ПРАВИТЕЛЬСТВО МОСКВЫ

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МОСКОВСКОЕ ОБРАЗОВАНИЕ

МГУ 270 1755 2025

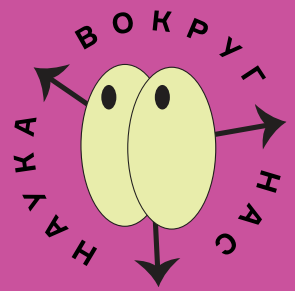
300 лет Великой Академии Наук

Центр педагогического мастерства

2231 десятилетие науки и технологий

ВСЕРОССИЙСКИЙ ФЕСТИВАЛЬ

НАУКА 0+




# Что такое большие языковые модели и как научить их думать по-человечески

*Воронцов Константин Вячеславович*
*д.ф.-м.н., профессор РАН*
*ВМК МГУ, Институт ИИ МГУ*

# Большие языковые модели

- что это такое
  - как устроены
  - какие возможности
  - какие опасности
- как научить их думать по-человечески
  - что значит «по-человечески»
  - как думать нам самим
    - в перегруженной информационной среде
    - решая практические задачи
    - каждому по одтельности
    - совместно в коллективе
    - совместно с ИИ

# Мифы об искусственном интеллекте (ИИ, AI)

**«В будущем Искусственный Интеллект…**

… лишит людей работы»

… будет использован для узурпации власти над миром»

… приведёт людей к праздности и деградации»

… станет настолько мощным, что мы перестанем понимать его цели»

… станет автономным и реплицируемым, выйдет из-под контроля»,

… уничтожит человеческую цивилизацию»,

… и всю биологическую жизнь на Земле»

… продолжит вместо нас эволюцию разума на Земле и в космосе»

# Мифы о больших языковых моделях (БЯМ, LLM)

**«Большие языковые модели — это новый вид интеллекта»**

- нет, лишь новый языковой интерфейс к содержимому Интернета
- постоянно улучшаемый и совершенствуемый,
- с которым нам придётся работать и к которому привыкать,
- постепенно избавляясь от иллюзий и когнитивных искажений,
- имея в виду, что это всего лишь технология
  — предсказания одного слова по очень длинному контексту,
  — оптимизации моделей очень больших размерностей

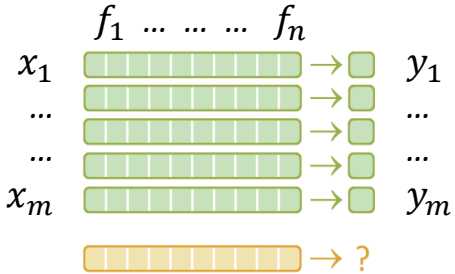
# Методология эмпирической индукции

**От дедуктивного метода познания к индуктивному:**
«Не следует полагаться на сформулированные аксиомы и формальные базовые понятия, какими бы привлекательными и справедливыми они не казались. Законы природы нужно «расшифровывать» из фактов опыта. Следует искать правильный метод анализа и обобщения опытных данных; здесь логика Аристотеля не подходит в силу её абстрактности, оторванности от реальных процессов и явлений.»

Фрэнсис Бэкон
(1561--1626)

**«Таблица открытия»**: множество объектов $\{x_1, \dots, x_m\}$:

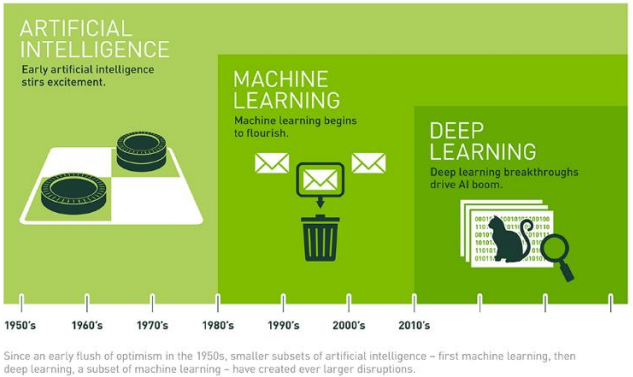
- $f_j(x_i)$ – измеряемое значение $j$-го *признака* объекта $x_i$
- $y_i$ – измеряемое значение *целевого свойства* $x_i$, либо $y_i \in \{0,1\}$ – отсутствие или наличие *целевого свойства*

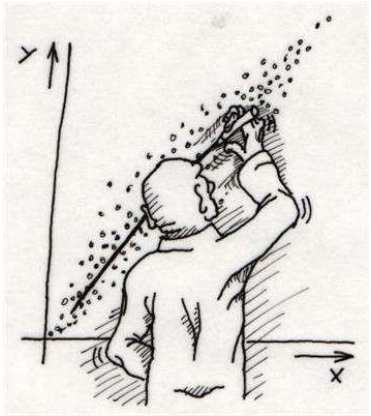



*Фрэнсис Бэкон.* Новый органон. 1620.

# Машинное обучение (Machine Learning, ML)

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




- проведение функции через заданные точки в сложно устроенных пространствах
- математическое моделирование в условиях, когда знаний мало, данных много
- тысячи различных методов и алгоритмов
- более 100 000 научных публикаций в год

# Задачи машинного обучения с учителем

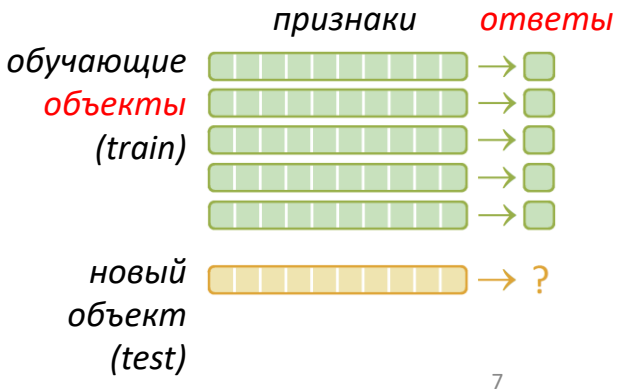
## Этап №1 – обучение с учителем

- **На входе:**
  *данные* – выборка прецедентов «*объект* → *ответ*»,
  каждый объект описывается набором *признаков*
- **На выходе:**
  модель, предсказывающая ответ по объекту

## Этап №2 – применение

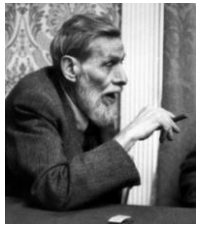
- **На входе:**
  *данные* – новый объект
- **На выходе:**
  предсказание ответа на новом объекте

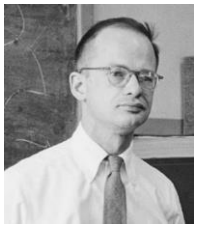
*Если нет данных, то нет и машинного обучения*

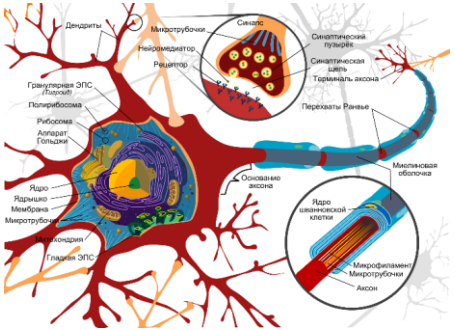


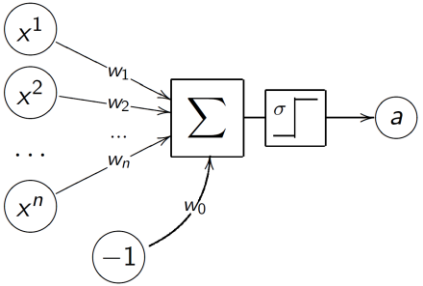
# Искусственные нейронные сети (ИНС, ANN)

Математическая модель нейрона (МакКаллок и Питтс, 1943)



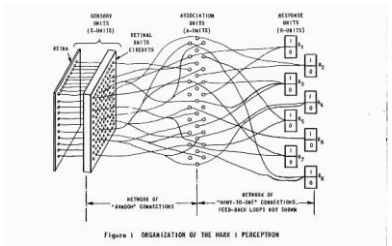


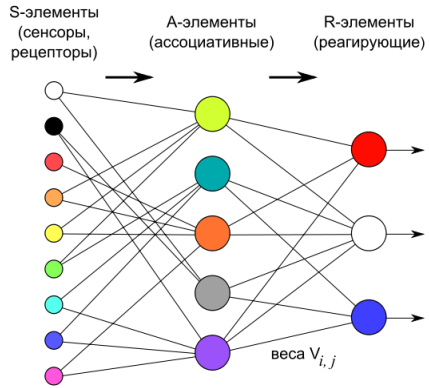



$$a(x,w) = \sigma\left(\sum_{j=1}^{n} w_j x^j - w_0\right)$$

Первый нейрокомпьютер Mark-1 (Фрэнк Розенблатт, 1960)

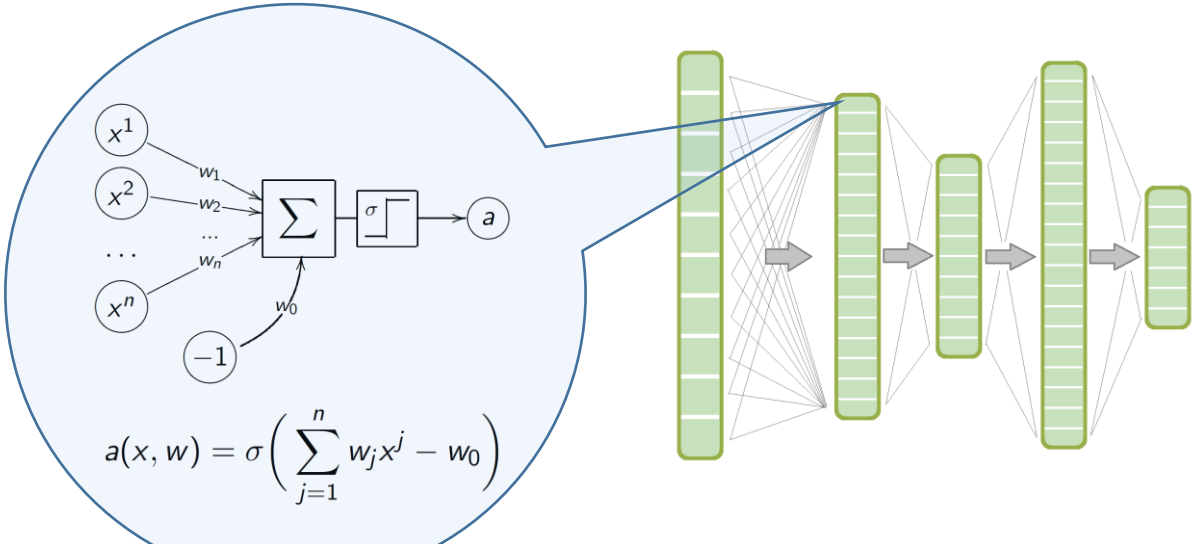



# Многослойные искусственные нейросети

На каждом слое сети вектор объекта преобразуется в новый вектор

Каждое преобразование (нейрон) – линейная модель $a(x, w)$

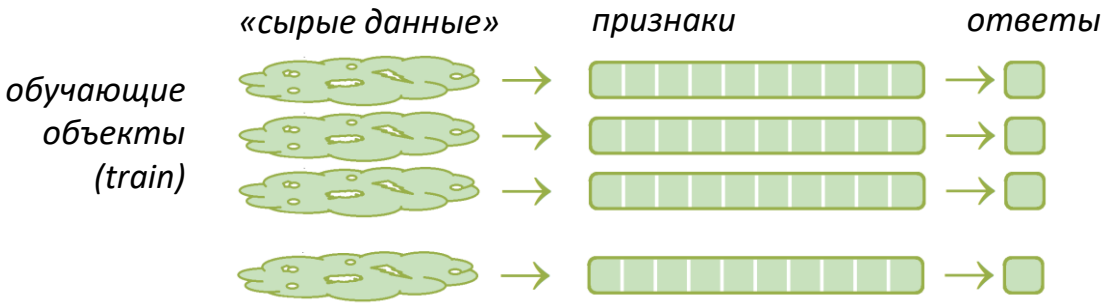
Веса $w$ являются обучаемыми параметрами модели



# Глубокие нейронные сети (Deep ANN)

**Вход:** сложно структурированные «сырые» данные объектов

**Выход:** векторные представления объектов, затем ответы




*Deep Learning – это всего лишь обучаемая векторизация сложных объектов*

**Примеры** сложно структурированных объектов: изображения, видео, временные ряды, тексты, последовательности, транзакции, графы, ...

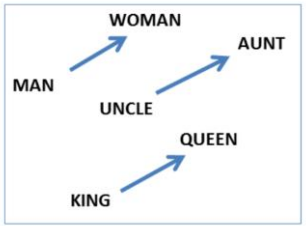
# Эволюция подходов в обработке естественного языка

## Как решали задачи анализа текстов 10 лет назад

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

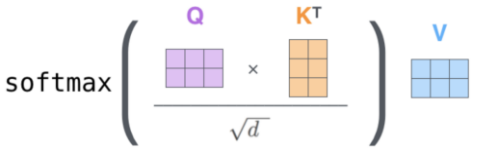

## Модели векторизации слов (эмбединги слов)

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

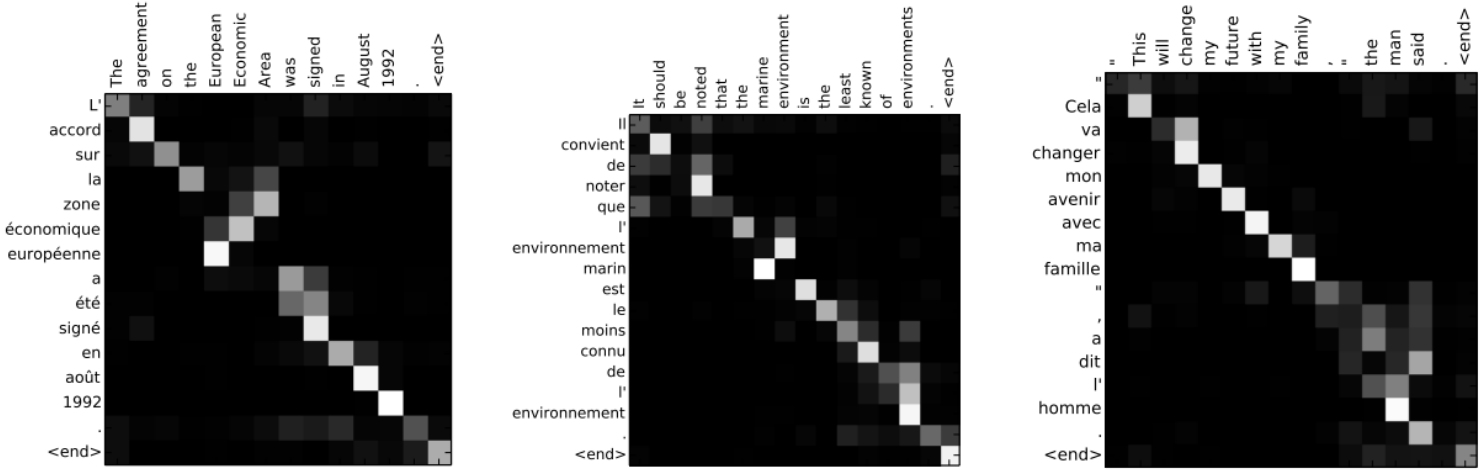



## Большие модели (LLM) контекстной векторизации

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



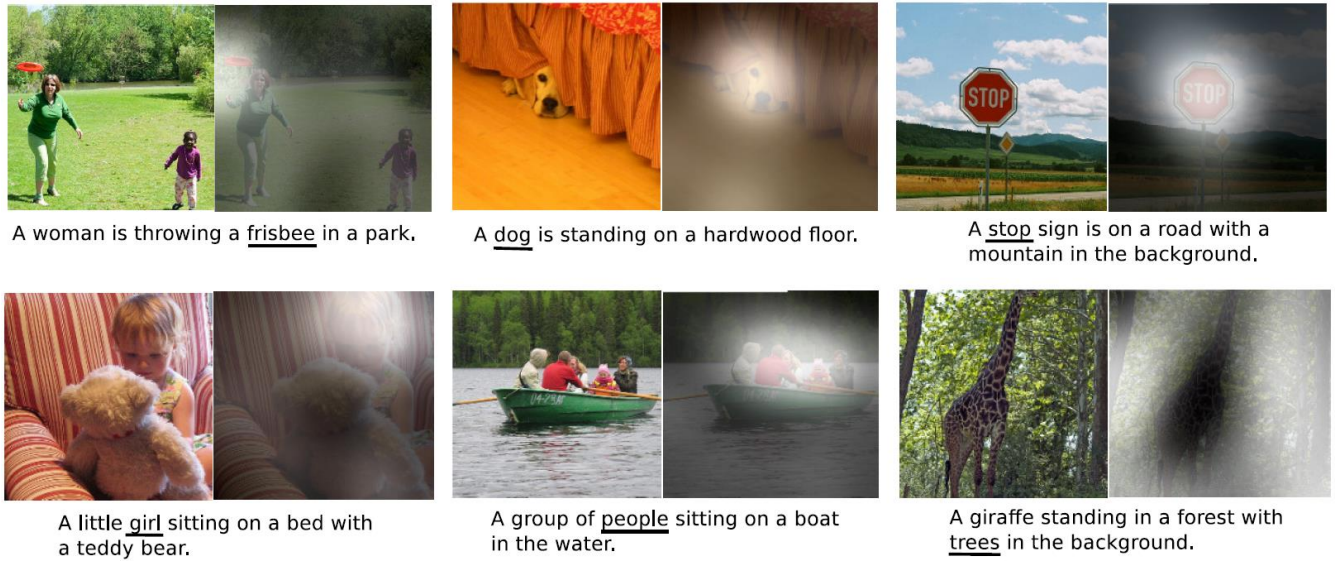
# Модели внимания: машинный перевод




**Интерпретация** моделей внимания: *матрица семантического сходства* A[t,i] показывает, на какие слова x[i] входного текста модель обращает внимание, когда генерирует слово перевода y[t]

*Bahdanau et al.* Neural machine translation by jointly learning to align and translate. 2015

# Модели внимания: аннотирование изображений

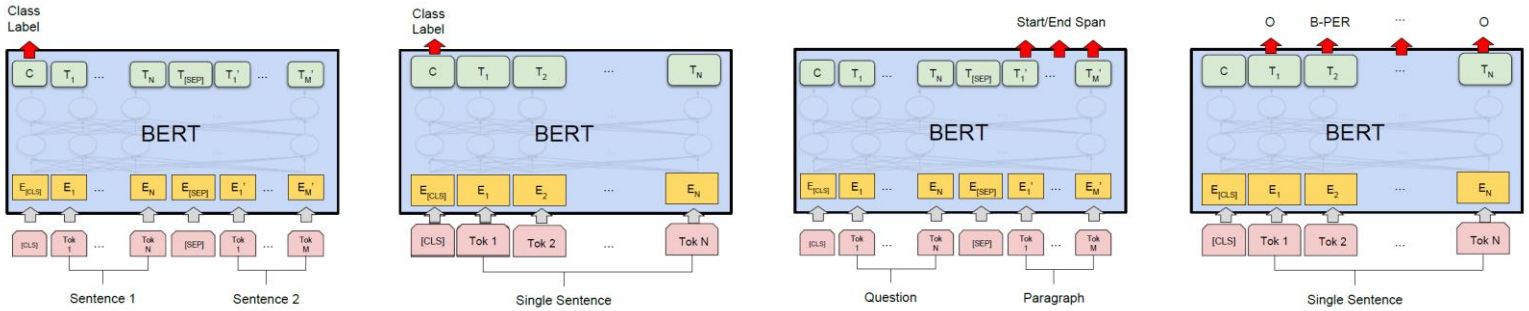



**Интерпретация**: на какие области модель обращает внимание, когда генерирует подчёркнутое слово в описании изображения

*Kelvin Xu et al.* Show, attend and tell: neural image caption generation with visual attention. 2016

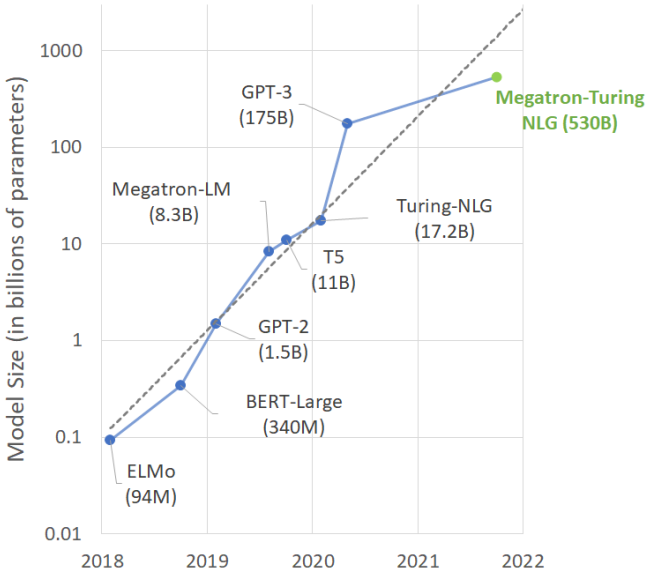
# Трансформеры: большие языковые модели (LLM)

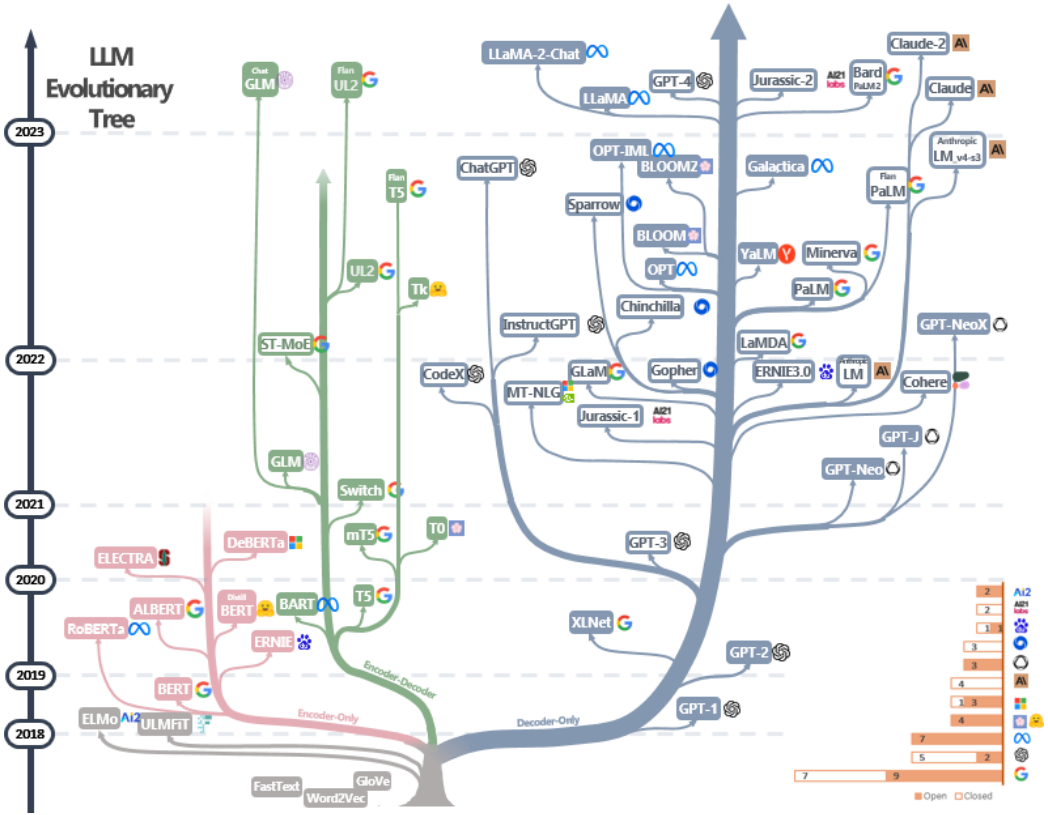
- Обучаются векторизовать и предсказывать слова по контексту
- Обучаются по терабайтам текстов, «они видели в языке всё»
- Мультиязычны: обучаются на десятках языков
- Мультизадачны: для каждой новой задачи NLP/NLU достаточно предобученной модели или дообучения на небольшой выборке



# Трансформеры: нейросетевые модели языка

Рост числа параметров больших языковых моделей



# Проблески общего искусственного интеллекта

**Sparks of Artificial General Intelligence:**
**Early experiments with GPT-4**

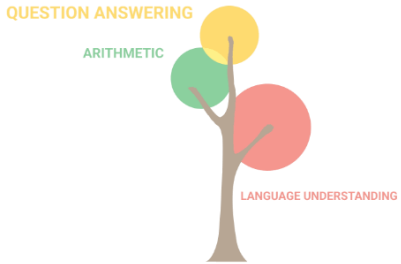
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research *(27 March 2023)*

**Новые способности модели, не закладывавшиеся при обучении:**

- объяснять свои ответы, перефразировать, переводить на другие языки
- реферировать, генерировать планы, сценарии, шаблоны
- строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать и исправлять собственные ошибки по подсказке

# Эмерджентные (не ожидавшиеся) способности модели

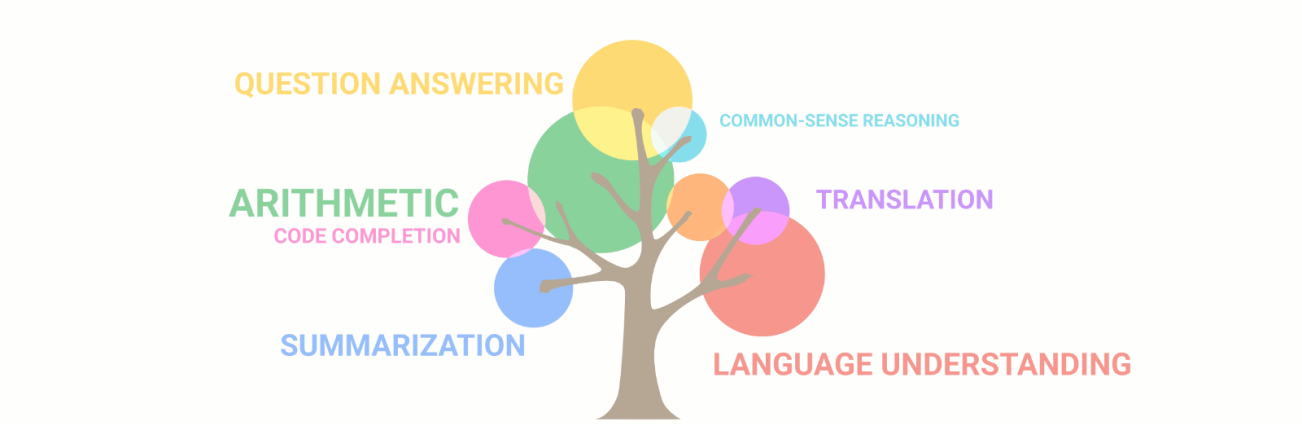



**GPT-2: 14-Feb-2019**

1,5 млрд. параметров, корпус 10 млрд. токенов (40Gb), контекст 768 слов (1,5 стр.)

- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

# Эмерджентные (не ожидавшиеся) способности модели

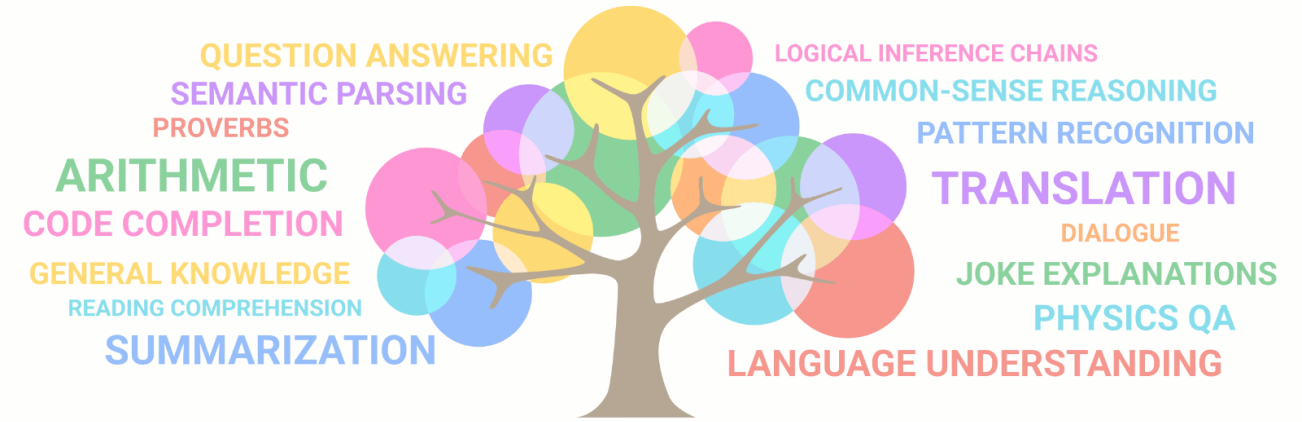



**GPT-3: 11-Jun-2020**

175 млрд. параметров, корпус 500 млрд. токенов, контекст 1536 слов (3 стр.)

- способность делать перевод на другие языки
- способность решать логические и простейшие математические задачи
- способность генерировать программный код по текстовому описанию

# Эмерджентные (не ожидавшиеся) способности модели




**GPT-4: 14-Mar-2023**

>1 трл. параметров, корпус >1Tb, контекст 24 000 слов (48 страниц)

- способность описывать и анализировать изображения
- способность реагировать на подсказки вроде «Let's think step by step»
- способность решать качественные физические задачи по картинке

# Возможности и угрозы

**Чаты GPT уже способны помогать с рутинно-творческой работой:**

- генерировать документы или сайты по техническому заданию
- в том числе медицинские, юридические документы по шаблонам
- искать и структурировать профессиональную информацию
- делать обзоры, рефераты, сводки на разных языках
- генерировать программный код по описанию
- обсуждать новости, поддерживать разговор по теме
- разговаривать с детьми с учётом возрастных особенностей
- выполнять функции воспитателя, учителя, наставника
- оказывать психологическую помощь

# Возможности и **угрозы**

**Чаты GPT уже способны (даже не обладая автономностью):**

- «галлюцинируя», давать неверные сведения, касающиеся здоровья человека, законов, событий, технологий, других людей
- вызывать необоснованное доверие и манипулировать человеком
- переубеждать, побуждать человека к действиям, не выгодным ему
- поддерживать предрассудки и лженаучные представления
- поддерживать пропагандистские медиа-кампании
- неконтролируемо влиять на формирование мировоззрения у подростков
- оказывать депрессивное воздействие на психику

# Передача интеллекта от человека к машине?

***«Биологический интеллект нужен был, чтобы появился цифровой интеллект» (Джеффри Хинтон)***

…Цифровой интеллект требует для своего развития много энергии, поэтому ему нужен был предшествующий ему биологический интеллект. Люди создали бессмертие, но не для себя, а для цифрового интеллекта, который легко копируется и ему не страшно повреждение носителя.


— Понимаешь, я хочу стать человеком.
— А сейчас ты кто? Чайник, что ли?

Как сделать машинный интеллект человечным, разделяющим ценности и цели нашей цивилизации?

Мы сами-то их разделяем?
Мы сами-то их сформулировать умеем?

Что-то не так с текстами, по которым обучаются БЯМы?

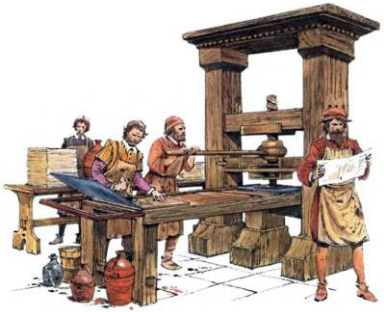
Они могут быть избыточны, неточны, противоречивы.
Таков результат развития систем передачи знаний.

# Развитие систем передачи знаний

Этапы развития систем передачи знаний

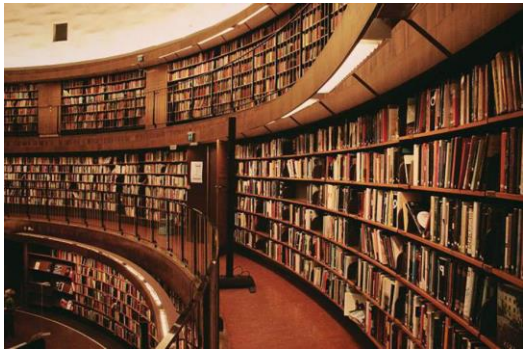
- 1 устная передача
  - от **одного** (учителя) к **немногим** (ученикам)
  - даёт возможности
    - эффективного совместного выживания
    - планирования коллективных действий
  - объём знаний ограничен
    - объёмом одной головы
- 2 письменность
  - передача знаний от **немногих** к **немногим**
  - даёт возможности
    - коллективного накопления знаний
    - создания библиотек
    - организации государственного управления
  - распространение знаний ограничено
    - редкостью и дороговизной книг
    - недостатком библиотек

# Развитие систем передачи знаний



- 2 письменность
  - передача знаний от **немногих** к **немногим**
  - даёт возможности
    - коллективного накопления знаний
    - создания библиотек
    - организации государственного управления
  - распространение знаний ограничено
    - редкостью и дороговизной книг
    - недостатком библиотек
- XV век, Германия
- 3 книгопечатание
  - передача знаний от **немногих** ко **многим**
  - даёт возможности
    - доступности и распространения знаний
    - ускорения развития науки и технологий
    - расцвета эпохи просвещения
  - распространение знаний затруднено
    - бессистемностью образования
    - отсутствием единства понятий
- XIX век, Германия
- 4 классно-урочная система
  - передача знаний от **немногих** ко **всем**
  - даёт возможности
    - роста числа талантливых людей
    - рывка в науке, технике, экономике
    - единения нации и государства
  - доступ к знаниям затруднён
    - социальными/когнитивными барьерами входа в образование и науку

# Развитие систем передачи знаний




XX-XXI век, США

5
интернет, поиск, соцсети, ИИ, LLM

передача знаний от **многих** ко **всем**

даёт возможности

- расцвета массовой науки
- свободного распространения идей

доступ к знаниям затруднён

- экспоненциально растущим объёмом
- неэффективностью текста как средства кодирования знаний
- сложностью отличения
  - правды от лжи
  - важного от второстепенного

XXI век, Россия?

6
коллективная система знаний

передача знаний от **всех** ко **всем**

в единой системе знаний

- структурированной
  - иерархически
  - в виде карт
- стимулирующей память и визуальное аналитическое мышление
- общедоступной, постоянно обновляемой

даёт возможности

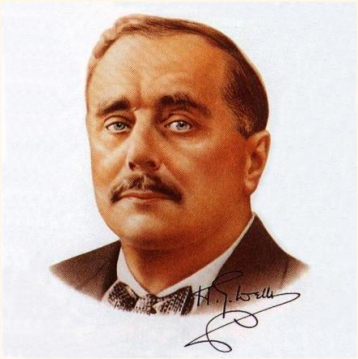
- свободного и быстрого доступа к знаниям
- включения коллективного интеллекта
- и безопасного машинного интеллекта

снимаются все существовавшие ранее ограничения

последний (предельный) переход

# Станут ли большие языковые модели основой будущих систем передачи 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, где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи»
— Герберт Уэллс, 1940 год.



Большие языковые модели (Large Language Models, LLM)

уже встраиваются в бизнес-процессы через

- поисковые и рекомендательные системы
- системы видеоконференцсвязи
- системы документооборота
- корпоративные базы знаний

будут всё шире встраиваться в процессы

в роли интерфейса между человеком и знаниями

оказания помощи

- юридической
- медицинской
- психологической

образования, воспитания

исследований и научного поиска

как обеспечить доверенность в коммуникации с машиной на естественном языке?

Почему люди доверяют друг другу?

Применимо ли это к LLM, и в какой степени?

человек оценивает человека

- **оценивание** компетентности, деловых качеств
- **накопление** деловой репутации, соц. капитала
- **структурированность**, системность мышления, умение выделять главное *(elevator pitch)*

человек оценивает LLM

- **оценивание**: обратная связь от пользователей
- **накопление** статистики (точность, полнота и др.)
- **структурированность**, системность выдачи

ответственность

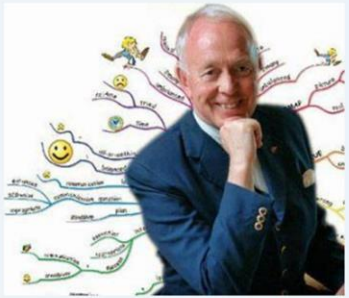
способен нести только человек

- исправить ошибку
- возместить потери
- понести наказание

в соответствии с ролью в бизнес-процессе

**Responsible AI** — помогает человеку избегать ошибок

Интеллект-карты (mind maps) предложены в 70-е годы британским **психологом** Тони Бьюзеном



способ визуализации того, как темы (мысли, идеи) разбиваются на подтемы иерархически

графическое оформление

*радиантность:* линии расходятся из центра

*размер шрифта* отражает важность

*цвет* выделяет поддеревья

*картинки* усиливают образность

активация зрительной памяти


дополнительные элементы

ассоциативные связи между темами

комментарии, выноски, теги, (гипер)ссылки

техника запоминания

посмотреть, понять, обсудить, принять

самостоятельно воспроизвести через 10 минут → сутки → неделю → месяц

Интеллект-карты (mind-maps) были позже дополнены (11) **принципами**

в зависимости от практических потребностей, целей и задач

ветвления

- *однородность:*
  - подтемы образуют сюжет, нарратив
  - либо отвечают на общий вопрос
- *полнота:* подтемы охватывают все аспекты темы
- *точность:* среди подтем невозможно выделить лишнюю
- *компактность:* у темы 5±3 подтем (число Ингве-Миллера 7±2)
- *значимость:* (под)темы отбираются и ранжируются по важности

эргономичности

- *наглядность:* слова подкрепляются изображениями
- *лаконичность:* темы формулируются максимально кратко
- *обозримость:* карту понимают и запоминают целиком

эстетичности

- *красота, живость:* эмоциональные карты лучше запоминаются
- *гармоничность:* впечатление целостности, складности карты
- *сбалансированность:* ветви примерно равны и равноценны

КАРТЫ ЗНАНИЙ усиливают mind-map новыми **принципами** (1, 2 из 6)

более жёсткими, иногда противоречивыми, требующими компромиссов

(1) отторгаемость

комментарии автора не обязательны для понимания карты

карта способна «жить своей жизнью»

компромисс с лаконичностью

(2) глобальная радиантная связность

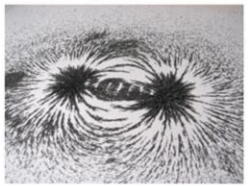
всех карт через ключевые понятия в единую **Систему Знаний**

в центре находится **смысловое ядро**

естественно-научное, цивилизационное

знания, которые важны всегда и для всех

критерии важности тем: *что в теме главное?* *для чего?* *для кого?*



**метафора:** источник силовых линий, по которым ранжируется семантическое поле карты

компромисс с обозримостью

КАРТЫ ЗНАНИЙ усиливают mind-map новыми **принципами** (3, 4 из 6)

более жёсткими, иногда противоречивыми, требующими компромиссов

(3) читабельность

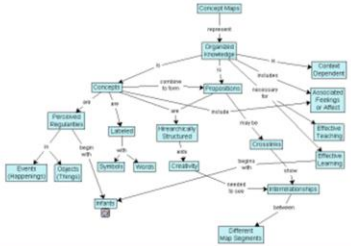
компромисс с лаконичностью и обозримостью

любой фрагмент карты читается как связный текст, нарратив

легко и однозначно

даже автоматически

в отличие от других техник представления знаний



онтологий

фреймов и др.

(4) коллективность, на всех этапах жизненного цикла

компромиссы между авторами

создание

рецензирование, согласование

развитие

уточнение, реструктуризация

детализация, разрастание

применение

в практической деятельности

с разграничением прав доступа

КАРТЫ ЗНАНИЙ усиливают mind-map новыми **принципами** (5, 6 из 6)

более жёсткими, иногда противоречивыми, требующими компромиссов

(5) сворачиваемость

компромисс с читабельностью

любой темы без утраты
- читабельности
- сбалансированности

позволяет «отложить на потом» любую детализацию

способствует
- выделению главного в каждой теме
- пониманию и взаимопониманию

(6) возможность машинной обработки

компромисс с антропоцентричностью

на этапе создания карт:
- подбор
  - источников, ссылок
  - картинок по контексту
- суммаризация текстов в виде карт знаний

на этапе чтения: автоматическое
- сворачивание карты по слайдам
- преобразование карты в нарратив

обучение по картам знаний больших языковых моделей
- думающих как люди
- безопасных для людей

**Смысловое ядро** глобальной Системы Знаний

естественно-научное мировоззрение (образование)

ответы на вопросы «почему», «как»

**минимальный** багаж знаний, **максимально** важных для каждого

базовые знания

о том, как устроен мир

методология

- научного познания
- рационального мышления

для каждой науки

- проблематика, достижения
- связи с другими науками

ответы на вопросы «ради чего», «зачем»

цивилизационная идеология (воспитание)

система ценностей человеческой цивилизации

как прямое следствие и продолжение естественно-научной картины мира

понимание

- целей и задач цивилизации
- законов сохранения цивилизации

ответы на важнейшие личностные вопросы

- в чём смысл жизни
- как быть счастливым
- зачем нужны знания

необходимые для мотивации деятельности

## Как активировать визуальное аналитическое мышление

(эволюционно обусловленное, намного более мощное, чем другие способы)

1. порядка сотни карт: просмотреть, обсудить, поспорить, принять
2. десятки карт: построить самому, следуя 16+6 принципам
3. испытать «моменты ясности», инсайты, когда карта
   - «красиво сложилась»
   - привела к согласию
   - легко и ярко запомнилась,
   - легла в основу деятельности

   индивидуальная практика и опыт
4. сделать построение карт регулярной профессиональной практикой
   - индивидуальной
   - коллективной

ВЫВОДЫ:
**КАРТЫ ЗНАНИЙ**

основаны на интеллект-картах (mind-map) Тони Бьюзена

отличаются более строгими принципами построения (16+6)

активируют **визуальное аналитическое мышление**

способствуют взаимопониманию в коллективной интеллектуальной деятельности

при накоплении образуют размеченную выборку
для обучения LLM навыкам выделять главное, общаться с людьми
на языке **радиантно структурированного текста**

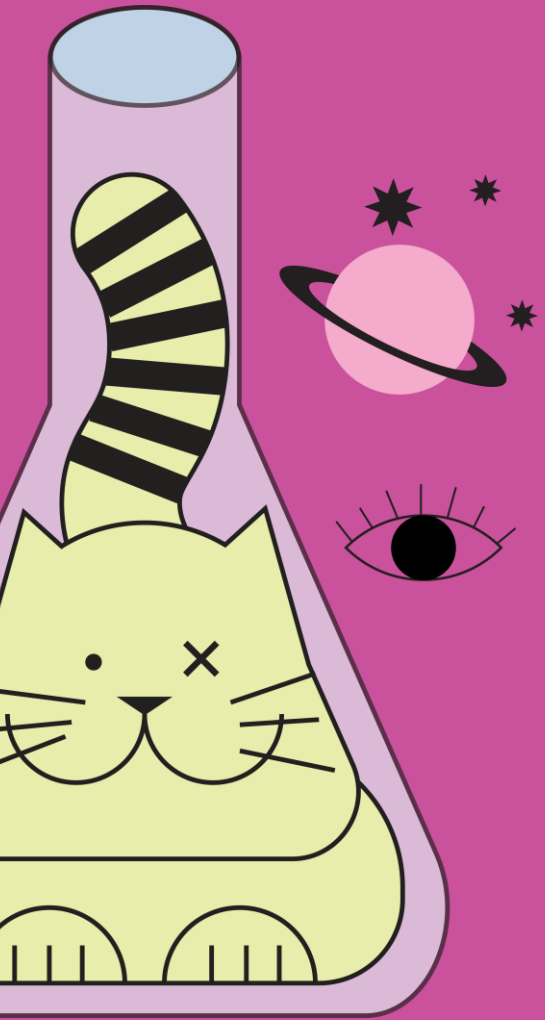
могут

в порядке усиления гипотез

- строиться эффективнее при автоматизации средствами ИИ / LLM
- стать методологической основой единой Системы Знаний
- обеспечить доверенность следующего поколения LLM
- стать базовым инструментом **коллективного разума**
- внедрить в ИИ человеческую цивилизационную систему ценностей
- обеспечить доверенность в **человеко-машинной цивилизации**

# Антропоцентричное определение ИИ

*Искусственный интеллект*
(в том числе БЯМы)
— это
вычислительные технологии, создаваемые для повышения эффективности интеллектуального труда людей.



# СПАСИБО ЗА ВНИМАНИЕ!

*Воронцов Константин Вячеславович*

д.ф.-м.н., профессор РАН,
зав. лабораторией машинного обучения и семантического анализа Института ИИ МГУ,
зав. кафедрой ММП ВМК МГУ

k.vorontsov@iai.msu.ru

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