



The Science of Data

Как перемножение матриц и первые производные
меняют нашу жизнь

Сергеев Дмитрий
Senior Data Scientist ŌURA

Кто сегодня рассказывает

- Руководитель курсов по машинному обучению OTUS
- Senior Data Scientist at ŌURA



NATIONAL RESEARCH
UNIVERSITY



Немного истории



1940-1950е

- Появление первых компьютеров
- Алан Тьюринг придумал **тест Тьюринга**
- Первая программа, которая могла **учиться** играть в шашки
- Франк Розенблатт создает модель Перцептрона – **первая искусственная нейросеть**



1960-1970е

- Разрабатывается всё больше моделей **машинного обучения**
- kNN, Decision Trees (CART)
- **Машинное обучение (ML)** и **Искусственный интеллект (AI)** начали разделяться как самостоятельные направления



1980-1990е

- **Стагнация ML**, люди сосредоточены на создании экспертных систем и систем, основанных на правилах
- Небольшой расцвет, а потом снова закат нейронных сетей
- **Kernel Trick** - новый прорыв для старой модели SVM
- **Ансамблирование моделей**, Random Forest, Boosting
- В самом конце тысячелетия, в 1998, появляется нейронная сеть, способная распознавать рукописные цифры (LeNet)



2000-2010е

- **Интернет!** Всё больше данных и больше вычислительных мощностей для их анализа
- **Ренессанс нейронных сетей:** обработка изображений, видео, звуков, текстов и т.д.
- Победа нейросети в игре Go (Google DeepMind 2014)
- **Генеративные модели:** DeepDream, DeepFakes, GPT-1/2/3/..., etc.
- **Облачные вычисления** - доступный ML в каждый дом



2020-2030е


Наша очередь творить историю ;)



Ссылки на крутые штуки

- <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>
- <https://thispersondoesnotexist.com>
- <https://app.inferkit.com/demo> (ex. talk to transformer)
- <https://www.nvidia.com/en-us/research/ai-playground/>





Распространенные заблуждения

“ЭйАй”, “БигДата”, и другие способы инфоцыганства



Mat Velloso @matvelloso · Nov 23, 2018

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

💬 206

↻ 9.3K

♥ 23.9K





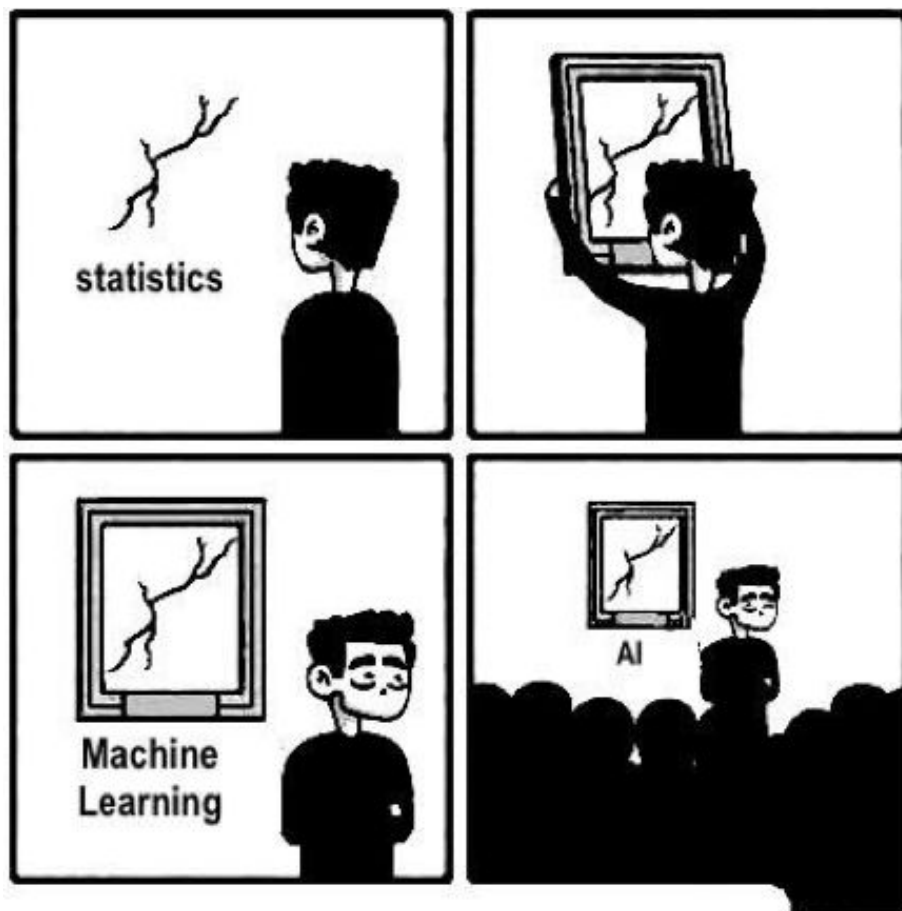
Basic Math

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix}$$

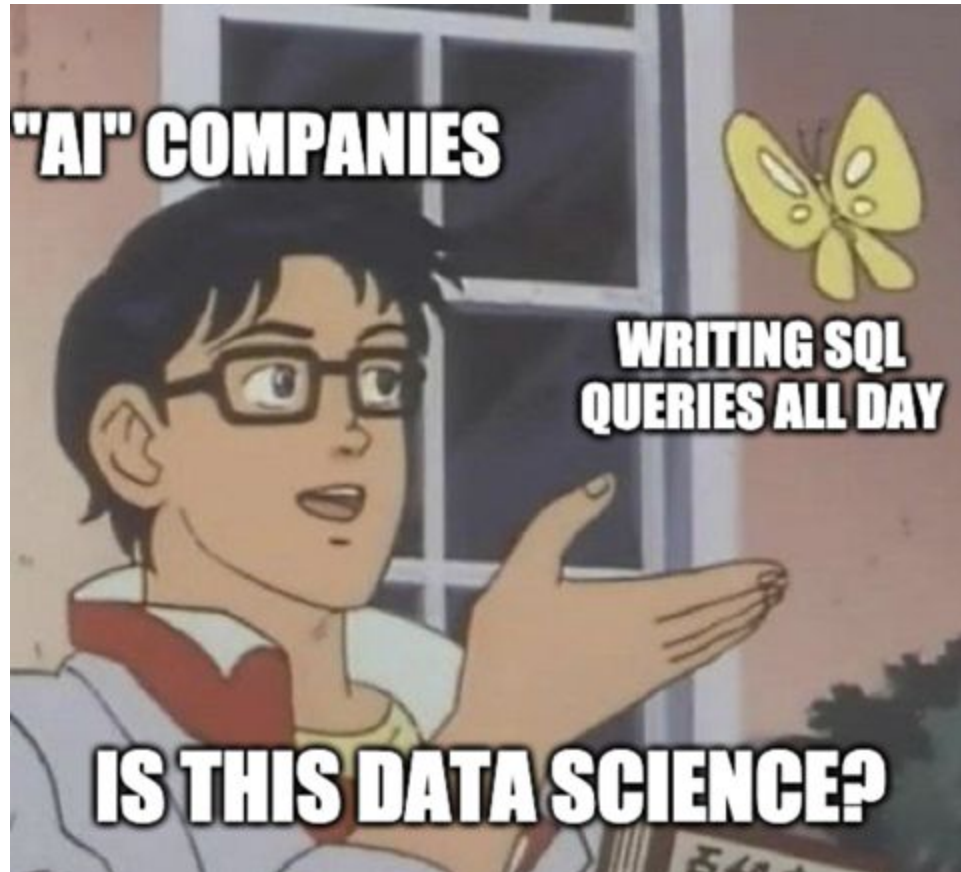
Dangerous Artificial Intelligence

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} * \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} * \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} * \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix}$$



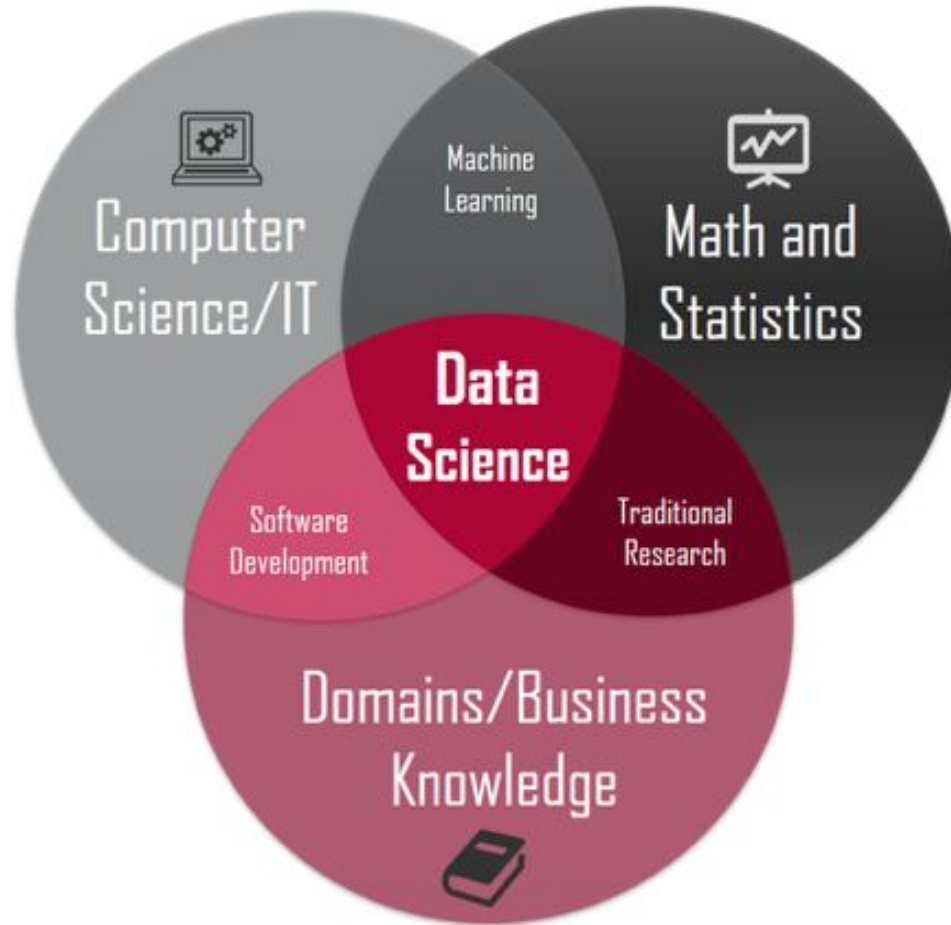








Что такое Data Science?



Our life is Data



Our life is Data

- Колоссальное количество данных генерируется каждый день
- <https://techjury.net/blog/how-much-data-is-created-every-day/>
- Цель науки о данных - **извлекать знания из данных**

Какие данные мы умеем генерировать и какие знания можем извлекать?






Мы читаем

- Обработка естественного языка (NLP)
Ступень 2, Модуль 2: Анализ текстовых данных



Мы читаем

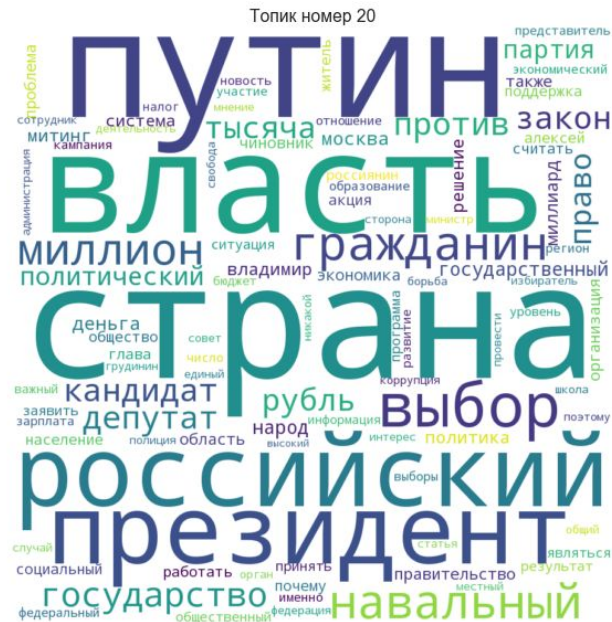
- Анализ тональности:
 - Как люди отзываются о моем продукте?
 - В каком отеле хорошие отзывы?
 - Находится ли этот человек в депрессии?

 <p>My experience so far has been fantastic!</p> <p>POSITIVE</p>	 <p>The product is ok I guess</p> <p>NEUTRAL</p>	 <p>Your support team is useless</p> <p>NEGATIVE</p>
---	--	---



Мы читаем

- Тематическое моделирование:
 - О чем пишут пользователи?
 - Что популярно в соцсетях?
 - О чем эта книга?
- Парочка моих проектов:
 - О чем говорят депутаты Госдумы
 - О чем говорят любители рэпа



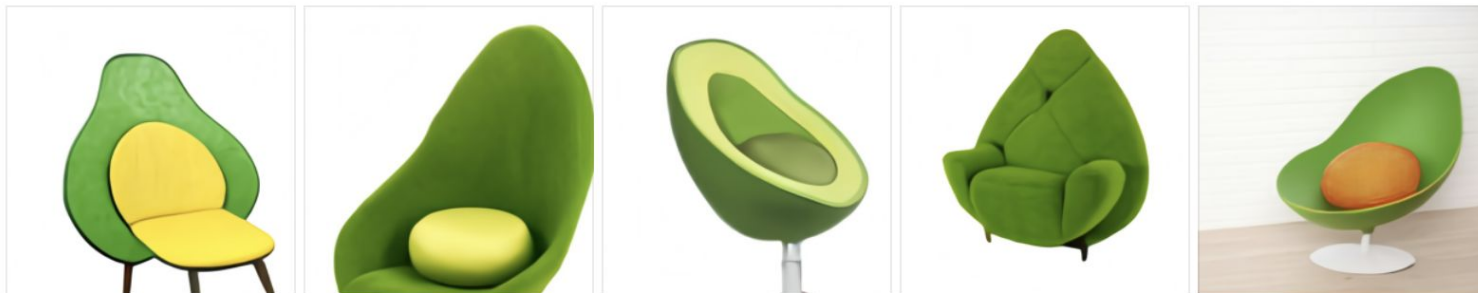
Мы читаем

- Генерация текстов, text-to-image:
 - <https://app.inferkit.com/demo>
 - <https://openai.com/blog/dall-e/>
 - [GPT-3 от Сбера](#)

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



Мы объединяемся

- Анализ социальных сетей (SNA), Анализ графов (Ступень 2, Модуль 1)
 - Как находить сообщества в соцсетях?
 - Как распространяются болезни?
 - Как оптимизировать транспортные маршруты в городе?
 - Как нейроны общаются в мозге?



Modelled COVID-19 spread



Мы рекомендуем

- Рекомендательные системы (Ступень 2, Модуль 4)
 - С этим товаром также покупают...
 - Рекомендация музыки, фильмов, книг, и т.д.

YouTube recommendations : here's a 45 min documentary on the fall of the Soviet Union

me, trying to fall asleep at 2:38am :



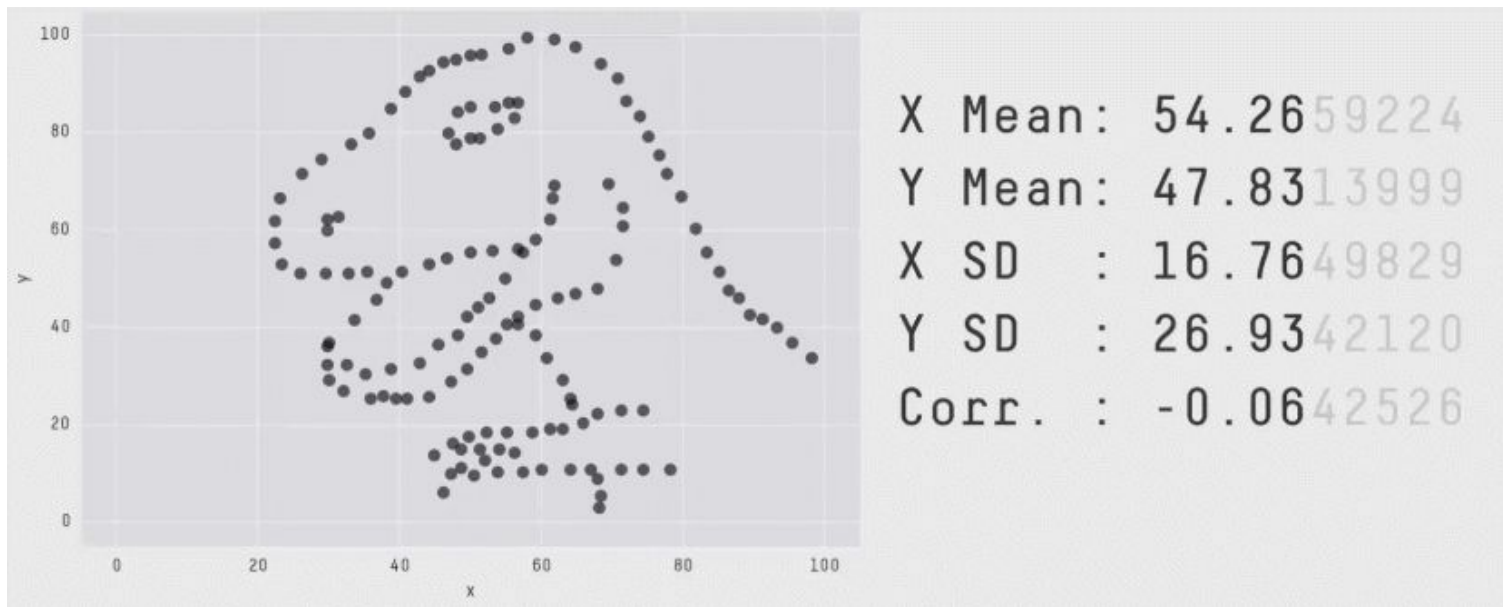
no one:

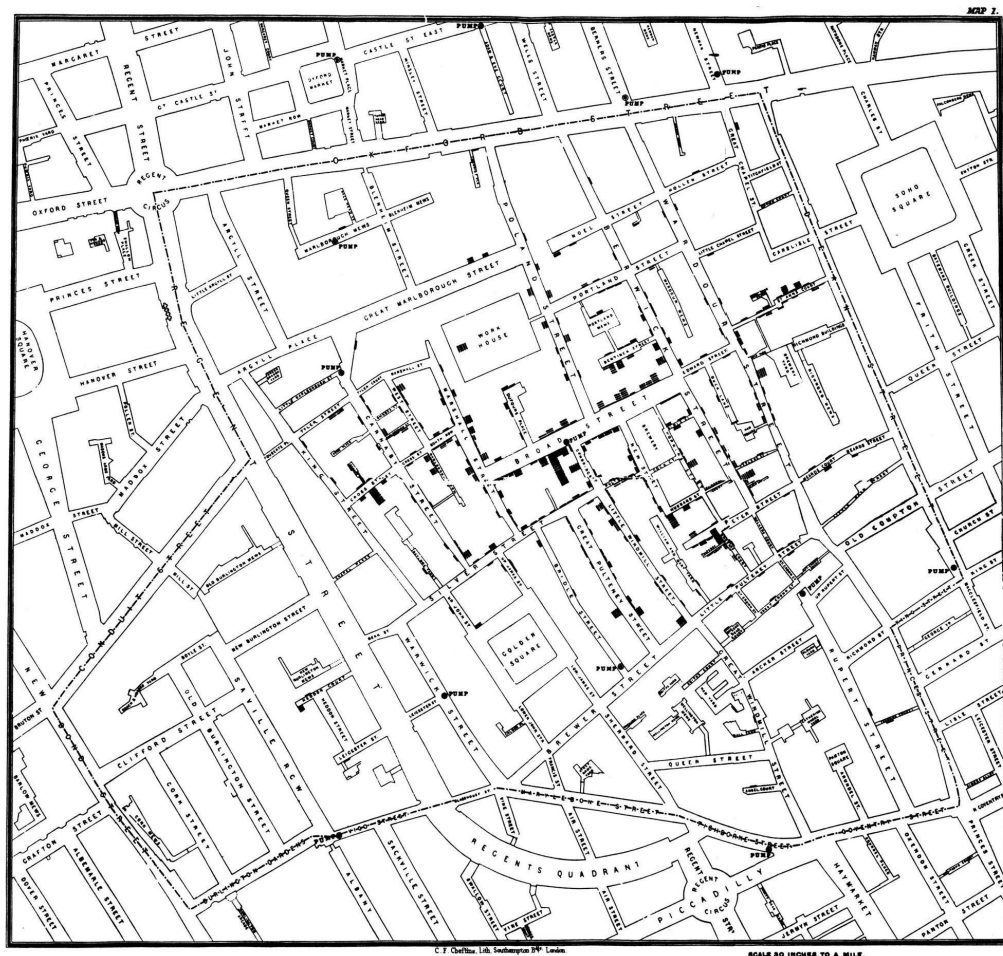
youtube
recommendations:



Мы видим

- Визуализация данных (Степень 1, Модуль 3)

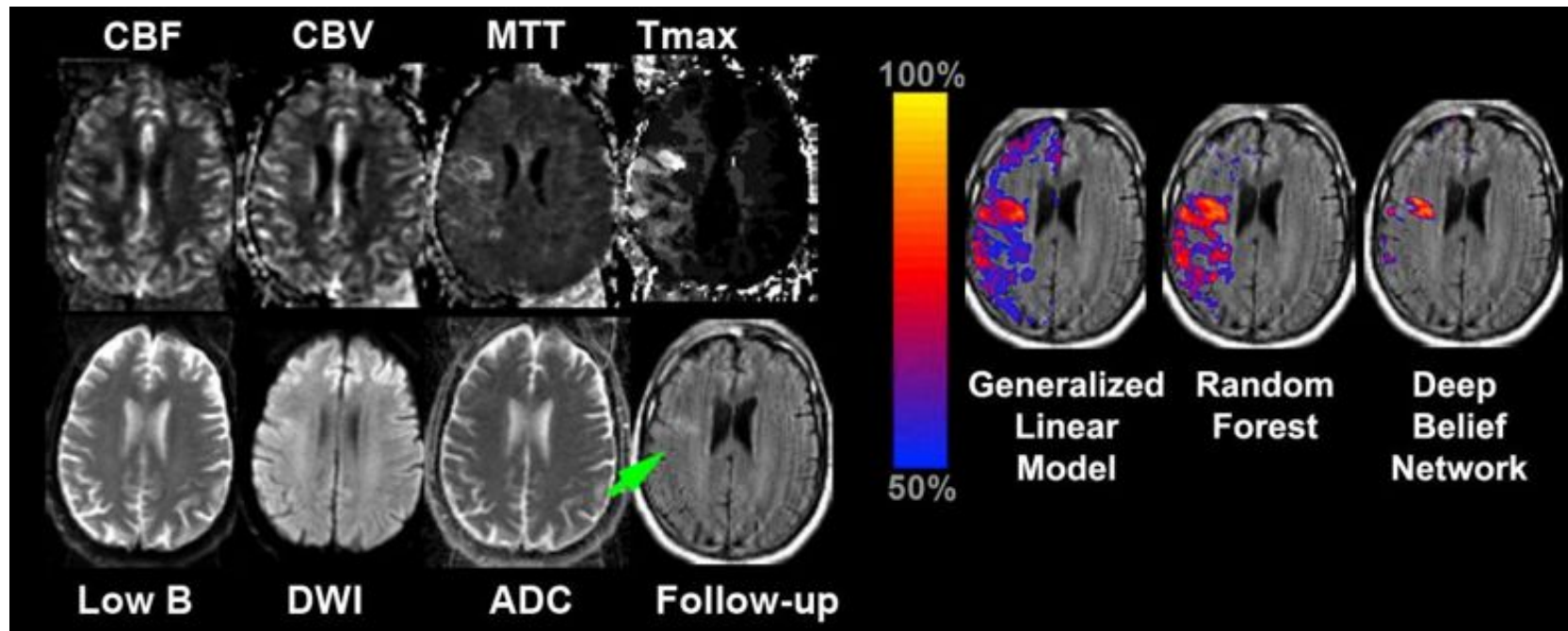




Мы видим

- Обработка изображений и видео, [компьютерное зрение \(CV\)](#)
- Генерация изображений и видео:
 - <https://thispersondoesnotexist.com/>
 - [Two Minute Papers on DeepFakes](#)

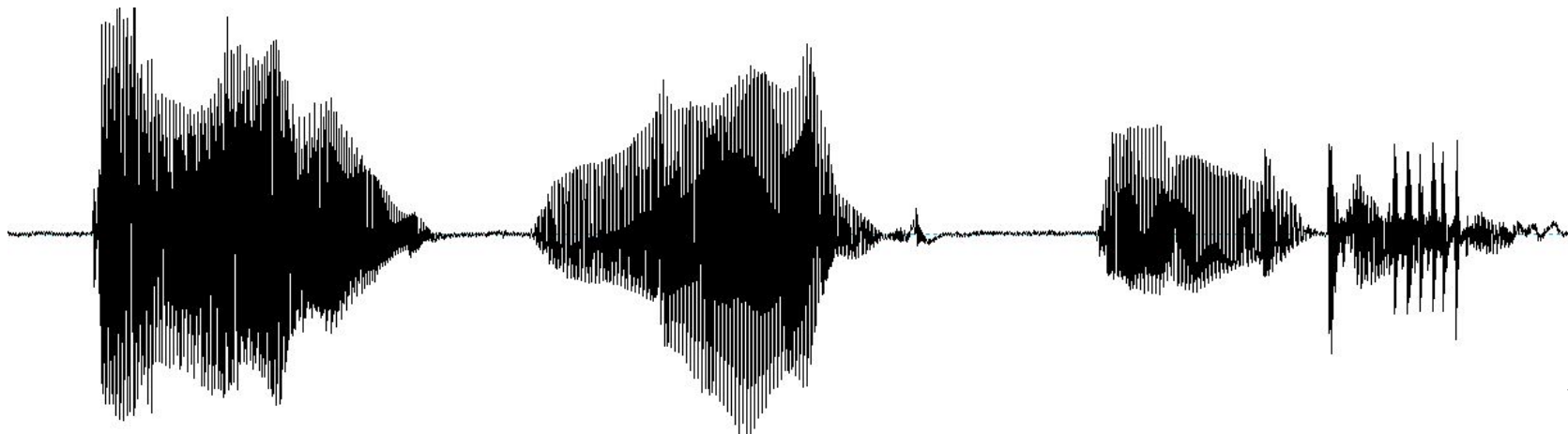






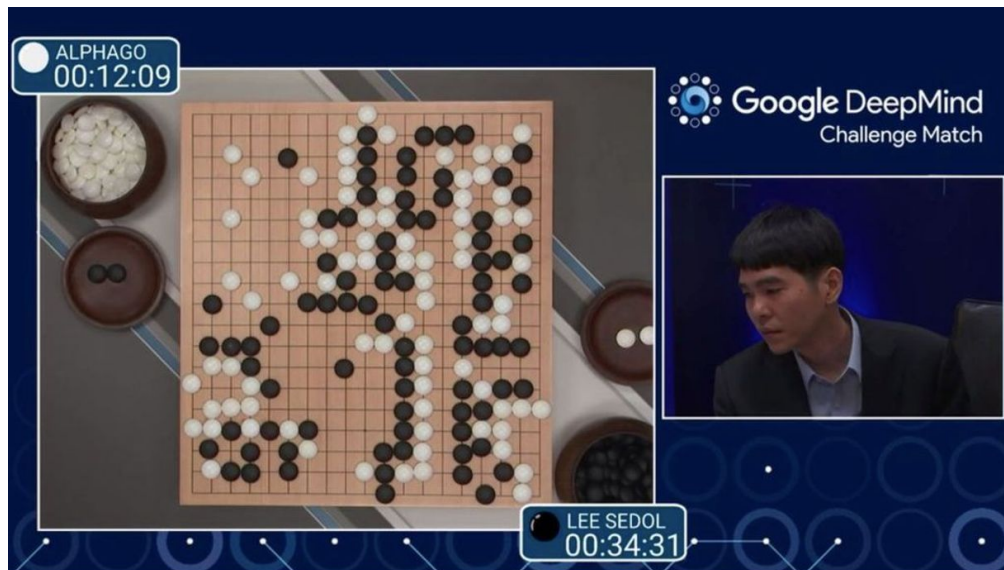
Мы говорим

- Аудио анализ, анализ сигналов, генерация речи
- Голосовые помощники - “Ok Google!”, “Open the pod bay doors Hal!”
- Анализ музыки - Spotify



Мы играем :D

- Обучение с подкреплением (RL)
[Machine Learning Advanced, Модуль 7](#)

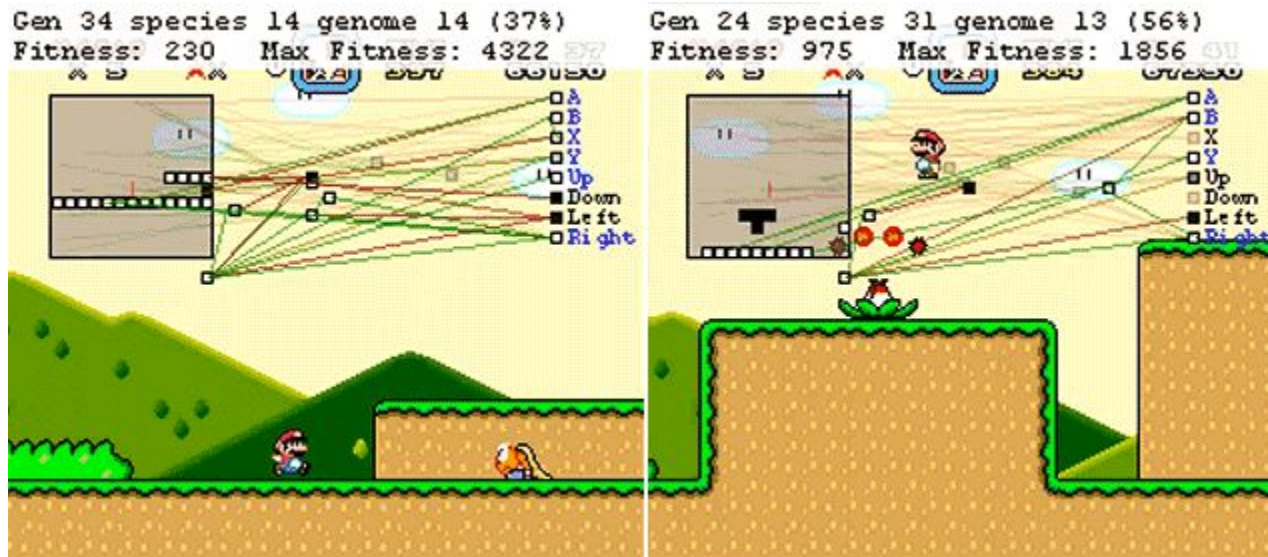


<https://deepmind.com/research/case-studies/alphago-the-story-so-far>



Мы играем :D

- Marl/O



Мы играем :D



Мы живём

- Анализ здоровья, превентивная медицина, мониторинг
 - Физическая активность
 - Сон
 - Питание
 - Психологическое состояние
 - И многое другое
- [Wearables in Medicine](#)





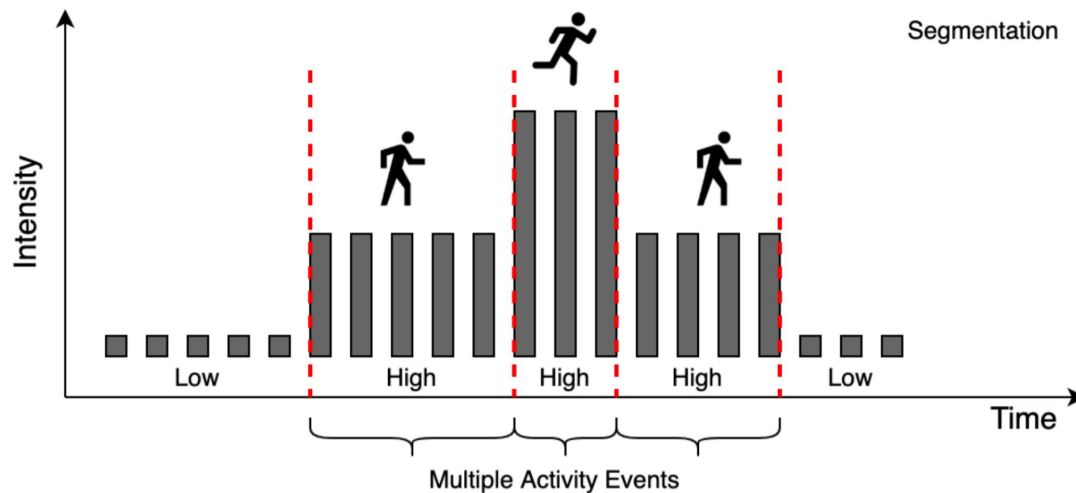




Мы живём

- Распознавание физической активности человека (HAR)

[Machine Learning Advanced, Модуль 3](#)



Почему это круто?



Плюсы/минусы?

- Пересечение разных дисциплин



Плюсы/минусы?

- Пересечение разных дисциплин
- Высокий спрос (и относительно низкая конкуренция)



Плюсы/минусы?

- Пересечение разных дисциплин
- Высокий спрос (и относительно низкая конкуренция)
- Разнообразие задач



Плюсы/минусы?

- Пересечение разных дисциплин
- Высокий спрос (и относительно низкая конкуренция)
- Разнообразие задач
- Развивающаяся индустрия



Плюсы/минусы?

- Пересечение разных дисциплин
- Высокий спрос (и относительно низкая конкуренция)
- Разнообразие задач
- Развивающаяся индустрия
- Самообразование



Плюсы/минусы?

- Пересечение разных дисциплин
- Высокий спрос (и относительно низкая конкуренция)
- Разнообразие задач
- Развивающаяся индустрия
- Самообразование
- State-of-the-art появляются каждый месяц



Плюсы/минусы?

- Пересечение разных дисциплин
- Высокий спрос (и относительно низкая конкуренция)
- Разнообразие задач
- Развивающаяся индустрия
- Самообразование
- State-of-the-art появляются каждый месяц
- Эффективность?





Этика и безопасность

Когда ИИ захватит мир?

Этика и безопасность

- Нет “плохого ИИ”, но есть не очень грамотные люди



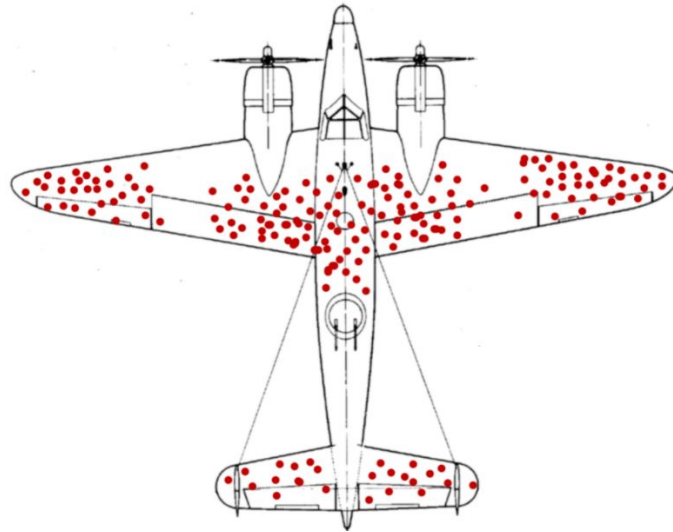
Этика и безопасность

- Data Privacy:
 - Как обеспечить безопасность данных?
 - Данные могут украсть, продать, использовать против вас
 - [General Data Protection Regulation \(GDPR\)](#)



Этика и безопасность

- Data Biases:
 - Насколько наши модели “честные”?
 - Что если смещение заложено в наших данных?
 - [Invisible Women: Exposing Data Bias In A World Designed For Men](#)



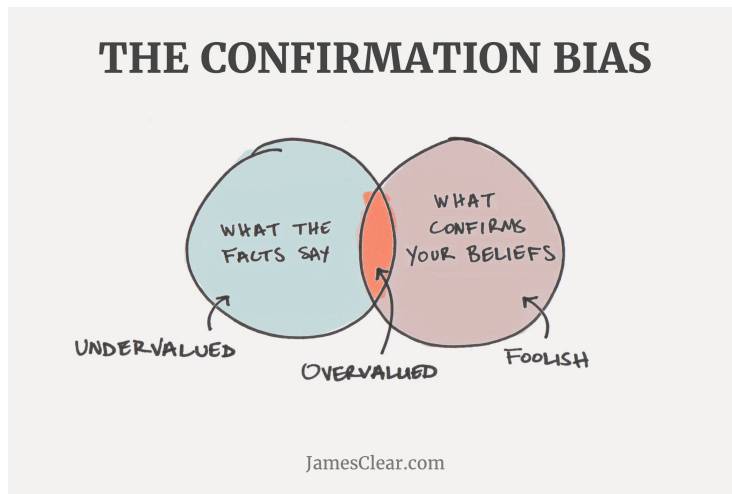
Этика и безопасность

- Fake news/videos/people:
 - Не верь глазам своим
 - <https://www.poynter.org/ifcn/anti-misinformation-actions/>



Этика и безопасность

- Информационные пузыри:
 - Как нам увидеть цельную картину?
 - Рекомендации показывают мне только то, что мне понравится
 - <https://fs.blog/2017/07/filter-bubbles/>



Этика и безопасность

- К счастью, мы знаем об этих проблемах и можем искать решения



Этика и безопасность

- К счастью, мы знаем об этих проблемах и можем искать решения
- Но пока что стоит соблюдать цифровую гигиену



Этика и безопасность

- К счастью, мы знаем об этих проблемах и можем искать решения
- Но пока что стоит соблюдать цифровую гигиену



The unexamined life is not worth living

Сократ (хотя никто точно не знает, так что fake news)



Мои контакты

- Mail: dmitry.sergeev@ouraring.com
- Telegram: @dmitryserg
- GitHub: <https://github.com/dmitryserg>
- LinkedIn: <https://www.linkedin.com/in/sergeyevdmitry/>



