Конференция Data Halloween

31 октября 2018 ● Москва, Центр цифрового лидерства SAP

Data Science: как наладить взаимодействие науки, бизнеса и образования

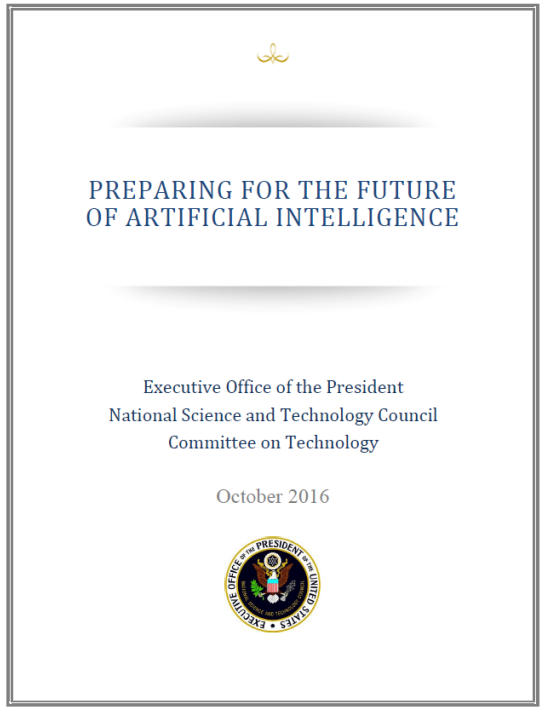
Воронцов Константин Вячеславович
(лаборатория машинного интеллекта МФТИ)

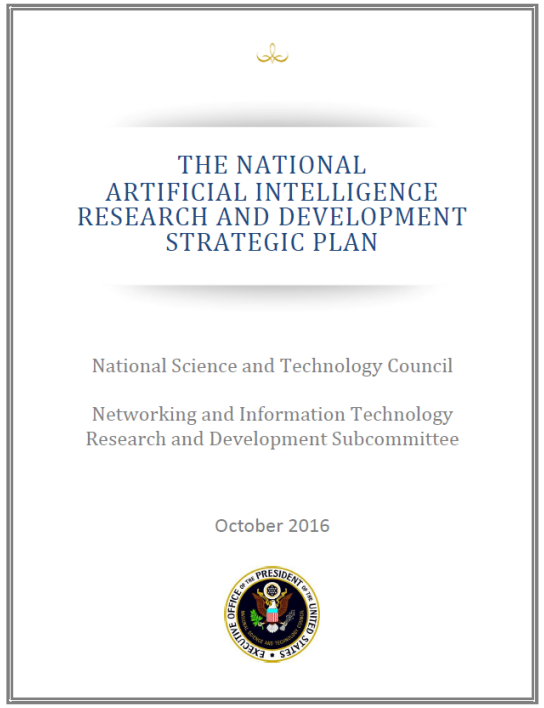
k.v.vorontsov@phystech.edu

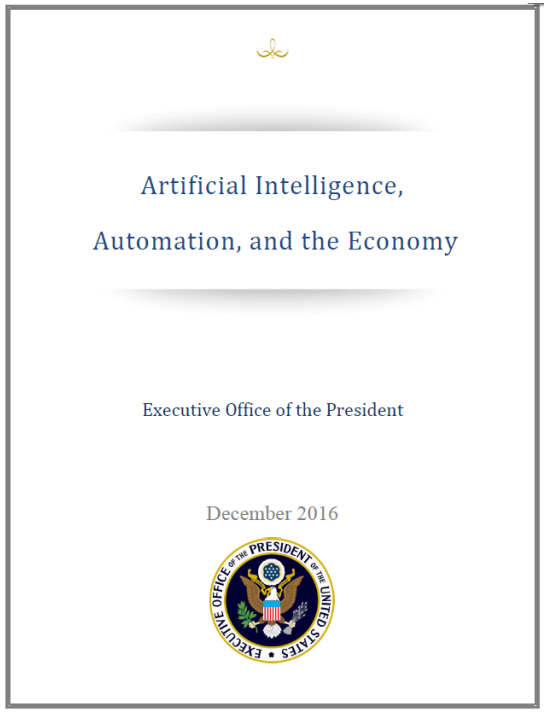
«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, *искусственном интеллекте* и *машинном обучении*» (2016)

Клаус Мартин Шваб,
президент Всемирного
экономического форума

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

По материалам отчёта «PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE»

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.

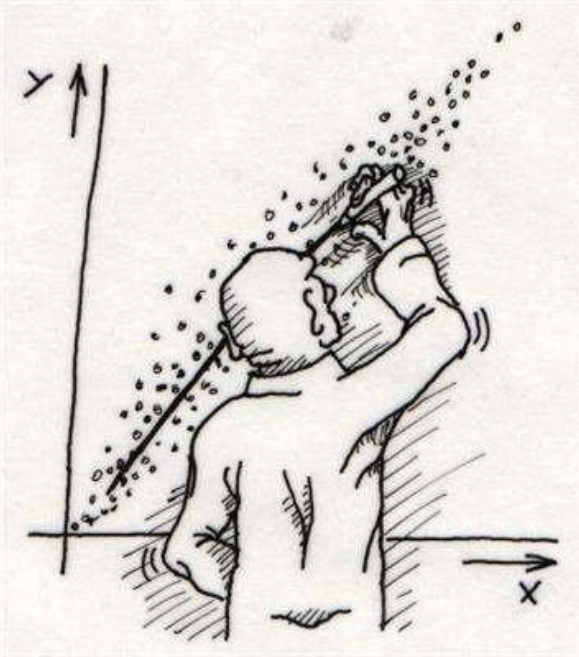
#14. Развивать образовательные программы по ИИ и курсы повышения квалификации для прикладных специалистов.

#20. Развивать международную кооперацию по ИИ.

#22. Учитывать взаимовлияние ИИ и кибербезопасности.

Машинное обучение (Machine Learning)

- одна из ключевых информационных технологий будущего
- наиболее успешное направление *искусственного интеллекта*, 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

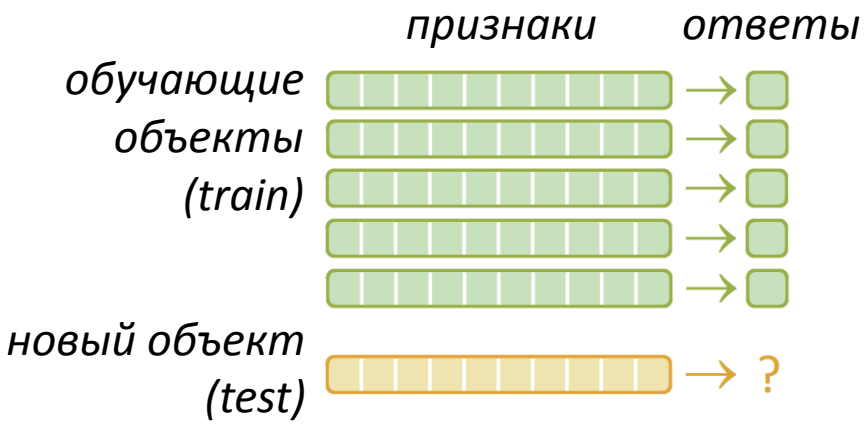
Этап №1 – обучение с учителем

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

Если нет данных,
то нет
и машинного
обучения

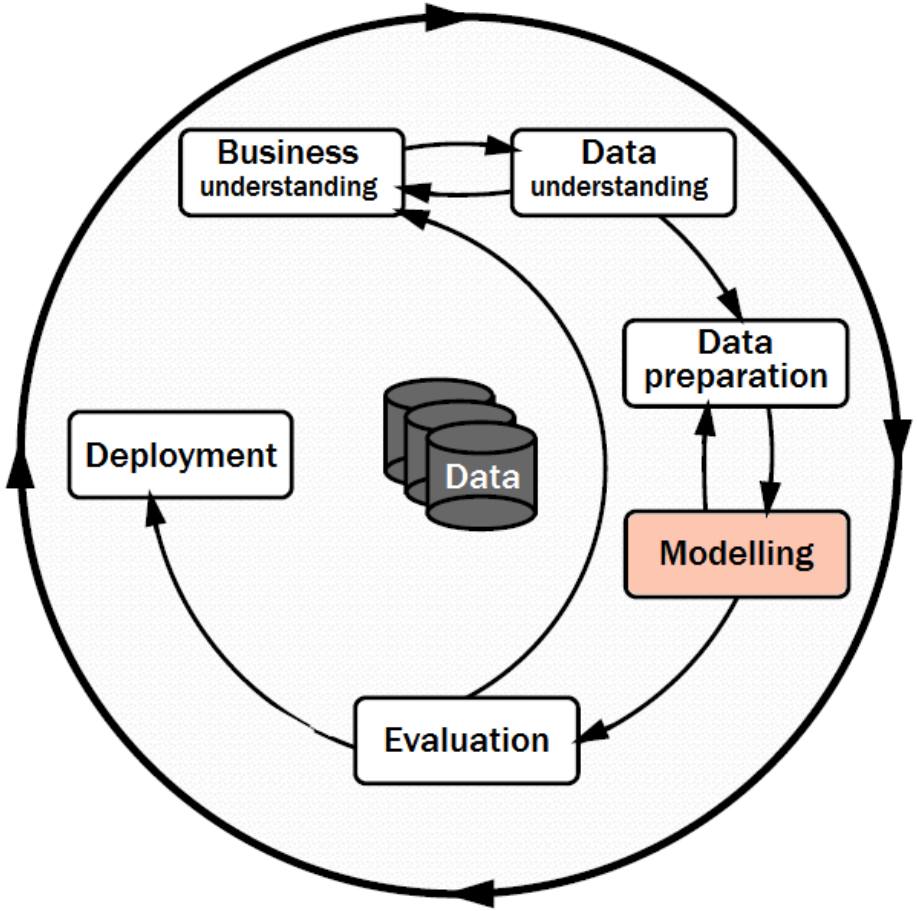
Этап №2 – применение

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



Стандартный процесс анализа данных

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



- понимание бизнес-задач
- понимание данных
- предобработка данных
- инженерия признаков
- построение моделей
- оптимизация параметров
- контроль переобучения
- оценивание качества решения
- внедрение и эксплуатация

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 проектов

Проблема №1: некомпетентный исполнитель

- не готов к преодолению сложностей реальных задач

Проблема №2: некомпетентный заказчик

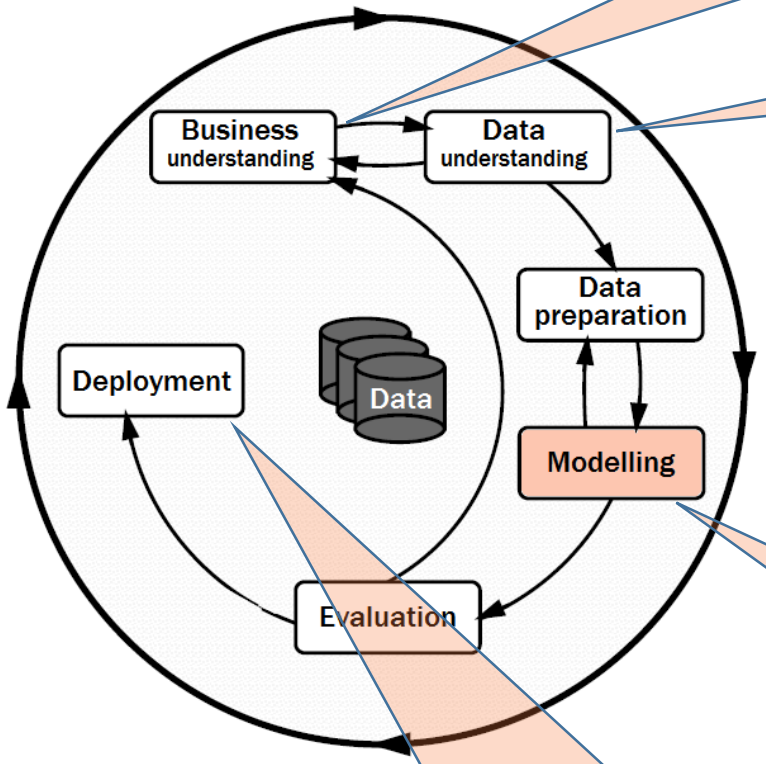
- ждёт чуда от искусственного интеллекта и больших данных
- не имеет численных критериев качества (KPI)
- не заботится о чистоте данных
- не готов пилотировать новые технологии
- не отличает простые задачи от сложных

Выход: инвестиции + образование

- цифровая трансформация бизнес-процессов
- введение контроля качества данных
- кооперация бизнеса, науки и образования
- проведение конкурсов на открытых данных

Для внедрения искусственного интеллекта придётся напрягать естественный

Факторы риска и точки приложения силы



Несовершенство бизнес-процессов сбора данных

Системный анализ и реорганизация процессов

Грязные данные

Конкурсы на открытых данных

Нехватка ресурсов на исследования

Нехватка квалификации

Кооперация науки, образования и бизнеса

Нехватка кадров

Создание новых методов для новых задач

Нестандартные задачи и модели

Платформы адаптивного обучения моделей

Задержки между получением данных, моделированием и внедрением

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

Кооперация бизнеса, науки и образования

Проблемы: • различия в целях • «некомпетентность» • дефицит доверия

Опыт кафедры «Интеллектуальные системы» МФТИ

- Практикум В.В.Стрижова (страница на www.MachineLearning.ru) более 700 индивидуальных студенческих проектов за 12 лет
- Начало сотрудничества – пилотный проект в рамках практикума (ограничение: старт пилота два раза в год, февраль и сентябрь)

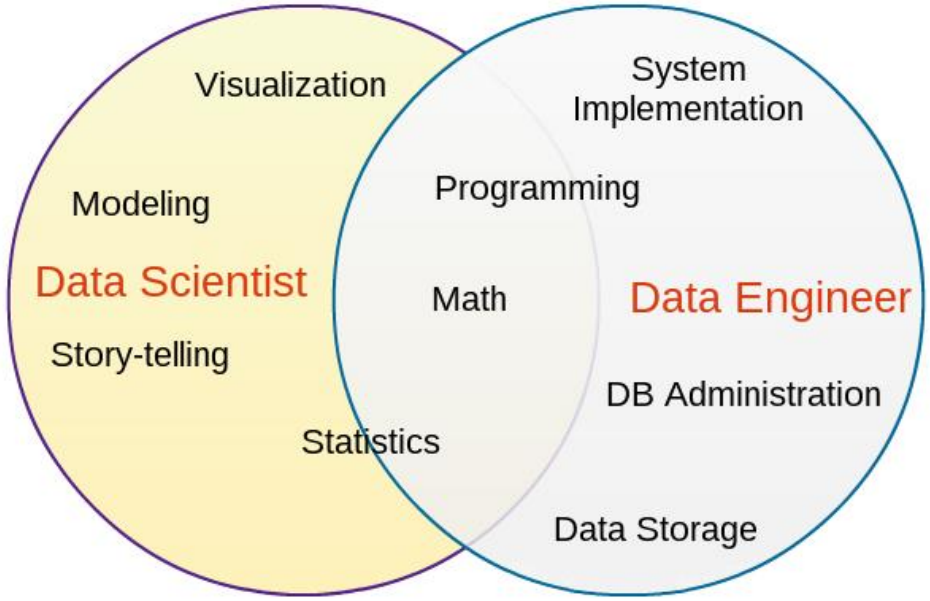
Шаги долгосрочного сотрудничества

- НИР/ОКР для университетской лаборатории
- формирование постоянной проектной группы, стажерская программа
- формирование образовательных курсов/модулей по решенным задачам
- открытие собственной лаборатории или кафедры
- тесная кооперация с собственным исследовательским отделом

Рынок труда в области анализа данных

Инженер по данным **(Data Engineer)**

- Понимает бизнес-процессы, порождающие данные
- Работает с сырыми 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
- Адекватно оценивает сложность задач и трудозатраты

Платформы адаптивного обучения

Обычная схема решения задач DS|ML|AI:

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

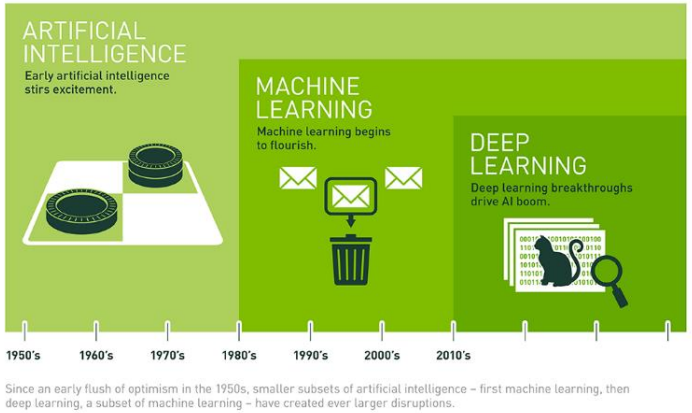
Будущее – за онлайновым машинным обучением:

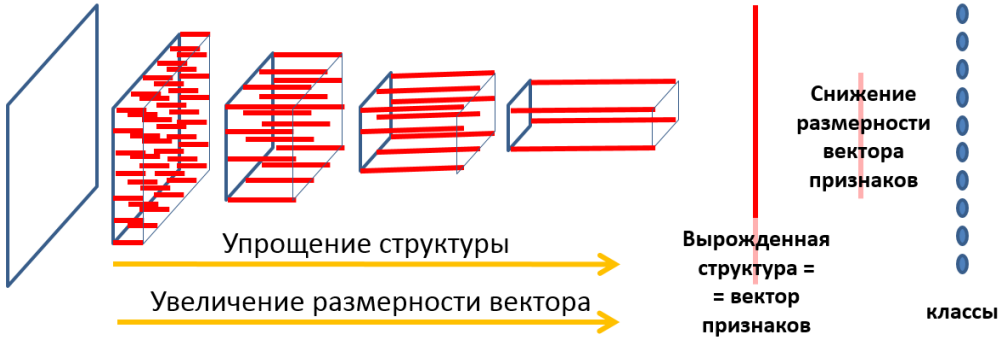
- Предобработка данных и дообучение моделей – налету
- Валидация моделей по совокупности критериев
- Адаптивная селекция и композиция моделей
- Работа аналитика – мониторинг, визуализация и доработка моделей

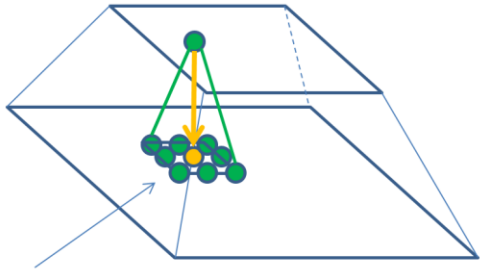
Новые методы:
вытеснит ли DL всё остальное машинное обучение?

Глубокие сети – это инструмент автоматизации извлечения признаков (Feature Extraction).

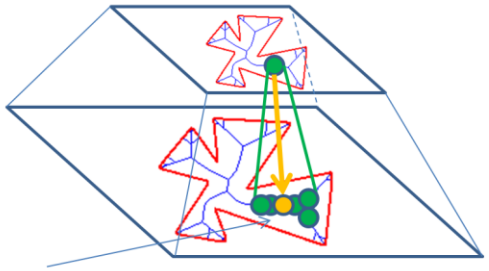
Ближайшее будущее: свёрточные сети обобщаются на любые данные с локальными структурами.







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



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

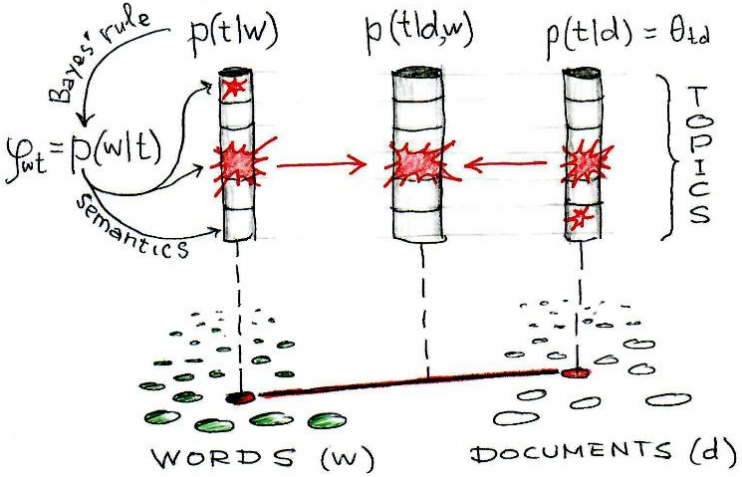
Визильтер Ю.В., Горбацевич В.С. Структурно-функциональный анализ и синтез глубоких конволюционных нейронных сетей. ММРО-2017.

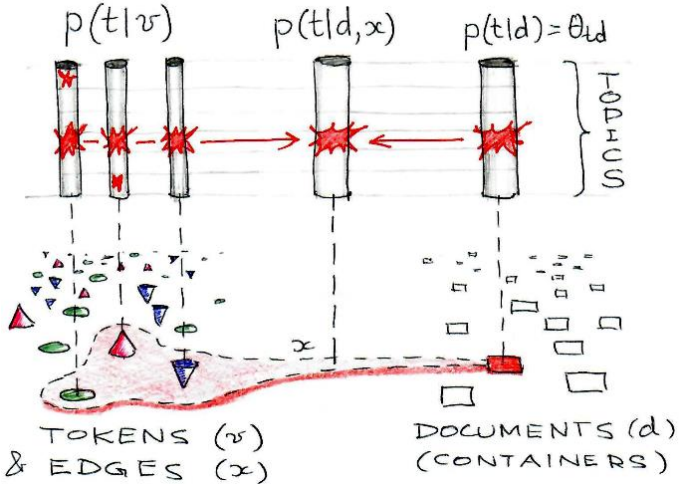
Новые методы:
векторизация сложных данных

Сложные данные: тексты, изображения, графы, гиперграфы, транзакции

Векторные представления объектов по наблюдаемым взаимодействиям:

- *неинтерпретируемые:* word2vec, doc2vec, node2vec, graph2vec, prod2vec, StarSpace,…
- *интерпретируемые:* тематические модели (Topic Modeling)



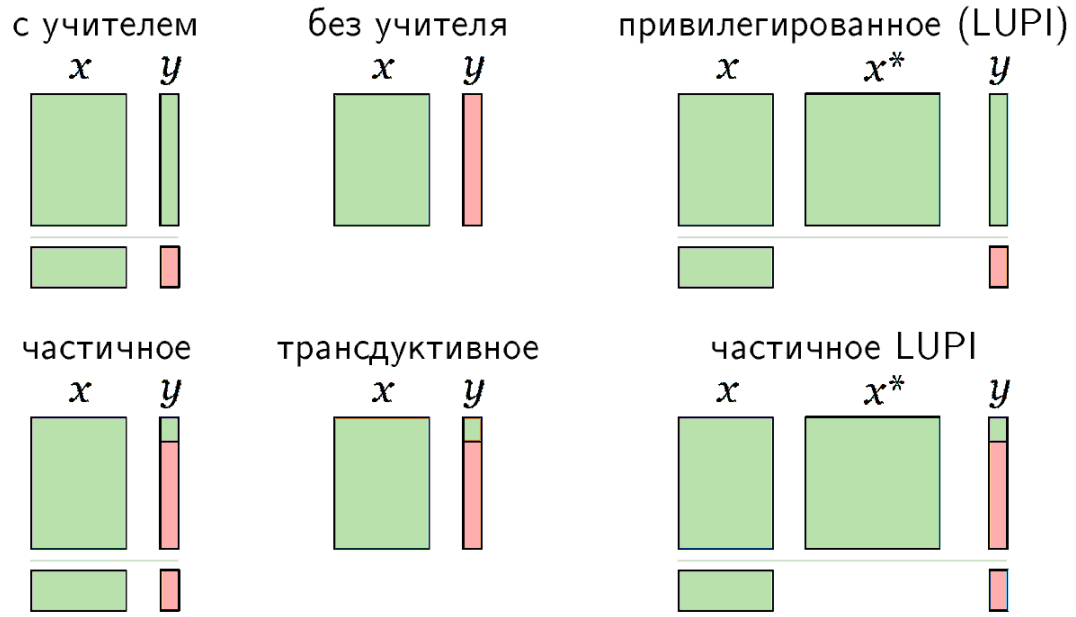


Воронцов К.В. Вероятностное тематическое моделирование: обзор моделей и аддитивная регуляризация. www.MachineLearning.ru. 2018.

Новые методы:
обучение с привилегированной информацией

Естественная модель обучения с учителем:
LUPI – Learning Using Priveleged Information
учитель даёт не только правильные ответы, но и объяснения

- На стадии обучения учитель сообщает важную информацию x^* об объектах обучения
- Но на стадии тестирования этой информации не будет



V.Vapnik, A.Vashist. A new learning paradigm: Learning Using Privileged Information. 2009.

Сухой остаток

1. Цифровизация бизнес-процессов и чистота данных
2. Популяризация задач через открытые данные и конкурсы
3. Кооперация бизнеса и науки: step-by-step
4. Создание новых методов под новые задачи

===

Воронцов Константин Вячеславович

k.v.vorontsov@phystech.edu

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.