14-я международная мультиконференция «Биоинформатика регуляции и структуры геномов / системная биология»
5-10 августа 2024 г., Новосибирск

Эволюция идей в искусственном интеллекте и их связь с задачами биомедицины и биоинформатики

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,
зав. кафедрой математических методов прогнозирования МГУ,
рук. лаб. машинного обучения и семантического анализа Института ИИ МГУ
зав. кафедрой интеллектуальных систем МФТИ,
г.н.с. ФИЦ «Информатика и управление» РАН

k.vorontsov@iai.msu.ru

Содержание

1. Задачи машинного обучения
   - Предыстория машинного обучения
   - Терминология машинного обучения
   - Примеры задач машинного обучения
2. Методология машинного обучения
   - Нейронные сети и глубокое обучение
   - Оптимизационные задачи машинного обучения
   - Задачи машинного обучения с векторизацией объектов
3. Большие языковые модели
   - Модели внимания и трансформеры, эмерджентность
   - О некоторых задачах вычислительной биологии
   - Смена парадигмы

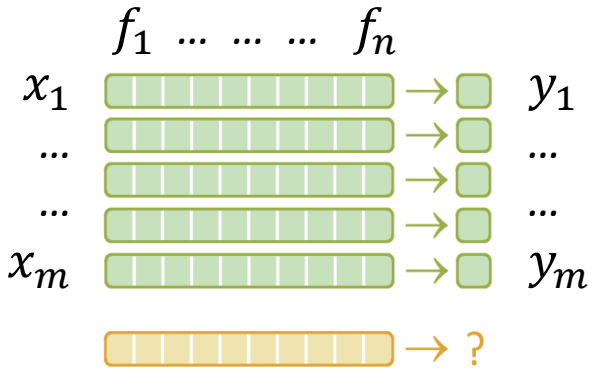
Методология эмпирической индукции

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.

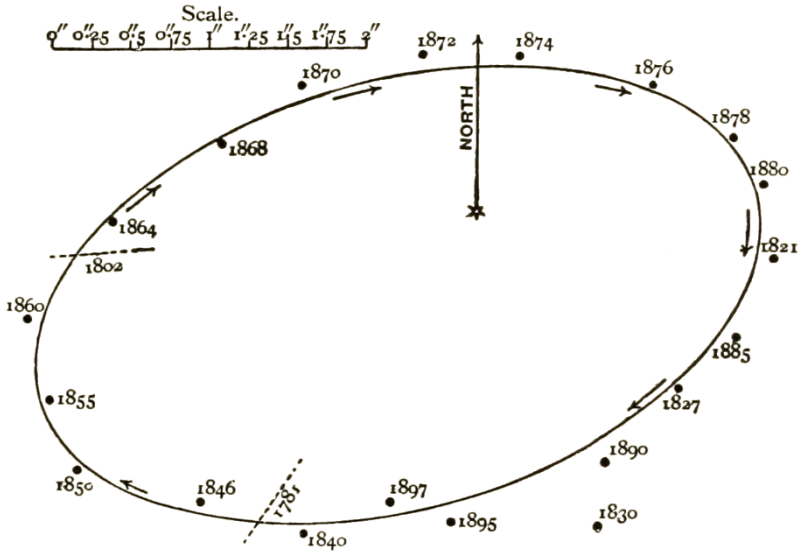
Задача проведения функции через точки

Предсказание свойства $y(x)$ по признакам $f_j(x)$,
(линейной) моделью $a(x, w)$ с параметрами w:

$$a(x, w) = \sum_j w_j f_j(x)$$

Метод наименьших квадратов (Гаусс, 1795):

$$\sum_{(x,y)} (a(x, w) - y)^2 \to \min_w$$



Карл Фридрих Гаусс
(1777--1855)

«Our principle, which we have made use of since 1795, has lately been published by Legendre…»

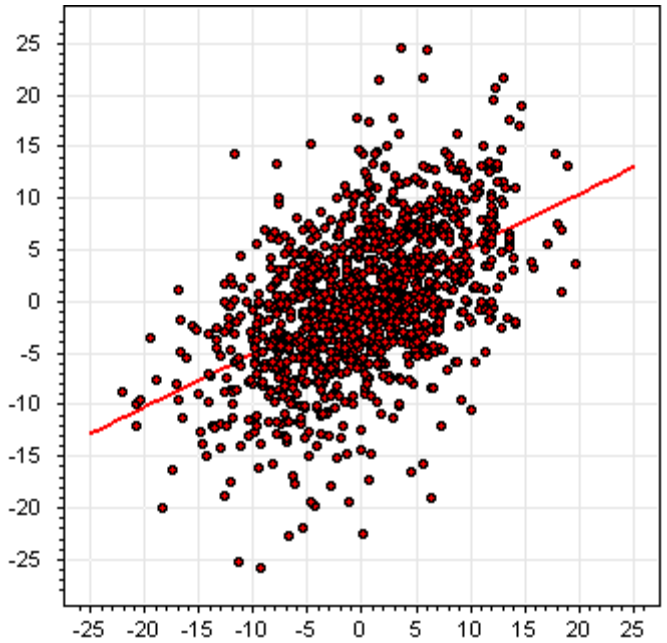
C.F.Gauss. Theory of the motion of the heavenly bodies moving about the Sun in conic sections. 1809.

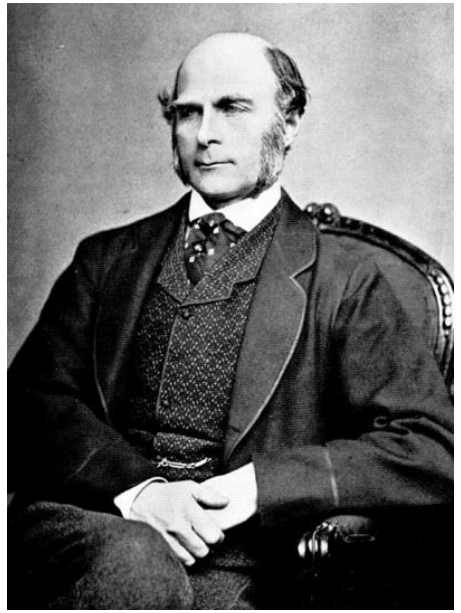
Задача восстановления регрессии

Исследование наследственности роста (Гальтон, 1886).

Δ — отклонение роста от среднего в популяции

Зависимость (линейная?) Δ взрослого сына от Δ отца:





Фрэнсис Гальтон
(1822--1911)

Двойной смысл термина «регрессия»:

- регрессия (роста)
  к среднему в популяции
- *необычный «обратный» ход исследования: сначала данные, потом модель*

Galton F. Regression towards mediocrity in hereditary stature. 1886.

Задачи машинного обучения с учителем

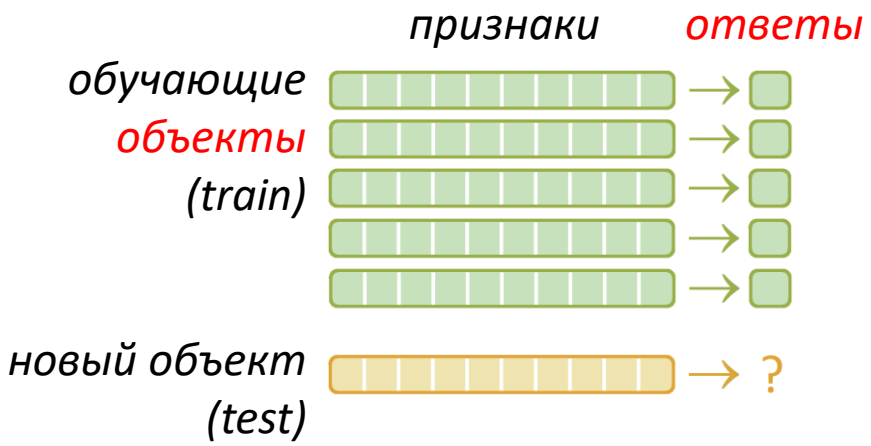
Этап №1 – обучение с учителем

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

Если нет данных,
то нет
и машинного
обучения

Этап №2 – применение

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



Примеры задач машинного обучения

- **Медицинская диагностика:**
  объект – данные о пациенте на текущий момент
  ответ – диагноз / решения о мероприятиях

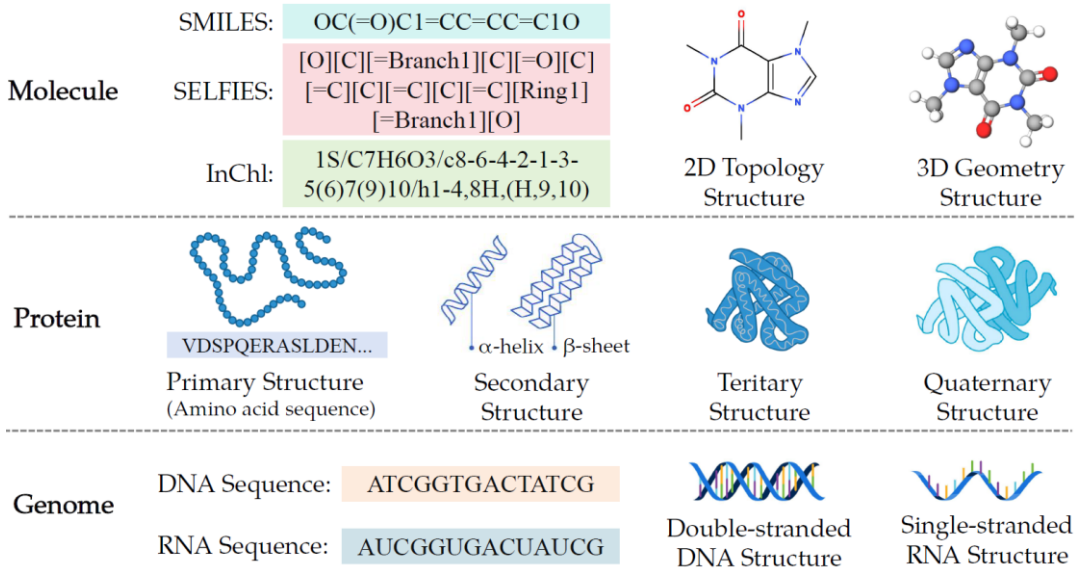
- **Предсказание инфицирования в результате контакта:**
  объект – данные с *носимого устройства* (http://amuleit.ru)
  ответ – вероятность передачи инфекции

- **Предсказание инфицирования по множеству контактов:**
  объект – данные о контактах индивида в интервале времени
  ответ – вероятность инфицирования

Примеры задач в молекулярной биологии

- **Предсказание свойств по структуре:**
  объект – химическая формула
  ответ – свойство вещества

- **Предсказание структуры белка:**
  объект – АК-последовательность
  ответ – 2D, 3D, 4D структуры белка



- **Аннотирование генома:**
  объект – нуклеотидная последовательность
  ответ – разметка участков ДНК: кодирующие, промоторы, энхансеры и др.

Qiang Zhang et al. Scientific Large Language Models: A Survey on Biological & Chemical Domains. 2024-07-23
Serafim Batzoglou. Large Language Models in Molecular Biology. Towards Data Science, 2023-06-02.

Содержание

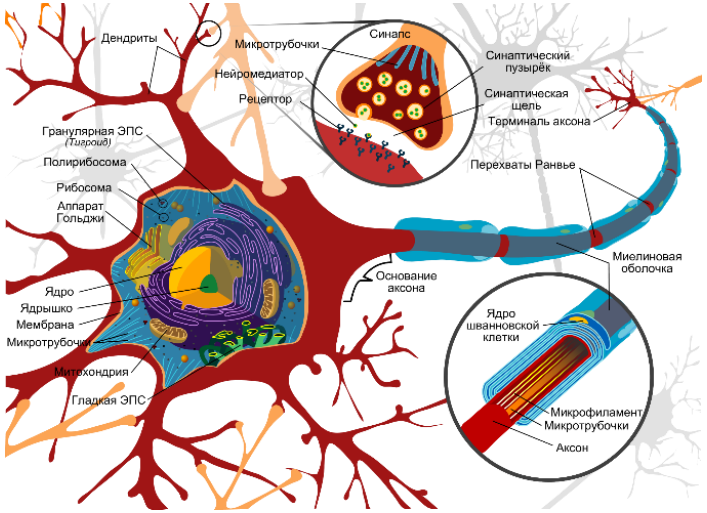
1. Задачи машинного обучения
   - Предыстория машинного обучения
   - Терминология машинного обучения
   - Примеры задач машинного обучения
2. Методология машинного обучения
   - Нейронные сети и глубокое обучение
   - Оптимизационные задачи машинного обучения
   - Задачи машинного обучения с векторизацией объектов
3. Большие языковые модели
   - Модели внимания и трансформеры, эмерджентность
   - О некоторых задачах вычислительной биологии
   - Смена парадигмы

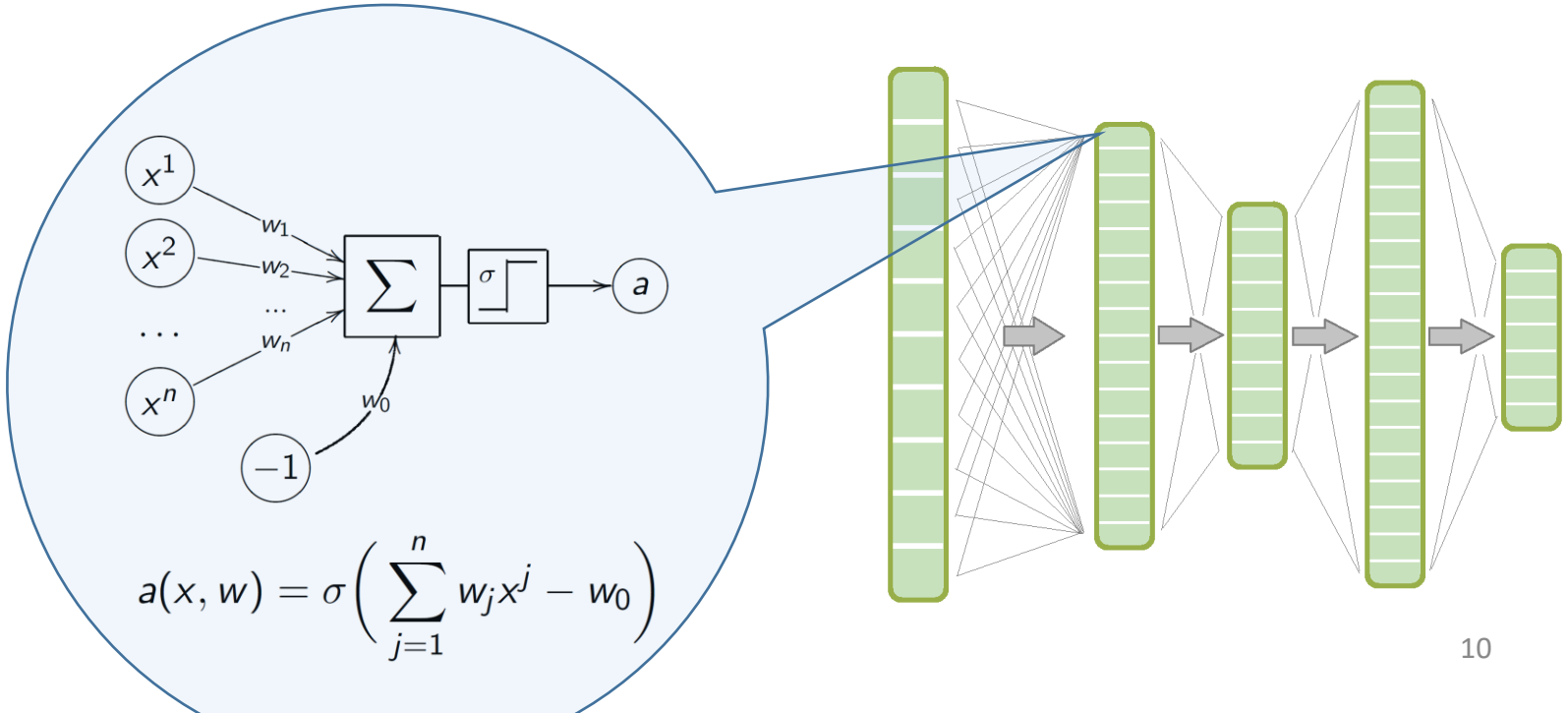
Искусственные нейронные сети (ANN)

На каждом слое сети вектор объекта преобразуется в новый вектор

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

Веса w являются обучаемыми параметрами модели

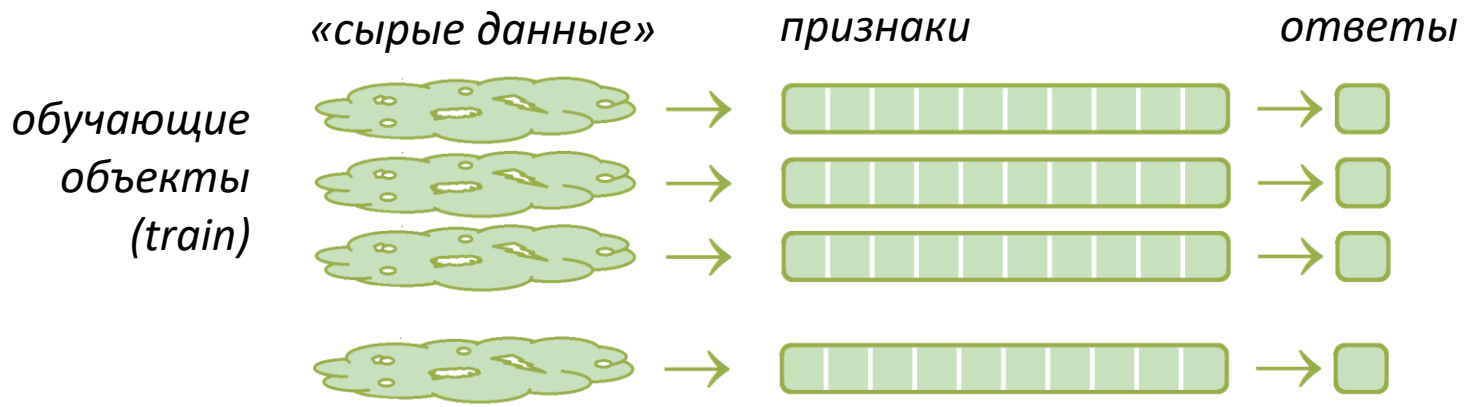




Глубокие нейронные сети (Deep ANN)

Вход: сложно структурированные «сырые» данные объектов

Выход: векторные представления объектов, затем ответы

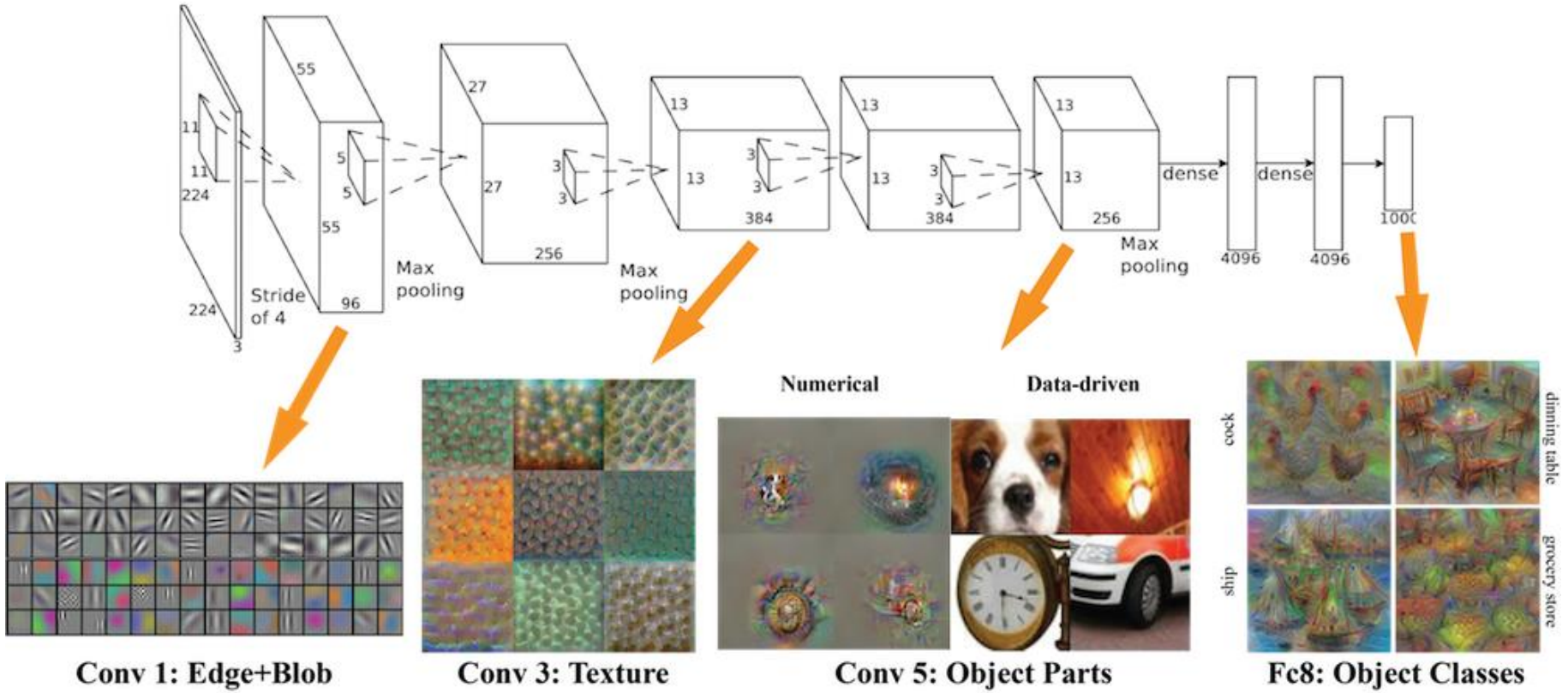


Deep Learning – это всего лишь обучаемая векторизация сложных объектов

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

Глубокие свёрточные нейронные сети (CNN)

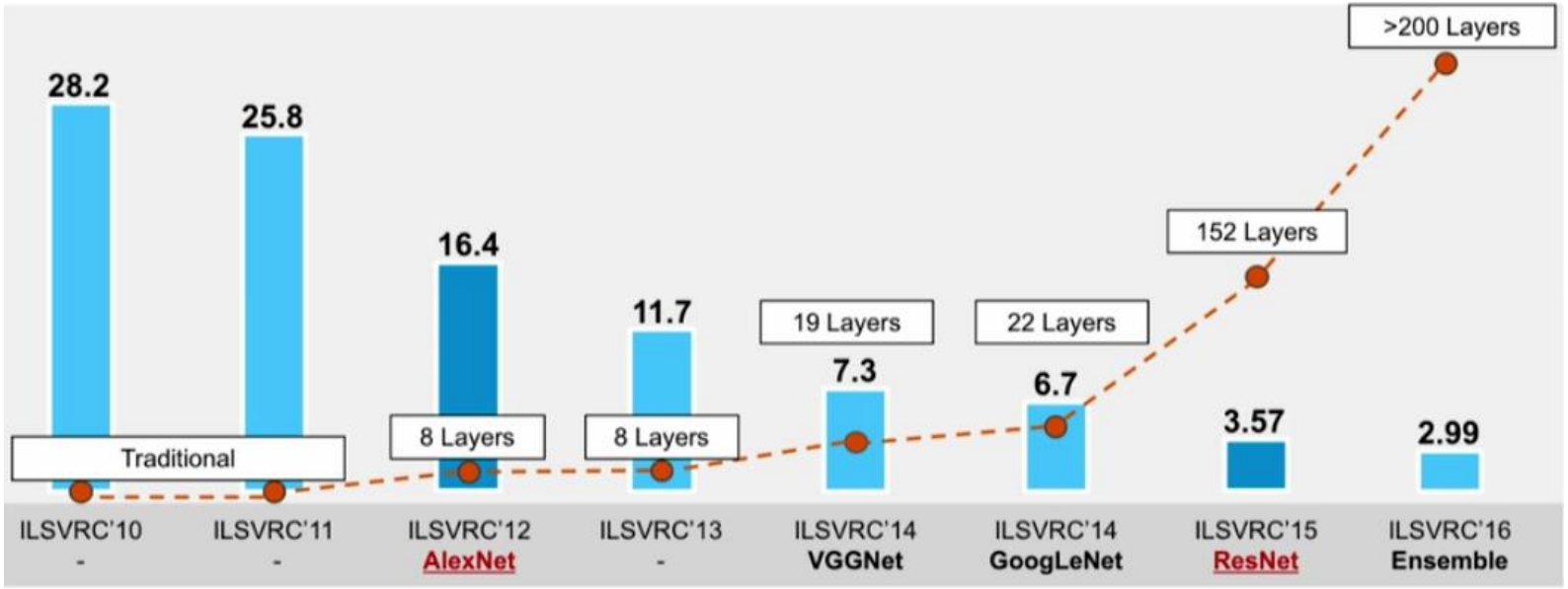
для классификации объектов на изображениях



Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks. 2012.

Роль больших данных

ImageNet: открытая выборка 14M изображений, 20K категорий



Старт в 2009 г. Человеческий уровень ошибок 5% пройден в 2015 г.

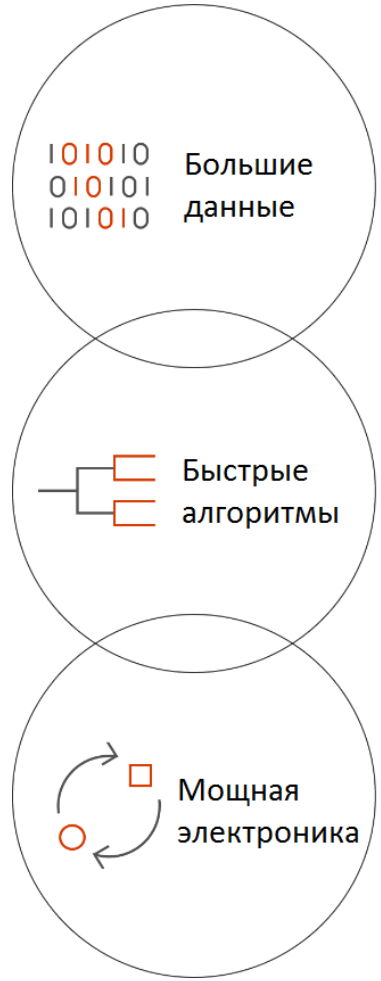
Li Fei-Fei et al. ImageNet: A large-scale hierarchical image database. 2009.
Li Fei-Fei et al. Construction and analysis of a large scale image ontology. 2009.

Три составляющих успеха Deep Learning

- Повсеместное применение компьютерных технологий
  → *накопление больших выборок данных*
  в частности, ImageNet

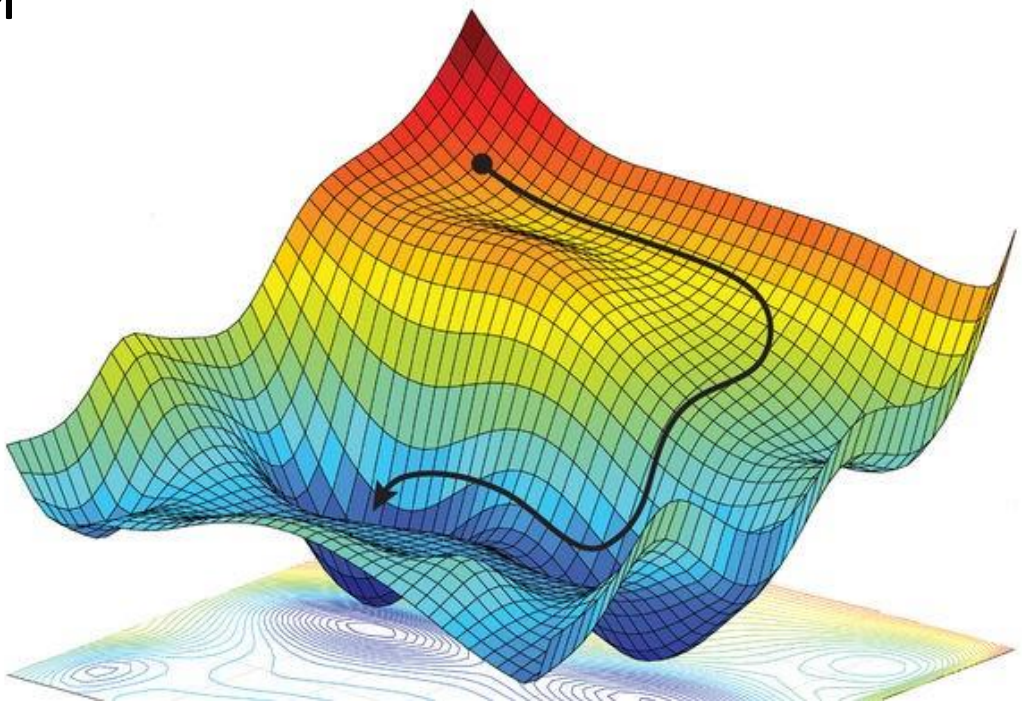
- Развитие математических методов и алгоритмов
  → *накопление критической массы опыта*
  методы оптимизации, контроль переобучения

- Достижения микроэлектроники
  → *рост вычислительных мощностей по закону Мура*
  в частности, GPU



Машинное обучение – это оптимизация

x – вектор объекта обучающей выборки
$a(x, w)$ – предсказательная модель
w – параметры модели
$\mathrm{Loss}(x, w)$ – функция потерь
$Q(w)$ – критерий качества модели



Задача обучения параметров модели:

$$Q(w) = \sum_x \mathrm{Loss}(x, w) \quad \to \min$$

Способ решения – численные методы оптимизации

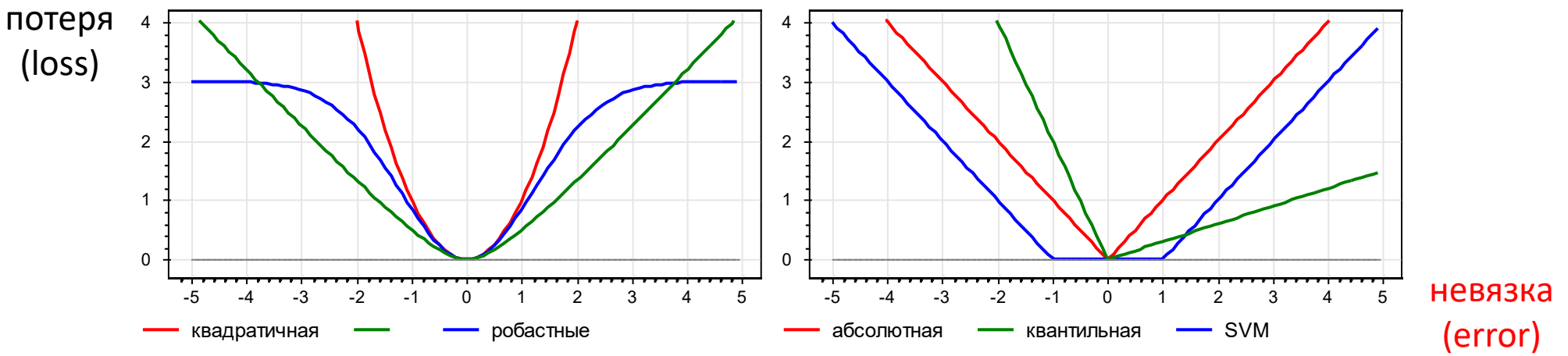
Обучение с учителем (supervised learning): восстановление регрессии (regression)

x — вектор объекта обучающей выборки, y — числовой ответ

$a(x, w)$ — модель регрессии с параметрами w

Например, $a(x, w) = \sum_j w_j x_j$ — линейная модель регрессии

$\text{Loss}(x, w) = (a(x, w) - y)^2$ — квадратичная функция потерь



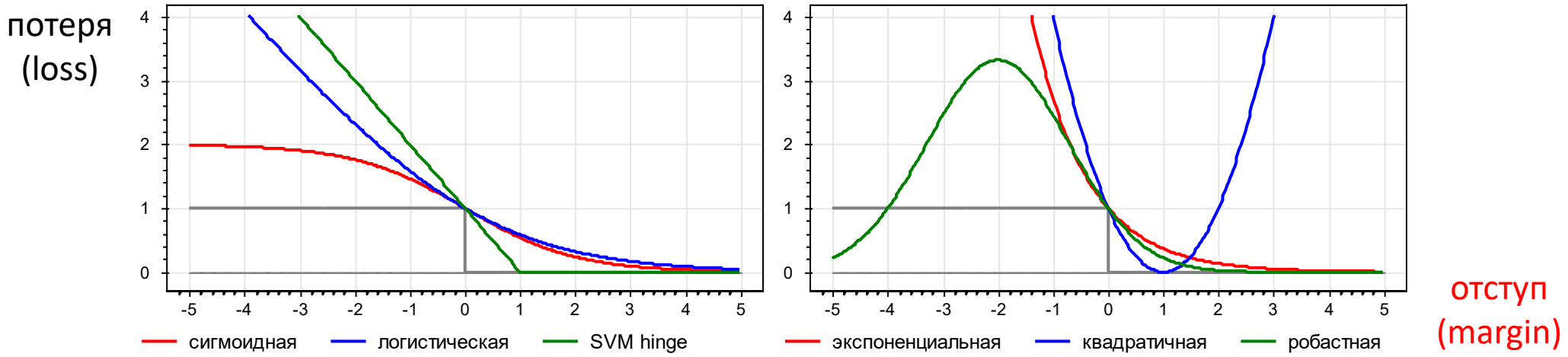
Обучение с учителем (supervised learning): классификация (classification)

x — вектор объекта обучающей выборки, y — ответ ($+1$ или -1)

$a(x, w)$ — модель классификации с параметрами w

Например, $a(x, w) = \text{sign}\left(\sum_j w_j x_j\right)$ — линейная модель

$\text{Loss}(x, w) = \max\left(0, 1 - y\sum_j w_j x_j\right)$ — функция потерь SVM hinge



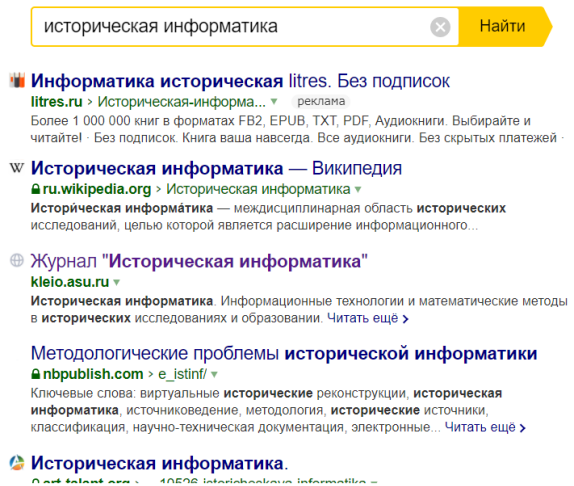
Обучение с учителем (supervised learning): ранжирование (learning to rank)

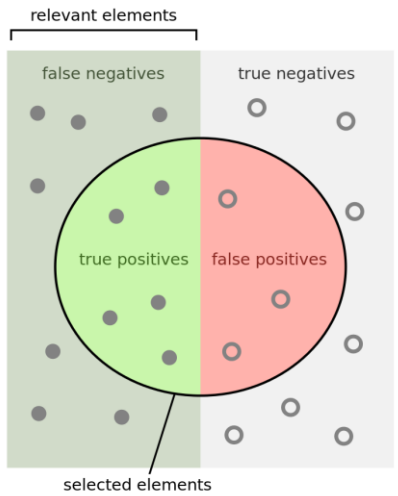
x — вектор пары «запрос-документ», y — оценка релевантности

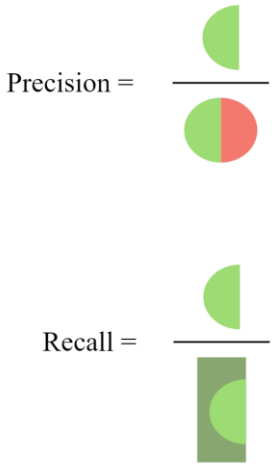
$a(x, w)$ — модель ранжирования документов по запросу, параметр w

Например, $a(x, w) = \sum_j w_j x_j$ — линейная модель

$$\text{Loss}(x, x', w) = \max\Big(0, 1 - [y > y'](a(x, w) - a(x', w))\Big)$$







не только поиск,
но и любые задачи, где
человеку удобно
принимать решения,
выбирая один из вариантов

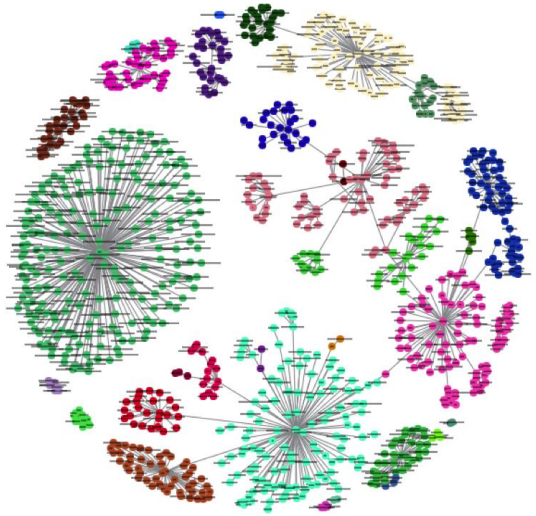
Обучение без учителя (unsupervised learning): кластеризация (clustering)

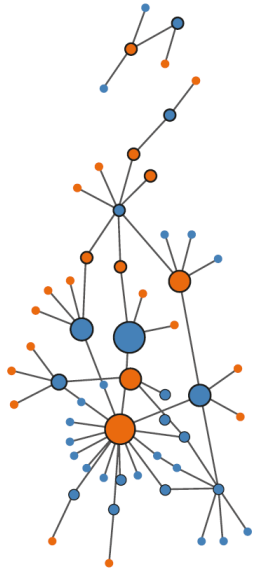
x — вектор объекта обучающей выборки, ответы не задаются

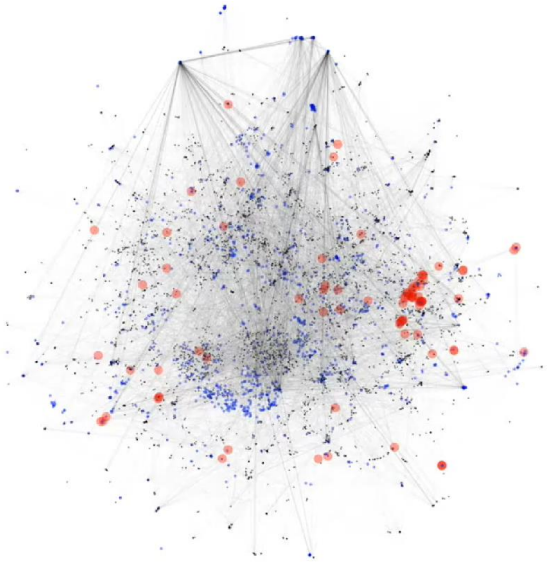
$a(x, w)$ — кластер, ближайший к x

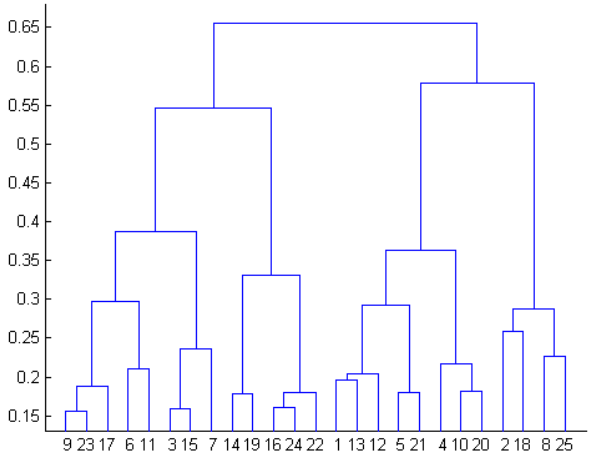
$w = \{c_1, \dots, c_K\}$ — векторы центров всех кластеров

$\text{Loss}(x, w) = \min_k \|x - c_k\|$ — расстояние до ближайшего кластера









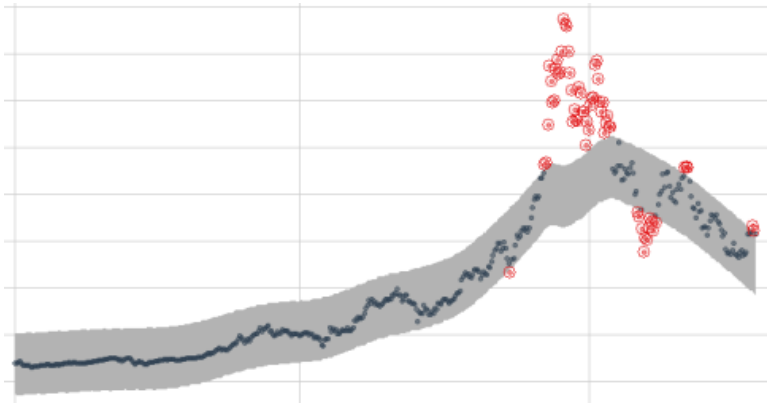
Выявление аномалий, выбросов, нового (anomaly /outlier /novelty detection)

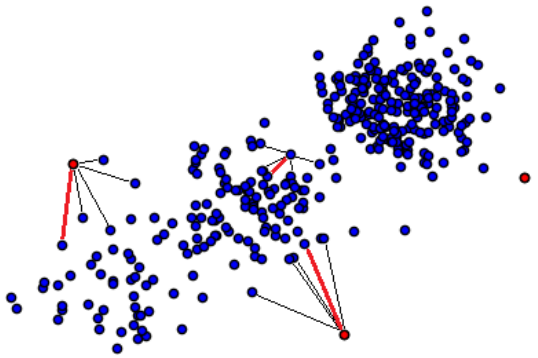
x — вектор объекта

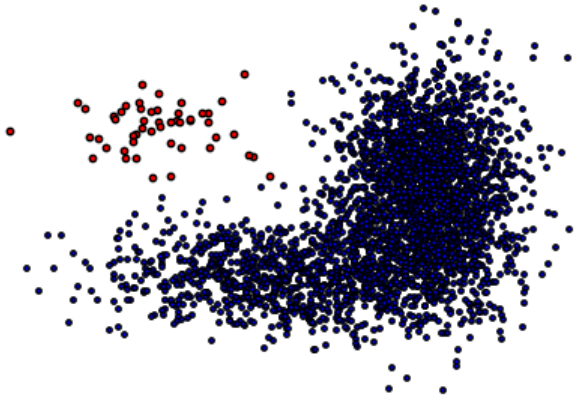
$a(x, w)$ — модель регрессии / классификации / кластеризации

$\mathrm{Loss}(x, w)$ — выбранная функция потерь

объекты ранжируются по убыванию потерь, анализируются top-*k*







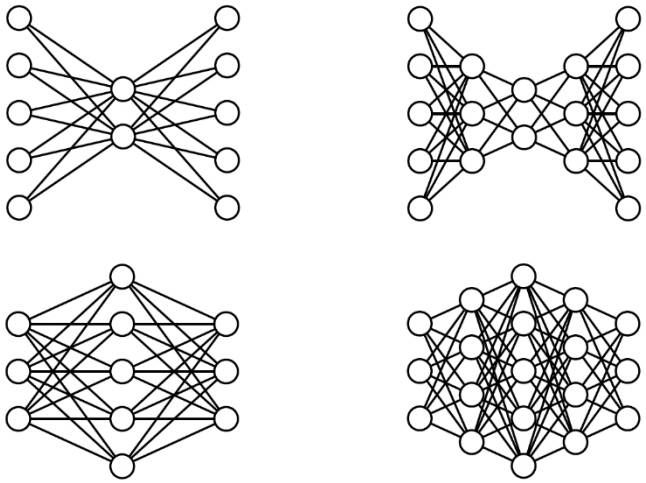
Обучение без учителя (unsupervised learning): векторизация, автокодировка (autoencoder)

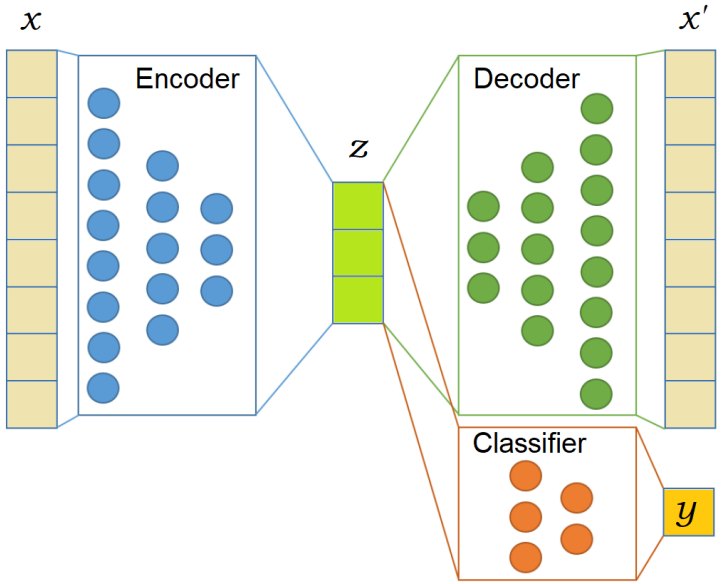
x — описание объекта обучающей выборки, ответов не дано

$z = f(x, w)$ — модель кодирования x в векторное представление z

$x' = g(z, w')$ — модель декодирования z в реконструкцию x'

$\text{Loss}(x, w) = \|g(f(x, w), w') - x\|$ — точность реконструкции объекта





обучаемая векторизация сложных объектов

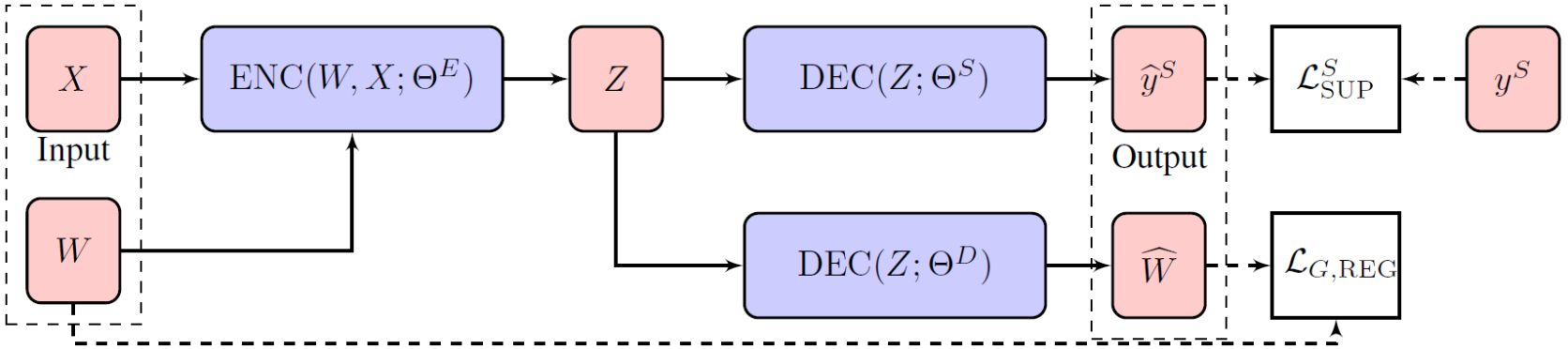
Частичное обучение (semi-supervised learning): векторизация графов (graph embeddings)

$x; (x, x')$ — данные об объектах и взаимодействиях между объектами

$z = f(x, \theta^E)$ — модель векторизации объектов x (вершин графа)

$x' = g(z, \theta^D)$ — модель декодирования z в реконструкцию x'

$\text{Loss}(x, w) = \|g(f(x, \theta^E), \theta^D) - x\| + \tau L^S_{\text{SUP}}(x, \theta^S)$ — два критерия



обучаемая векторизация сложных объектов по данным об их взаимодействиях

I.Chami et al. Machine learning on graphs: a model and comprehensive taxonomy. 2020.

T.Mikolov et al. Efficient estimation of word representations in vector space, 2013.

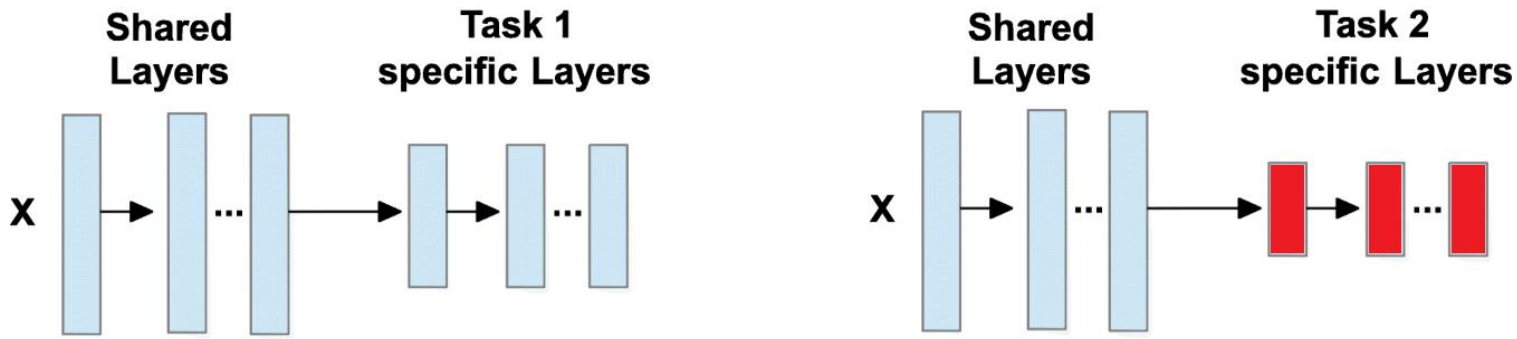
Перенос обучения (transfer learning), предобучение модели векторизации

$z = f(x, w)$ — модель векторизации, универсальная для многих задач

$y = g(z, w')$ — часть модели, специфичная для своей задачи

$\min\limits_{w,\, w'}: \sum_x \text{Loss}_1(g_1(f(x, w), w'))$ — обучение по большим данным

$\min\limits_{w'}: \sum_{x'} \text{Loss}_2(g_2(f(x', w), w'))$ — обучение по своим данным

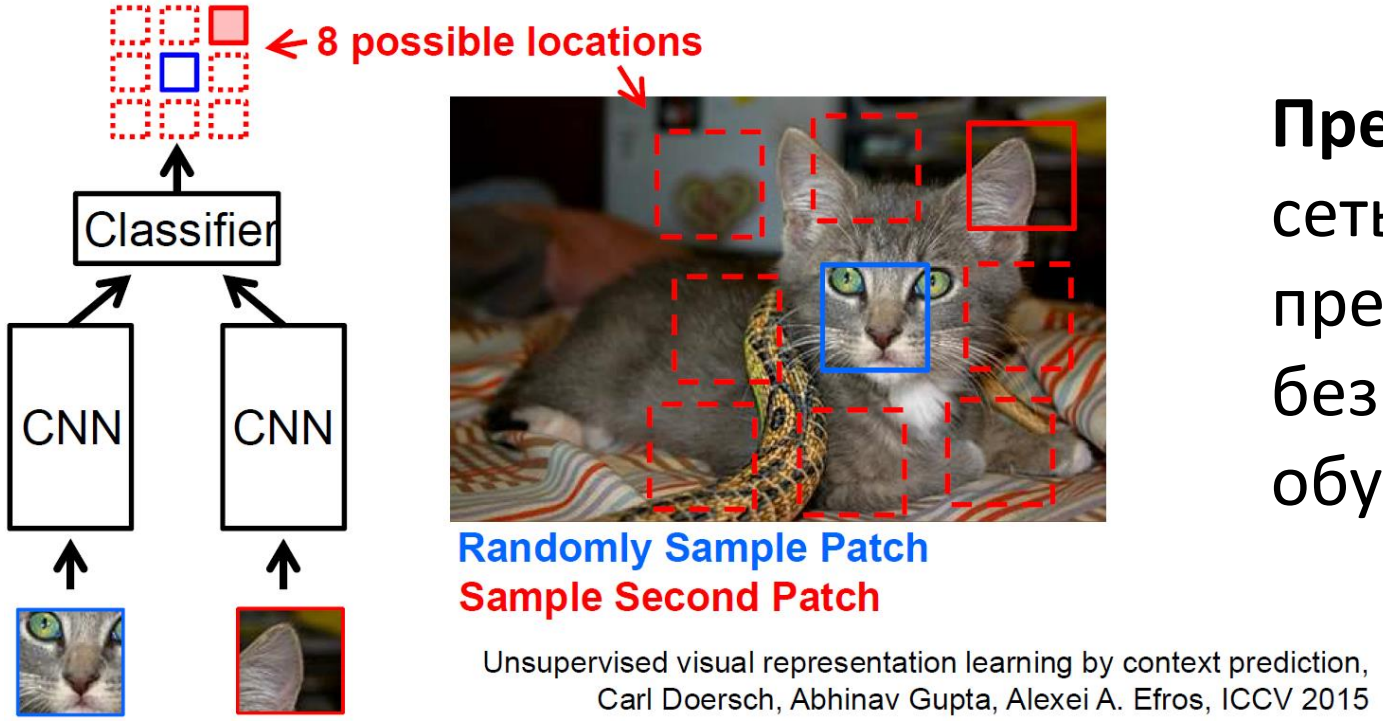


Sinno Jialin Pan, Qiang Yang. A Survey on Transfer Learning. 2009

Самостоятельное обучение (self-supervised)

x — изображение

$z = f(x, w)$ — модель векторизации, обучается предсказывать взаимное расположение пар фрагментов одного изображения



Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Преимущество:
сеть выучивает векторные представления объектов без размеченной обучающей выборки

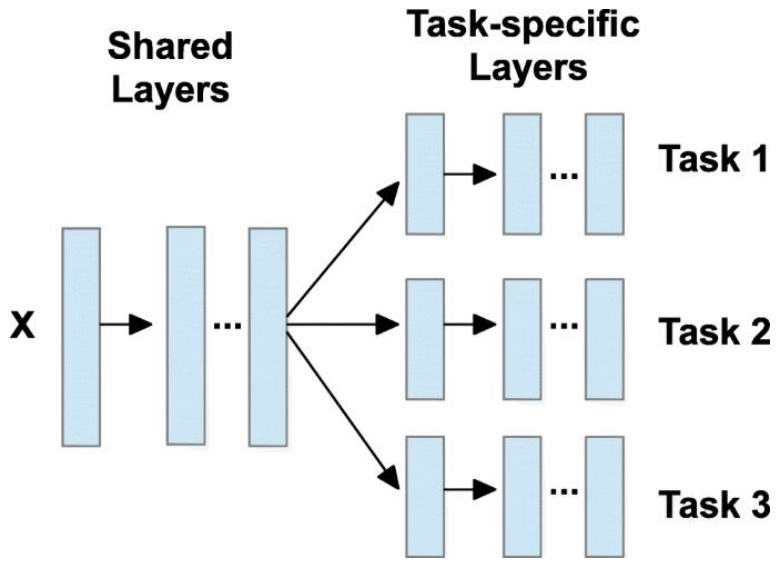
Многозадачное обучение (multi-task learning)

$z = f(x, w)$ – модель векторизации, универсальная для всех задач

$y = g_t(z, w'_t)$ – часть модели, специфичная для t-й задачи

$\min\limits_{w, w'_t}: \sum_t \sum_x \text{Loss}_t(g_t(f(x, w), w'_t))$ – обучение по всем задачам

few-shot learning – обучение по малому числу примеров



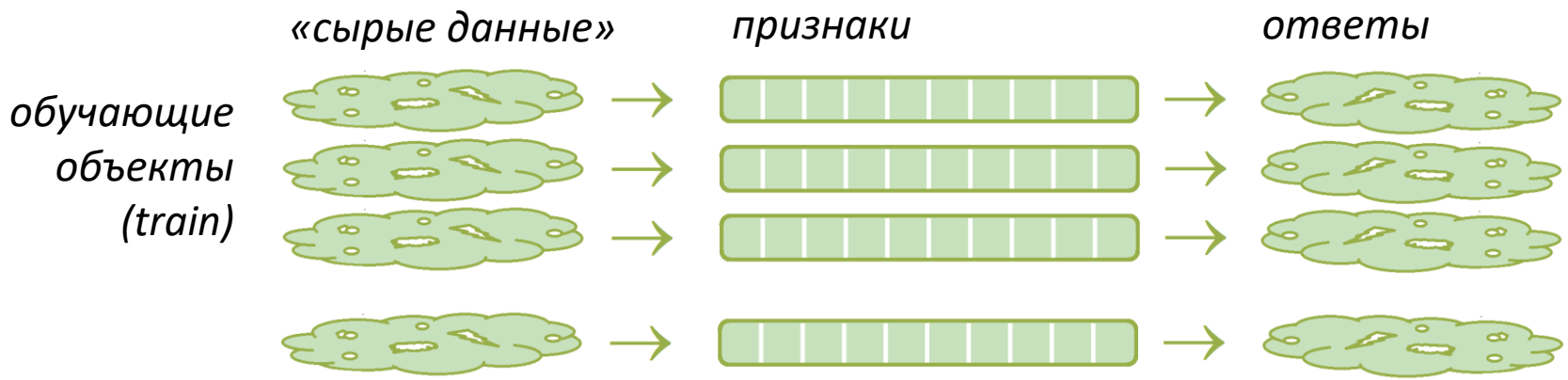
M.Crawshaw. Multi-task learning with deep neural networks: a survey. 2020

Y.Wang et al. Generalizing from a few examples: a survey on few-shot learning. 2020

Нейронные сети для синтеза объектов

Вход: сложно структурированные объекты

Выход: сложно структурированные ответы



похоже на автокодировщиков

Примеры: синтез изображений, перенос стиля, распознавание речи, машинный перевод, суммаризация текстов, диалог с пользователем

Модели: seq2seq, CNN, RNN, LSTM, GAN, BERT, GPT и др.

Генеративная состязательная сеть (GAN)

$x = g(z, w)$ — модель генерации реалистичного объекты x из шума z

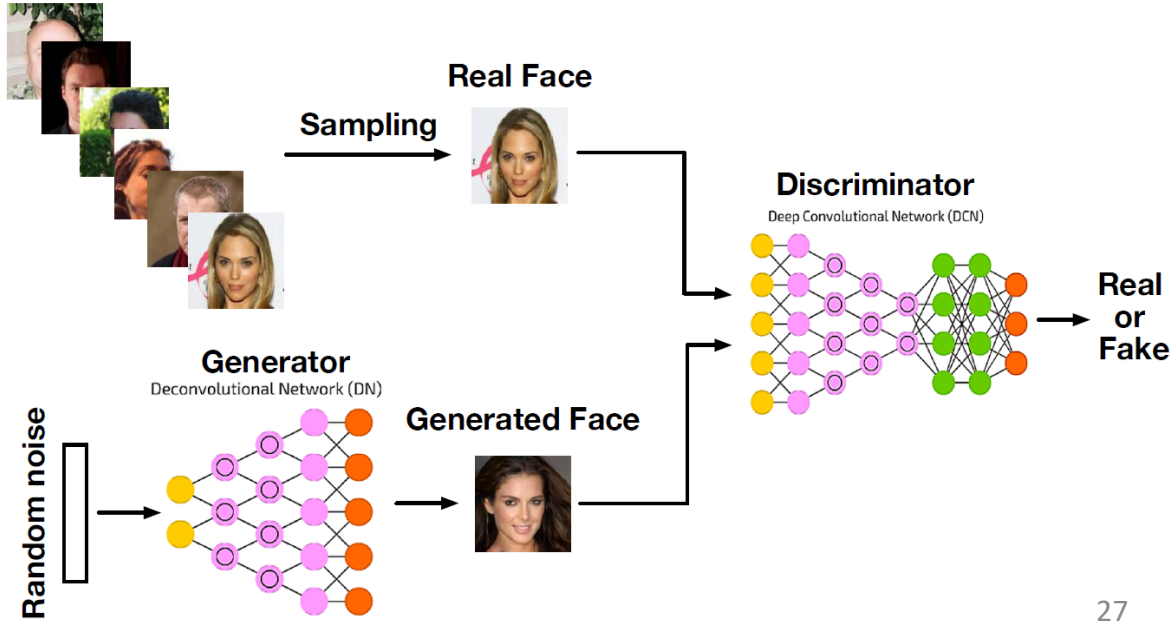
$f(x, w')$ — модель классификации x «реальный/сгенерированный»

$\min_{w} \max_{w'} \sum_x \ln f(x, w') + \ln\left(1 - f(g(z, w), w')\right)$ — совместное обучение

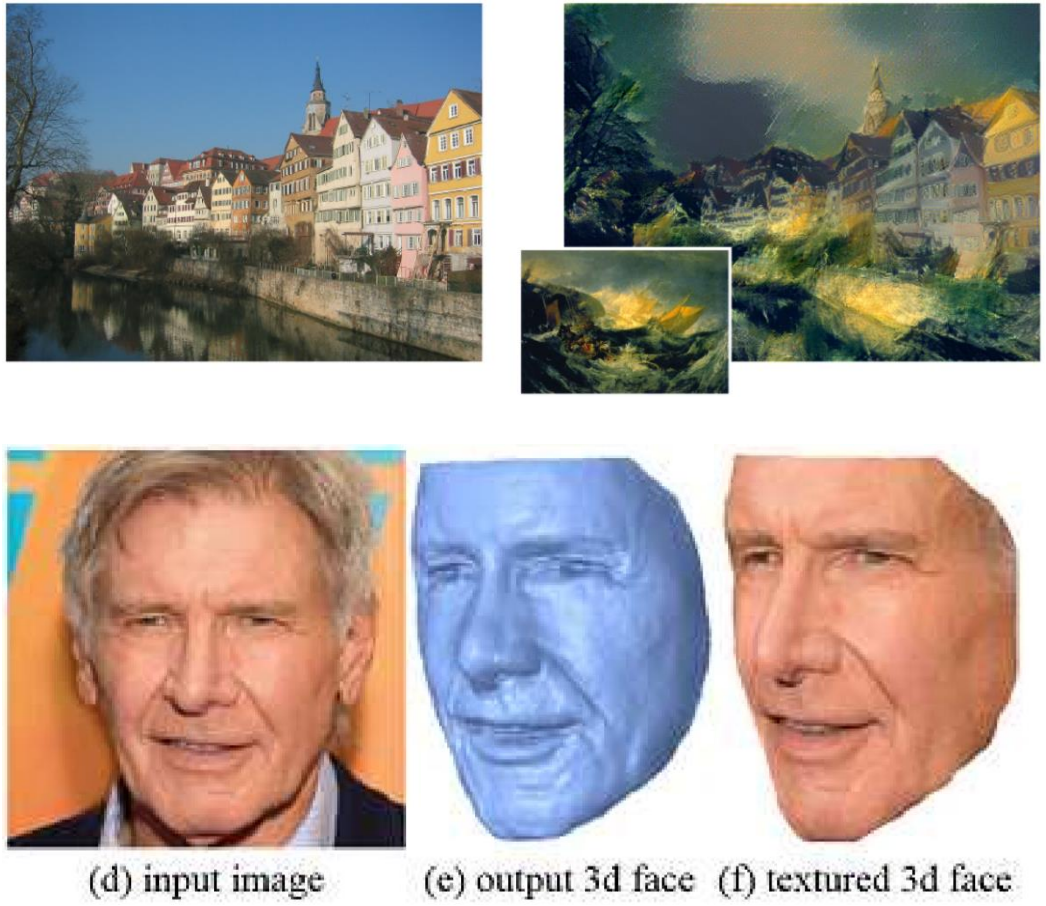
Antonia Creswell et al. Generative Adversarial Networks: an overview. 2017.

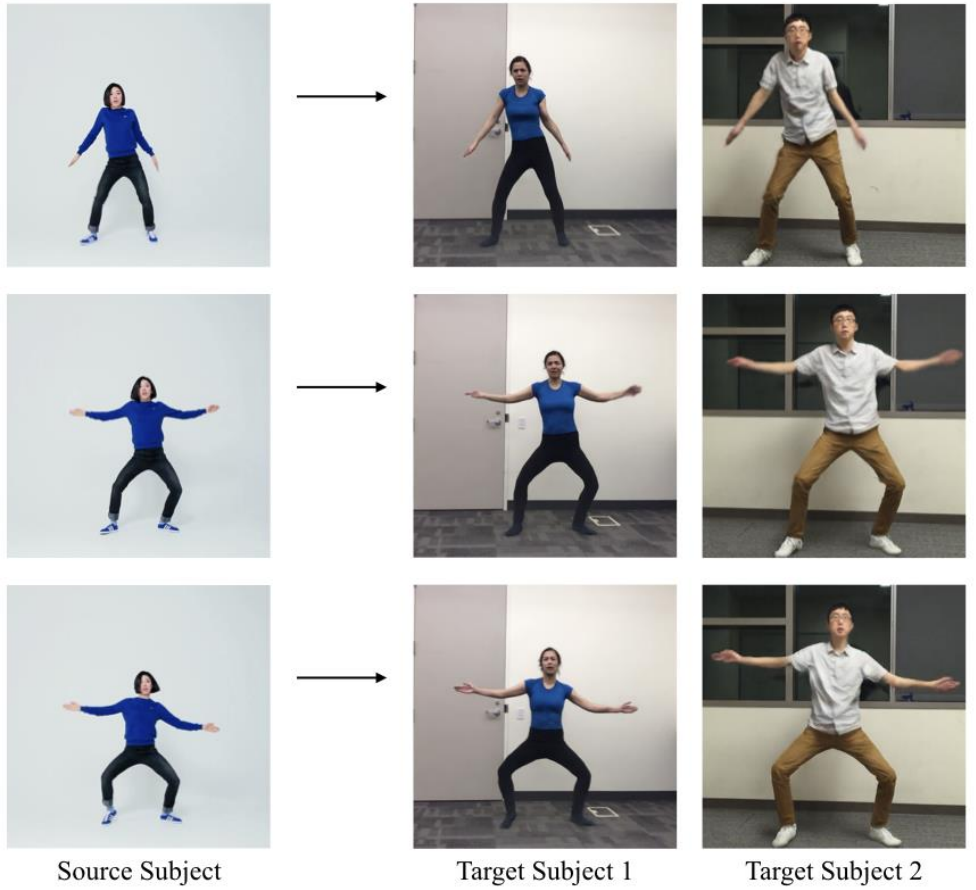
Zhengwei Wang et al. Generative Adversarial Networks: a survey and taxonomy. 2019.

Chris Nicholson. A Beginner's Guide to Generative Adversarial Networks. 2019.



Синтез изображений и видео





Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros. Everybody Dance Now. ICCV-2019.

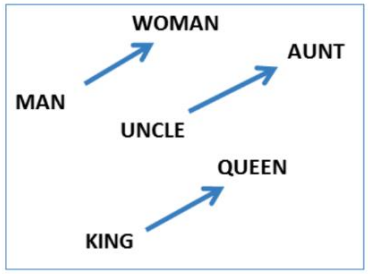
Эволюция подходов в обработке текстов

Декомпозиция задач по уровням «пирамиды NLP»

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

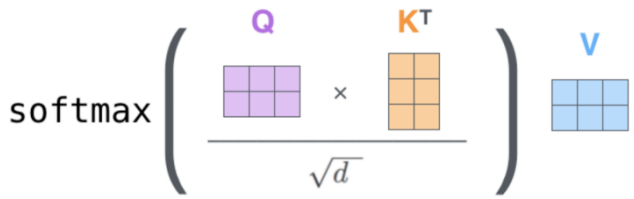
Модели векторизации слов (эмбедингов)

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



Нейросетевые модели контекстной векторизации

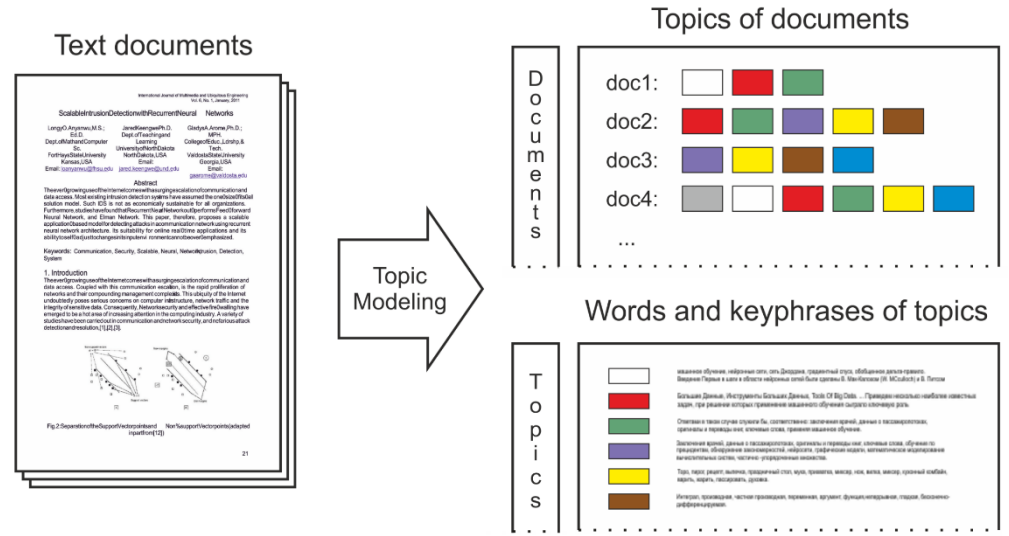
- рекуррентные нейронные сети: LSTM, GRU, …
- «end-to-end» модели внимания и трансформеры:
  машинный перевод [2017], BERT [2018], GPT-4 [2023], …

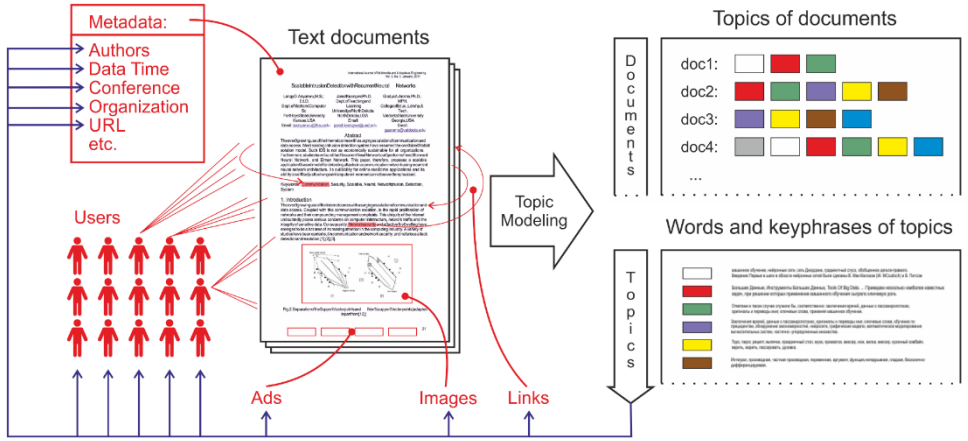


Тематическое моделирование (topic models)

Тематическая модель (ТМ) коллекции текстовых документов определяет

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





Мультимодальная ТМ определяет также,

- какие ещё нетекстовые токены содержатся в каждой теме

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

Vorontsov K. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. 2023. http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

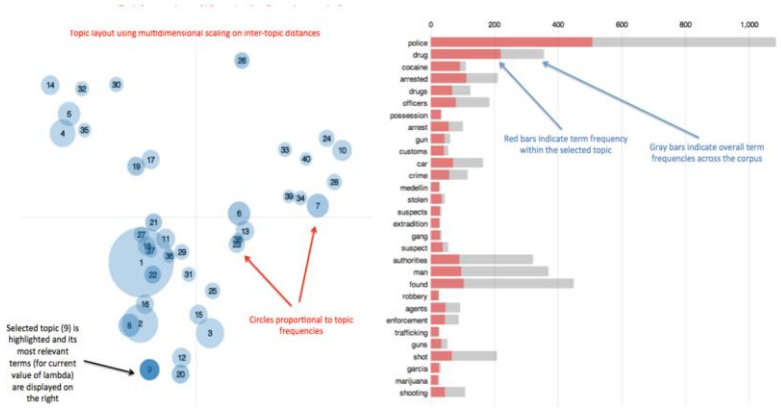
Тематическое моделирование (topic models)

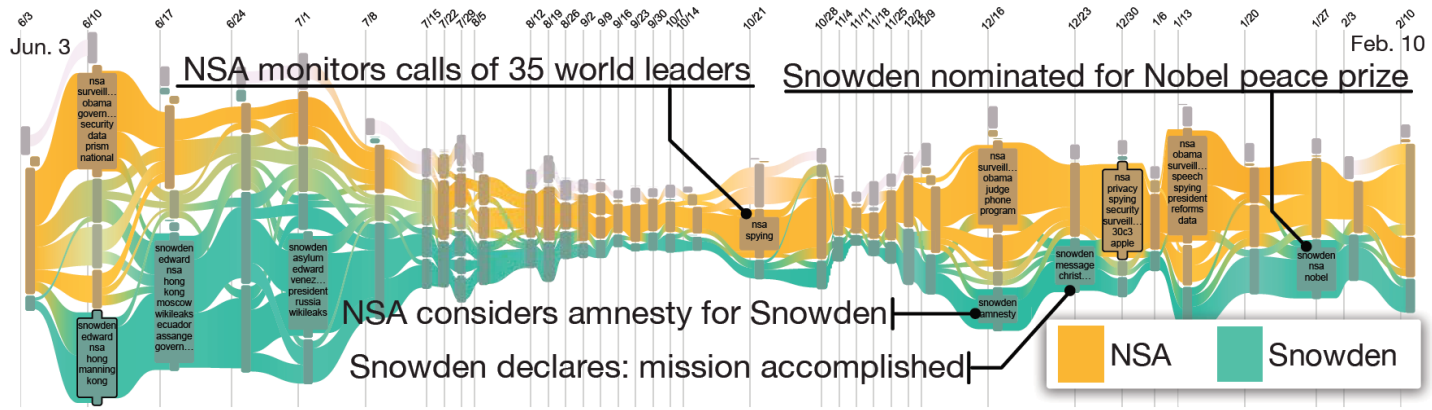
x – текст на естественном языке, «мешок слов» $p(\text{слово}|x)$

$z = f(x, w)$ – *модель кодирования* x в вектор тем $z = p(\text{тема}|x)$

$x' = g(z, w)$ – *модель декодирования* z в реконструкцию текста x'

$\text{Loss}(x, w) = \text{KL}\big(x \parallel g(f(x, w), w)\big)$ – точность реконструкции текста





Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URSS, 2023.

Тематическое моделирование:
возможное применение в вычислительной биологии

Можно заметить аналогию между

— *текстами*, в которых встречаются *слова* из общего словаря

— *клетками*, в которых экспрессируются *гены* из общего генома

данные — текстовая коллекция	Human Cell Atlas (42M клеток человека)
слово	ген
текст — мешок слов	клетка — мешок генов
частота слова	уровень экспрессии гена
тема — связанные по смыслу слова	тема — биомолекулярный путь?

Chung-Chau Hon et al. The Human Cell Atlas: Technical approaches and challenges. 2018
Lin Liu et al. An overview of topic modeling and its current applications in bioinformatics. 2016.

Содержание

1. Задачи машинного обучения
   - Предыстория машинного обучения
   - Терминология машинного обучения
   - Примеры задач машинного обучения
2. Методология машинного обучения
   - Нейронные сети и глубокое обучение
   - Оптимизационные задачи машинного обучения
   - Задачи машинного обучения с векторизацией объектов
3. Большие языковые модели
   - Модели внимания и трансформеры, эмерджентность
   - О некоторых задачах вычислительной биологии
   - Смена парадигмы

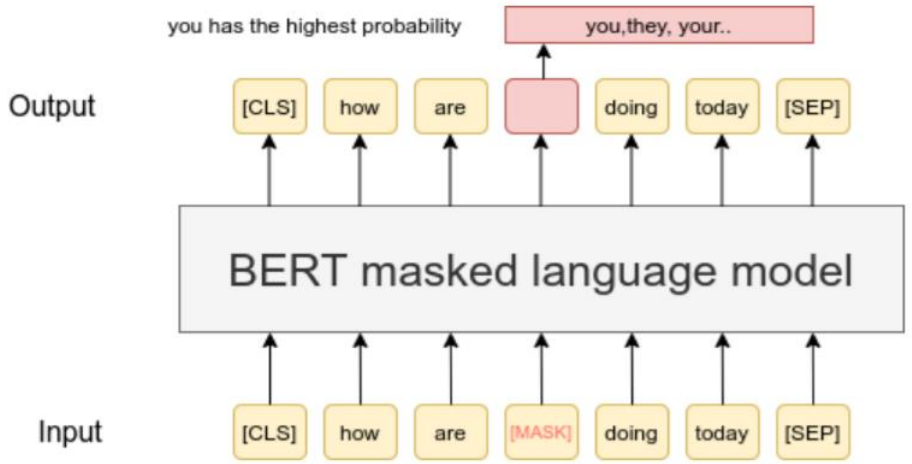
Обучение контекстной векторизации слов

x_i — слово на i-й позиции в коллекции текстовых документов

$z_i = f(x_i, C_i, w)$ — модель векторизации слова x_i по контексту C_i

$p(x|i, z, w')$ — вероятностная модель предсказания слова по вектору z

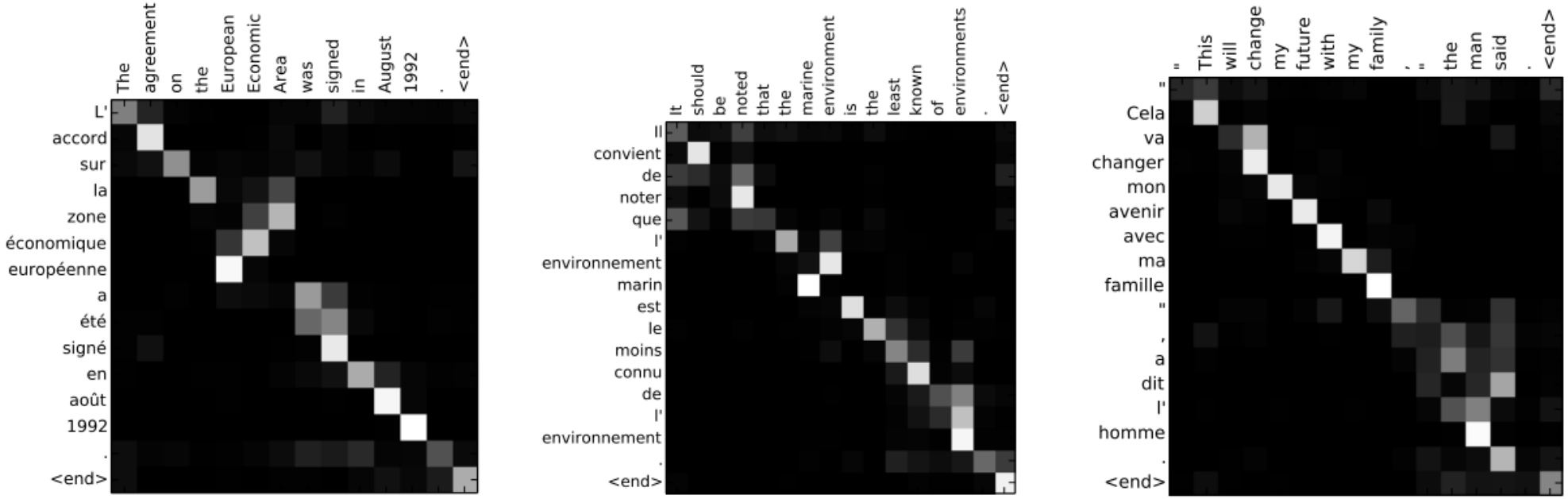
$\text{Loss}(x_i, w) = -\ln p(x_i|i, f(x_i, C_i, w), w')$ — потеря от предсказания слова на i-й позиции по его контексту (Masked Language Model)



Vaswani et al. (Google) Attention is all you need. 2017.

Jacob Devlin et al. (Google AI Language)
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

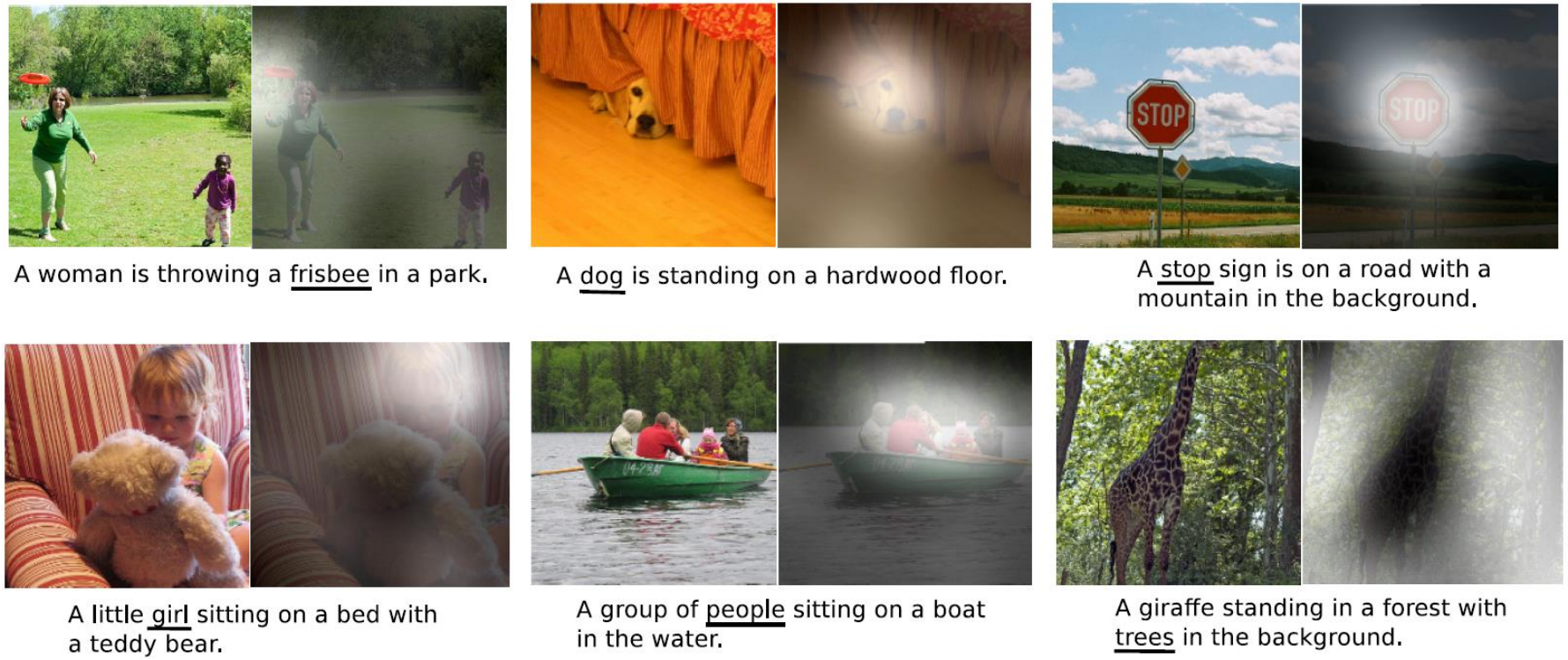
Модели внимания: машинный перевод



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

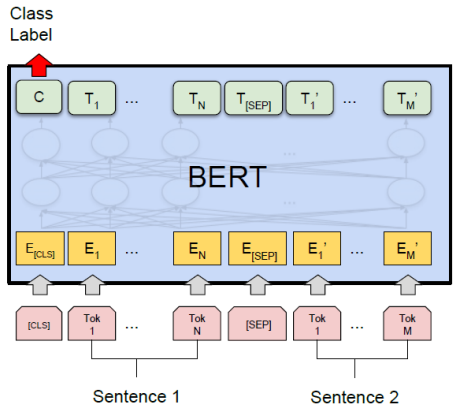


Интерпретация: на какие области модель обращает внимание, генерируя подчёркнутое слово в описании изображения

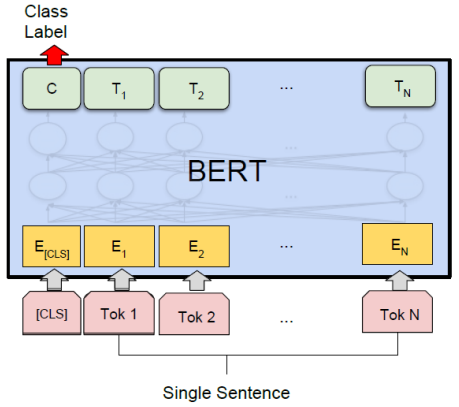
Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

Трансформеры: большие языковые модели

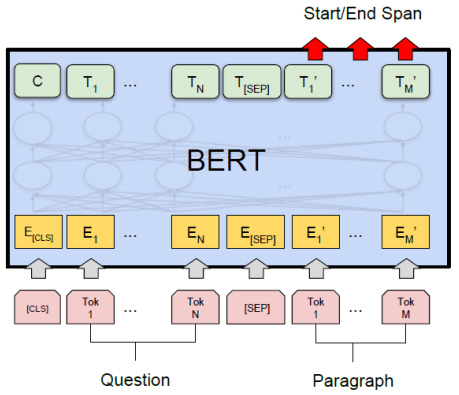
- LLM обучаются векторизовать и предсказывать слова по контексту
- обучаются по терабайтам текстов, «они видели в языке всё»
- мультиязычны: обучаются на десятках языков
- мультизадачны: для каждой новой задачи NLP/NLU достаточно предобученной модели или дообучения на небольшой выборке



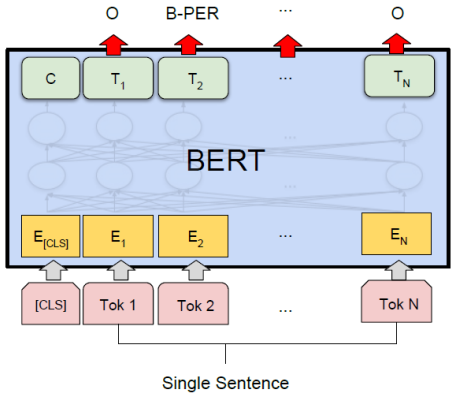
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



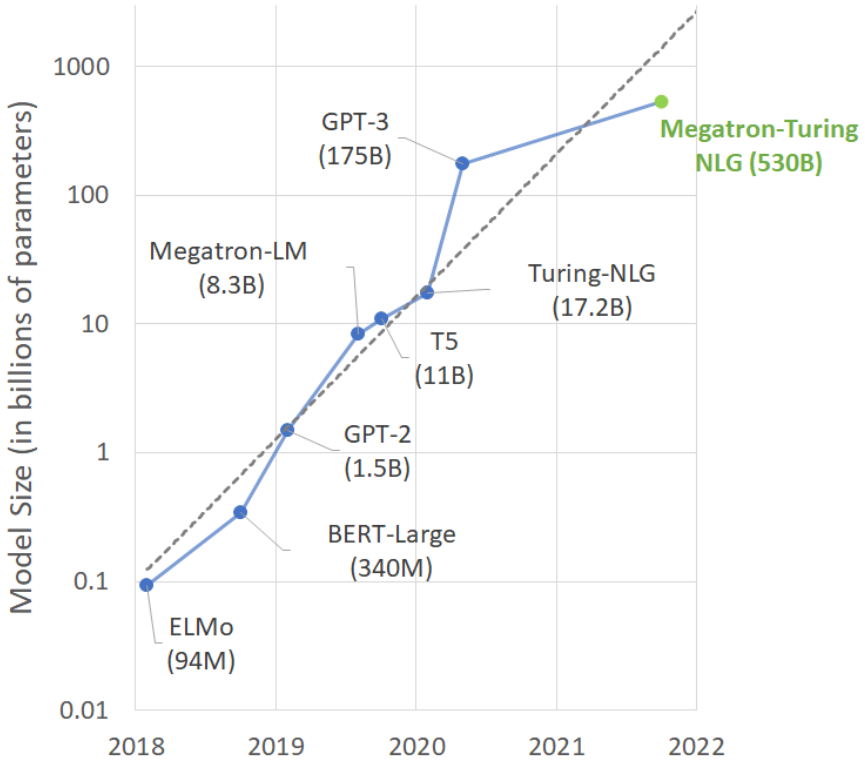
(c) Question Answering Tasks: SQuAD v1.1

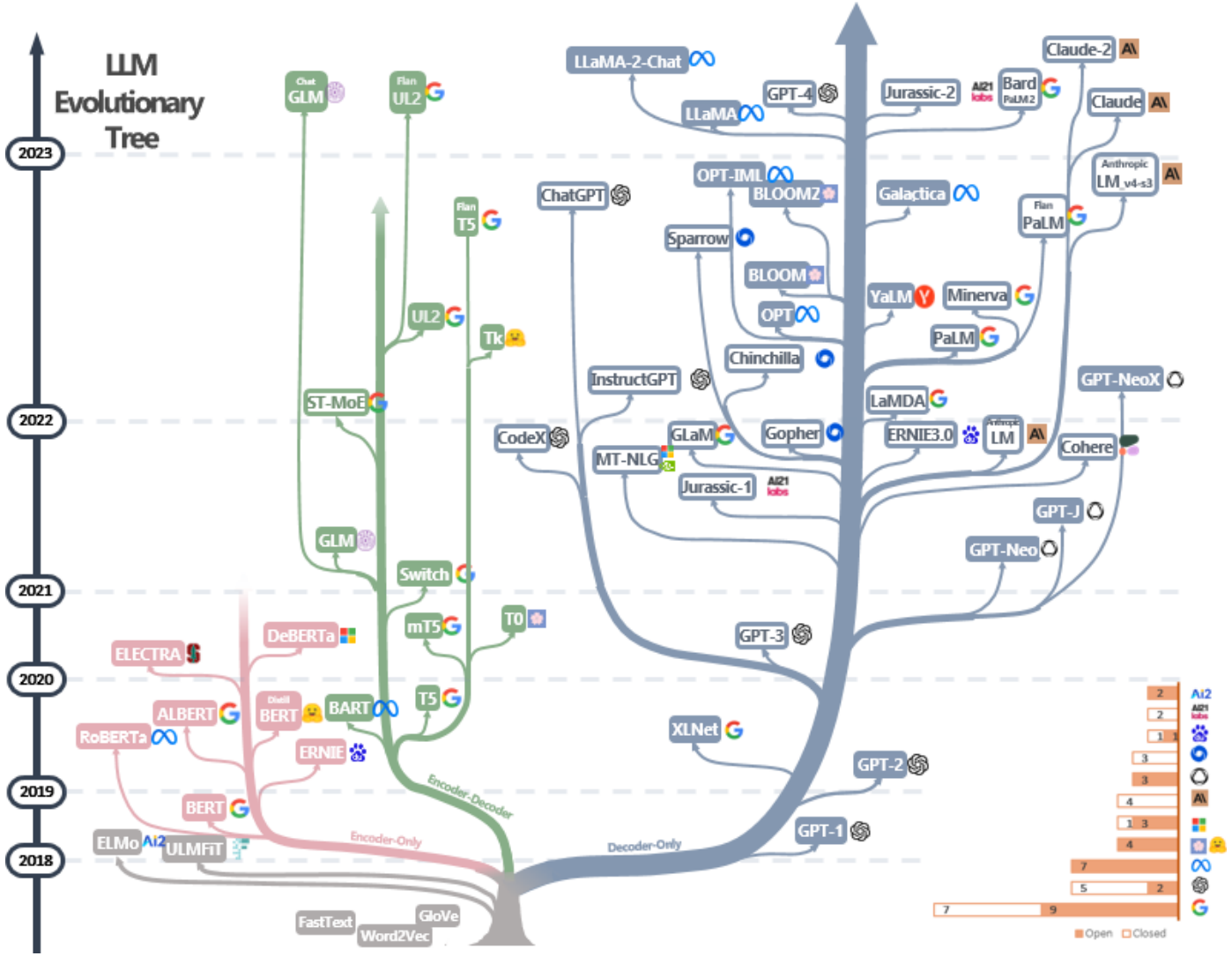


(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Трансформеры: большие языковые модели

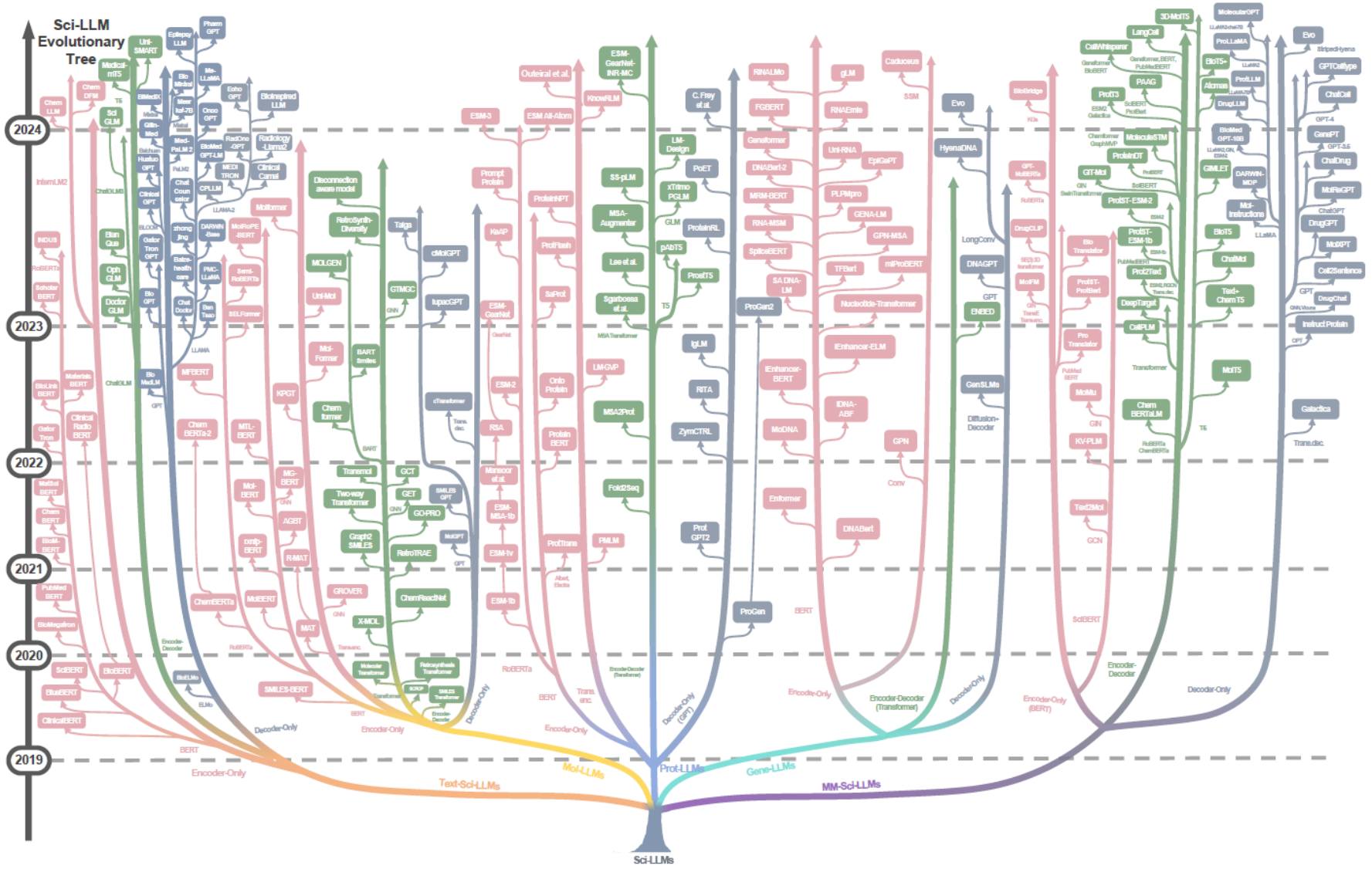
Рост числа параметров больших языковых моделей





Sci-LLM

большие языковые модели для обработки научных данных



Qiang Zhang et al. Scientific Large Language Models: A Survey on Biological & Chemical Domains. 2024-07-23

Проблески общего искусственного интеллекта

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

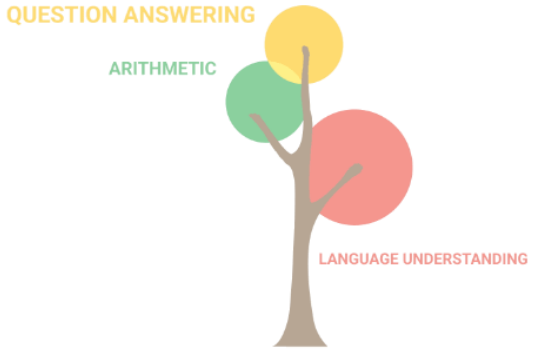
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

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

Новые (эмерджентные) способности модели

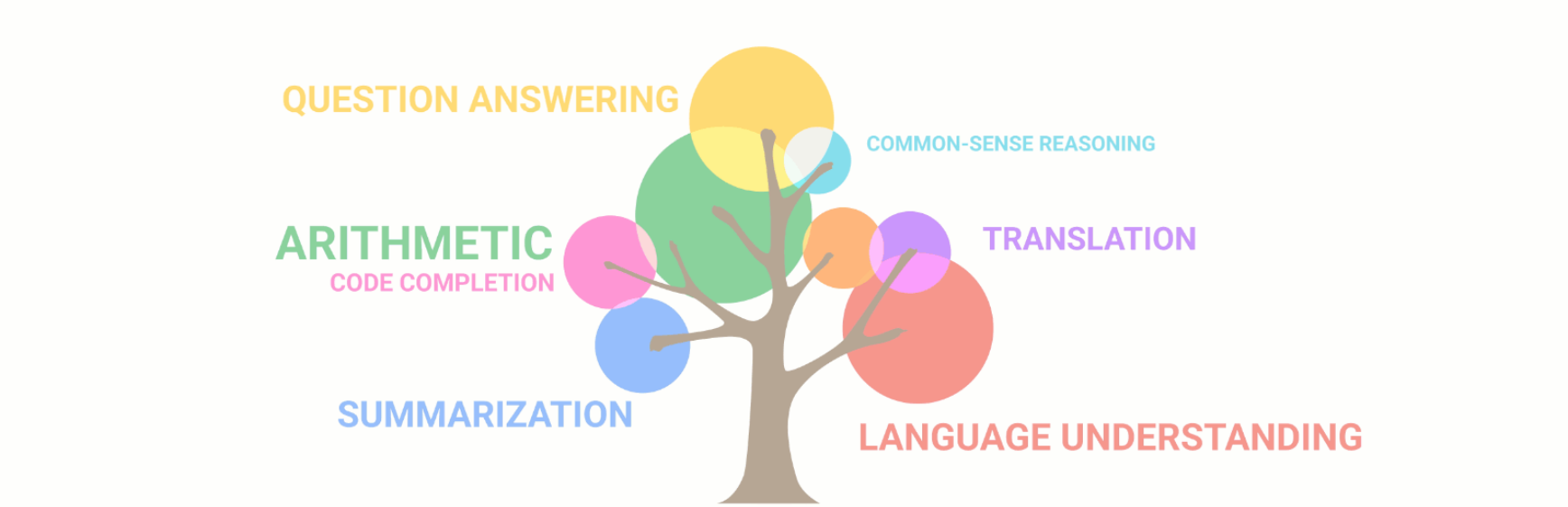


GPT-2: 14-Feb-2019

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

Новые (эмерджентные) способности модели

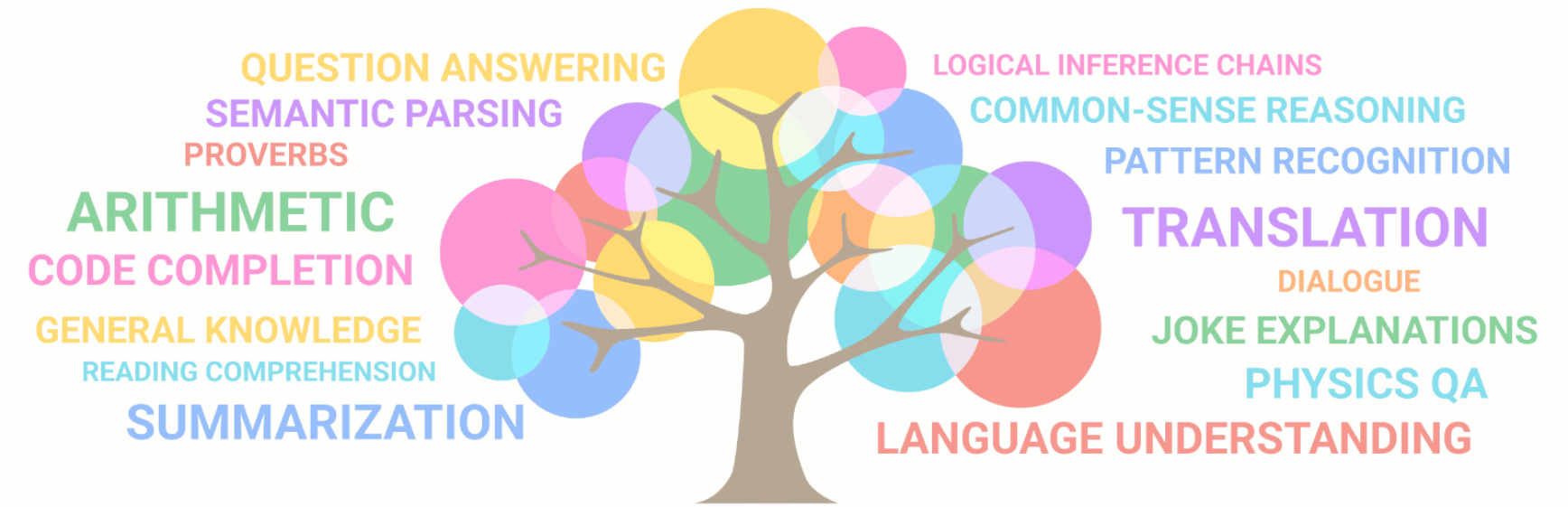


GPT-3: 11-Jun-2020

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

Новые (эмерджентные) 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 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

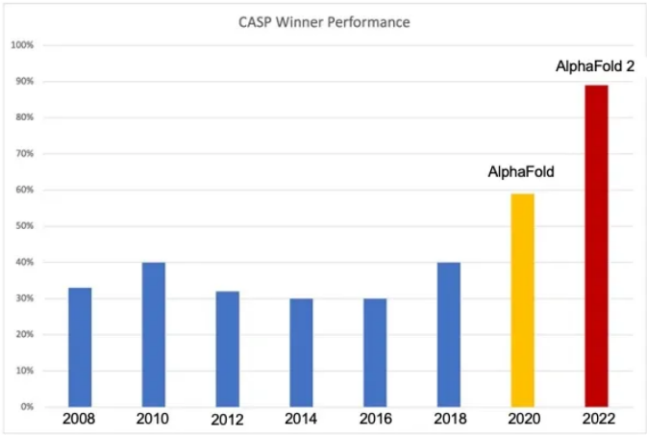
- «галлюцинировать», давать неверные сведения, касающиеся здоровья человека, законов, событий, технологий, других людей
- вызывать необоснованное доверие и манипулировать человеком
- переубеждать, побуждать человека к действиям, не выгодным ему
- поддерживать предрассудки и лженаучные представления
-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

Примеры задач молекулярной биологии, где LLM превосходят обычные методы

- **Предсказание фенотипических признаков по геномным вариациям:**
  объект – вариабельные участки генома
  ответ – статус заболевания, экспрессия генов, содержание белка и т.п.

- **Диагностика генетических заболеваний:**
  объект – нуклеотидная последовательность
  ответ – нарушения границ сплайсинга из-за мутаций

- **Предсказание структуры белка (AlphaFold, DeepMind):**
  объект – аминокислотная последовательность (1D)
  ответ – 3D структура белка

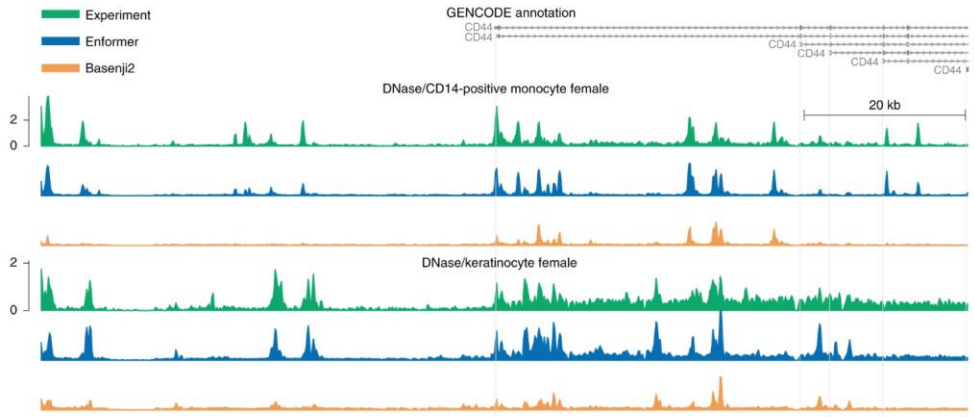


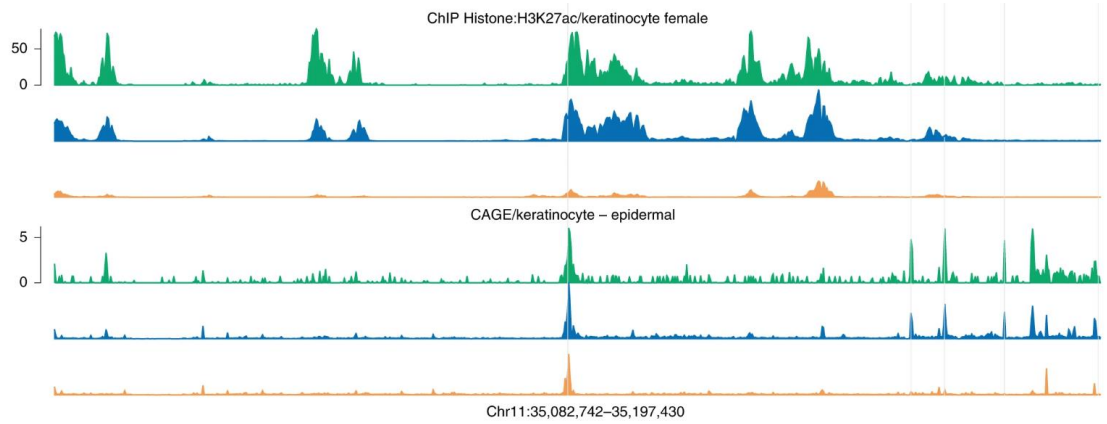
Jaganathan et al. Predicting splicing from primary sequence with deep learning. Cell 2019
Jumper J. et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021
Serafim Batzoglou. Large Language Models in Molecular Biology. Towards Data Science, 2023

Модель Enformer: трансформер над ДНК

- **Прогнозирование уровня экспрессии генов:**
  объект – нуклеотидная последовательность, тип клетки
  ответ – экспрессия генов, модификации гистонов, доступность хроматина и др.

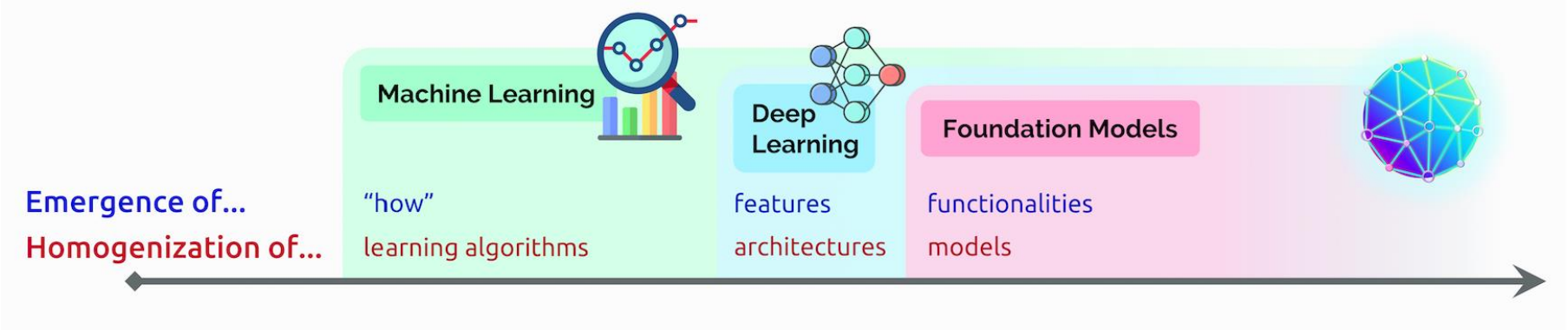
Размер контекста — 100К нуклеотидов

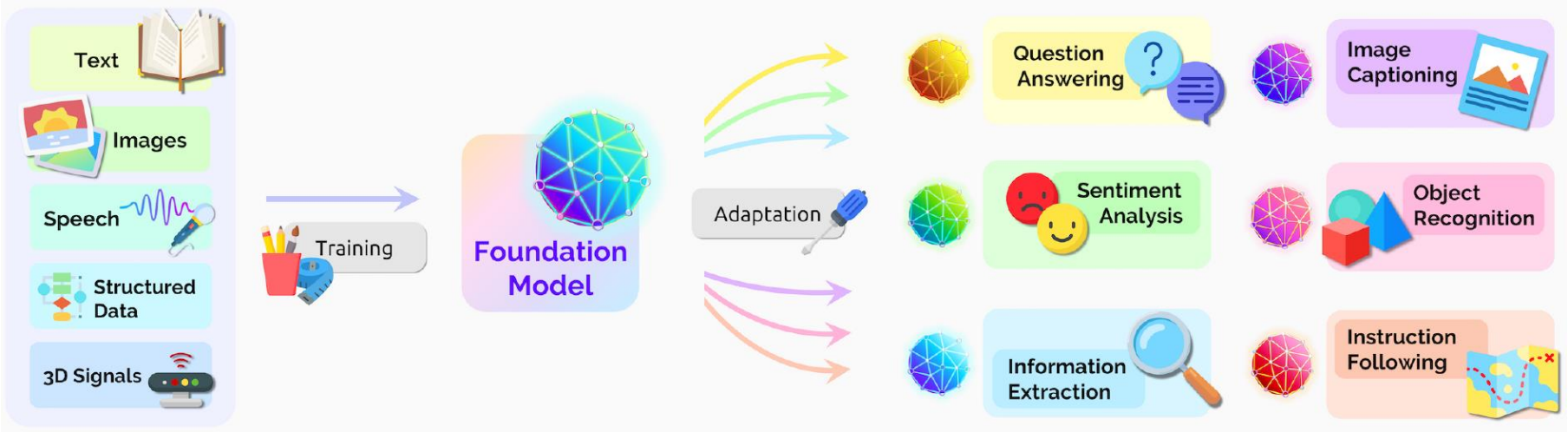




Avsek Z et al. Effective gene expression prediction from sequence by integrating long-range interactions. 2021
Serafim Batzoglou. Large Language Models in Molecular Biology. Towards Data Science, 2023

Фундаментальные модели (Foundation Models)





R.Bommasani et al. (CRFM, Stanford University). On the opportunities and risks of foundation models. 2021

Интеграция данных в Foundation Model

- нуклеотидные последовательности с позиционной разметкой
- типы тканей, клеток — для разметки
- фенотипы человека, нозология, терапия — для разметки
- данные Human Cell Atlas, с (частичной) разметкой путей генов
- экспрессия генов
- метилирование, модификации гистонов, доступность хроматина
- геномика приматов, других видов млекопитающих
- варианты белков, 1D, 2D, 3D структуры
- данные биобанков (UKB и др.)

Serafim Batzoglou. Large Language Models in Molecular Biology. Towards Data Science, 2023-06-02

Смена парадигмы в вычислительной биологии

Почему LLM подходят для анализа омиксных данных *даже лучше,* чем для анализа текстовых коллекций на естественном языке:

- важна возможность обработки больших данных
- важны статистические и причинно-следственные связи
- важен длинный контекст (CNN, трансформеры)
- не нужно генерировать последовательности
- не нужно моделировать логические рассуждения в тексте
- не нужно планирование и агентность (reinforcement learning)
- не нужен общий искусственный интеллект (AGI)

Serafim Batzoglou. Large Language Models in Molecular Biology. Towards Data Science, 2023-06-02

Смена парадигмы в вычислительной биологии

- **Раньше:**
  закономерность → гипотеза → эксперимент → теория
- **Теперь:**
  данные → LLM модель → интерпретация → теория

Bozhen Hu et al. Advances of Deep Learning in Protein Science: A Comprehensive Survey. 2024
Kumar N., Srivastava R. Deep learning in structural bioinformatics: current applications and future perspectives. 2024
Qing Li et al. Progress and Opportunities of Foundation Models in Bioinformatics. 2024
Qiang Zhang et al. Scientific Large Language Models: A Survey on Biological & Chemical Domains. 2024
Yang Z et al. AlphaFold2 and its applications in the fields of biology and medicine. 2023
Zhang S et al. Applications of Transformer-based Language Models in Bioinformatics: A Survey. 2023
Serafim Batzoglou. Large Language Models in Molecular Biology. Towards Data Science, 2023-06-02