



# Введение в анализ данных

3 октября 2019 г.

## DATA PREPARATION

### DATA CLEANING      TRANSFORMATION

INCONSISTENT DATATYPES



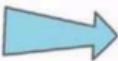
talend

MISSPELLED ATTRIBUTES



informatica

MESSING AND DUPLICATE VALUES

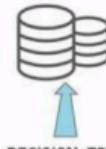


## EXPLORATORY DATA ANALYSIS



DEFINES AND REFINES  
THE SELECTION OF FEATURE  
VARIABLES THAT WILL BE USED  
IN THE MODEL DEVELOPMENT

## DATA MODELING



NAIVE BAYES



KNN

## VISUALIZATION AND COMMUNICATION



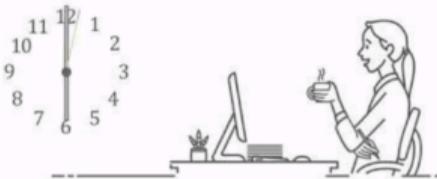
# WHAT IS DATA SCIENCE?



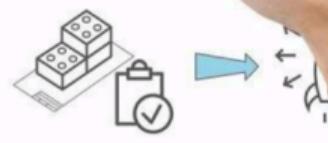
DATA ACQUISITION  
- WEB SERVERS  
- LOGS  
- DATABASES  
- APIs  
- ONLINE REPOSITORIES



WHY?...WHY?...WHY?...



## DEPLOYS AND



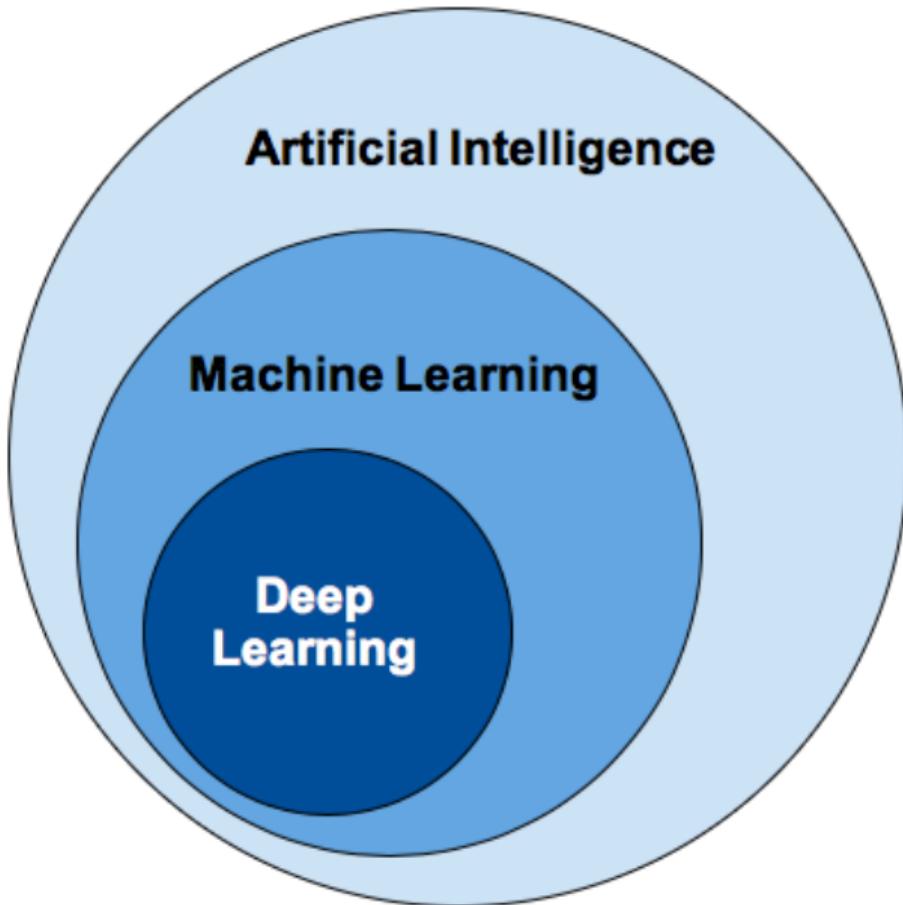
AI, ML and DL

---

# Что это за понятия?

---

- Artificial Intelligence (Искусственный интеллект)
- Machine Learning (Машинное обучение)
- Deep Learning (Глубокое обучение)



 Andrew Chen Retweeted



**Mat Veloso** @matveloso · Nov 22



Difference between machine learning and **AI**:

If it is written in Python, it's probably machine learning

If it is written in **PowerPoint**, it's probably **AI**



166



6.6K



19K



Show this thread

# Machine Learning

---

# Зачем обучать машины и что для этого нужно?

---

- данные
- признаки
- алгоритм

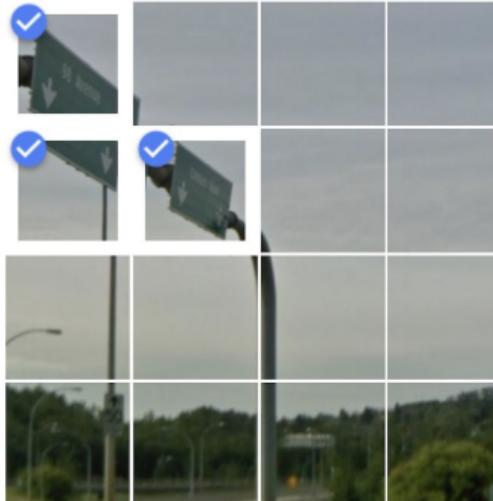


# Классическое Обучение



[https://vas3k.ru/blog/machine\\_learning//](https://vas3k.ru/blog/machine_learning//)

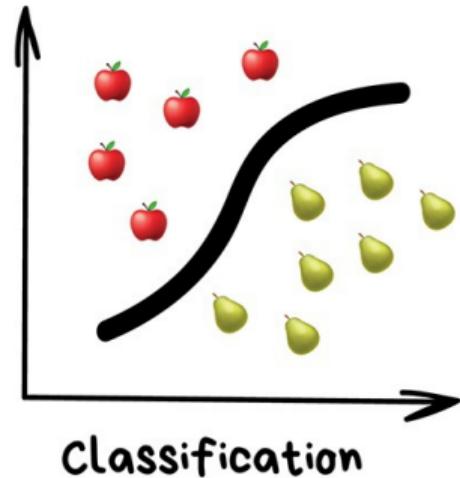
Select all squares with  
**street signs**  
If there are none, click skip



VERIFY

# Задача классификации

- спам-фильтр
- определение подозрительных транзакций
- кредитный scoring
- определение языка
- анализ тональности
- оценка состояния человека по ЭЭГ



# Credit Score

Excellent

Good

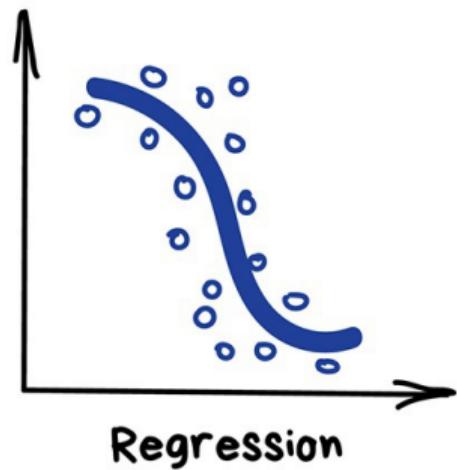
Fair



Показатель	Значение (или диапазон) показателя	Скоринг - балл
<i>Пол</i>	женщина	25
	мужчина	20
<i>Возраст</i>	меньше 30 лет	30
	30-45 лет	35
	больше 45 лет	28
<i>Образование</i>	среднее	22
	среднее специальное	28
	неоконченное высшее	30
	высшее	40
<i>Трудовой стаж</i>	до 1 года	16
	1-5 лет	19
	5-10 лет	24
	более 10 лет	31
<i>Семейное положение</i>	холост	20
	женат/замужем	30
<i>Наличие авто</i>	есть	40
	нет	20

# Задача регрессии

- прогноз стоимости ценных бумаг
- анализ спроса
- прогнозирование температуры воздуха
- прогнозирование цены дома
- прогнозирование объема потребления электроэнергии



# Нейронные сети

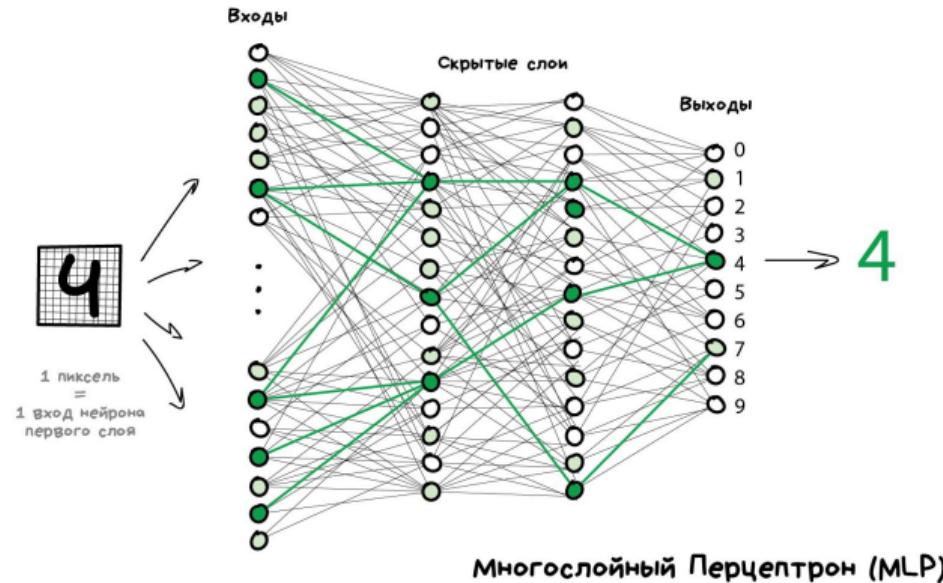
---

# Какие задачи решают?

---

- компьютерное зрение
- распознавание речи
- поиск(web search)
- распознавание лиц
- генерация текста
- обучение с подкреплением

# Нейронные сети



[https://vas3k.ru/blog/machine\\_learning//](https://vas3k.ru/blog/machine_learning//)



# Генерация текста

Iteration: 100

tyntd-iafhatawiaoihrdemot lytdws e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e  
plia tkldrgd t o idoe ns,smtt h ne etie h,hregtrs nigtike,aoaenns lng

Iteration: 300

"Tmont thithey" fomesscerliund  
Keushey. Thom here  
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome  
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."

Iteration: 700

Aftair fall unsuch that the hall for Prince Velzonski's that me of  
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort  
how, and Gogition is so overelical and ofter.

Iteration: 2000

"Why do what that day," replied Natasha, and wishing to himself the fact the  
princess, Princess Mary was easier, fed in had oftened him.  
Pierre aking his soul came to the packs and drove up his father-in-law women.

# Обучение с подкреплением



Каспаров играет против Deep Blue в 1997 году

# Обучение с подкреплением

---

- [https://www.youtube.com/watch?time\\_continue=9&v=qv6UVOQ0F44](https://www.youtube.com/watch?time_continue=9&v=qv6UVOQ0F44)
- <https://www.youtube.com/watch?v=Aut32pR5PQA&t=77s>

Немного про нас

---

НОВИЧКИ В  
ПРОГРАММИРОВАНИИ

AI И  
МАШИННОЕ ОБУЧЕНИЕ

АЛГОРИТМЫ

СТРУКТУРЫ  
ДАННЫХ

OOП

HELLO WORLD



# Что мы научимся делать?

---

- поймём как должны выглядеть данные
- научимся работать с данными, начиная от их сбора и заканчивая передачей в модель
- будем красиво визуализировать наши данные
- разберемся как работает линейная регрессия
- посмотрим на задачу классификации и обучим логистическую регрессию
- немножко поработаем с графовой кластеризацией

Финиш

---