Форсайт-сессия учителей информатики «Взгляд в будущее»

Искусственный интеллект и анализ данных – профессия будущего

Воронцов Константин Вячеславович

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

k.v.vorontsov@phystech.edu

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

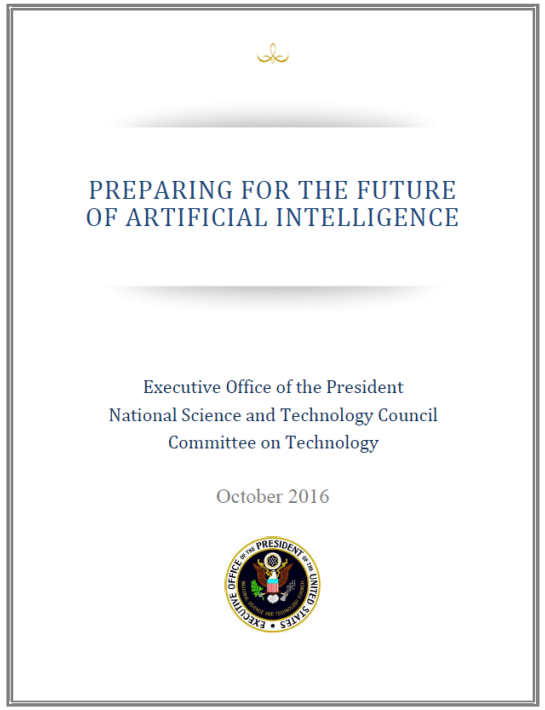
Клаус Мартин Шваб,
президент Всемирного
экономического форума

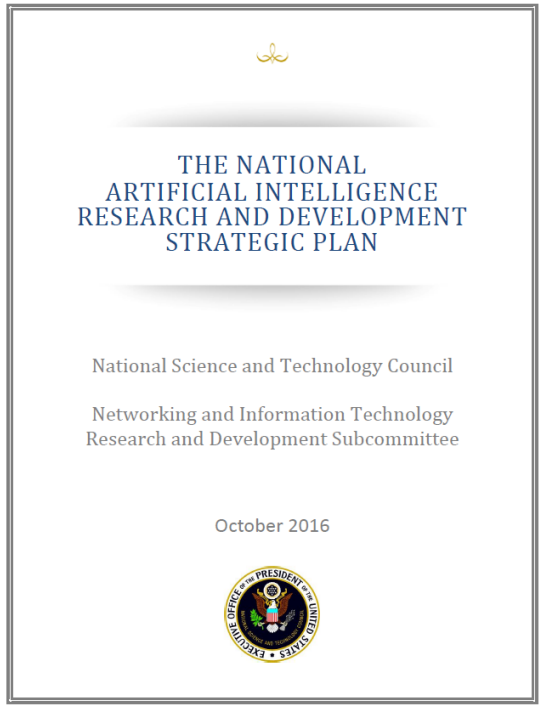
Технологии ИИ, которые меняют мир

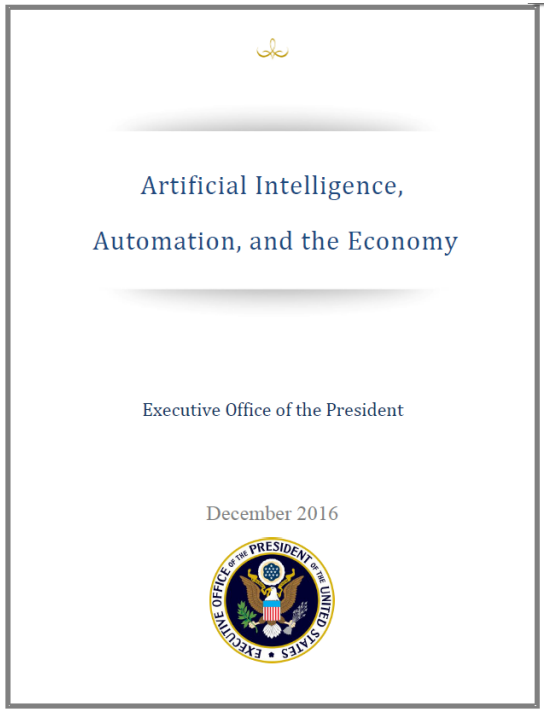
Яндекс Найти

Google

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.

Бум искусственного интеллекта

1997: IBM Deep Blue обыграл чемпиона мира по шахматам

2005: Беспилотный автомобиль: DARPA Grand Challenge

2006: Google Translate – статистический машинный перевод

2011: 40 лет DARPA CALO привели к созданию Apple Siri

2011: IBM Watson победил в ТВ-игре «Jeopardy!»

2011–2018: ImageNet: 25% → 2,5% ошибок против 5% у людей

2015: Фонд OpenAI в $1 млрд. Илона Маска и Сэма Альтмана

2016: DeepMind, OpenAI: динамическое обучение играм Atari

2016: Google DeepMind обыграл чемпиона мира по игре го

2017: OpenAI обыграл чемпиона мира по компьютерной игре Dota 2

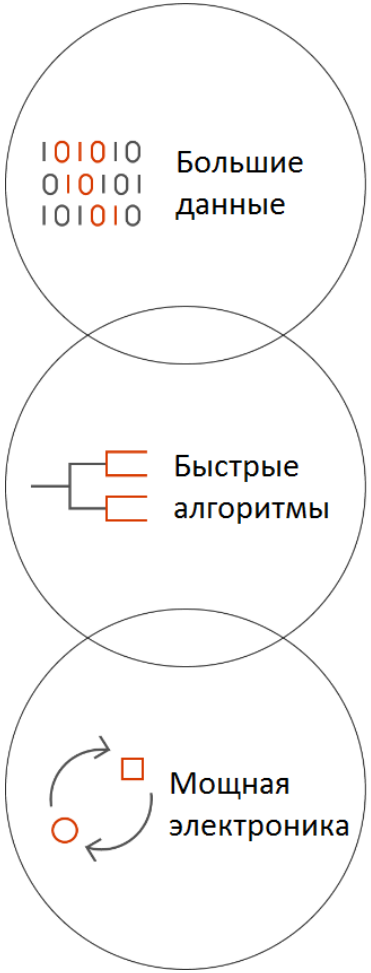
Три предпосылки этого бума

– три перехода количества в качество:

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

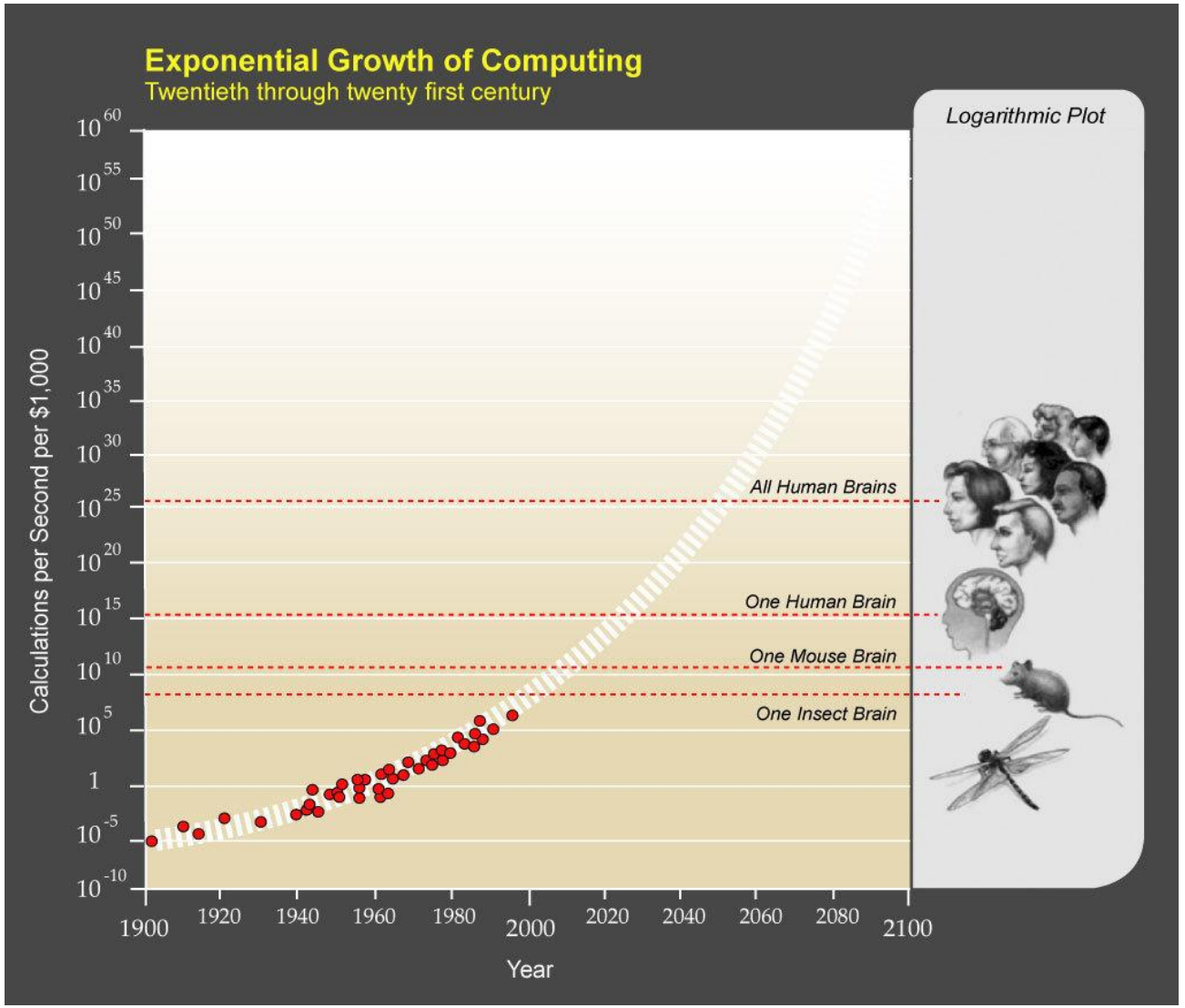
- Развитие математических методов и алгоритмов
  → *накопление критической массы опыта*
  в частности, Deep Neural Networks

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



Закон Мура

Закон ускоряющейся отдачи
(Рэймонд Курцвейл)



Глубокие нейронные сети обеспечили прорыв в компьютерном зрении

ImageNet: открытая выборка
15M размеченных изображений

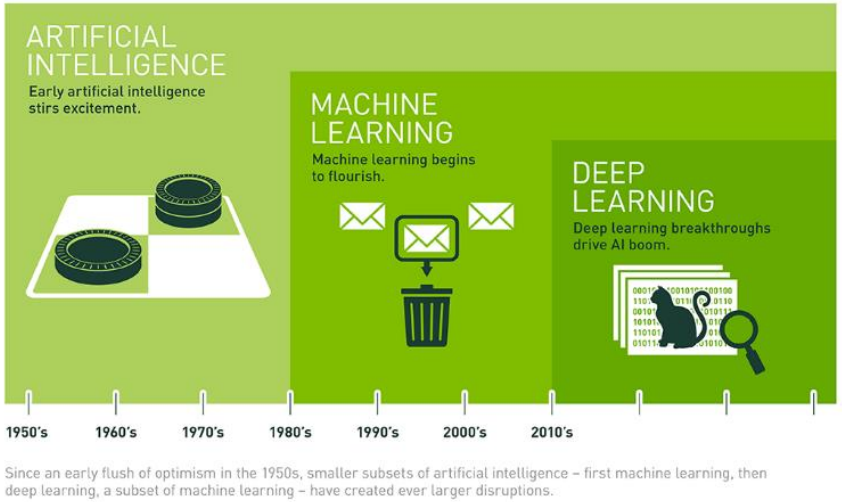
Google: Распознавание кадров с котами на видео из Youtube

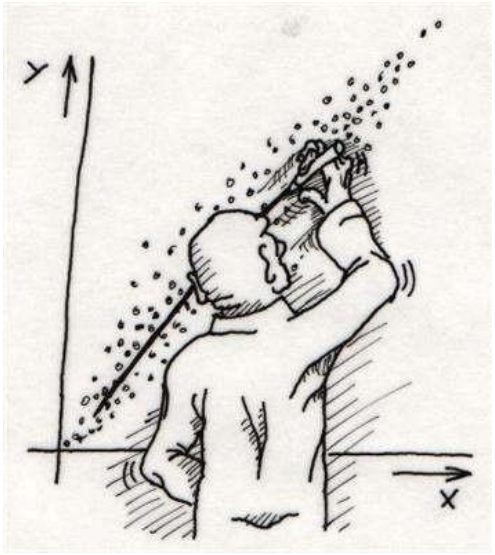
Машинное обучение, большие данные «и много других страшных слов»

- Статистический анализ данных (Statistical Data Analysis)
- Искусственный интеллект (Artificial Intelligence) 1955
- Распознавание образов (Pattern Recognition)
- Машинное обучение (Machine Learning) 1959
- Статистическое обучение (Statistical Learning)
- Интеллектуальный анализ данных (Data Mining) 1989
- Машинный интеллект (Machine Intelligence) 2000
- Бизнес-аналитика (Business Intelligence, Business Analytics)
- Предсказательная аналитика (Predictive Analytics) 2007
- Большие данные (Big Data) 2008
- Аналитика больших данных (Big Data Analytics)
- Наука о данных (Data Science) 2011

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

Основная задача машинного обучения

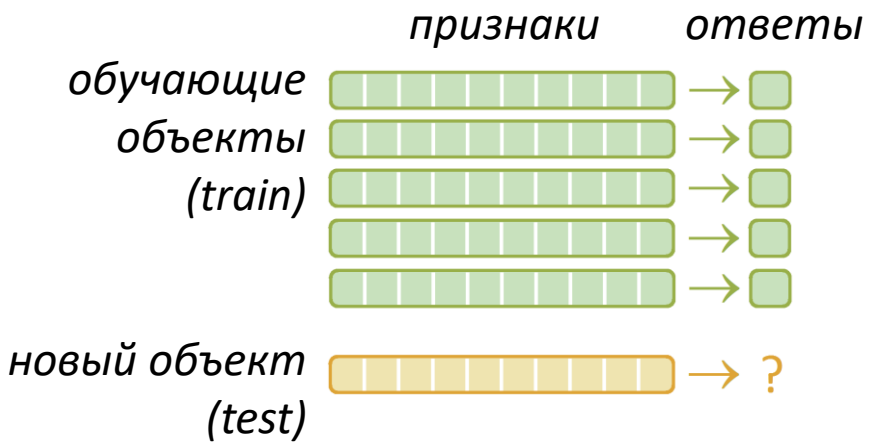
Этап №1 – обучение с учителем

- **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

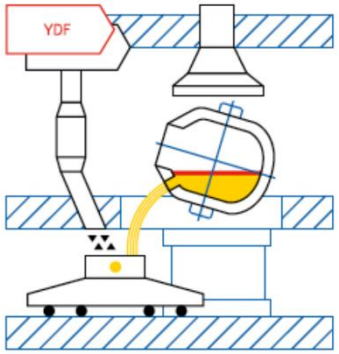


Примеры задач машинного обучения

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

- **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 машинного обучения

- **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

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

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*

Несколько выводов и вопросов

- Анализ данных – повсюду
- Анализ данных – удобная точка входа во многие науки и профессии (данные легко «потрогать руками»)
- Анализ данных всегда междисциплинарен (математика + информатика + предмет)

но…

- Где взять данные и постановки задач, чтобы учиться и учить?
- Возможно ли применять методы, не зная, как они устроены?
- Что из анализа данных возможно преподавать в школе?

Открытые данные

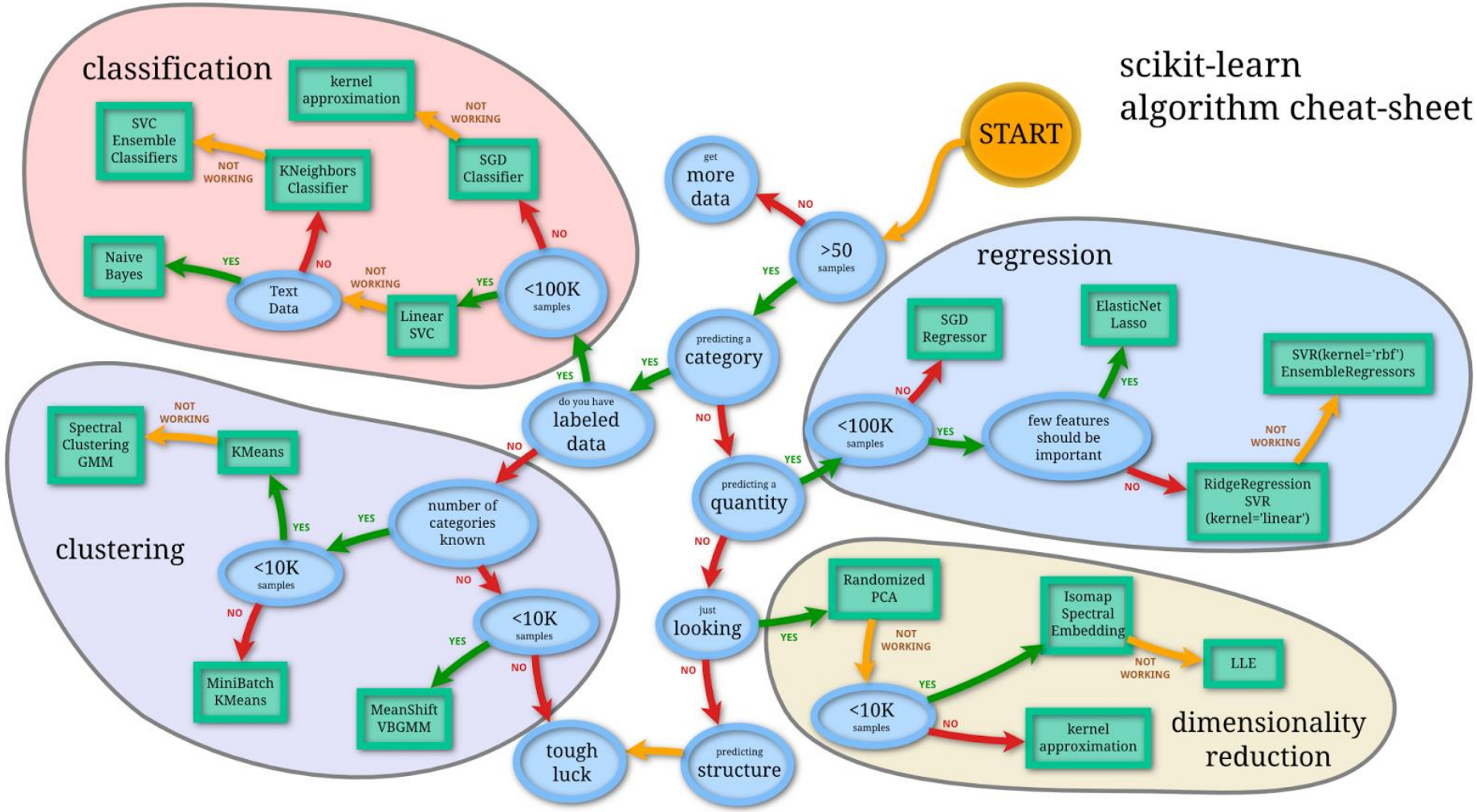
Выгоды открытых данных

- *для государства:* новые сервисы, кооперация бизнеса и науки
- *для индустрии:* бенчмаркинг, стандартизация, популяризация
- *для компаний:* подбор исполнителей, сокращение издержек и рисков
- *для университетов:* интеграция практических задач в учебный процесс
- *для исследователей:* проверка новых теорий и технологий в деле
- *для студентов:* получение опыта, наработка портфолио

Конкурсы анализа данных

- www.NetflixPrize.com (2006-2009) – первый крупный конкурс, $1 млн.
- www.kaggle.com – наиболее известная в мире платформа
- DataRing.ru – отечественная конкурсная платформа

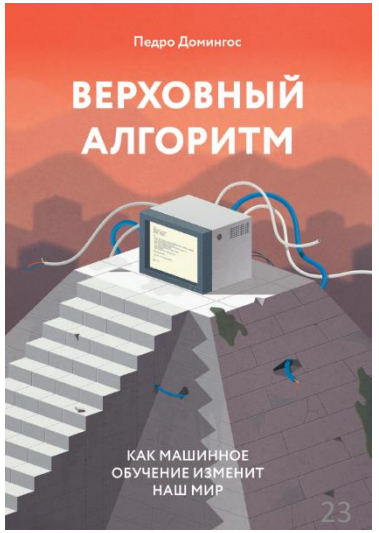
Задачи и методы машинного обучения



Основные школы машинного обучения

- *Символизм* – поиск логических закономерностей
- *Коннекционизм* – обучаемые нейронные сети
- *Эволюционизм* – адаптивная оптимизация сложных моделей
- *Байесионизм* – оценивание распределений над параметрами
- *Аналогизм* – «близким объектам близкие ответы»

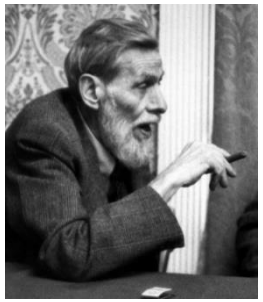
\+ *Композиционизм* – кооперация моделей

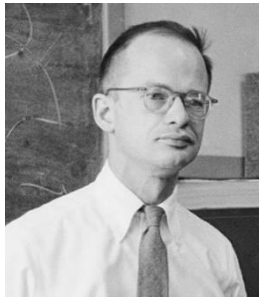


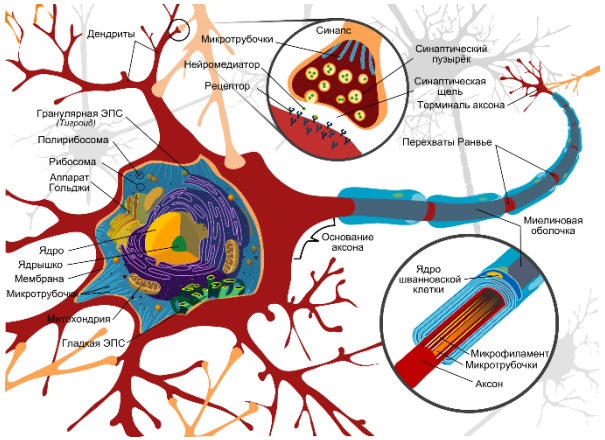
Педро Домингос. «Верховный алгоритм». 2016.

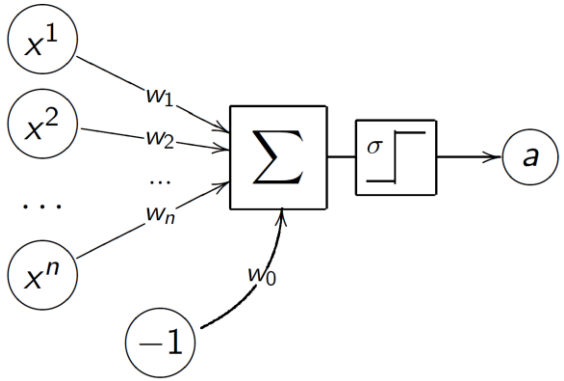
Что такое «искусственные нейронные сети»

Математическая модель нейрона (МакКаллок и Питтс, 1943)



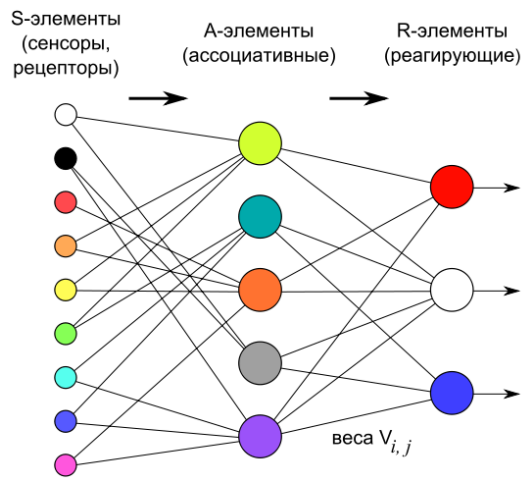




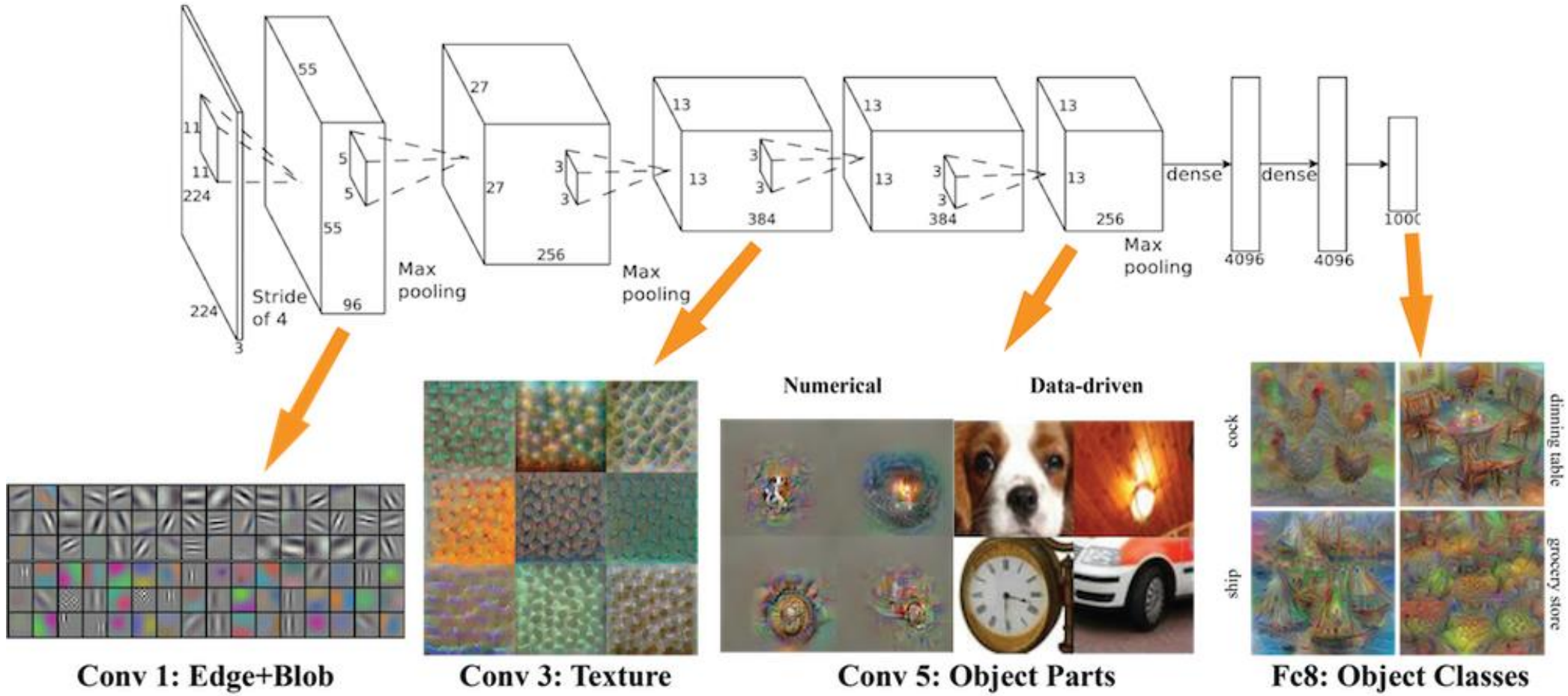


$$a(x, w) = \sigma\left(\sum_{j=1}^{n} w_j x^j - w_0\right)$$

Первый нейрокомпьютер Mark-1 (Фрэнк Розенблатт, 1960)



Что такое «глубокие нейронные сети»

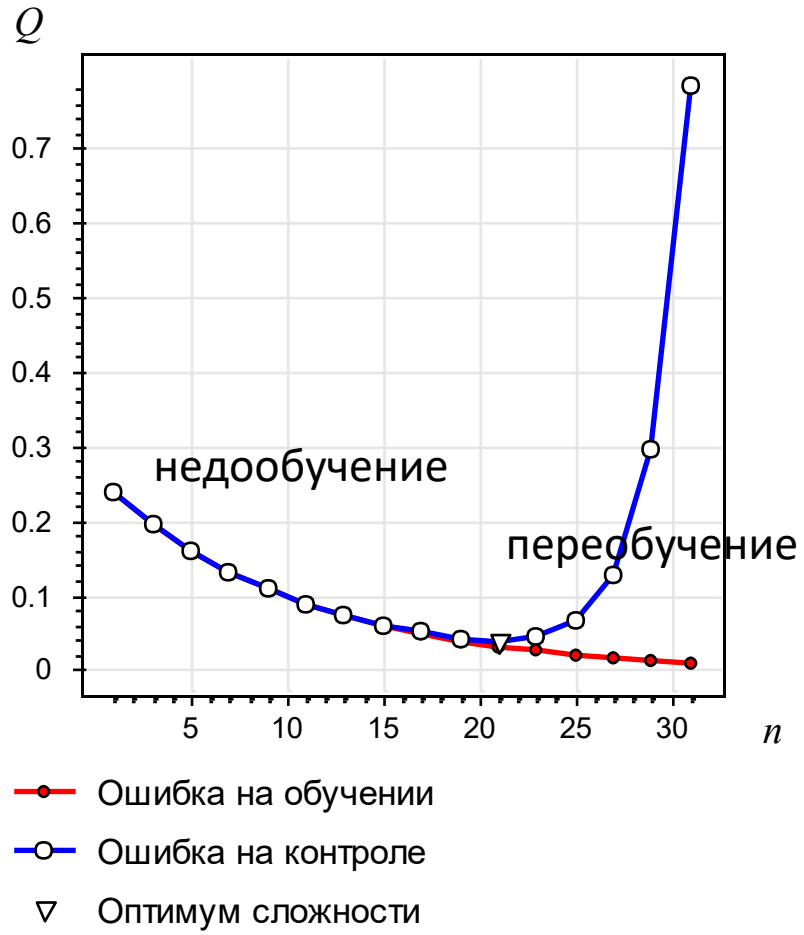


Что такое «переобучение»

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

Часто используемые внешние критерии:

- hold-out
- (q-fold) cross-validation, leave-one-out
- out-of-sample, out-of-time



Рекомендуемая литература

- *Домингос П.* Верховный алгоритм. 2016.
- *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.
- *Бенджио И., Гудфеллоу Я., Курвилль А.* Глубокое обучение. ДМК-Пресс, 2018.
- *Николенко С., Кадурин А., Архангельская Е.* Глубокое обучение. Питер, 2018.
- *Воронцов К. В.* Лекции по машинному обучению. www.MachineLearning.ru, 2004-2018.
- *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2014.
- *Bishop C. M.* Pattern Recognition and Machine Learning. - Springer, 2006.