



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

ВВЕДЕНИЕ В ОБЛАСТЬ DATA SCIENCE

Теванян Элен

Москва 2019

ЭЛЕН ТЕВАНИЯН



X5 RETAIL GROUP



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Сейчас:

- Data Scientist
- Преподаватель
- Исследователь-биоинформатик
- Аспирант

В предыдущих сериях:



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

- Магистр'18 (Анализ данных в биологии и медицине)
- Бакалавр'15 (Прикладная математика)

ФРИЦ

Альфа·Банк

Организация курса

САМЫЕ ВАЖНЫЕ ССЫЛКИ

http://wiki.cs.hse.ru/Информационный_менеджмент:_Введение_Data_Science

https://fulyankin.github.io/HSE_Data_Culture/

ОЦЕНИВАНИЕ КУРСА

$$O_{\text{итог}} = \max\{0.7 \cdot O_{\text{накопленная}} + 0.3 \cdot O_{\text{Экзамен}}, \\ 0.5 \cdot O_{\text{накопленная}} + 0.5 \cdot O_{\text{Экзамен}}\}$$

$$O_{\text{накопленная}} = 0.1 \cdot O_{\text{DataCamp}} + 0.1 \cdot O_{\text{Семинары}} + \\ + 0.2 \cdot O_{\text{Самостоятельные}} + 0.2 \cdot O_{\text{Кейс}} + \\ + 0.2 \cdot O_{\text{дз1}} + 0.2 \cdot O_{\text{дз2}}$$

КОНТАКТЫ

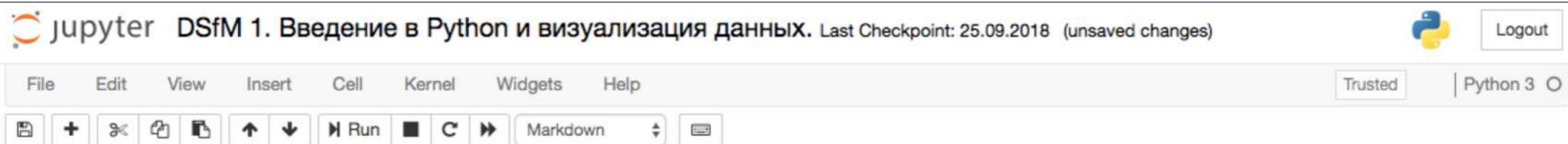
elentevanyan@gmail.com

tg: @elentevanyan

ЧТО БУДЕМ УМЕТЬ ПО ОКОНЧАНИЮ КУРСА

1. Немножко питонить
2. Делать описательный анализ данных
3. Разбираться в двух классах задач машинного обучения
4. Выбирать правильные метрики
5. Делать простые модели
6. Измерять результаты изменений

JUPYTER-NOTEBOOK



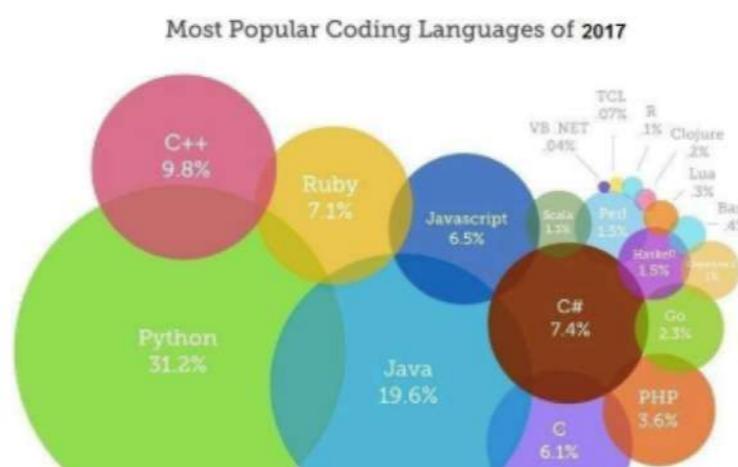
0. Где мы сейчас?

Jupyter Notebook - интерактивная среда для запуска программного кода в браузере. Удобный инструмент для анализа данных, который используется многими специалистами по data science.



1. Python

Python - это свободный интерпретируемый объектно-ориентированный расширяемый встраиваемый язык программирования очень высокого уровня (Г.Россум, Ф.Л.Дж.Дрейк, Д.С.Откидач "Язык программирования Python").



ЕСЛИ ПОЛЬЗУЕМСЯ ЛИЧНЫМИ НОУТАМИ

Ставим анаконду с питоном 3.7

<https://www.anaconda.com/distribution/#download-section>



Data Science

Результатов: примерно 2 450 000 000 (0,37 сек.)

Очный курс «Data Scientist» | Обучение в Нетологии. 18+ | netology.ru

Реклама www.netology.ru/ ▾ 8 (800) 301-39-69

Data Science, Machine learning, Python - 6 месяцев обучения, старт 16 октября. Эксперты-преподаватели. Практические занятия. Диплом об окончании.

📍 Варшавское шоссе 1с6, Москва

Онлайн обучение Data Science | От Mail.Ru Group | GeekBrains.ru

Реклама www.geekbrains.ru/data/science ▾

Стань специалистом в области больших данных с гарантированным трудоустройством! Научим решать настоящие бизнес-задачи. Онлайн обучение: 2-4 занятия в неделю. Обеспеченное будущее. Онлайн обучение с нуля. От Mail.Ru Group. Государственная лицензия.

GeekUniversity · Бесплатные вебинары · Обучение IT-Профессиям · Отзывы студентов о курсах

📍 Ленинградский проспект, 39, строение 79, Москва

Видео



Data Science In 5 Minutes | Data Science For Beginners | What Is Data ...
Simplilearn
YouTube - 4 дек. 2018 г.



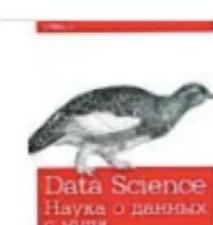
What REALLY is Data Science? Told by a Data Scientist
Joma Tech
YouTube - 23 июн. 2018 г.



Learn Data Science in 3 Months
Siraj Raval
YouTube - 30 окт. 2018 г.

Результаты по запросу "д..."

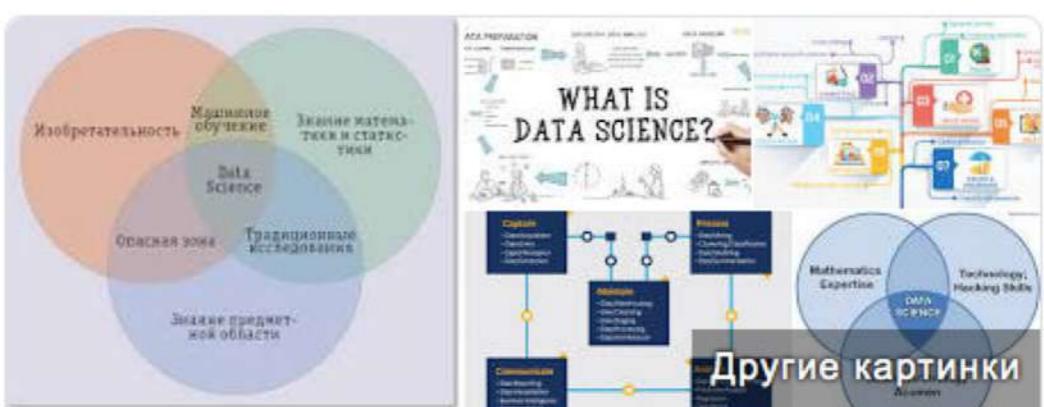
Реклама



(0+) Data Science. Наука о данных с нуля

827 ₽
Labirint.Ru

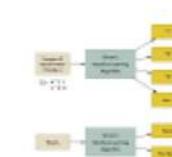
→ Другие результаты в Google



Наука о данных

Наука о данных — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме.
[Википедия](#)

Похожие запросы



Машинное обучение



Аналитика



Информатика



Искусственный интеллект



Финансы

Ещё 10+

Оставить отзыв

DATA

Data Scientist: The Sexiest Job of the 21st Century

by [Thomas H. Davenport](#) and [D.J. Patil](#)

FROM THE OCTOBER 2012 ISSUE

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Data Scientist Salaries

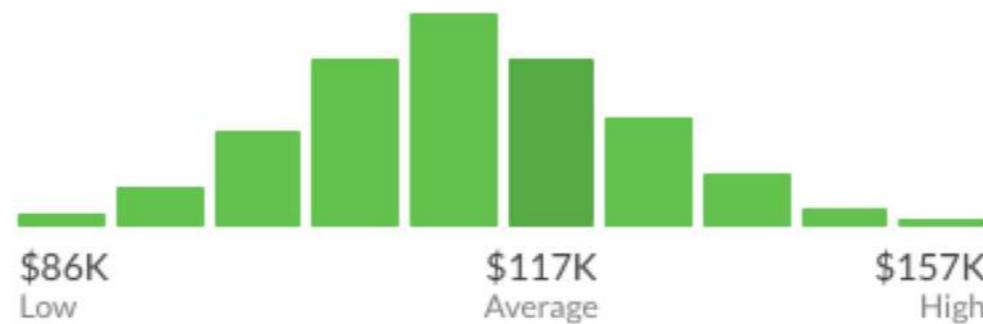
About This Data ?

4,354 Salaries Updated Apr 3, 2019

Industries ▼ Company Sizes ▼ Years of Experience ▼

Average Base Pay

\$117,345 /yr



Additional Cash Compensation ?

Average \$11,530

Range \$3,933 - \$26,784

How much does a Data Scientist make?

The national average salary for a Data Scientist is \$117,345 in United States. Filter by location to see... [More](#)

Salaries for Related Job Titles

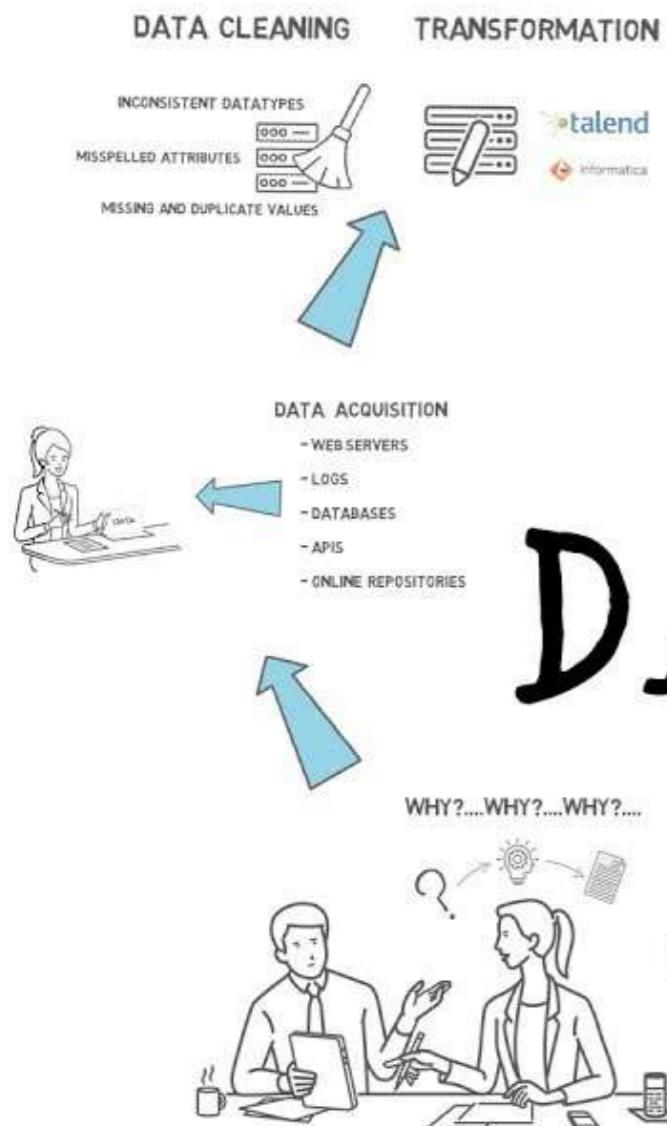
[Data Analyst](#) \$67K

[Quantitative Analyst](#) \$116K

[Senior Data Scientist](#) \$137K

https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_K00,14.htm

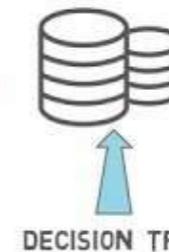
DATA PREPARATION



EXPLORATORY DATA ANALYSIS



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



WHAT IS DATA SCIENCE?



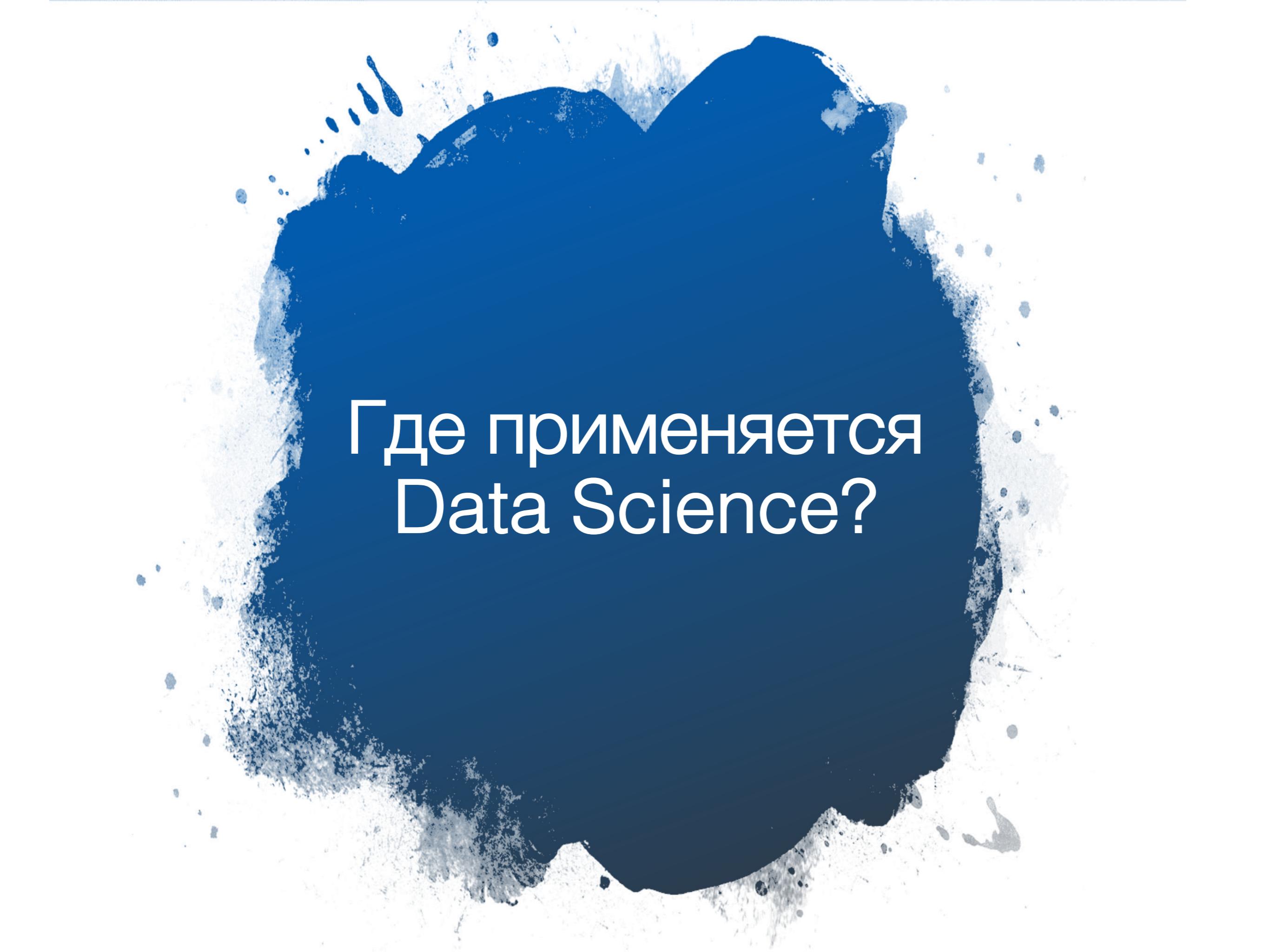
DEPLOYS AN



<https://www.youtube.com/watch?v=X3paOmcrTjQ>

DATA SCIENCE – ЭТО

практическая деятельность по обработке, анализу и
предоставлению данных



Где применяется
Data Science?

ПОИСК

Яндекс

что делать если скучно

Найти

что делать если скучно

что делать если скучно за компом

что делать если скучно дома

что делать если скучно на уроке

что делать если скучно трум трум

что делать если скучно в симс 4

что делать если скучно дома мне 12 лет девочке

что делать если скучно дома мне 10 лет девочке

что делать если скучно за компом ссылки на прикольные сайты

что делать если скучно с подругой дома

наука о данных (англ. *data science*, иногда даталогия — *datalogu*) — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных. Скрыть

Что такое **data science** и как это работает? | Ru...

[rb.ru](#) > Авторские колонки > [чтo-takoe-dt](#)

Термин **data science** на русский переводят как «наука о данных», а в профессиональной среде часто просто транслитерируют – «дата сайенс». Формально это набор некоторых взаимосвязанных дисциплин и методов из... [Читать ещё >](#)



Data Science Skills / Хабр

[habr.com](#) > post/271085/

Data Science также немного пересекается с такими областями деятельности как ... Data Science – это новая область деятельности, поэтому требования к Data Scientists еще не до конца сформированы. [Читать ещё >](#)

Дорога в Data Science глазами новичка

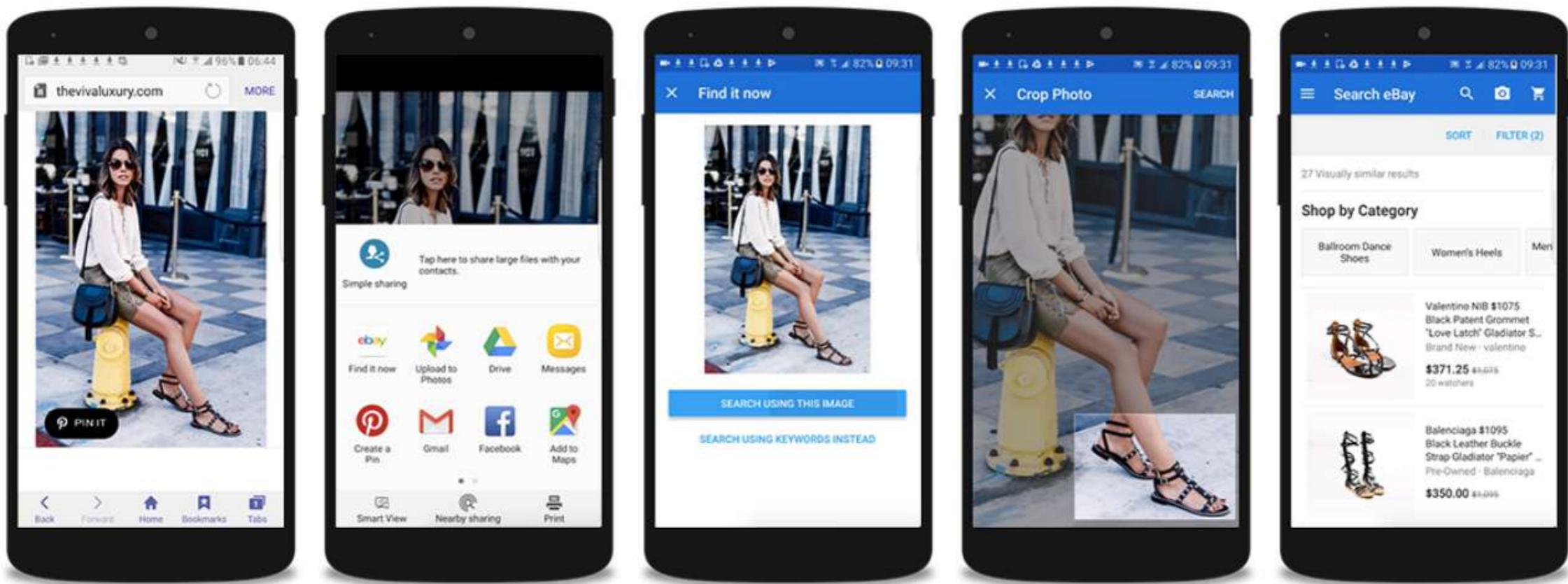
[pikabu.ru](#) > story/doroga_v_data_science_glaz...

Что такое Data Science? В 21 веке информация повсюду. Вы буквально не можете жить, не оставляя вокруг себя информационный след. [Читать ещё >](#)



Элен Теванян

ПОИСК ТОВАРОВ



ebay

РАНЖИРОВАНИЕ

Яндекс

наушники

Найти



плюс



Элен Теванян

Поиск

Картинки

Видео

Карты

Маркет

Новости

Эфир

Коллекции

Знатоки

Услуги

Ещё

1200x1200

700x700

600x600

694x694



• Купить Наушники в интернет-магазине М.Виде...

Наушники Bluetooth Наушники-вкладыши
mvideo.ru > naushniki/naushniki-3967

Наушники в интернет-магазине «М.Видео» представлены широким ассортиментом устройств. Цены варьируются от 390 до 109990 рублей. Читать ещё >



Нашёлся 791 млн результатов

3 млн показов в месяц

[Дать объявление](#) [Показать все](#)

w Наушники — Википедия

ru.wikipedia.org > Наушники

Стереофонические наушники (наушники) — два телефона с оголовьем, предназначенные для подключения к бытовым радиоэлектронным аппаратам. Читать ещё >



• Купить наушники в Москве, низкие цены на на...

svyaznoy.ru > catalog/audiovideo/1558

В интернет магазине Связной представлен широкий выбор наушников с онлайн-подбором совместимых брендов и моделей. В нашем каталоге Вы можете заказать... Читать ещё >



• Наушники купить наушники и гарнитуры... - Мо...

doctorhead.ru > Наушники

Лучшие наушники в интернет магазине Doctorhead.ru, широкий ассортимент, проводные и беспроводные наушники, для плеера и профессиональные.



OZON.ru | Количество излучателей в каждом на...

OZON.ru > catalog/1196662/

Наушники и гарнитуры. 2364 товара. ... Наушники TWS Наушники Беспроводные с микрофоном TWS I9S, TWS-I9S, белый. Читать ещё >

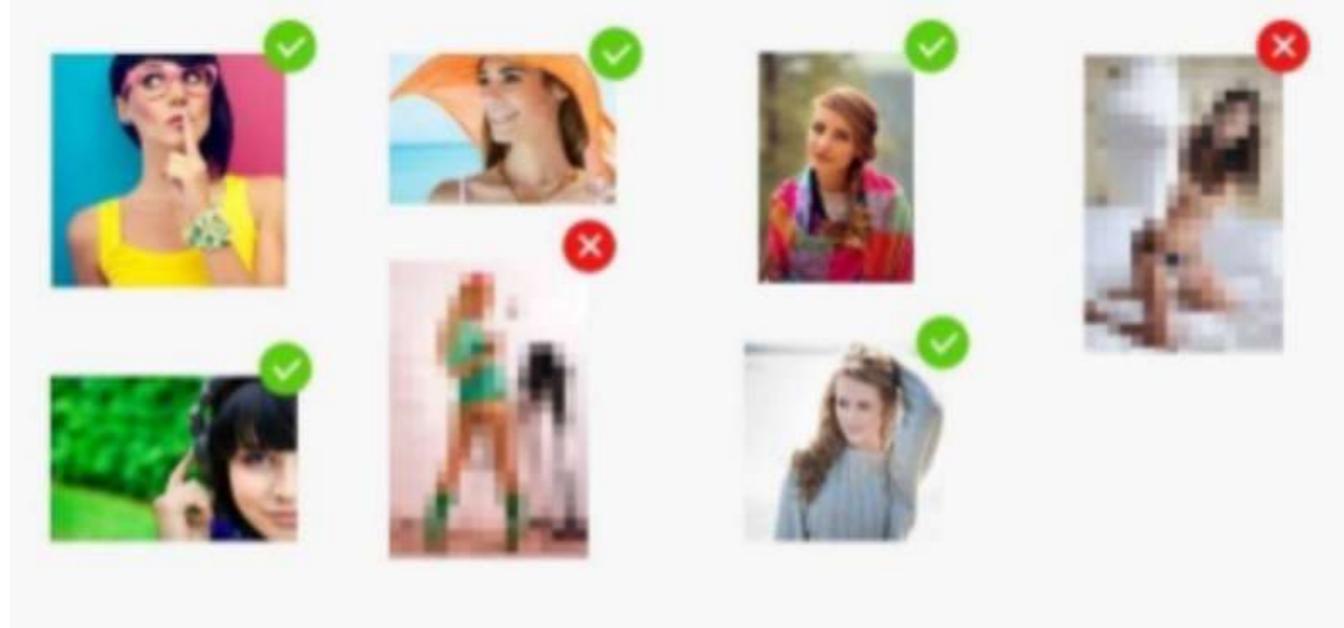


• Наушники и Bluetooth-гарнитуры на Маркете

Яндекс.Маркет > Наушники и Bluetooth-гарнитуры

Наушники и Bluetooth-гарнитуры — купить по выгодной цене с доставкой. 6641 моделей в проверенных интернет-магазинах: популярные новинки и лидеры продаж. Поиск по параметрам, удобное сравнение моделей и цен на Яндекс.Маркете.

МОДЕРАЦИЯ СЕРВИСОВ И САЙТОВ



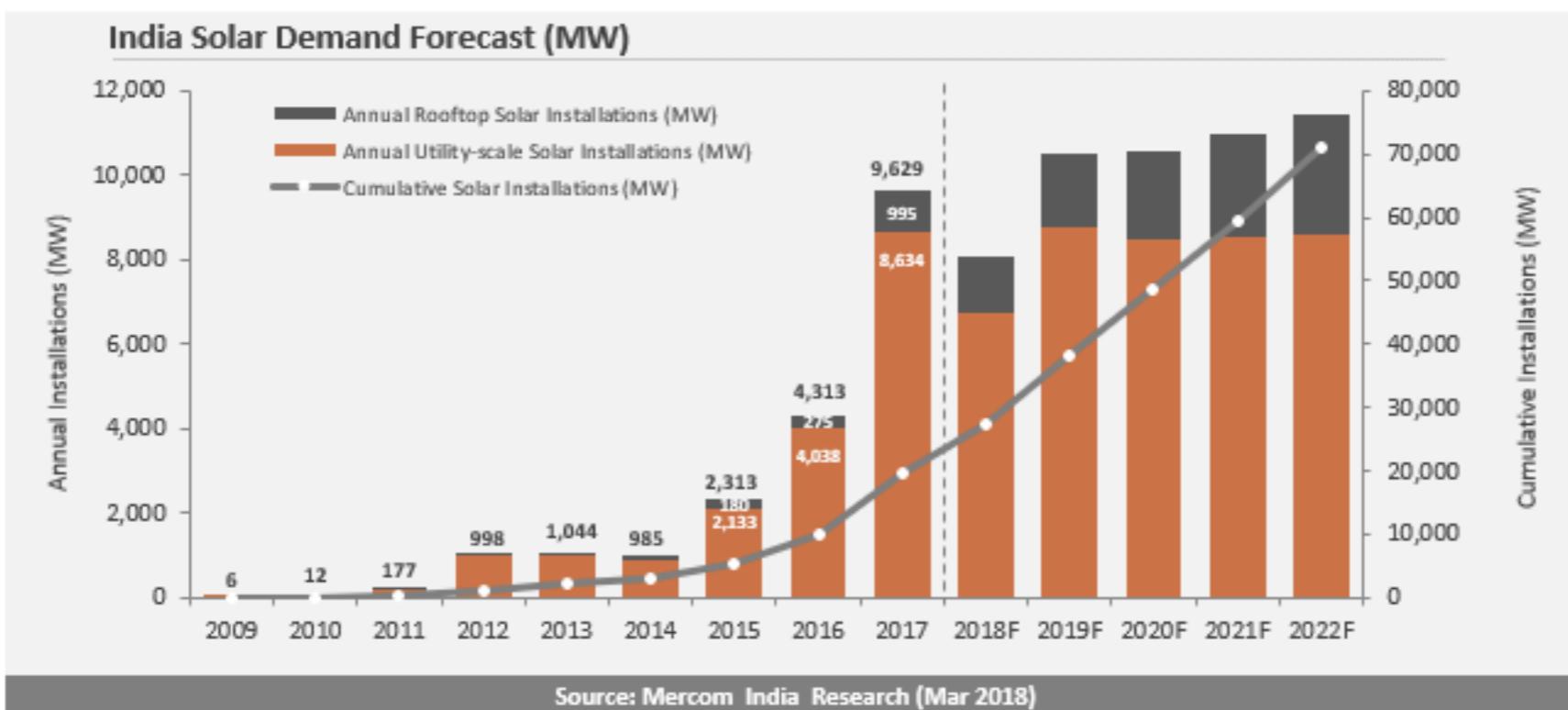
КРЕДИТНЫЙ СКОРИНГ



ОТТОК КЛИЕНТОВ



ПРОГНОЗИРОВАНИЕ СПРОСА



ПРЕДСКАЗАНИЕ ПРОБОК



РЕКОМЕНДАЦИИ

Похожие товары



Canon EOS 200D Kit

от 32 390 ₽

1 отзыв 109 предложений

Любительская зеркальная
фотокамера

Байонет Canon EF/EF-S

Объектив в комплекте, модель
уточняйте у продавца

Матрица 25.8 МП (APS-C)

Цвет:

Цены 109



Canon EOS 750D Kit

от 32 650 ₽

5 отзывов 133 предложения

Любительская зеркальная
фотокамера

Байонет Canon EF/EF-S

Объектив в комплекте, модель
уточняйте у продавца

Матрица 24.7 МП (APS-C)

Цвет:

Цены 133



Canon EOS M10 Kit

от 21 250 ₽

4 отзыва 89 предложений

Фотокамера с поддержкой
сменных объективов

Байонет Canon EF-M

Объектив в комплекте, модель
уточняйте у продавца

Матрица 18.5 МП (APS-C)

Цвет:

Цены 89



до -30%



Canon EOS M6 Kit

от 38 300 ₽

до 84 940 ₽

2 отзыва 105 предложений

Фотокамера с поддержкой
сменных объективов

Байонет Canon EF-M

Объектив в комплекте, модель
уточняйте у продавца

Матрица 25.8 МП (APS-C)

Цвет:

Цены 105



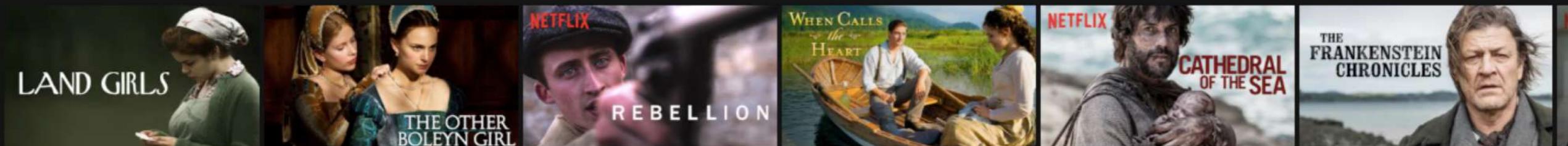
ПЕРСОНАЛИЗИРОВАННЫЕ РЕКОМЕНДАЦИИ



Watch It Again



Because you watched Outlander



Critically-acclaimed TV Dramas ›



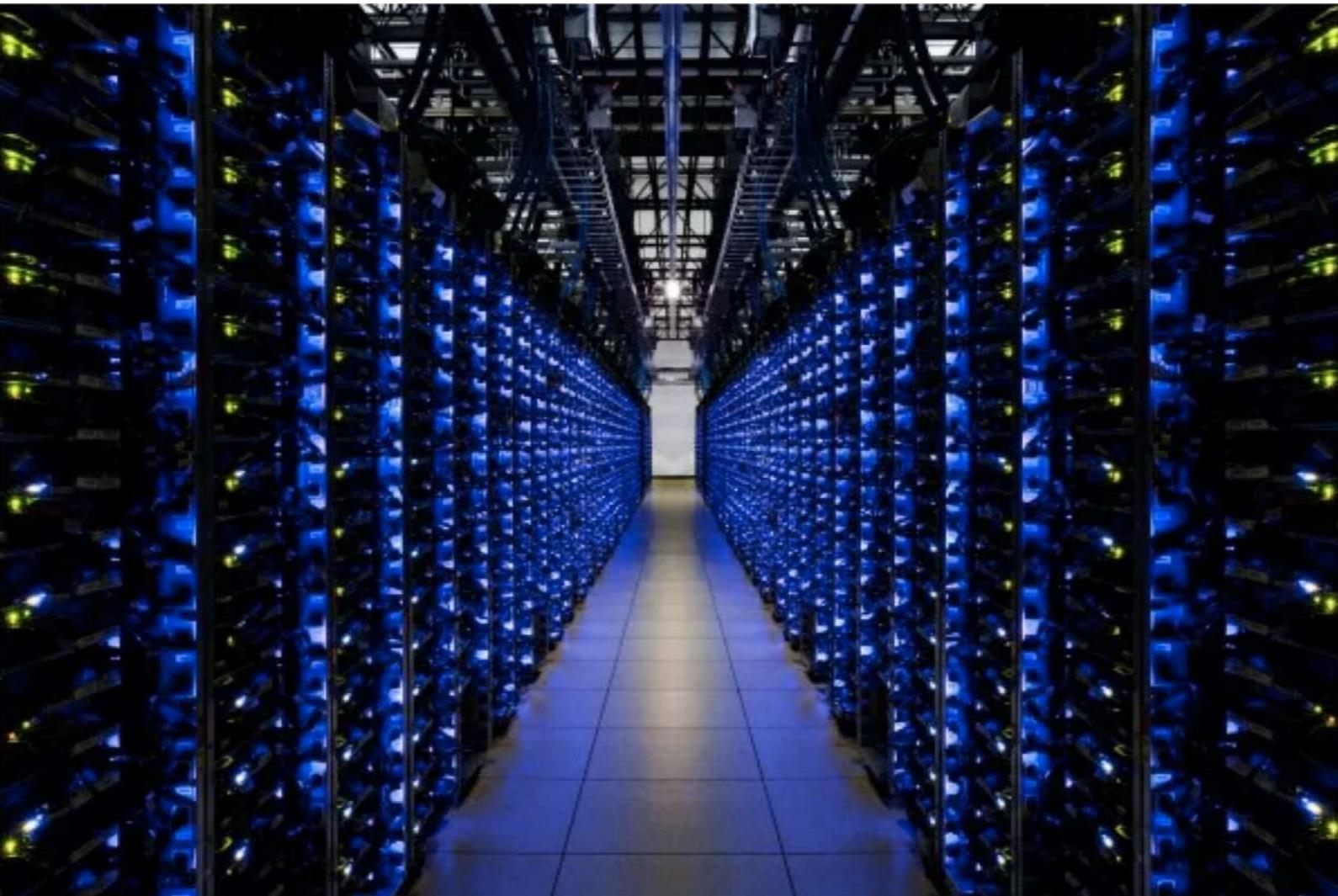
Romantic TV Dramas



ВЫПЛАВКА СТАЛИ

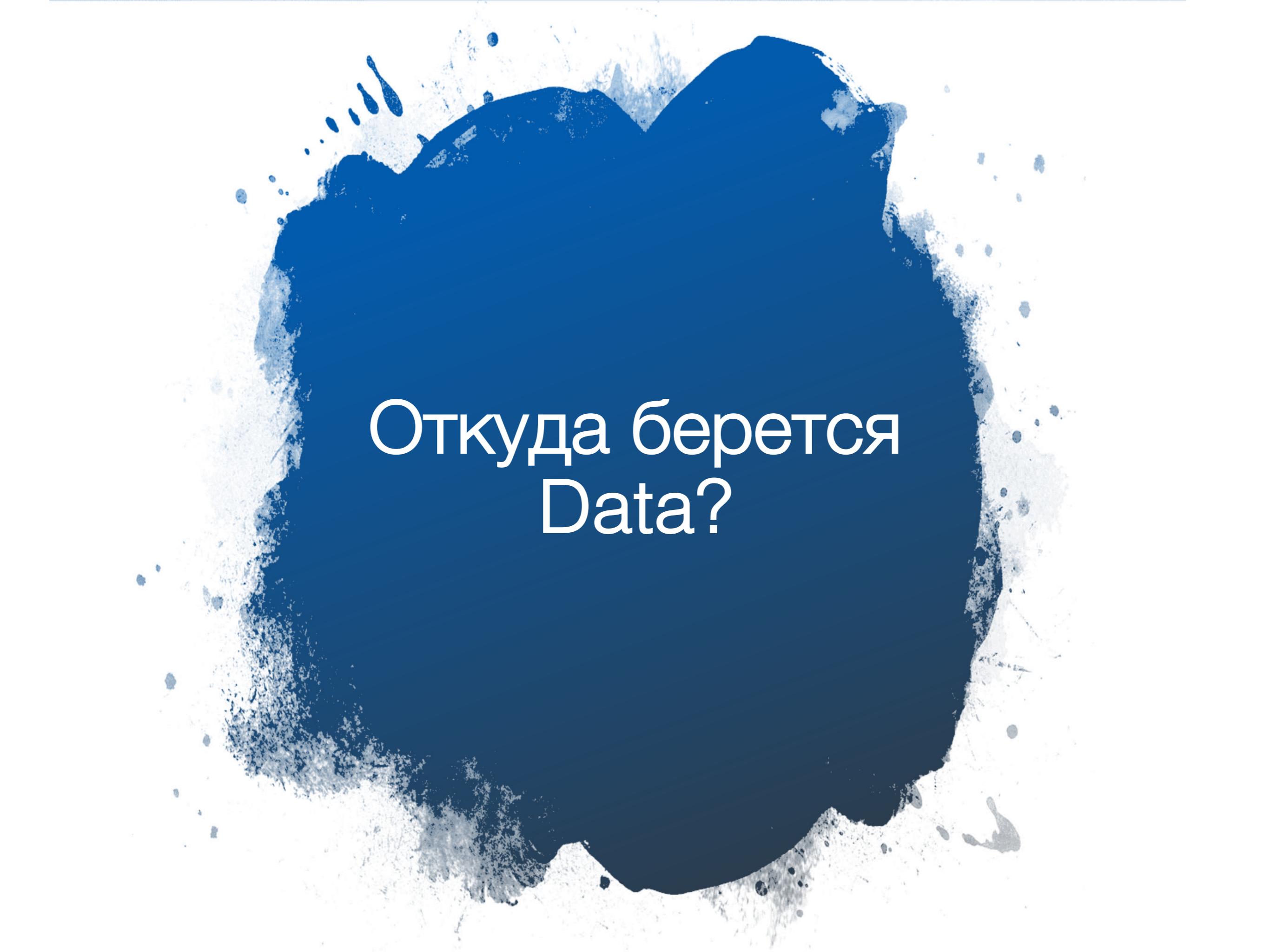


ОХЛАЖДЕНИЕ ДАТА-ЦЕНТРА



РАСПОЗНАВАНИЕ ЛИЦ





Откуда берется
Data?

ПОИСКОВЫЕ СЕРВИСЫ

Яndex

Google

СОЦИАЛЬНЫЕ СЕТИ



VKontakte



Twitter, Inc.



Instagram

...

БАНКИ И ТРАНЗАКЦИИ

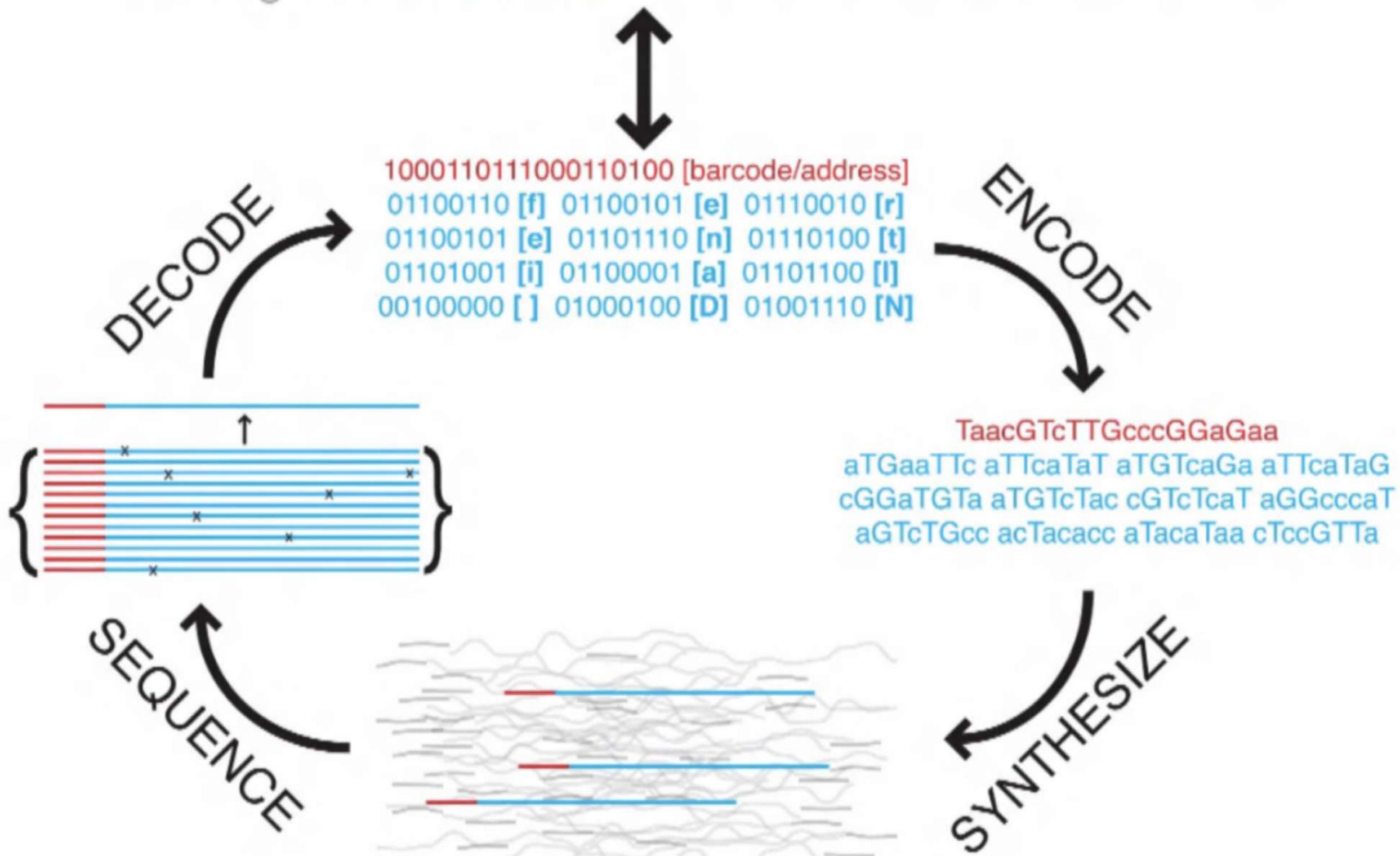


ТЕЛЕКОМ

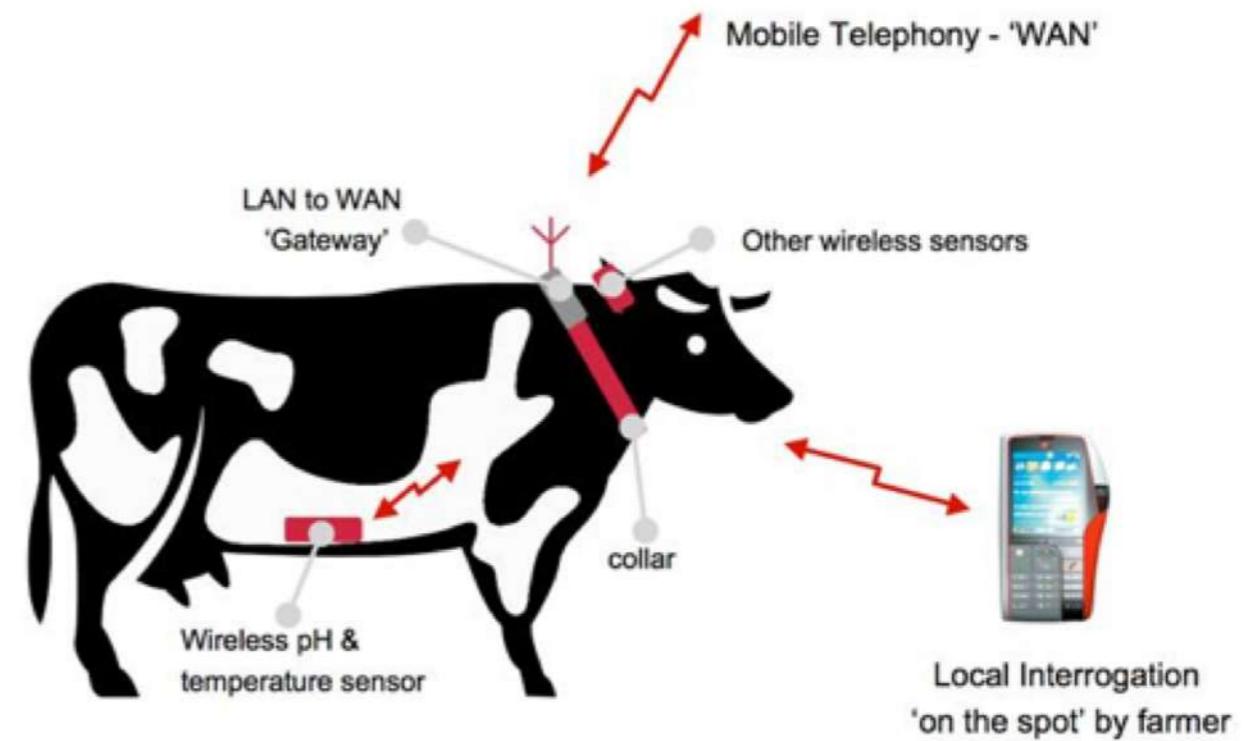


БИОЛОГИЯ

Decoding self-referential DNA that encodes these notes.



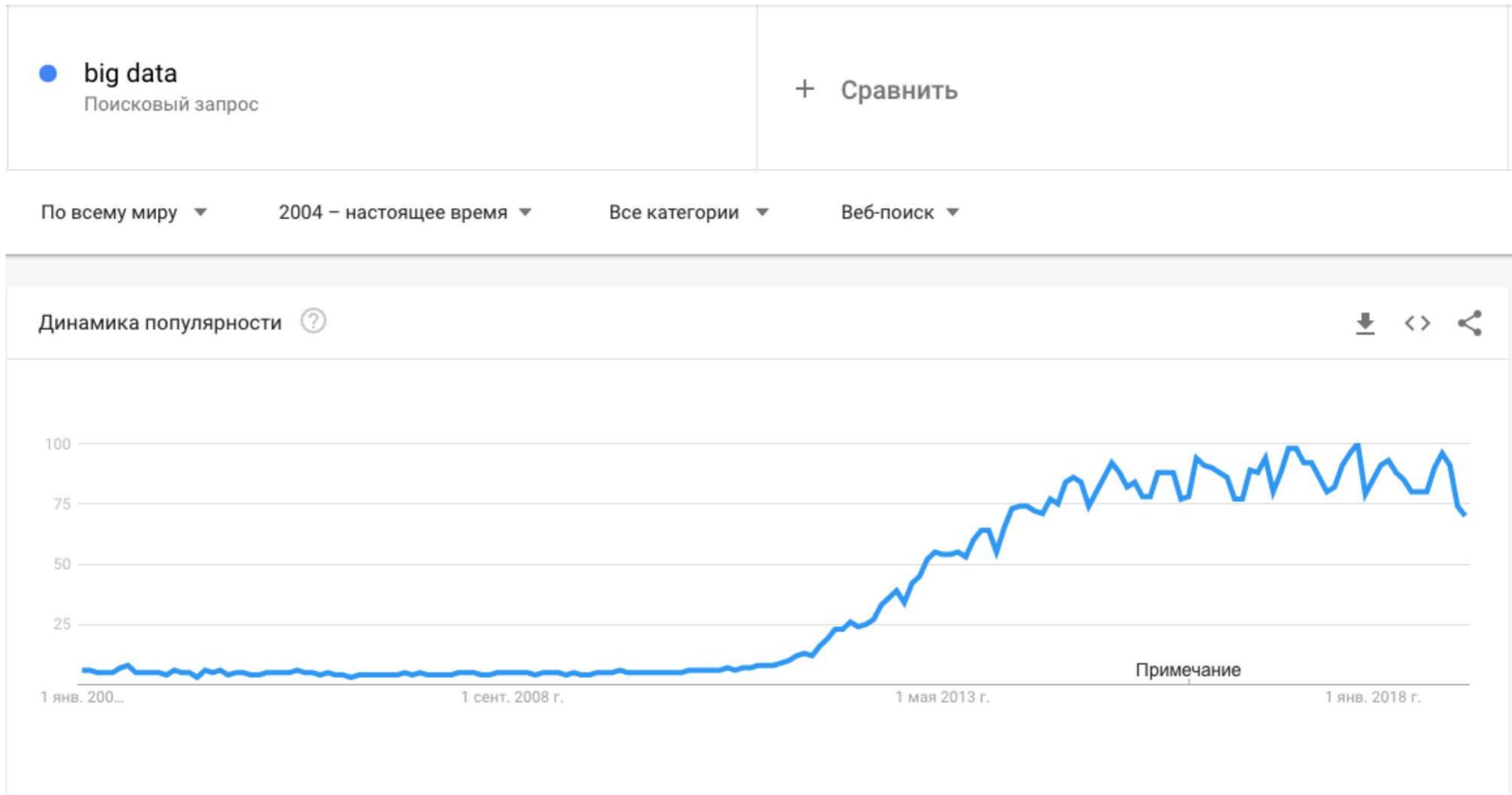
ИНТЕРНЕТ ВЕЩЕЙ





Big Data

BIG DATA



BIG DATA

- Это вообще все данные
- Это данные, превышающие определенный объем
(больше 100ГБ, 500ГБ, 1 ТБ, 10 ТБ и т.д.)
- Это данные, которые невозможно обработать на одном компьютере
- Термин не существует, придуман маркетологами



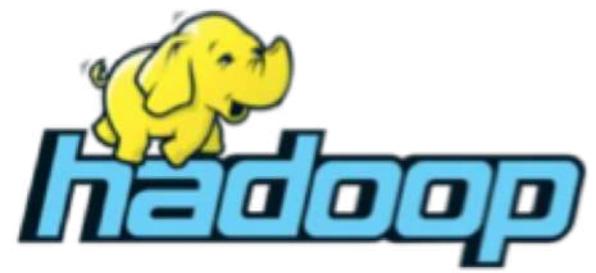
BIG DATA

- Это подходы, методы и инструменты для хранения и обработки структурированных и неструктурированных данных

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Education
41	Yes	Travel_Rarely	1102	Sales		1	2 Life Science
49	No	Travel_Frequently	279	Research & Development		8	1 Life Science
37	Yes	Travel_Rarely	1373	Research & Development		2	2 Other
33	No	Travel_Frequently	1392	Research & Development		3	4 Life Science
27	No	Travel_Rarely	591	Research & Development		2	1 Medical
32	No	Travel_Frequently	1005	Research & Development		2	2 Life Science
59	No	Travel_Rarely	1324	Research & Development		3	3 Medical

browser_id	utc_millisecs	Page	Pageviews
1. 3ef08ba3-6ec6-4ed1-a5b3-0c902cde2dd4	1436528937532	/chicken.htm	1 (16.67%)
2. 3ef08ba3-6ec6-4ed1-a5b3-0c902cde2dd4	1436528986774	/chicken.htm	1 (16.67%)
3. 3ef08ba3-6ec6-4ed1-a5b3-0c902cde2dd4	1436528994724	/chicken.htm	1 (16.67%)
4. ce452a5d-b5e3-446d-adb2-73cebfa750a5	1436529019861	/chicken.htm	1 (16.67%)
5. ce452a5d-b5e3-446d-adb2-73cebfa750a5	1436529037408	/chicken.htm	1 (16.67%)
6. ce452a5d-b5e3-446d-adb2-73cebfa750a5	1436529088962	/chicken.htm	1 (16.67%)

BIG DATA





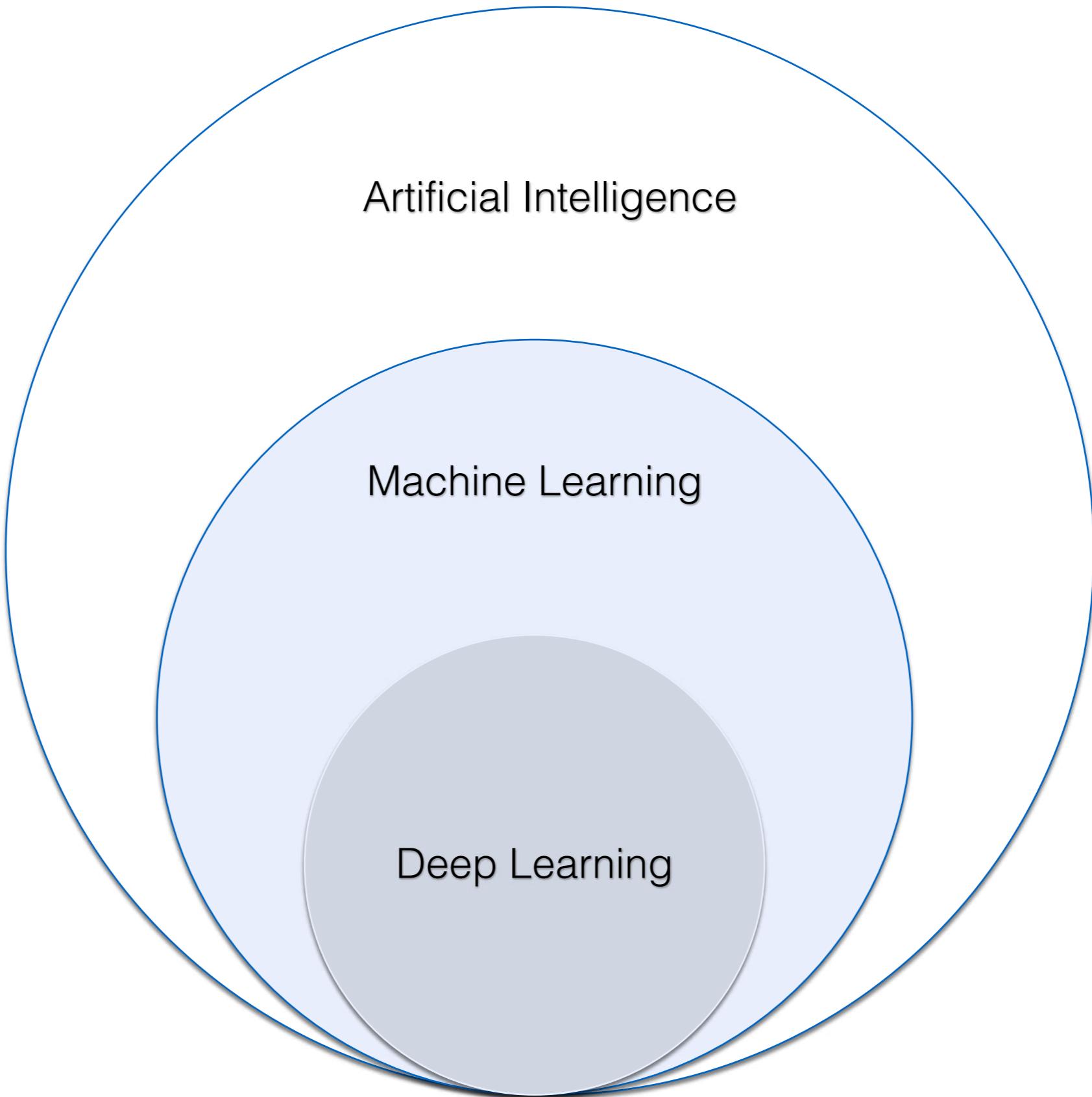
AI, ML, DL

КАК СВЯЗАНЫ ЭТИ ПОНЯТИЯ?

Artificial
Intelligence
(Искусственный
интеллект)

Machine
Learning
(Машинное
обучение)

Deep Learning
(Глубокое
обучение)



AI, ML, DL

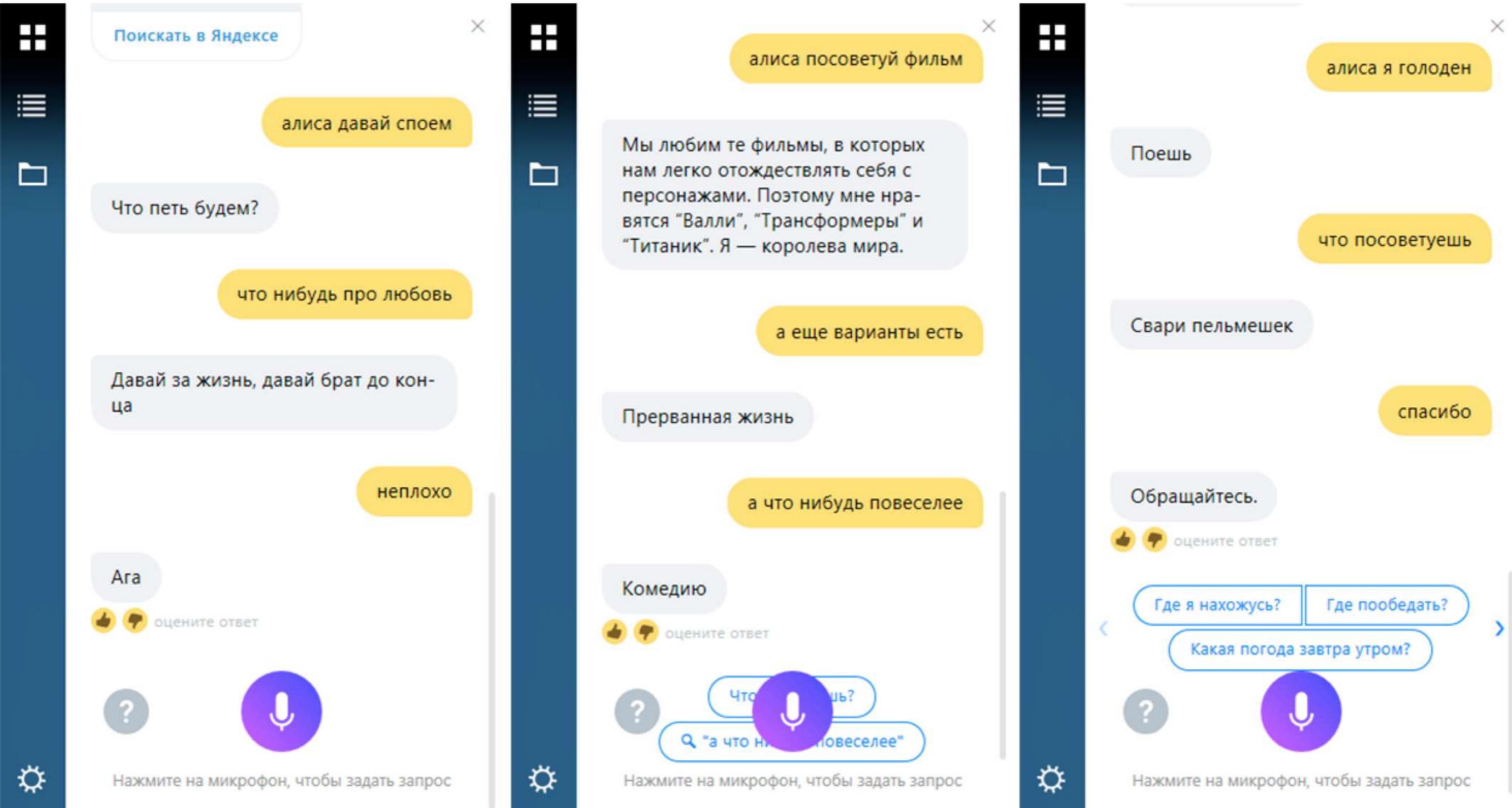
- Искусственный интеллект – широкая область, в которой изучают процесс принятия решений.
- Машинное обучение – подобласть искусственного интеллекта, в которой на основании данных машины учатся принимать решения без прямого, явного программирования по сценариям.
- Глубокое обучение – подобласть машинного обучения, сфокусированная на нейронных сетях.

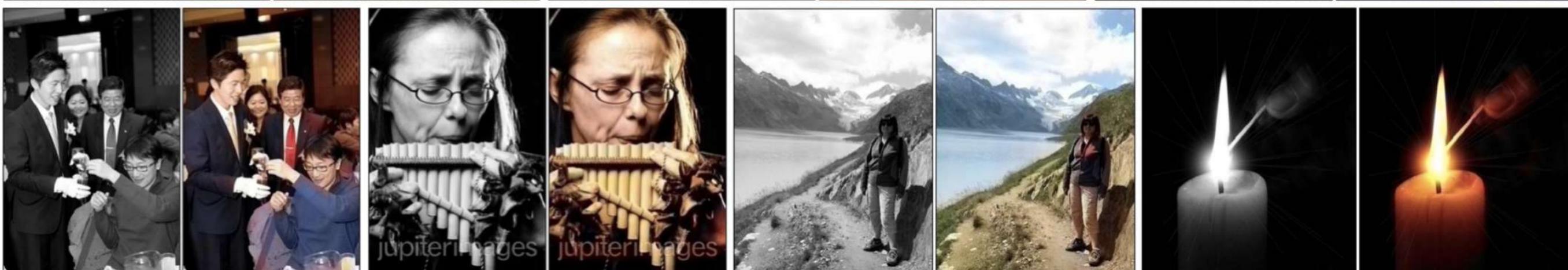
AI, ML, DL

- Может решить задачу:
 - На **входе** объект наблюдения
 - На **выходе** ответ
 - Пример: человек → давать кредит?
- Общий ИИ – решает любую задачу
- Специализированный ИИ – решает узкую задачу









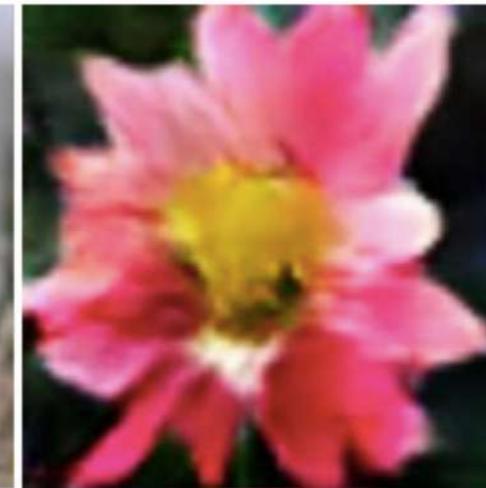
This bird is white with some black on its head and wings, and has a long orange beak



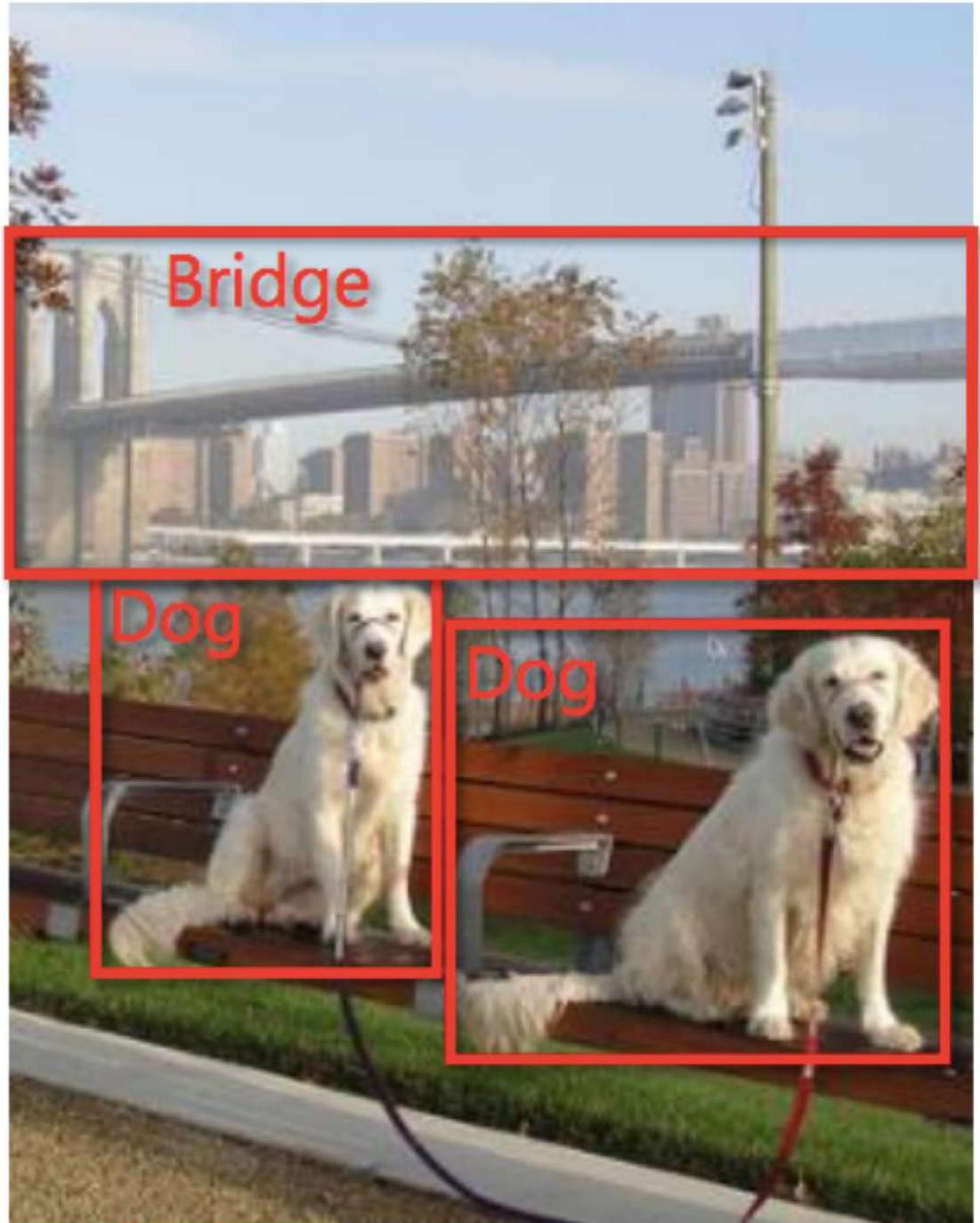
This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face



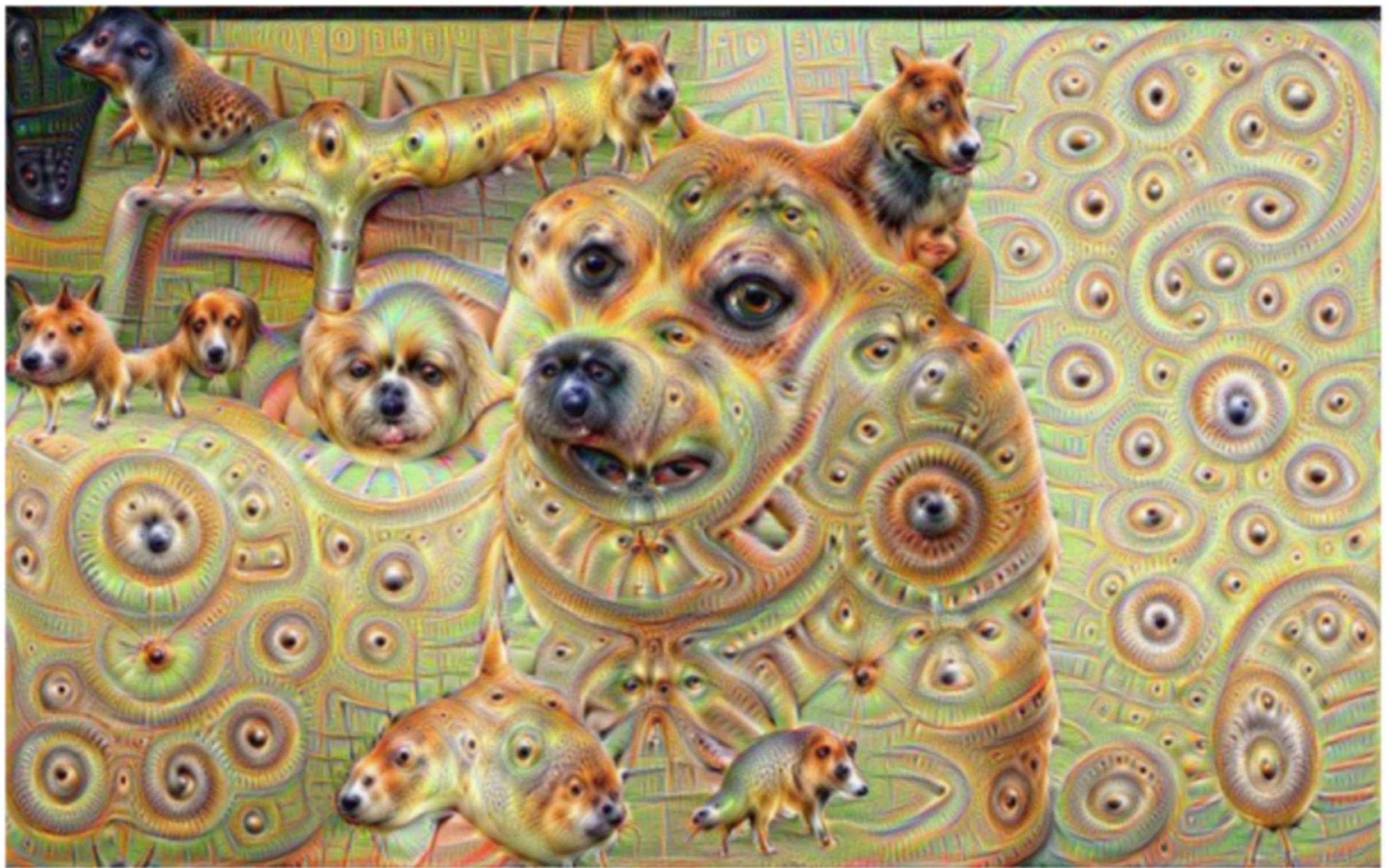
This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments



✓ Dog
✓ Bridge







Основные понятия в ML

ПРИМЕР ЗАДАЧИ

- Для улучшения эффективности диспетчерских служб такси важно знать, когда водитель закончит один заказ и будет готов принять следующий.
- Оценка длительности текущей поездки – один из факторов эффективного распределения заказов.
- Как оценить длительность поездки?

ТЕРМИНОЛОГИЯ

- x (**sample**) – объект, для которой хотим делать предсказания
 - Поездки
 - y (**target**) – ответ, целевая переменная, т.е. То, что хотим предсказать
 - Длительность поездки
-
- $(x_i, y_i)_{i=1}^{\ell}$ – обучающая выборка, прецеденты, т.е. все объекты, для которых известны значения целевого признака
 - ℓ – размер выборки.

ПРИЗНАКИ

- Компьютер умеет работать с числовой информацией
- Объекты характеризуются числовой информацией – признаками, факторами, «фичами» (от англ. features)
- m – число признаков
- $x = (x^1, \dots, x^m)$

ПРИЗНАКИ ДЛЯ ЗАДАЧИ

ПРИЗНАКИ ДЛЯ ЗАДАЧИ

- Временные
- Географические
- Погодные
- Маршруты
- Пассажиры

ПРИЗНАКИ ДЛЯ ЗАДАЧИ

- Временные
 - Дата и время посадки
 - Дата и время высадки
- Географические
 - Ширина и долгота места посадки
 - Ширина и долгота места высадки
- Погодные
 - Осадки: дождь, снег, шторм
 - Сила осадков
- Маршруты
 - Наиболее быстрые маршруты
 - Скорость по маршрутам
- Пассажиры
 - Число пассажиров

ОБУЧЕНИЕ

- $a(x)$ – алгоритм/модель
- Это функция, предсказывающая ответ для любого объекта

ОБУЧЕНИЕ

- $a(x)$ – алгоритм/модель
- Это функция, предсказывающая ответ для любого объекта
- Алгоритм предсказал 100 минут, а поездка длилась 83 минут. Хорошее ли предсказание или плохое?

ОБУЧЕНИЕ

- $a(x)$ – алгоритм/модель
- Это функция, предсказывающая ответ для любого объекта
- Алгоритм предсказал 100 минут, а поездка длилась 83 минут. Хорошее ли предсказание или плохое?
- Функция потерь (ошибок) – мера корректности алгоритма
- Для нашей задачи можно использовать среднеквадратическую ошибку Mean Square Error:

$$MSE = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

ОБУЧЕНИЕ

- Функция потерь – один из важнейших компонентов при анализе данных и должна соответствовать бизнес требованиям.

ОБУЧЕНИЕ

- Функция потерь – один из важнейших компонентов при анализе данных и должна соответствовать бизнес требованиям.
- Есть прецеденты
- Определен функционал качества
- Есть параметризованное семейство алгоритмов:
«Если время после α часов, то длительность заказа сокращается на 10%»

ОБУЧЕНИЕ

- Функция потерь – один из важнейших компонентов при анализе данных и должна соответствовать бизнес требованиям.
- Есть прецеденты
- Определен функционал качества
- Есть параметризованное семейство алгоритмов:
«Если время после α часов, то длительность заказа сокращается на 10%»

Обучение – поиск оптимальных алгоритмов с точки зрения функционала качества.

ПРЕДСКАЗАНИЕ ЦЕНЫ НА ТОВАР

- Задача: товар → цена
- x_i – объект, для которого строим предсказания (i -ый товар)
- y_i – целевая переменная (цена на i -ый товар)
- (x_i, y_i) – прецедент
- Обучающая выборка – набор всех прецедентов

Как решить эту задачу?

Найти алгоритм $a(x)$: $a(x_i) \approx y_i$

ПРЕДСКАЗАНИЕ ЦЕНЫ НА ТОВАР

- Прецедент – запись в таблице

Номер товара	Стоймость производства одной единицы в рублях	Стоймость упаковки одной единицы товара в рублях	Срок производства одной единицы товара в днях	Вес товара в граммах	Цена в рублях
1	5000	300	14	327	13499

x_i

y_i

Номер товара	Стоймость производства одной единицы в рублях	Стоймость упаковки одной единицы товара в рублях	Срок производства одной единицы товара в днях	Вес товара в граммах	Цена в рублях
1	5000	300	14	327	13499
2	4312	500	12	588	8847
3	6438	270	10	1020	16485

Обучающая выборка

ПРЕДСКАЗАНИЕ ЦЕНЫ НА ТОВАР

- Алгоритм (модель) – это формула, учитывающая характеристики объекта
- Формулы могут быть любыми

w_1^* стоимость производства

w_1^* стоимость производства²

w_1^* стоимость производства + w_2^* вес товара

- Вид формулы задает класс алгоритмов
- В процессе обучения ищем оптимальные w_1, w_2

ПРЕДСКАЗАНИЕ ЦЕНЫ НА ТОВАР

Номер товара	Стоимость производства одной единицы в рублях (р)	Цена в рублях	Предсказание модели $a(x) = 2.4 * p$	Ошибка модели $(a(x_i) - y_i)^2$
1	5000	13499	12000	2247001
2	4312	8847	10348,8	2255403,24
3	6438	16485	15451,2	1068742,44

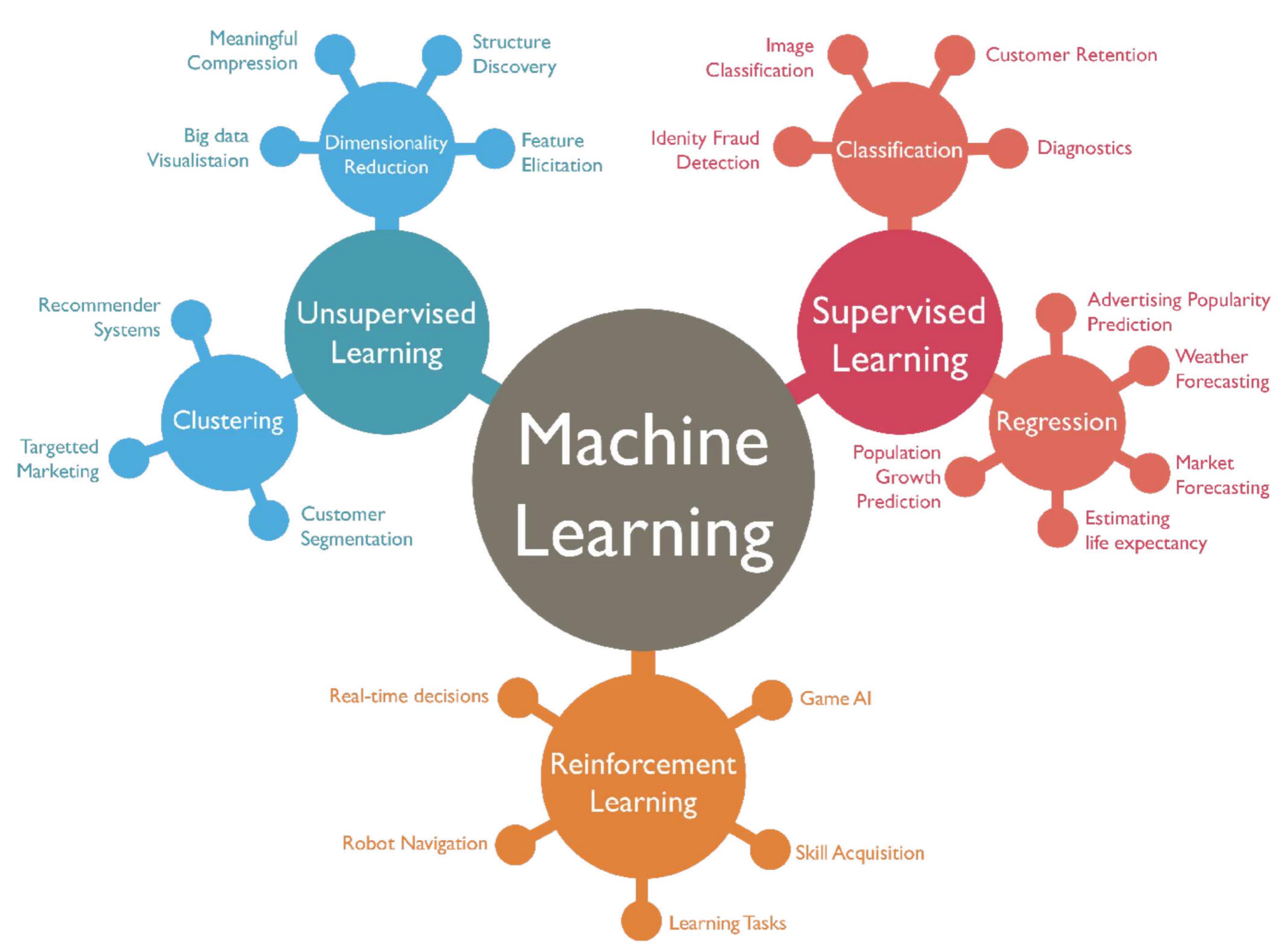
$$MSE = 1857048,89$$

$$RMSE = 1362,73$$

Номер товара	Стоимость производства одной единицы в рублях (р)	Цена в рублях	Предсказание модели $a(x) = 2.5 * p$	Ошибка модели $(a(x_i) - y_i)^2$
1	5000	13499	12500	998001
2	4312	8847	10780	3736489
3	6438	16485	16095	152100

$$MSE = 1628863,33$$

$$RMSE = 1276,27$$



ПОДХОДЫ К ОБУЧЕНИЮ

- Обучение с учителем
 - Классификация
 - Регрессия
 - Ранжирование
- Обучение без учителя
 - Кластеризация
 - Уменьшение размерности
- Обучение с частичным привлечением учителя
- Обучение с подкреплением

КЛАССЫ ЗАДАЧ

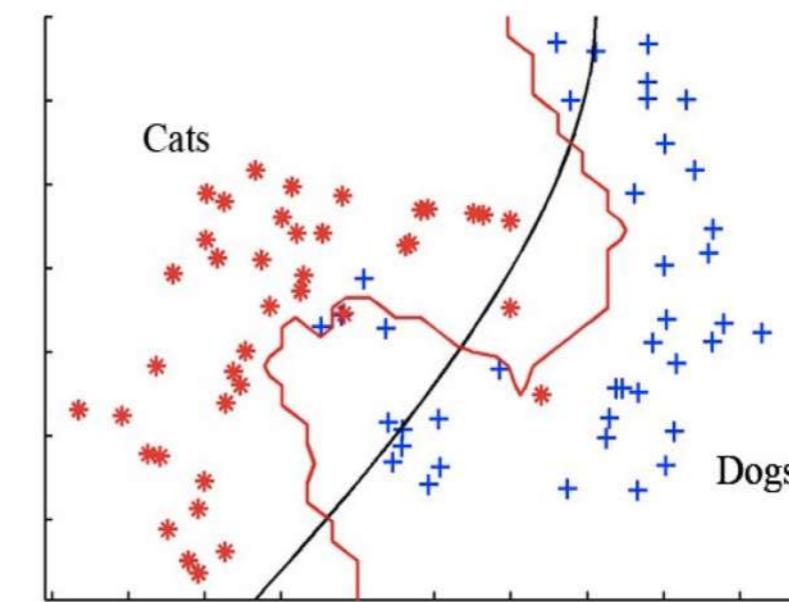
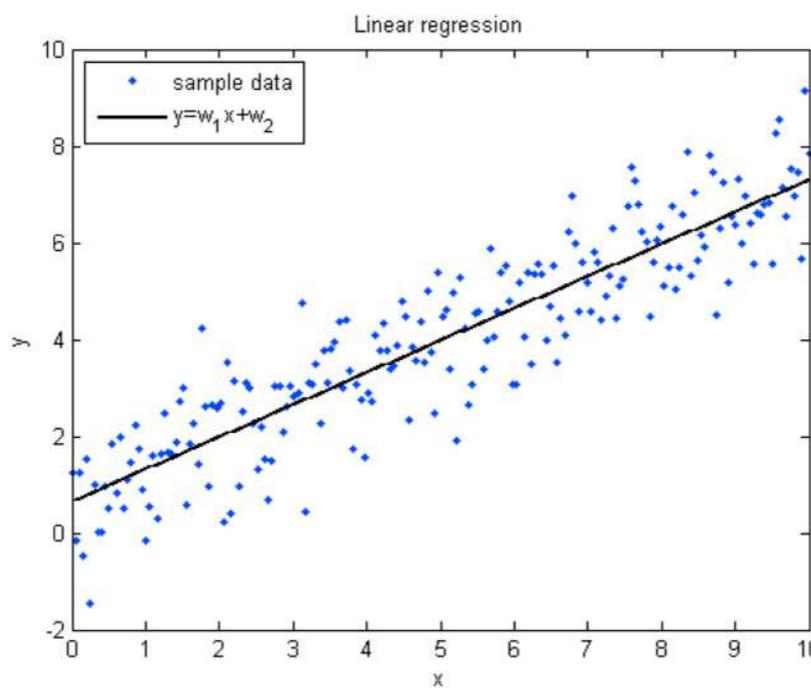
Регрессия

Классификация

Обучение с учителем

Вещественная
целевая переменная

Конечное множество
ответов

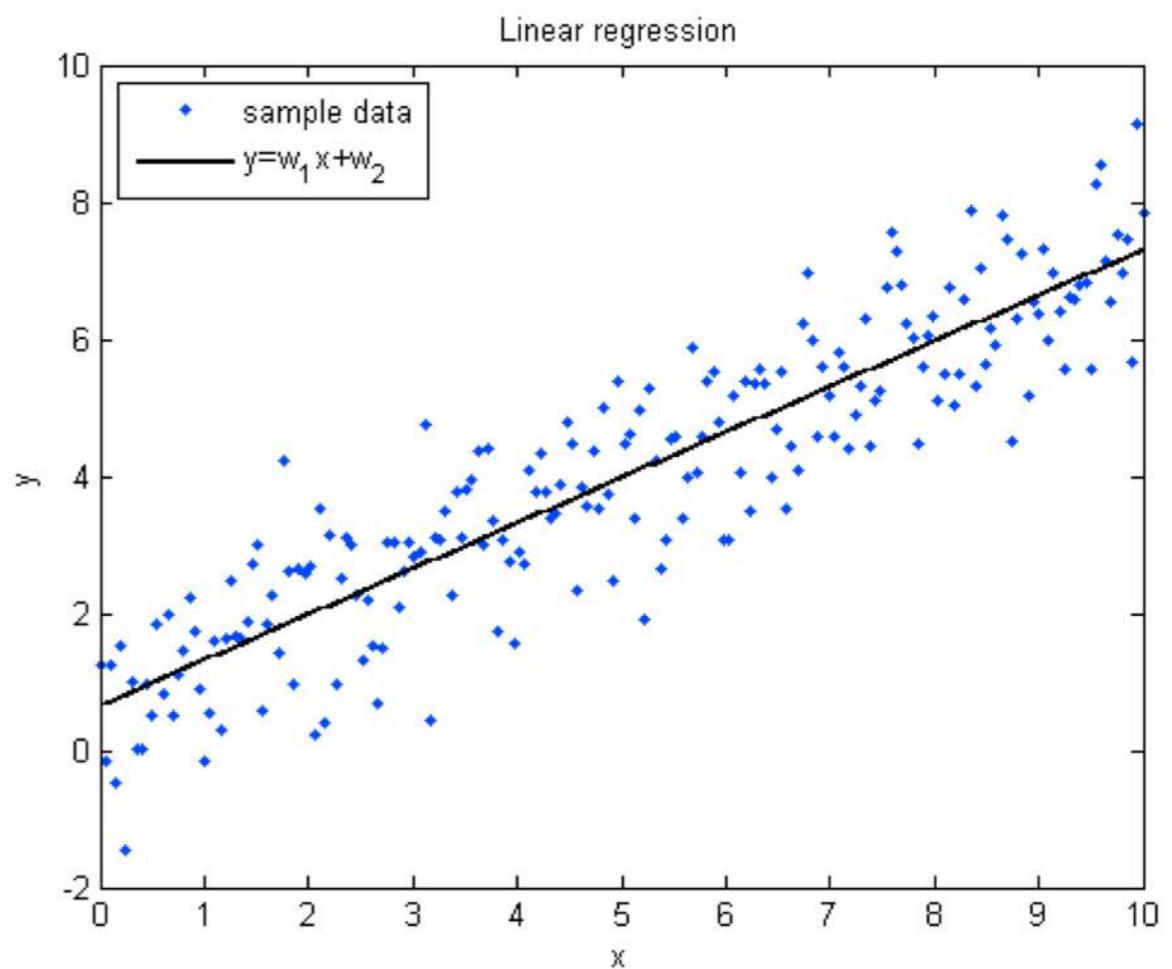


РЕГРЕССИЯ. ПРИМЕРЫ

- Прогнозирование цены дома
- Прогнозирование заработной платы по описанию вакансии
- Прогнозирование спроса на товар в ближайшую неделю
- Прогнозирование уровня экспрессии гена
- Прогнозирование температуры воздуха
- Прогнозирование суммы компенсаций по страховке
- Прогнозирование объема потребления электроэнергии

РЕГРЕССИЯ

- Есть обучающая выборка, в которой объекты представлены признаковым описанием и есть значение целевой переменной
- Целевое значение: любое действительное число
- Задача: найти алгоритм, который спрогнозирует для любого объекта его целевое значение



КЛАССИФИКАЦИЯ. ПРИМЕРЫ

- Предсказание пола для неизвестного пользователя
- Определение типа документа
- Определение языка документа
- Определение эмоционального окраса отзыва
- Вероятность ухода сотрудника/клиента
- Предсказание типов писем: спам/не спам
- Определение объектов на фотографии
- Оценка состояния человека по ЭЭГ

КЛАССИФИКАЦИЯ

- Есть обучающая выборка, в которой объекты представлены признаковым описанием и дана целевая переменная – метка класса.
 - Метод обучения с учителем – требуются прецеденты (размеченная выборка)
 - Задача: найти алгоритм, который каждому нового объекту будет присваивать метку класса.
-
- Классов может быть много.
 - Бинарная классификация – 2 класса
 - Многоклассовая классификация – 3 и более.



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ