

# Машинное обучение

Лекция 1

Основные понятия, термины, подходы, инструменты

*Власов Кирилл Вячеславович*



2018

# О курсе

**Три больших блока:**

1. Линейный модели, SVM, Naive Bayes

# О курсе

## Три больших блока:

1. Линейный модели, SVM, Naive Bayes
2. Деревья решений, Ансамбли, Бустинг

# О курсе

## Три больших блока:

1. Линейный модели, SVM, Naive Bayes
2. Деревья решений, Ансамбли, Бустинг
3. Радемахеровская сложность, статистическое обучение и прочее

# О курсе

**Три больших блока:**

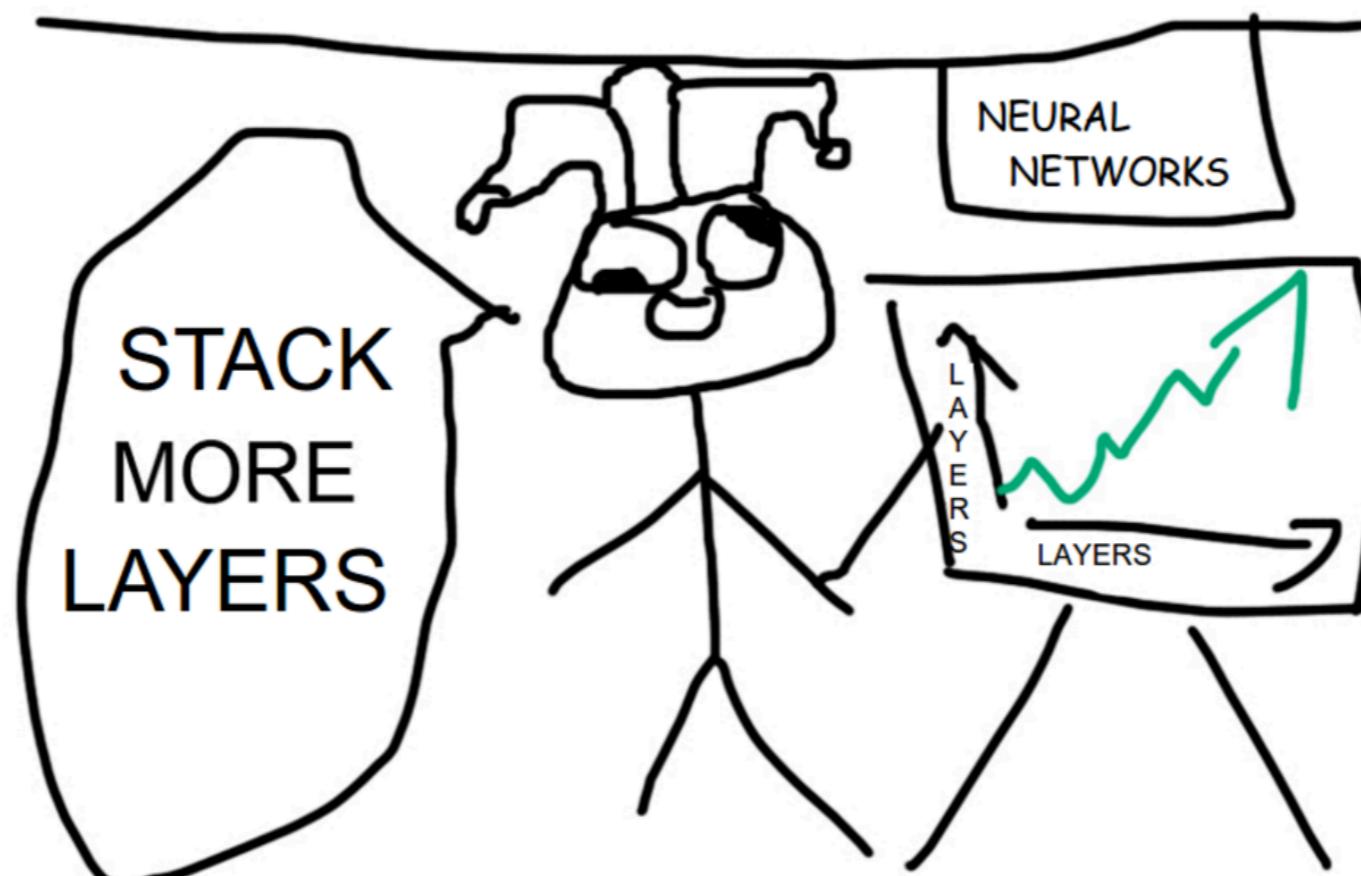
1. Линейный модели, SVM, Naive Bayes
  2. Деревья решений, Ансамбли, Бустинг
  3. Радемахеровская сложность, статистическое обучение и прочее
-

# О курсе

**Три больших блока:**

1. Линейный модели, SVM, Naive Bayes
  2. Деревья решений, Ансамбли, Бустинг
  3. Радемахеровская сложность, статистическое обучение и прочее
- 

*Deep Learning!*



# О курсе

**I**  
семестр

**II**  
семестр

# О курсе

**I**  
семестр

**Дифференциальный зачет**

**II**  
семестр

**Экзамен**

# О курсе

**I**  
семестр

**Дифференциальный зачет**

**II**  
семестр

**Экзамен**

**Видео-лекции**

# О курсе

**I**  
семестр

**Дифференциальный зачет**

**II**  
семестр

**Экзамен**

**Видео-лекции**

**Живые семинары**

**15:20 / 16:55**

# О курсе

**I**  
семестр

**Дифференциальный зачет**

**II**  
семестр

**Экзамен**

**Видео-лекции**

**Живые семинары**

**15:20 / 16:55**

**Quizzes! - Перед каждым семинаром**

# О курсе

**I**  
семестр

**Дифференциальный зачет**

**II**  
семестр

**Экзамен**

**Видео-лекции**

**Живые семинары**

**15:20 / 16:55**

**Quizzes! - Перед каждым семинаром**

**7 домашних заданий с дедлайнами**

# О курсе

**I**  
семестр

**Дифференциальный зачет**

**II**  
семестр

**Экзамен**

**Видео-лекции**

**Живые семинары**

**15:20 / 16:55**

**Quizzes! - Перед каждым семинаром**

**7 домашних заданий с дедлайнами**

**2 in-class competitions в течении всего курса**

**ФИВТ**



 **МФТИ**

The logo features a blue stylized wave or pulse icon followed by the text 'МФТИ' in a bold, blue, sans-serif font.

**<https://github.com/ml-mipt>**

**<https://github.com/ml-dafe>**



**[https://github.com/ml-dafe/ml\\_mipt\\_dafe\\_major](https://github.com/ml-dafe/ml_mipt_dafe_major)**



# 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!**

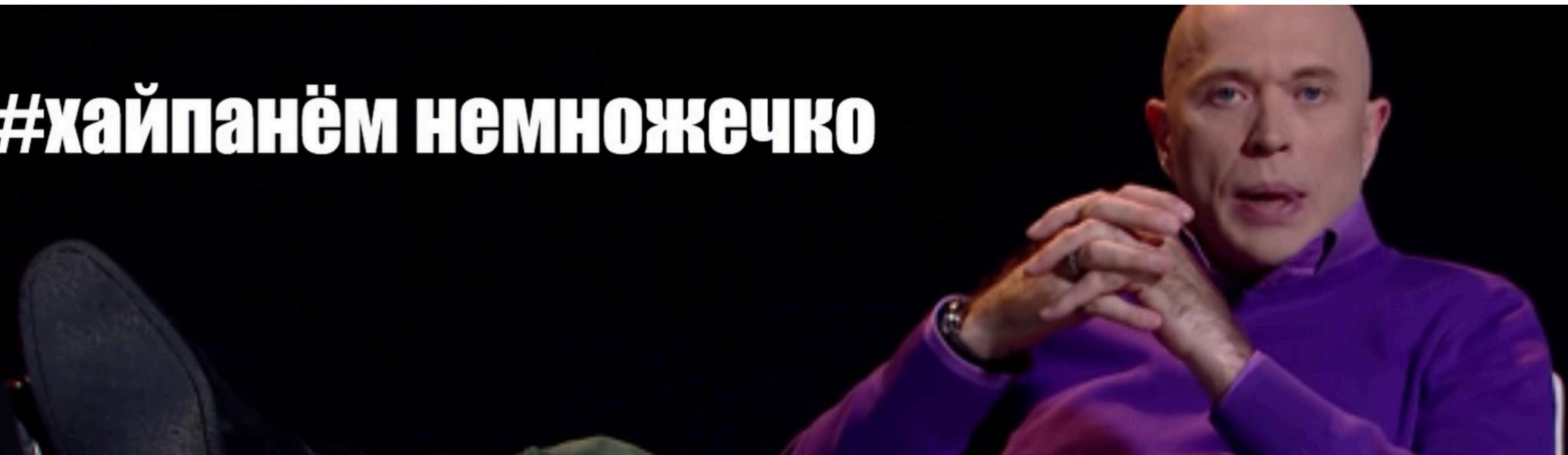
Компании собирают  
БигДату и пытаются  
извлекать из нее пользу

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!**

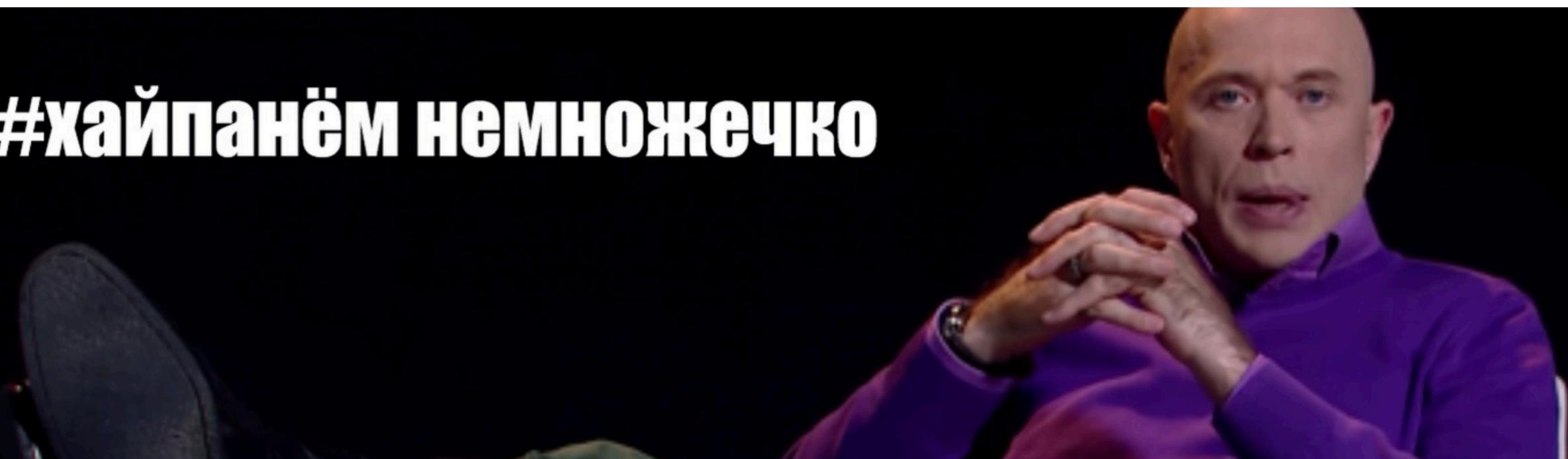
Компании собирают  
БигДату и пытаются  
извлекать из нее пользу

4.

**Превосходство ИИ!**

В некоторых задачах ИИ  
справляется с решением  
лучше человека

#хайпанём немножечко



# Искусственный интеллект



1997

*Deep Blue выиграл чемпиона мира по шахматам Гарри Каспарова.*



2016

*AlphaGo выиграла матч у профессионала Ли Седоль*

# Примеры в авиации

**ИИ победил человека в симуляции воздушного боя**

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

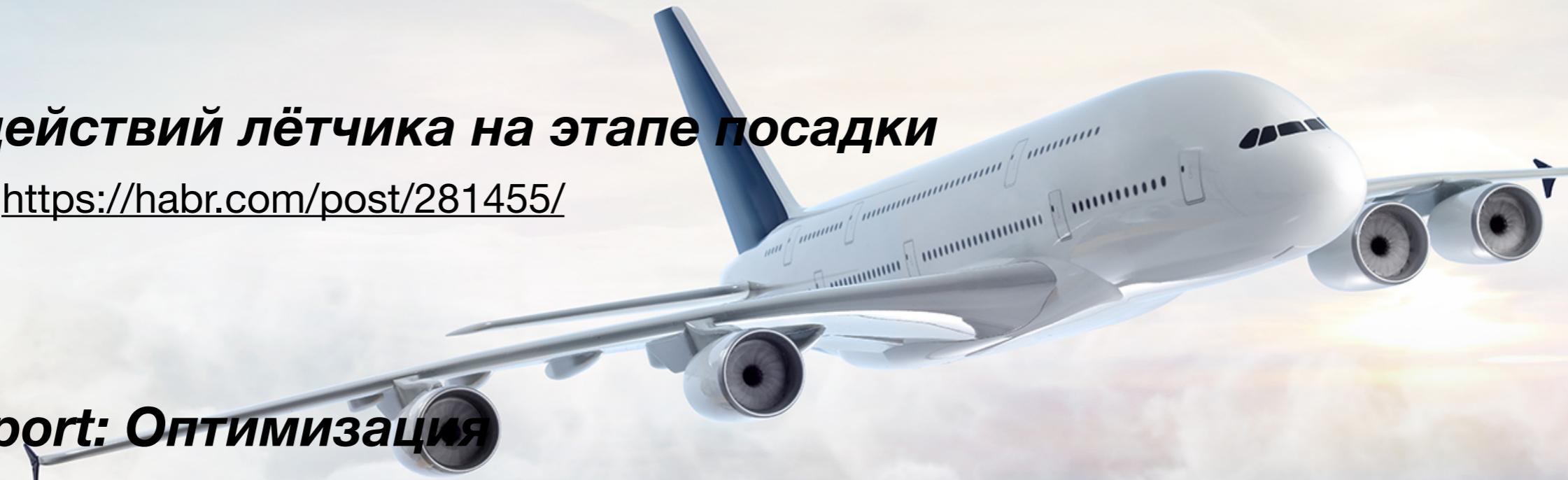
**Оценка действий лётчика на этапе посадки**

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

**Dubai airport: Оптимизация выходов на посадку и полетов**

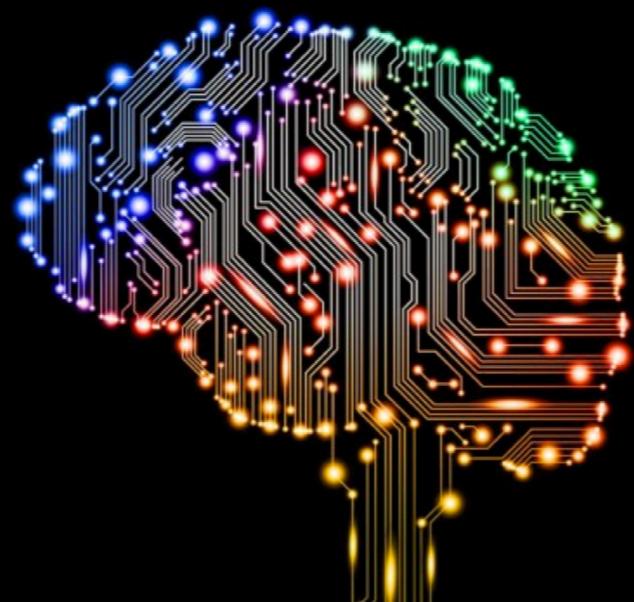
**Rolls-Royce: Диагностика самолетов**

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



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

© 1956 – Джон Маккарти



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'

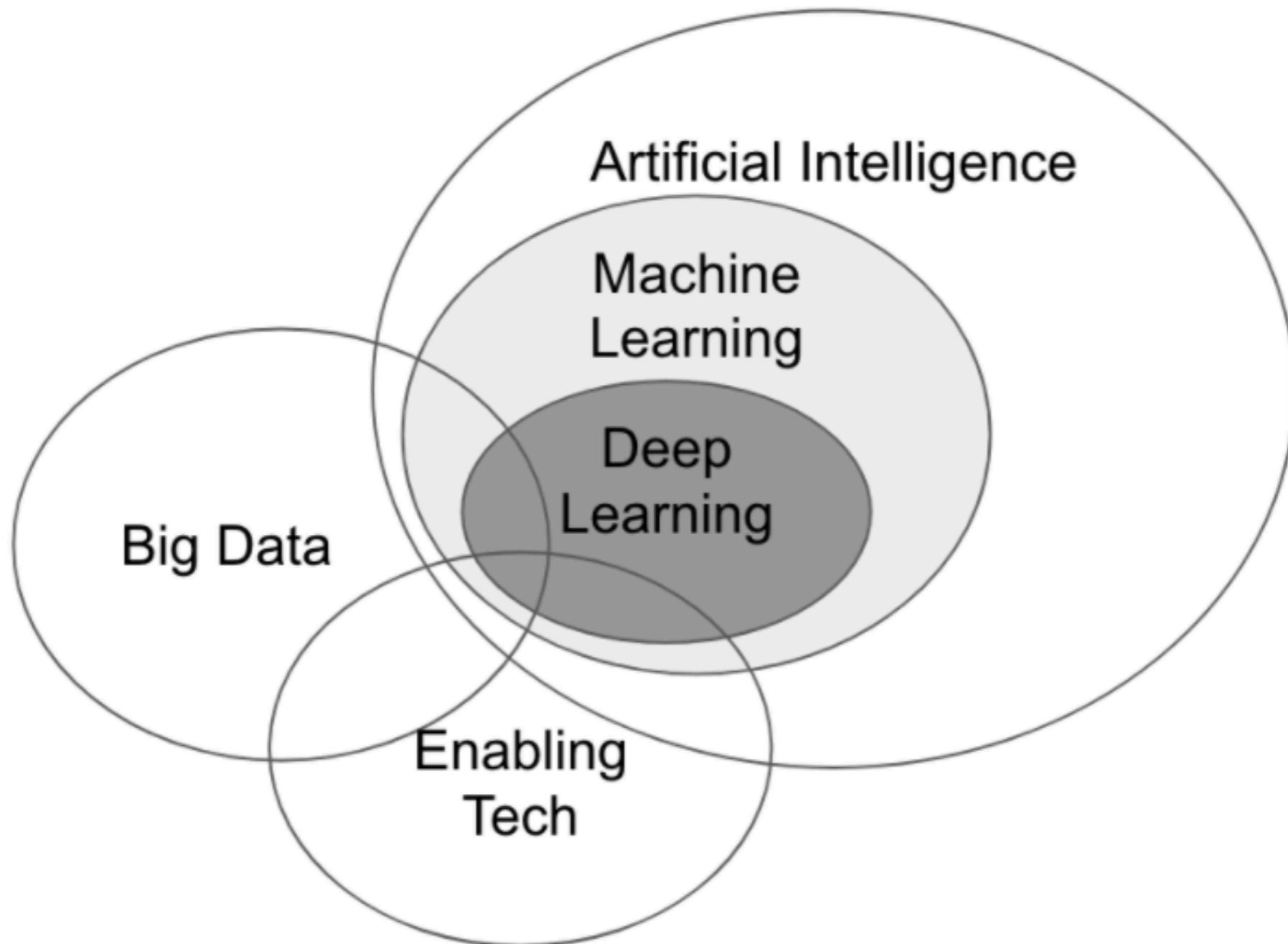
© 2002 – Rodney Brooks  
*the director of MIT's Artificial Intelligence Laboratory*

# Искусственный интеллект

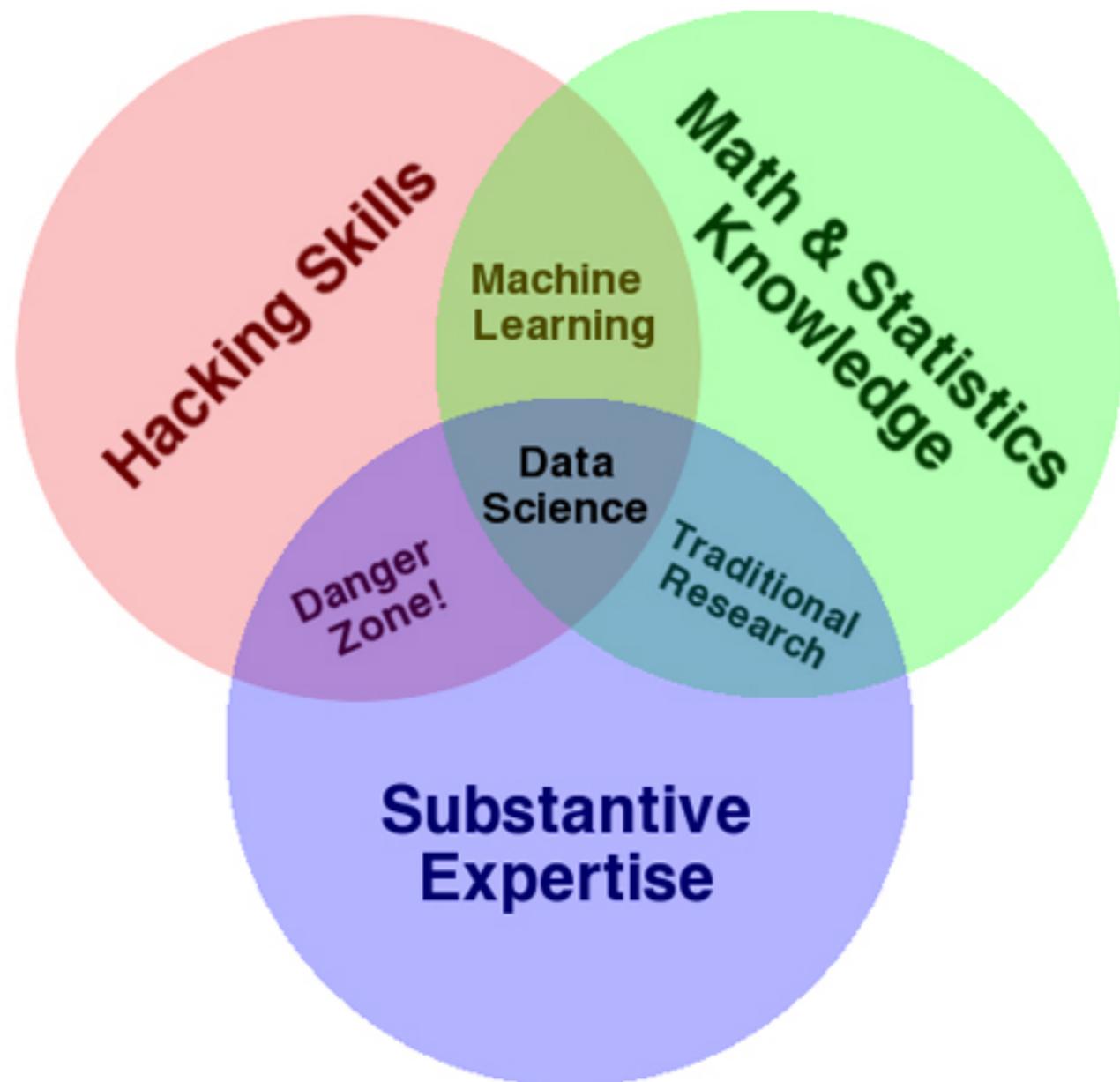
И.И.



# Место машинного обучения в области ИИ



# 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



- Задача классификации
  - Задача регрессии
  - Задача ранжирования
  - Задача прогнозирования
- Задача кластеризации
  - Задача поиска ассоциативных правил
  - Задача фильтрации выбросов
  - Задача построения доверительной области
  - Задача сокращения размерности
  - Задача заполнения пропущенных значений

**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

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

Объекты (прецеденты)

# 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

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

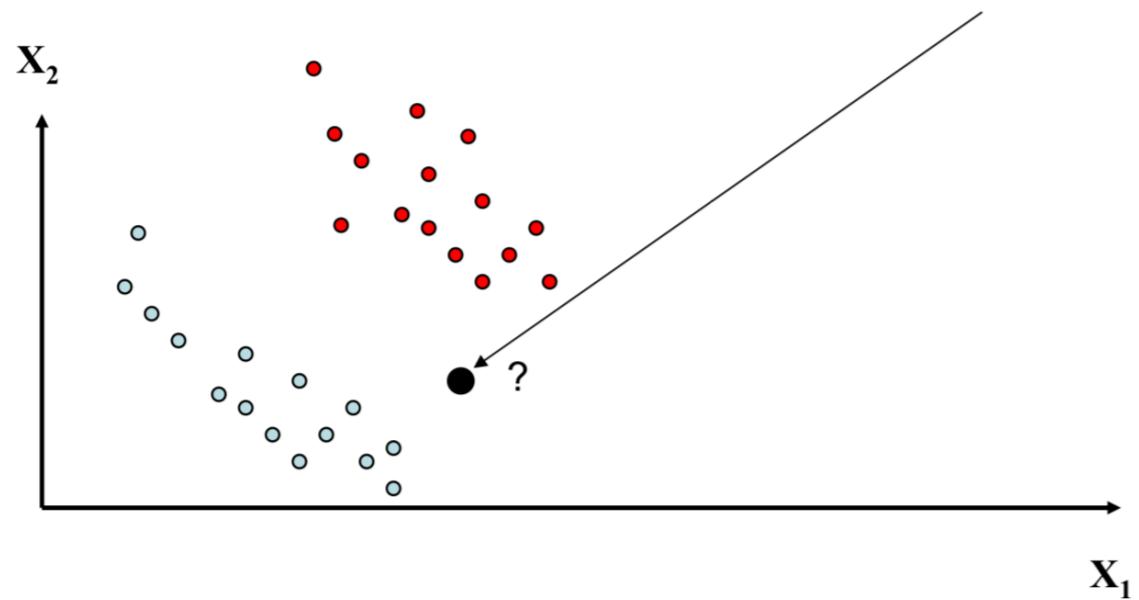
Задать такую функцию  $f(x)$  от вектора признаков  $x$ , которое выдает ответ для любого возможного наблюдения  $x$

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

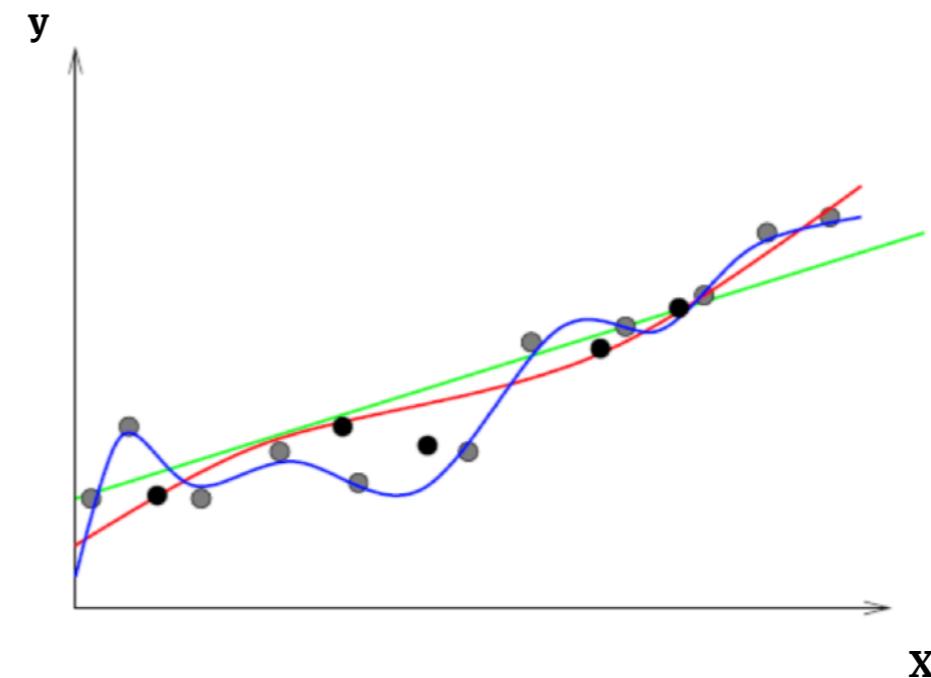
**Основная гипотеза МО:** Схожим объектам соответствуют схожие объекты

# Формальная постановка задачи

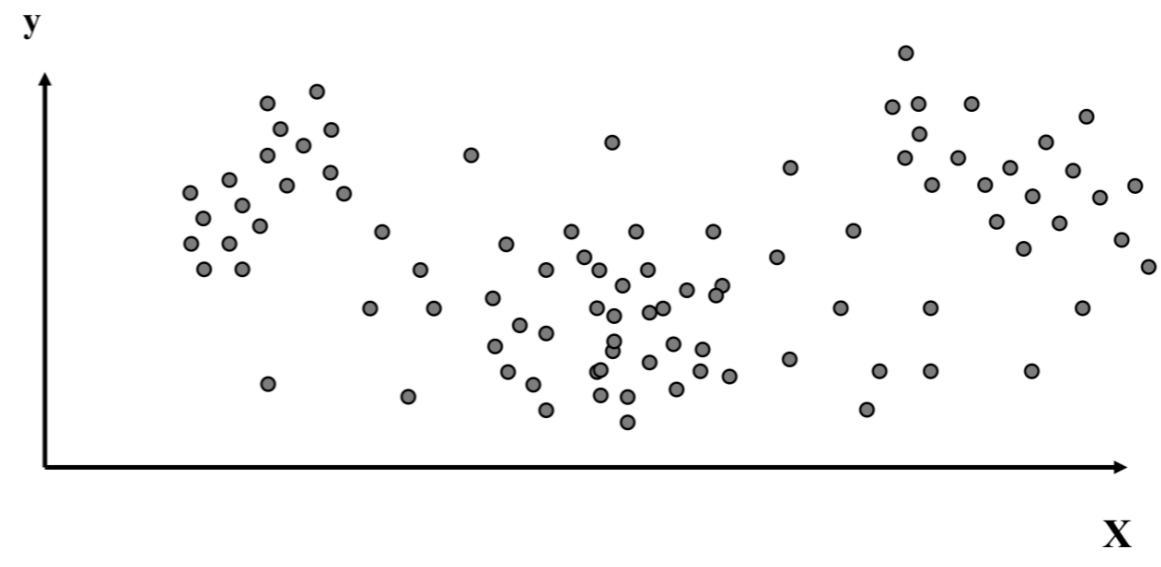
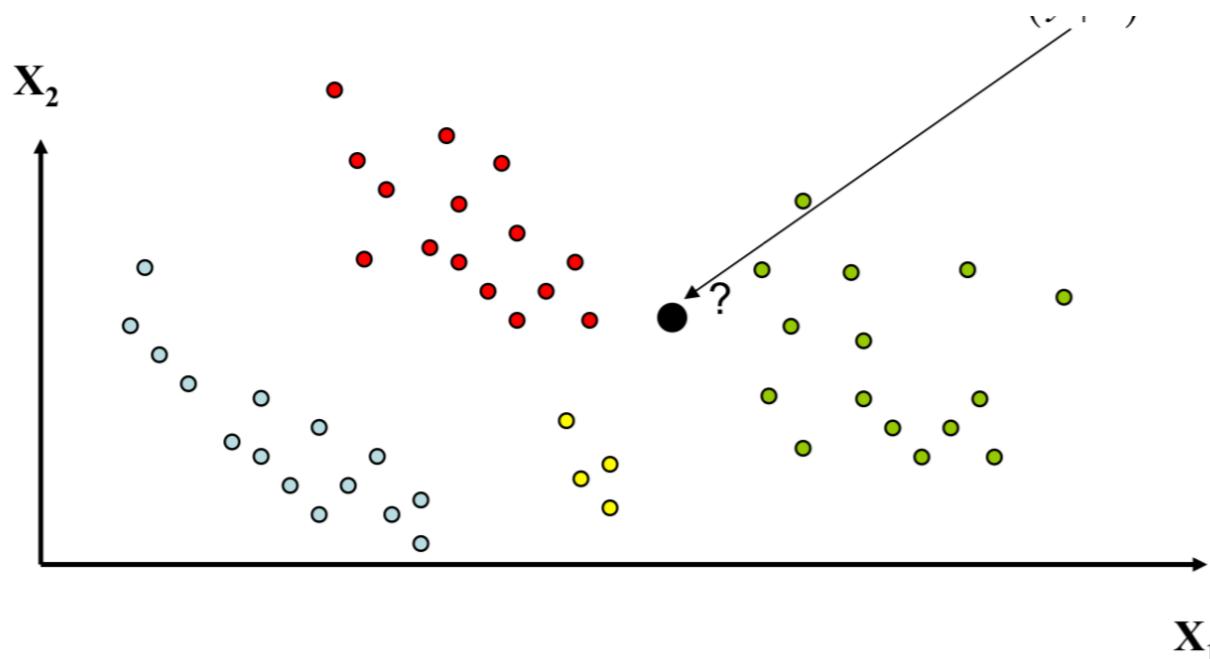
**Классификация**



**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

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

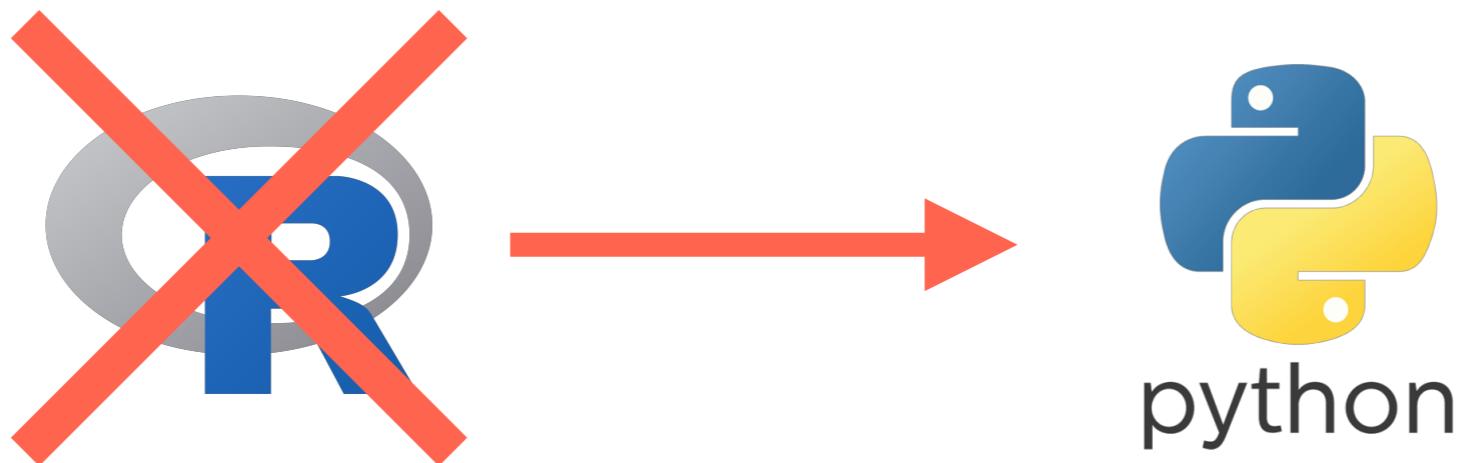
В нашем примере:  
Расстояние

# Инструменты и библиотеки



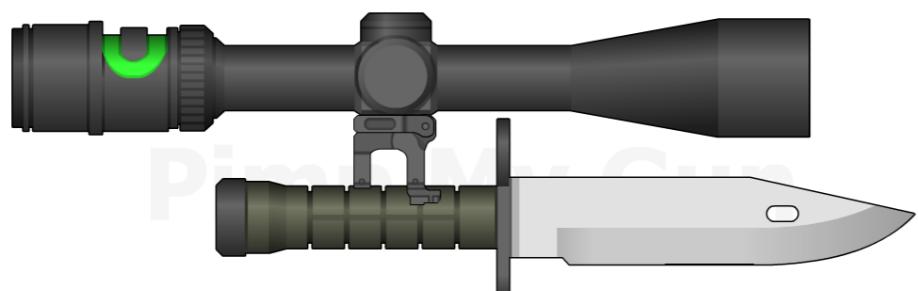
Подробнее про R vs Python: <https://habr.com/company/piter/blog/263457/>

# Инструменты и библиотеки



# Инструменты и библиотеки

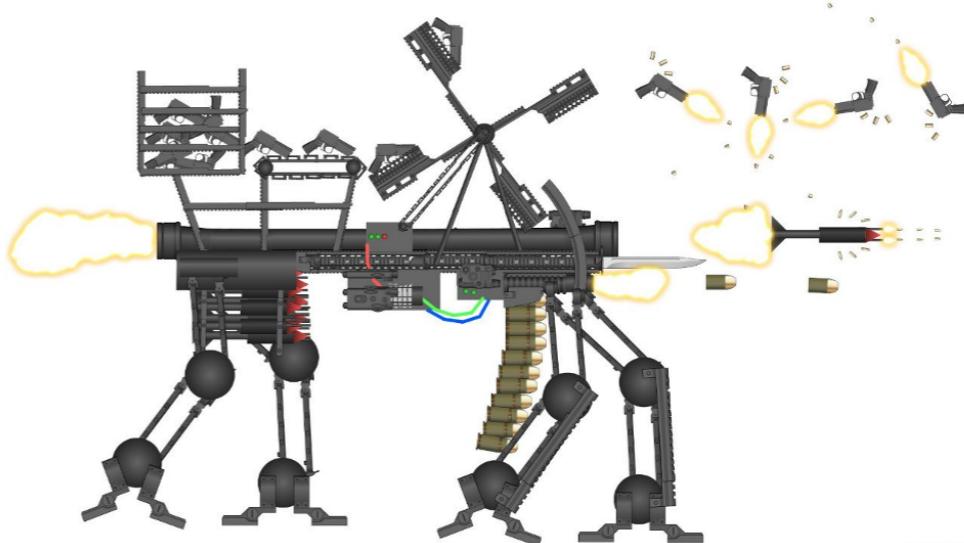
Assembly



C++



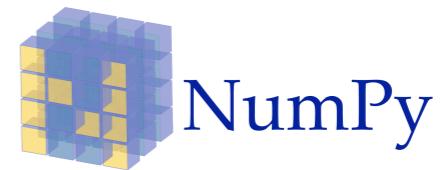
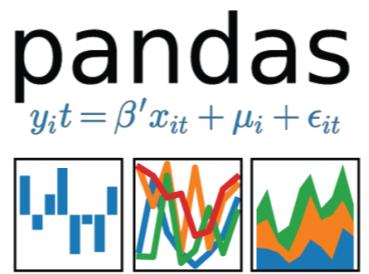
Python



C



# Инструменты и библиотеки



**XGBoost**



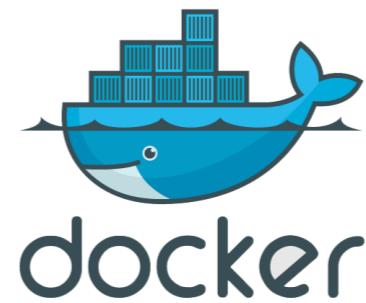
**P Y T  R C H**



# Инструменты и библиотеки



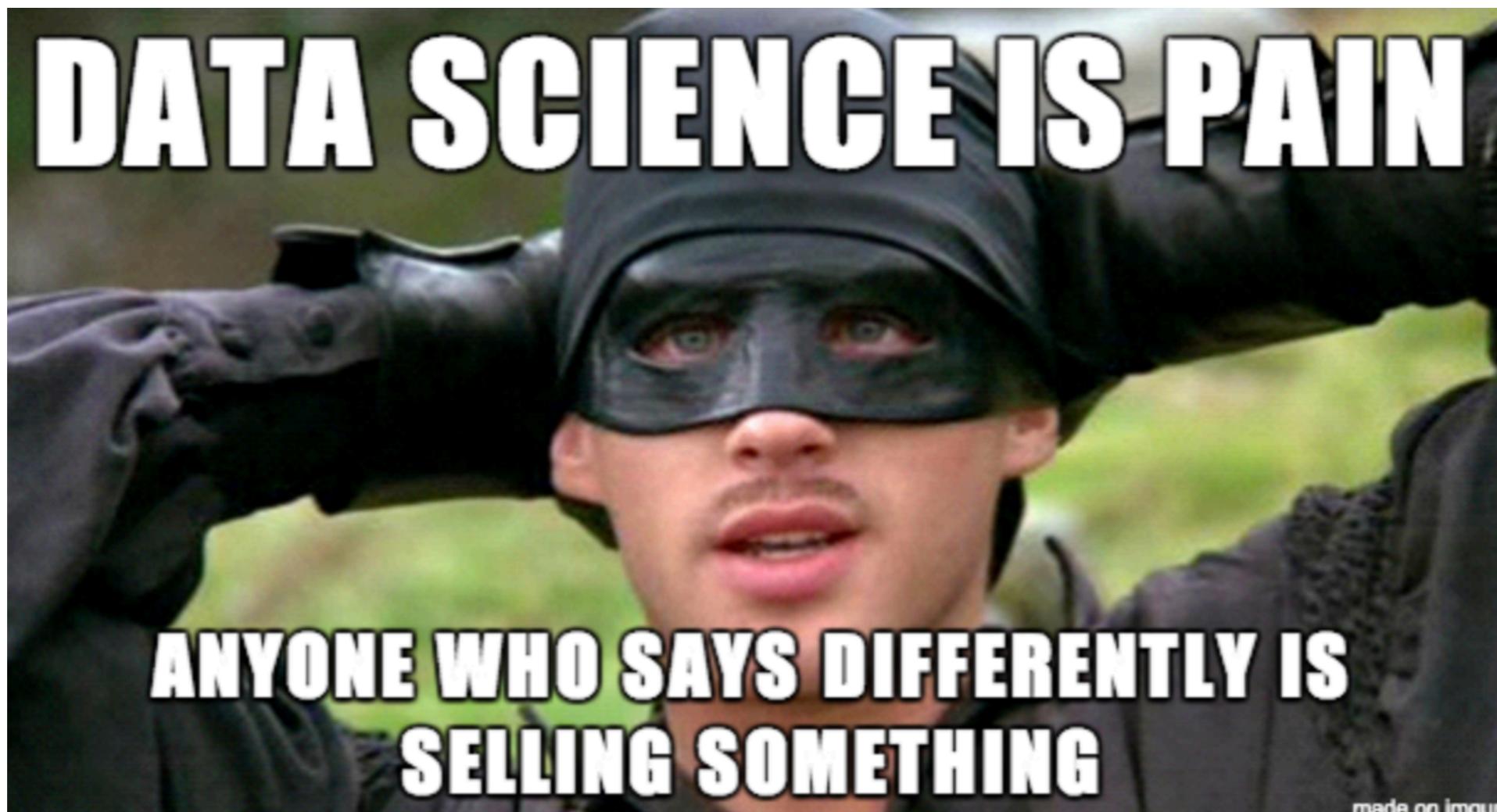
<https://www.anaconda.com/download/>



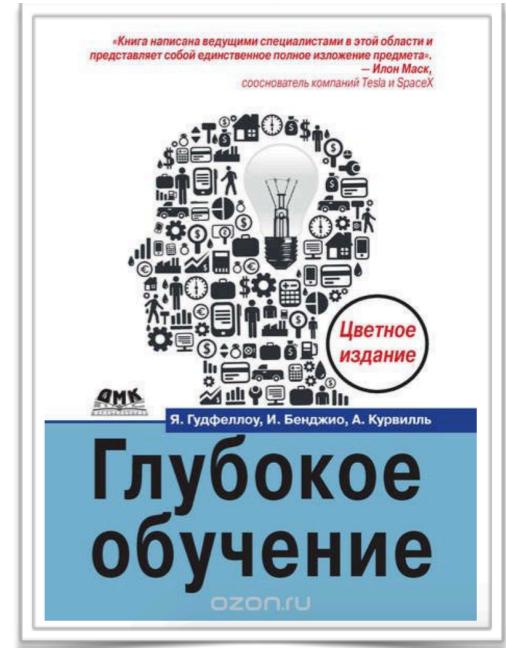
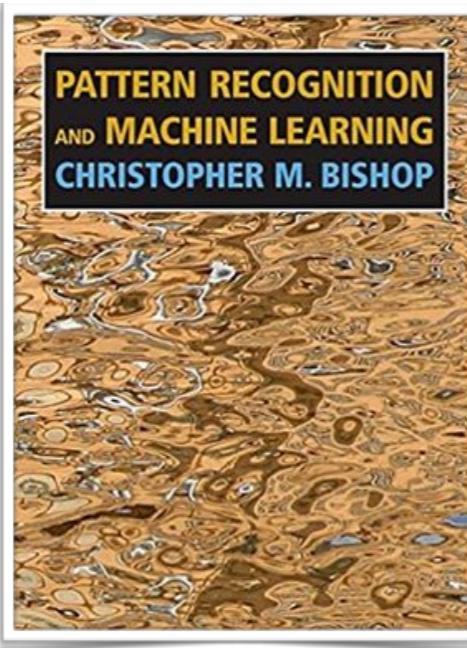
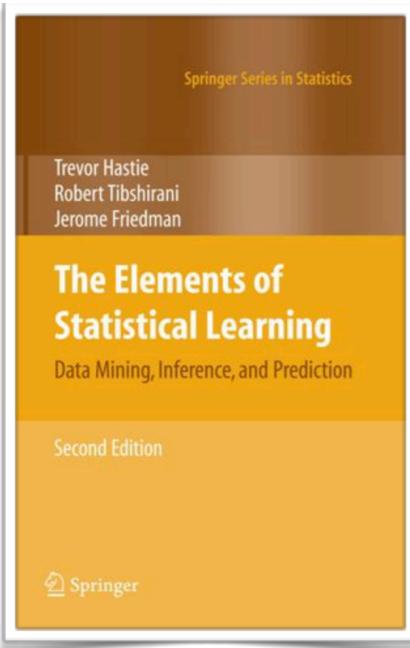
<https://www.docker.com/products/docker-desktop>

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

Just for fun



# 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>