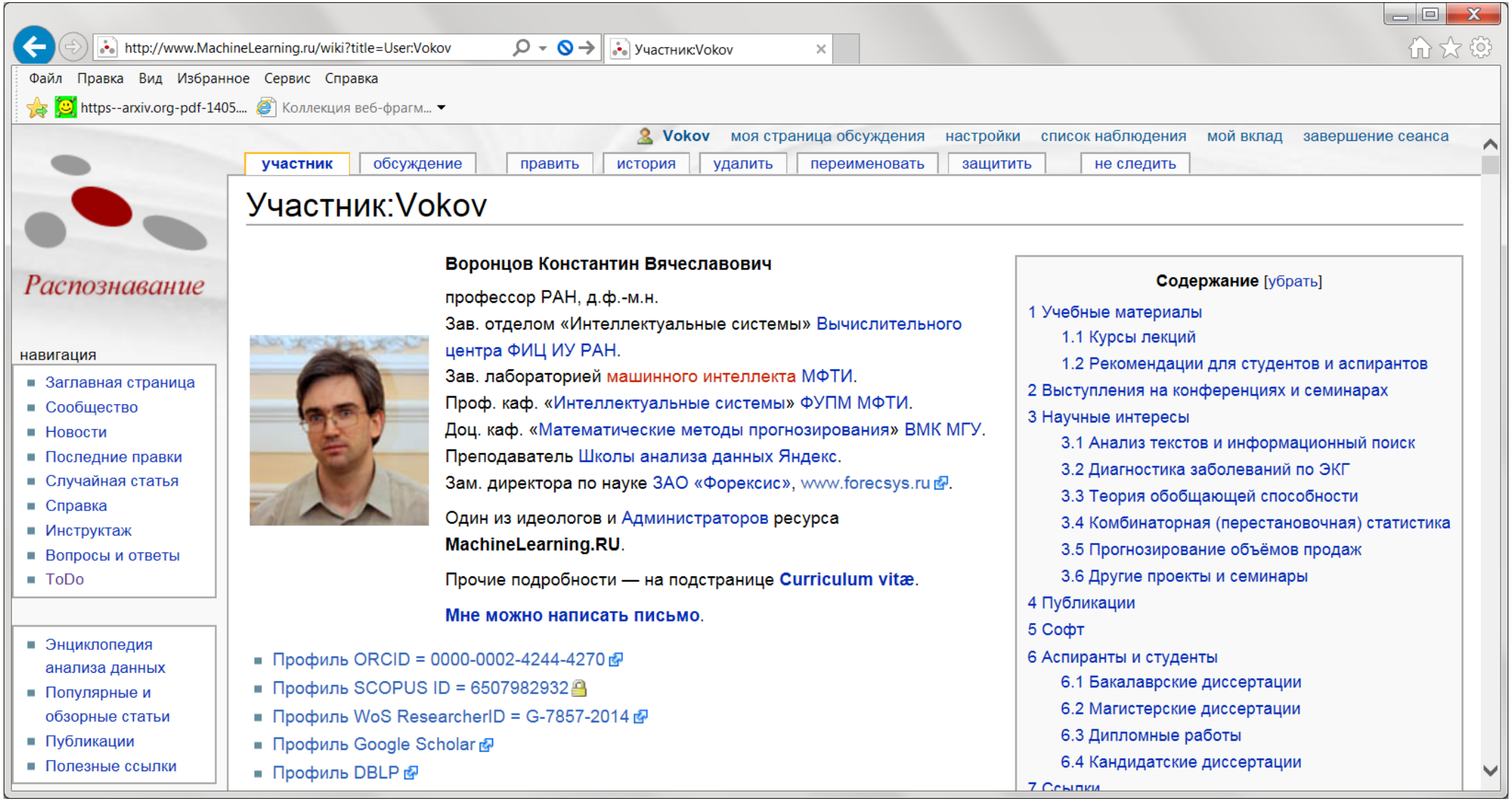
Искусственный интеллект и машинное обучение: практические шаги в цифровую экономику

Воронцов Константин Вячеславович

● Московский Физико-Технический Институт ●
● Вычислительный Центр им. А.А.Дородницына ФИЦ ИУ РАН ●
● ШАД Яндекс ●

voron@forecsys.ru

Докладчик: *Воронцов Константин Вячеславович*



Искусственный интеллект и машинное обучение: практические шаги в цифровую экономику

1. Знакомство с машинным обучением
   - Основные понятия
   - Примеры задач
2. Как оно менялось само
3. Как оно теперь меняет мир
4. Как решать задачи
5. Практические шаги

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

Клаус Мартин Шваб,
президент Всемирного
экономического форума

Машинное обучение – это …

- одна из ключевых информационных технологий будущего
- наиболее успешное направление искусственного интеллекта, вытеснившее экспертные системы и инженерию знаний
- математическое моделирование в сложно формализуемых областях, когда данных много, знаний мало
- проведение функции через заданные точки в сложно устроенных пространствах
- тысячи алгоритмов на стыке математической статистики и численных методов оптимизации
- около 100 000 научных публикаций в год

На полях:

машинное обучение – это, прежде всего, математические технологии

Основная задача машинного обучения

Этап №1 – обучение с учителем

- **На входе:**
  данные – выборка прецедентов «объект → ответ»
- **На выходе:**
  алгоритм, по любому объекту предсказывающий ответ

Этап №2 – применение

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

На полях:

Если нет данных,
то нет
и машинного обучения

Примеры задач машинного обучения

- **Медицинская диагностика:**
  объект – данные о пациенте на текущий момент
  ответ – диагноз / лечение / риск исхода
- **Распознавание 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 – вероятность дефолта, решение о выдаче кредита
- **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 товара 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:**
  объект – данные о паре «пользователь, товар»
  ответ – оценка вероятности, что пользователь купит товар

Примеры задач машинного обучения

- **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 – изображение предмета в видеопоследовательности
  ответ – решение (объехать, остановиться, игнорировать)

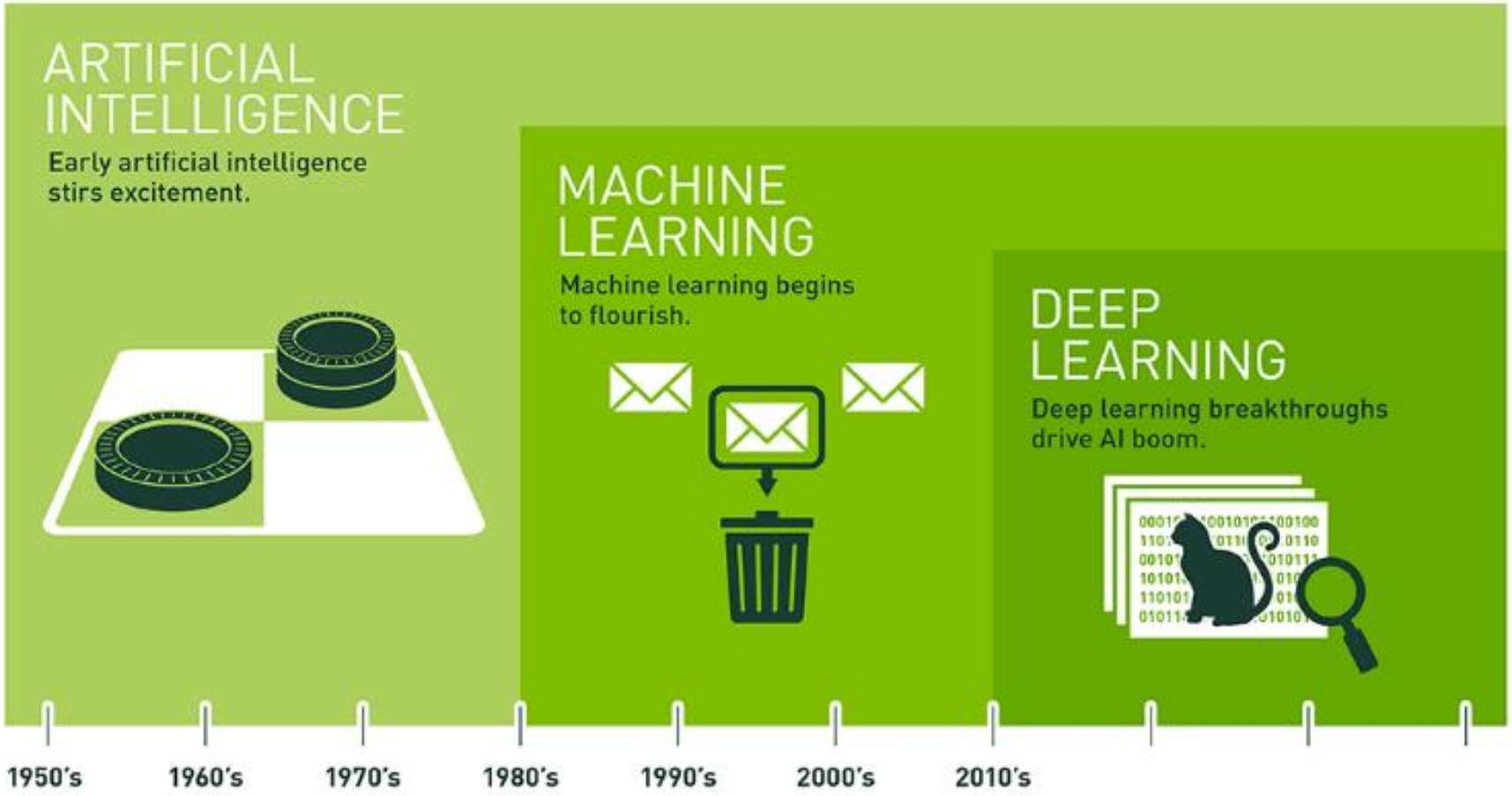
На полях:

Прогресс в этих областях связан с большими данными

Искусственный интеллект и машинное обучение: практические шаги в цифровую экономику

1. Знакомство с машинным обучением
2. Как оно менялось само
   - Эволюция идей и направлений ИИ
   - Бум искусственного интеллекта 201X
   - Три предпосылки этого бума
3. Как оно теперь меняет мир
4. Как решать задачи
5. Практические шаги

Эволюция искусственного интеллекта



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

На полях:

Глубокое обучение – одна из новейших технологий машинного обучения

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
- 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

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!»

2011–2015: ImageNet: 25% $\rightarrow$ 3,5% 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

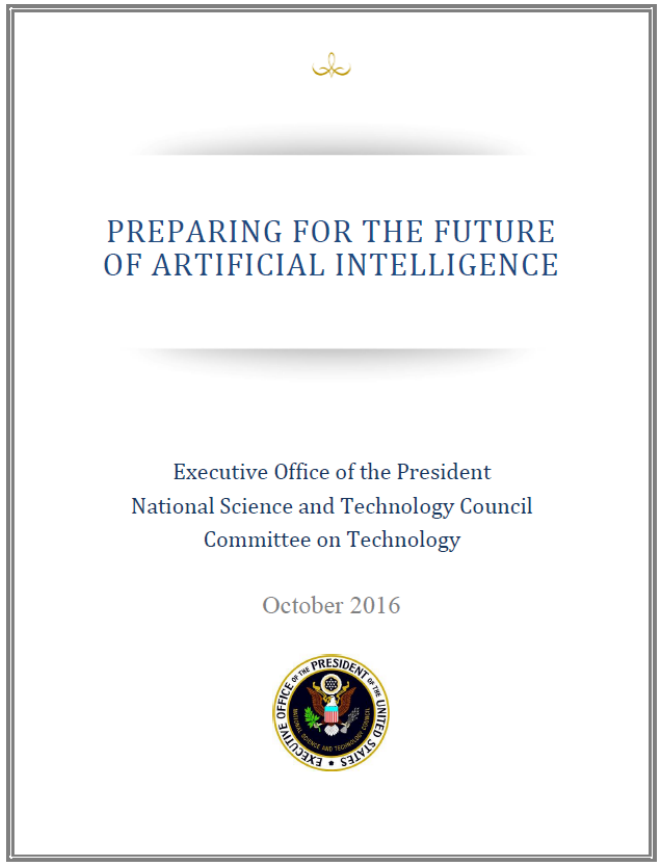
– три перехода количества в качество:

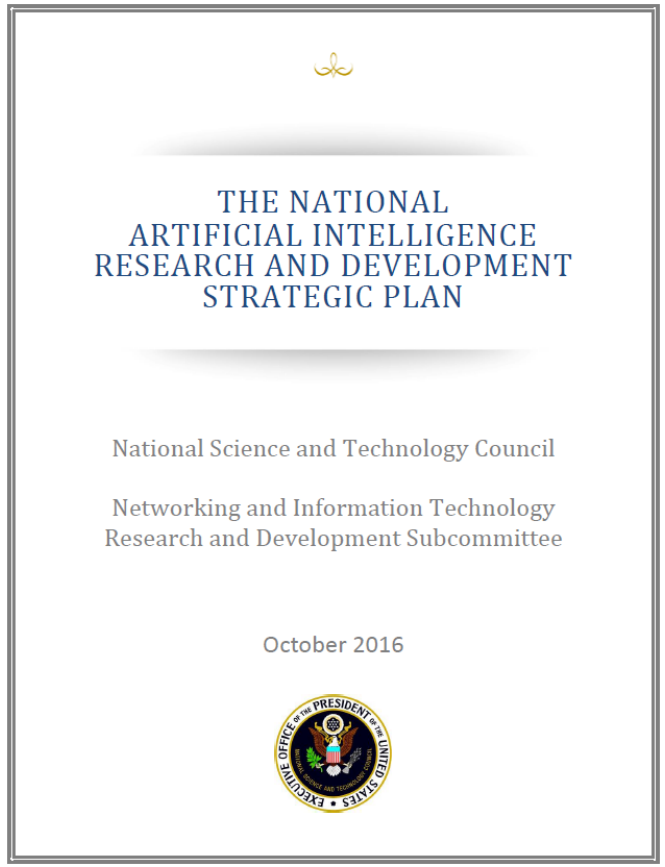
- Повсеместность и доступность компьютерных технологий
  $\rightarrow$ *Накопление больших выборок данных*
- Постепенное развитие математических методов и эвристик
  $\rightarrow$ *Накопление критической массы опыта*
- Достижения микроэлектроники
  $\rightarrow$ *Рост вычислительных мощностей по закону Мура*

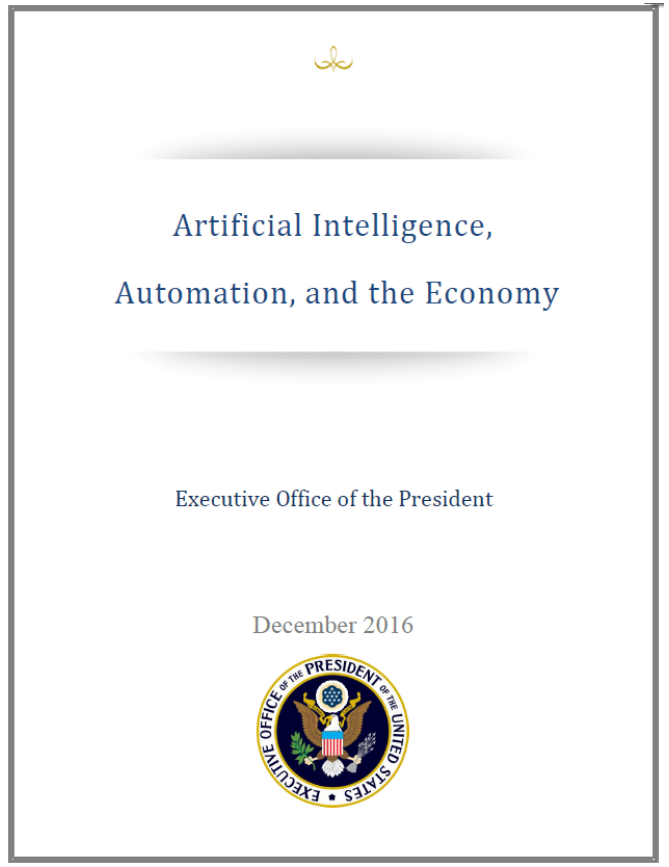
Искусственный интеллект и машинное обучение: практические шаги в цифровую экономику

1. Знакомство с машинным обучением
2. Как оно менялось само
3. Как оно теперь меняет мир
   - Отчёты Белого Дома США конца 2016 г.
   - Открытые данные
   - Открытый код
4. Как решать задачи
5. Практические шаги

Отчёты Белого дома США, октябрь 2016







Основные выгоды ИИ

- Автоматизация и сокращение издержек повсеместно
- Автономный транспорт и роботизация
- Оптимизация логистики и цепей поставок
- Оптимизация энергетических и транспортных сетей
- Сенсорные сети, мониторинг сельского хозяйства
- Информационные сервисы и распределённая экономика
- Персональная медицина, улучшение клинических практик
- Персональные образовательные траектории, социальная инженерия
- Автономные системы вооружений

Государственное финансирование ИИ

Правительство США выделяет на исследования по ИИ

- $1.1 млрд. в 2015
- $1.2 млрд. в 2016

(без учёта вложений в ИИ со стороны частных компаний США).

Согласно отчёту, для США необходимо в 2-4 раза больше.

$1 млрд. выделяет Южная Корея,
$1 млрд. выделяет компания Toyota,
$1 млрд. инвестиций привлекла компания Илона Маска OpenAI.

7 стратегий R&D в области ИИ

1. Долгосрочные инвестиции в исследования в области ИИ
2. Разработка эффективных человеко-машинных систем ИИ
3. Исследование этических, юридических и социальных аспектов ИИ
4. Обеспечение безопасности, надёжности и доверия к системам ИИ
5. Развитие открытых данных и средств разработки ИИ
6. Развитие стандартов и платформ для тестирования ИИ
7. Подготовка квалифицированных кадров в области ИИ

«Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future»

Некоторые из 23 рекомендаций

#1. Государственным и коммерческим организациям: активнее развивать партнёрство с научными коллективами для эффективного использования данных.

#2. Развивать стандарты открытых данных для привлечения научного сообщества к решению задач.

#11. Вести постоянный мониторинг исследований ИИ в мире.

#13. Поддерживать фундаментальные исследования в области ИИ.

#20, #21. Развивать международную кооперацию по ИИ.

#22. Учитывать взаимовлияние ИИ и кибербезопасности.

Открытые данные для ИИ

- Выгоды открытых данных
  - привлечение научного сообщества к решению задач
  - получение оценок предельно достижимого качества решения
  - поиск талантов, выявление центров компетенции
  - стимулирование развития прикладной науки и образования
- Культура открытых данных
  - преодоление ментальных барьеров
  - подготовка данных: очистка, отбор, агрегирование, деперсонификация
  - подготовка условий конкурсов и тендеров
- Конкурсы анализа данных
  - www.NetflixPrize.com (2006-2009) – первый крупный конкурс, $1 млн.
  - www.kaggle.com – самая известная конкурсная платформа
  - DataRing.ru – отечественная конкурсная платформа

Искусственный интеллект и машинное обучение: практические шаги в цифровую экономику

1. Знакомство с машинным обучением
2. Как оно менялось само
3. Как оно теперь меняет мир
4. Как решать задачи
   - Особенности реальных данных
   - Особенности постановок задач
   - Этапы решения в стандарте CRISP-DM
   - Типология задач машинного обучения
5. Практические шаги

Особенности данных и постановок задач

Свойства реальных данных:

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

На полях:

Самая большая беда – «грязные» данные

Особенности данных и постановок задач

Идеальный заказчик:

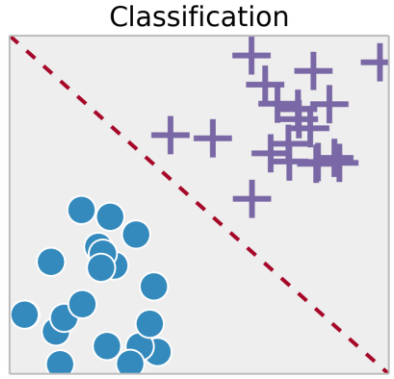
- знает точно, чего хочет
- имеет численные критерии качества (KPI)
- заботится о качестве своих данных
- готов пилотировать новые технологии
- понимает ограничения готовых методов, и что чудес не бывает
- видит уникальность задачи, когда нужна новая математика

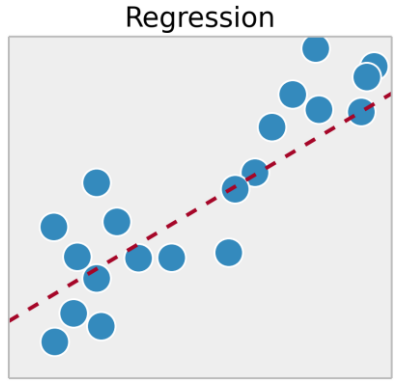
На полях:

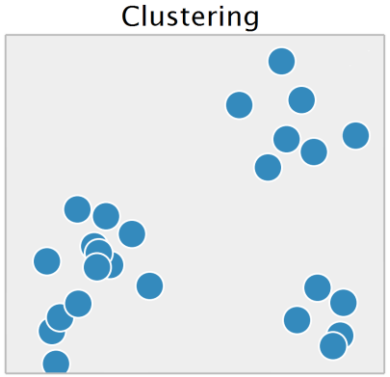
обучение машин заставляет думать людей

Типология задач машинного обучения

- Обучение с учителем (supervised learning)
  - классификация (classification)
  - регрессия (regression)
  - ранжирование (learning to rank)
  - прогнозирование (forecasting)
- Обучение без учителя (unsupervised learning)
  - кластеризация (clustering)
  - поиск ассоциативных правил (association rule learning)
  - восстановление плотности (density estimation)
  - одноклассовая классификация (anomaly detection)
- Частичное обучение (semi-supervised learning)
  - Обучение с положительными примерами (PU-learning)







Типология задач машинного обучения

- Предварительная обработка (data preparation)
  - извлечение признаков (feature extraction)
  - отбор признаков (feature selection)
  - восстановление пропусков (missing values)
- Обучение представлений (representation learning)
  - обучение признаков (feature learning)
  - обучение многообразий (manifold learning)
  - анализ главных компонент (principal component analysis)
  - матричные разложения (matrix factorization)
  - тематическое моделирование (topic modeling)
  - коллаборативная фильтрация (collaborative filtering)

Типология задач машинного обучения

- Обучение глубоких сетей (deep learning)
- Обучение выявлению связей (relational learning)
- Динамическое обучение (online/incremental learning)
- Обучение с подкреплением (reinforcement learning)
- Активное обучение (active learning)
- Обучение с противником (adversarial learning)
- Привилегированное обучение (learning with privileged information)
- Обучение с переносом опыта (transfer learning)
- Мета-обучение (meta-learning)

Что должен уметь менеджер в области Data Science

- Видеть возможности применения машинного обучения
- Ставить задачи в виде ДНК (Дано-Найти-Критерий)
- Разбираться в методах на уровне «возможности–ограничения»
- Организовывать бизнес-процессы для сбора чистых данных
- Организовывать открытые конкурсы анализа данных
- Запускать пилотные проекты
- Знать экспертное сообщество
- Адекватно оценивать сложность задачи и трудозатраты

Искусственный интеллект и машинное обучение: практические шаги в цифровую экономику

1. Знакомство с машинным обучением
2. Как оно менялось само
3. Как оно теперь меняет мир
4. Как решать задачи
5. Практические шаги

Практические шаги в цифровую экономику

- Образование. Знания и таланты – это нефть XXI века
  - элитарное образование: поддержка лидирующих школ и университетов
  - инженерное образование: курсы переподготовки в области анализа данных
  - развитие национальной платформы онлайн-образования
  - задача СМИ: пропаганда науки, технологий, стратегий личного успеха
- Наука и инновации
  - создание реестра отечественных разработок
  - создание сетевой инфраструктуры, объединяющей центры компетенции, проектные научно-исследовательские группы, заказчиков и инвесторов
  - выработка стратегий в открытом диалоге между властью и учёными
- Формирование рынков данных
  - стимулирование партнёрства компаний и научных коллективов
  - продвижение открытых конкурсов анализа данных во всех отраслях
  - развитие облачных сервисов и центров коллективного пользования

Полезные ссылки

- ods.ai – *Open Data Science*: русскоязычное сообщество индустрии данных
- www.MachineLearning.ru – русскоязычная вики
- www.kdNuggets.com – главный сайт датамайнеров
- www.DataScienceCentral.com – 72 000 датамайнеров
- www.kaggle.com – конкурсы анализа данных
- archive.ics.uci.edu/ml – UCI ML Repository (349 datasets)
- ru.coursera.org/learn/machine-learning – курс Эндрю Ына
- ru.coursera.org/learn/vvedenie-mashinnoe-obuchenie – курс машинного обучения от ВШЭ и ШАД Яндекс
- ru.coursera.org/specializations/machine-learning-data-analysis – специализация от МФТИ и ШАД Яндекс

Рекомендуемая литература

- *Домингос П.* Верховный алгоритм. 2016. 336 с.
- *Коэльо Л. П., Ричарт В.* Построение систем машинного обучения на языке Python. 2016. 302 с.
- *Мерков А. Б.* Распознавание образов. Введение в методы статистического обучения. 2011. 256 с.
- *Мерков А. Б.* Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
- *Воронцов К. В.* Лекции по машинному обучению. www.MachineLearning.ru. 2004-2017.
- *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2014. 739 p.
- *Bishop C. M.* Pattern Recognition and Machine Learning. - Springer, 2006. 738 p.