Наука и бизнес в одном FLACONe: возгонка цифровой экономики

Воронцов Константин Вячеславович

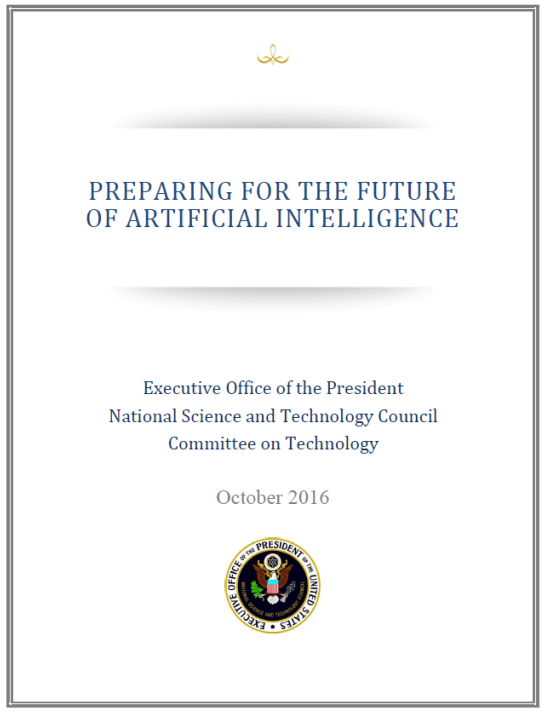
● Московский Физико-Технический Институт ●
● Сбербанк ● ФИЦ ИУ РАН ● ШАД Яндекс ●

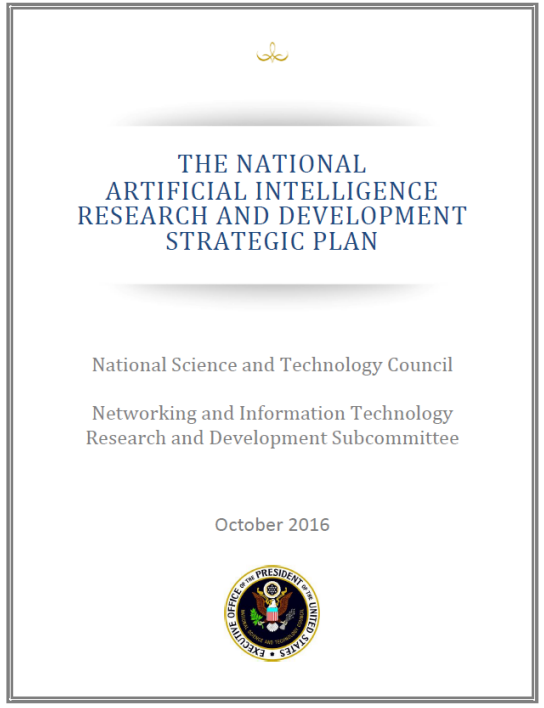
k.v.vorontsov@phystech.edu

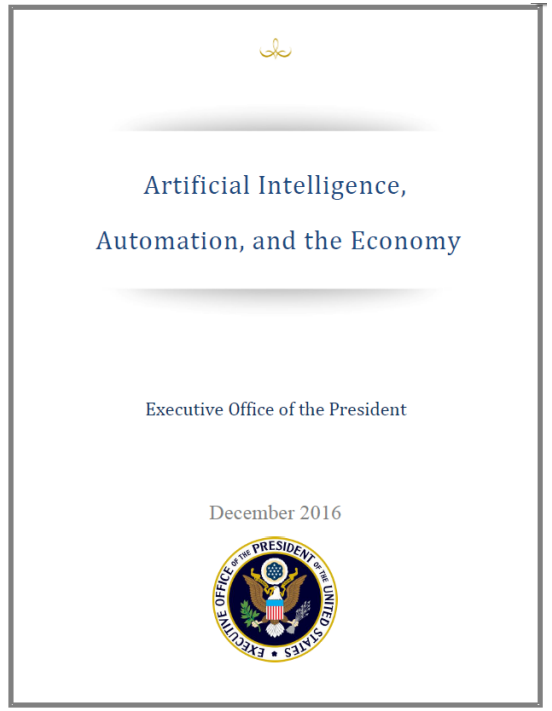
«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, *искусственном интеллекте* и *машинном обучении*» (2016)

Клаус Мартин Шваб,
президент Всемирного
экономического форума

Отчёты Белого дома США, май-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
- Автоматизация банковских услуг и посреднической деятельности
- Информационные сервисы и распределённая экономика
- Персональная медицина, улучшение клинических практик
- Персональные образовательные траектории, социальная инженерия
- Автономные системы вооружений

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. 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.

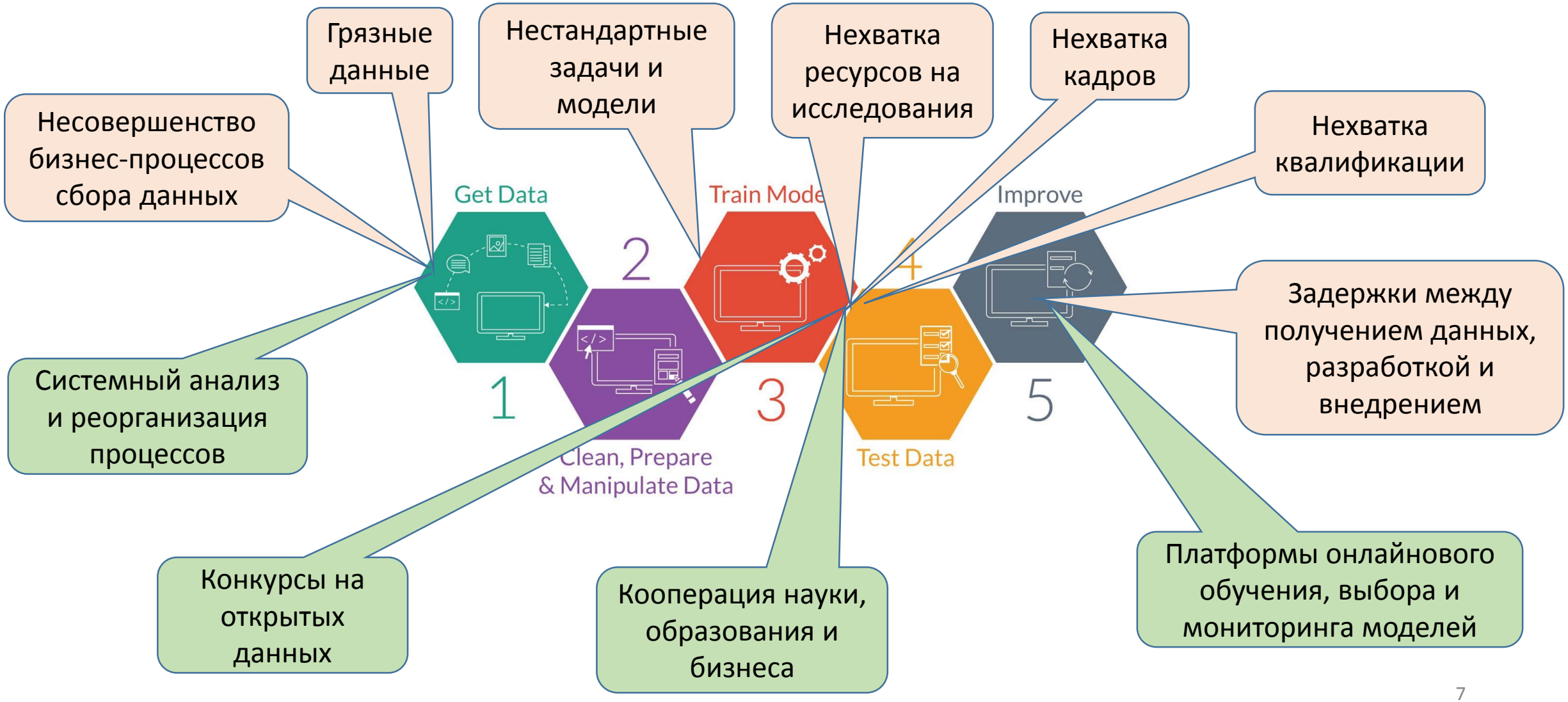
#8. Инвестировать в разработку систем автоматического управления воздушным трафиком.

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

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

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

Факторы риска и точки приложения силы



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! ☹

Особенности реальных постановок задач

В реальных приложениях заказчик, как правило, …

- не знает точно, чего хочет («чтобы было хорошо»)
- не имеет численных критериев качества (KPI)
- не заботится о чистоте данных
- не готов пилотировать новые технологии
- не понимает ограничений готовых методов
- не отличает простые задачи от сложных

Выход есть – становиться более цивилизованными!

- наращивать экспертизу: учиться + привлекать учёных
- менять бизнес-процессы получения данных
- вводить контроль качества данных
- открывать данные, проводить конкурсы

Для внедрения искусственного интеллекта придётся напрягать естественный

Открытые данные для ИИ

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 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

- консалтинг по подготовке данных и условий конкурса
- очистка, отбор, агрегирование, деперсонификация данных

Платформы онлайнового обучения

Обычная схема решения задач DS|ML|AI:

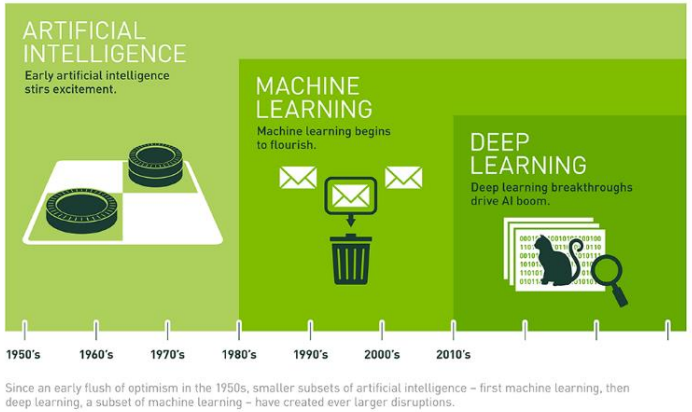
- Забираем данные из промышленной системы (долго!)
- Стоим модели, экспериментируем в удобной для нас среде
- Переносим модели обратно в пром (долго!)

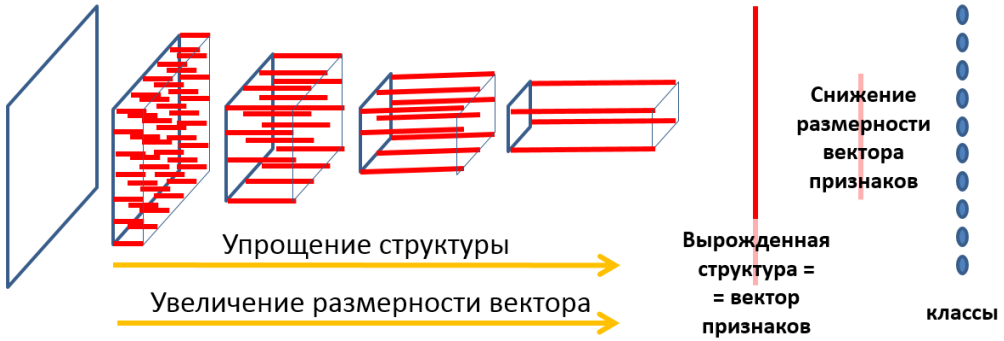
Будущее – за онлайновым машинным обучением:

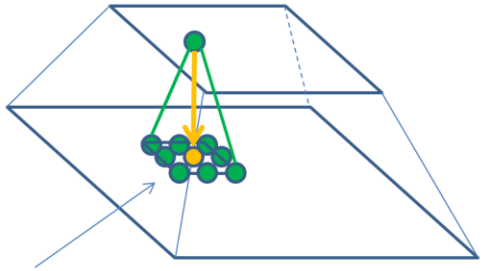
- Предобработка данных и дообучение моделей – налету
- Модель включается только после периода валидации
- Модели валидируются и отбираются по совокупности критериев
- Работа аналитика – постоянный мониторинг качества моделей

Вытеснит ли DL весь остальной ML?

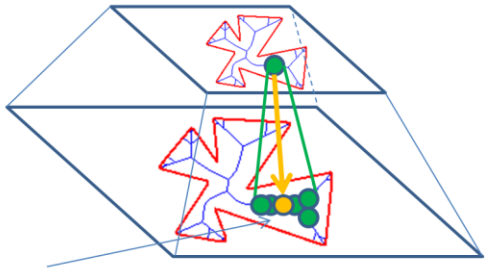
Глубокие сети – это инструмент автоматизации извлечения признаков (Feature Extraction).
Ближайшее будущее: свёрточные сети обобщаются на любые данные с локальными структурами.







Прямоугольное окно заданного размера с центром в заданной точке +
+ операция свёртки по окну



Локальная окрестность, определяемая для любой вершины графа +
+ операция свёртки по окрестности

Визильтер Ю.В., Горбацевич В.С. Структурно-функциональный анализ и синтез глубоких конволюционных нейронных сетей. ММРО-2017.

Менеджер в области Data Science должен…

- 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

Сухой остаток

- Чистота данных и бизнес-процессы
- Конкурсы на открытых данных
- Кооперация бизнеса и науки
- Образовательные проекты
- Правильное разделение труда
- Онлайновое обучение в проме

===

Воронцов Константин Вячеславович

k.v.vorontsov@phystech.edu