

# Машинное обучение: базовые концепции машинного обучения

MADE academy  
Эмели Драль

# About me



- **Co-founder & CTO** Evidently AI
- Ex **Chief Data Scientist** at Yandex Data Factory and Mechanica AI
- Co-founder of **Data Mining in Action**, largest offline data science course in Russia
- Co-author of two **Coursera** specializations in data science with > 100K students
- Lecturer at **Harbour.Space University**, GSOM MBA

**50+**

Industrial applications of  
machine learning

# Программа курса

Курс состоит из **3х** блоков:

1. **Базовые** концепции машинного обучения
2. **Алгоритмы** машинного обучения
3. **Прикладное** машинное обучение

# Базовые концепции машинного обучения

1. Виды обучения, виды задач, базовые концепции
2. Простые алгоритмы: логика построения и связь с математикой
3. Оценка качества в машинном обучении

**Результат изучения:** разбираетесь в **видах обучения**, понимаете логику работы **базовых алгоритмов**, можете **валидировать модели**

# Алгоритмы машинного обучения

1. Обучение с учителем: линейные модели
2. Обучение с учителем: модели на основе деревьев и композиции
3. Обучение с учителем: нейросетевые модели
4. Обучение без учителя: обзор методов
5. (optional) Рекомендательные системы
6. (optional) Обучение с подкреплением

**Результат изучения:** разбираетесь в деталях методов, можете **применять на практике**

# Прикладное машинное обучение

1. Предпроектное исследование: от постановки задачи до оценки потенциального эффекта
2. Оптимизация модели: feature engineering, pipelines, fall-backs, hybrid models
3. Валидация модели: качество, стабильность, несмещенность, калибровка, интерпретируемость
4. Чек-лист data scientist: классические ошибки при работе на проектов и минимизация рисков

**Результат изучения:** можете **работать в индустрии** под руководством старшего специалиста

# Система оценки

В курсе 4 домашних задания:

1. **Базовые концепции машинного обучения** – 1 задание (20 баллов)
2. **Алгоритмы машинного обучения** – 2 задания (25 и 30 баллов)
3. **Прикладное машинное обучение** – 1 задание (25 баллов)

Итоговая оценка определяется суммой баллов, полученных за задания:

- от 60 до 70 - зачленено
- от 71 до 85 - хорошо
- от 86 и выше - отлично

# Наши цели

- Изучить машинное обучение на уровне, достаточном для работы с технологией в качестве смежного специалиста
- Научиться строить и валидировать стандартные модели
- Наработать базу для изучения более специализированных областей машинного обучения
- Разобраться в нюансах применения машинного обучения на практике для того, чтобы работать под руководством старшего специалиста

# Базовые концепции ML

1. Области применения
2. Базовые концепты
3. Виды обучения
4. Постановка задач обучения
5. Жизненный цикл модели

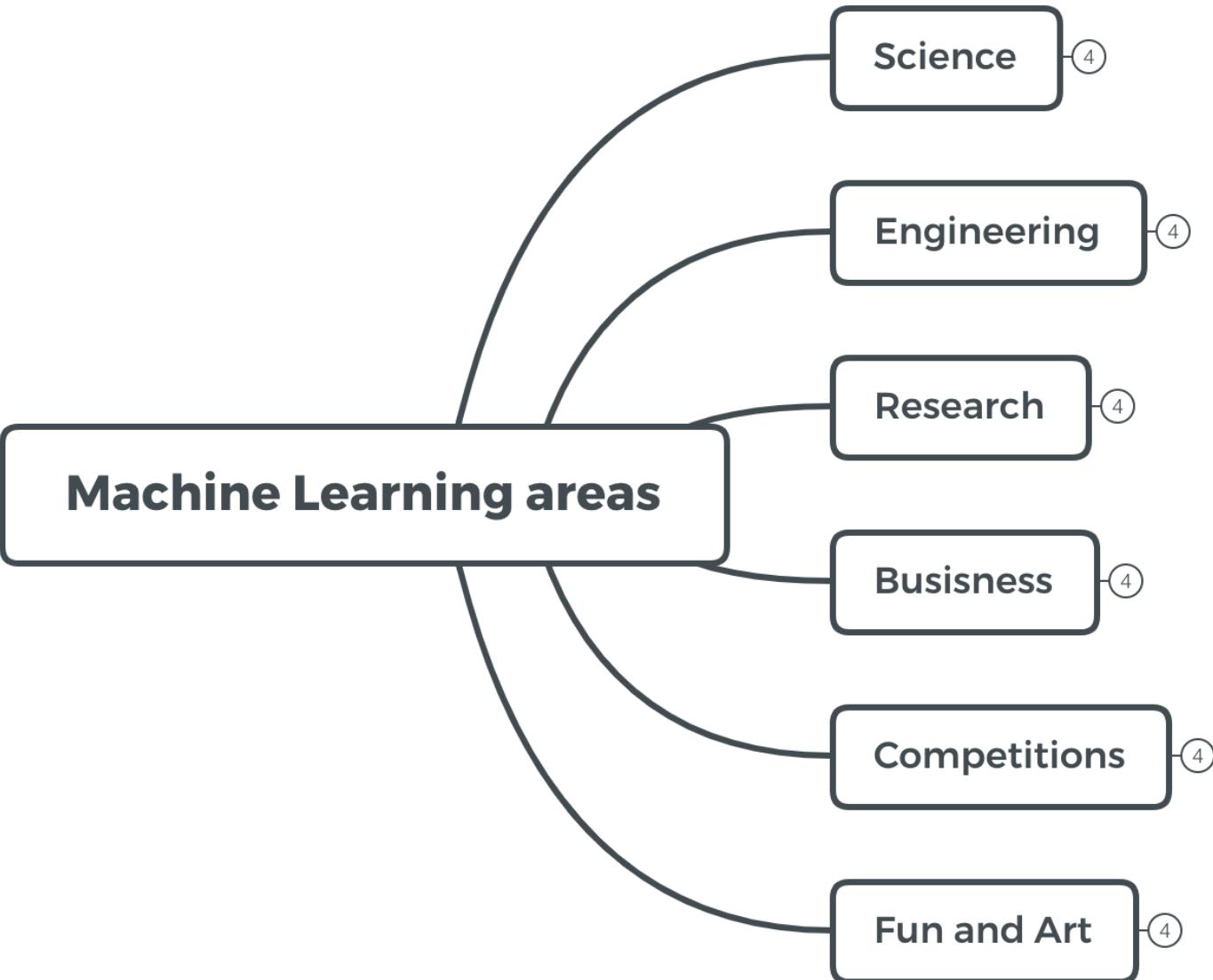
# Области применения машинного обучения

## Области применения

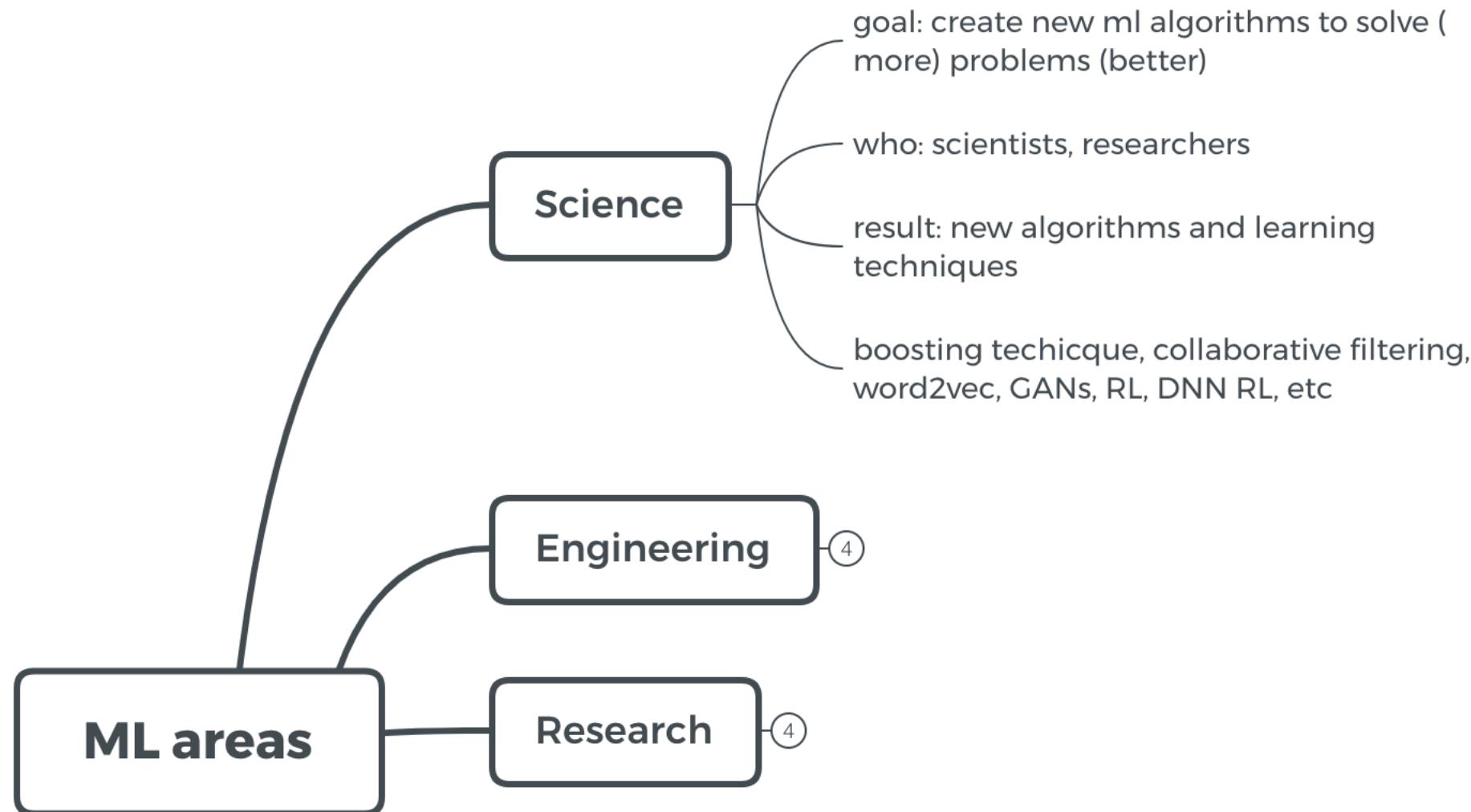


В каких сферах есть место для применения машинного обучения?

# Области применения



# Области применения



# Области применения

## BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

Google AI Language

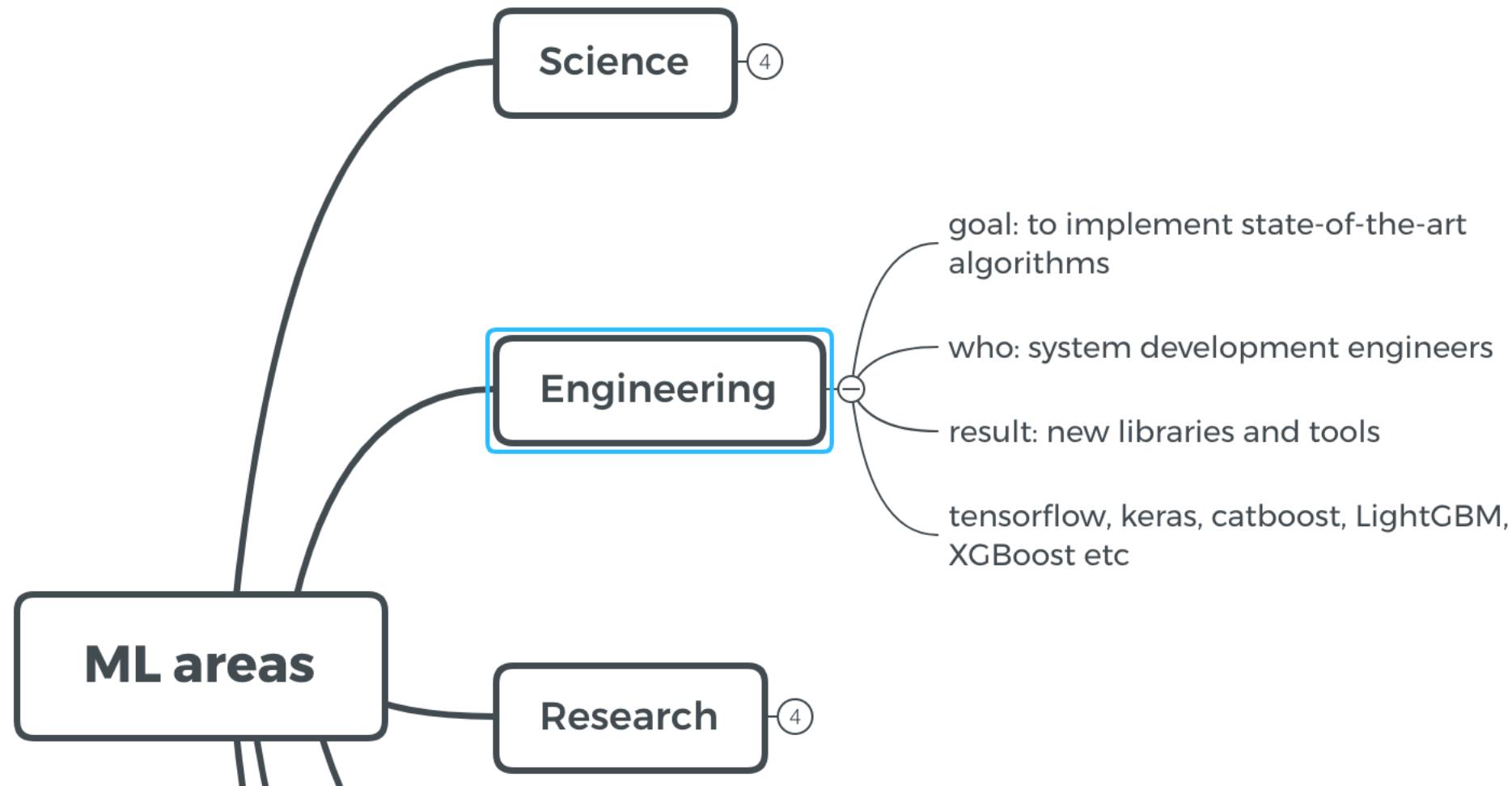
{jacobdevlin, mingweichang, kentonl, kristout}@google.com

### Abstract

We introduce a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

# Области применения



# Области применения

Home > Overview of CatBoost

## Overview of CatBoost

CatBoost is a machine learning algorithm that uses gradient boosting on decision trees. It is available as an open source library.

Key features:



### Training

[Training](#)

[Training on GPU](#)

[Python train function](#)

[Cross-validation](#)

[Overfitting detector](#)

[Pre-trained data](#)

[Categorical features](#)

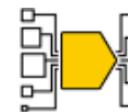
[Text features](#)



### Model analysis

[Feature importances](#)

[Object importances](#)



### Applying models

[Regular prediction](#)

[C and C++](#)

[Java](#)

[Rust](#)

[Calculate metrics](#)

[Staged prediction](#)

[Applying the model in ClickHouse](#)



### Metrics

[Implemented metrics](#)

[User-defined metrics](#)

# Области применения



FEATURES    DOC

# Open-source Version Control System for Machine Learning Projects



Download  
(Mac OS)



Watch video  
How it works

# Области применения



great\_expectations



Greetings! Have any questions about using Great

# Welcome to Great Expectations

Always know what to expect from your data

Great Expectations helps data teams eliminate pipeline debt, through data testing, documentation, and profiling.

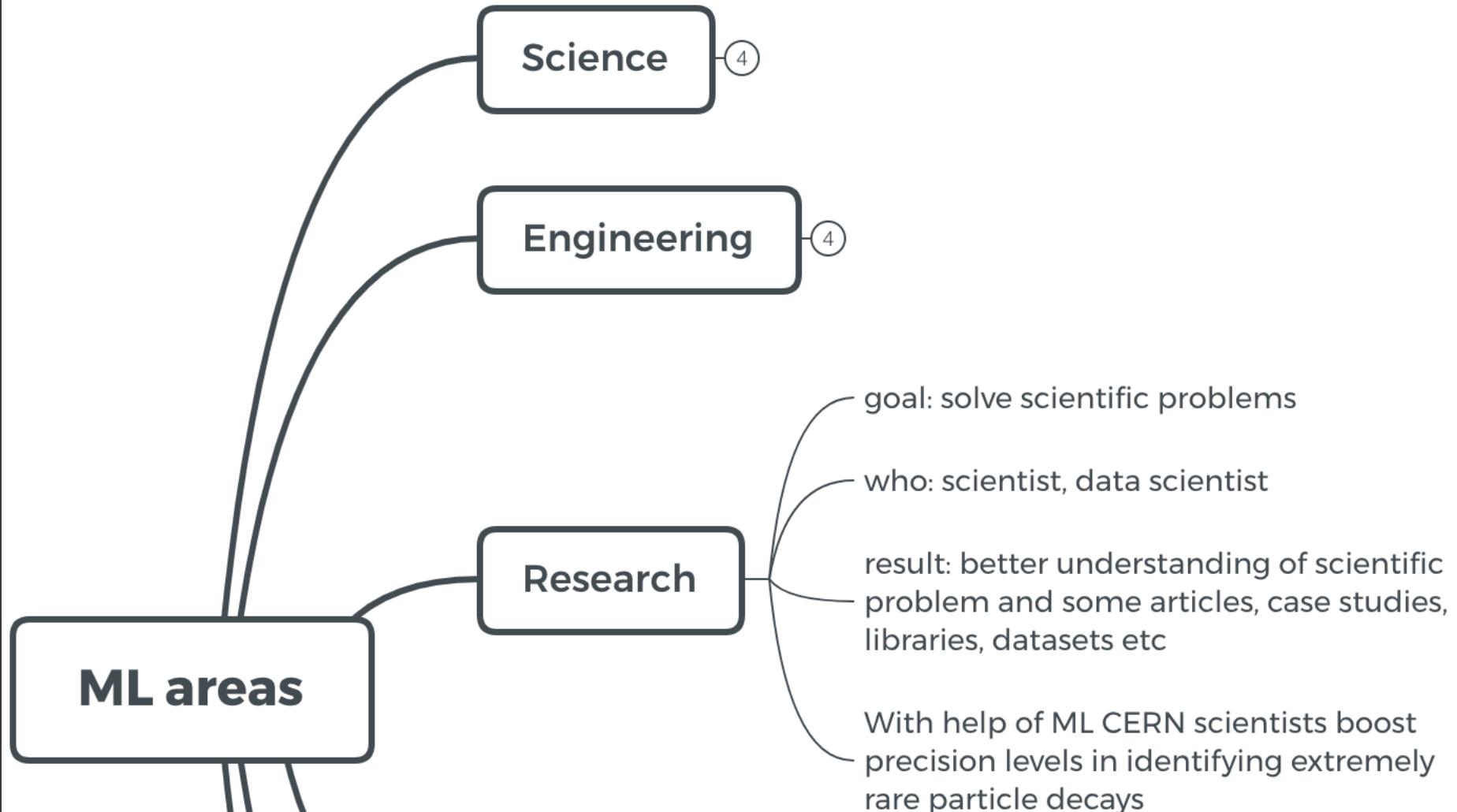


 Join us on GitHub

 2680

We're open source. Get involved!

# Области применения

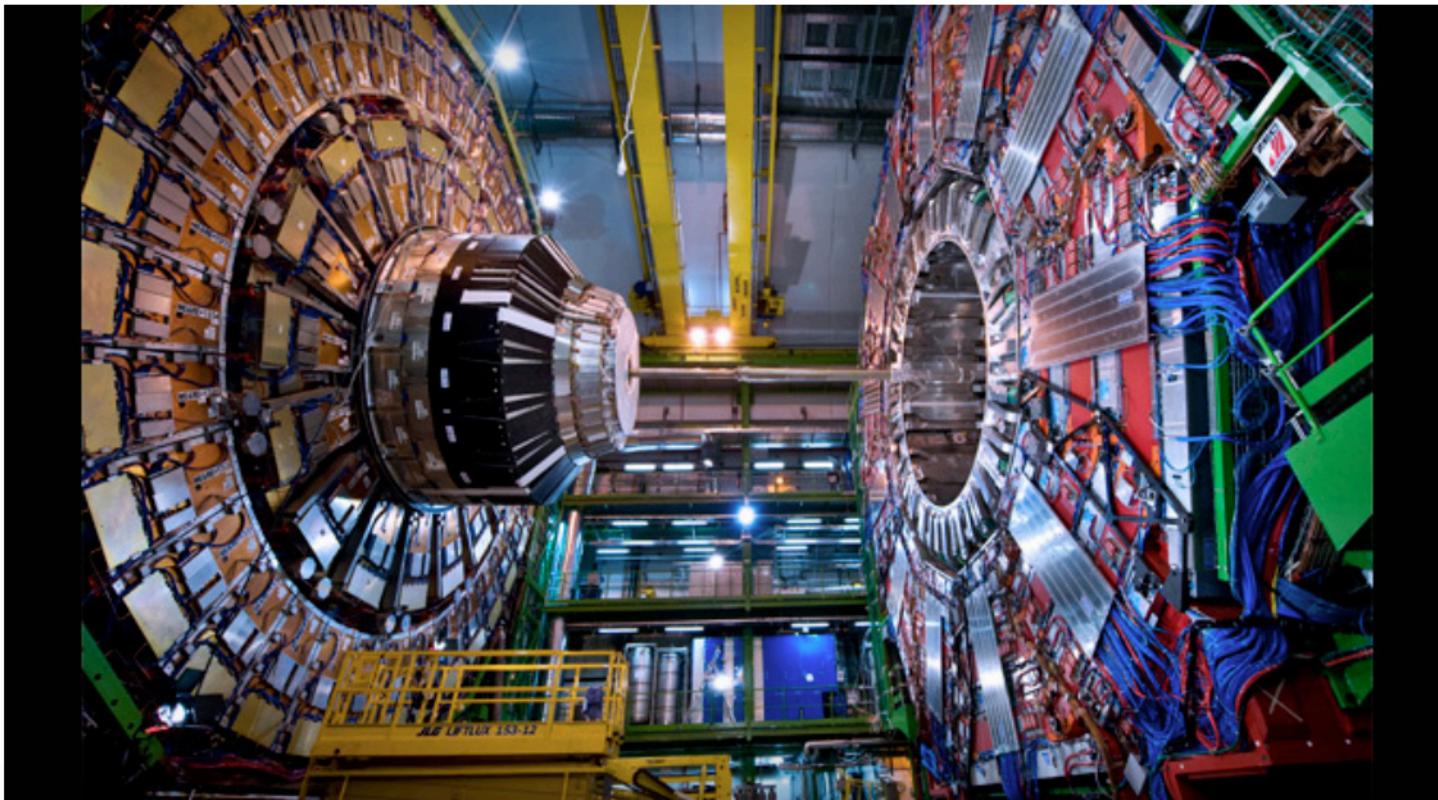


# CERN boosts its search for antimatter with Yandex's MatrixNet search engine tech

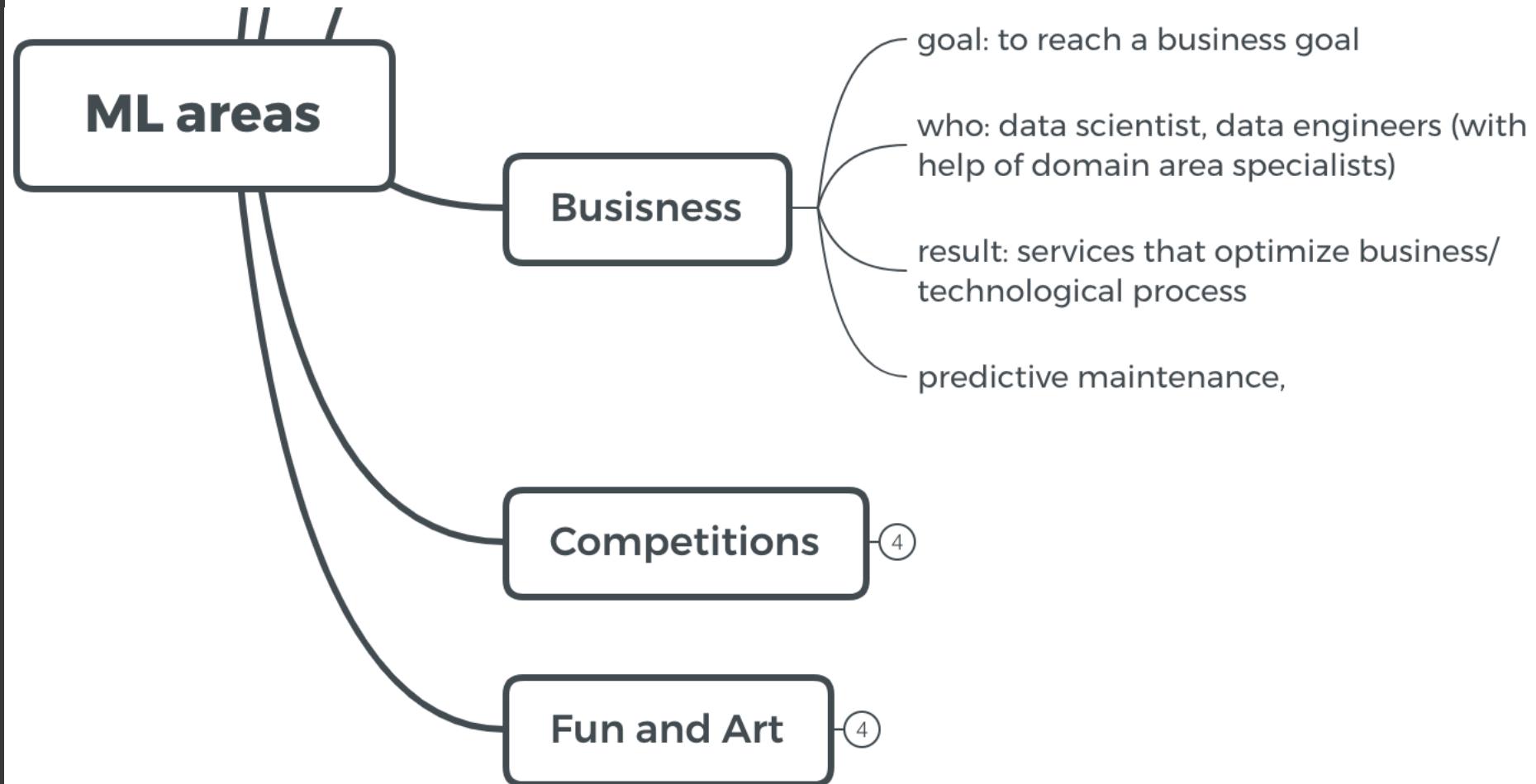
By Tim Verry on February 1, 2013 at 8:36 am | [5 Comments](#)



## Области применения



# Области применения



# Области применения

# Google



Search Google or type a URL



# Области применения

NETFLIX

Home Characters TV Shows Movies Latest My List

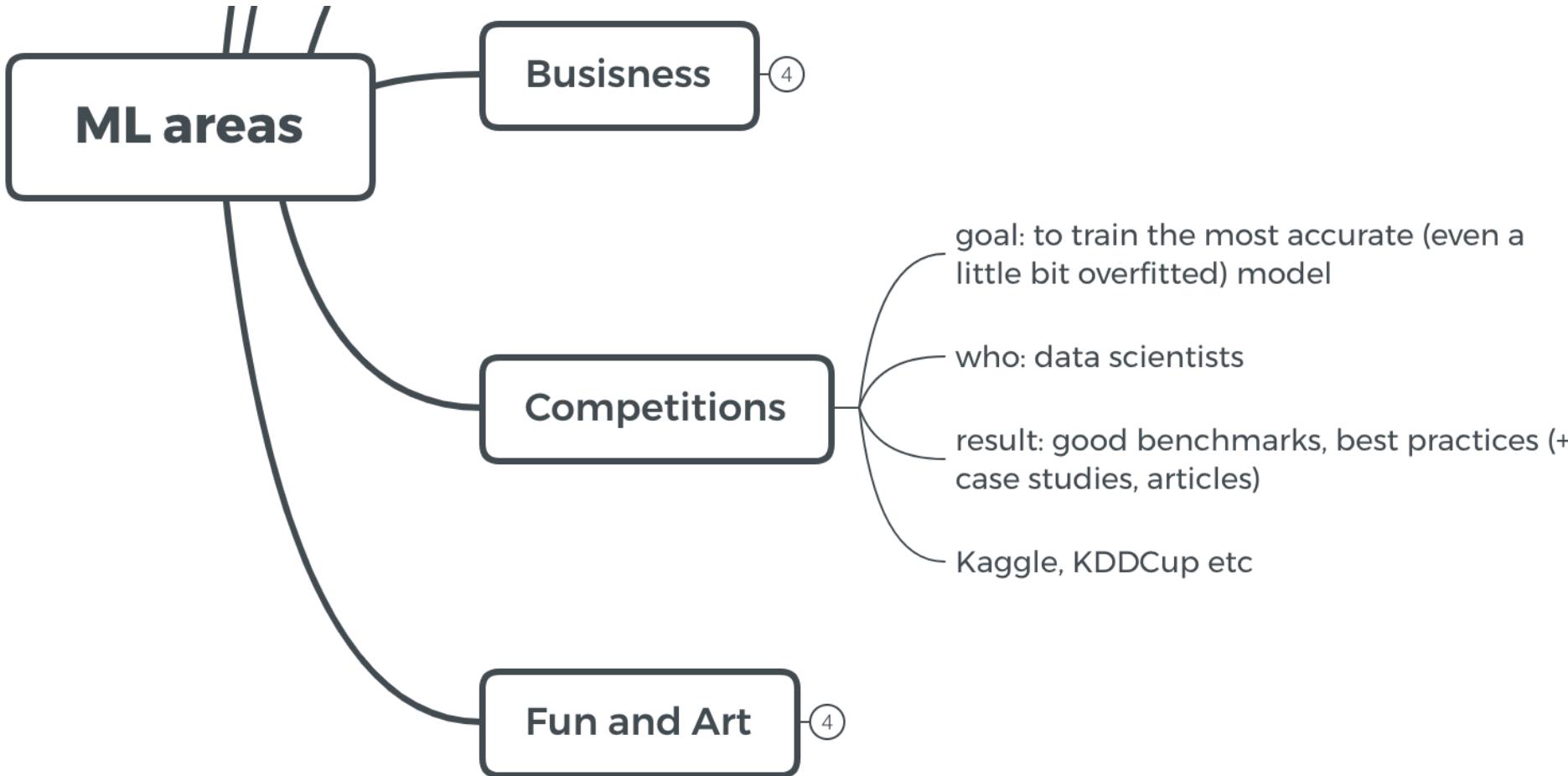
Everyone's Watching



Animated



# Области применения



# Области применения

# Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [Documentation](#) or learn about [InClass competitions](#).



## New to Kaggle? Start here!

Our Titanic Competition is a great first challenge to get started.



### Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics  
Getting Started • Ongoing • 18123 Teams

## All Competitions

[Active](#)   [Completed](#)   [InClass](#)



### OSIC Pulmonary Fibrosis Progression

Predict lung function decline  
Featured • 11d to go • Code Competition • 1939 Teams



# Области применения

## Boosters.pro - сила данных

Крупнейшая в России и Восточной Европе платформа для проведения кейс-контестов по анализу данных.

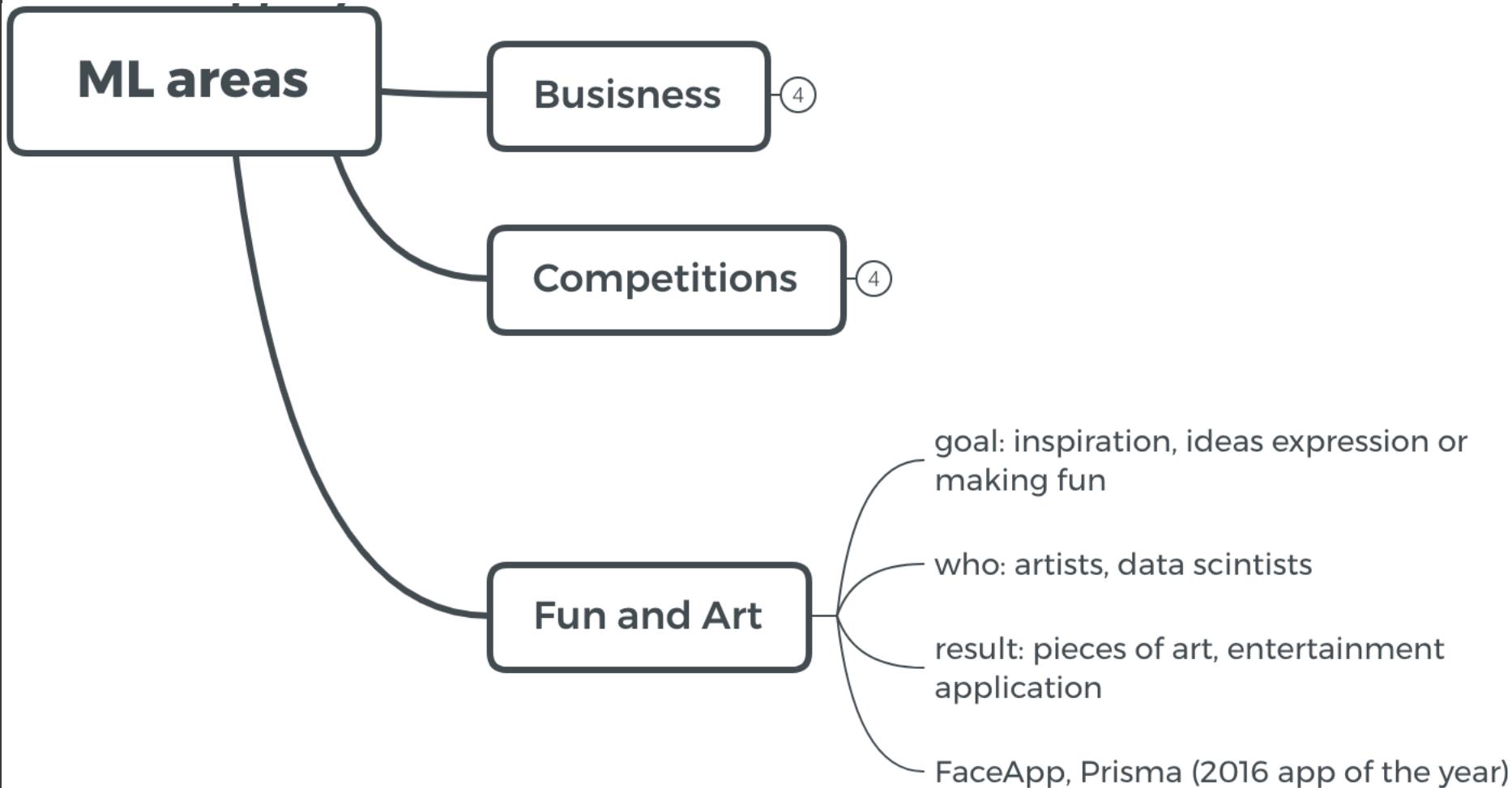
**15**  
проведённых  
контестов

**1700**  
активных  
пользователей

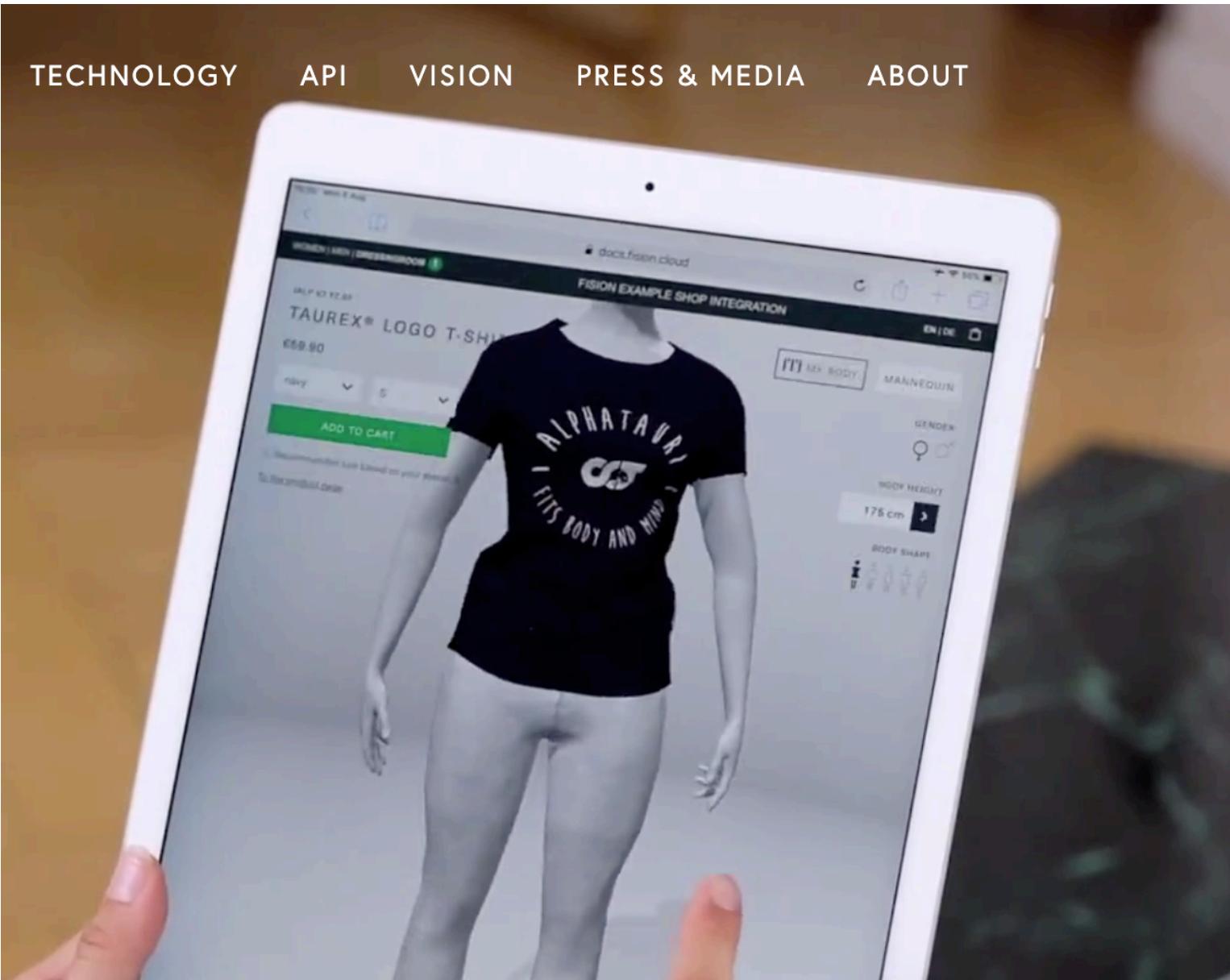
**7200**  
зарегистрированных  
пользователей

Зарегистрироваться

# Области применения



# Области применения



# Области применения



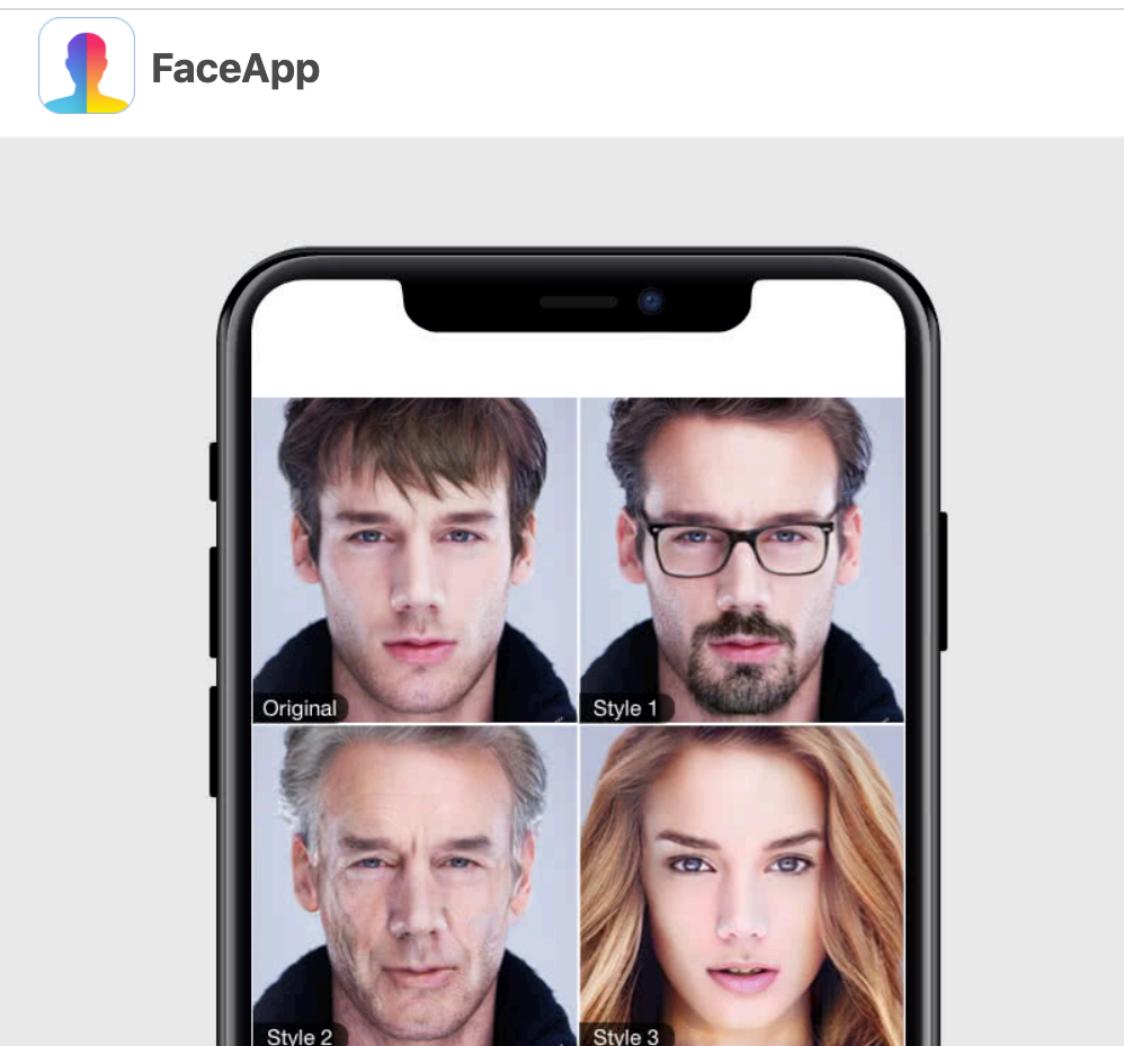
APP OF  
THE YEAR 2016  
APP STORE



BEST APP OF 2016  
GOOGLE PLAY



# Области применения



App Store  
**BEST OF 2017**

Google Play  
BEST OF 2017 AWARD

# Области применения

## To take away:

1. Работа специалистов по анализу данных очень сильно различаются в разных областях: задачи, навыки, инструменты разные.
2. Каждый специалист может выбрать подходящую сферу в зависимости от того, чем действительно интересно заниматься!
3. С другой стороны изучать ML можно очень по-разному: участие в соревнованиях, pet projects, исследовательские группы, стажировки.

# Области применения

## To take away:

1. Работа специалистов по анализу данных очень сильно различаются в разных областях: задачи, навыки, инструменты разные.
2. Каждый специалист может выбрать подходящую сферу в зависимости от того, чем действительно интересно заниматься!
3. С другой стороны изучать ML можно очень по-разному: участие в соревнованиях, pet projects, исследовательские группы, стажировки.

В нашем курсе мы фокусируемся на **индустриальном применении** машинного обучения.

# Базовые концепты

## Базовые концепты

# Объекты и ответы

- $x$  – объект
- $y$  – ответ или значение

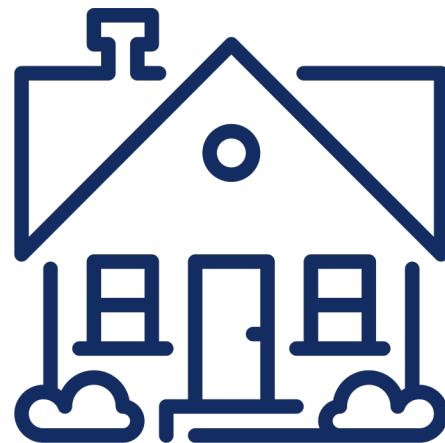
- $X$  – множество объектов
- $Y$  – множество ответов

Базовые  
концепты

## Объекты и ответы



250k \$



375k \$



179k \$

## Базовые концепты

# Признаки объектов (features)

- $f_1 \dots f_n$  – признаки, описывающие объект
- $x = (f_1, f_2, \dots, f_n)$
- $x$  - вектор размера  $n$ , описывающий объект с помощью признаков

## Базовые концепты

# Признаки объектов (features)



250k \$

- 270m<sup>2</sup>
- 2 спальни
- 2 ванные
- 1 парковочное место

...



375k \$

- 420m<sup>2</sup>
- 4 спальни
- 3 ванные
- 2 парковочных места

...



179k \$

- 120m<sup>2</sup>
- 1 спальни
- 1 ванные
- 0 парковочных место

...

## Базовые концепты

# Выборка (dataset)

- $X = (x_i, y_i)_{i=1,l}$
- $X$  – выборка

- $(x_i, y_i)$  – элемент выборки, пара (объект, ответ)
- $x_i = (f^1_i, f^2_i, \dots, f^n_i)$
- $f^k_i$  – значение признака  $k$  на объекте  $i$

## Базовые концепты

# Выборка (dataset)

Количество ванных комнат	Количество спален	Стоимость
Площадь		
2	2	250k \$
420m <sup>2</sup>	3	375k \$
120m <sup>2</sup>	1	179k \$

## Базовые концепты

# Обучающая выборка

Количество ванных комнат	Количество спален	Стоимость
Площадь		
2	2	250k \$
420m <sup>2</sup>	3	375k \$
120m <sup>2</sup>	1	179k \$

## Базовые концепты

# Тестовая выборка

	Площадь	Количество ванных комнат	Количество спален	Стоимость
	270m <sup>2</sup>	2	2	250k \$
	420m <sup>2</sup>	3	4	375k \$
	120m <sup>2</sup>	1	1	179k \$

# Базовые концепты

## Модель

- $a: X \rightarrow Y$
- $a(x) = y$
- $A$  – семейство моделей

# ФУНКЦИЯ ПОТЕРЬ

- $Q(a, X)$  – ошибки модели  $a(x)$  на выборке  $X$

Базовые  
концепты

# Базовые концепты

## Базовые концепты

Объекты и признаки:

- $x$  – объект
- $y$  – ответ
- $(f_1, f_2 \dots f_n)$  – признаки, описывающие объекты

Модель:

- $a: X \rightarrow Y$
- $a(x) = y$
- $A$  – семейство моделей

- $X$  – пространство объектов
- $Y$  – пространство ответов

Оценка качества

- $Q(a, X)$  – ошибки модели  $a(x)$  на группе объектов  $X$

## Базовые концепты

# Логика построения модели

1. Определяем объекты
2. Формулируем задачу:  
на какой вопрос касательно объектов мы хотим  
ответить?
3. Определяем признаки, собираем выборку
4. Строим модель
5. Оцениваем её качество

# Виды обучения

# Виды обучения

## Классификация постановок задач

- По типу задач
- По виду обучения
- По алгоритмам
- По области применения
- Смешенная

# Виды обучения

## По типу задач

- Обучение с учителем/Supervised learning
- Обучение без учителя/Unsupervised learning
- Частичное обучение/Semi-supervised learning
- Обучение с подкреплением/ Reinforcement learning

# Виды обучения

## По виду обучения

- Classic learning
- Active learning
- Online learning
- Transfer learning
- ...

# Виды обучения

## По алгоритмам

- Ensemble learning
- Deep learning
- Bayesian learning
- ...

# Виды обучения

## По данным и задачам

- Tabular data
- Timeseries data
- Computer vision
- Natural language processing
- Cognitive technologies
- Recommender systems
- Learning to rank
- ...

# Виды обучения

## Смешенная (особенно грустно)

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning
- Deep learning
- Active learning
- Online learning
- Transfer learning
- ... you name it =(

# Виды обучения

## По типу задач



- Обучение с учителем/Supervised learning
- Обучение без учителя/Unsupervised learning
- Частичное обучение/Semi-supervised learning
- Обучение с подкреплением/ Reinforcement learning

В курсе мы фокусируемся на **обучении с учителем**, но так или иначе затронем все виды обучения.

# Постановка задач обучения

# Задачи машииного обучения

## Обучение с учителем

- Задача – найти верный ответ для каждого объекта:
  - метку класса или вероятность в случае задачи классификации
  - численное значение в случае регрессии
- Нам доступны верные ответы на достаточно большой группе объектов для обучения
- Мы учим модель находить закономерности между значениями признаков и ответами

# Задачи машииного обучения

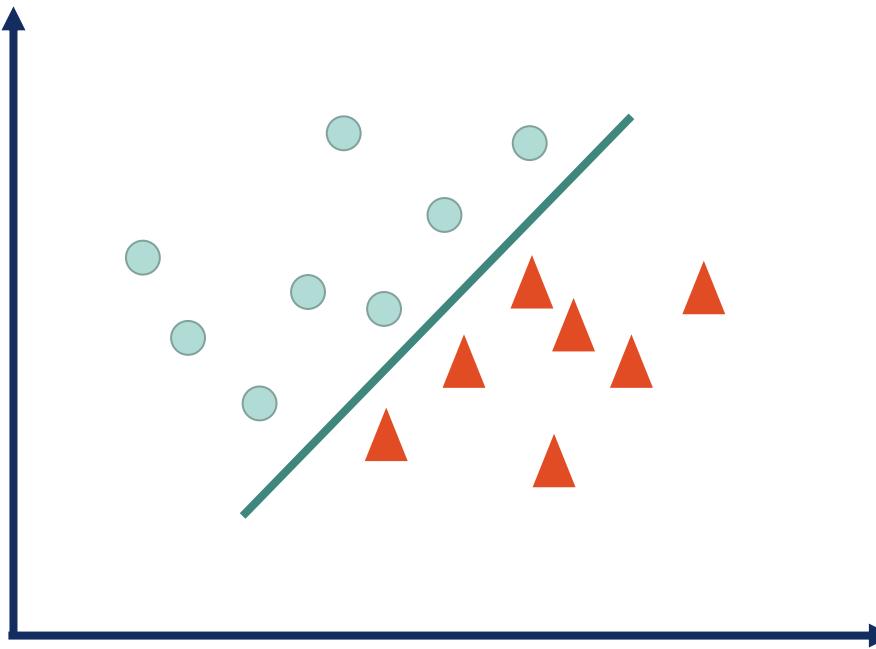
## Обучение с учителем

- После обучения мы применяем обученную модель к внешним данным – данным, которые не входили в обучающую выборку
- Объекты из тестовой выборки могут существенно отличаться от объектов в обучающей выборке
- Чем больше верных ответов мы получаем (особенно в тестовой выборке) – тем лучше обучена модель

# Задачи машииного обучения

## Обучение с учителем

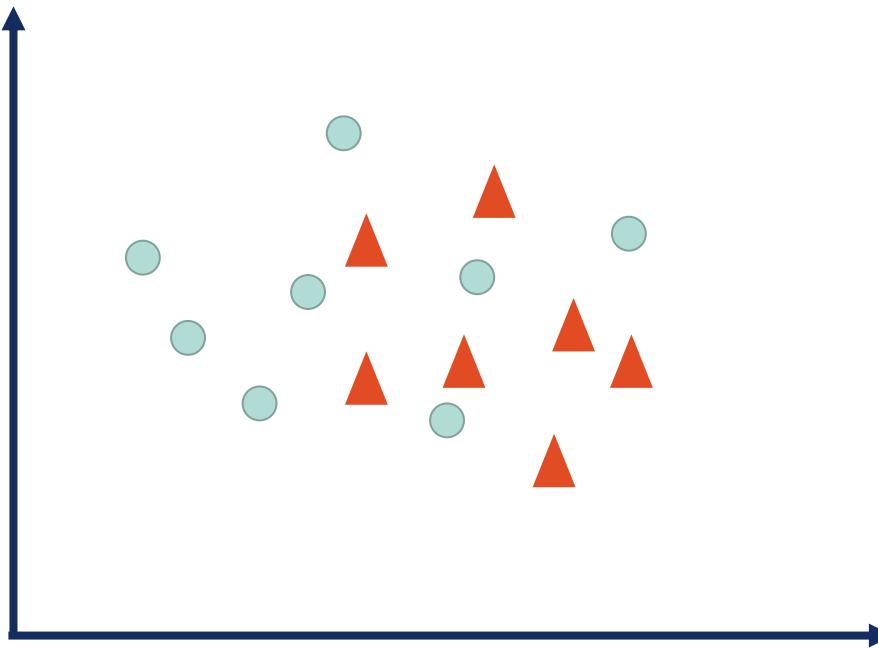
### Классификация



# Обучение с учителем

Что если задачу решить не получается?

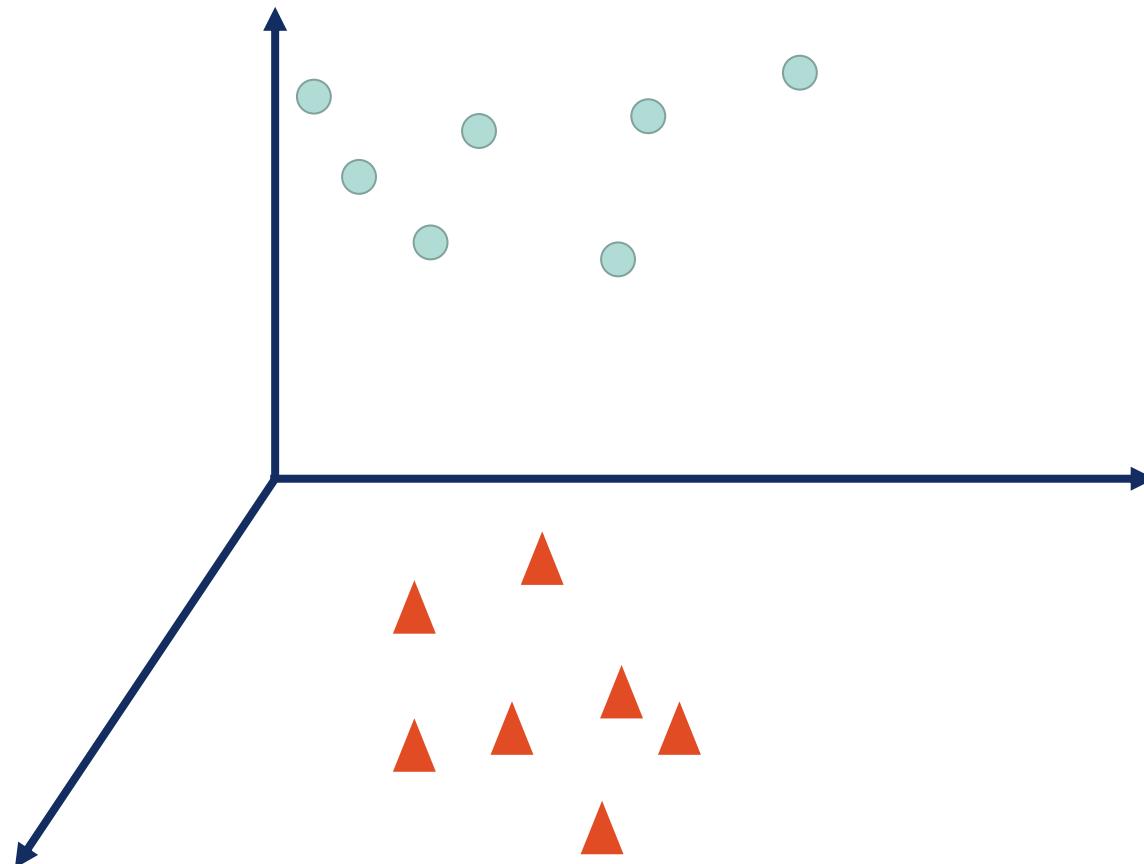
Задачи  
машииного  
обучения



# Задачи машииного обучения

## Обучение с учителем

Что если задачу решить не получается?



# Задачи машинного обучения

## Обучение с учителем

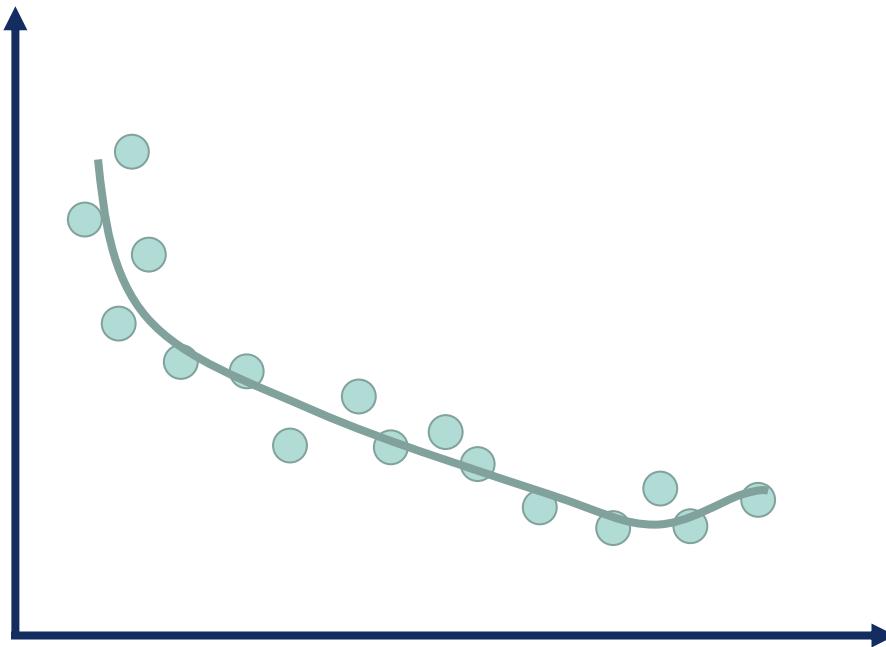
Регрессия



# Задачи машинного обучения

## Обучение с учителем

Регрессия



# Задачи машииного обучения

## Обучение с учителем

### Ранжирование

#### [Learning to rank - Wikipedia](#)

[https://en.wikipedia.org/wiki/Learning\\_to\\_rank](https://en.wikipedia.org/wiki/Learning_to_rank) ▾

Learning to rank or machine-learned ranking (MLR) is the application of machine learning, .... Often a learning-to-rank problem is reformulated as an optimization problem with respect to one of these metrics. Examples of ranking quality ...

[Applications](#) · [Feature vectors](#) · [Evaluation measures](#) · [Approaches](#)

#### [Ranking \(information retrieval\) - Wikipedia](#)

[https://en.wikipedia.org/wiki/Ranking\\_\(information\\_retrieval\)](https://en.wikipedia.org/wiki/Ranking_(information_retrieval)) ▾

Ranking of query results is one of the fundamental **problems** in information retrieval (IR), the scientific/engineering discipline behind search engines. Given a ...

#### [\[PDF\] Statistical Ranking Problem](#)

<https://web.stanford.edu/group/mmds/slides/zhang-mmds.pdf> ▾

Statistical Ranking Problem. Tong Zhang. Yahoo! Inc. New York City. Joint work with. David Cossack. Yahoo! Inc. Santa Clara ...

#### [Problem & Preference Ranking | SSWM](#)

[www.sswm.info/content/problem-preference-ranking](http://www.sswm.info/content/problem-preference-ranking) ▾

Problem/Preference Ranking is a participatory technique that allows analysing and identifying problems or preferences stakeholder share in order to implement ...

# Задачи машииного обучения

## Обучение без учителя

Задача – разделить объекты на группы (кластеры) таким образом, чтобы:

- группы соответствовали исходной структуре данных
- объекты внутри одной группы были схожи
- объекты из разных группы существенно различались

# Задачи машинного обучения

## Обучение без учителя

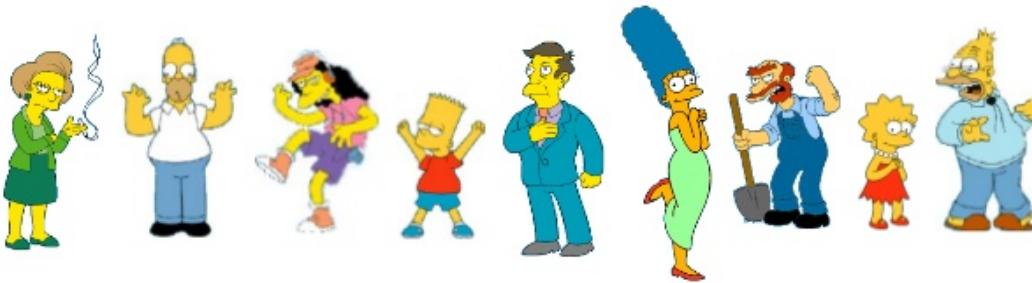
В чем сложность:

- нам не известна исходная структура данных
- ответы для обучения не доступны
- не известно даже количество групп

# Задачи машинного обучения

## Кластеризация

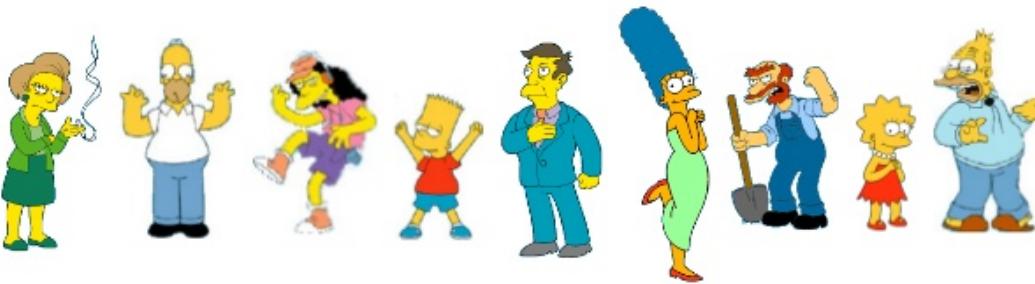
What is a natural grouping among these objects?



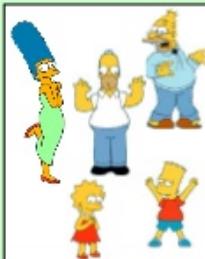
# Задачи машинного обучения

# Кластеризация

What is a natural grouping among these objects?



Clustering is subjective



Simpson's Family



School Employees



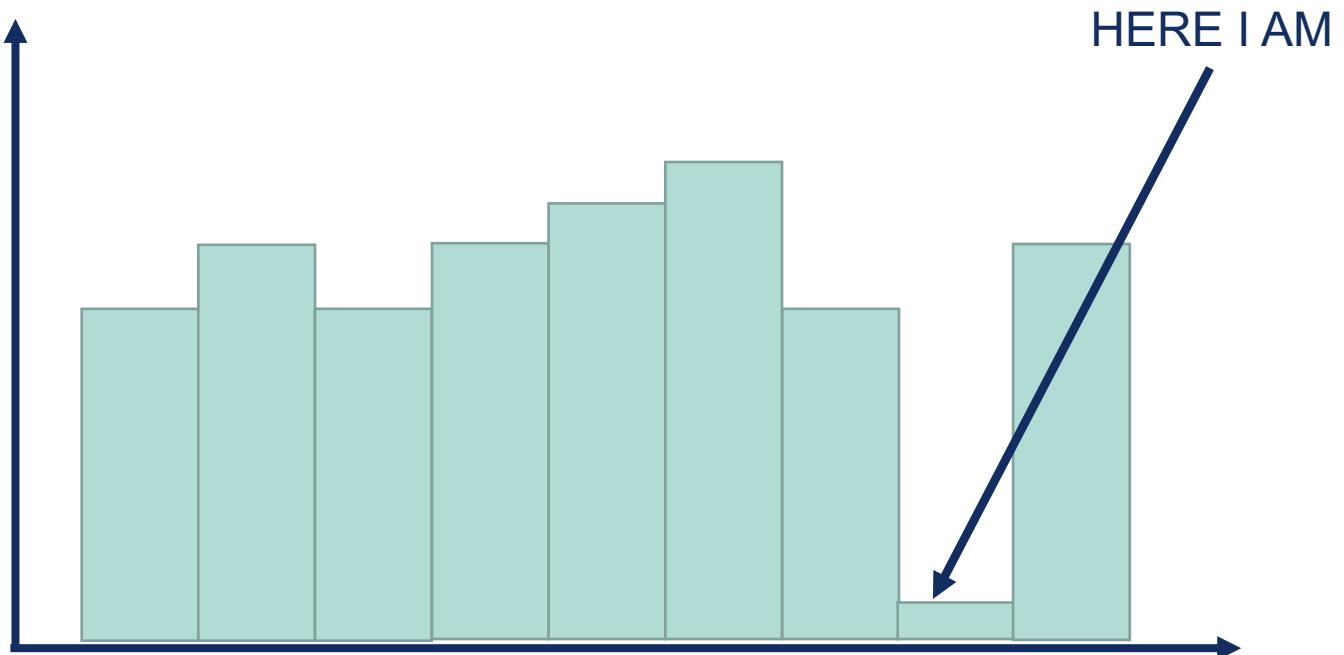
Females



Males

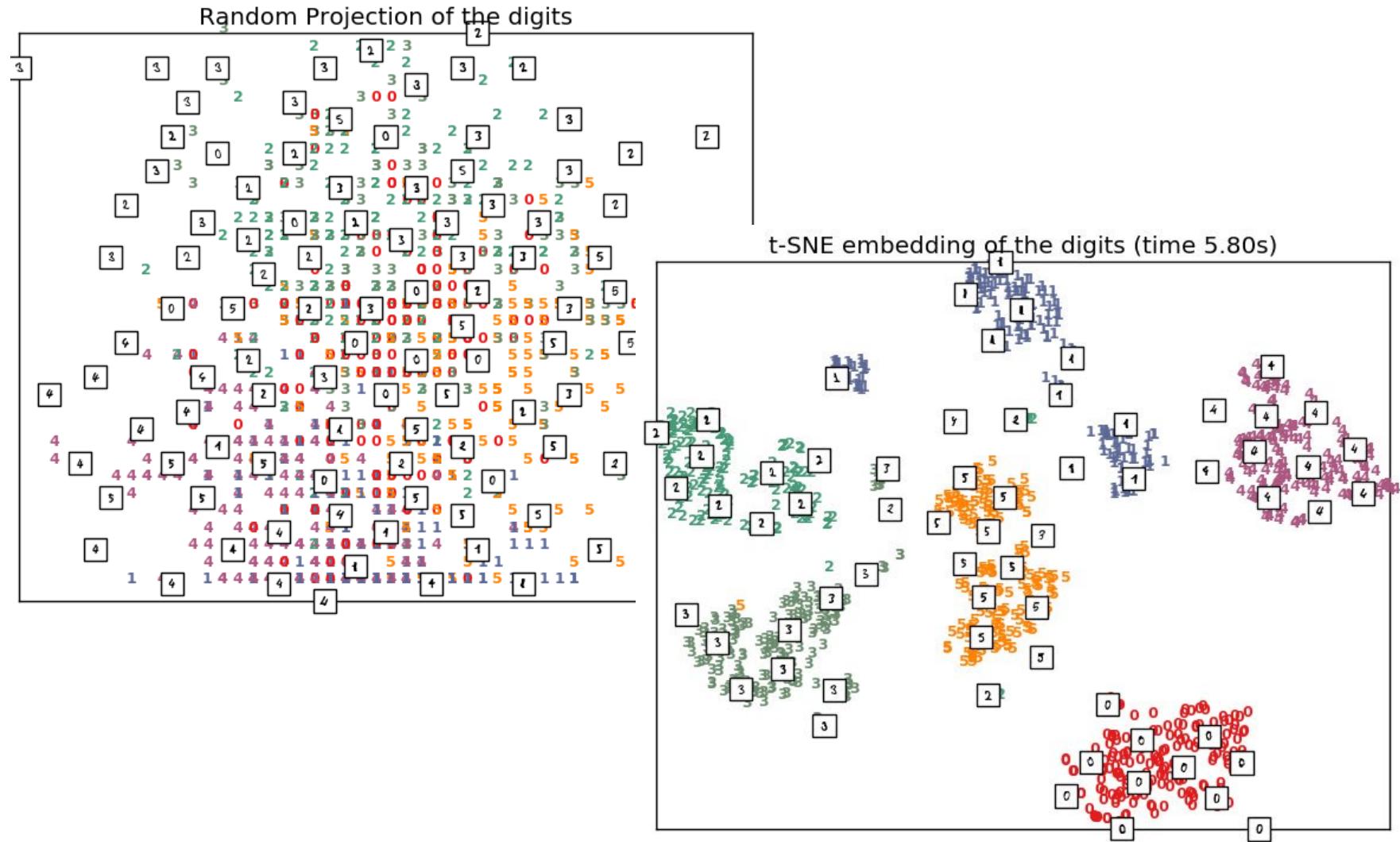
# Задачи машииного обучения

# Детектирование аномалий



# Понижение размерности

Задачи  
машииного  
обучения



# Задачи машинного обучения

## Частичное обучение

- На границе между обучением с учителем и обучением без учителя
- Нам доступны ответы на небольшой группе объектов
- Для большинства объектов ответы не доступны

# Задачи машииного обучения

## Обучения с подкреплением

- Модели доступен ограниченный набор действий
- Модель взаимодействует с динамической средой для получения обратной связи в ответ на выбранное действие
- Обратная связь: штраф или награда

# Задачи машииного обучения

## Обучения с подкреплением

- Модели доступен ограниченный набор действий
- Модель взаимодействует с динамической средой для получения обратной связи в ответ на выбранное действие
- Обратная связь: штраф или награда
- Чаще всего получение обратной связи осложненно: долго, дорого, вычислительно затратно
- Модель ограничена в получении обратной связи (в единицу времени)

# Обучение с подкреплением

- Обучение игре в шахматы, ГО
- Симуляторы движения

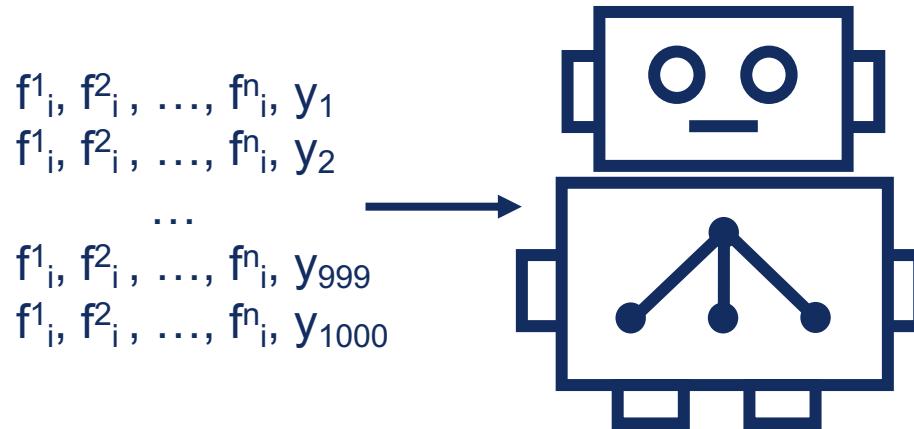
Задачи  
машинного  
обучения

# Жизненный цикл модели

# Жизненный цикл модели

## Модель

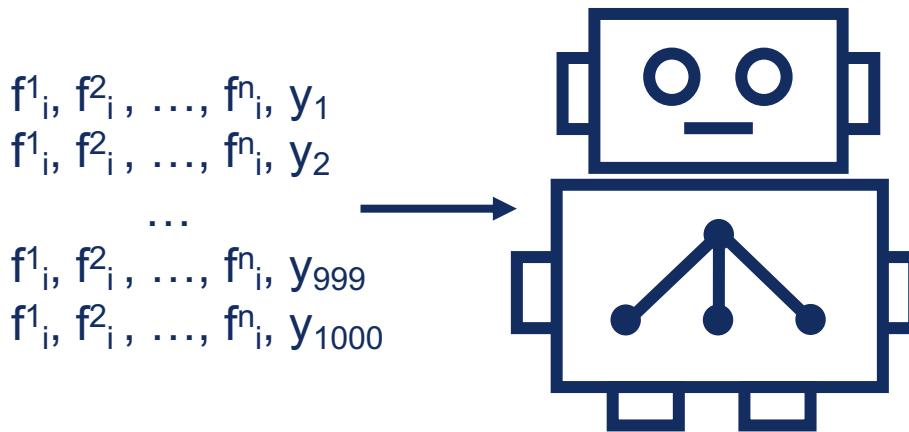
Обучение модели:



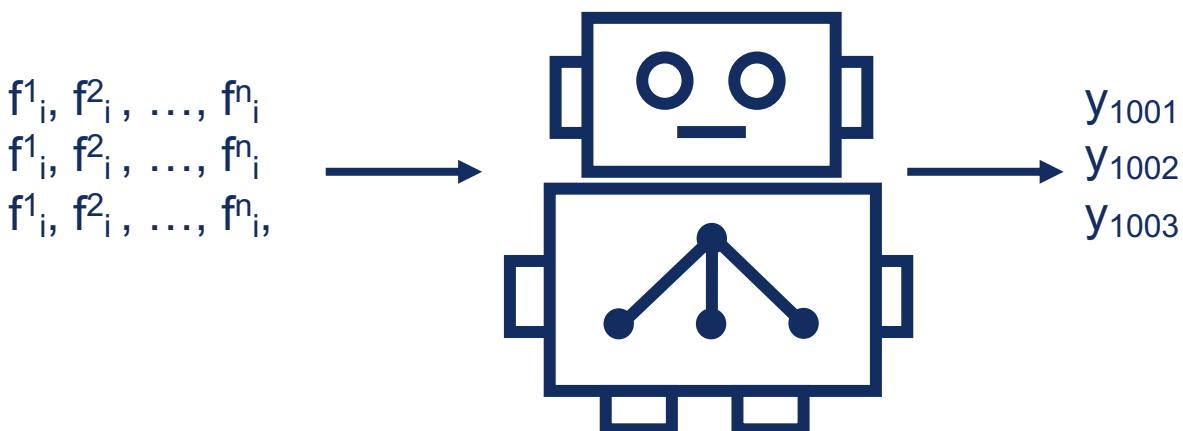
# Жизненный цикл модели

## Модель

Обучение модели:



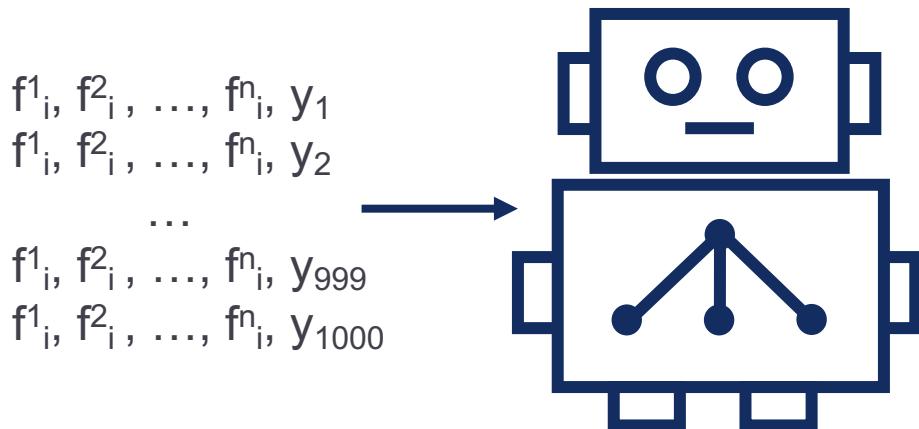
Применение модели:



# Жизненный цикл модели

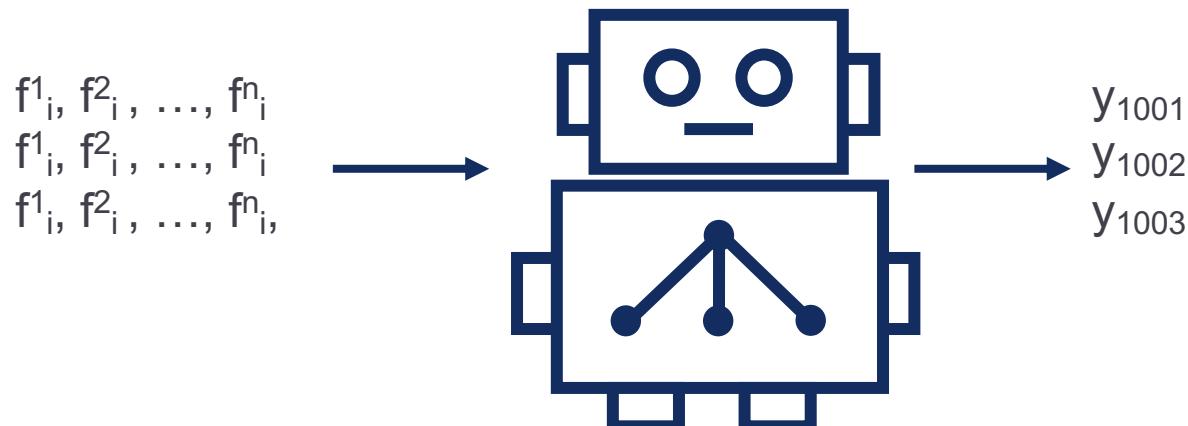
## Модель

Обучение модели:



$Q_{\text{train}}(a, X)$  – ошибки модели  $a(x)$  на обучающей выборке

Применение модели:



$Q_{\text{test}}(a, X)$  – ошибки модели  $a(x)$  на тестовой выборке

# Жизненный цикл модели

## Модель

- $Q_{\text{train}}(a, X)$  – ошибки модели  $a(x)$  на обучении
- $Q_{\text{test}}(a, X)$  – ошибки модели  $a(x)$  на teste

?

О чём говорят:

- Ошибка на обучении высокая?
- Ошибка на обучении низкая?

# Жизненный цикл модели

## Модель

- $Q_{\text{train}}(a, X)$  – ошибки модели  $a(x)$  на обучении
- $Q_{\text{test}}(a, X)$  – ошибки модели  $a(x)$  на teste

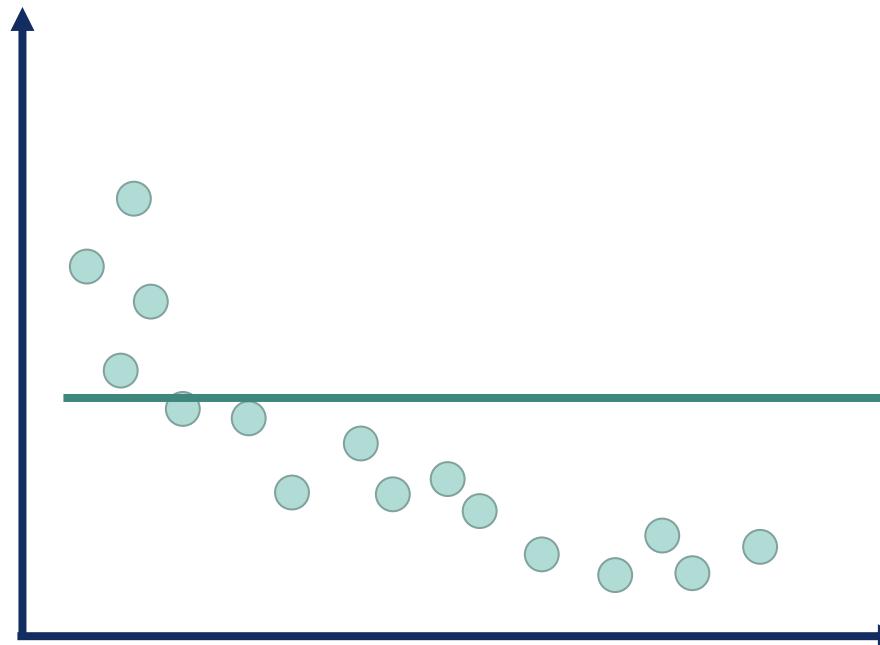
?

О чём говорят:

- Ошибка на обучении высокая?
- Ошибка на обучении низкая?
- Ошибка на обучении немного ниже, чем на teste?
- Ошибка на обучении существенно ниже, чем на teste?

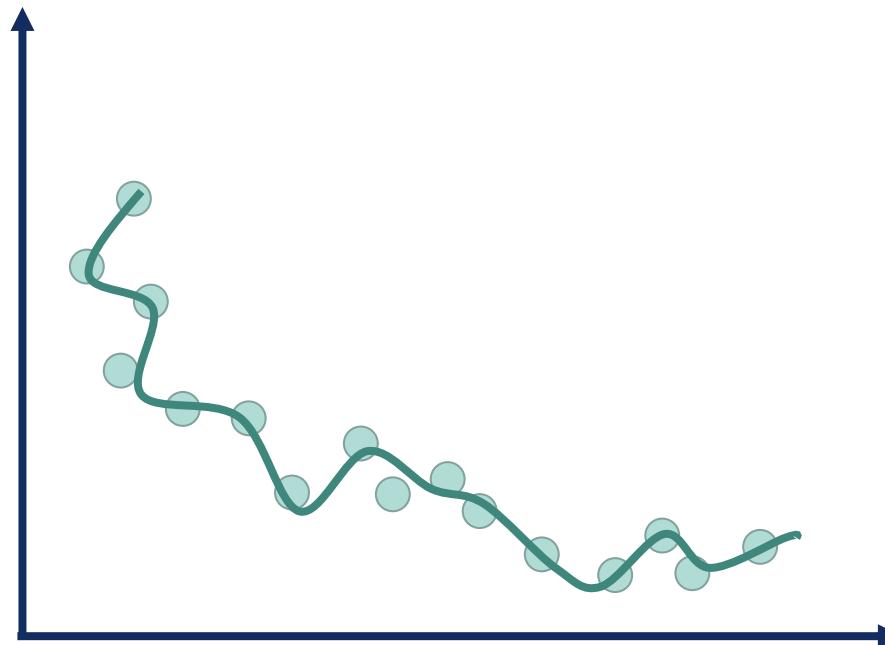
# Жизненный цикл модели

## Недообучение



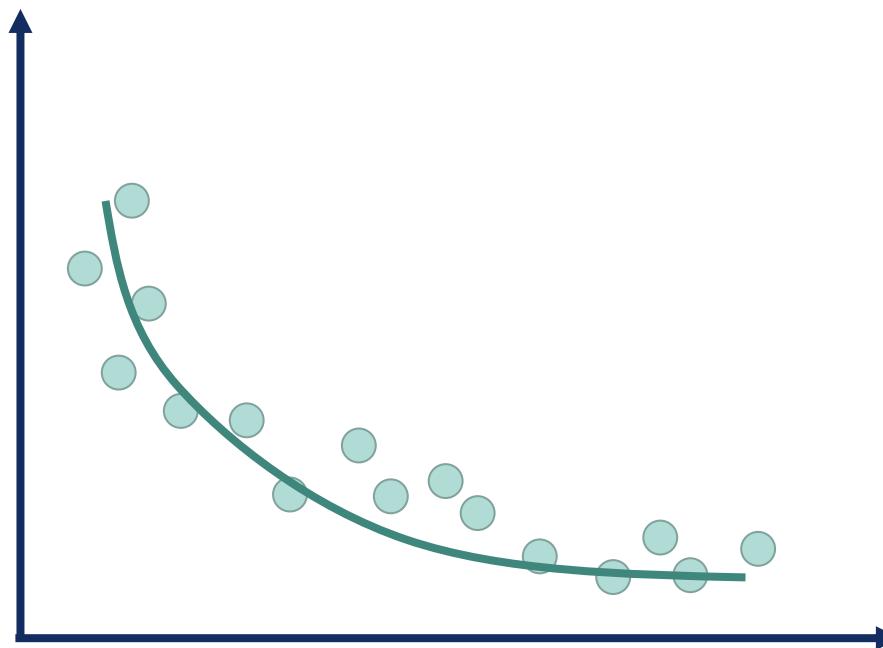
# Жизненный цикл модели

## Переобучение



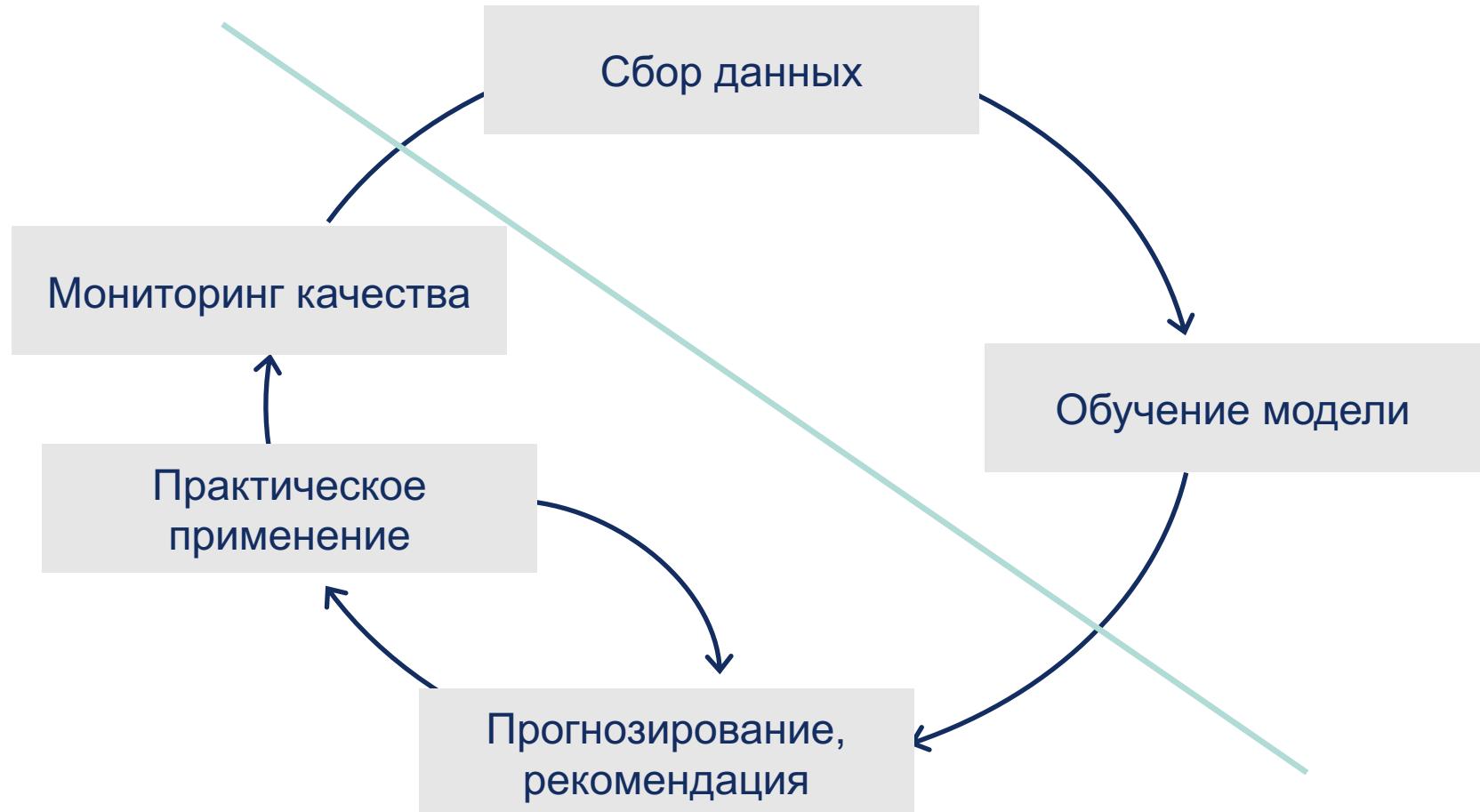
# Жизненный цикл модели

# Качественная модель



# Жизненный цикл модели

## Индустриальное применение



# Машинное обучение: базовые концепции машинного обучения

Спасибо!  
Эмели Драль