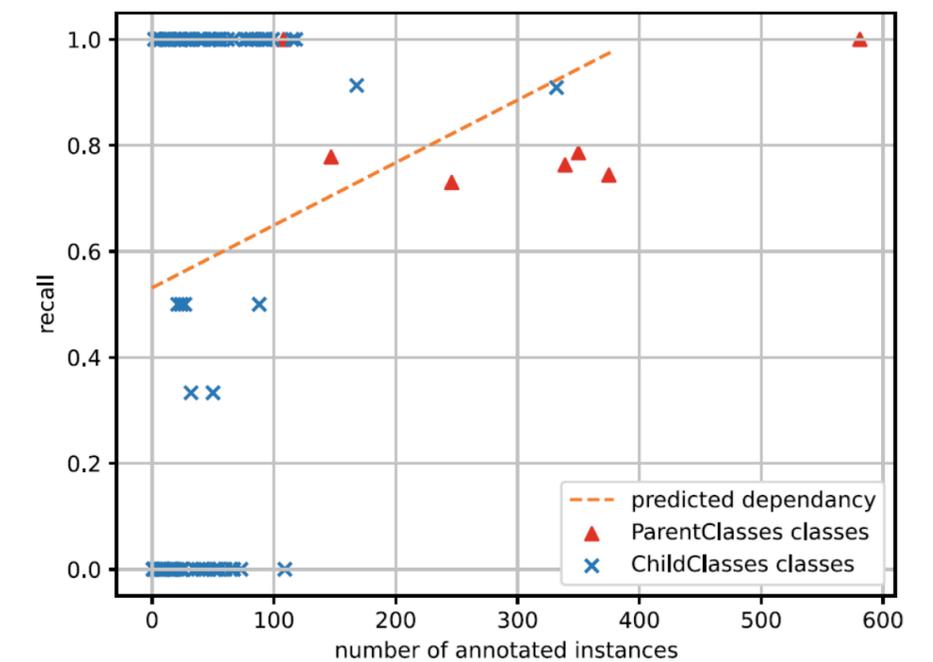
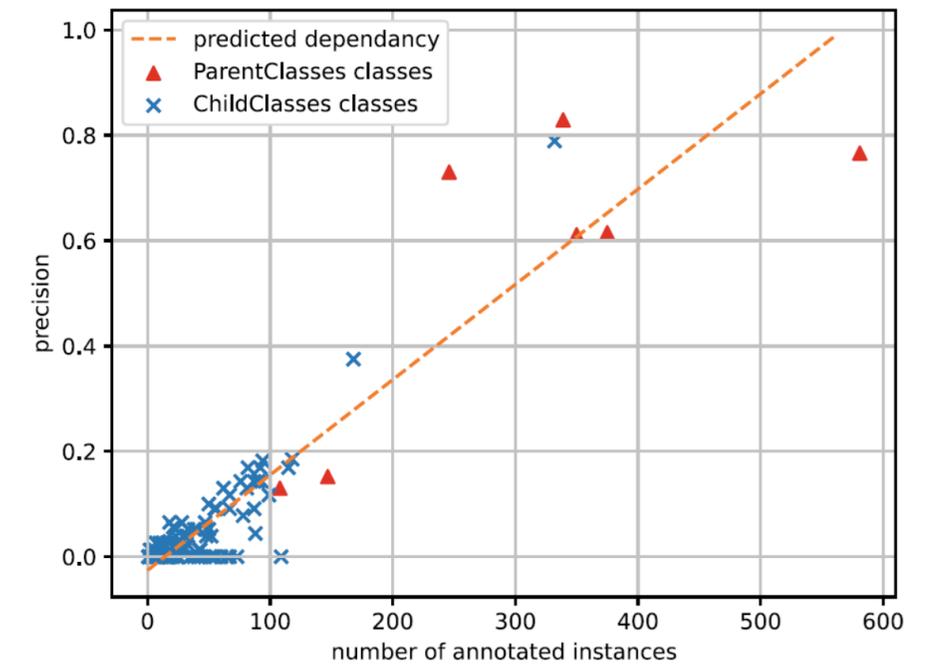
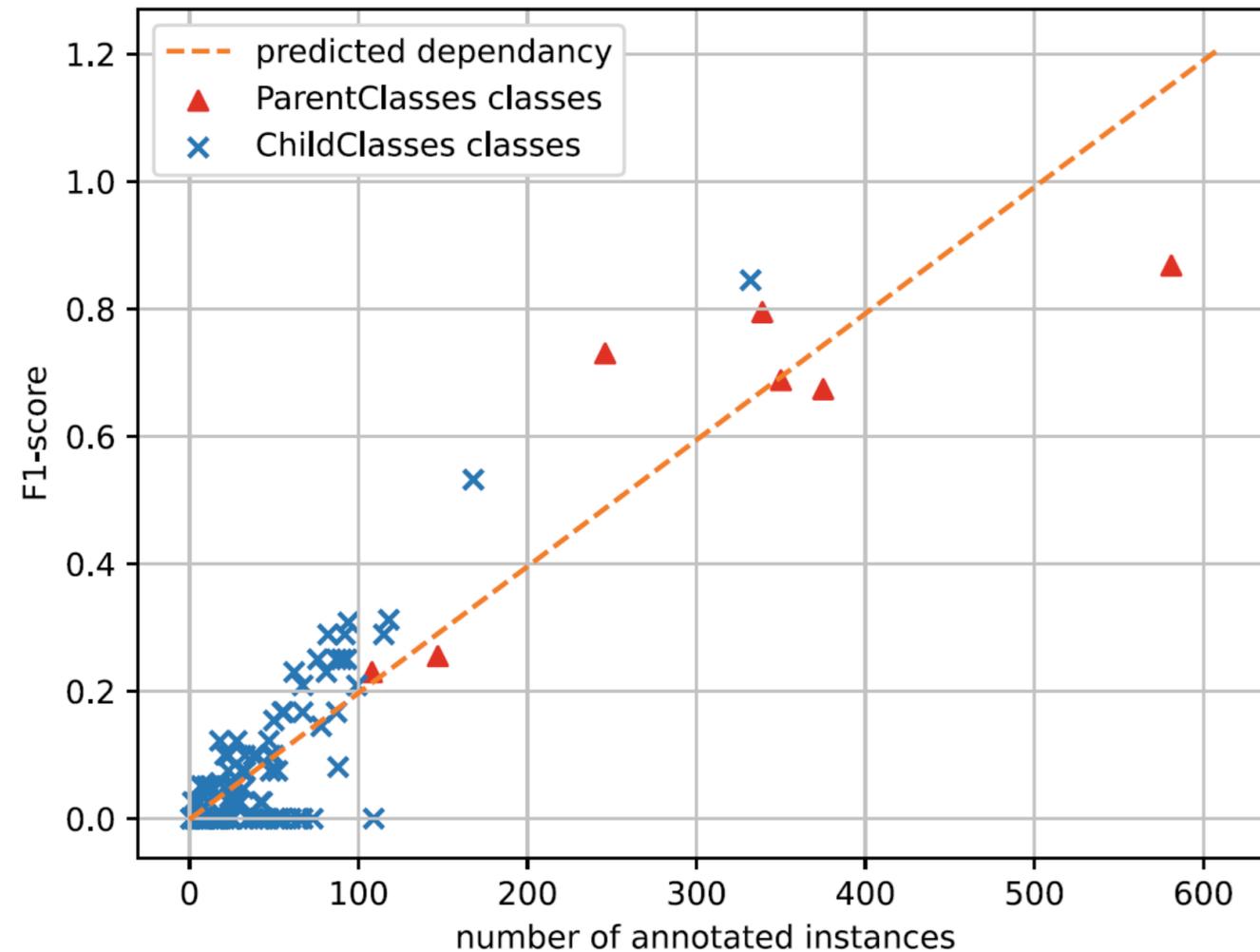
40%



1	Материальные ценности	1569
2	Жизнь	478
3	Историческая память и преемственность поколений	329
4	Правосознание (гражданская активность, гражданственность)	267
5	Политические ценности	246
6	Семья	243
7	Этничность	237
8	Культура и искусство	236
9	Права и свободы	235
10	Патриотизм	233

1. *Rink Olga, Lobachev Viktor, Vorontsov Konstantin*. Detecting human values and sentiments in large text collections with a context-dependent information markup: a methodology and math. HCII 2024. Lecture Notes in Computer Science series (in print). Cham: Springer.
2. *Vorontsov K.V., Gladchenko I.A., Lobachev V.A., Mamontova A.V., Rink O.L., Shabelskaya N.K.* Methodology for detecting human values in large text collections // Bulletin of St. Petersburg University. International relations. 2024

# Зависимость точности, полноты и F1 от объёма выборки, по классам



Rink O.L., Maysuradze A.I., Fedorov A.M., Ischenko R.V., Korchagina A.V., Tabachenkov A.M., Tsybanov I.A., Vorontsov K.V. Automated detection of human values in texts: ML challenges and performance benchmarks. 2025.

# Список публикаций

1. *Rink O.L., Lobachev V.A., Vorontsov K.V.* Detecting human values and sentiments in large text collections with a context-dependent information markup: a methodology and math. HCII 2024.
2. *Vorontsov K.V., Gladchenko I.A., Lobachev V.A., Mamontova A.V., Rink O.L., Shabelskaya N.K.* Methodology for detecting human values in large text collections // Bulletin of St. Petersburg University. International relations. 2024.
3. *Rink O. L., Vorontsov K. V., Shabelskaya N. K.* Uncovering positivism, negativism, and conflict in large text collections. Material values and the code of «noble maidens» // EDN ADEUQP, April 18–20, 2024. – P. 137-139.
4. *Maysuradze A.I., Rink O.L., Fedorov A.M., Tabachenkov A.M., Vorontsov K.V.* Does annotating multi-spans improve classification in large text collections? 2024ICSAI (China & IEEE). Lecture Notes in Computer Science series, 2024.
5. *Vorontsov K.V., Gladchenko I.A., Lobachev V.A., Mamontova A.V., Rink O.L., Shabelskaya N.K.* Developing an Open Interdisciplinary Classifier of Human Values by means of Annotating Multi-fragments // Bulletin of St. Petersburg University. International relations. 2025.
6. *Rink O.L., Maysuradze A.I., Fedorov A.M., Ischenko R.V., Korchagina A.V., Tabachenkov A.M., Tsybanov I.A., Vorontsov K.V.* Automated detection of human values in texts: ML challenges and performance benchmarks. 2025.

# Бенчмарк для универсальных моделей разметки

Состоит из датасетов с задачами разметки текстов с максимально широким покрытием различных доменов.

Все датасеты приведены к универсальному формату данных разметки для демонстрации универсальности подхода.

Бенчмарк содержит:

- 21 датасет
- 17 типов задач

**Мультиасессорная разметка** встречается в некоторых датасетах, однако **авторы пренебрегают решением проблемы противоречивости разметок** и используют стандартные для области критерии оценивания качества.



# Датасеты для бенчмарка универсальных моделей разметки

Dataset	Task Type	Level	Paper Count
Kaggle NER Corpus (Bos et al., 2017)	NER	Span Level-1	25
	POS Tagging	Span Level-1	5
MultiCoNER (Malmasi et al., 2022)	NER	Span Level-1	85
RuTermEval Dialogue (Mamontova et al., 2025)	NER	Span Level-1	8
SWDA (Stolcke et al., 2000)	Dialogue Act	Span Level-2	228
RuSentNE (Golubev et al., 2023)	NER	Span Level-1	5
	Semantic analysis	Span Level-2	5
	Opinion tuple extraction	Element Level	2
ADE (Gurulingappa et al., 2012)	Text classification	Span Level-1	10
	NER	Span Level-1	150
	RE	Element Level	43
	Coreference resolution	Element Level	1
DDI corpus (Herrero-Zazo et al., 2013)	NER	Span Level-1	102
	RE	Element Level	131
PcMSP (Yang et al., 2022)	Text classification	Span Level-1	1
	NER	Span Level-1	5
	RE	Element Level	5
ChemProt (Krallinger M., 2017)	Text classification	Span Level-1	1
	NER	Span Level-1	3
	RE	Element Level	70
NERRE (Dagdelen et al., 2024)	NER	Span Level-1	10
	RE	Element Level	11
NEREL (Loukachevitch et al., 2023)	NER	Span Level-1	9
	RE	Element Level	2
	Extraction of multi-spans with named entities	Element Level	1
RURED (Gordeev et al., 2020)	NER	Span Level-1	3
	RE	Element Level	6
	Entity linking	Element Level	1

Scienc (Luan et al., 2018)	NER	Span Level-1	137
	RE	Element Level	134
	Coreference resolution	Element Level	113
CONLL 2012 Ontonotes (Pradhan et al., 2013)	NER	Span Level-1	87
	Semantic Role labeling	Element Level	25
	Coreference resolution	Element Level	15
RuSuperGLUE's RWSD task (Shavrina et al., 2020)	Relation classification	Element Level	7
MERA's RWSD task (Fenogenova et al., 2024)	Relation classification	Element Level	5
MERA's Ruethics task (Pradhan et al., 2013)	Relation classification	Element Level	1
SemEval 2010 Task 8 (Hendrickx et al., 2019)	Relation classification	Element Level	8
SemEval-2018 Task 7 (Buscaldi et al., 2017)	NER	Span Level-1	14
	RE	Element Level	57
	Relation classification	Element Level	57
	Knowledge Graph Construction	Element Level	3
UpGreat READ//ABLE (READ//ABLE, 2017-2025)	Extraction and classification of spans with errors	Span Level-2	2
	Extraction of multi-spans with errors	Element Level	1
	Annotation of text spans with comments	Extended Level	1
Human Values dataset (Rink et al., 2024)	Text classification	Span Level-2	1
	Extraction of spans with human values	Span Level-2	2
	Extraction of elements with human values	Element Level	2
	Semantic analysis of elements with human values	Element Level	2
	Annotation of text markups with comments	Extended Level	2

# Подытожим. Автоматизация контент-анализа для цифровых гуманитарных исследований

## Технологии LLM дополняют схему исследований в контент-анализе:

1. Остаётся:
  - разработка рубрикатора, инструкций разметчика, организация разметки
2. Остаётся:
  - выбор репрезентативного корпуса, ручная разметка
3. Добавляется:
  - обучение универсальной модели разметки по размеченному корпусу,
  - масштабирование: автоматическая разметка больших (всех?) данных,
  - обоснование качества автоматической разметки
4. Остаётся:
  - качественный и количественный анализ размеченного корпуса

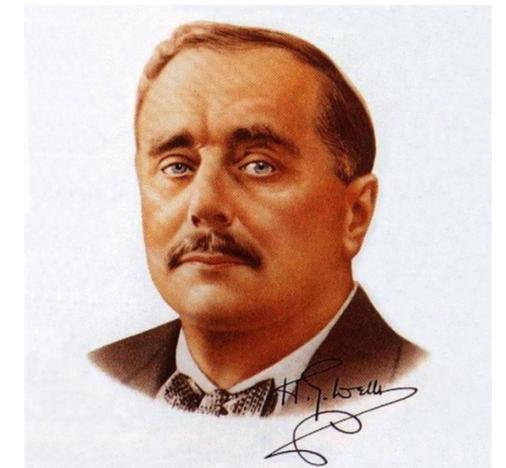
## Машинное обучение и семантический анализ

- тематическое моделирование в ДН
- автоматизация контент-анализа
- «мастерская знаний»: концепция, цели, задачи

# Концепция «Мастерской знаний»

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в своеобразной **мастерской ума**, где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи.» – Герберт Уэллс, 1940

(An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: a **depot (склад с мастерской)** where **knowledge** and ideas are received, sorted, summarized, digested, clarified and compared – Herbert Wells, 1940)



Теперь технологии IR/NLP/LLM позволяют ставить и решать такие задачи

# Что такое «знания»



мудрость  
(wisdom)

самое главное:  
смыслы, ценности, цели, задачи



знания  
(knowledge)

информация, структурированная  
для удобства понимания и  
практического использования



информация  
(information)

результат обработки и  
анализа данных



данные  
(data)

зарегистрированные факты  
окружающей реальности

Технологии больших языковых моделей (Large Language Model, LLM) позволяют выделять знания и идеи из текста и систематизировать их

# Эволюция подходов в обработке естественного языка

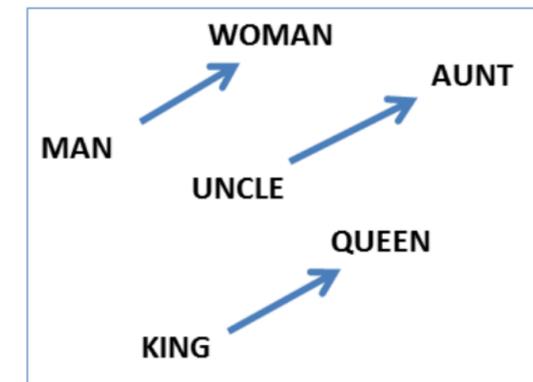
## Декомпозиция задач по уровням «пирамиды NLP»

- морфологический анализ, лемматизация, опечатки,...
- синтаксический анализ, выделение терминов, NER,...
- семантический анализ, выделение фактов, тем,...



## Векторные представления слов (embeddings)

- модели дистрибутивной семантики: word2vec [Mikolov, 2013], FastText [Bojanowski, 2016],...
- тематические модели LDA [Blei, 2003], ARTM [2014],...



## Большие языковые модели (БЯМ, LLM)

- рекуррентные нейронные сети: LSTM, GRU,...
- «end-to-end» модели внимания, трансформеры: Google NMT [2016], BERT [2018], GPT-3 [2020], GPT-4 [2023],...

$$\text{softmax} \left( \frac{\begin{matrix} \mathbf{Q} & \mathbf{K}^T \\ \begin{matrix} \square & \square \\ \square & \square \end{matrix} & \begin{matrix} \square & \square \\ \square & \square \end{matrix} \end{matrix}}{\sqrt{d}} \right) \begin{matrix} \mathbf{V} \\ \begin{matrix} \square & \square \\ \square & \square \end{matrix} \end{matrix}$$

# От поиска информации к «Мастерской знаний»

## Недостатки обычного поиска:

- как искать новые знания?
- что делать с найденным?

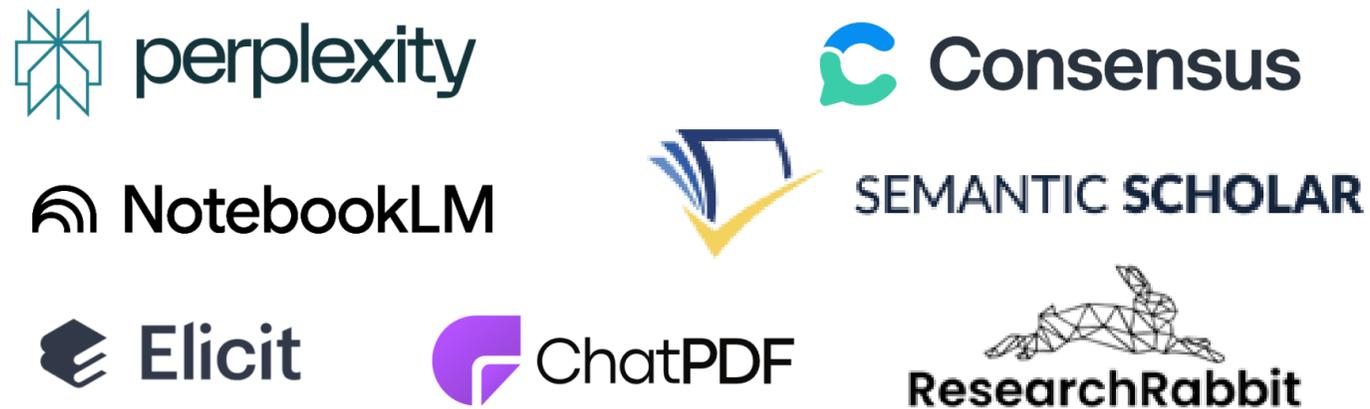


**Мастерская знаний** – инструментарий для работы с текстовыми источниками **на всём жизненном цикле** научного проекта:

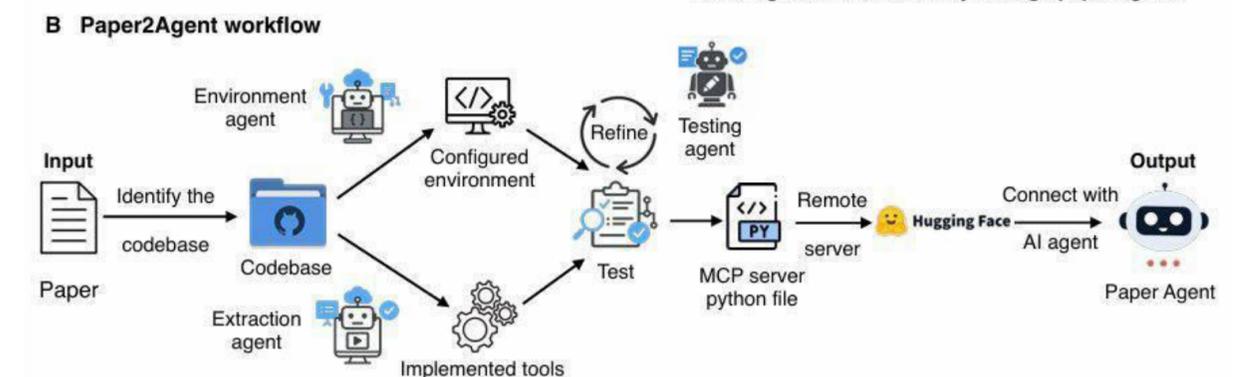
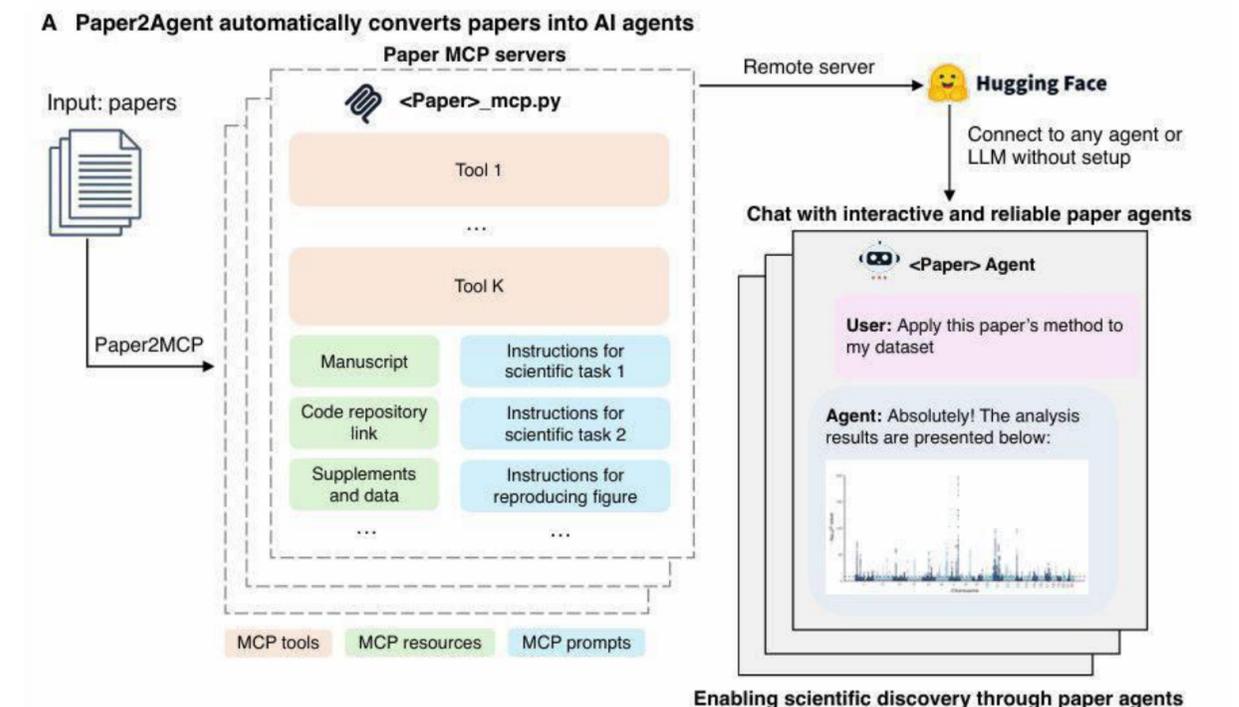
- ищу текстовые документы – чтобы сохранять их и накапливать
- накапливаю – чтобы их перечитывать, анализировать, понимать
- понимаю – чтобы получать, обрабатывать, систематизировать *знания*
- систематизирую – чтобы применять и передавать *знания и мудрость*

Теперь технологии IR/NLP/LLM позволяют ставить и решать такие задачи

# Научный поиск на основе LLM и ИИ-агентов



**Paper2Agent** — интерактивный ИИ-агент  
<https://github.com/jmiao24/Paper2Agent>



## Открытые проблемы ИИ-систем:

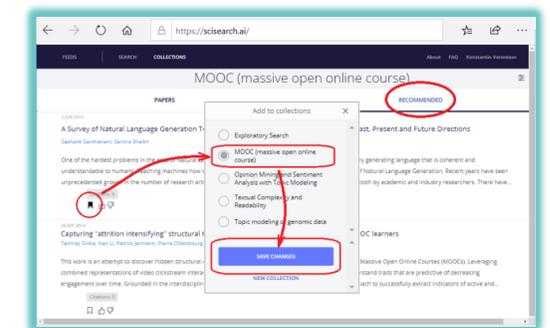
- как зафиксировать долгосрочный тематический поисковый интерес?
- как выделять и как обновлять знания?
- как обеспечить ясность представления знаний — «посмотрел и всё понял»?
- как включить «коллективный разум»?

# Концепция сервисов «Мастерской знаний»

**Подборка** текстов фиксирует тематику поискового интереса пользователя или группы  
**Расширенная подборка:** + результаты поиска семантически близких текстов

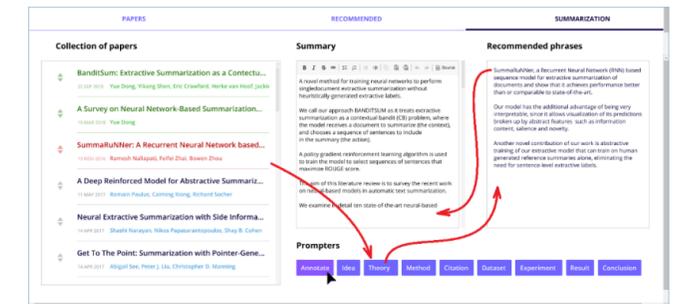
## Поисково-рекомендательные сервисы:

- поиск семантически близких документов по **подборке**
- контекстный поиск по фрагменту документа из **подборки**
- мониторинг новых документов по тематике **подборки**



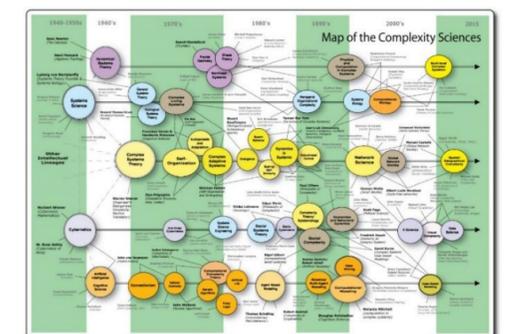
## Аналитические сервисы:

- полуавтоматическое реферирование **подборки**
- тематизация, картирование, онтологизация **подборки**
- хронологизация, выявление трендов по тематике **подборки**
- контент-анализ, сбор и анализ фактов из документов **подборки**



## Коммуникативные сервисы:

- совместное обсуждение, анализ, использование **подборок**
- создание нового контента в соавторстве на основе **подборки**

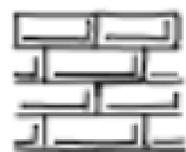


# Декларация принципов «Мастерской знаний»

- 1. Тематичность.** Рабочая среда пользователя образуется тематическими подборками
- 2. Текстуальность.** Знания содержатся в текстах, написанных людьми для людей
- 3. Коллегиальность.** Формы представления знаний служат взаимопониманию в группе
- 4. Креативность.** Пользователи создают в среде свой контент, информационный продукт
- 5. Доверенность.** Меньше генерации, больше экстракции, источников, ссылок
- 6. Антропоцентричность.** Интенсификация творчества людей — цель, а не средство
- 7. Когнитивность.** Представление знаний учитывает особенности восприятия и памяти
- 8. Мультиязычность.** Автоматический перевод с языков источника на язык пользователя
- 9. Расширяемость.** Платформа МЗ поддерживает возможность добавления сервисов
- 10. Открытость.** Базовые функции общедоступны ради устранения цифрового неравенства
- 11. Экономичность.** Чтобы сделать мир умнее, сначала сделать монетизируемый продукт
- 12. Социоцентричность.** Проектируя систему, предвидеть социальные практики и эффекты

## Миссия Мастерской Знаний

— устранять *барьеры* между человеком и знанием



технологические

из-за избыточности, неструктурированности, ненадёжности информации



КОГНИТИВНЫЕ

из-за ограниченности наших возможностей запоминания, понимания, анализа

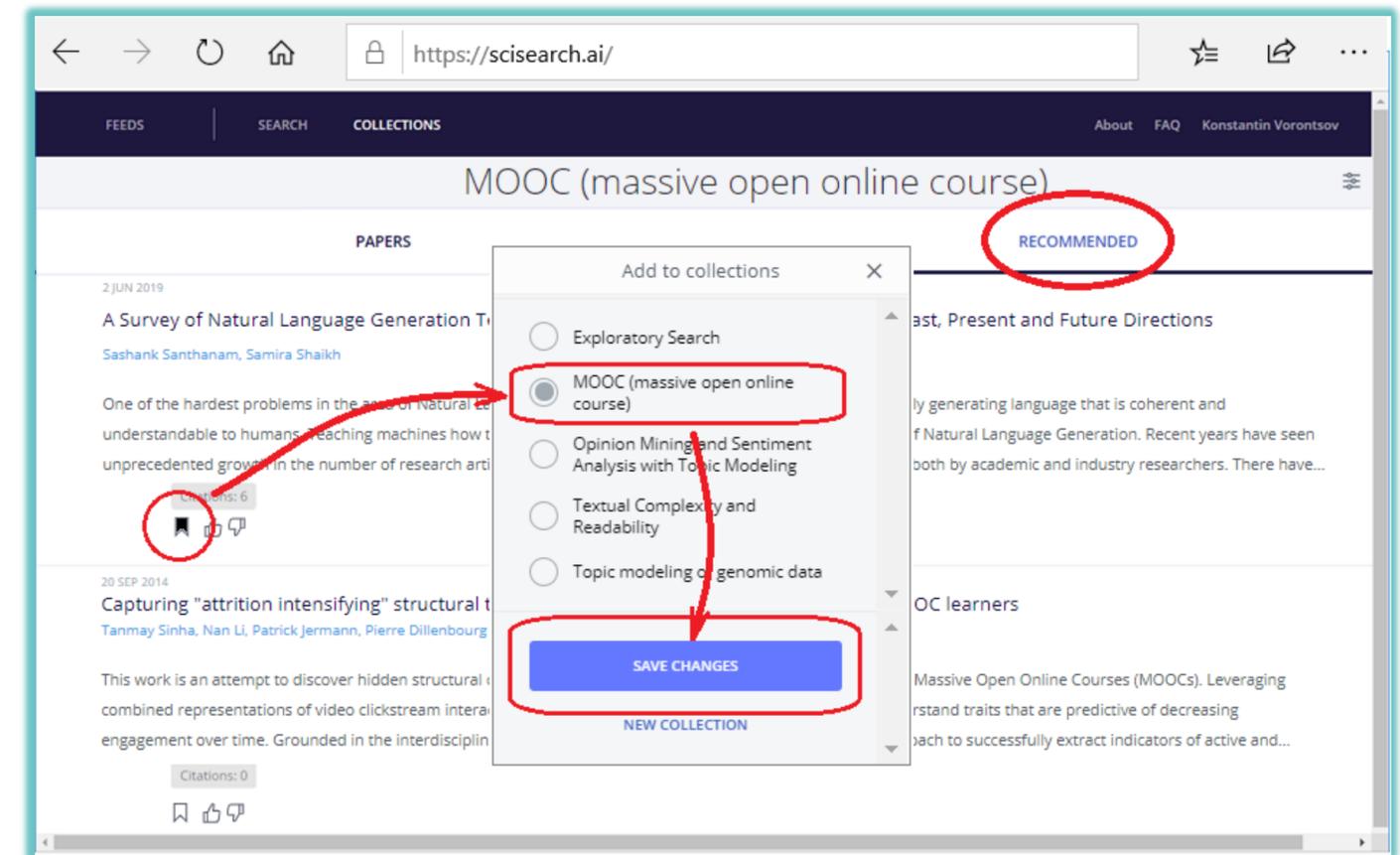
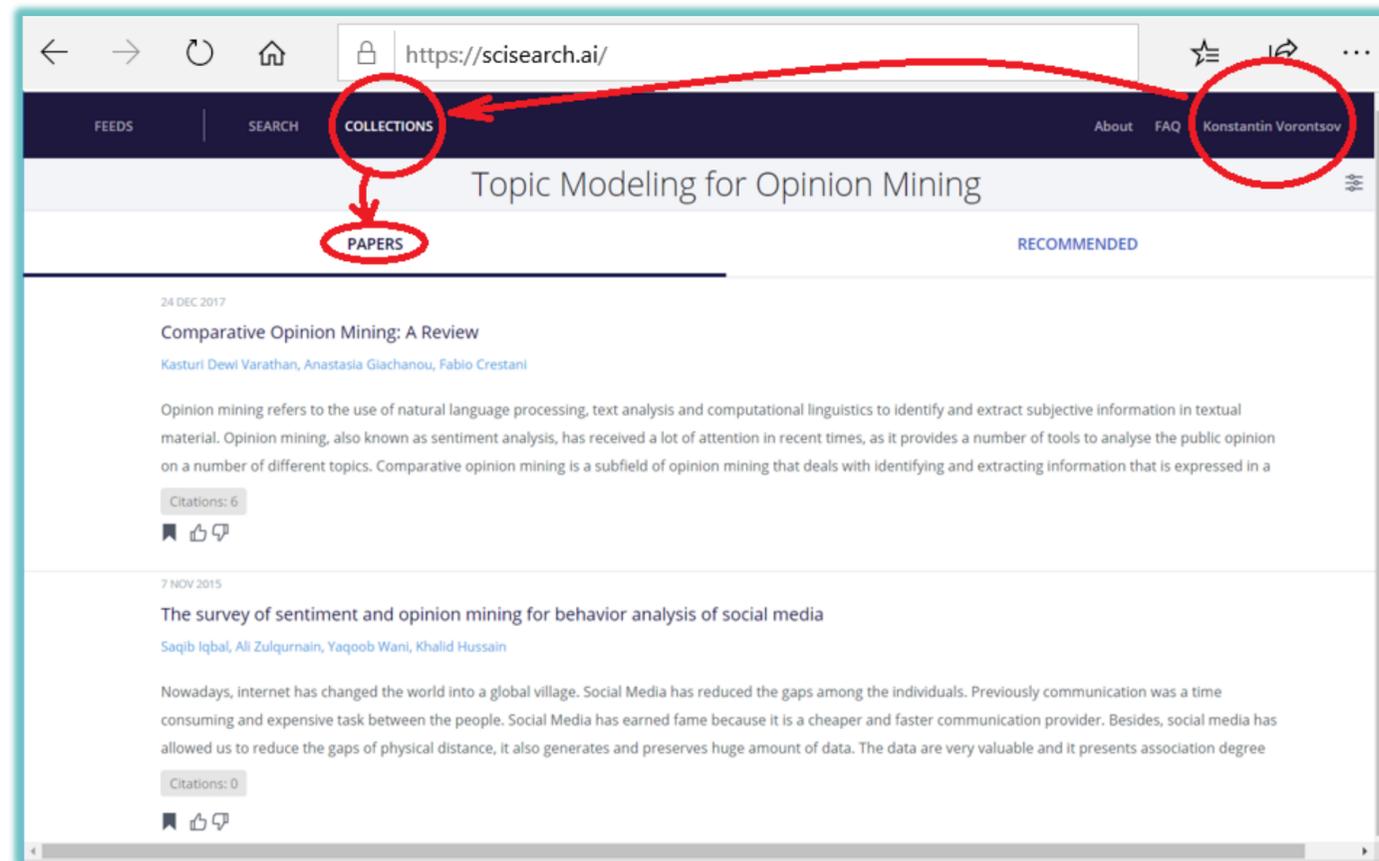


коммуникативные

из-за различий в мотивациях, уровне компетенций, социальном и служебном положении

# Сервис поиска и ранжирования рекомендаций

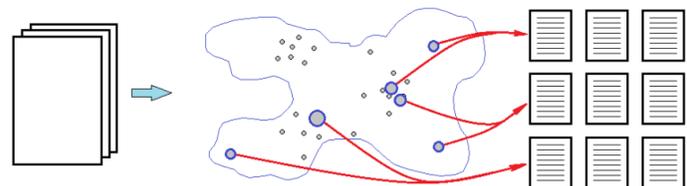
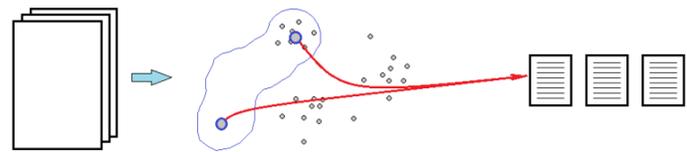
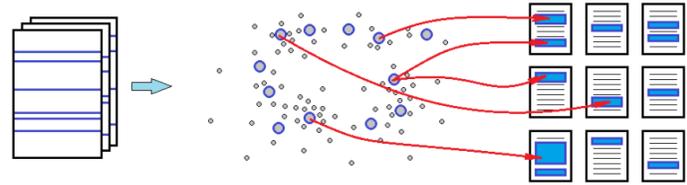
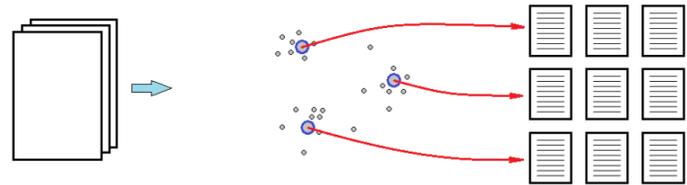
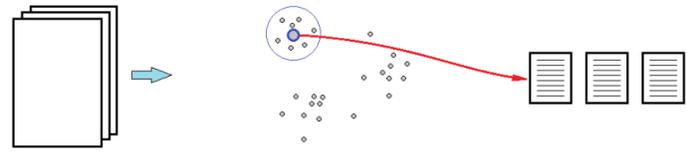
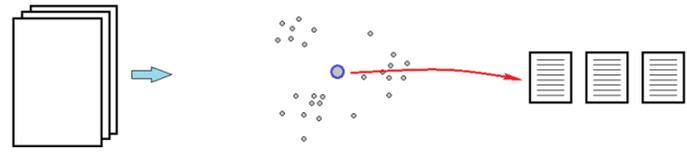
**Цель:** помочь пользователю быстро собрать тематическую подборку по своей информационной потребности, бегло знакомясь с документами



Герасименко Н.А., Ватолин А.С., Янина А.О., Воронцов К.В. SciRus: легкий и мощный мультязычный энкодер для научных текстов // Доклады РАН, 2024, том 520

Ватолин А.С., Герасименко Н.А., Янина А.О., Воронцов К.В. RuSciBench: открытый бенчмарк для оценки семантических векторных представлений научных текстов на русском и английском языках // Доклады РАН, 2024, том 520

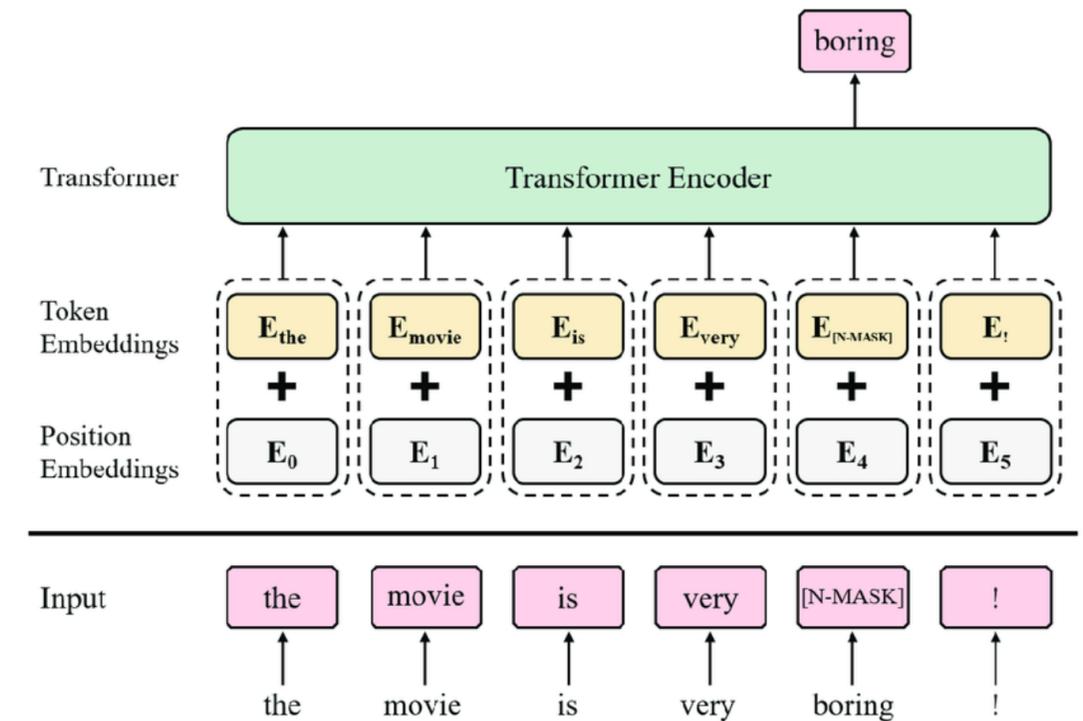
# Стратегии векторного поиска документов



1. Поиск по среднему вектору **подборки** (самая простая, но не самая удачная стратегия)
2. Поиск по документу из **подборки** или нескольким близким к нему документам
3. Разбиение **подборки** на кластеры и поиск по центральным документам кластеров
4. Разбиение документов **подборки** на сегменты и поиск по сегментам документов
5. Поиск по документам смежной тематики для документа или части документов **подборки**
6. Поиск по тематике, смежной для всей **подборки**

# Большие языковые модели научных текстов

- **SciBERT (2019)** *Beltagy et al.*  
SciBERT: A pretrained language model for scientific text
- **SPECTER (2020)** *Cohan et al.*  
SPECTER: Document-level representation learning using citation-informed transformers
- **LaBSE (2020)** *Feng et al.*  
Language agnostic BERT sentence embedding
- **MPNet (2020)** *Song et al.*  
MPNet: Masked and permuted pre-training for language understanding
- **SPECTER-2 (2022)** *Singh et al.*  
SciRepEval: A multi-format benchmark for scientific document representations
- **SciNCL (2022)** *Ostendorff et al.*  
Neighborhood contrastive learning for scientific document representations with citation embeddings
- **mE5 (2024)** *Wang et al.*  
Multilingual E5 text embeddings: A technical report. 2024.



# Модель SciRus: мотивации исследования

**Модель должна быть применима** в русскоязычных сервисах **для экстрактивных задач**: поиска, рекомендации, классификации, анализа текстовых документов — в различных сервисах и приложениях («Мастерская знаний», eLibrary.ru, научные электронные библиотеки)

## Требования к модели:

- минимальный размер (23М параметров)
- при качестве, сопоставимом с лучшими (SOTA) моделями
- возможность вычисления эмбедингов без GPU
- мультиязычность: английский, русский, **китайский** и др.
- возможность дообучения модели по данным о цитировании
- оценивание качества — по стандартным + новым benchmark-ам

# Данные для обучения модели научных текстов

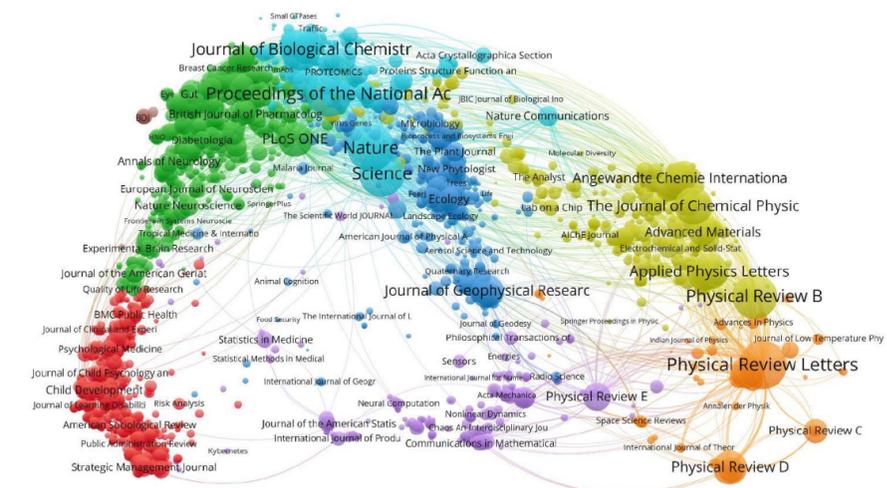
## Данные для обучения — title+abstract:

- **S2ORC — Semantic Scholar Open Research Corpus**  
30M (12B токенов), 85% en, 2% ru
- **eLibrary:**  
8.5M (2B токенов) ru  
5.2M (1.2B токенов) ru+en
- **ScienceChina (title+abstract):**  
5M аннотаций (0.85 токенов) en+zh



## Данные для дообучения:

- **S2AG — Semantic Scholar Academic Graph**  
источники: Crossref, PubMed, Unpaywall и др.  
2.5B связей цитирования



# Методики оценивания моделей (benchmarks)

## SciDocs: 6 задач

- классификация статей по MeSH / по тематике
- предсказание цитирования / со-цитирования
- предсказание пользовательской активности, рекомендации статей

## SciRepEval: 24 задачи, вкл. SciDocs (кроме рекомендаций):

- классификация, регрессия, сходство, поиск,
- подбор рецензента для статьи, разрешение неоднозначности авторов

## RuSciBench: 14 задач

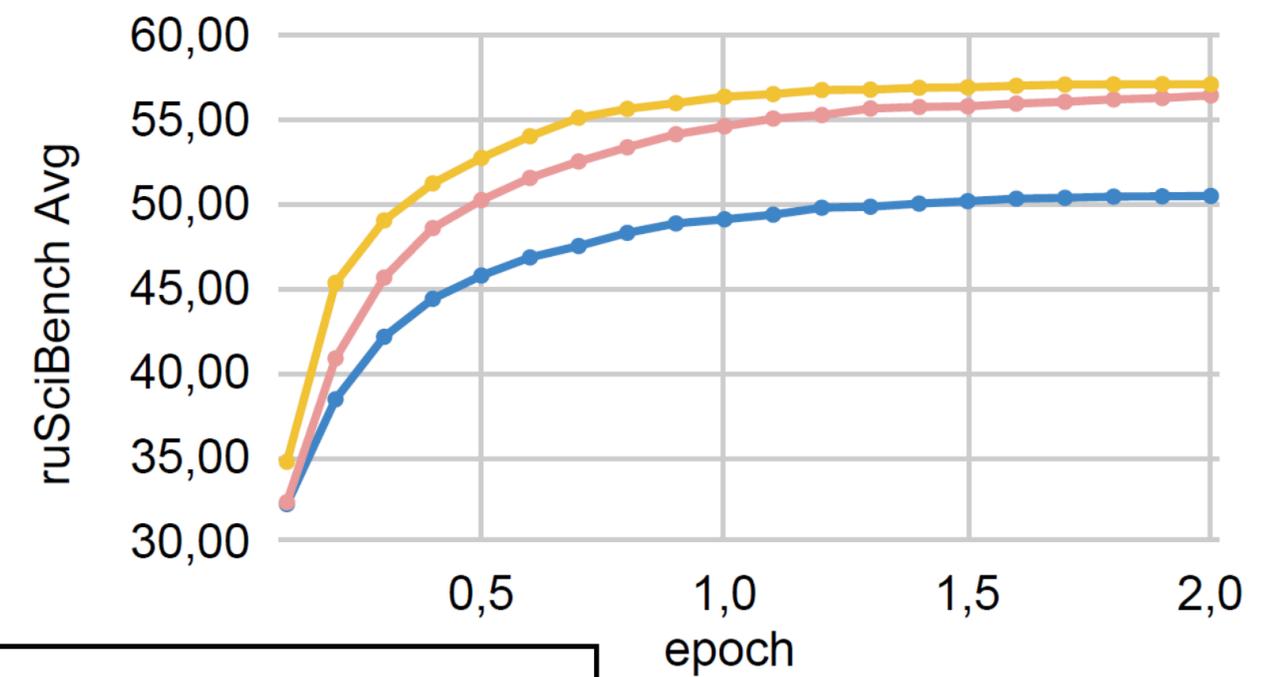
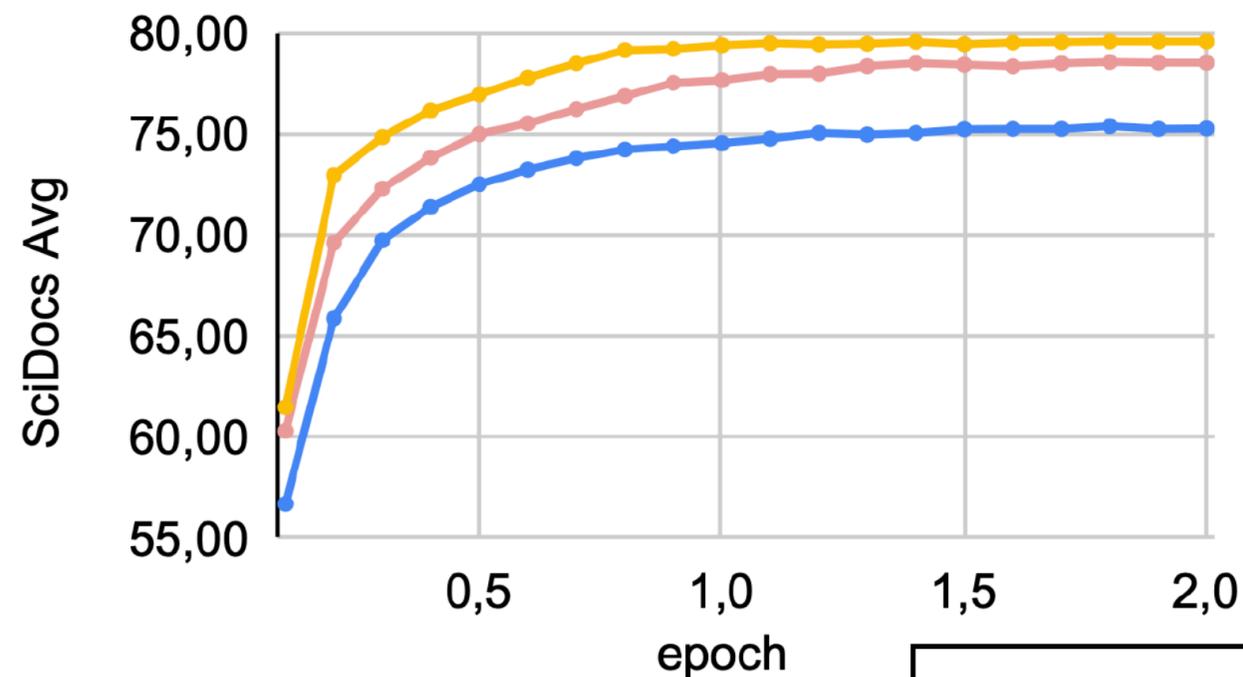
- 6 задач классификации OECD/ГРНТИ по аннотации ru / en / ru+en
- 4 задачи кросс-язычного поиска ru→en / en→ru / zh→en / en→zh
- 2 задачи предсказания цитирования / социтирования
- 2 задачи регрессии: предсказание года и цитируемости публикации



# Этап 1: предобучение модели SciRus-tiny (MSU)

**Архитектура** RoBERTa (Y.Liu et al., 2019), случайная инициализация:  
**tiny** (sz=23M, dim=312), **small** (sz=61M, dim=768), **base** (sz=85M, dim=1024)

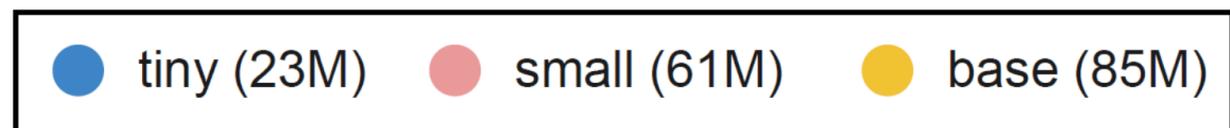
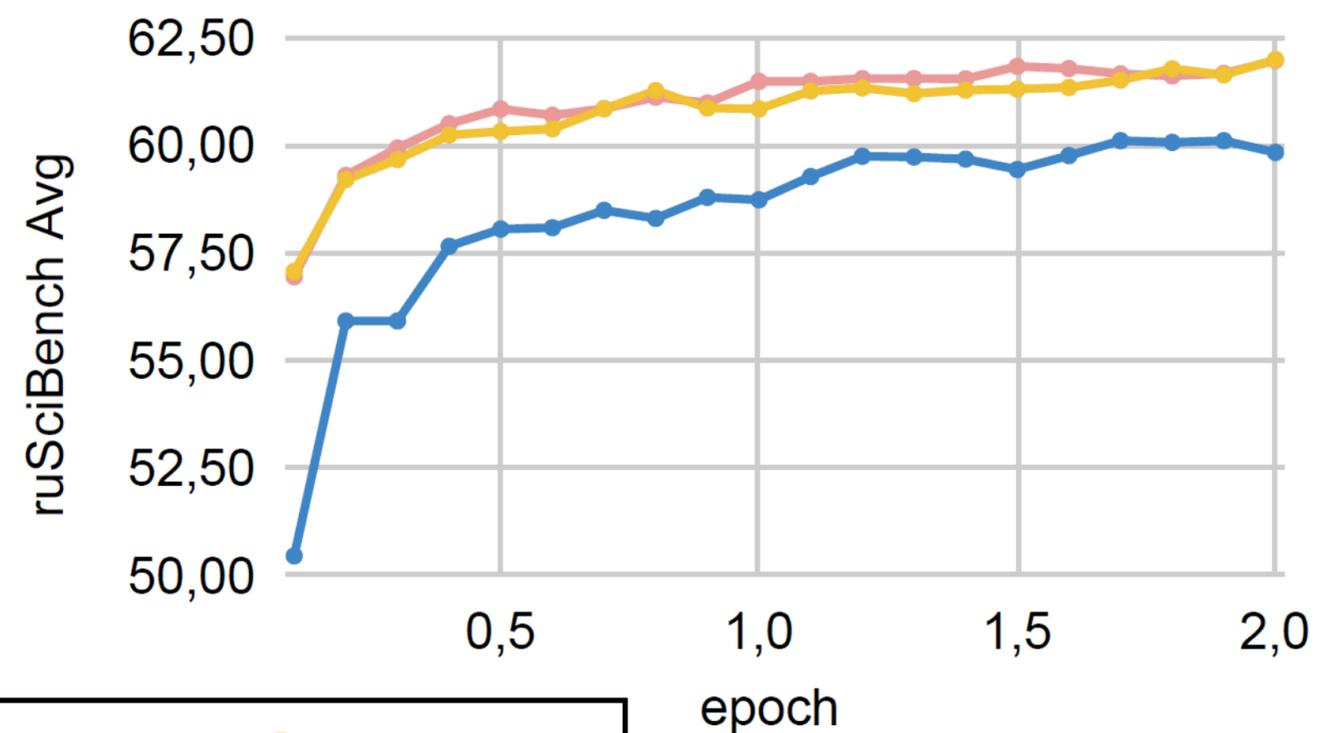
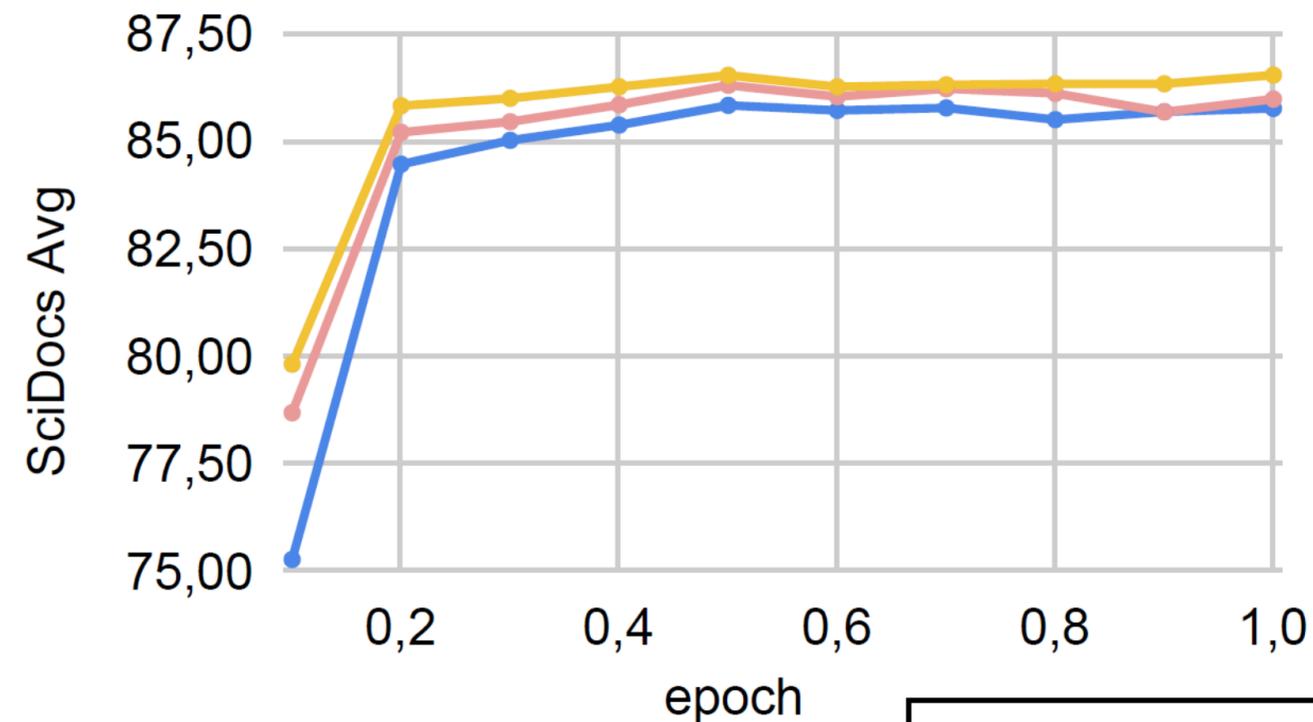
- критерий маскированного языкового моделирования MLM
- две эпохи обучения
- Avg — F1-мера, усреднённая по всем задачам бенчмарка



# Этап 2: дообучение на парах title-abstract

**Критерий:** сблизать эмбединги в контрастных парах название/аннотация, ru/en

- 30.6M пар из S2AG
- 17.8M пар из eLibrary



# Этап 3: дообучение на парах cite-cocite

**Критерий:** сблизать эмбединги пары документов (А,В) при цитировании:

«cite» — статья А цитирует статью В

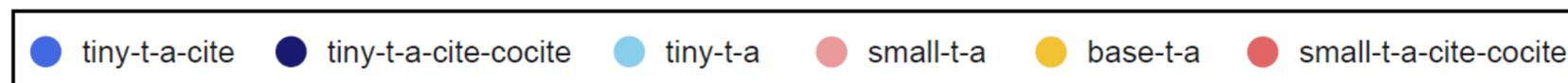
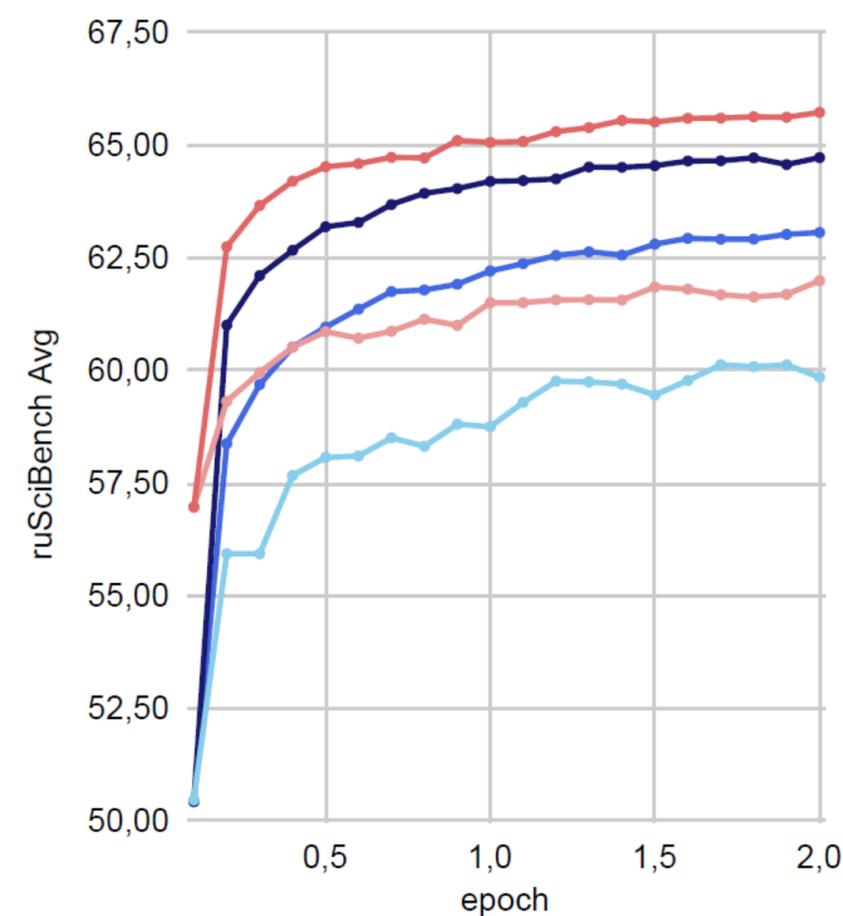
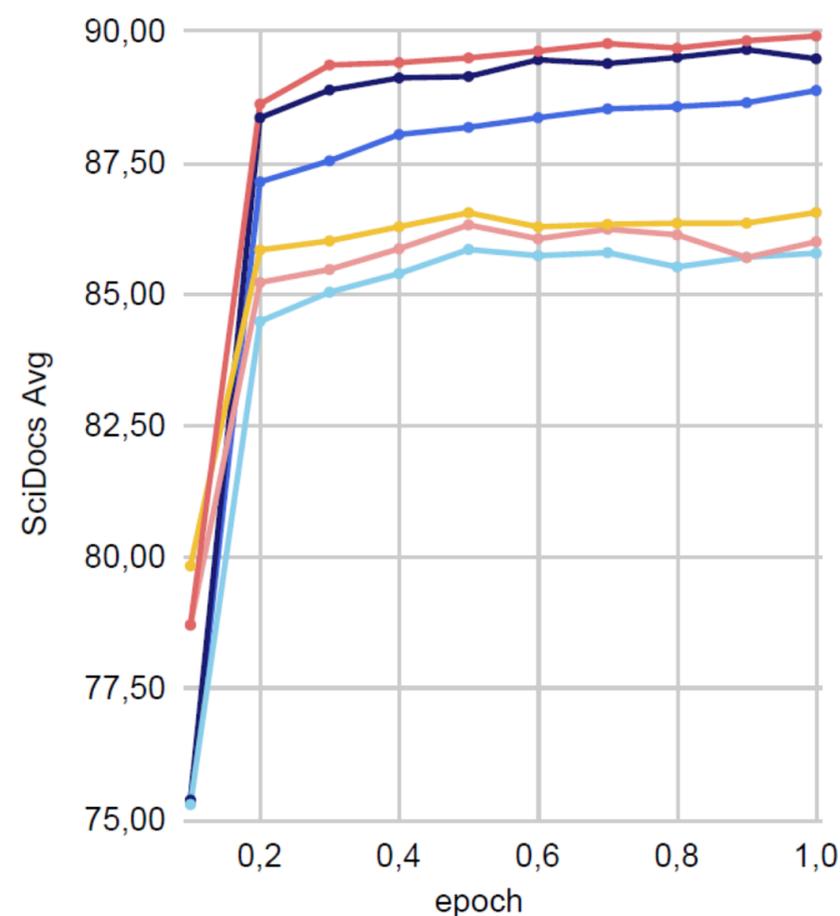
«co-cite» — третья статья С цитирует статьи А и В

## S2AG:

- 13.3М пар cite
- 62М пар co-cite

## eLibrary:

- 40М пар cite
- 33.7М пар co-cite



# Сравнение моделей по метрикам ruSciBench

 SOTA

model_name	Model size	elibrary_oecd_full	translation_search	
		macro_f1	ru_en recall@1	en_ru recall@1
e5-mistral-7b-instruct	7.11B	<b>67,28</b>	3,65	18,11
<b>scirus-tiny3.1</b>	<b>23M</b>	<b>65,40</b>	<b>97,40</b>	<b>98,80</b>
multilingual-e5-large	560M	63,70	<b>99,19</b>	<b>99,37</b>
<b>scirus-tiny2</b>	<b>23M</b>	<b>62,02</b>	<b>96,70</b>	<b>95,11</b>
multilingual-e5-base	278M	62,00	97,00	98,00
LaBSE	471M	60,21	98,31	97,20
LaBSE-en-ru	128M	60,05	98,26	96,93
paraphrase-multilingual-mpnet-base-v2	118M	60,03	66,33	78,18
FRED-T5-large	360M	59,80	22,25	0,79
distiluse-base-multilingual-cased-v1	135M	58,69	92,04	90,83
paraphrase-multilingual-MiniLM-L12-v2	118M	56,48	72,87	77,49
mfaq	280M	54,84	86,75	90,11
<b>scirus-tiny</b>	<b>23M</b>	<b>54,83</b>	<b>88,00</b>	<b>88,00</b>

- Сильнее модели, которая в ~20 раз больше
- Приблизились вплотную к SOTA, которую держит модель в ~300 раз больше

# Сравнение моделей по метрикам SciRepEval

Model name	Model size	SciDocs	Out-of-Train	In-Train
all-mpnet-base-v2	110M	<b>91,03</b>	50,2	53,12
scincl	110M	90,84	<b>51,8</b>	55,6
<b>scirus-tiny3.1</b>	<b>23M</b>	<b>90,1</b>	<b>50,08</b>	<b>57,2</b>
SPECTER	110M	89,10	50,6	54,7
e5-large-v2	335M	88,70		
e5-base	109M	88,58		
e5-base-v2	109M	88,43		
multilingual-e5-large	560M	87,53	49,32	55,65
e5-small-v2	33.4M	86,99		
multilingual-e5-base	278M	86,91		
e5-mistral-7b-instruct 4byte	7.11B	86,03		
<b>scirus-tiny2</b>	<b>23M</b>	<b>84,21</b>		
sentence-transformers/LaBSE	471M	80,78		
e5_pretrain_longer_240000_similarity_step_5581	23M	80,51		
cointegrated/rubert-tiny2	29.4M	71,60		
allenai/scibert_scivocab_uncased	110M	69,04		
<b>scirus-tiny</b>	<b>23M</b>	<b>67,92</b>		
nreimers/MiniLM-L6-H384-uncased (e5-small-v2 pretrain)	33.4M	65,68		

 SOTA  
(In-Train)

- **Топ-3 в SciDocs и Out-of-Train** (конкуренты в ~5 раз больше), SOTA в In-Train

# Выводы по результатам сравнения моделей

## 1. Размер и качество модели в сравнении с SciNCL

- меньше параметров: 23М против 110М
- меньше размерность эмбедингов: 312 против 768
- больше контекст: 1024 против 512
- сопоставимое качество (SciDocs Avg): 90.10 против 91.03

## 2. Контрастивное дообучение на парах title-abstract

- улучшает все метрики, особенно кросс-языковой поиск

## 3. Контрастивное дообучение на парах cite / cocite

- компенсирует недостаточность кросс-языковых данных

## 4. Open Source

- бенчмарк интегрирован в MTEB (Multilingual Text Embedding Benchmark)



Данные на Huggingface 🙌

---

Герасименко Н.А., Ватолин А.С., Янина А.О., Воронцов К.В. SciRus: легкий и мощный мультязычный энкодер для научных текстов. Доклады РАН, 2024

Ватолин А.С., Герасименко Н.А., Янина А.О., Воронцов К.В. RuSciBench: открытый бенчмарк для оценки семантических векторных представлений научных текстов на русском и английском языках. Доклады РАН, 2024.

K.Enevoldsen, ..., A. Vatolin et.al. MMTEB: Massive Multilingual Text Embedding Benchmark. 2025.

# Первое внедрение (2024)



«Разработанная в рамках данного проекта модель уже широко используется в **Научной электронной библиотеке** для решения целого ряда задач, связанных с оценкой тематической близости научных документов. Уже протестирован специалистами полезный сервис для ученых, позволяющий *для заданной статьи или подборки статей найти тематически похожие документы*, как среди всего массива [eLIBRARY.RU](https://elibrary.ru) (более 55 млн. научных публикаций), так и только среди новых поступлений. Важной для нас особенностью данной модели является ее мультязычность, поскольку **Научная электронная библиотека** содержит документы на различных языках.»

— *Геннадий Еременко, генеральный директор НЭБ*

---

Научная электронная библиотека, портал eLIBRARY.RU. Пресс-релиз 24-04-2024: «Открыт поиск близких по тематике публикаций с применением нейросети МГУ для анализа научных текстов.»

[https://elibrary.ru/projects/news/search\\\_similar\\\_publ.asp](https://elibrary.ru/projects/news/search\_similar\_publ.asp)

# Сервис полуавтоматического реферирования

**Цель:** автоматизировать написание реферата по подборке, попутно помогая пользователю освежить и систематизировать свои знания (non-linear reading)

The screenshot displays a web interface for semi-automatic summarization. It is organized into three main panels: **PAPERS**, **RECOMMENDED**, and **SUMMARIZATION**.  
1. **PAPERS**: A 'Collection of papers' list on the left. The selected paper is 'SummaRuNner: A Recurrent Neural Network based...' by Ramesh Nallapati, Feifei Zhai, and Bowen Zhou (dated 13 NOV 2016).  
2. **RECOMMENDED**: A 'Summary' section for the selected paper. It includes a rich text editor with a toolbar and a 'Promoters' section with buttons for 'Annotate', 'Idea', 'Theory', 'Method', 'Citation', 'Dataset', 'Experiment', 'Result', and 'Conclusion'.  
3. **SUMMARIZATION**: A 'Recommended phrases' section on the right, containing three paragraphs of text extracted from the paper. Red arrows show the flow from the paper list to the summary and then to the recommended phrases.



**Суфлёры-экстракторы про:**

- проблему, идею,
- теорию, метод, модель,
- эксперимент, датасет,
- результаты, выводы,
- достоинства, недостатки,

...

А. Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

С. Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

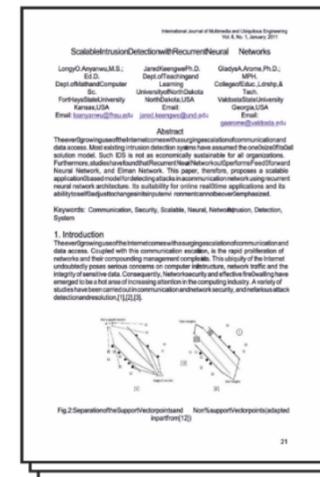
# Сервис тематизации

**Цель:** «разложить по полочкам» подборку — выявить тематическую кластерную структуру, дать представление о каждой теме

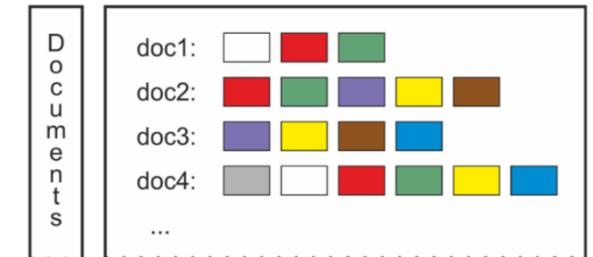
**Тематическая модель** определяет для каждой темы:

- слова, термины, фразы,
- название и суммаризацию для каждого документа:
- состав тем и сегментацию по темам для подборки в целом:
- визуализацию кластерной структуры
- тематический спектр, иерархию

Text documents

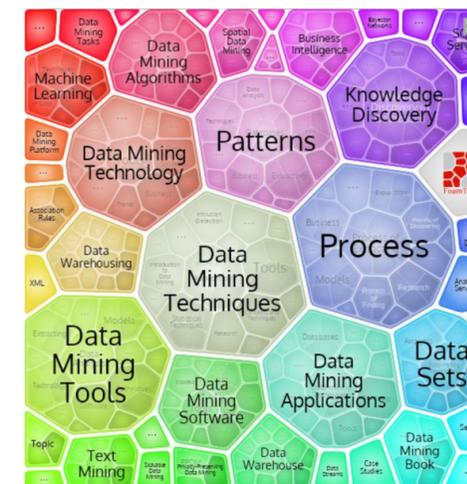


Topics of documents



Topic Modeling

Words and keyphrases of topics



Воронцов К.В. Вероятностное тематическое моделирование: Теория регуляризации ARTM и библиотека с открытым кодом BigARTM. — Москва: URSS. — 2025. — 224 с..

<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

# Сервис поиска и анализа трендов

**Цель:** автоматизировать выявление в подборке новых научных тем, момента их появления, терминологии, интервала роста

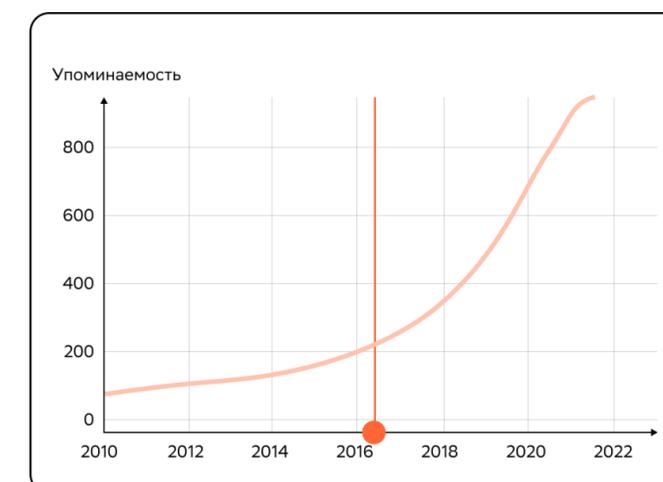
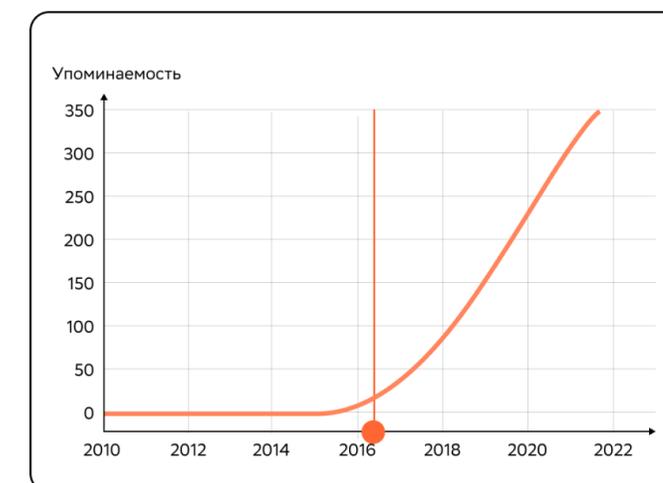
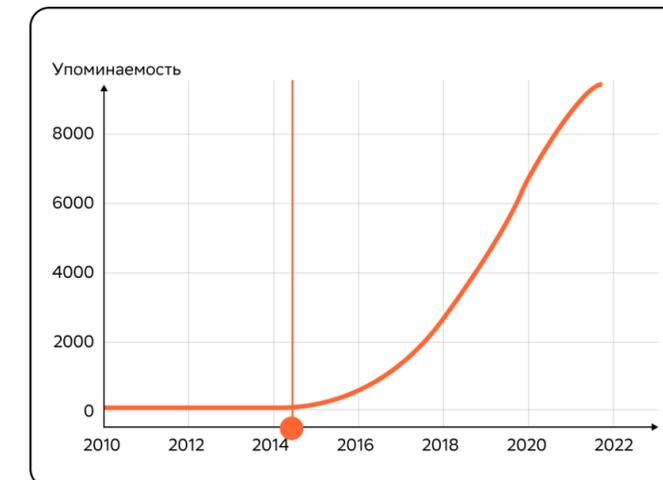
*Темпоральная тематическая модель* дообучается без учителя (без размеченных данных) последовательно на месячных интервалах

**Для валидации модели** экспертами отобраны 87 трендовых тем из области Data Science

**Результат:** >60% тем детектируются в течение года после появления темы

---

*Герасименко Н. А., Чернявский А. С., Никифорова М. А., Никитин М. Д., Воронцов К. В.*  
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях // Доклады РАН. 2022.



# Сервис хронологизации

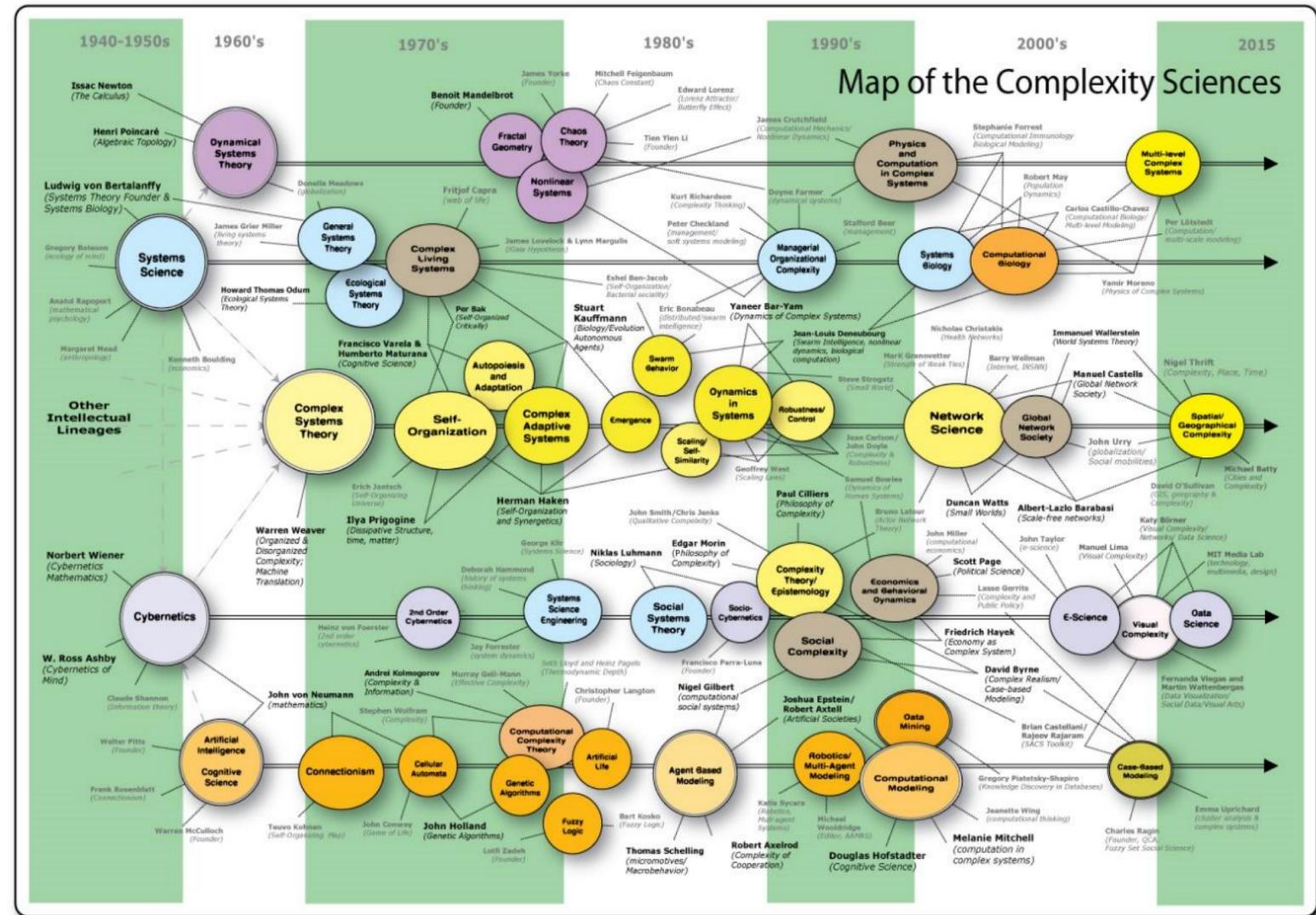
**Цель:** показать развитие во времени основных долгосрочных тем подборки, обозначив ключевые вехи, идеи и их авторов

## Трёхуровневая тематическая иерархия:

- научные направления
- научные теории
- научные школы и учёные

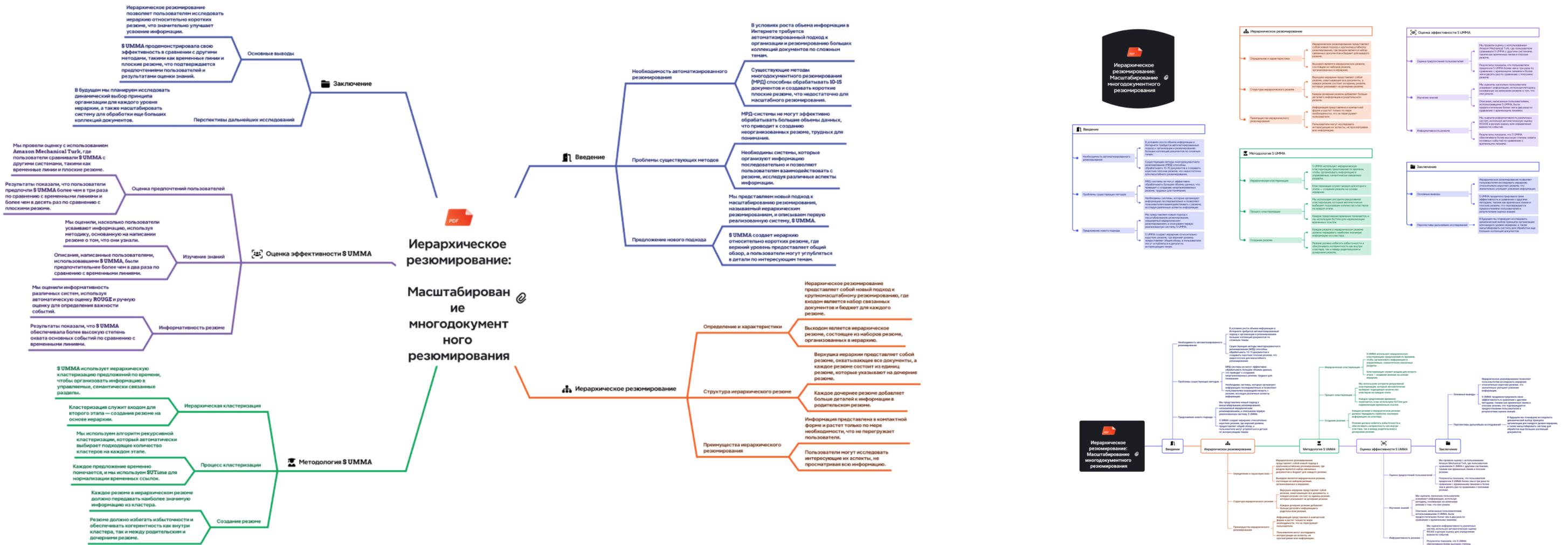
## Оси на карте:

- время × спектр тем
- читабельность,
- релевантность и др.



# Сервис картирования

**Цель:** автоматизировать структурирование знаний в виде mind-map, путём выделения главных идей и разделения их на главные подыдеи иерархически



# От интеллект-карт (mind-maps) к картам знаний



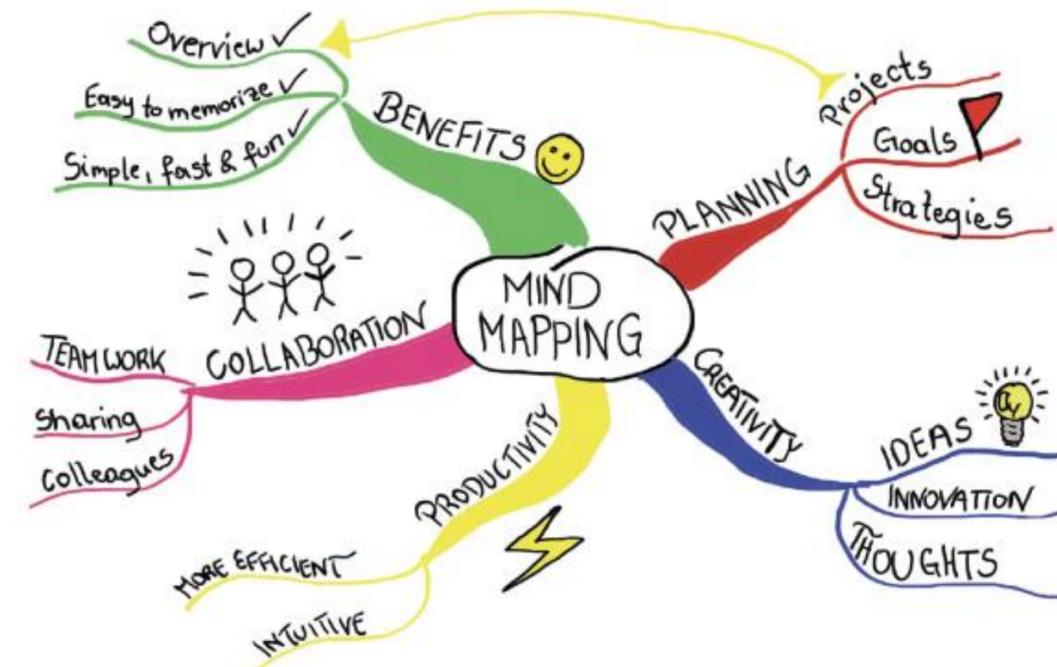
## Интеллект-карты (mind maps)

текстографическое отображение того, как темы (мысли, идеи) разбиваются на подтемы иерархически

максимально близкое к тому, как мы храним знания у себя в головах



предложены в 70-е годы британским **психологом** Тони Бьюзеном



# От интеллект-карт (mind-maps) к картам знаний



нацелены на  
повышение  
эффективности

конспектирования

понимания

запоминания

систематизации

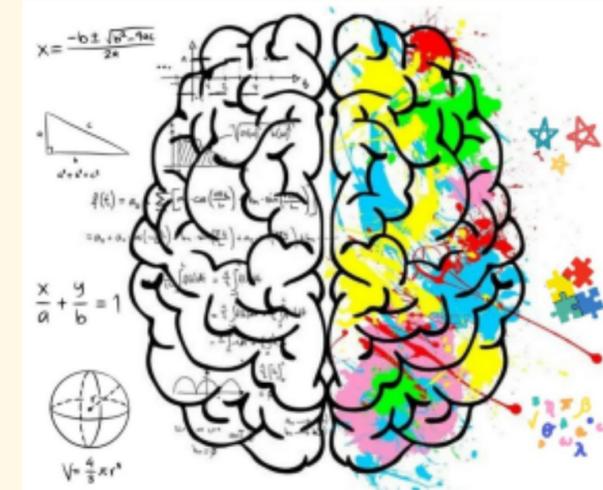
поиска консенсуса



техника  
запоминания

посмотреть, понять, обсудить, договориться, принять

самостоятельно воспроизвести через  
10 минут → сутки → неделю → месяц



благодаря активизации обеих  
полушарий мозга, учёта особенностей  
восприятия, мышления, памяти

# От интеллект-карт (mind-maps) к картам знаний



16 принципов построения интеллект-карт



графическое оформление

для активации зрительной памяти

*радиантность*: линии расходятся из центра

*размер шрифта* отражает важность тем и подтем

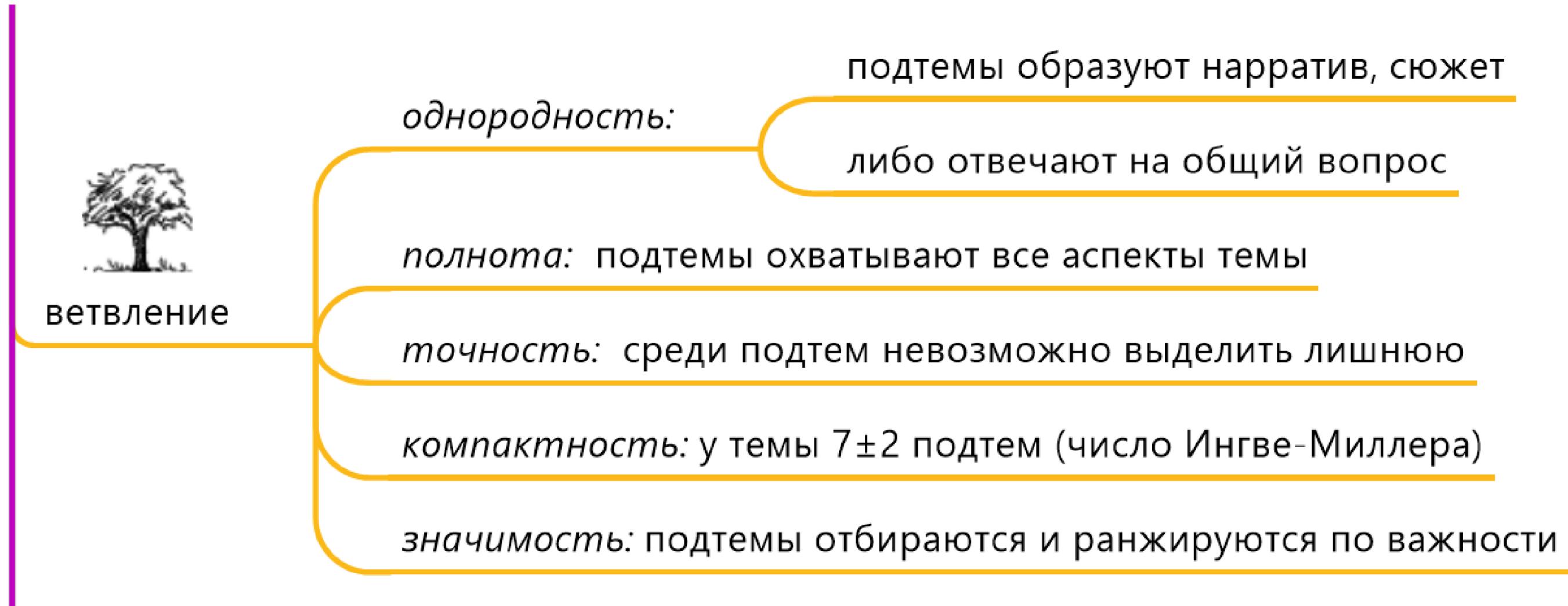
*цвет* выделяет поддеревья

*картинки* усиливают образность

*дополнение* связями, выносками, ссылками

# От интеллект-карт (mind-maps) к картам знаний

(16 принципов построения интеллект-карт)



# От интеллект-карт (mind-maps) к картам знаний

(16 принципов построения интеллект-карт)



эргономика

*наглядность:* фразы подкрепляются изображениями

*лаконичность:* темы формулируются максимально кратко

*обозримость:* карту понимают и запоминают целиком



эстетика

*красота, живость:* эмоции способствуют запоминанию

*гармоничность:* впечатление целостности, сложности карты

*сбалансированность:* ветви примерно равны и равноценны

# От интеллект-карт (mind-maps) к картам знаний



6 принципов, усиливающих интеллект-карты до **карт знаний**



(1) читабельность

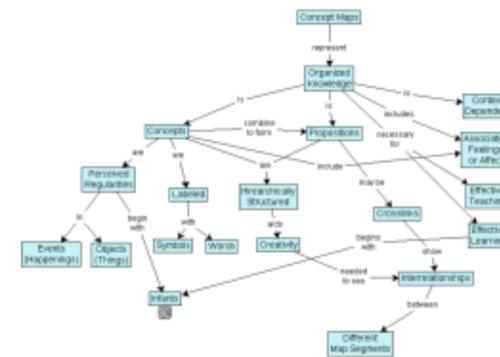
компромисс с лаконичностью и обзорностью

любой фрагмент карты читается как нарратив

легко и однозначно

даже автоматически

в отличие от других способов представления знаний

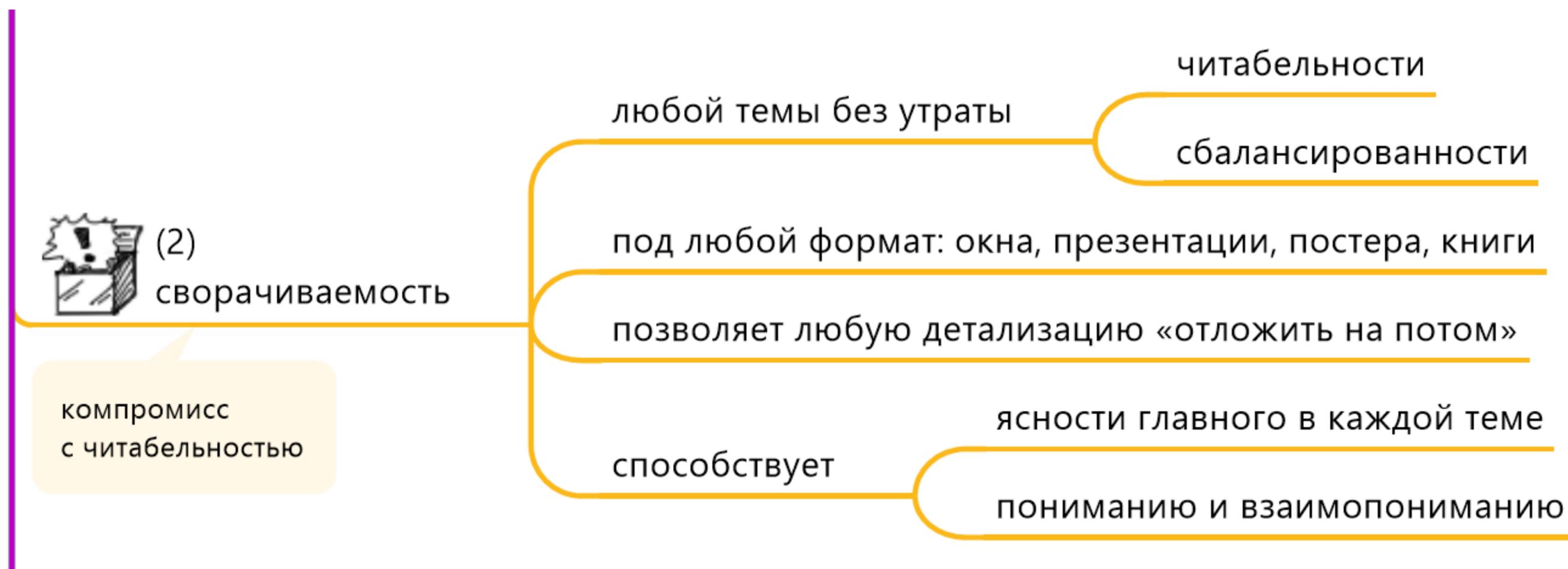


онтологий

фреймов и др.

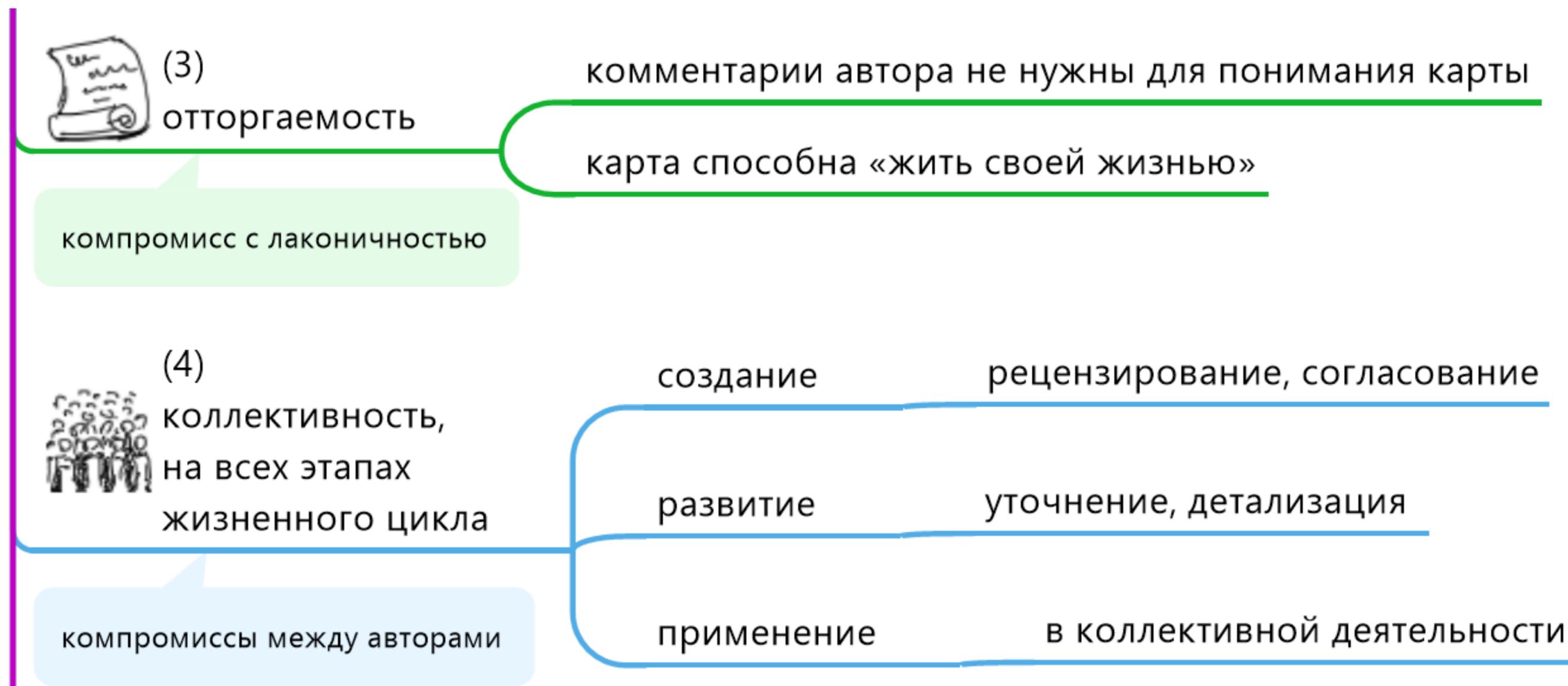
# От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до карт знаний)



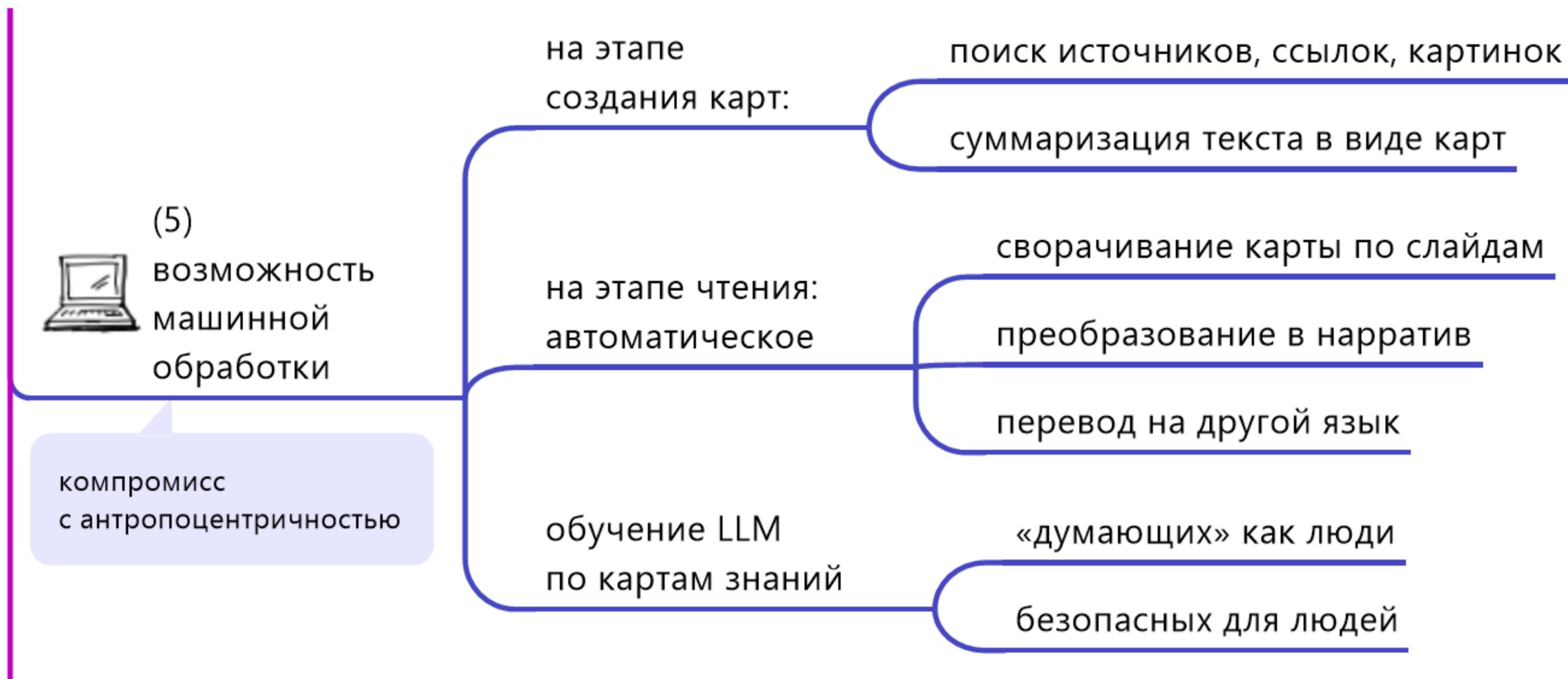
# От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до карт знаний)



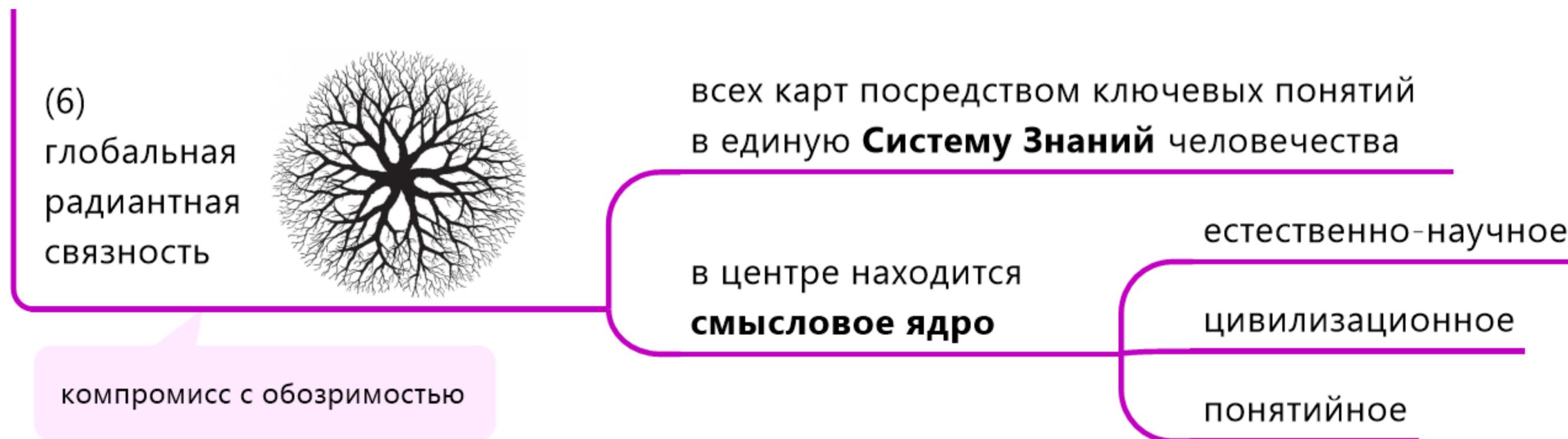
# От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до карт знаний)



# От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до карт знаний)



# Подытожим. Карты знаний

- **Представление знаний**, универсальное для человека и машины
  - 16+6 принципов реализуются через промпты для LLM
- **Перспективный инструмент «коллективного разума»**, развивающий навыки работы с научной информацией:
  - во всём выделять главное ( $7\pm 2$ ),
  - делать это быстро, формулировать лаконично,
  - достигая в команде единства понимания <sup>77</sup>целей, идей, смыслов
- **Обучение ИИ по тексто-графическим представлениям** потребует:
  - освоить картирование знаний (индивидуально и в коллективе)
  - накапливать обучающие выборки и бенчмарки



## Как активировать визуальное аналитическое мышление (эволюционно обусловленное, намного более мощное)

**1** порядка сотни карт: просмотреть, обсудить, поспорить, принять

**2** десятки карт: построить самому, следуя 16+6 принципам

**3** испытать «моменты ясности»,  
инсайты, когда карта



индивидуальная практика и опыт

«красиво сложилась»

привела к согласию

легко и ярко запомнилась,

легла в основу деятельности

**4** сделать построение карт регулярной  
профессиональной практикой



индивидуальной

коллективной

# Подытожим. Мастерская знаний

**Мастерская знаний — масштабная адаптивная концепция:**

пусть «мастерских знаний» будет много и разных

**Миссия:** устранять барьеры между человеком и знанием,  
не пытаюсь передать интеллектуальный труд человека машине

**Реализация:** в основном экстрактивные методы  
(LLM-энкодеры, векторный поиск, ранжирование, тематические модели),  
генеративные LLM — по принципу «минимальной достаточности»

**Конец технократии?** Пора проектировать информационные системы  
не просто как инженерно-технические — «сделал потому, что мог»,  
а как социально-технические (во благо людей и человеческой цивилизации),  
предвидя возникающие социальные практики и долгосрочные эффекты

# Антропоцентричное определение ИИ

*Искусственный интеллект* —

вычислительные технологии,  
создаваемые для повышения  
производительности созидательного  
интеллектуального труда людей

**не** замена человека

**не** «загадочный новый тип разума»

**не** повод уподобиться Богу, чтобы  
«творить по образу и подобию Своему»



# Спасибо за внимание!



## ***Воронцов Константин Вячеславович***

д.ф.-м.н., профессор РАН,  
зав. кафедрой ММП ВМК МГУ,  
зав. лаб. МОСА Института ИИ МГУ,  
зав. кафедрой ИС и кафедрой МОЦГ МФТИ,  
г.н.с. ФИЦ «Информатика и управление» РАН

[k.vorontsov@iai.msu.ru](mailto:k.vorontsov@iai.msu.ru)

<http://www.MachineLearning.ru/wiki?title=User:Vokov>

## **Научный семинар ИПУ РАН**

«Проблемы управления знаниями»

руководители:

академик РАН Д.А.Новиков,

проф. РАН К.В.Воронцов

