МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ)

Кафедра 402

Курсовая работа по теме

Data science.

Работу выполнил:

Студент группы

М4О-304Ск-18

Бубенцов Александр Алексеевич.

Москва

2020

**Содержание.**

1. **Появление Data science…………………………….…..........3**
2. **Характеристики больших данных……….…………....…..3**
3. **Основа Data science.……………………………………….....4**
4. **Типы данных………………………………………………....5**
5. **Процесс Data science……………………………………...….6**
6. **Где используется Data science……………………………....7**
7. **Заключение……………………………………...………..…..7**
8. **Список источников………………………………………..8**

**Data science.**

**Появление Data science.**

Data science - наука о проблемах обработки, представления и анализа данных в цифровом виде. Данная наука объединяет в себе методы обработки данных в условиях параллелизма и больших данных.

Начала зарождаться в момент учреждения комитета по данным для науки и техники, термин data science в свою очередь впервые был применен Peter Naur в своей книге под названием Concise Survey of Computer Methods 1974 года.

В 1990 годах data science начала активно развиваться и стала считаться дисциплиной, а в 2000 года стала общепризнанной благодаря профессору Университета Пердью Уильяму Кливленду, он опубликовал статью где предложил план изучения некоторых технических аспектов, которые по его мнению должны были преподаваться отдельной дисциплиной.

С 2011 приобрела высокую популярность за счет публикаций в журналах O’Reilly и конференций EMC (Dell EMC с 2015-2016 года)

**Характеристики больших данных.**

1. Объем - сколько данных содержит объект
2. Разнообразие - различие данных разного типа
3. Скорость - скорость генерации данных
4. Достоверность - точность данных

Данные характеристики отличают обычные данные от больших данных, которые встречаются в традиционных средствах управления данными.

поэтому изменения привнесенные данными характеристиками влияют в первую очередь на сбор данных, хранение, обслуживание, поиск и обмен.

**Основа Data science.**

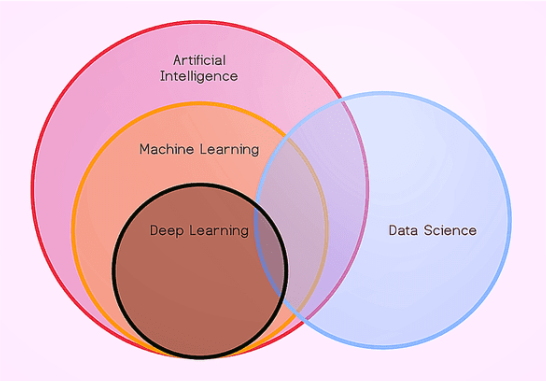


Рис.1

Ис­кус­ствен­ный ин­тел­лект (Ar­ti­fi­cial In­tel­li­gence) — данная область фокусируется на создании машин, думающих и работающих как люди. Её воз­ник­но­ве­ние связа­но с по­яв­ле­ни­ем ма­шин Ала­на Тью­рин­га в 1936 году. К сожалению на данный момент, искусственный интеллект все еще не может полностью заменить человека.

Ма­шин­ное обу­че­ние (Ma­chine learn­ing) — создание модели для более удобного извлечения данных. Машинное обучение происходит самостоятельно основываясь на какие либо данные

Глу­бо­кое обу­че­ние (Deep learn­ing) — происходит создание многочисленных нейронных систем по причине не справления с задачей машинного обучения . глубокое обучение получила свое название за счет того, что некоторые математические операции происходят глубоко в скрытых слоях нейронных сетей.

Боль­шие дан­ные (Big Data) — ра­бо­та с боль­шим объемом ча­сто неструктурированных дан­ных. Спе­ци­фи­ка сфе­ры — это инстру­мен­ты и си­сте­мы, спо­соб­ные выдерживать вы­со­кие нагруз­ки.

**Типы данных**

1. Структурированные
2. Неструктурированные
3. Машинные
4. На естественном языке
5. Графовые
6. Аудио, видео и графика
7. Потоковые

Структурированные.

Структурированные данные зависят от модели данных и хранятся в фиксированных ячейках внутри документа. Такая информация чаще всего хранится в таблицах наподобие EXCEL. SQL язык предназначенный для управления и обращения к структурированным данным.

Неструктурированные.

Неструктурированный данный в отличии от структурированных сложно подогнать под какую либо модель данных, поэтому их содержании имеет переменный характер и часто зависит от контекста.

Машинный данные.

К машинным данным можно отнести информацию которая генерируется компьютером без вмешательства человека в данный процесс. На данный момент машинная информация является основным источником данных.

На естественном языке.

Данный на естественном языке являются разновидностью неструктурированных данных, эти данные сложны в обработке, т.к данный вид требует лингвистики и специальных методов data science.

Графовые

Данный тип данных представляется как математические графы данная математическая структура нужна для последующего моделирования попарных отношений между объектами данных. Графовые данные чаще всего используются в социальных сетях.

Аудио, видео и графика.

Крайне сложный вид данных для data science, т.к если рассматривать анализ картинки системе будет сложно определить что на ней находится

Потоковые

Потоковые данные могут принимать форму любого из вышеперечисленных типов данных, но в отличии от них данные поступают большими массивами при происхождении какого нибудь события.

**Процесс Data science.**

Сначала назначается цель исследования, делается проект в котором указывается цель исследования и нужные ресурсы. После чего происходит сбор данных которые указываются в проекте, данные могут быть найдены самостоятельно или получены от 3-го лица. Следующий этап подготовка данных, на этом этапе данные проходят проверку ради улучшения их качества и в последующем применения их на дальнейших этапах, данный этап состоит их 3-х этапов: очистка, преобразование и интеграция. После чего происходит исследование данных, на этой стадии углубляются в полученные данные, для лучшего понимания и исследования их взаимодействий. Предпоследний этап происходит модуляция данных для получения ответа на поставленную задачу и последний этап представление результатов заказчику и если надо автоматизировать процесс.

**Где используется Data science.**

1. В бизнесе, прогноз на спрос.
2. В банках,
3. В транспортных компаниях, для выстраивания оптимального маршрута.
4. В страховых компаниях, для оценки вероятности страхового случая
5. На производстве, для прогноза сбоя оборудования
6. В метеослужбах, для прогнозов погоды
7. В IT - сфере, создание ботов и поиск информации
8. В сельском хозяйстве, для прогноза на урожай

**Заключение**

Данная отрасль крайне актуальна, и востребована в последнее время, а так как она стала крайне актуальна относительно недавно, существует большой дефицит в кадрах. За счет того, то Data science охватывает Big Data и Ar­ti­fi­cial In­tel­li­gence данная отрасль будет актуальна еще долгое время.

Список источников.

1. <https://www.osp.ru/os/2012/05/13016245> (дата обращения 27.12.2020)
2. [https://blog.skillfactory.ru/perevod/what\_is\_ds](https://blog.skillfactory.ru/perevod/what_is_ds/) (дата обращения 27.12.2020)
3. <https://www.tadviser.ru/index.php> (дата обращения 27.12.2020)
4. <https://www.osp.ru/cio/2018/03/13054275> (дата обращения 27.12.2020)
5. <https://skillbox.ru/media/code/chto_takoe_data_science_i_kto_takoy_data_scientist/> (дата обращения 27.12.2020)

## Силен,Мейсман.Основы Data Science и Big Data.Python и наука о данных.2017

1. <https://netology.ru/blog/01-2020-gid-po-data-science> (дата обращения 27.12.2020)

Ссылка на картинку.

1. https://netology.ru/blog/wp-content/uploads/2020/01/860.png