МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»
МАИ

Реферат

По дисциплине:

«Информационные технологии»

На тему:

Data Science – Наука о данных

Выполнил:
Студент группы М4В-301б
Коротков В.В.
Руководитель:
Терехин А. Г.

Г. Москва
2020

Содержание

1. история

Наука о данных зародилась во второй половине 20-го века, намного раньше чем термин «большие данные», который стал популярен с 2010-х гг. Первое упоминание этого понятия датируется 1974 годом, когда вышла книга Петера Наура. В этой публикации Data Science определяется как дисциплина по изучению жизненного цикла цифровых данных, от момента их появления до преобразования и использования в других областях знаний. Тем не менее, широкое употребление этот термин получил лишь в 1990-е годы, а общепризнанным стал только в начале 2000-х. В частности, в 2002 году междисциплинарный Комитет по данным для науки и техники начал выпускать журнала CODATA Data Science Journal, а в январе 2003 года вышел первый номер The Journal of Data Science Колумбийского университета. Следующая волна интереса к DS возникла при популяризации понятия Big Data, с 2010 года, когда вычислительные мощности даже бытовых компьютеров стали позволять работать с большими объемами данных. Примерно с этого же времени стали проводиться многочисленные профессиональные конференции, а университеты по всему миру включили эту дисциплину в свои учебные курсы, разработав соответствующие образовательные программы. Сегодня Data Science активно применяется в широком спектре прикладных областей деятельности: от астрономии до медицины, включая коммерческие кейсы: маркетинг, ритейл, менеджмент, финансовый анализ, предиктивная аналитика чрезвычайных ситуаций и т.д.

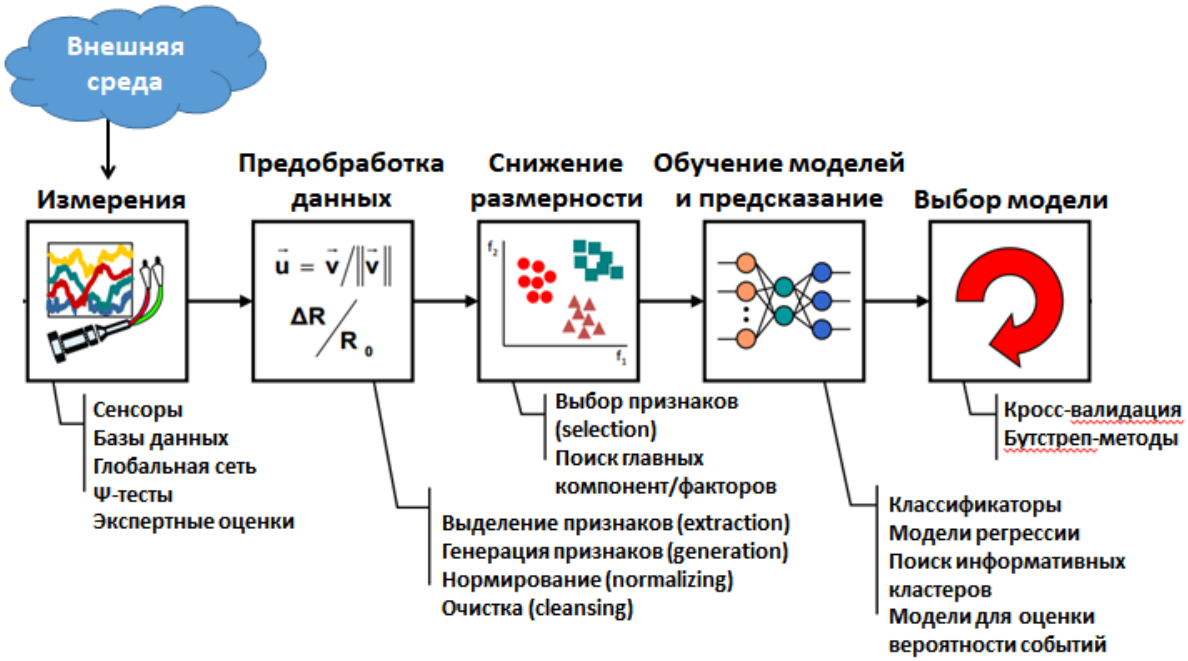
2. Что такое Data Science?

Data Science — это набор конкретных дисциплин из разных направлений, отвечающих за анализ данных и поиск оптимальных решений на их основе. Раньше этим занималась только математическая статистика, затем начали использовать машинное обучение и искусственный интеллект, которые в качестве методов анализа данных к матстатистике добавили оптимизацию и computer science (то есть информатику, но в более широком смысле, чем это принято понимать в России)

Три основных составляющих Data Science:

1) **Организация данных —** хранение и форматирование. Также сюда входят практики управления данными (Data Management).
2) **Агрегация данных** — объединение исходных данных в новое представление и/или пакет.
3) **Доставка данных** — обеспечение доступа к массивам агрегированных данных.

3. Data Science - как это работает?

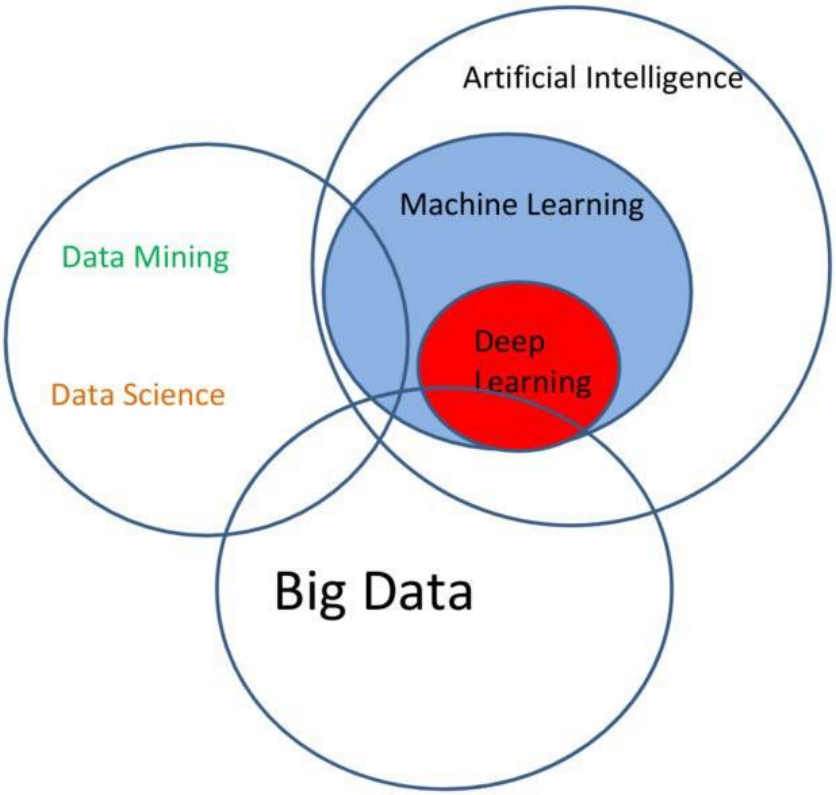


Сырые данные изначально беспорядочны и запутаны, собраны из различных источников и непроверенных записей. Не очищенные данные могут скрыть правду, зарытую глубоко в биг дате, и ввести в заблуждение аналитика.
Дата майнинг — это процесс очистки больших данных и подготовки их последующему анализу или использованию в алгоритмах машинного обучения. Дата майнеру нужно обладать исключительными распознавательными качествами, чудесной интуицией и техническими умениями для объединения и трансформирования огромного количества данных.
Исследователи применяют техники машинного обучения, чтобы автоматизировать решение некоторых задач. Эти системы очень нужны для работы с некоторыми очень сложными проектами. Например, чтобы узнать в какой стране живут самые счастливые люди, ученые определяли улыбки на фотографиях, загруженных в Instagram.

Традиционные риски Data Science проектов:

1) Высокая стоимость реализации проекта приведет к финансовым потерям (не окупится)
2) Отсутствие подробной отчетности по проекту не позволит отчитаться о потраченных средствах или принять правильное решение о продолжении проекта
3) Внедрение закрытого алгоритма или программы («Черный ящик») сделает невозможным дальнейшее изменение или модернизацию проекта внешними или внутренними ресурсами

4. Важные инструменты для работы с данными



4.1 Big Data

Big Data — это различные инструменты, подходы и методы обработки как структурированных, так и неструктурированных данных, которые позволяют использовать эти данные для решения конкретных задач и достижения целей. Используя анализ Big Data, розничные продавцы смогут заранее узнать, какие продукты будут хорошо продаваться, телекоммуникационные компании смогут предсказать, захочет ли клиент сменить оператора и когда это произойдёт, а страховые компании смогут оценить, насколько безопасно их клиенты управляют автомобилем. Среди прочего, анализ Big Data позволяет нам лучше понимать и прогнозировать эпидемии болезней и находить самые эффективные способы лечения.

4.2 Машинное обучение

Цитируя Тома Митчела: Машинное обучения изучает вопрос создания программ, способных улучшаться в процессе обучения.

Машинное Обучение носит междисциплинарный характер и использует, среди прочего, методы из области информатики, статистики и искусственного интеллекта.

Основной областью исследований в Машинном Обучении являются алгоритмы, которые способны обучаться и запоминать и могут применяться в различных областях науки и бизнеса.

4.3 Data Mining (Сбор и интеллектуальный анализ данных)

Файяд, Пятецкий-Шапиро и Смайт дают следующее определение Data Mining:

«Применение специальных алгоритмов для извлечения шаблонов из данных. В интеллектуальном анализе данных акцент делается на применение алгоритмов, а не на сами алгоритмы.»

Мы можем определить взаимосвязь машинного обучения и Data Mining следующим образом: интеллектуальный анализ данных — это процесс, в ходе которого алгоритмы МО используются в качестве инструментов для извлечения потенциально ценных шаблонов, содержащихся в наборах данных.

4.4 Deep Learning

Deep Learning — относительно новый термин, однако существовавший ещё до резкого роста повышения внимания к науке о данных.

Deep Learning — это процесс применения технологий глубоких нейронных сетей — архитектур нейронных сетей с несколькими скрытыми уровнями — для решения поставленных задач.

По сути это Data Mining, в котором используются архитектуры глубоких нейронных сетей — особого типа алгоритмов машинного обучения.

4.5 Искусственный интеллект

Искусственный интеллект — научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования тех видов человеческой деятельности, которые традиционно считаются интеллектуальными.

Исследования, связанные с ИИ, высокотехнологичны и узкоспециализированны. Одной из ключевых задач искусственного интеллекта является программирование компьютеров, которые демонстрируют такие способности, как понимание, рассуждение, решение проблем, восприятие, обучение, планирование и т. д.

Основные составляющие ИИ — машинное обучение, инженерия знаний (knowledge engineering) и робототехника.

5. Big Data≠Data Science

Big Data – это:

1) ETL\ELT
2) Технологии хранения больших объемов структурированных и не структурированных данных
3) Технологии обработки таких данных
4) Управление качеством данных
5) Технологии предоставления данных потребителю

Data Science – это:

1) Распознавание видео
2) Распознавание текстов
3) Распознавание речи
4) Построение рекомендательных моделей
5) Сегментация
6) Кластеризация и т.д.

6. Data Science в реалиях производства

1) Сложный и длительный во времени процесс
2) Требуется глубокое понимание предметной области
3) Разная частота съема данных и не все оцифровано
4) Нет сквозного контроля и фиксации событий тех.процесса
5) Доверие к модели со стороны технологов и операторов
6) Для проверок модели требуются эксперименты с данными реального времени на производстве

7. Заключение

1) Чем больше данных, тем сложнее их анализ.
2) Наука о данных — это знания о выводимых данных, отбор, подготовка и анализ.
3) Машинное обучение применяется для сбора и анализа массивов данных.
4) Дата майнинг — это процесс очистки больших данных и подготовки их к последующему анализу.

Принимая во внимание перечисленные научные области, концепции, и инструменты, мы можем заключить, что Data Science — это наше будущее. Наука о данных изменит мир, и сильно.