

Машинное обучение

Лекция 1

Основные понятия, термины, подходы, инструменты

Власов Кирилл Вячеславович



2018

Коммуникация



bit.ly/open_ml_dafe_chat



Содержание курса

- Постановка задачи и их типы. Основные инструменты. Подходы и методы
- Задача классификации, простые методы классификации.
- Деревья решений для задач классификации и регрессии.
- Линейные модели классификации и регрессии.
- Композиции алгоритмов: бэггинг, случайный лес, бустинг.
- Нейронные сети
- Обучение без учителя



Что такое машинное обучение?

Давайте решим задачу

Мальчик на санках едет с горки. Масса мальчика вместе с санками составляет 40 кг, угол наклона горы 30° . Найдите ускорение, которым съезжает мальчик, если коэффициент трения скольжения равен 0,2.

Давайте решим задачу

Мальчик на санках едет с горки. Масса мальчика вместе с санками составляет 40 кг, угол наклона горы 30° . Найдите ускорение, с которым съезжает мальчик, если коэффициент трения скольжения равен 0,2.

Дано:

$$m = 40 \text{ кг}$$

$$\alpha = 30^\circ$$

$$\mu = 0,2$$

$$a - ?$$

$$m\vec{a} = \vec{N} + m\vec{g} + \vec{F}_{\text{тр}}$$

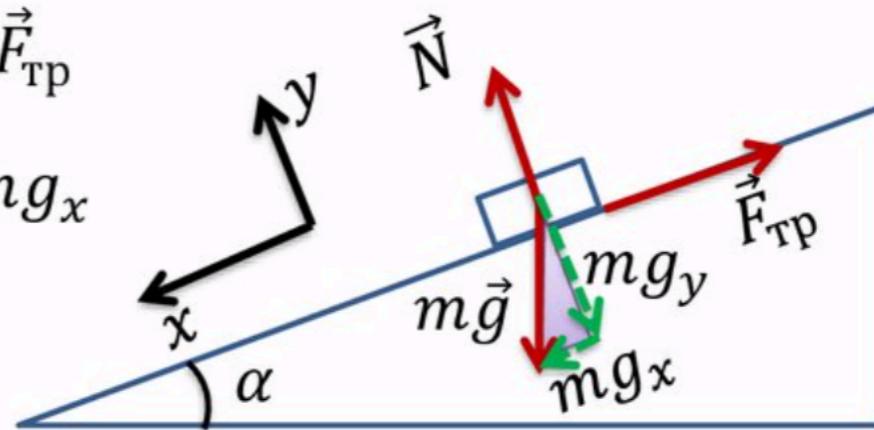
$$X: ma = -F_{\text{тр}} + mg_x$$

$$Y: 0 = N - mg_y$$

$$N = mg_y$$

$$F_{\text{тр}} = \mu N = \mu mg_y$$

$$ma = mg_x - \mu mg_y$$

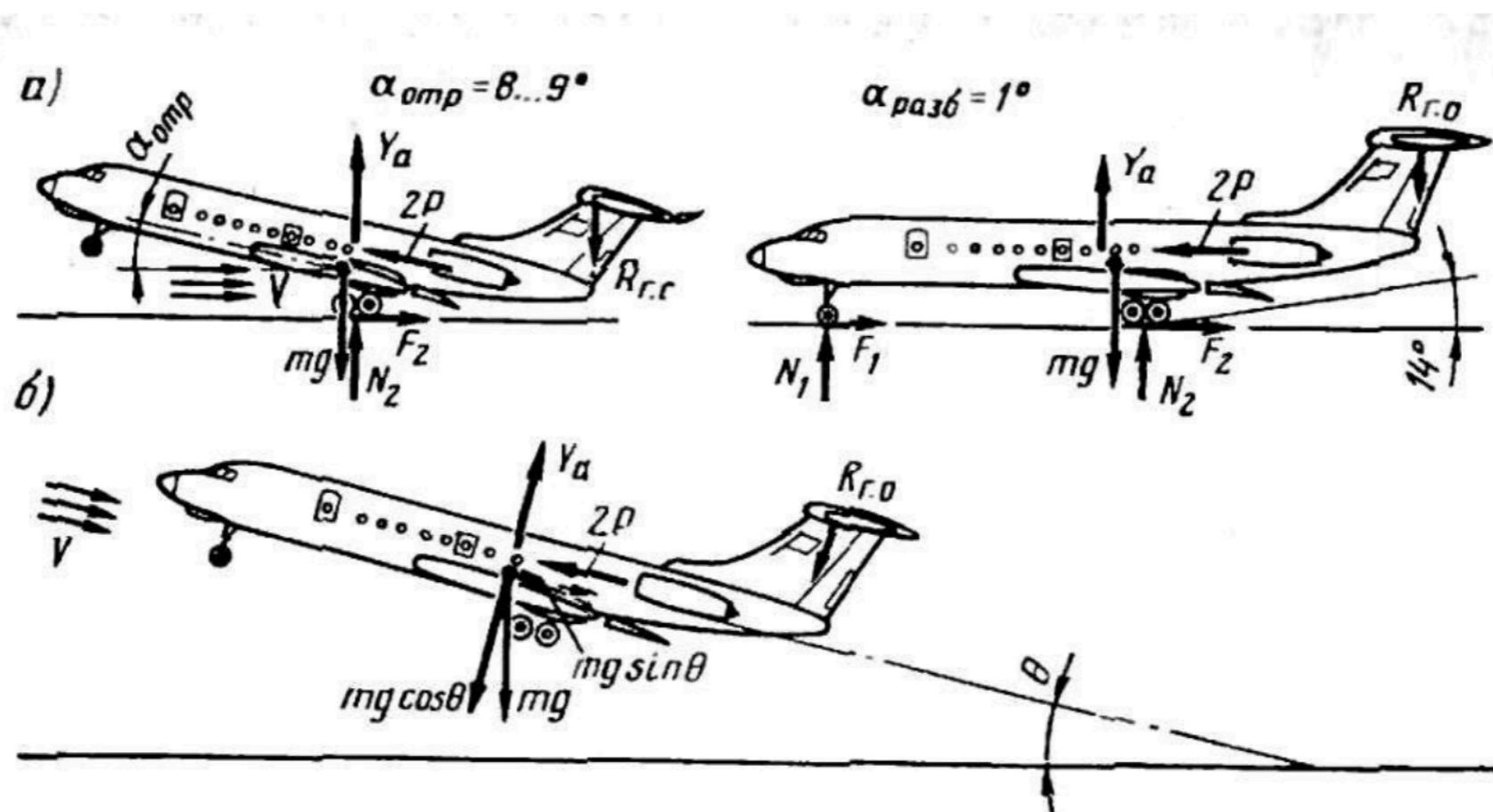


$$mg_x = mg \sin \alpha$$

$$mg_y = mg \cos \alpha$$

А что если система сложнее?

А что если система сложнее?



Давайте решим задачу

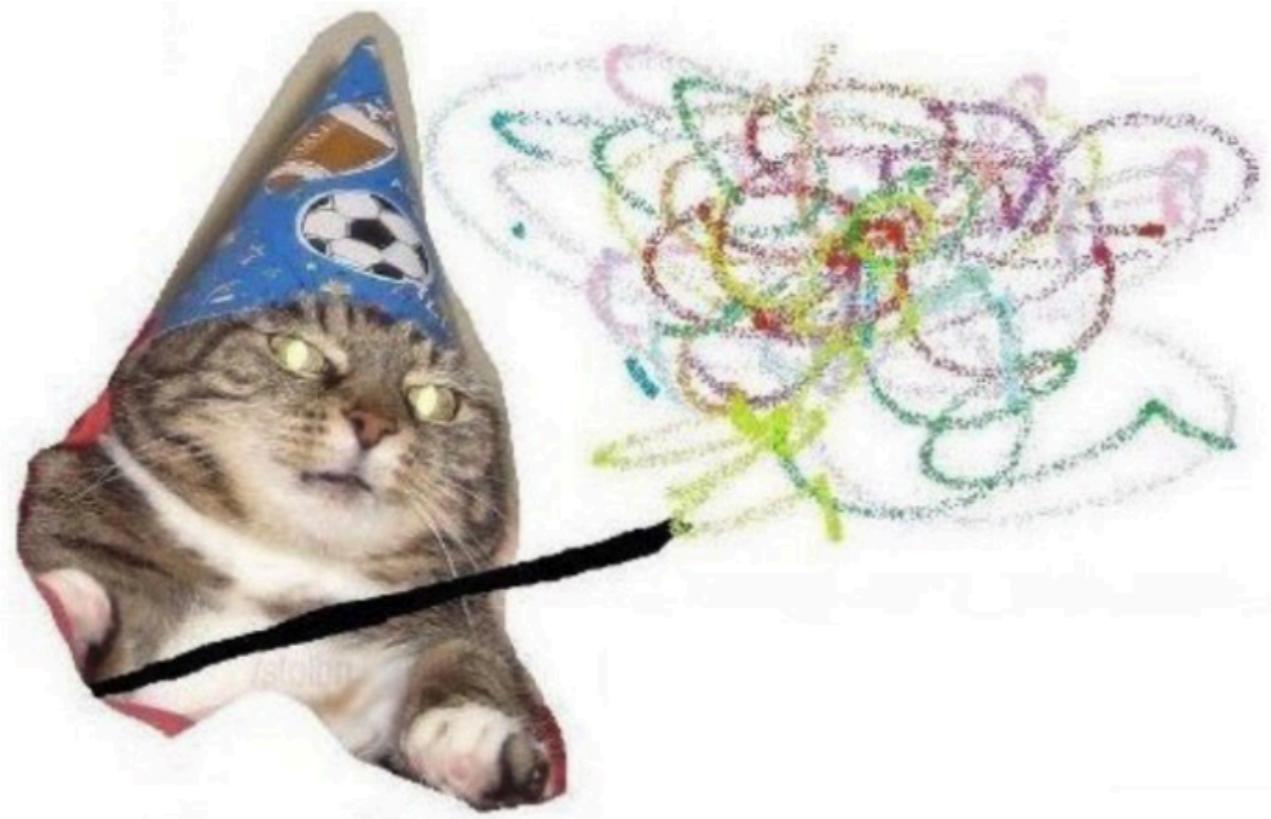
Дано:

- Сложная система
- 1000 или даже 5000 параметров
- Мы не имеем представления, как эти параметры влияют на наше решение

Давайте решим задачу

Дано:

- Сложная система
- 1000 или даже 5000 параметров
- Мы не имеем представления, как эти параметры влияют на наше решение



Можем ли мы сделать так?

- провести миллион экспериментов
- показать результаты компьютеру
- Компьютер сам найдет все связи и зависимости
- а мы Будем получать ответы на любых новых данных

Data Scientist: The Sexiest Job of the 21st Century

© 2012 – Harvard Business Review

Причина шумихи и ажиотажа

1.

Быстрее, выше, сильнее!

Вычислительные
мощности компьютеров
стали в разы больше и они
только растут

2.

Инфраструктура

Появилась много решений
и фреймворков для
анализа данных.
Большинство open-source

3.

Больше данных!

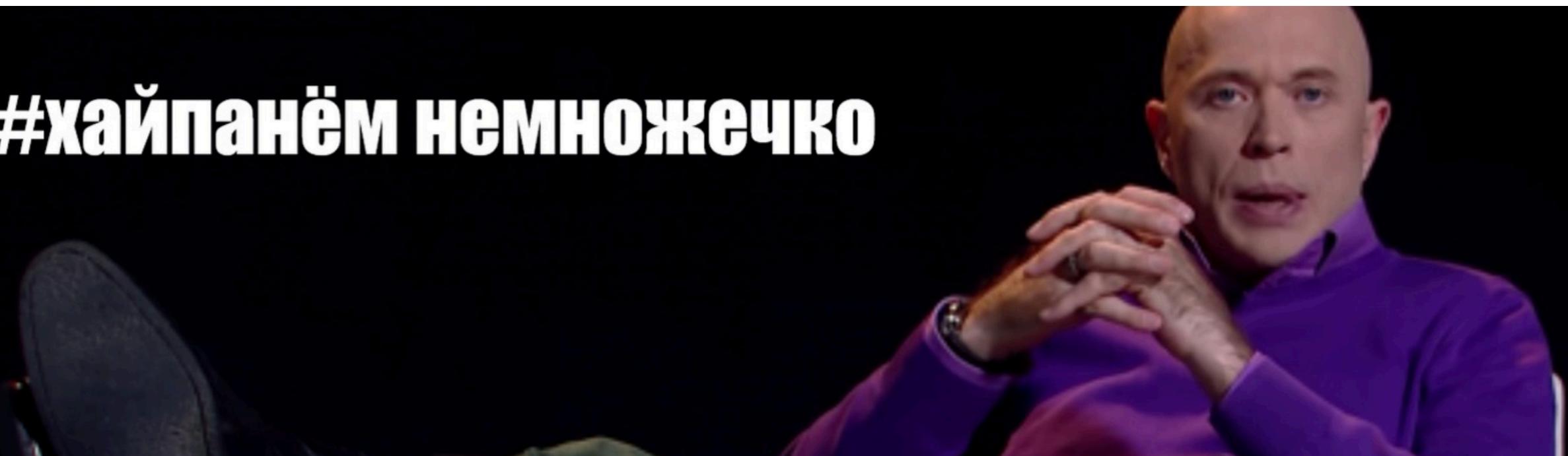
Компании собирают
БигДату и пытаются
извлекать из нее пользу

4.

Превосходство ИИ!

В некоторых задачах ИИ
справляется с решением
лучше человека

#хайпанём немножечко



Откуда столько данных?



On average, each cow generates about 200 megabytes of information a year.

© 2010 – *The Economist*,
Augmented business

Причина шумихи и ажиотажа

1.

Быстрее, выше, сильнее!

Вычислительные
мощности компьютеров
стали в разы больше и они
только растут

2.

Инфраструктура

Появилась много решений
и фреймворков для
анализа данных.
Большинство open-source

3.

Больше данных!

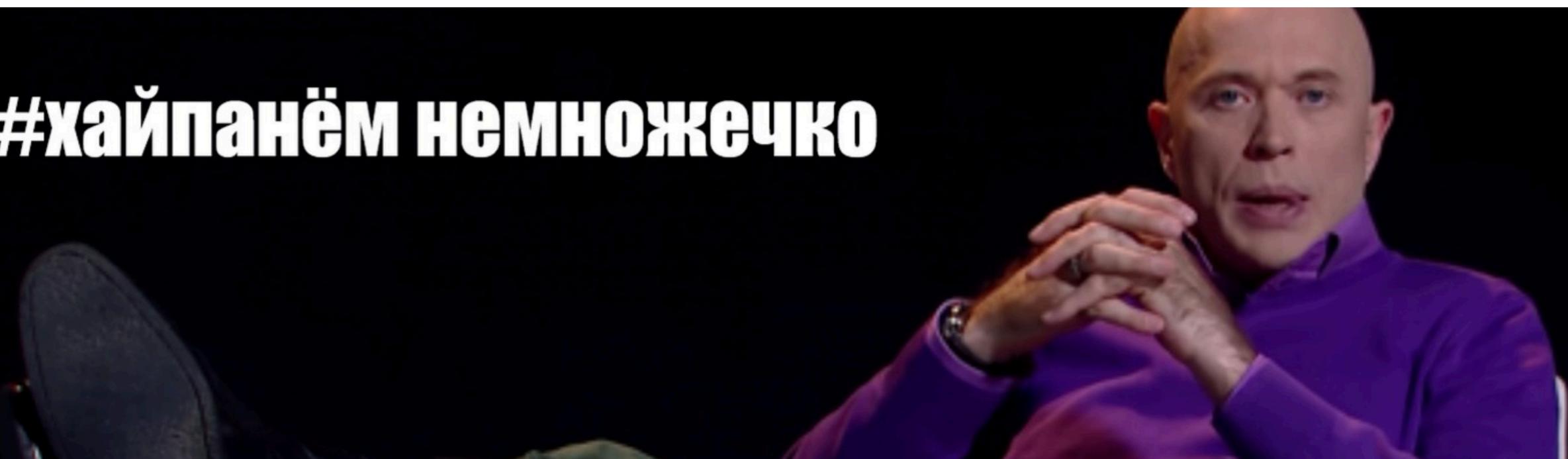
Компании собирают
БигДату и пытаются
извлекать из нее пользу

4.

Превосходство ИИ!

В некоторых задачах ИИ
справляется с решением
лучше человека

#хайпанём немножечко



Примеры в авиации

ИИ победил человека в симуляции воздушного боя

Подробнее: <https://habr.com/post/395525/>

Оценка действий лётчика на этапе посадки

Подробнее: <https://habr.com/post/281455/>

Dubai airport: Оптимизация выходов на посадку и полетов

Rolls-Royce: Диагностика самолетов

Подробнее: <http://azure.rbc.ru/business-practices/kak-sekonomit-na-diagnostike-samoletov/>



Искусственный интеллект лучше человека?



1997

Deep Blue выиграл чемпиона мира по шахматам Гарри Каспарова.



2016

AlphaGo выиграла матч у профессионала Ли Седоль

Искусственный интеллект лучше человека?



2019

*AlphaStar от DeepMind
Выиграл 10:1 команду
Team Liquid.*

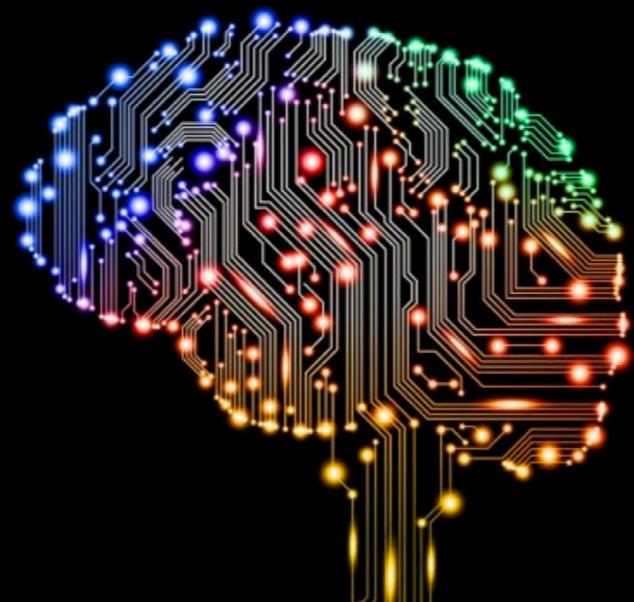
Искусственный интеллект лучше человека?

И.И.

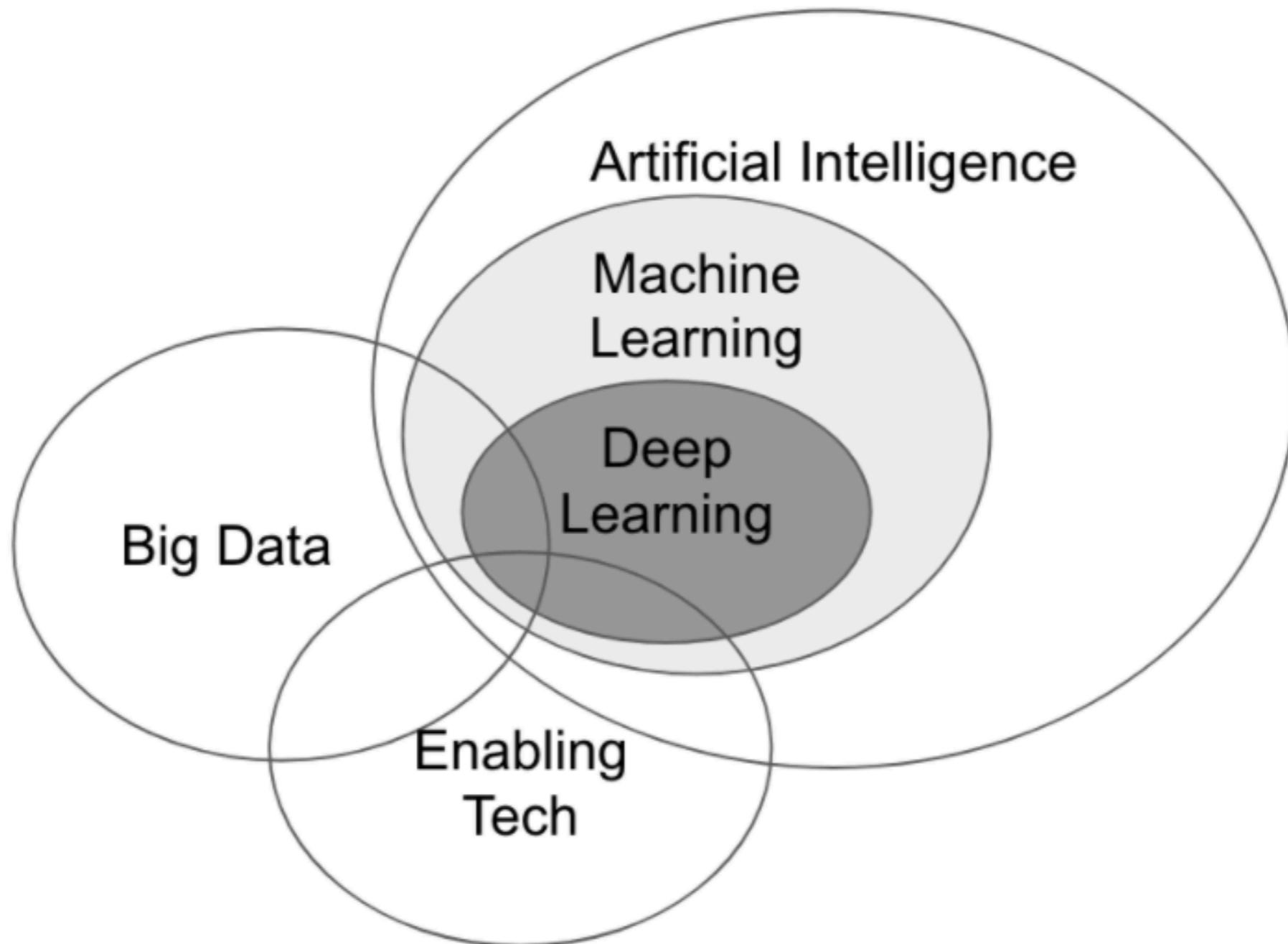


Искусственный интеллект — способность интеллектуальных машин выполнять творческие функции, которые традиционно считаются прерогативой человека. Также этим термином обозначают науку и технологию создания интеллектуальных машин.

© 1956 – Джон Маккарти



Место машинного обучения в области ИИ

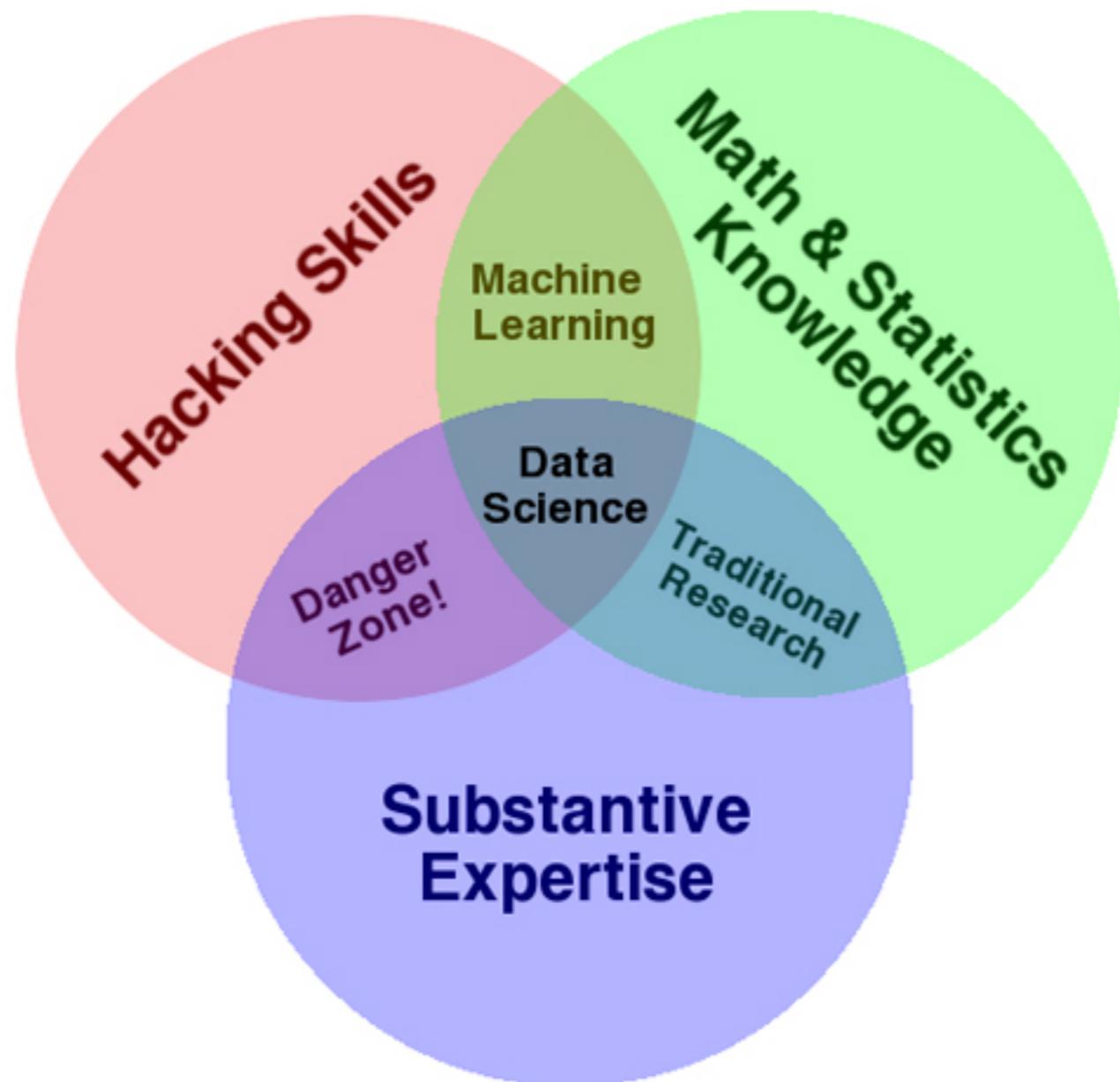


The background of the slide is a black and white aerial photograph of the Massachusetts Institute of Technology (MIT) campus. The image shows a dense cluster of buildings, including several large domes and modern structures, interspersed with green lawns and trees. The perspective is from above, looking down at the university grounds.

We used to joke that AI means
'almost implemented'

© 2002 – Rodney Brooks
the director of MIT's Artificial Intelligence Laboratory

Место машинного обучения в области ИИ



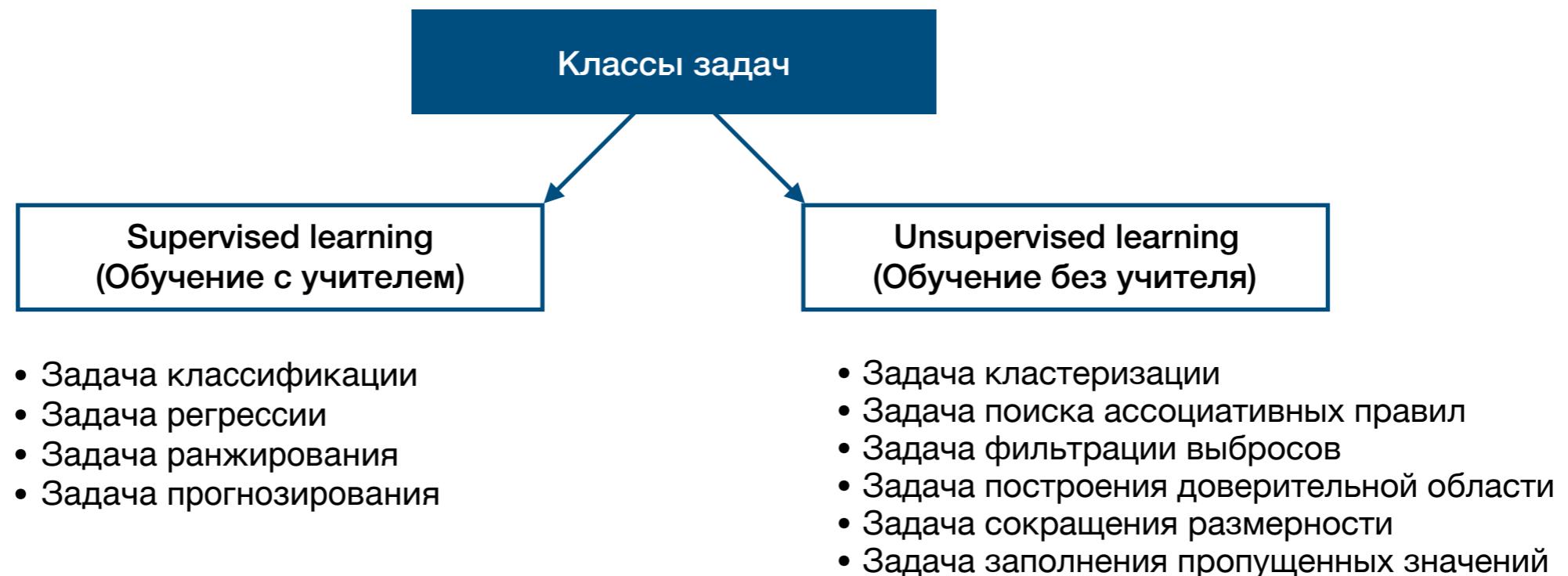
Определение

Машинное обучение – это процесс, в результате которого машина (компьютер) способна показывать поведение, которое в нее не было явно заложено (запрограммировано).

Артур Самуэль, 1959

Компьютерная программа обучается при решении какой-то задачи из класса Т, если ее производительность, согласно метрике Р, улучшается при накоплении опыта Е.

Том Митчелл, 1997



Какие еще бывают классы задач:

Semi-supervised learning (Частичное обучение)

Reinforcement learning (Обучение с подкреплением)

Трансдуктивное обучение

Динамическое обучение

Активное обучение

Метаобучение

Постановка задачи

Задача: восстановить сложную зависимость по конечному числу примеров



Обучающая выборка

Матрица «объекты–признаки»

Датасет о задержках рейсов более 15 минут.

| | Month | DayofMonth | DayOfWeek | DepTime | UniqueCarrier | Origin | Dest | Distance | dep_delayed_15min |
|---|-------|------------|-----------|---------|---------------|--------|------|----------|-------------------|
| 0 | c-8 | c-21 | c-7 | 1934 | AA | ATL | DFW | 732 | N |
| 1 | c-4 | c-20 | c-3 | 1548 | US | PIT | MCO | 834 | N |
| 2 | c-9 | c-2 | c-5 | 1422 | XE | RDU | CLE | 416 | N |
| 3 | c-11 | c-25 | c-6 | 1015 | OO | DEN | MEM | 872 | N |
| 4 | c-10 | c-7 | c-6 | 1828 | WN | MDW | OMA | 423 | Y |

Источник: <https://www.transtats.bts.gov>

Обучающая выборка

Матрица объекты–признаки

Датасет о задержках рейсов более 15 минут.

| | Month | DayofMonth | DayOfWeek | DepTime | UniqueCarrier | Origin | Dest | Distance | dep_delayed_15min |
|---|-------|------------|-----------|---------|---------------|--------|------|----------|-------------------|
| 0 | c-8 | c-21 | c-7 | 1934 | AA | ATL | DFW | 732 | N |
| 1 | c-4 | c-20 | c-3 | 1548 | US | PIT | MCO | 834 | N |
| 2 | c-9 | c-2 | c-5 | 1422 | XE | RDU | CLE | 416 | N |
| 3 | c-11 | c-25 | c-6 | 1015 | OO | DEN | MEM | 872 | N |
| 4 | c-10 | c-7 | c-6 | 1828 | WN | MDW | OMA | 423 | Y |

Признаки

Источник: <https://www.transtats.bts.gov>

Обучающая выборка

Матрица «объекты–признаки»

Датасет о задержках рейсов более 15 минут.

| | Month | DayofMonth | DayOfWeek | DepTime | UniqueCarrier | Origin | Dest | Distance | dep_delayed_15min |
|---|-------|------------|-----------|---------|---------------|--------|------|----------|-------------------|
| 0 | c-8 | c-21 | c-7 | 1934 | AA | ATL | DFW | 732 | N |
| 1 | c-4 | c-20 | c-3 | 1548 | US | PIT | MCO | 834 | N |
| 2 | c-9 | c-2 | c-5 | 1422 | XE | RDU | CLE | 416 | N |
| 3 | c-11 | c-25 | c-6 | 1015 | OO | DEN | MEM | 872 | N |
| 4 | c-10 | c-7 | c-6 | 1828 | WN | MDW | OMA | 423 | Y |

Объекты (прецеденты)

Источник: <https://www.transtats.bts.gov>

Обучающая выборка

Матрица «объекты–признаки»

Датасет о задержках рейсов более 15 минут.

| | Month | DayofMonth | DayOfWeek | DepTime | UniqueCarrier | Origin | Dest | Distance | dep_delayed_15min |
|---|-------|------------|-----------|---------|---------------|--------|------|----------|-------------------|
| 0 | c-8 | c-21 | c-7 | 1934 | AA | ATL | DFW | 732 | N |
| 1 | c-4 | c-20 | c-3 | 1548 | US | PIT | MCO | 834 | N |
| 2 | c-9 | c-2 | c-5 | 1422 | XE | RDU | CLE | 416 | N |
| 3 | c-11 | c-25 | c-6 | 1015 | OO | DEN | MEM | 872 | N |
| 4 | c-10 | c-7 | c-6 | 1828 | WN | MDW | OMA | 423 | Y |

Целевая переменная

Источник: <https://www.transtats.bts.gov>

Формальная постановка задачи

Дана обучающая выборка (объекты независимы):

$$X_m = \{ (x_1, y_1), \dots, (x_m, y_m) \}$$

Для задачи регрессии - Целевая переменная задана вещественным числом

$$(x_1, y_1) \in \mathbb{R}^m \times \mathbb{Y}, \mathbb{Y} = \mathbb{R}$$

Для задачи классификации - Целевая переменная задана конечным числом меток

$$(x_1, y_1) \in \mathbb{R}^m \times \mathbb{Y}, \mathbb{Y} = \{-1; 1\}$$

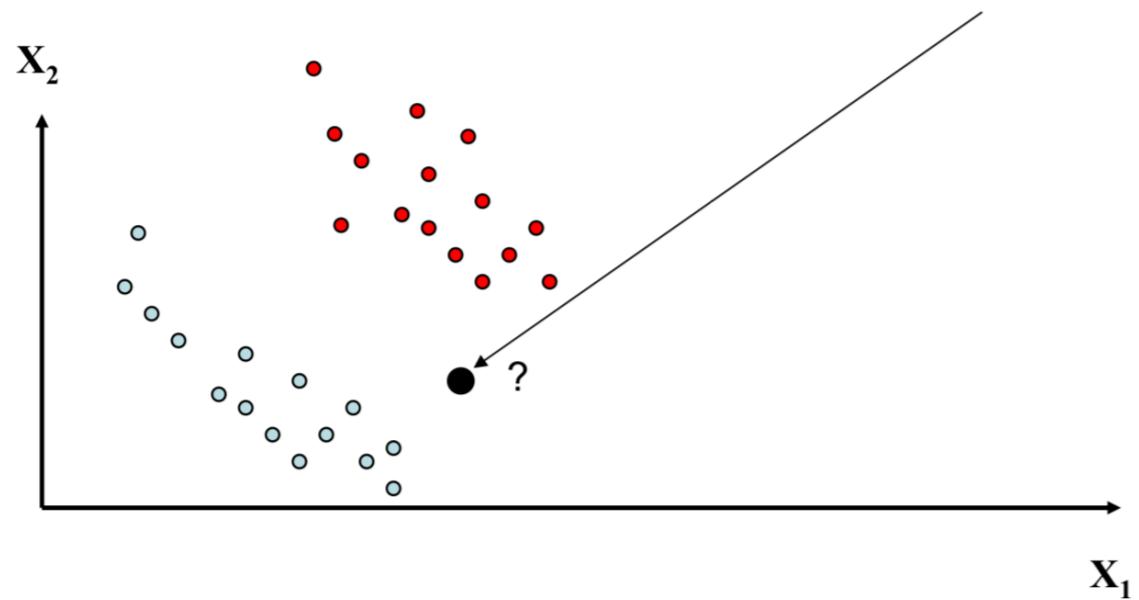
Задать такую функцию $f(x)$ от вектора признаков x , которое выдает ответ для любого возможного наблюдения x

$$f(x): \mathbb{X} \rightarrow \mathbb{Y}$$

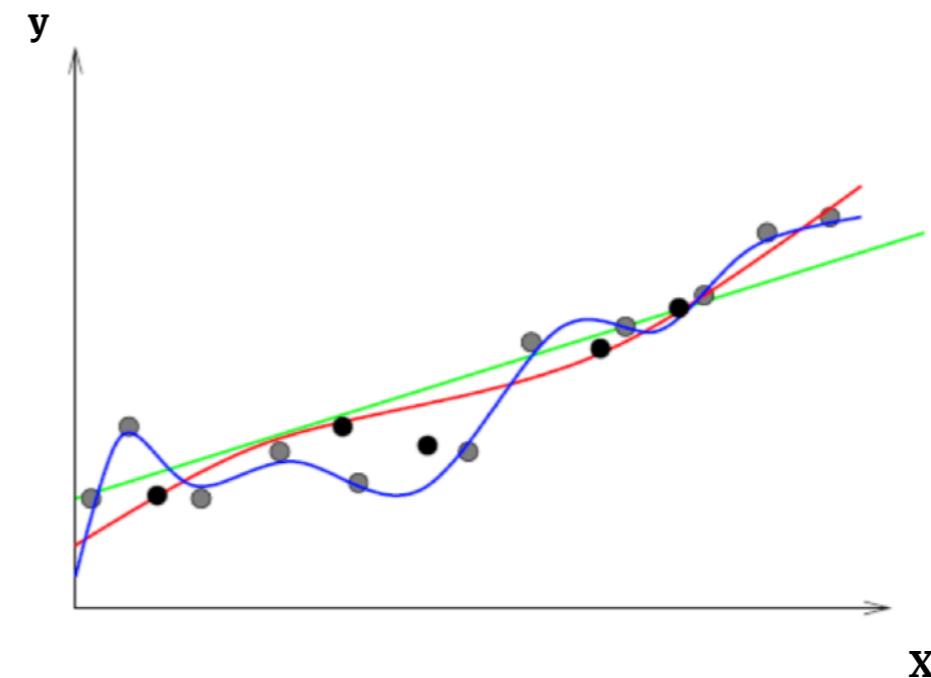
Основная гипотеза МО: Схожим объектам соответствуют схожие объекты

Формальная постановка задачи

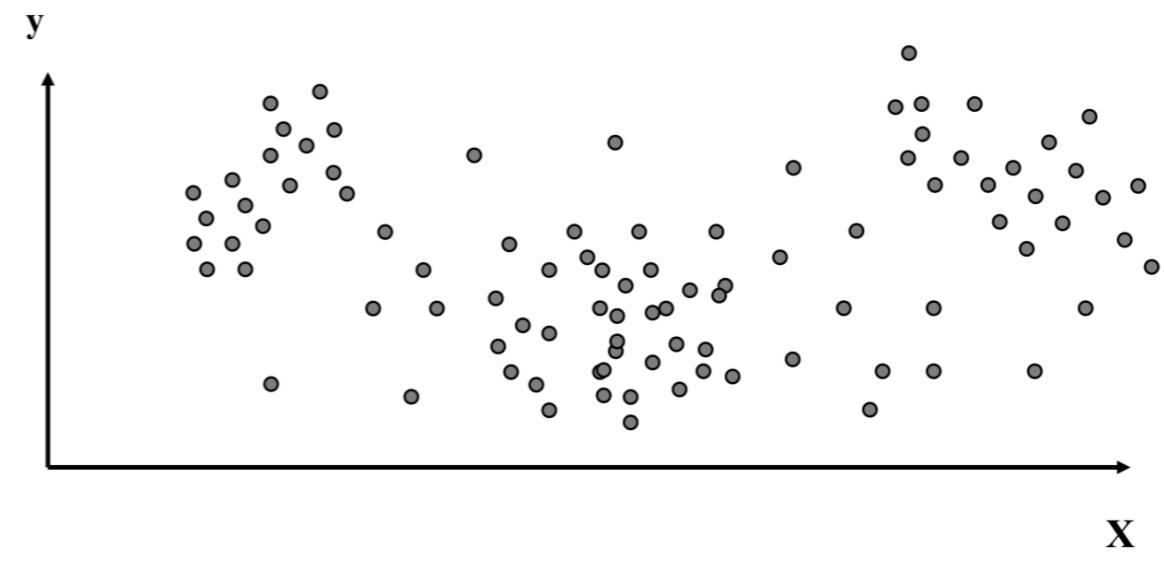
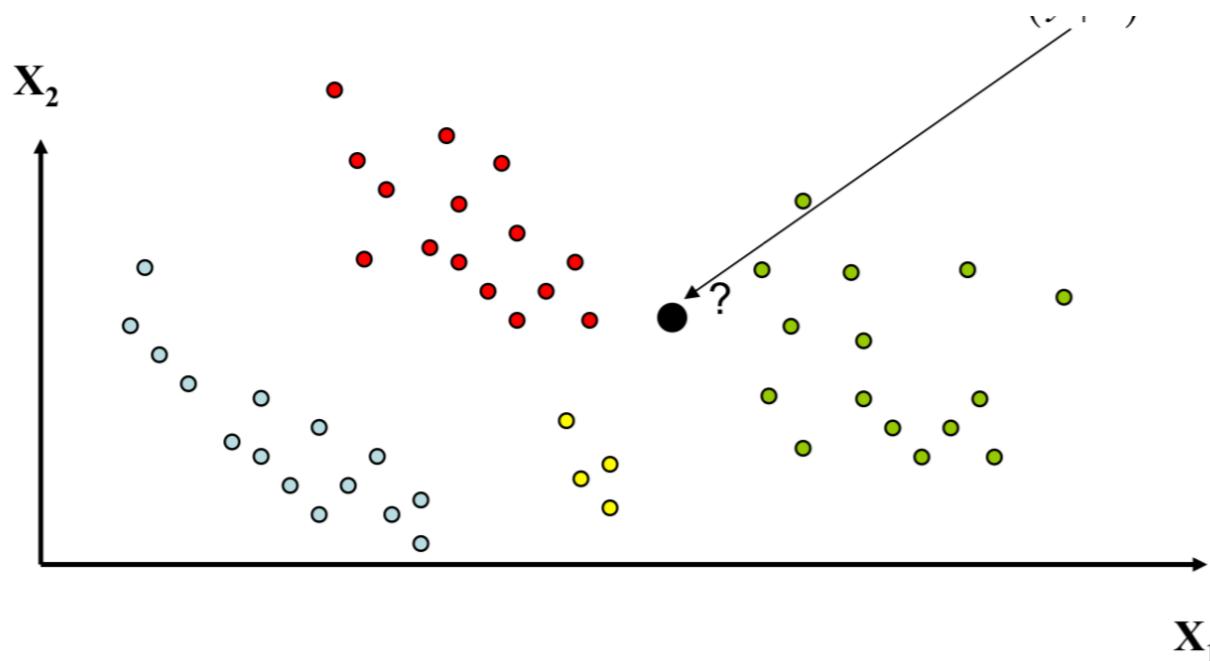
Классификация



Восстановление регрессии



Кластеризация



Признаки

Признаковое описание объекта - Вектор:

$$x_i = \{d_1, d_2, d_3, \dots, d_n\}$$

Множество значений признака

$$d_j \in D_j$$

Бинарные признаки

$$D_j = \{0, 1\}$$

В нашем примере:
Целевая переменная

Категориальные признаки

D_j - упорядоченное множество

В нашем примере:
Локация отправления
Локация прибытия

Вещественные признаки

$$D_j = \mathbb{R}^m$$

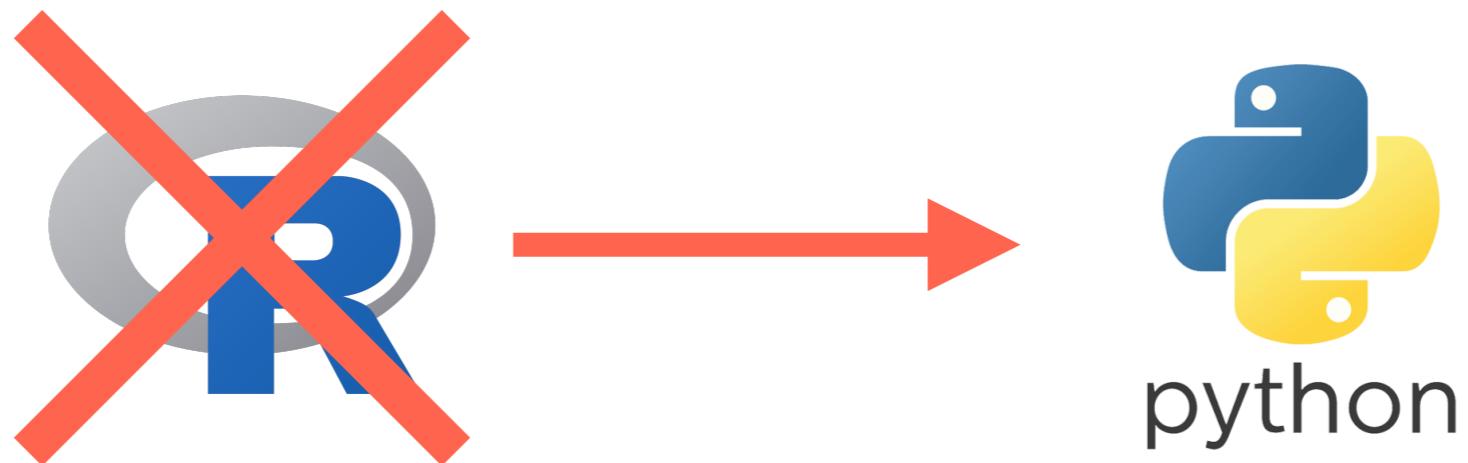
В нашем примере:
Расстояние

Инструменты и библиотеки



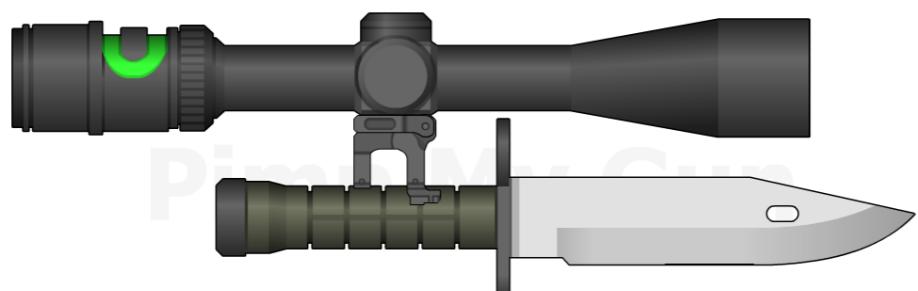
Подробнее про R vs Python: <https://habr.com/company/piter/blog/263457/>

Инструменты и библиотеки



Инструменты и библиотеки

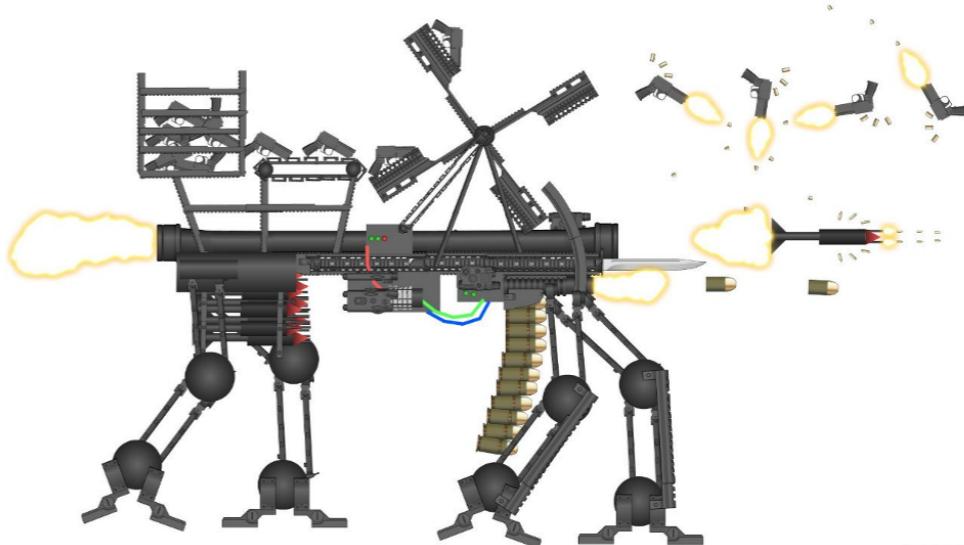
Assembly



C++



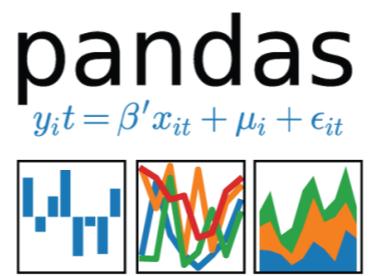
Python



C



Инструменты и библиотеки



XGBoost



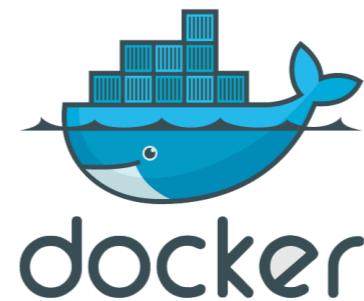
PYTORCH



Инструменты и библиотеки



<https://www.anaconda.com/download/>



<https://www.docker.com/products/docker-desktop>

<https://hub.docker.com/r/vlasoff/ds/>

Формальная постановка задачи

Дана обучающая выборка (объекты независимы):

$$X_m = \{ (x_1, y_1), \dots, (x_m, y_m) \}$$

Для задачи регрессии - Целевая переменная задана вещественным числом

$$(x_1, y_1) \in \mathbb{R}^m \times \mathbb{Y}, \mathbb{Y} = \mathbb{R}$$

Для задачи классификации - Целевая переменная задана конечным числом меток

$$(x_1, y_1) \in \mathbb{R}^m \times \mathbb{Y}, \mathbb{Y} = \{-1; 1\}$$

Задать такую функцию $f(x)$ от вектора признаков x , которое выдает ответ для любого возможного наблюдения x

$$f(x): \mathbb{X} \rightarrow \mathbb{Y}$$

Основная гипотеза МО: Схожим объектам соответствуют схожие объекты

Метод k ближайших соседей (kNN, *k Nearest Neighbours*)

Гипотеза компактности: если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

- Вычислить расстояние до каждого из объектов обучающей выборки
- Отобрать k объектов обучающей выборки, расстояние до которых минимально
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей



`sklearn.neighbors.KNeighborsRegressor`

`(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2,
metric='minkowski', metric_params=None, n_jobs=None, **kwargs)`

`sklearn.neighbors.KNeighborsClassifier`

`(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2,
metric='minkowski', metric_params=None, n_jobs=None, **kwargs)`

Метод k ближайших соседей (kNN, *k Nearest Neighbours*)

Гипотеза компактности: если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

- Вычислить расстояние до каждого из объектов обучающей выборки
- Отобрать k объектов обучающей выборки, расстояние до которых минимально
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей



`sklearn.neighbors.KNeighborsRegressor`

`(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None, **kwargs)`

`sklearn.neighbors.KNeighborsClassifier`

`(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None, **kwargs)`

Выбор числа соседей k

При $k=1$ алгоритм ближайшего соседа неустойчив к шумовым выбросам: он даёт ошибочные классификации не только на самих объектах-выбросах, но и на ближайших к ним объектах других классов.

При $k=l$, наоборот, алгоритм чрезмерно устойчив и вырождается в константу. Таким образом, крайние значения нежелательны.

Метод k ближайших соседей (kNN, *k Nearest Neighbours*)

Выбор метрики

Евклидово расстояние (“euclidean”)

$$\sqrt{\sum (x - y)^2}$$

Расстояние городских кварталов «манхэттенское расстояние» (“manhattan”)

$$\sum |x - y|$$

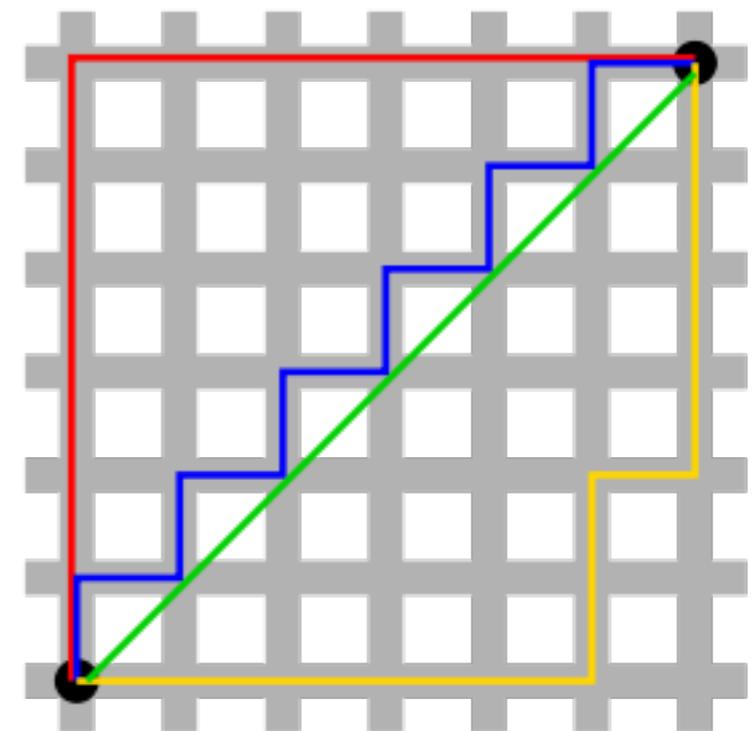
Расстояние Чебышева “chebyshev”

$$\max(x - y)$$

Расстояние Минковского “minkowski”

$$\left(\sum |x - y|^p \right)^{\frac{1}{p}}$$

расстояния с параметром p равным 1 (расстояние городских кварталов) или 2 (евклидова метрика).
 $p = \infty$ метрика обращается в расстояние Чебышёва.



Метод k ближайших соседей (kNN, *k Nearest Neighbours*)

Нормирование признаков

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$z \in [0,1]$$

Стандартизация признаков

$$z = \frac{x - \mu}{\sigma}$$

где

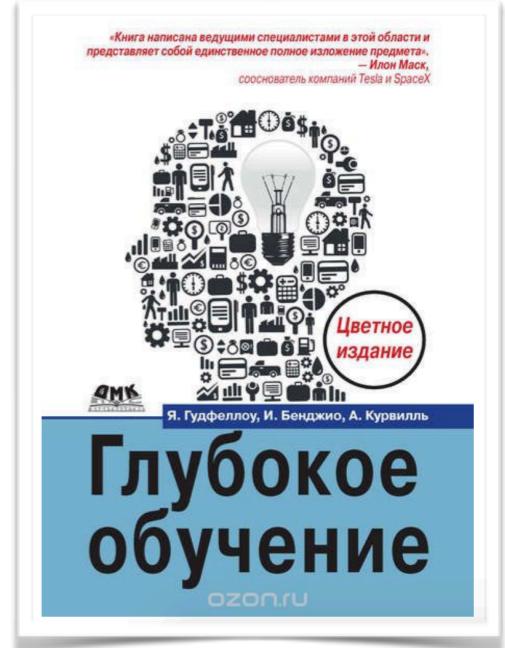
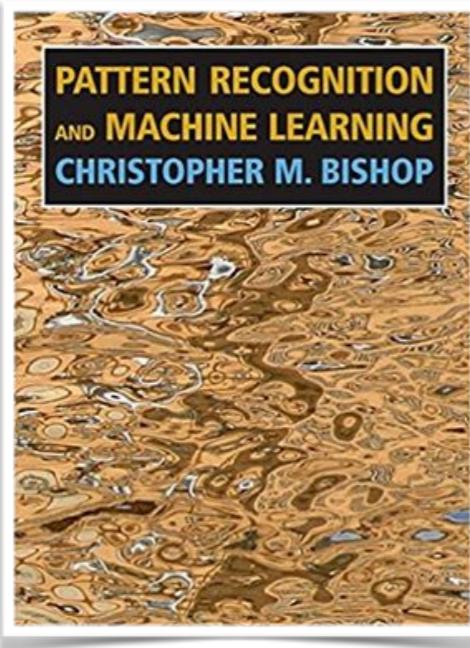
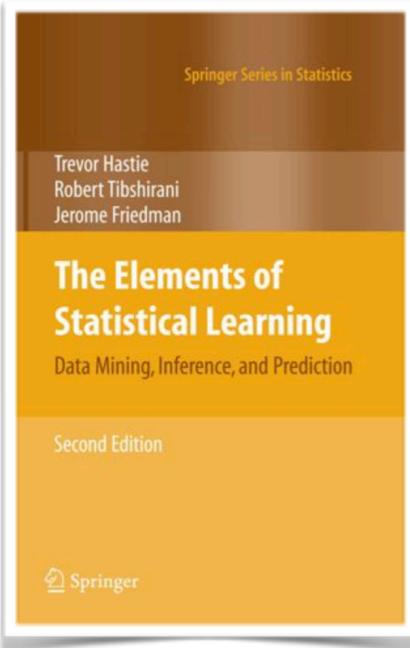
μ - Среднее

σ - Стандартное отклонение

Проклятие размерности

В пространстве высокой размерности все объекты примерно одинаково далеки друг от друга; выбор ближайших соседей становится практически произвольным.

Литература, курсы, ссылки



Курсы:

- Открытый курс машинного обучения (ODS)
- Специализация МФТИ и Яндекс на Coursera
- Machine Learning от Andrew Ng
- Введение в машинное обучение от Яндекса и ВШЭ

Ссылки:

- <https://github.com/ml-mipt>
- <https://www.openml.org>
- <https://opendatascience.slack.com>
- <https://www.kaggle.com>

Just for fun

