AI

ITAM
AI Lab MISIS

Лекция 1

Что такое машинное обучение?
Соревнования по машинному обучению.
Основные термины, литератера и источники.

AI - Иванов Арсений

Мой TG-канал

- 3 курс Прикладная математика
- сотрудник SberRecSys - ZVUK
- занимаюсь research’ем в SberAI

Изучаю и работаю в GNN, RecSys
Libs: Pandas, NumPy, SciPy, PyTorch, PyG, DgL

Коммуникации

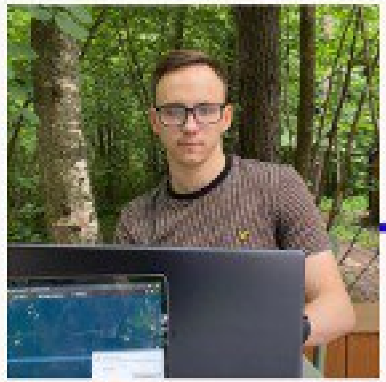
Сайт MISIS AI Lab

Чат в TG

TG-канал

Кто мы? И чем мы занимаемся?



- Изучение ML, Q/A, общение
- Data-Science чемпионаты
- Хакатоны
- Research и публикация статей

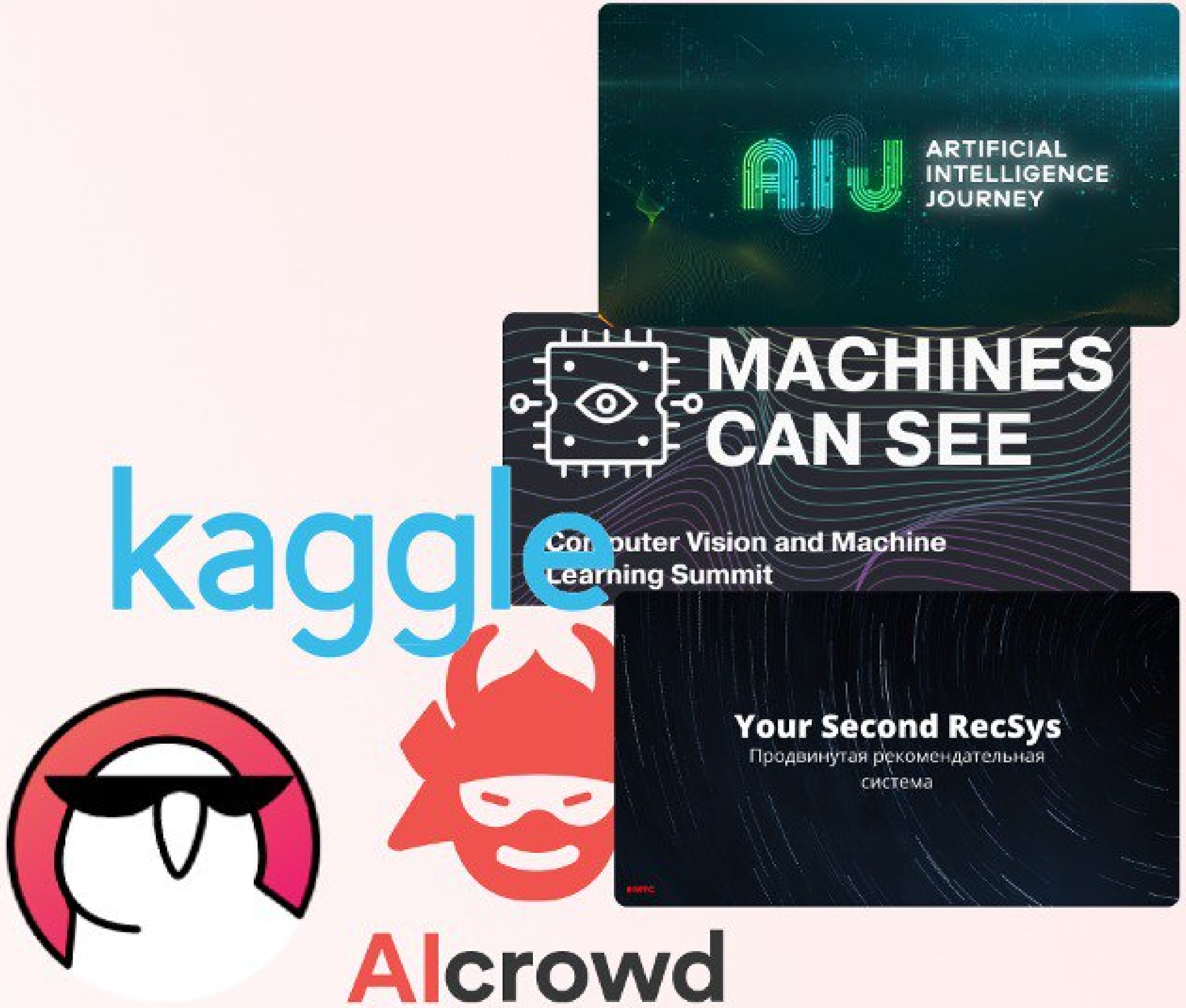
AI Lab MISIS

Artificial intelligence
Machine Learning, Data Science

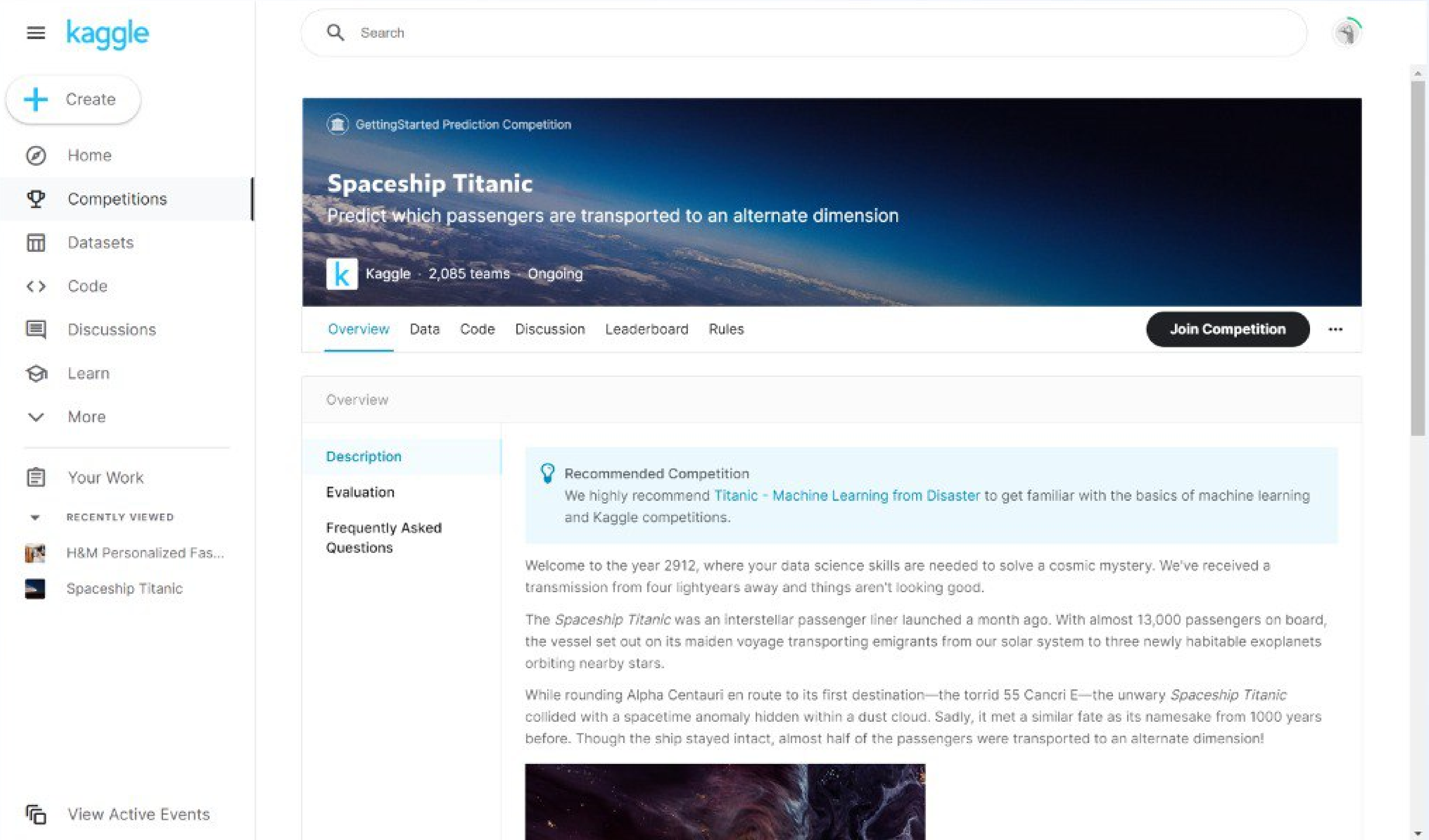
Data-Science чемпионаты

DS-competitions

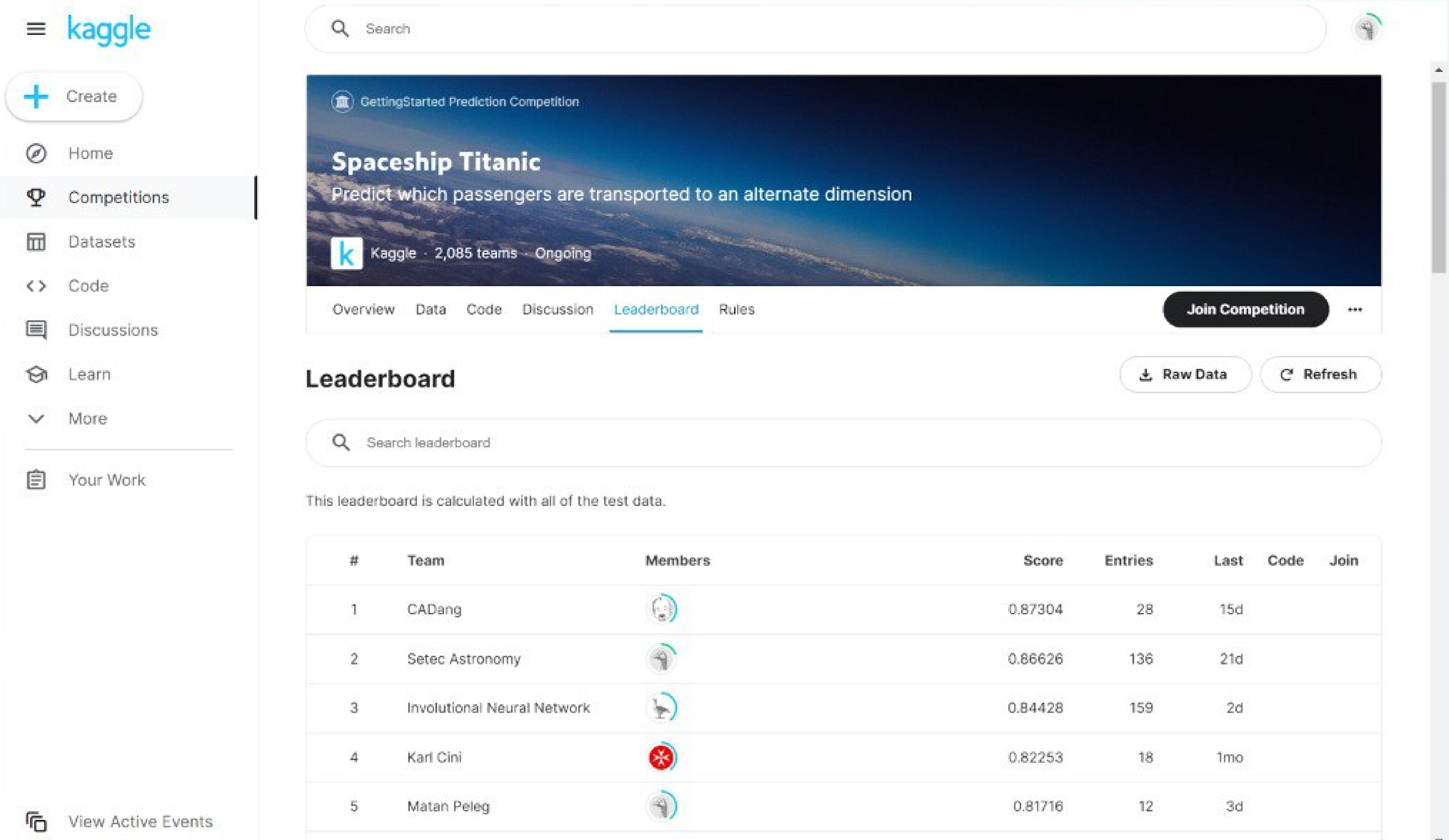
1. Работа с данными
2. Приближенные к рабочим задачи
3. Leader-board, math, ds



Kaggle

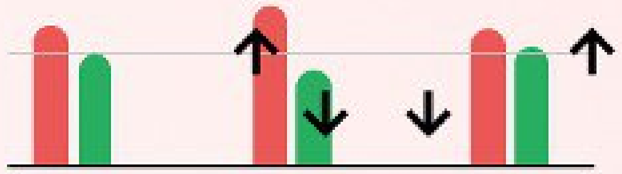

kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
RECENTLY VIEWED
H&M Personalized Fas...
Spaceship Titanic
View Active Events
GettingStarted Prediction Competition
Spaceship Titanic
Predict which passengers are transported to an alternate dimension
Kaggle · 2,085 teams · Ongoing
Overview
Data
Code
Discussion
Leaderboard
Rules
Join Competition
Overview
Description
Evaluation
Frequently Asked Questions
Recommended Competition
We highly recommend Titanic - Machine Learning from Disaster to get familiar with the basics of machine learning and Kaggle competitions.
Welcome to the year 2912, where your data science skills are needed to solve a cosmic mystery. We've received a transmission from four lightyears away and things aren't looking good.
The Spaceship Titanic was an interstellar passenger liner launched a month ago. With almost 13,000 passengers on board, the vessel set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars.
While rounding Alpha Centauri en route to its first destination—the torrid 55 Cancri E—the unwary Spaceship Titanic collided with a spacetime anomaly hidden within a dust cloud. Sadly, it met a similar fate as its namesake from 1000 years before. Though the ship stayed intact, almost half of the passengers were transported to an alternate dimension!

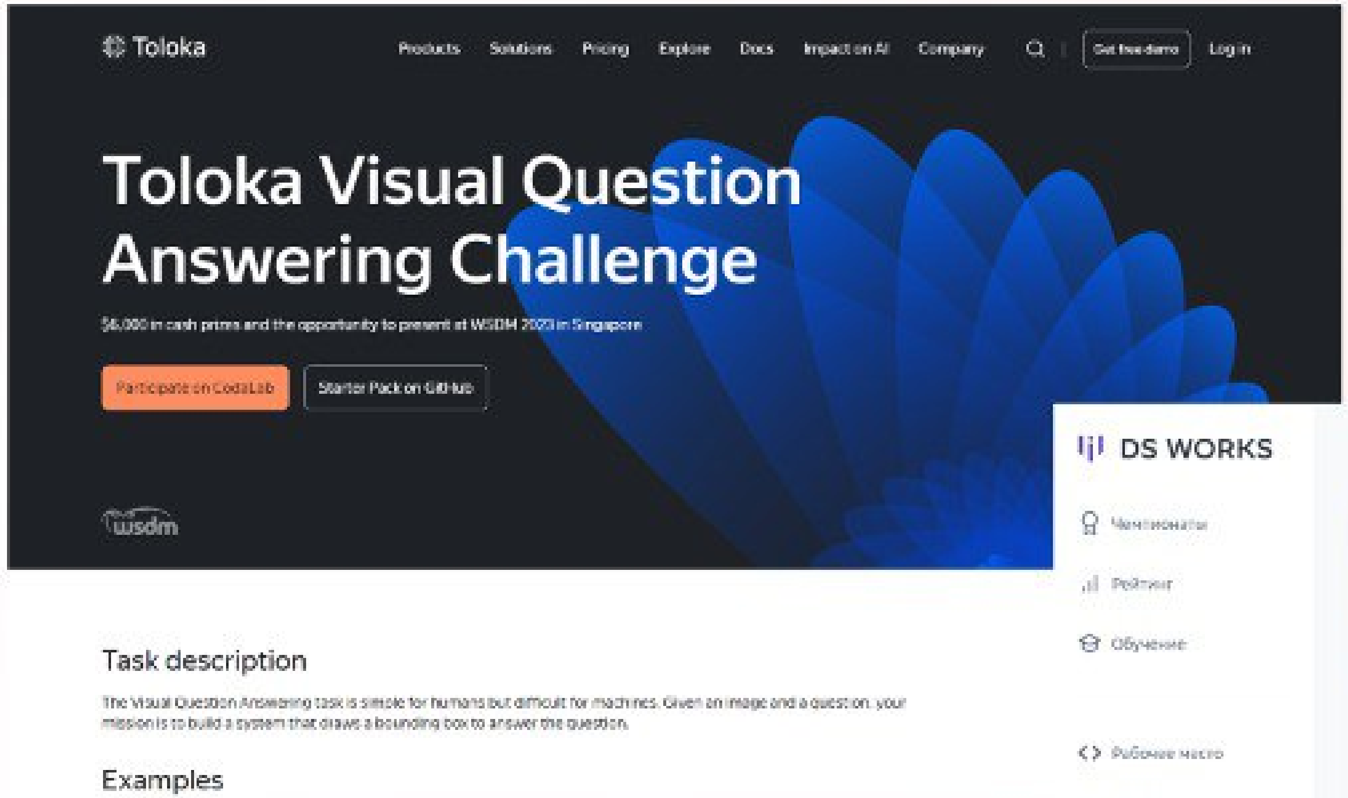
Kaggle


kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
View Active Events
GettingStarted Prediction Competition
Spaceship Titanic
Predict which passengers are transported to an alternate dimension
Kaggle · 2,085 teams · Ongoing
Overview
Data
Code
Discussion
Leaderboard
Rules
Join Competition
Leaderboard
Raw Data
Refresh
Search leaderboard
This leaderboard is calculated with all of the test data.
Team Members Score Entries Last Code Join
1 CADang 0.87304 28 15d
2 Setec Astronomy 0.86626 136 21d
3 Involutional Neural Network 0.84428 159 2d
4 Karl Cini 0.82253 18 1mo
5 Matan Peleg 0.81716 12 3d

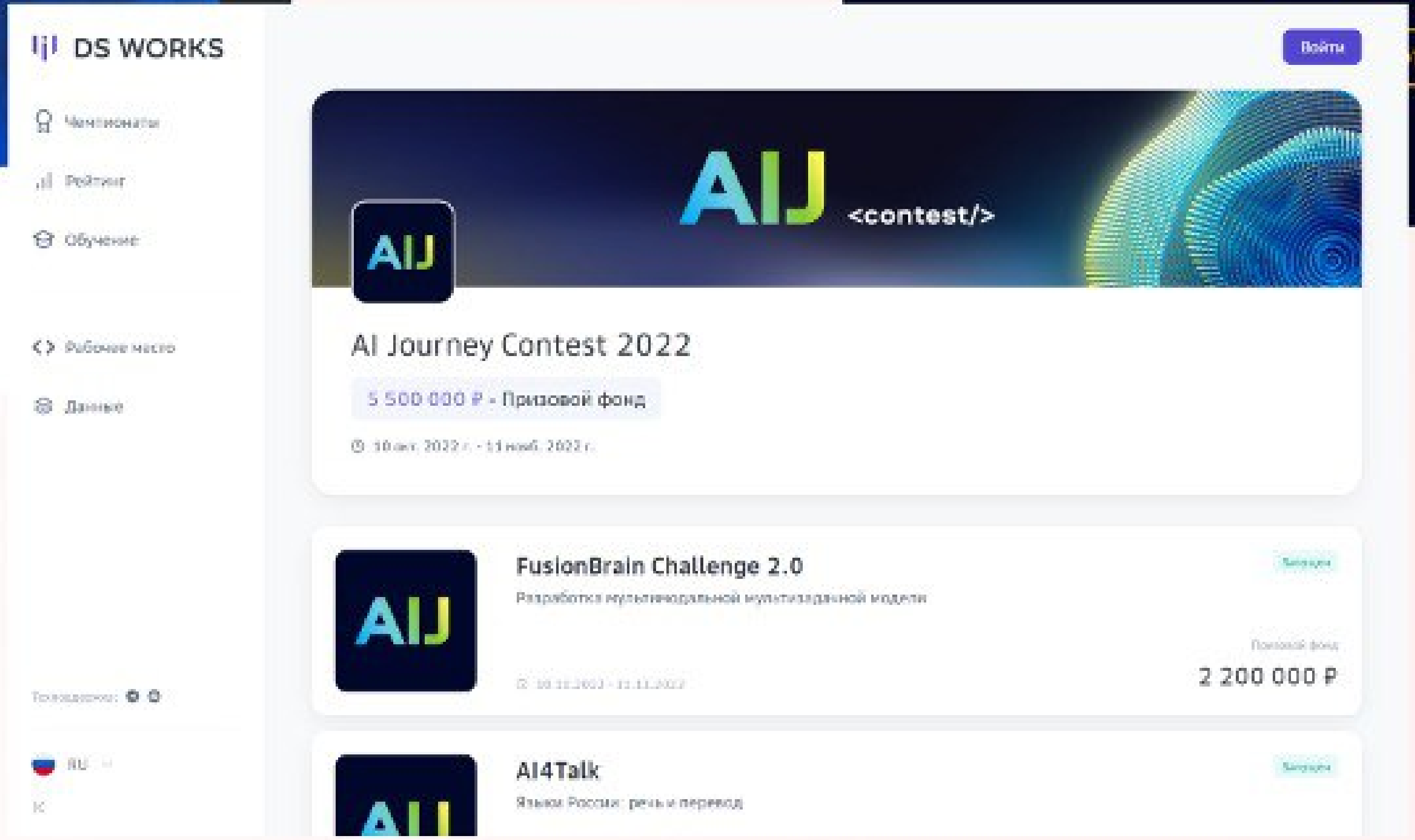
AI Lab MISIS

DS-competitions
Right Now





AIJ



YaCup

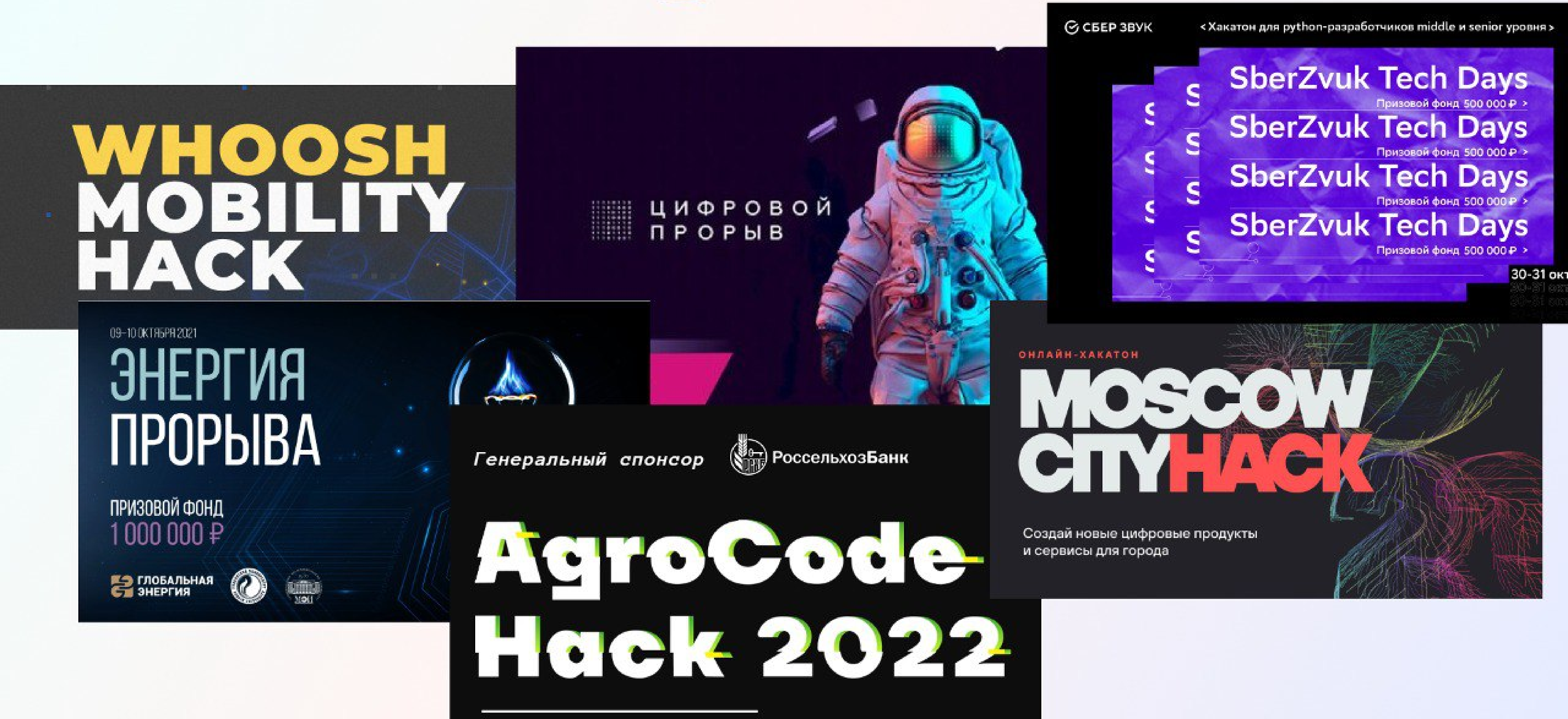
TOLOKA

Artificial intelligence
Machine Learning, Data Science

Data-Science чемпионаты

Хакатоны

Hackatons 😁


WHOOSH
MOBILITY
HACK
ЦИФРОВОЙ
ПРОРЫВ
СБЕР ЗВУК
< Хакатон для python-разработчиков middle и senior уровня >
SberZvuk Tech Days
Призовой фонд 500 000 ₽ >
30-31 окт
09-10 ОКТЯБРЯ 2021
ЭНЕРГИЯ
ПРОРЫВА
ПРИЗОВОЙ ФОНД
1 000 000 ₽
ГЛОБАЛЬНАЯ ЭНЕРГИЯ
ОНЛАЙН-ХАКАТОН
MOSCOW
CITYHACK
Создай новые цифровые продукты
и сервисы для города
Генеральный спонсор
РоссельхозБанк
AgroCode
Hack 2022

Hackatons
Agora-Hack

AGORA HACK 19 — 21 АВГУСТА 2022

TFIDF + SVD + KNN

На референсах - фаворитный метод обучения модели

Тренировка = 230мс+3сек на SVD
Векторизация = 1 сек

Плюсы:
- быстрый инференс
- нужно тренировать только на эталонах
- меньшая размерность векторов
- занимает мало места
- модель устойчива к выбросам

0.963

ACCURACY

Использование данной конфигурации модели показывает значительно высокую оценку

БЫСТРАЯ ВЕКТОРИЗАЦИЯ

Легковесный алгоритм и быстрое обучение модели

A

АВТОМАТИЗАЦИЯ

Оптимально и автоматизировано, поскольку в TFIDF используются sparse-матрицы

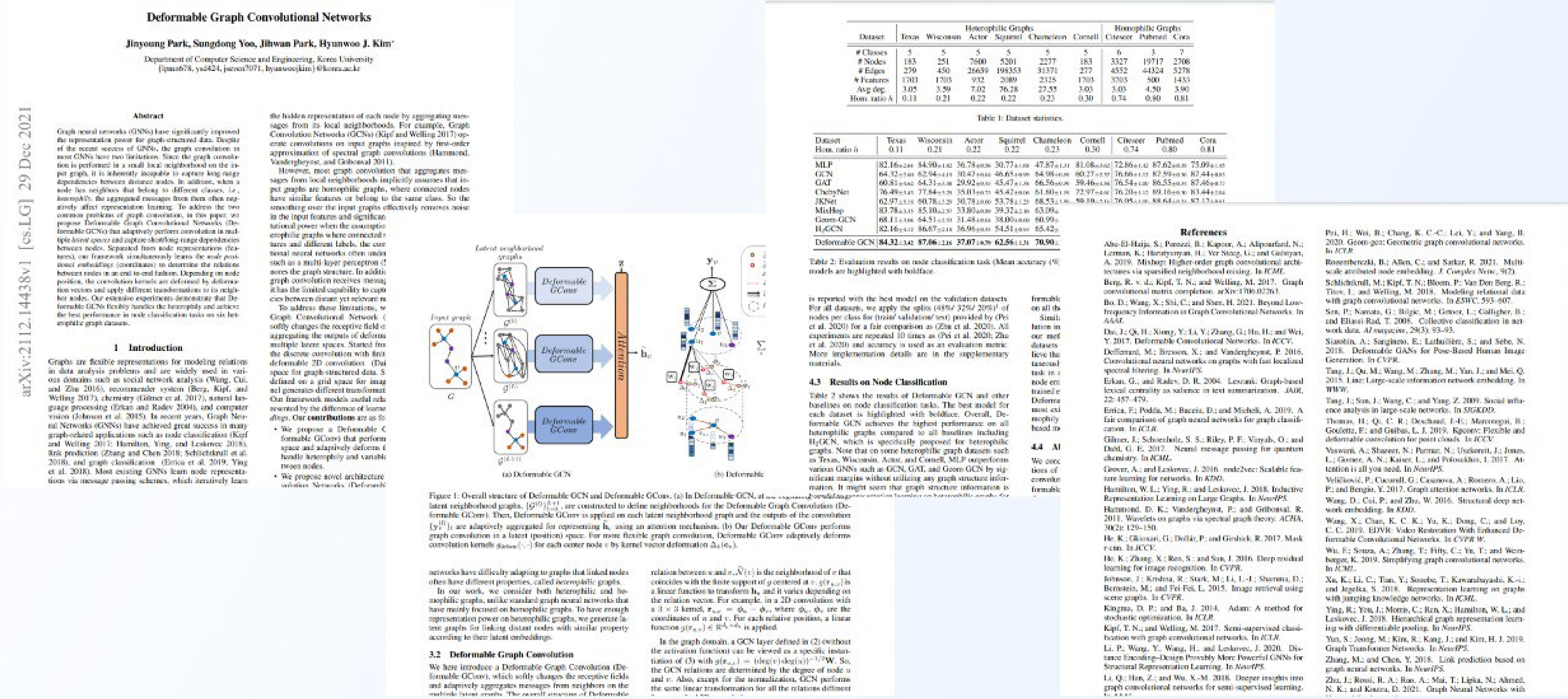
“Разработка системы с web-интерфейсом для сопоставления характеристик товаров маркетплейса с их эталонными значениями”

Artificial intelligence
Machine Learning, Data Science

- Data-Science чемпионаты
- Хакатоны
- Научные исследования

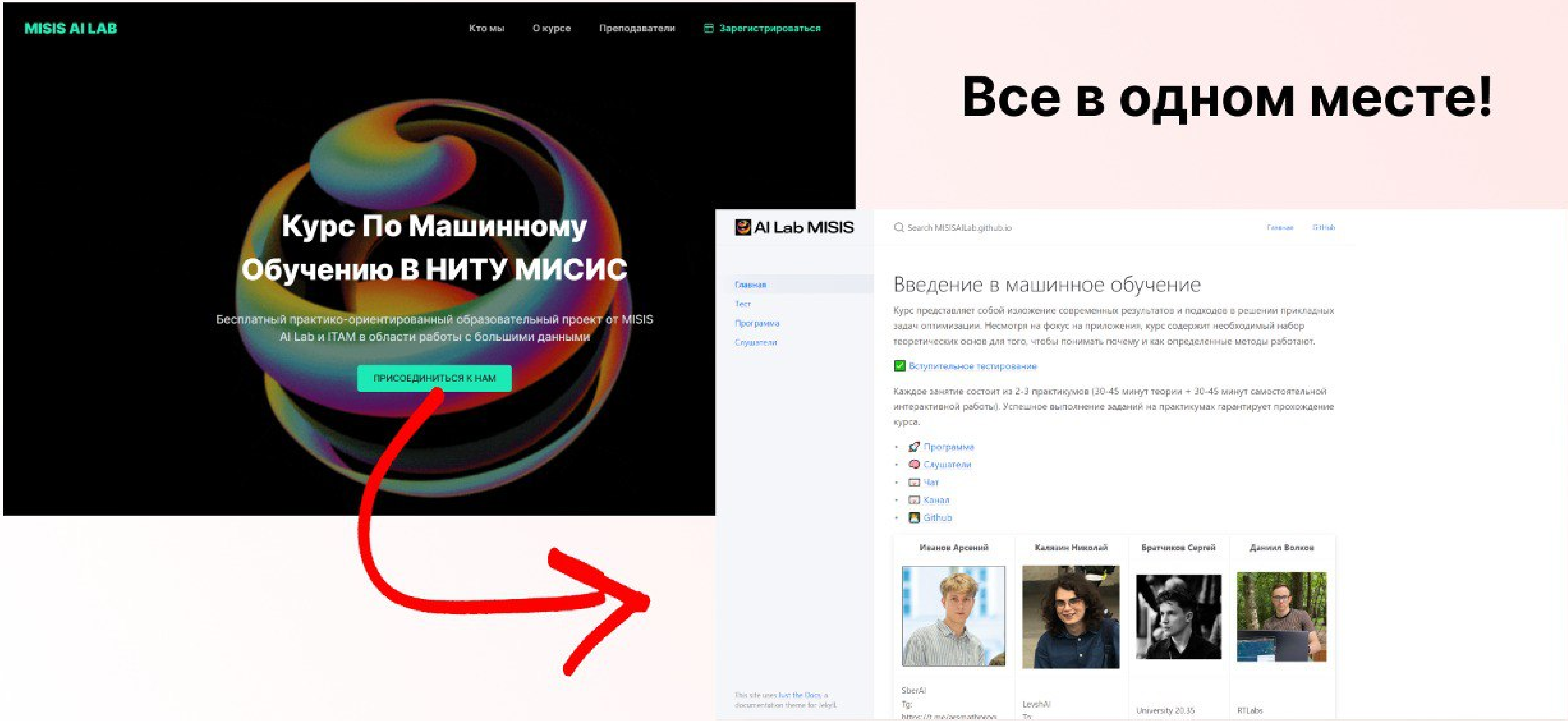
Research and science

AI Lab MISIS



Machine Learning course

AI Lab MISIS



Все в одном месте!

Николай Калязин

- 3 курс Прикладная математика
- ex сотрудник SberDevices
- работаю в стартапе LevshAI

Изучаю и работаю в CV, NLP
Libs: Pandas, NumPy, SciPy, PyTorch, TensorFlow

Machine Learning - is....

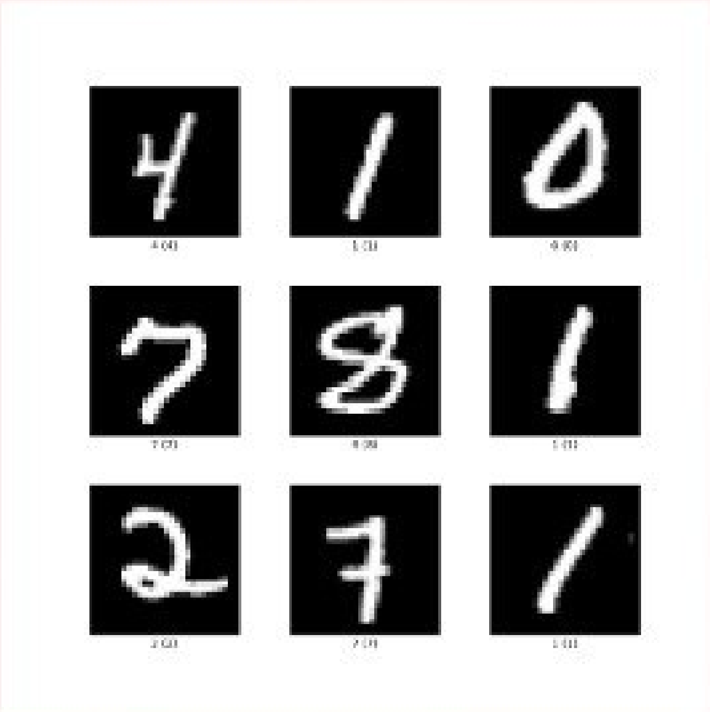
1. “Машинное обучение - это наука, изучающая алгоритмы, которые способны извлекать закономерности из ограниченного количества примеров и улучшаться за счет опыта”

2. “Алгоритм машинного обучения - это такой алгоритм, который способен обучаться на данных.”

Говорят, что программа обучается на опыте E относительно некоторого класса задач T и меры качества P, если качество на задачах из T, измеренное с помощью P, возрастает с ростом опыта E.

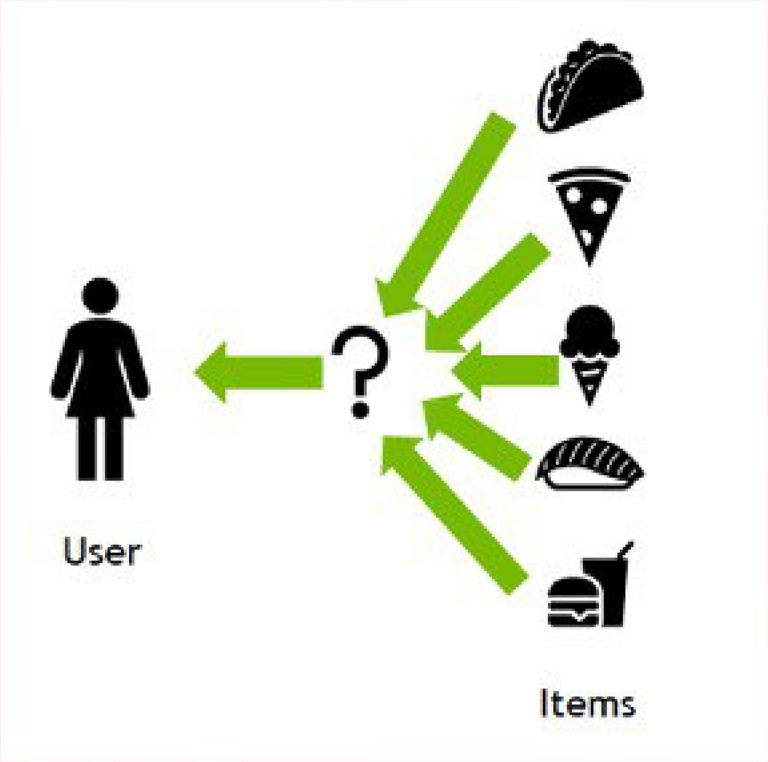
Tom Michael Mitchell

Machine learning



Задача: Распознование символов
Мера: Процент правильно распознаных
Опыт: База, размеченныхвручную, символов

Задача: Игра в шашки, шахматы, го
Мера: Процент побед
Опыт: Игра программы против себя



Задача: Рекомендация товаров, услуг, видео
Мера: Процент успешных рекомендаций
Опыт: Список товаров, просмотренных, купленных, оцененных пользователями

ML THESAURUS

Name

Age

Statistics

Python

Eye color

Native language

Target (mark)

Target (passed)

ML thesaurus

Observation (or datum, or data point) is one piece of information.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

In many cases the observations are supposed to be ***i.i.d.***

- ***independent***
- ***identically distributed***

Dataset

Machine Learning Tasks

Unsupervised learning

Group and interpret data based only on input data.

Supervised learning

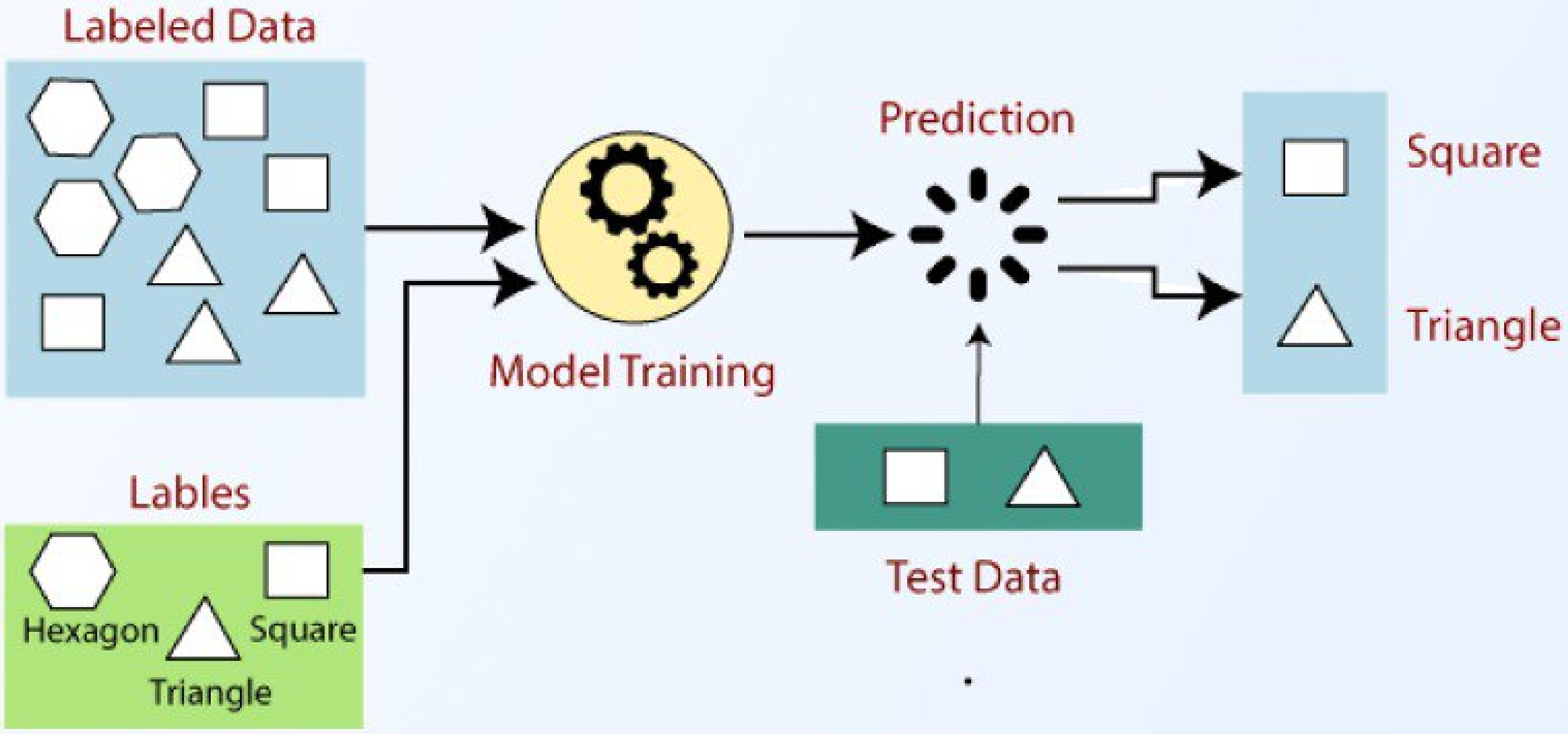
Develop predictive model based on both input and output data

Supervised learning

Добавим чуточку математических обозначений:

Пусть у нас есть датасет Xtrain={(x1, y1), ... (xn, yn). Где xi - объекты, yi- таргеты. Как уже было сказано, мы хотим построить отображение f: X → Y, где X - пространство объектов, а Y - пространство таргетов.

У нас есть объекты, мы хотим восстановить целевую зависимость, чтобы уметь выдавать ответы и на новые объекты тоже.



Supervised target learning tasks

Между собой задачи в обучении с учителем отличаются пространством таргетов.

Многоклассовая классификация

Бинарная классификация

Классификация

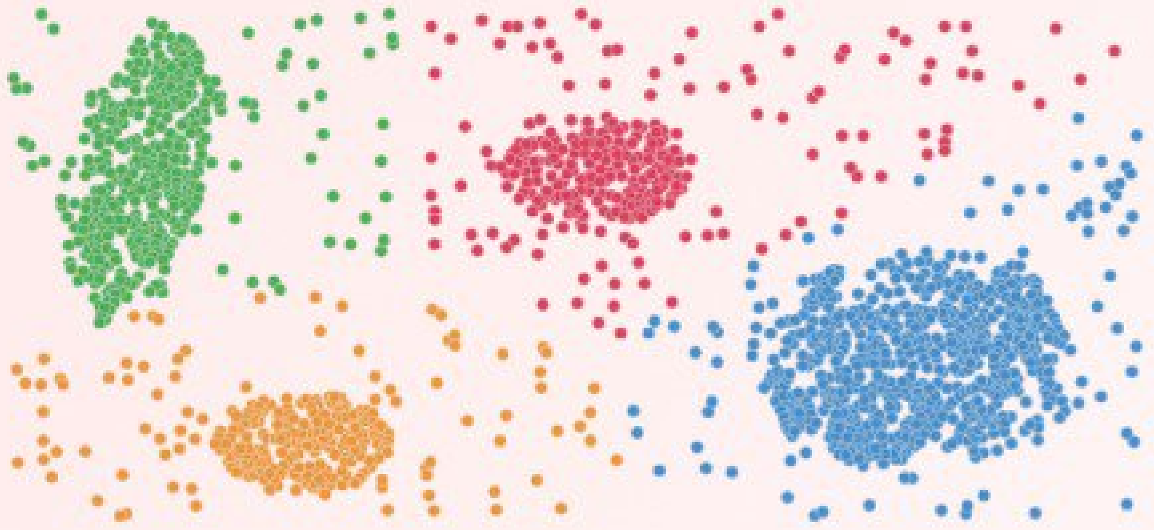
Мультилейбл классификация

Регрессия

Модель должна предсказать не класс, а какое-то действительное число. То есть Y = R для одномерной регрессии и Y = R^n для многомерной.

Unsupervised learning

Кластеризация



Понижение размерности