Исторические исследования в контексте науки о данных: информационные ресурсы, аналитические методы и цифровые технологии

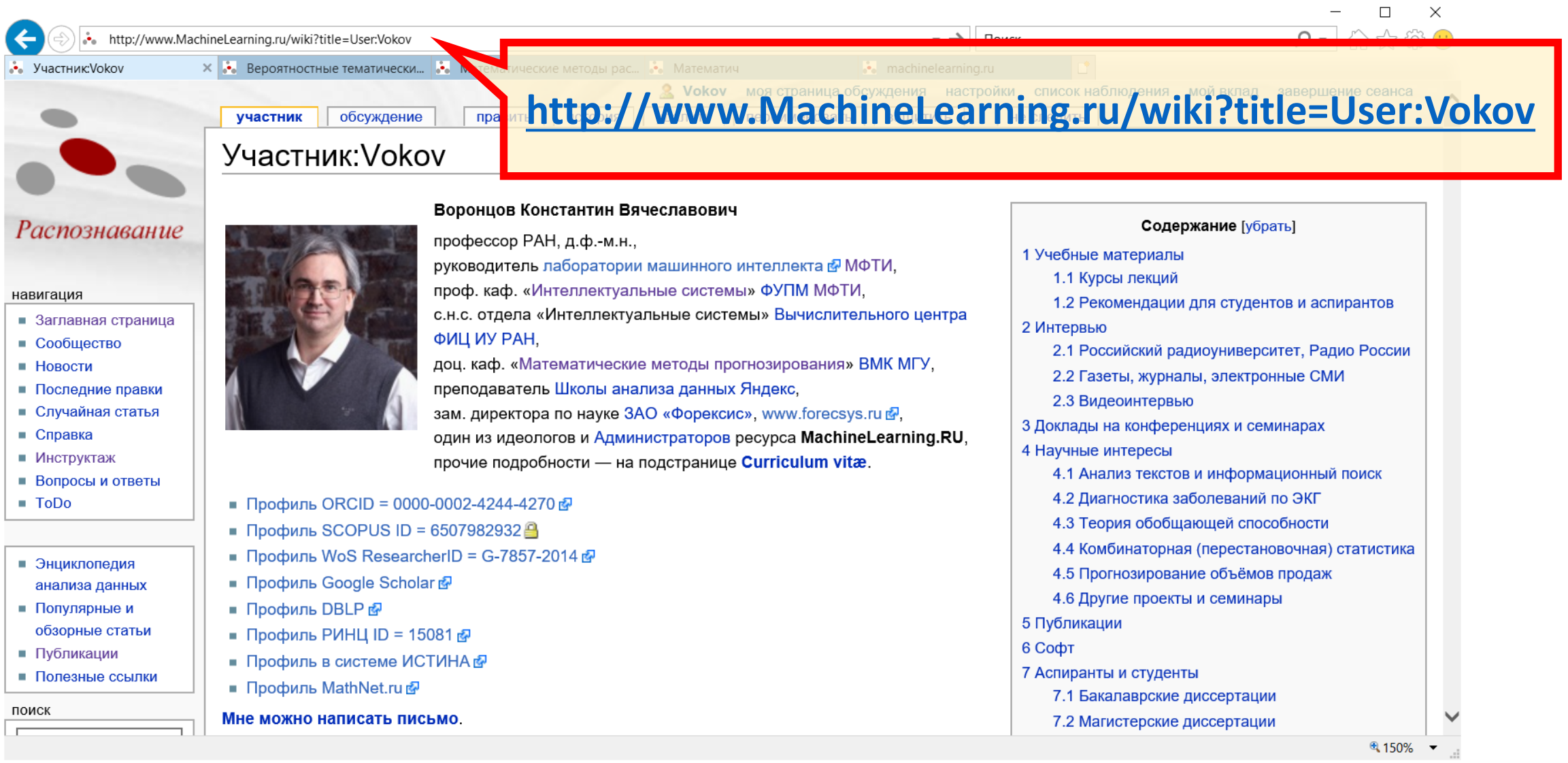
О методологии машинного обучения

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,
руководитель лаборатории Машинного интеллекта МФТИ

k.v.vorontsov@phystech.edu

Докладчик: *Воронцов Константин Вячеславович*



О методологии машинного обучения

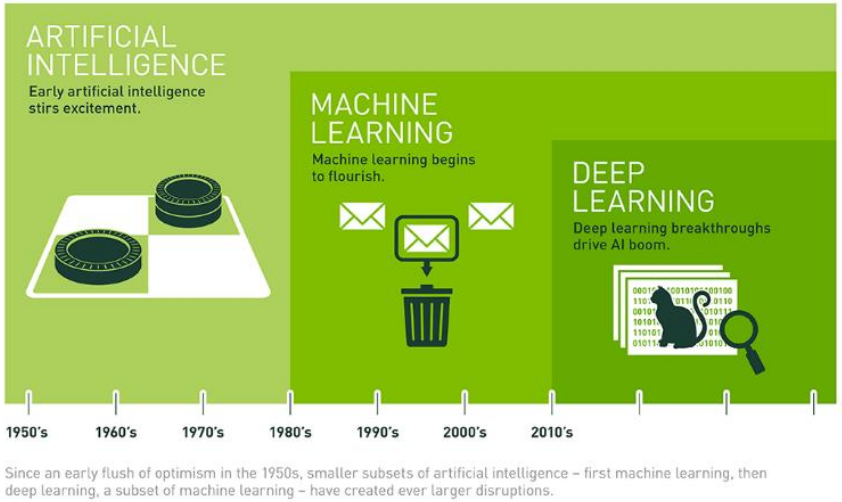
1. Задачи машинного обучения
   - Бум искусственного интеллекта и нейронных сетей
   - Постановки задач и терминология машинного обучения
   - Примеры задач машинного обучения
2. Методология машинного обучения
   - Нейронные сети и глубокое обучение
   - Обучение как задача оптимизации
   - Типология и методология машинного обучения
3. Проблемы и перспективы применения
   - Особенности практического применения DS/AI/ML
   - Необходимые условия применения DS/AI/ML
   - Мифы об искусственном интеллекте

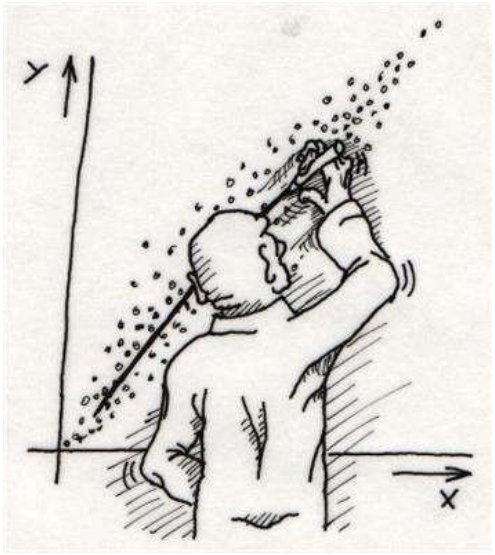
«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, *искусственном интеллекте* и *машинном обучении*» (2016)

Клаус Мартин Шваб,
президент Всемирного
экономического форума

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
- около 100 000 научных публикаций в год

Задача машинного обучения с учителем

Этап №1 – обучение (train)

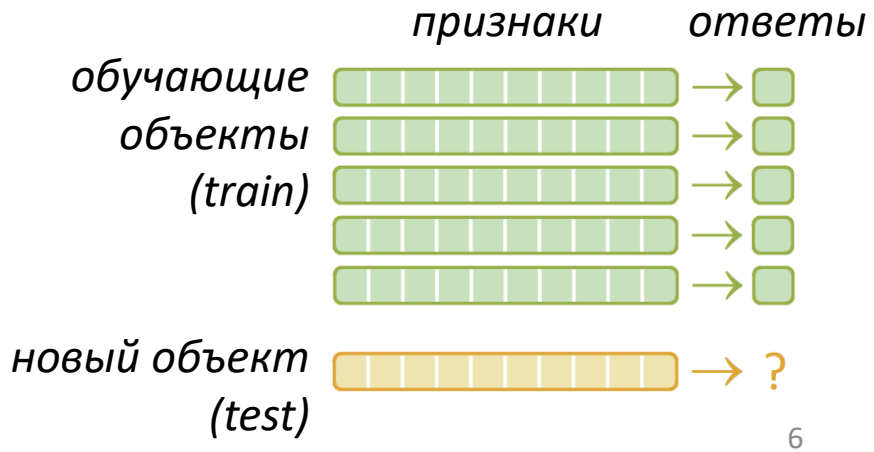
- **На входе:**
  данные – выборка пар «*объект* → *ответ*»,
  каждый объект описывается набором *признаков*
- **На выходе:**
  модель, предсказывающая ответ по объекту

Задача поставлена,
если у неё есть «ДНК»:

- ***Дано***
- ***Найти***
- ***Критерий***

Этап №2 – применение (test)

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

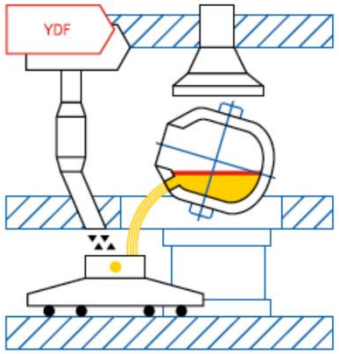


Примеры задач машинного обучения

- **Медицинская диагностика:**
  объект – данные о пациенте на текущий момент
  ответ – диагноз / лечение / риск исхода

- **Поиск месторождений полезных ископаемых:**
  объект – данные о геологии района
  ответ – есть/нет месторождение

- **Управление технологическими процессами:**
  объект – данные о сырье и управляющих параметрах
  ответ – количество/качество полезного продукта



Примеры задач ML в бизнесе

- **Кредитный скоринг:**
  объект – данные о заёмщике
  ответ – решение по кредиту & вероятность дефолта

- **Предсказание оттока клиентов:**
  объект – данные о клиенте на момент времени t
  ответ – уйдёт ли клиент к моменту времени $t + \Delta$

- **Прогнозирование объёмов продаж:**
  объект – данные о продажах на момент времени t
  ответ – объём спроса в интервале от t до $t + \Delta$

Примеры задач ML в интернет-сервисах

- **Информационный поиск в Интернете:**
  объект – данные о паре «запрос и документ»
  ответ – оценка релевантности документа запросу

- **Продажа рекламы в Интернете:**
  объект – данные о тройке «пользователь, страница, баннер»
  ответ – оценка вероятности клика (CTR, Click-Through Rate)

- **Рекомендательные системы в Интернете / TV:**
  объект – данные о паре «пользователь, товар / фильм»
  ответ – оценка вероятности покупки / просмотра

Примеры задач ML в LegalTech

- **Поиск схожей судебной практики:**
  объект – текст иска, акта или обращения заявителя
  ответ – ранжированный список схожих дел

- **Рекомендательный сервис:**
  объект – пара «описание дела, профиль юриста/фирмы»
  ответ – ранжированный список консультантов

- **Предсказание судебного решения:**
  объект – описание дела, документы по делу
  ответ – вероятность выиграть дело

Примеры задач с невекторными данными

- **Статистический машинный перевод:**
  объект – предложение на естественном языке
  ответ – его перевод на другой язык

- **Перевод речи в текст:**
  объект – аудиозапись речи человека
  ответ – текстовая запись речи

- **Компьютерное зрение:**
  объект – динамика сцены в видеопоследовательности
  ответ – решение (объехать, остановиться, игнорировать)

*Прогресс в этих областях связан с «**большими данными**» (англ. «Big Data»)*

…очень важное уточнение:

***С аккуратными** большими данными*

О методологии машинного обучения

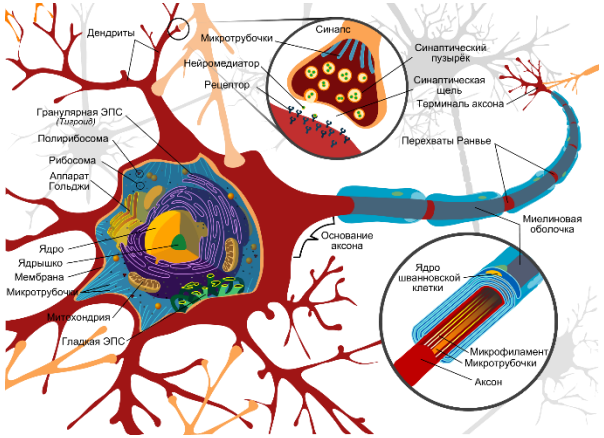
1. Задачи машинного обучения
   - Бум искусственного интеллекта и нейронных сетей
   - Постановки задач и терминология машинного обучения
   - Примеры задач машинного обучения
2. Методология машинного обучения
   - Нейронные сети и глубокое обучение
   - Обучение как задача оптимизации
   - Типология и методология машинного обучения
3. Проблемы и перспективы применения
   - Особенности практического применения DS/AI/ML
   - Необходимые условия применения DS/AI/ML
   - Мифы об искусственном интеллекте

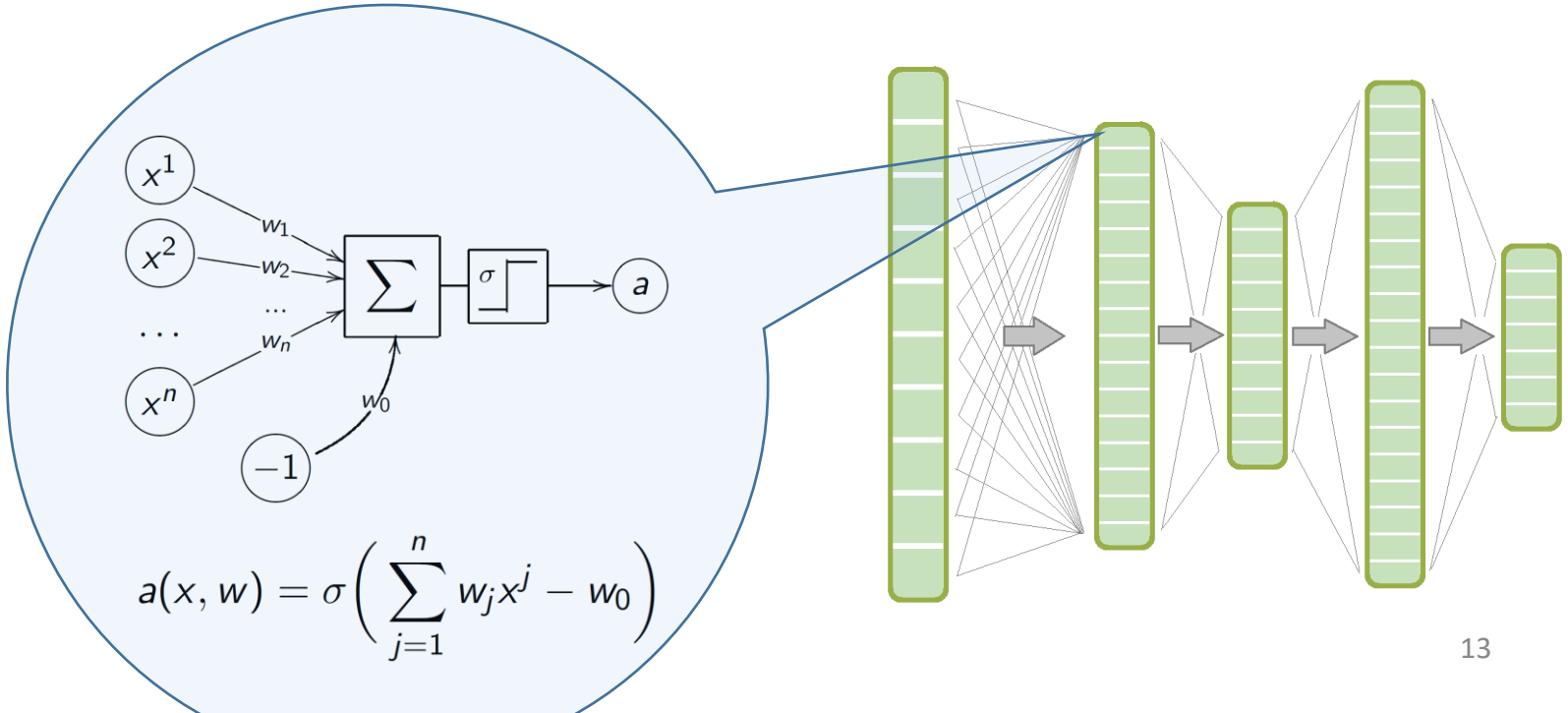
Искусственные нейронные сети

На каждом слое сети вектор объекта преобразуется в новый вектор

Эти преобразования обучаемые, их параметры входят в w

Каждое преобразование (нейрон) – взвешенная сумма признаков

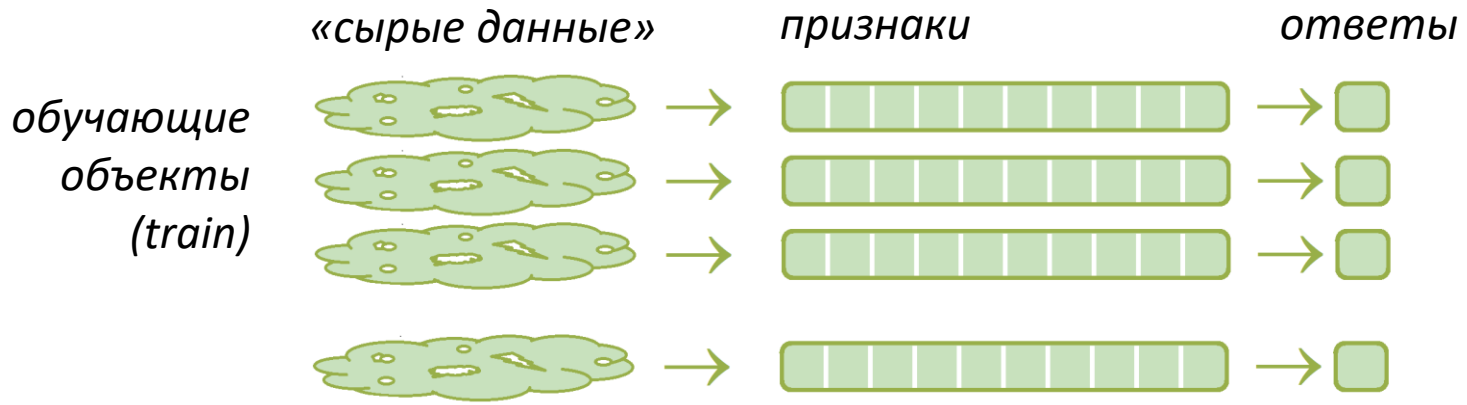




Глубокие нейронные сети

Вход: сложно структурированные «сырые» данные объектов

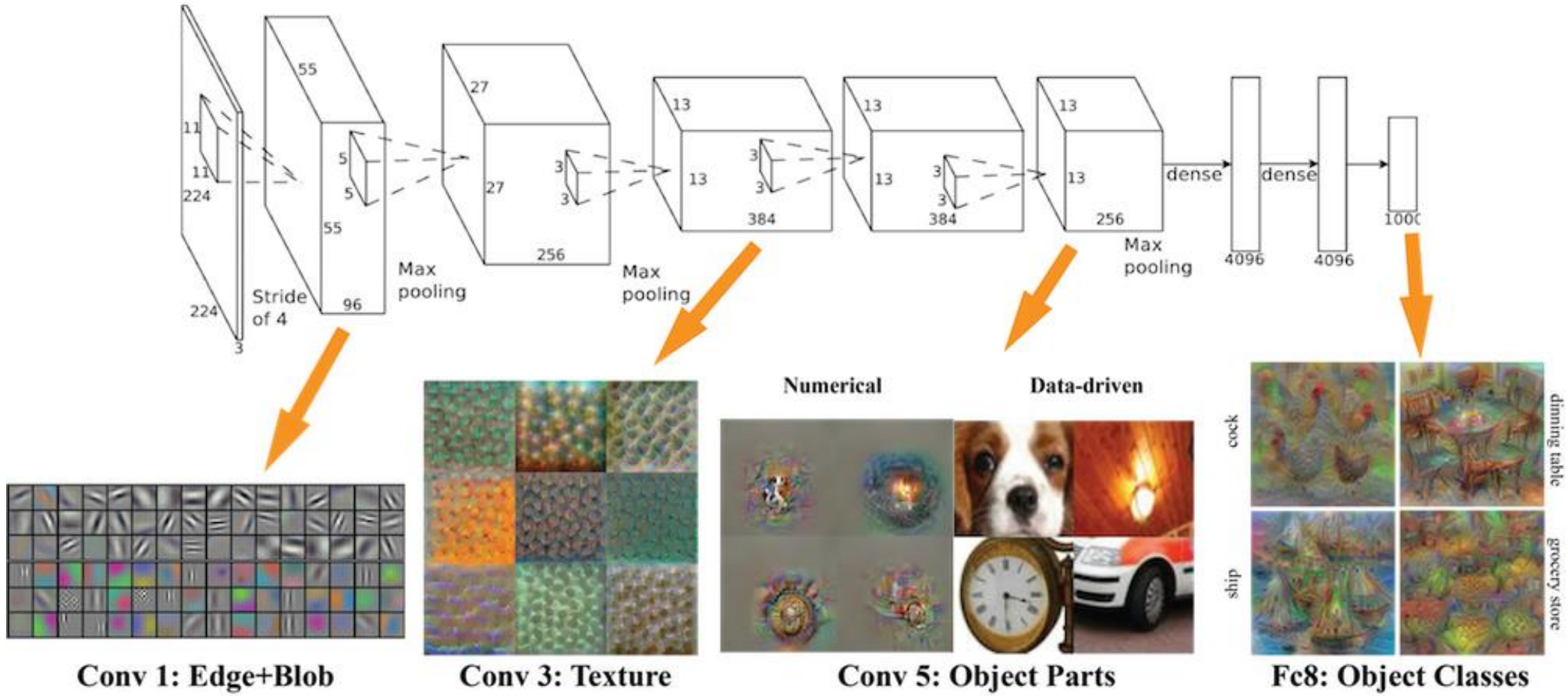
Выход: векторные представления объектов и ответы



Deep Learning – это всего лишь обучаемая векторизация сложных объектов

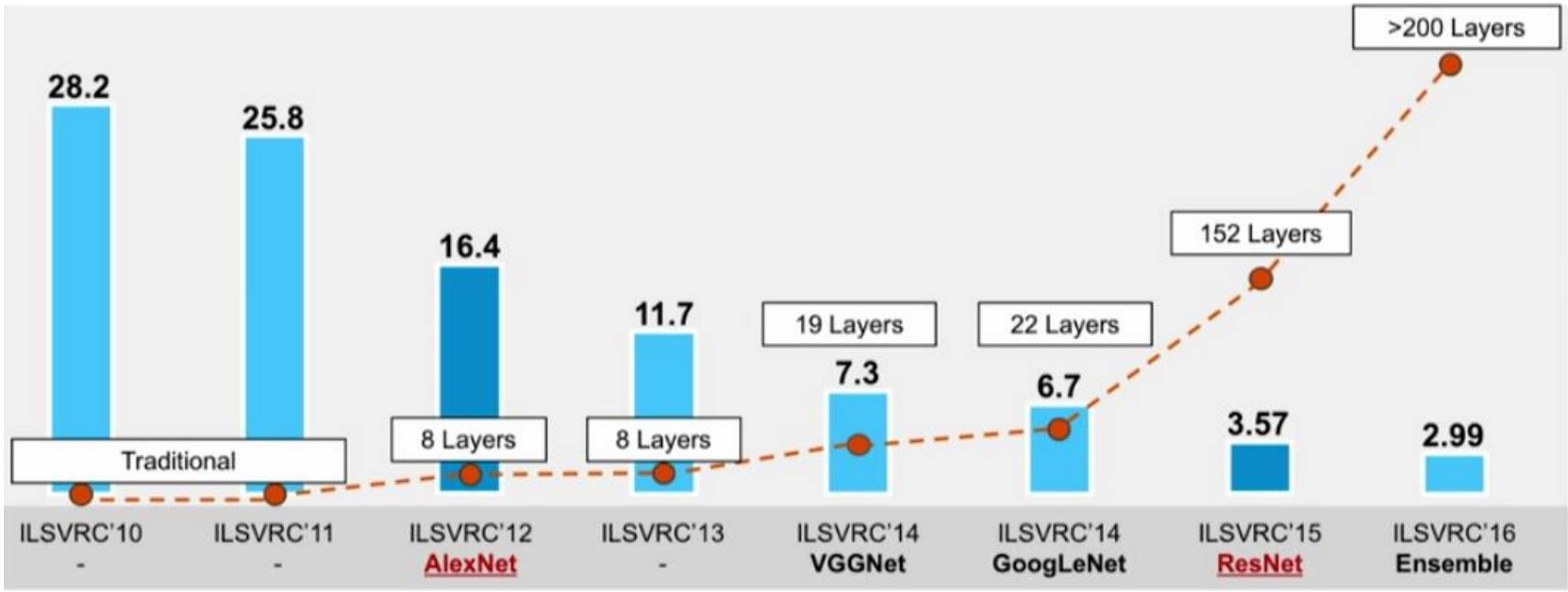
Примеры сложно структурированных объектов:
тексты, изображения, видео, временные ряды, транзакции, графы, ...

Глубокие свёрточные нейронные сети
для классификации изображений



Роль больших данных

ImageNet: открытая выборка 14M изображений, 20K категорий

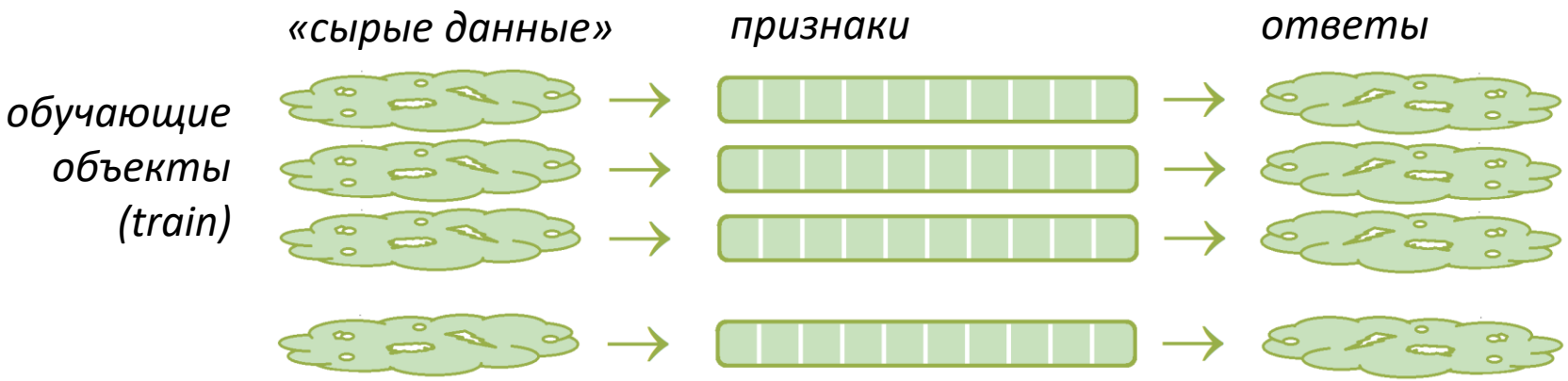


Старт в 2009 г. Человеческий уровень ошибок 5% пройден в 2015 г.

Нейронные сети для синтеза объектов

Вход: сложно структурированные объекты

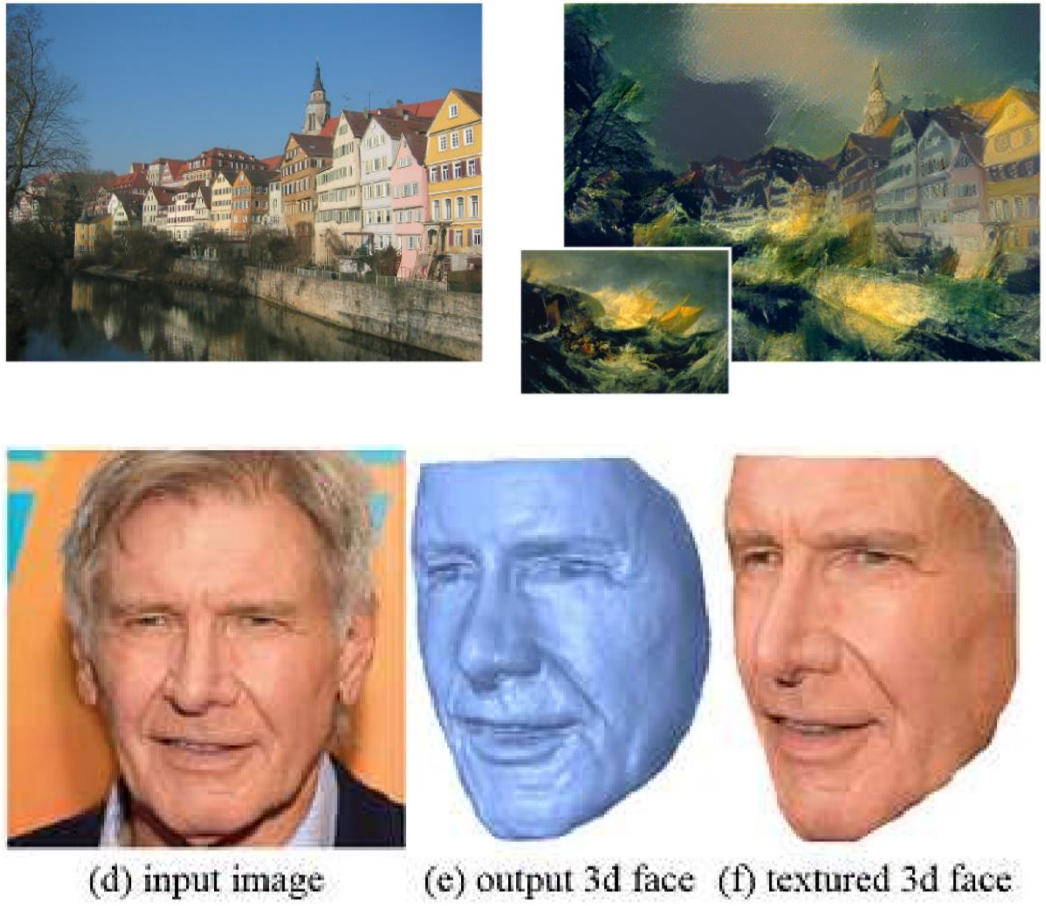
Выход: сложно структурированные ответы

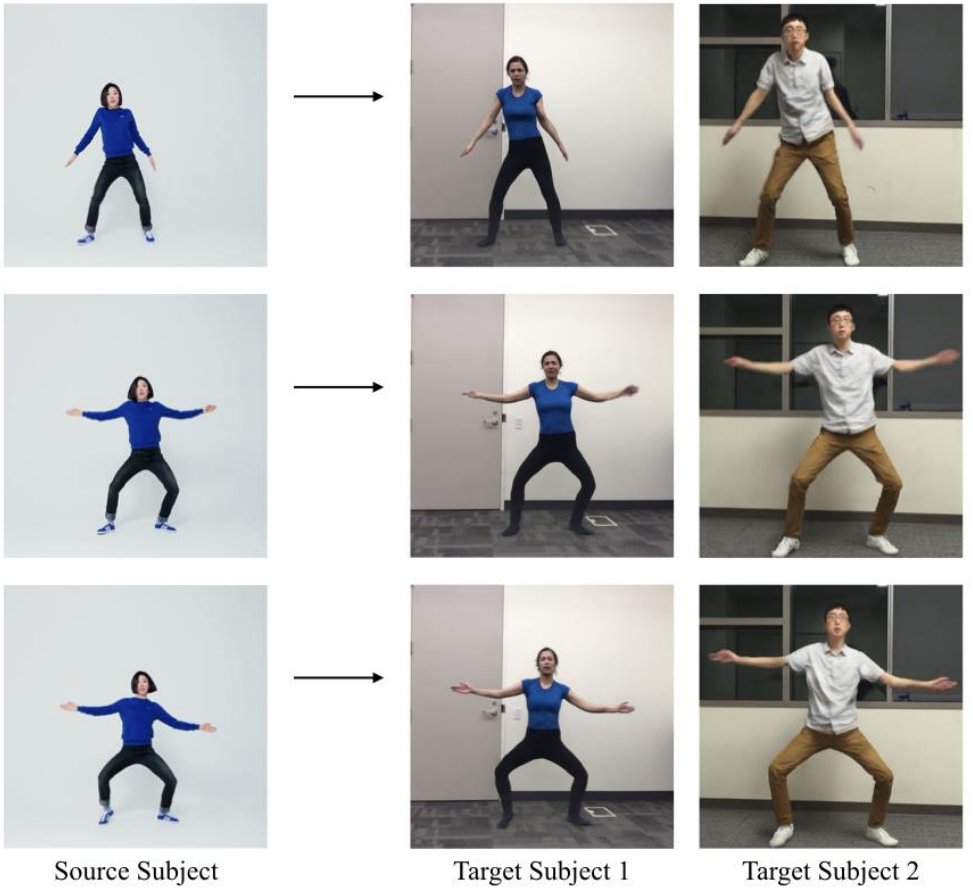


Примеры: синтез изображений, перенос стиля, машинный перевод, суммаризация текстов

Модели: seq2seq, CNN, RNN, LSTM, GAN, BERT, GPT-3 и др.

Синтез изображений и видео





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

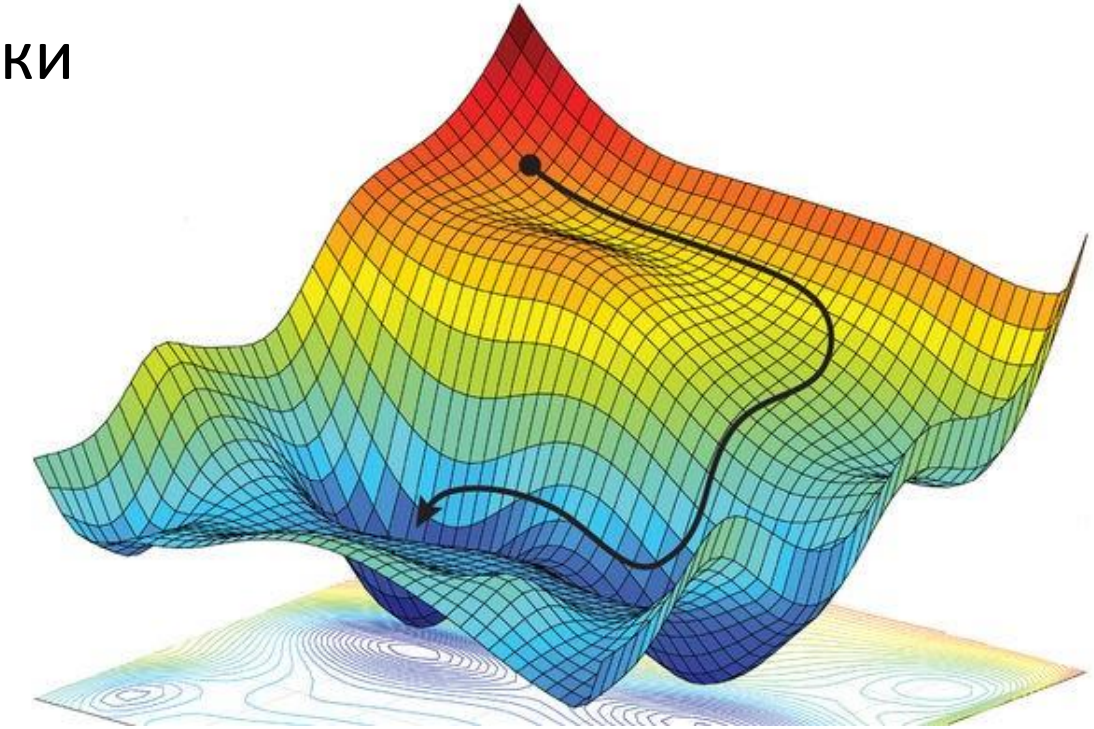
Машинное обучение – это оптимизация

x – вектор объекта обучающей выборки

w – параметры модели

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

$Q(w)$ – критерий качества модели



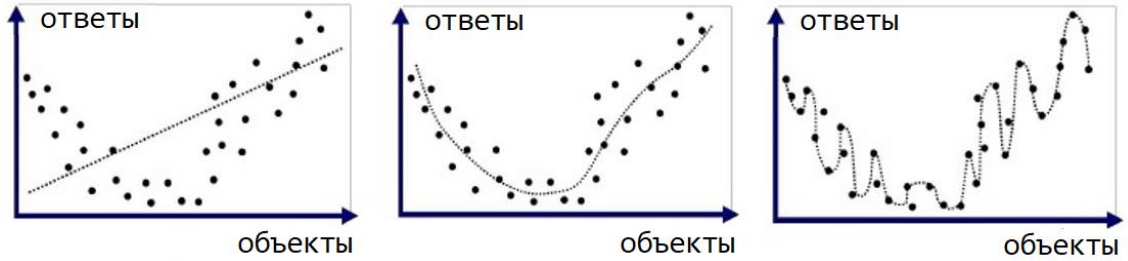
Задача на этапе обучения модели:

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

Способ решения – численные методы оптимизации

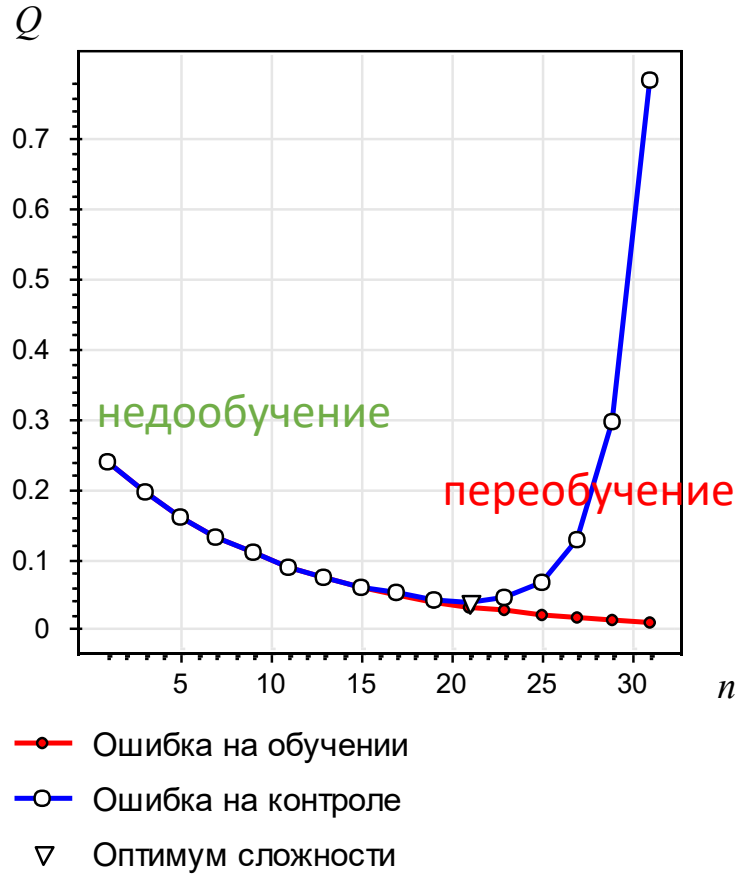
Переобучение – основная трудность ML

Причина переобучения – избыточная сложность модели



недообучение переобучение

- **Внутренние критерии:**
  для оптимизации параметров модели
- **Внешние критерии:**
  для оценивания обобщающей способности модели и контроля *переобучения*



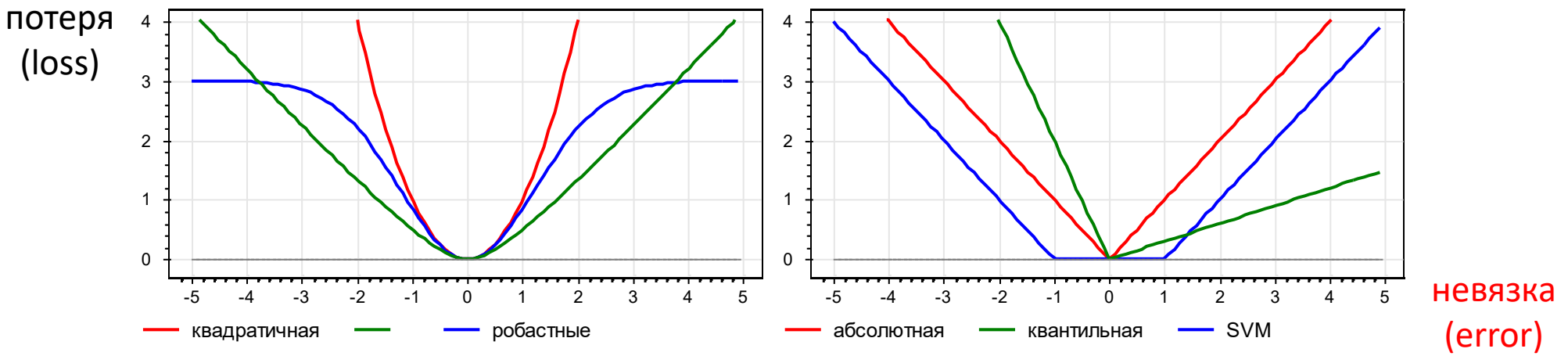
Восстановление регрессии (regression)

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



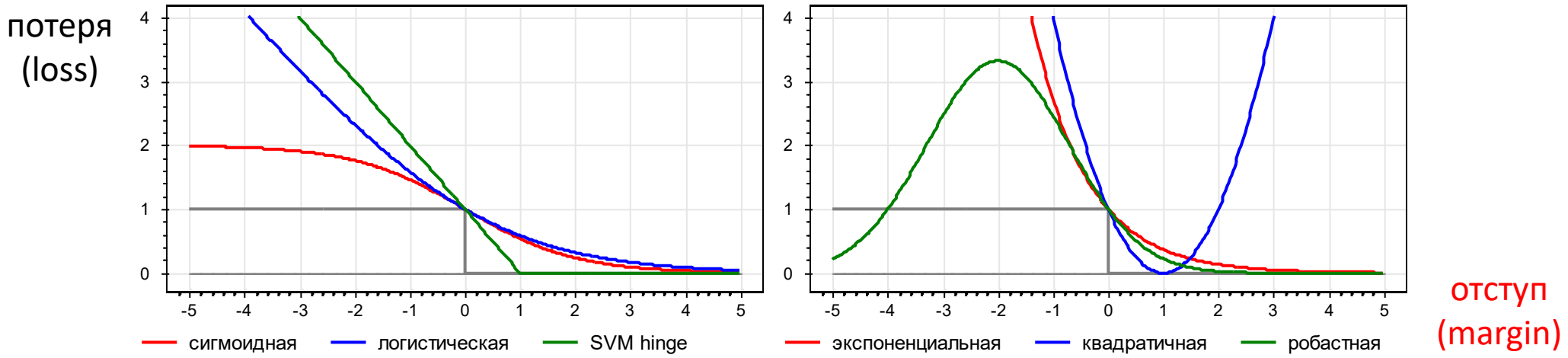
Классификация (classification)

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

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



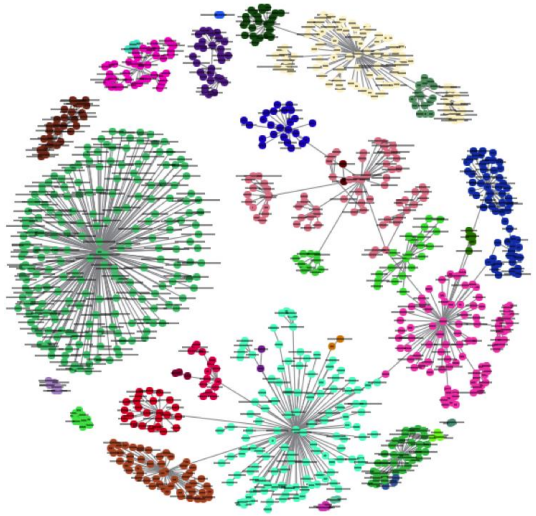
Кластеризация (clustering)

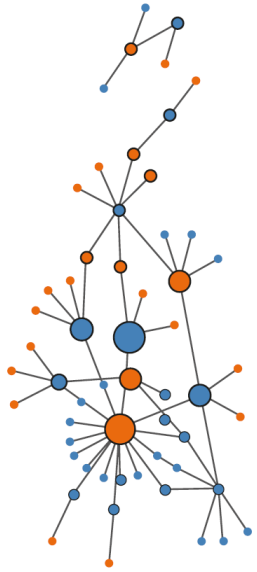
x – вектор объекта обучающей выборки, ответов не дано

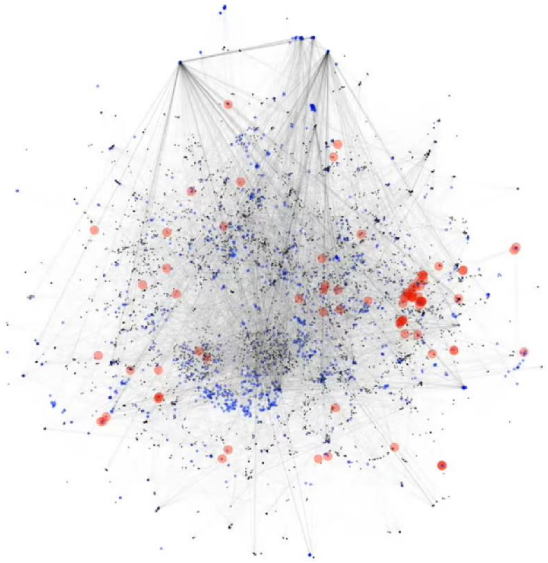
$a(x, w)$ – ближайший к x центр кластера

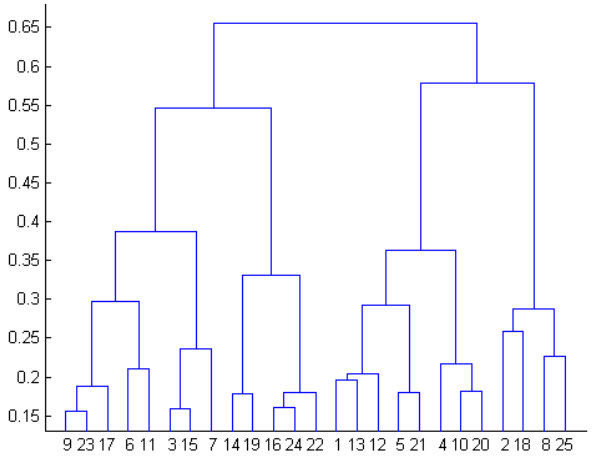
$w = \{c_1, \dots, c_K\}$ – векторы центров всех кластеров

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









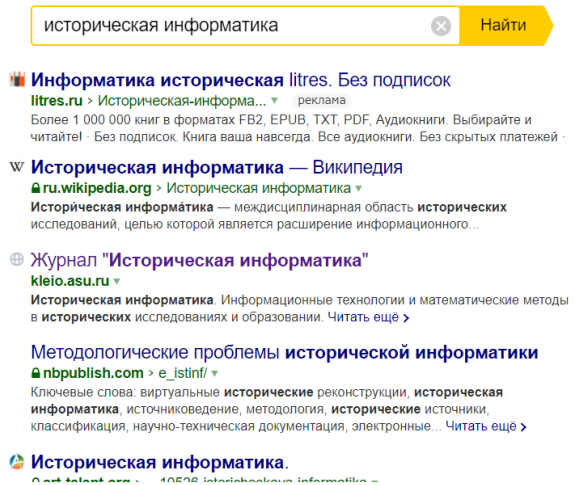
Ранжирование (learning to rank)

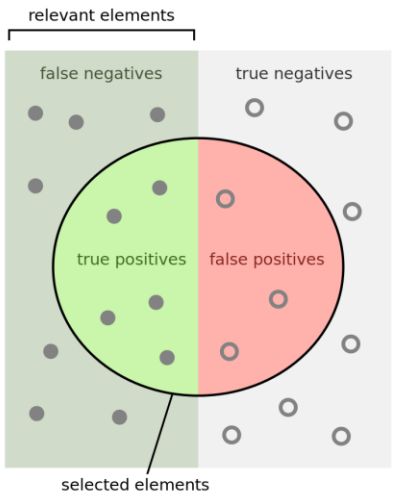
x – вектор пары «запрос-документ», y – оценка релевантности

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

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

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





$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

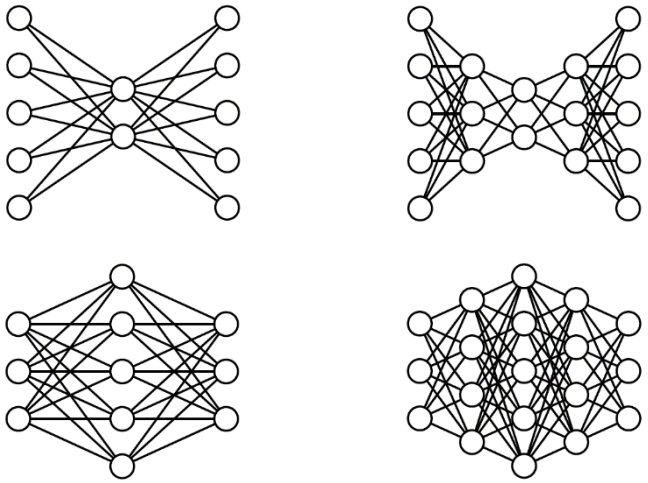
Векторизация объектов (autoencoders)

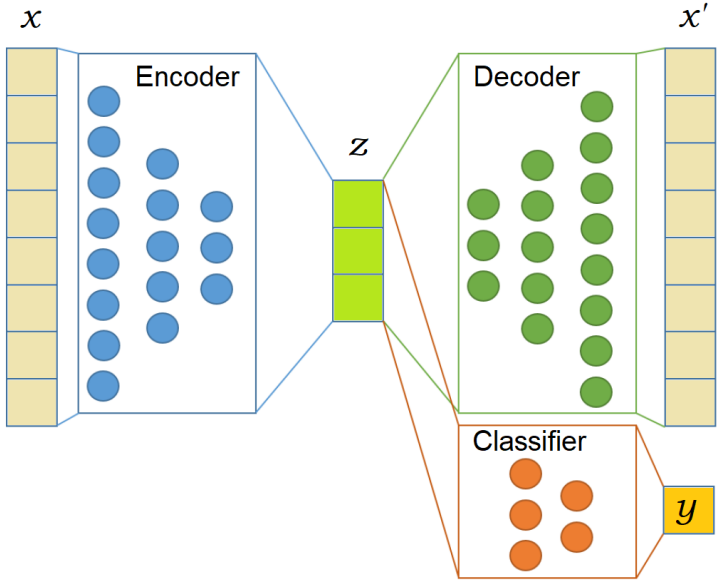
x – вектор объекта обучающей выборки, ответов не дано

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

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

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





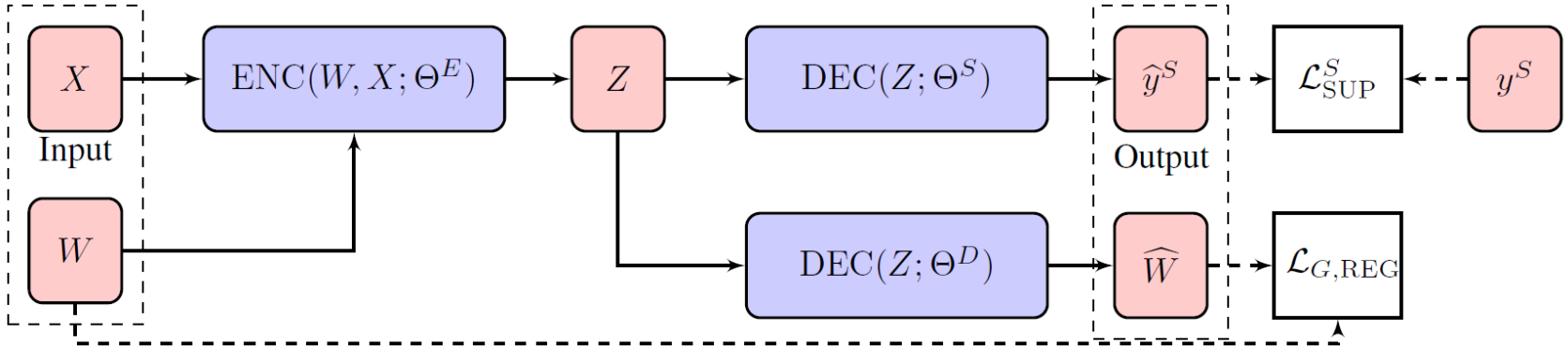
Векторизация графов (graph embeddings)

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

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

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

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



T.Mikolov et al. Ecient estimation of word representations in vector space, 2013.
I.Chami et al. Machine learning on graphs: a model and comprehensive taxonomy. 2020.

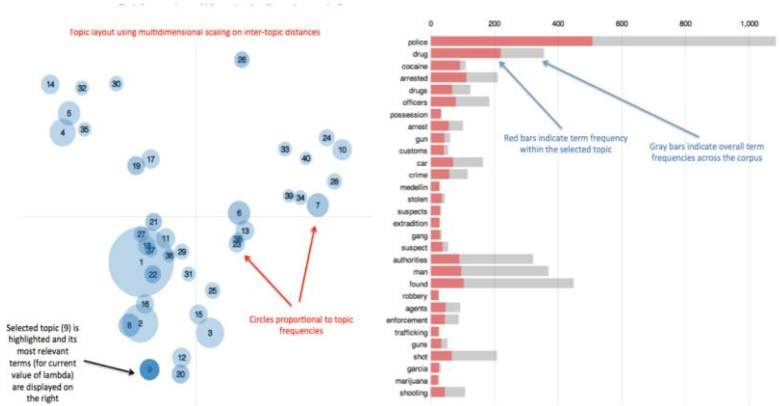
Тематизация текстов (topic modelling)

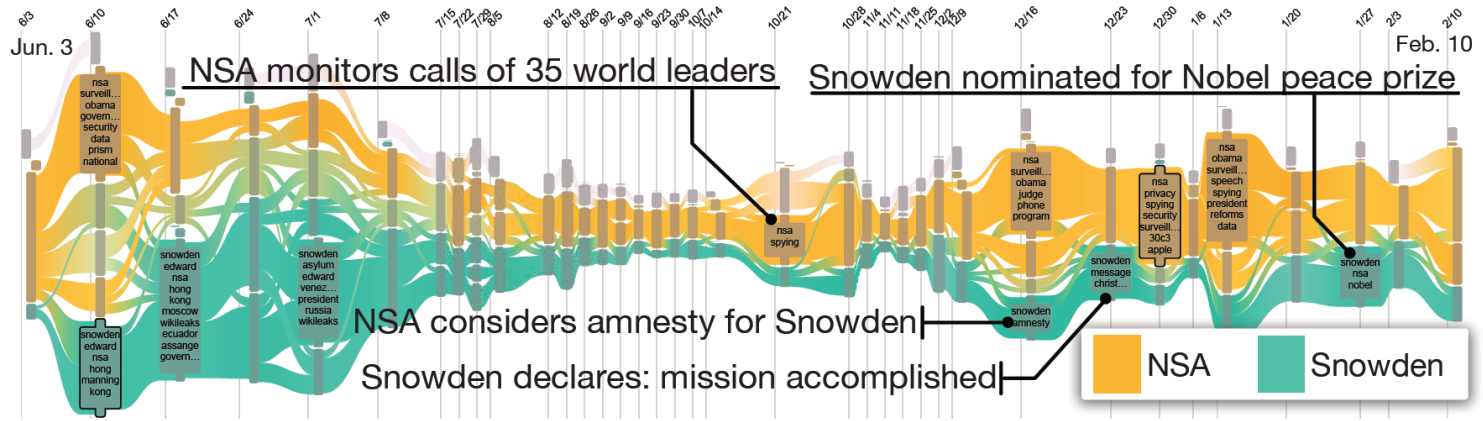
x – текстовый объект, последовательность слов или «мешок слов»

$z = f(x, w)$ – модель распределения текста x по темам $z = p(t|x)$

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

$\text{Loss}(x, w) = \|x'(w) - x\|_{KL}$ – точность реконструкции текста





Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. 2020
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Тематический поиск (exploratory search)

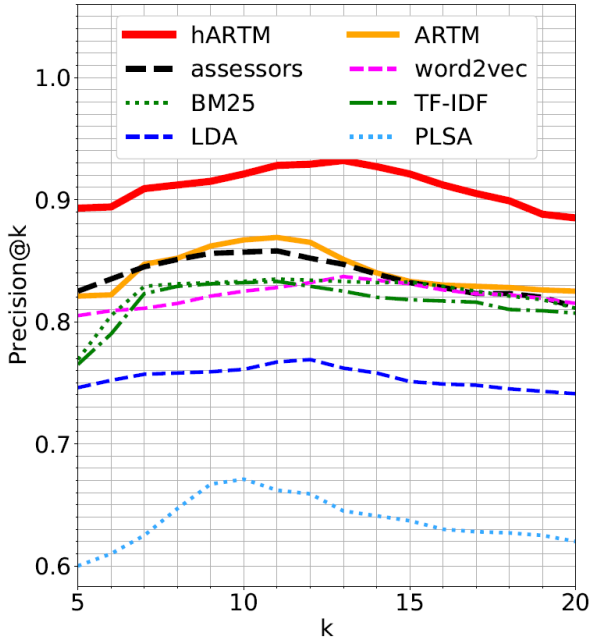
Схема эксперимента:

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

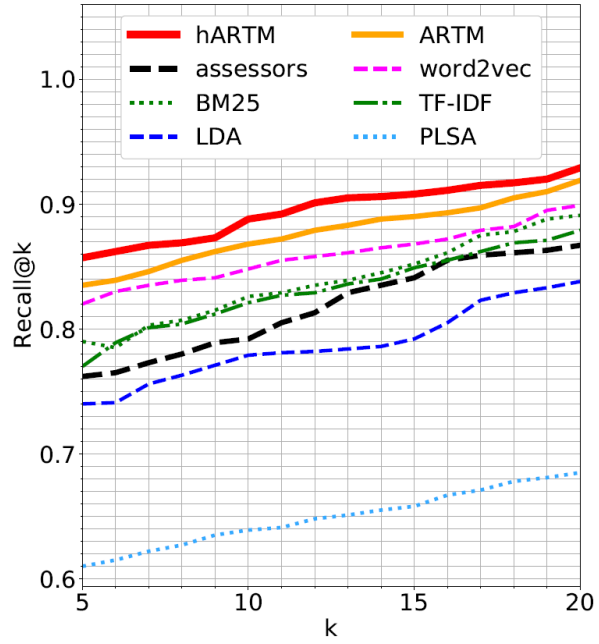
(170K текстов на русском) (750K текстов на английском)

Оценки качества поиска:

точность (precision@k)



полнота (recall@k)



Ianina A., Vorontsov K. Regularized Multimodal Hierarchical Topic Model for Document-by-Document Exploratory Search, 2019.
Ianina A., Golitsyn L., Vorontsov K. Multi-objective topic modeling for exploratory search in tech news. 2017.

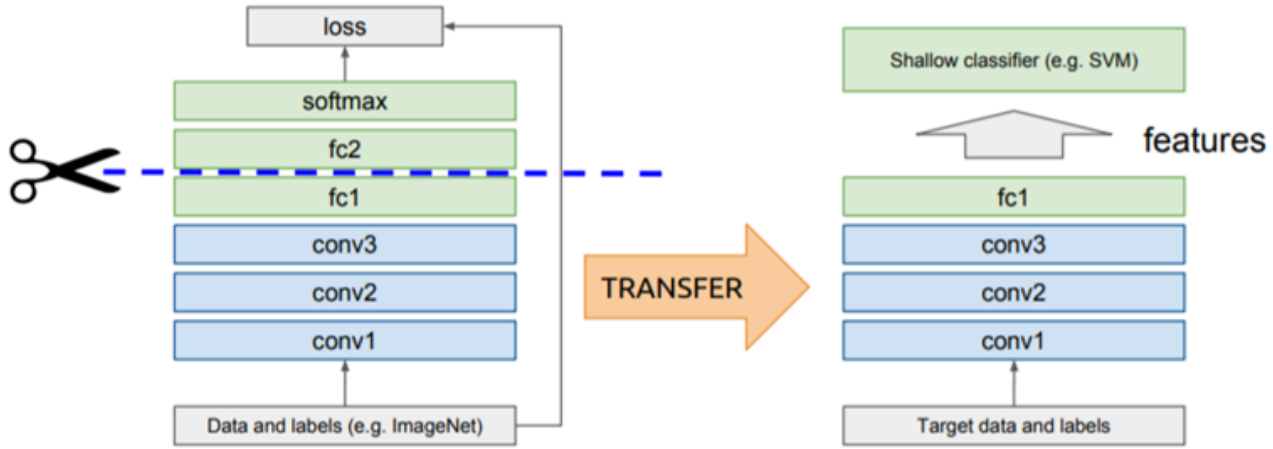
Перенос обучения (transfer learning)

$f(x, w)$ – часть модели, универсальная для всех задач

$g(x, w')$ – часть модели, специфичная для каждой конкретной задач

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

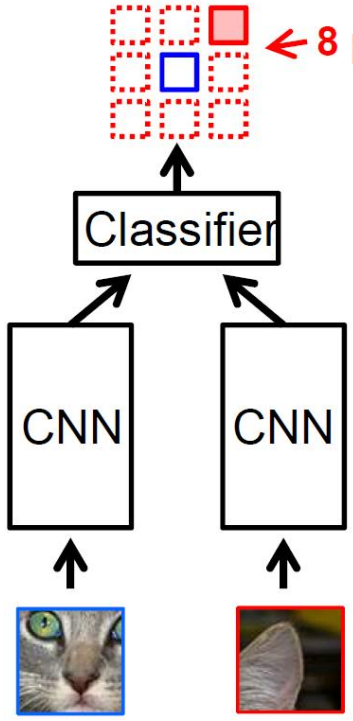
$\min_{w'} \sum_{x'} \text{Loss}_2(f(x', w^*), g_2(x', w'))$ – обучение по своим данным

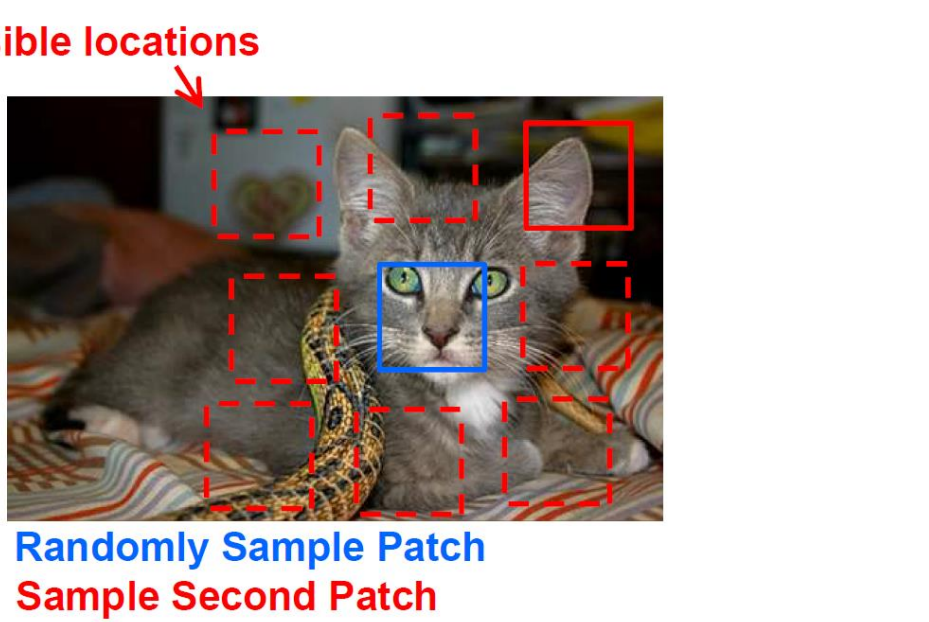


Самостоятельное обучение (self-supervised)

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

Преимущество: сеть выучивает векторные представления объектов без размеченной обучающей выборки





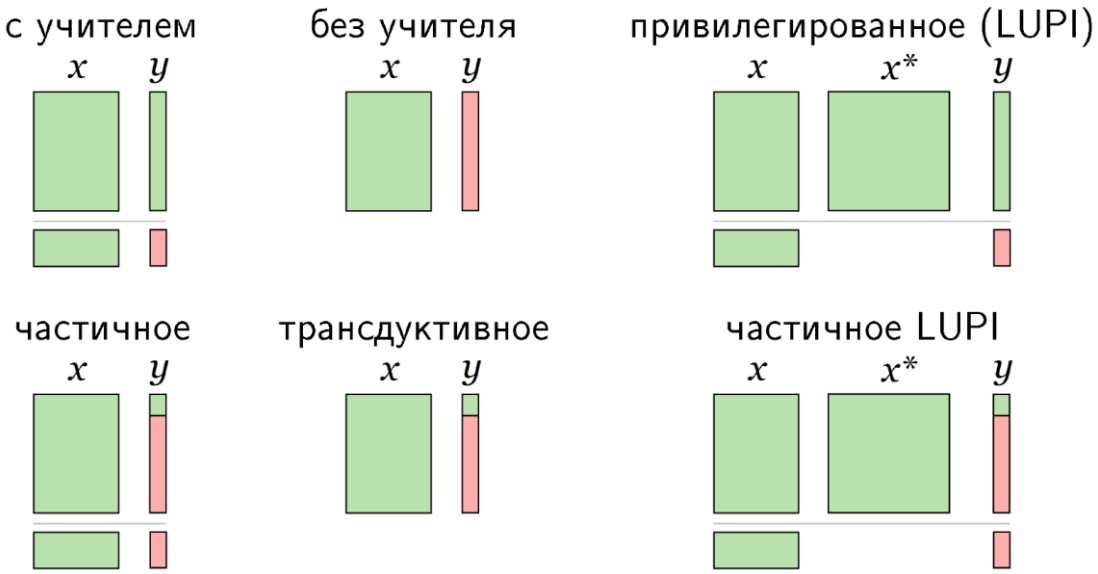
Unsupervised visual representation learning by context prediction, Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Обучение с привилегированной информацией

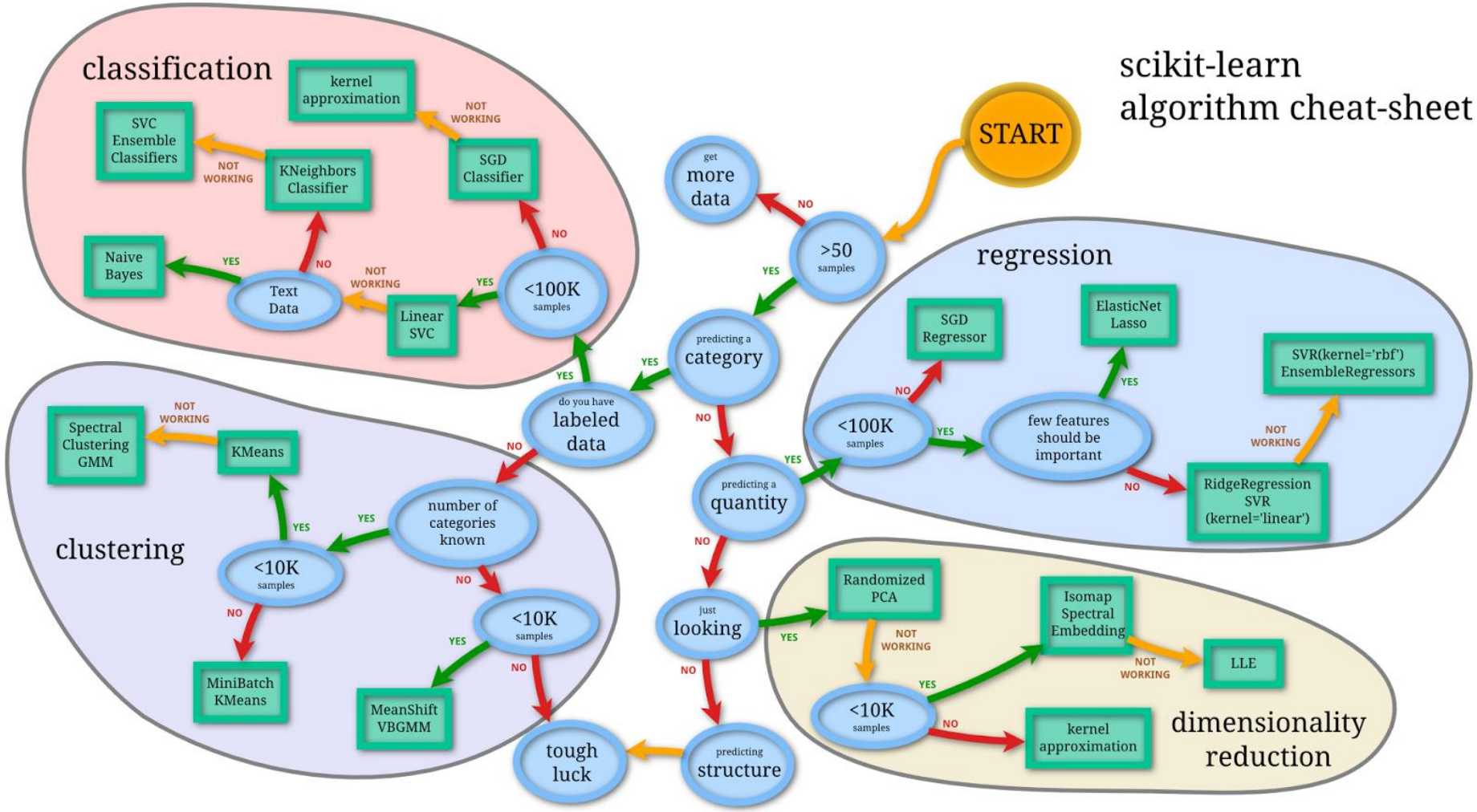
$g(x, x^*, w')$ – модель-учитель, имеет доступ к x^*

$f(x, w)$ – модель-ученик, учится повторять ошибки учителя

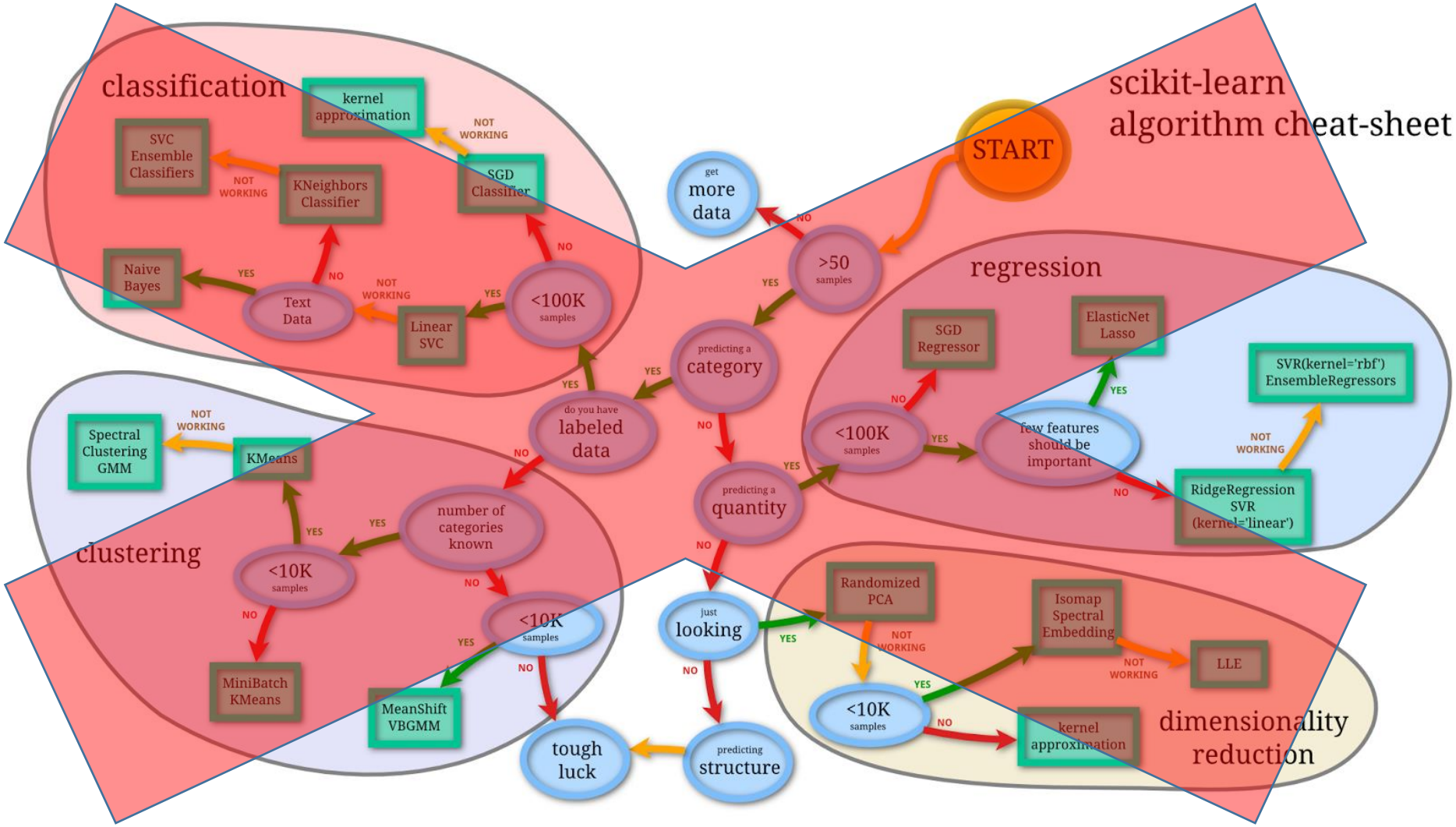
$$\min_{w,w'} \sum_{x} \text{Loss}(f(x,w), y) + \text{Loss}(g(x, x^*, w'), y) + \text{Loss}(f(x,w), g(x, x^*, w'))$$



Задачи и методы машинного обучения



Избегайте пользоваться чужими шпаргалками!



О методологии машинного обучения

1. Задачи машинного обучения
   - Бум искусственного интеллекта и нейронных сетей
   - Постановки задач и терминология машинного обучения
   - Примеры задач машинного обучения
2. Методология машинного обучения
   - Нейронные сети и глубокое обучение
   - Обучение как задача оптимизации
   - Типология и методология машинного обучения
3. Проблемы и перспективы применения
   - Решение прикладных задач на практике
   - Необходимые условия применения DS/AI/ML
   - Мифы об искусственном интеллекте

Особенности реальных данных

В реальных приложениях данные бывают …

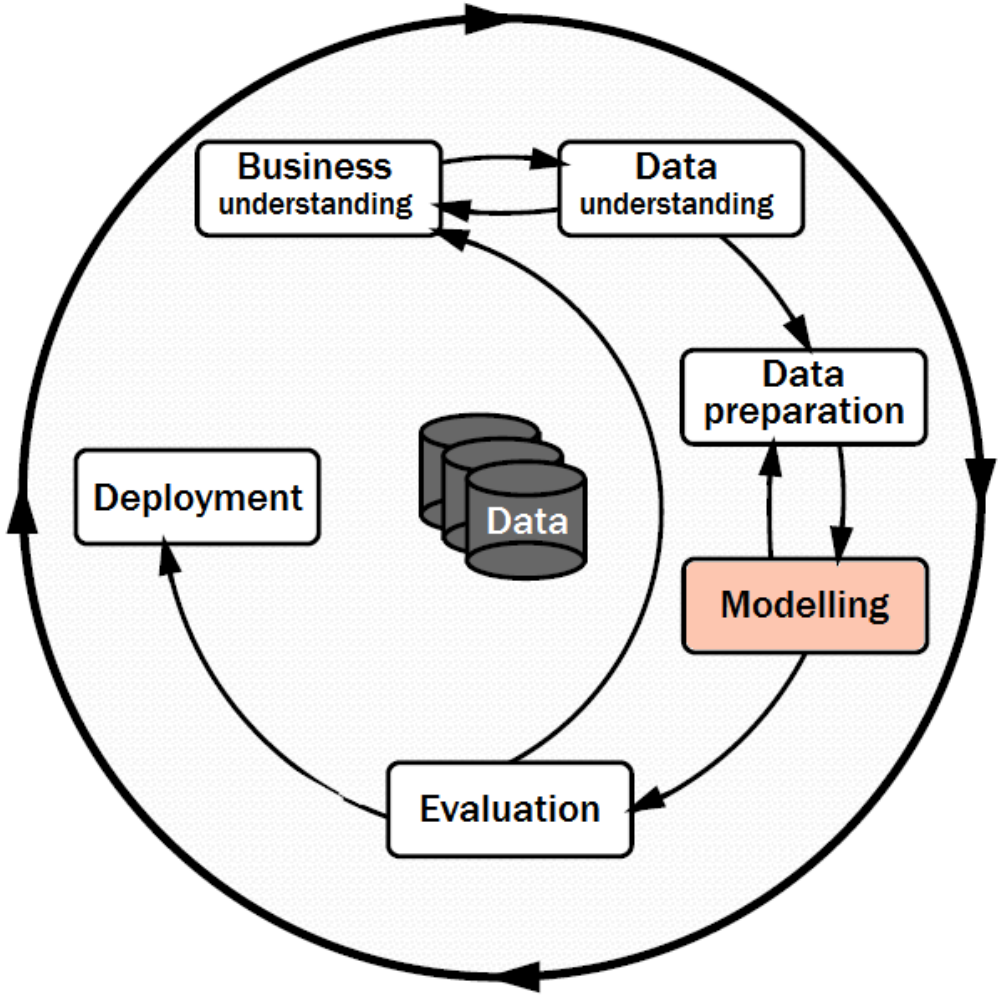
- разнородные (признаки измерены в разных шкалах)
- неполные (признаки измерены не все, имеются пропуски)
- неточные (признаки измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- неструктурированные (нет признаковых описаний)
- «грязные» (ошибочные, грубо не соответствующие истине)

со всем этим можно работать ☺

но только не с грязными данными! ☹

Этапы решения задач ML/DS/AI

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



- понимание прикладной задачи
- понимание данных
- предобработка данных
- инженерия признаков
- построение моделей
- оптимизация параметров
- контроль переобучения
- (кросс-)валидация решения
- внедрение и эксплуатация

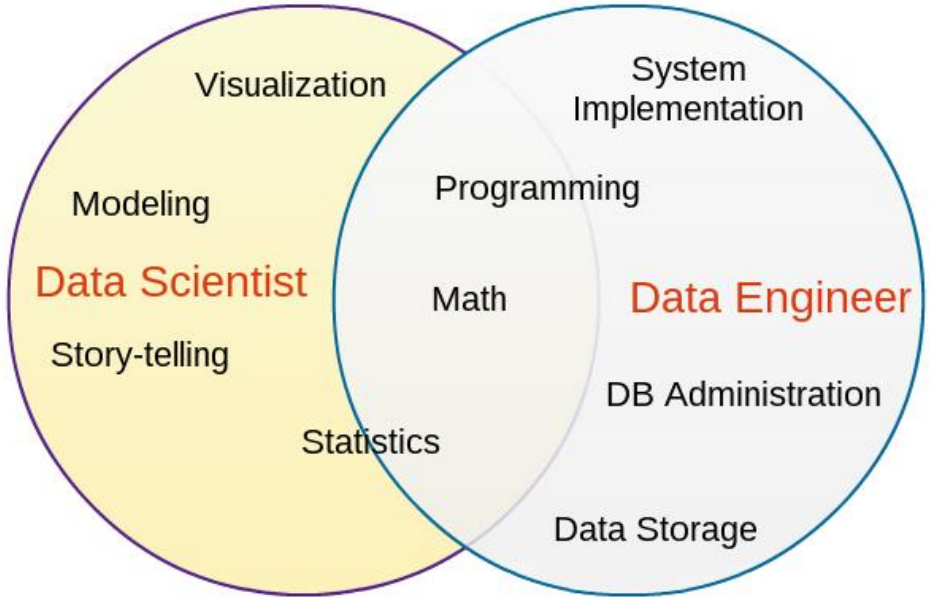
Необходимые условия применения ИИ

- **Полнота, чистота, достоверность данных**
  - Автоматизация и цифровизация процессов, порождающих данные
  - Контроль качества данных (цифровой двойник или «цифровое чучело»?)
  - Трудовая и технологическая дисциплина при работе с данными
- **Культура постановки задач**
  - Предметная экспертиза вместо «абстрактной веры во всемогущий ИИ»
  - Понимание целей анализа и их формализация в критериях качества
  - Готовность пилотировать новые технологии (кто на деле «data-driven»?)
- **Культура анализа данных**
  - Владение средствами визуализации и понимания данных
  - Грамотный и тщательный анализ ошибок при выборе моделей
  - Умение находить «простые но гениальные» решения

Рынок труда в области анализа данных

Инженер по данным **(Data Engineer)**

- Понимает бизнес-процессы, порождающие данные
- Работает с данными в различных форматах
- Визуализирует, понимает, очищает, готовит данные



Исследователь данных **(Data Scientist)**

- Моделирует, строит признаки (feature engineering)
- Выбирает модели и методы, оценивает решения
- Ходит по кругу CRISP-DM

Менеджер проектов **по анализу данных**

- Организует бизнес-процессы сбора и очистки данных
- Видит бизнес задачи и формализует их в терминах «Дано-Найти-Критерий»
- Организует открытые конкурсы и пилотные проекты
- Реалистично оценивает сложность задач и трудозатраты

Миф №1

«Сейчас наблюдается прорыв в области Искусственного Интеллекта»

- Нет, это лишь прорыв в технологиях глубоких нейронных сетей
- Точнее, в технологиях эффективного решения задач численной оптимизации больших размерностей
- Прикладной потенциал этих технологий намного шире «ИИ»

Миф №2

«Скоро будет создан Общий Искусственный Интеллект (AGI)»

- Пока создаётся лишь функциональный ИИ = Имитация Интеллекта
- Artificial Intelligence – это мечта учёных, поэтичное название перспективного научного направления, придуманное в 1955 г.

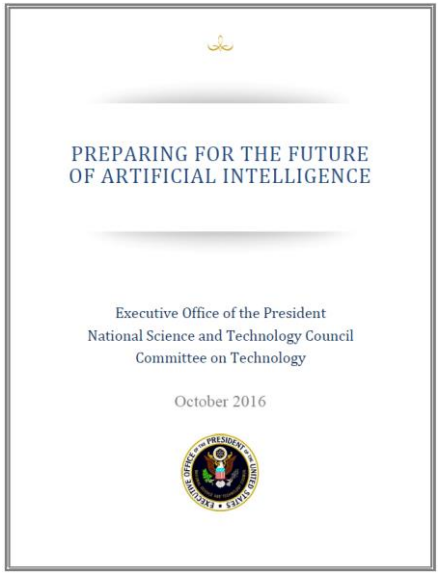
Отличия ИИ от естественного биологического интеллекта:

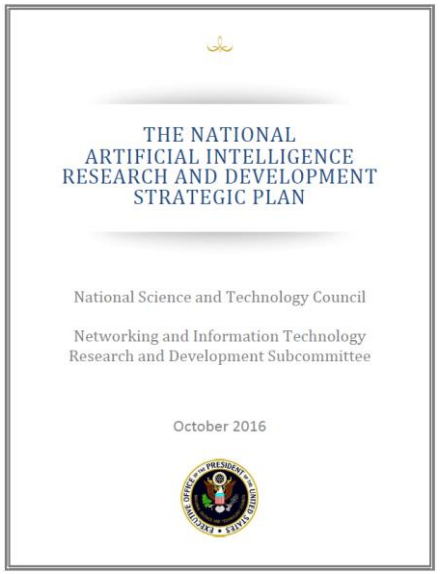
- Мы обучаемся не по выборкам, а на основе объяснений учителей, воспитания, опыта, коммуникации, то есть в естественной среде
- Мы строим картину мира и имеем целеполагание
- У нас 80 млрд. нейронов, и они устроены намного сложнее

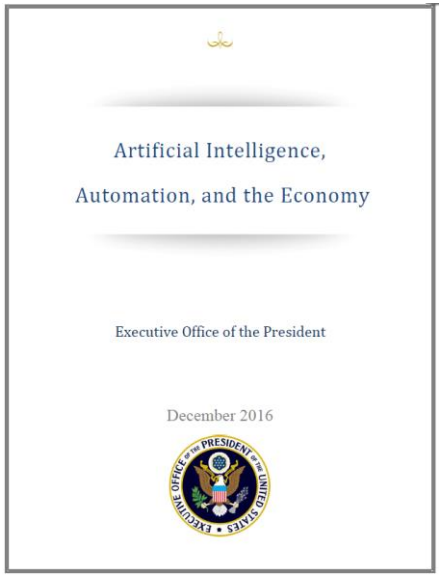
Миф №3

«Тот, кто станет лидером в сфере ИИ, будет властелином мира»

Отчёты Белого дома США, май-октябрь 2016:
«Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future»







Основные выгоды ИИ

- Сокращение издержек и повышение производительности труда
- Автоматизация банковских и финансовых услуг (FinTech)
- Автоматизация юридических услуг (LegalTech)
- Автоматизация посреднической деятельности, распределённая экономика
- Роботизация производств, автономный транспорт
- Оптимизация логистики и цепей поставок
- Оптимизация энергетических и транспортных сетей
- Сенсорные сети, мониторинг сельского хозяйства
- Персональная медицина, улучшение клинических практик
- Персональные образовательные траектории, социальная инженерия
- Автономные системы вооружений

Некоторые из 23 рекомендаций

#1. Организации должны активно развивать партнёрство с научными коллективами для эффективного использования данных.

#2. В приоритетном порядке развивать стандарты *открытых данных* для привлечения научного сообщества к решению задач.

#8. Инвестировать в разработку систем автоматического управления воздушным трафиком.

#11. Вести постоянный мониторинг развития ИИ в других странах.

#13. Приоритетно поддерживать фундаментальные и долгосрочные исследования в области искусственного интеллекта.

#14. Развивать образовательные программы по ИИ и курсы повышения квалификации для прикладных специалистов.

#20. Развивать международную кооперацию по ИИ.

#22. Учитывать взаимовлияние ИИ и кибербезопасности.

Миф №4

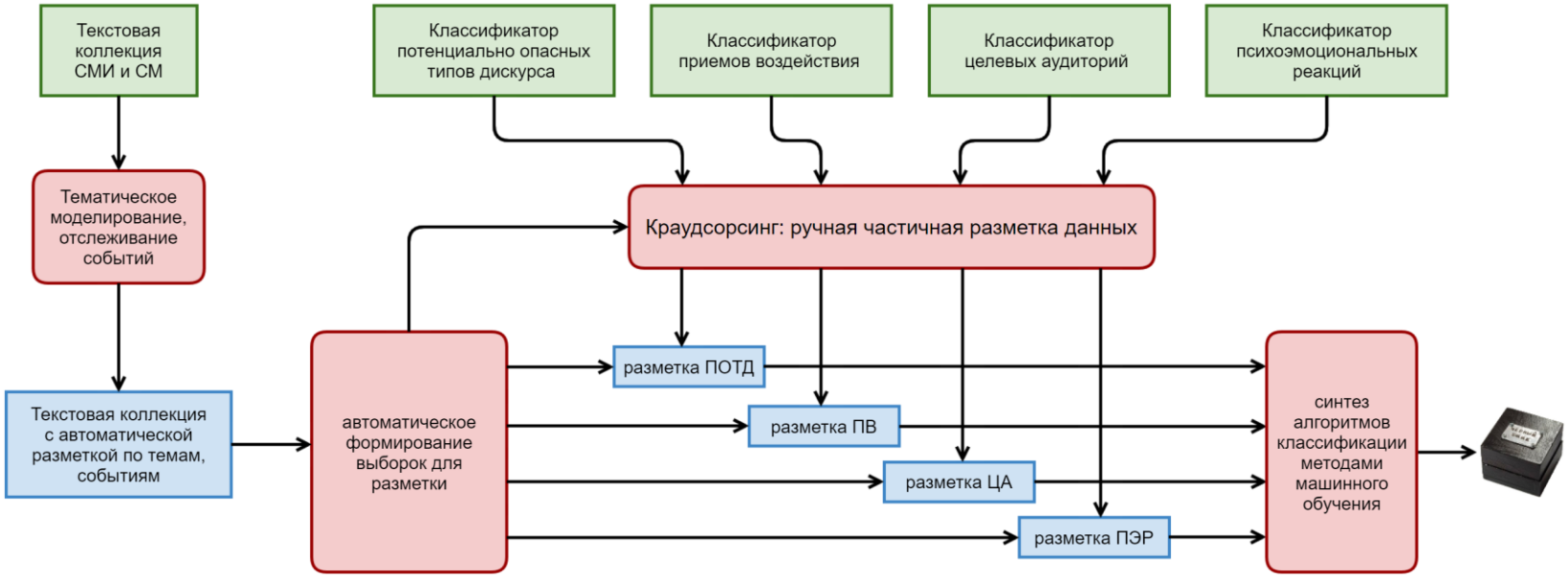
«ИИ является альтернативой классическому моделированию»

- То и другое – восстановление законов природы по данным
- ML: много данных, мало предметных знаний
- ML: универсальные аппроксиматоры для решения разных задач
- ML: модель может работать хорошо, но не понятно, почему
- ML: данные могут быть неполные, неточные, разнородные, …

- **Нет чётких различий! Граница размыта!**

Наши проекты в области анализа текстов

- Инструменты тематического моделирования BigARTM, TopicNet
- Поисково-рекомендательная система [https://arxiv-search.mipt.ru]
- Выявление потенциально опасных типов дискурса, приемов воздействия, целевых аудиторий и их психоэмоциональных реакций



Рекомендуемая литература

- *Бенджио И., Гудфеллоу Я., Курвилль А.* Глубокое обучение. ДМК-Пресс, 2018.
- *Николенко С. И., Кадурин А. А., Архангельская Е. О.* Глубокое обучение. Питер, 2018.
- *Воронцов К. В.* Лекции по машинному обучению. www.MachineLearning.ru, 2004-2020.
- *Коэльо Л. П., Ричарт В.* Построение систем машинного обучения на языке Python. 2016.
- *Мерков А. Б.* Распознавание образов. Введение в методы статистического обучения. 2011.
- *Мерков А. Б.* Распознавание образов. Построение и обучение вероятностных моделей. 2014.
- *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2014.
- *Bishop C. M.* Pattern Recognition and Machine Learning. - Springer, 2006.
- *Домингос П.* Верховный алгоритм. 2016.