



ІІТМО

Анализ графовых данных и глубокое обучение

Азимов Рустам
Высшая школа цифровой культуры



ІІТМО

Информация о курсе

Познакомиться с методами машинного и глубокого обучения для работы с графовыми данными:



- **Построение эмбедингов для вершин:** DeepWalk, Node2Vec
- **Графовые нейронные сети (GNN):** GCN, GraphSAGE, GAT...
- **Анализ графов знаний (Knowledge graphs):** TransE, Query2box
- **Практическое применение GNN**

Пререквизиты

Желательны, но не обязательны знания в следующих областях:



- Машинное и глубокое обучение
- Теория графов и алгоритмы
- Теория вероятностей
- Математическая статистика

Желательные навыки программирования:

- Python
- sklearn, PyTorch



PyG  

- [PyG](#) (PyTorch Geometric)
 - Самая популярная библиотека для GNN
- [GraphGym](#) - платформа для проектирования GNN
 - Реализовано множество модулей GNN
 - Упрощенный подбор гиперпараметров
 - Гибкая кастомизация
- [NetworkX](#) - полезная библиотека для различных манипуляций над графами и сетевого анализа

Полезные ссылки

- [GitHub репозиторий курса](#)
- [Graph Representation Learning Book](#) - Will Hamilton
- [CS224W: Machine Learning with Graphs](#)



Оценивание курса

ІІТМО

- Практические домашние задания — 4 шт.
 - На программирование, максимум по 10 баллов
 - Дедлайн 1-2 недели
- Проект (групповой на 2-3 человека)
 - Практическое применение графовых нейронных сетей, максимум по 20 баллов каждому участнику
- Максимум за курс можно набрать 60 баллов
- Экзамен (устный)



Правила сдачи заданий

- Домашние задания следует посылать на почту rsazimov@itmo.ru
 - Ссылка на блокнот (colab, GitHub)
 - В теме письма указать: “Анализ графовых данных с использованием глубокого обучения”
- Также работа должна содержать: ФИО студента, группу, список выполненных пунктов задания
- Если ваша работа зачтена до дедлайна, то вы получаете все заработанные баллы
- Если ваша работа зачтена *после* дедлайна, то вы получаете половину заработанных баллов



- Защита проекта проходит на последнем занятии
- Проект должен содержать решение с помощью графовых нейронных сетей интересной задачи на реальных данных
- Проект следует оформить в виде презентации или блокнота с подробным пояснением всех этапов
 - Обзор предметной области
 - Постановка задачи
 - Описание данных
 - Предлагаемое решение
 - Эксперименты и выводы



- Экзамен будет состоять из двух вопросов по пройденным темам
- Для допуска к экзамену необходимо набрать хотя бы 25 баллов
- Набрав от 30 до 39 баллов можно получить за экзамен 3 автоматом
- Набрав от 40 до 49 баллов можно получить за экзамен 4 автоматом
- Набрав 50 и более баллов можно получить за экзамен 5 автоматом



Пересдача

ІТМО

- Для допуска к пересдаче необходимо набрать хотя бы 25 баллов
- При этом штрафы за просроченные дедлайны уже не будут учитываться
- Для получения автомата те же правила сохраняются





ІІТМО

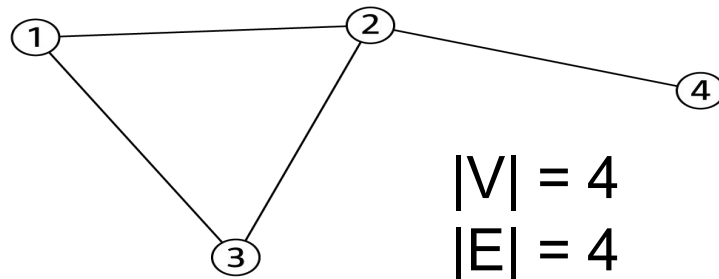
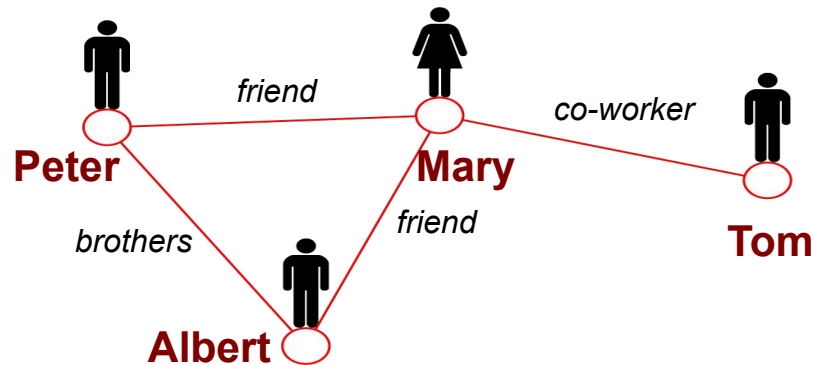
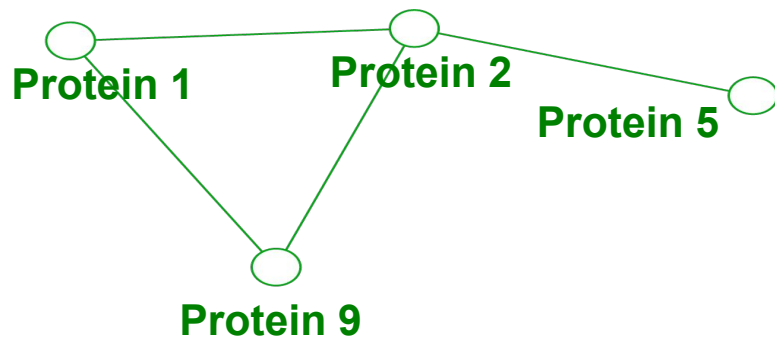
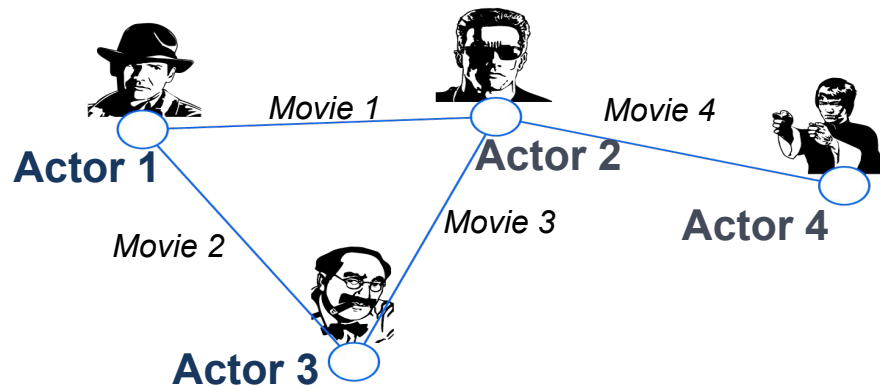
**Машинное обучение
для анализа графов**

Почему графы?

- Графы - это универсальный язык для описания и анализа сущностей с отношениями/взаимодействиями
- Данные во многих областях естественным образом представляются в виде графов



Графы - универсальный язык



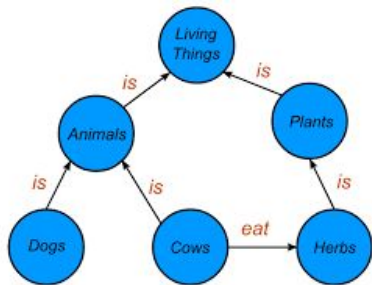
$$|V| = 4$$
$$|E| = 4$$

Графы в реальной жизни

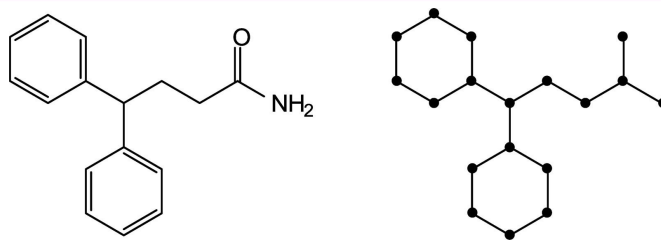
ІТМО



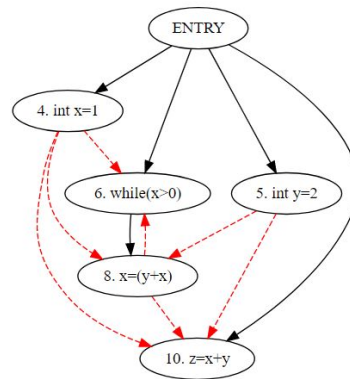
Социальные сети



Графы знаний



Молекулы



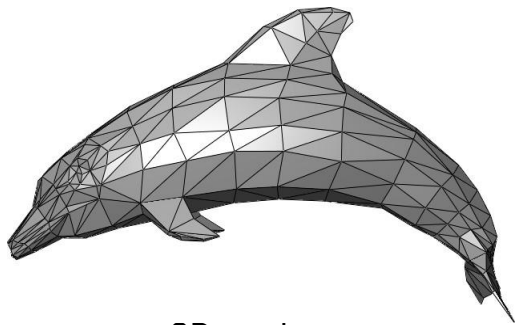
Программы

Графы в реальной жизни

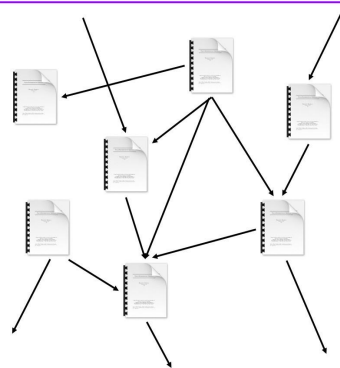
ІТМО



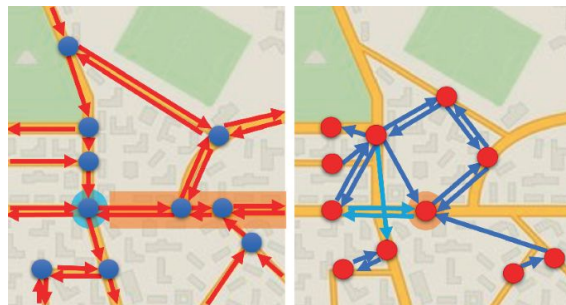
Компьютерные сети



3D графика



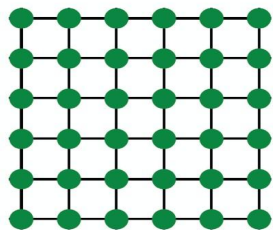
Граф цитирований



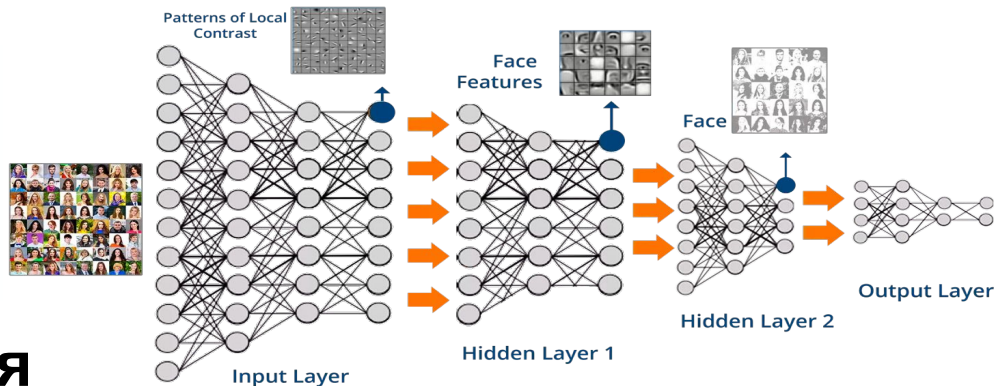
Навигация



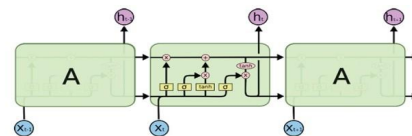
Современные ML инструменты



Изображения

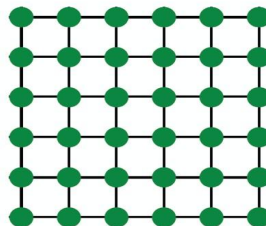


Текст/Аудио

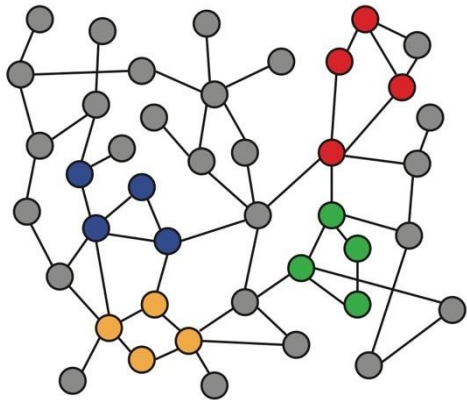


- Инструменты для глубокого обучения изначально спроектированы для анализа лишь подмножества графов

- Последовательности
- Сетки

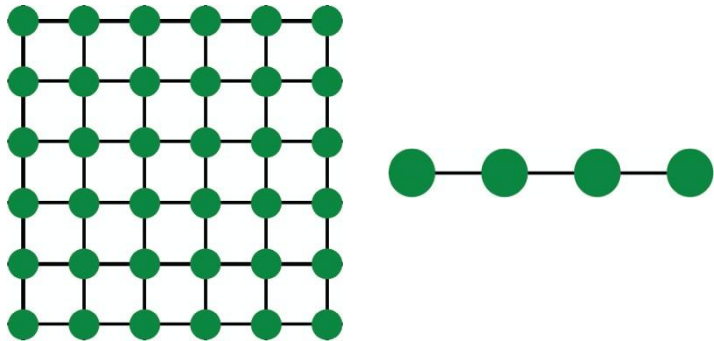


Сложность анализа графов

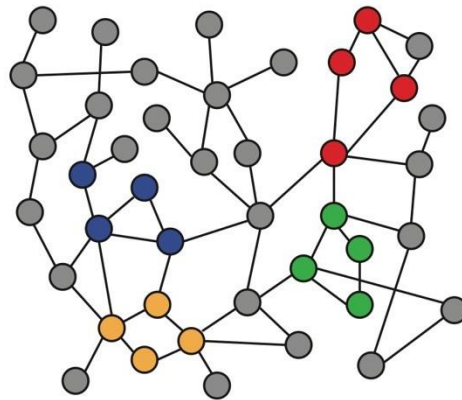


- Сложная топологическая структура (нет пространственной локальности как в сетках)
- Нету начальной вершины или порядка обхода графа
- Граф часто динамичный
- Вершины могут иметь мультимодальные признаки

VS



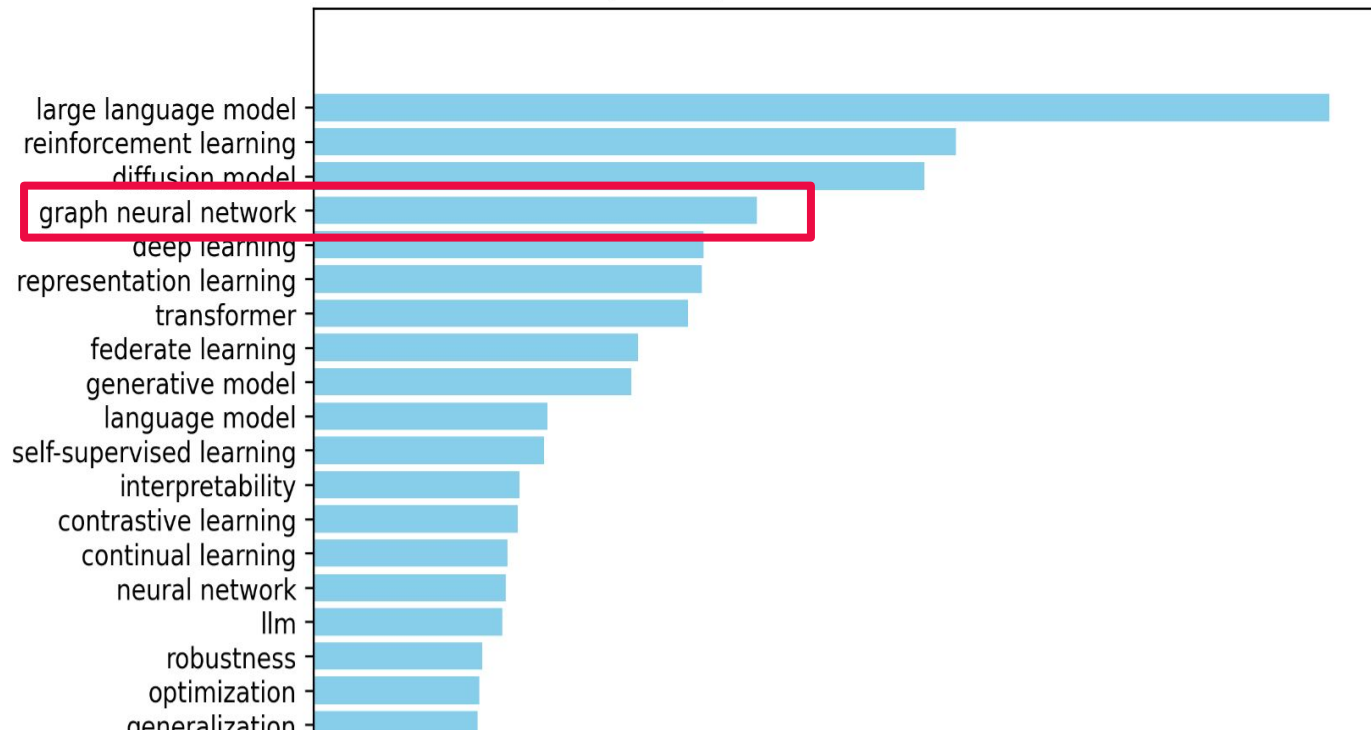
- Как мы можем спроектировать нейронные сети намного более общего применения?
 - Произвольные графы



Популярность GNN

ICLR 2024

Top 50 Keywords after Lemmatization





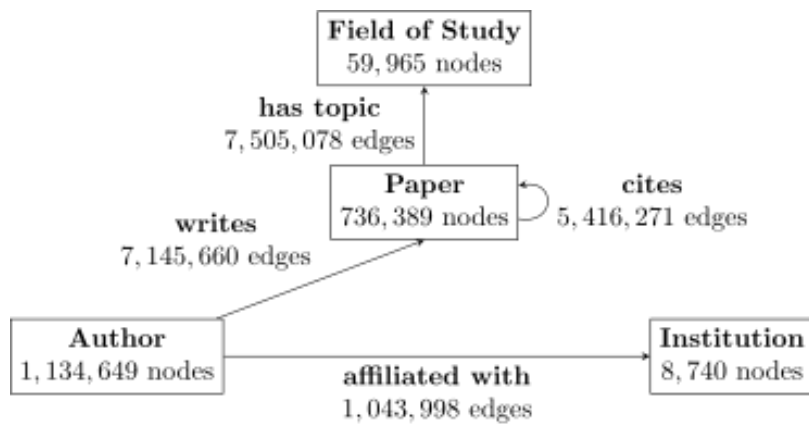
ІТМО

**Выбор
представления
графовых данных**

- Во многих областях данные могут быть представлены в виде гетерогенного графа $G = (V, E, R, T)$
 - Вершины $v_i \in V$
 - Рёбра $(v_i, r, v_j) \in E$
 - Типы вершин $T(v_i)$
 - Типы отношений $r \in R$
 - У вершин и рёбер могут быть атрибуты/признаки



Пример гетерогенного графа



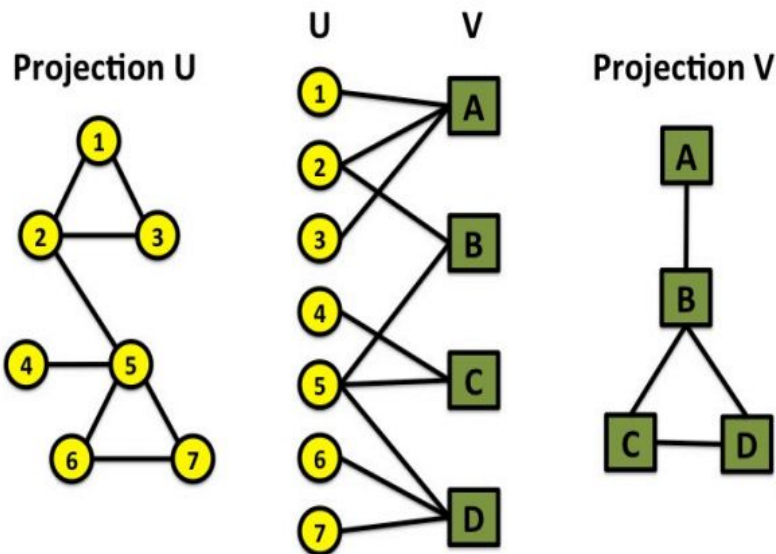
- [ogbn-mag](#) (Microsoft Academic Graph)
- Типы вершин: ***author***, ***paper***, ***institution*** и ***field of study***
- Типы рёбер: ***writes***, ***affiliated with***, ***cites*** и ***has topic***

Выбор подходящего представления ИТМО

- Как построить граф?
 - Что сделать вершинами?
 - Что сделать рёбрами?
 - Направленный vs ненаправленный
 - Нужны ли веса на рёбрах?
 - Какие типы вершин/рёбер?
 - Какие признаки хранятся в вершинах/рёбрах?
 - Нужен ли особый вид графа?
- От сделанного выбора зависит природа вопросов, на которые можно будет ответить в результате анализа графа



Двудольные графы



- Примеры двудольных графов
 - Авторы-Статьи
 - Пользователи-Фильмы
 - Покупатели-Товары
- Можно провести дополнительные рёбра и получить новые графы
 - Соавторы
 - Пользователи/покупатели со схожими вкусами

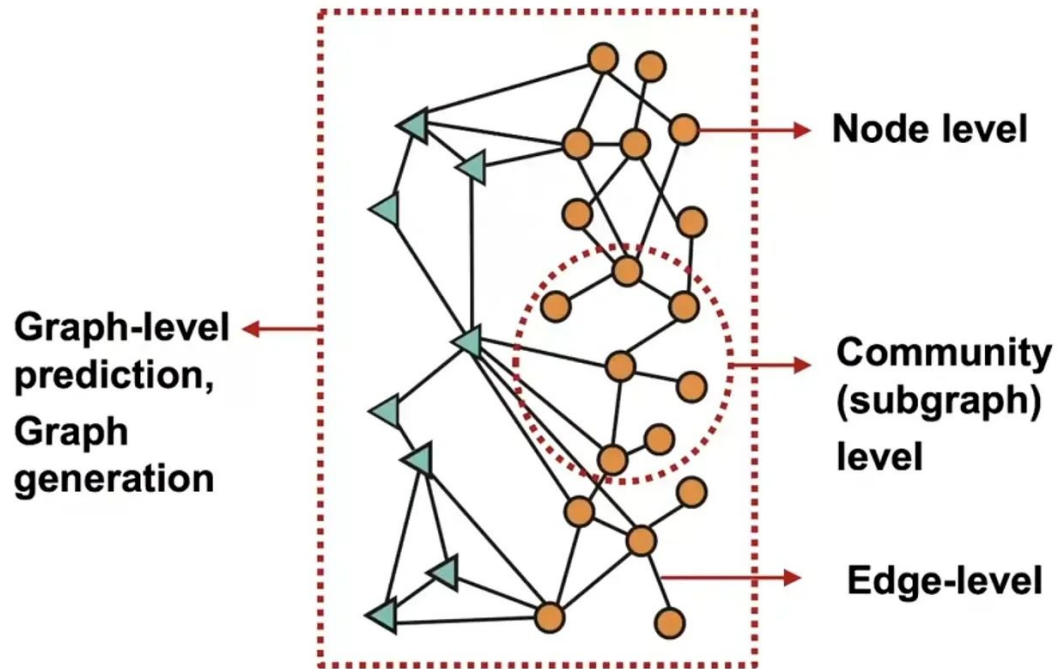




ІТМО

**Виды задач
машинного обучения
на графах**

Виды graph ml задач



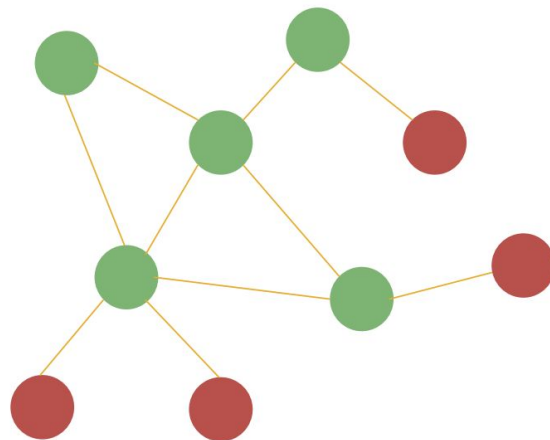
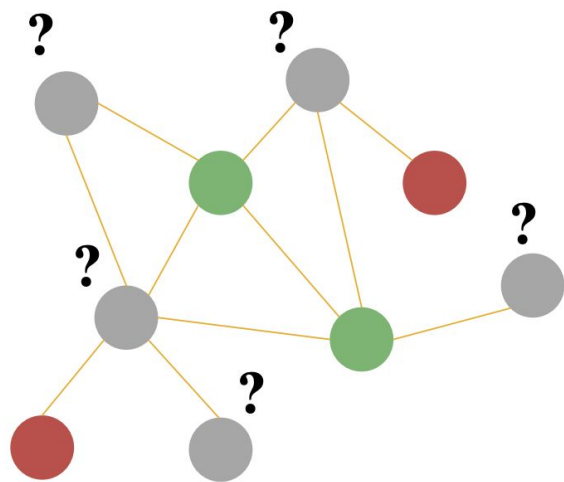


IITMO

Node-level tasks

Node classification

VITMO



Node classification

- Предсказываем признаки отдельных вершин
- Например, категоризация
 - Покупателей
 - Товаров
 - Транзакций
 - Лекарств



Protein folding

- Белки, составленные из аминокислот, под действием магнитных и прочих воздействий сворачиваются в сложные 3D фигуры
- От этого зависят многие важные биологические функции
 - Взаимодействие лекарств с белками и изменение процессов в организме для выздоровления

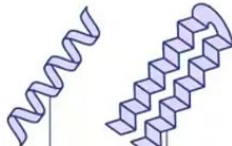


Every protein is made up of a sequence of amino acids bonded together



Amino acids

These amino acids interact locally to form shapes like helices and sheets



Alpha helix
Pleated sheet

These shapes fold up on larger scales to form the full three-dimensional protein structure



Pleated sheet
Alpha helix

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA



Protein folding

- Задача - предсказать 3D структуру белка, основываясь только на последовательности аминокислот



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



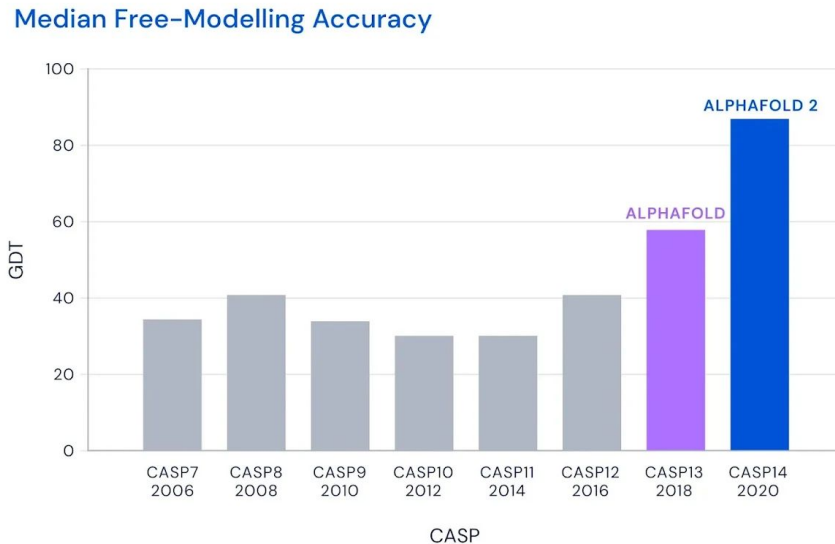
T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction

Image credit: [DeepMind](#)

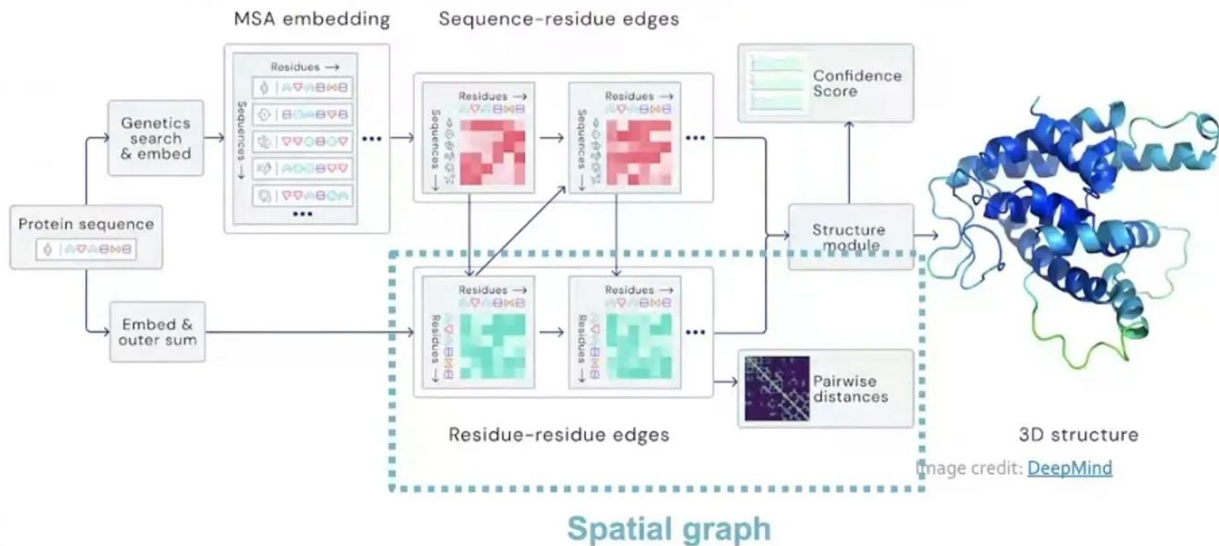
- Начиная с 1970-ых годов пытаются решить данную задачу
- Использование GNN позволило сделать прорыв и решить задачу с 90%

ТОЧНОСТЬЮ



AlphaFold

- Идея - представить белок в виде графа (пространственного графа)
- Вершины - аминокислоты
- Рёбра - пространственная близость аминокислот





IITMO

Edge-level tasks

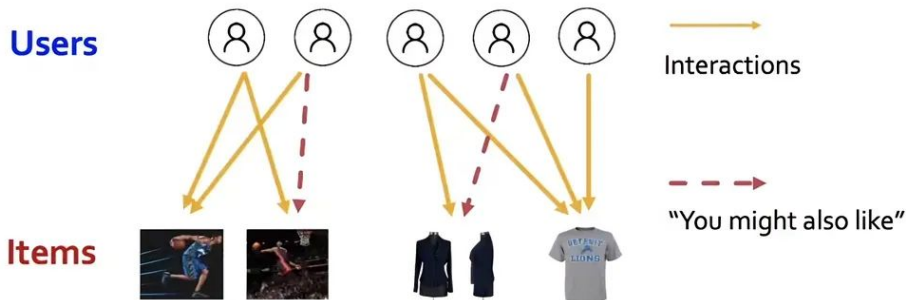
Link prediction

Task: Recommend related pins to users



Task: Learn node embeddings z_i such that

$$d(z_{cake1}, z_{cake2}) < d(z_{cake1}, z_{sweater})$$

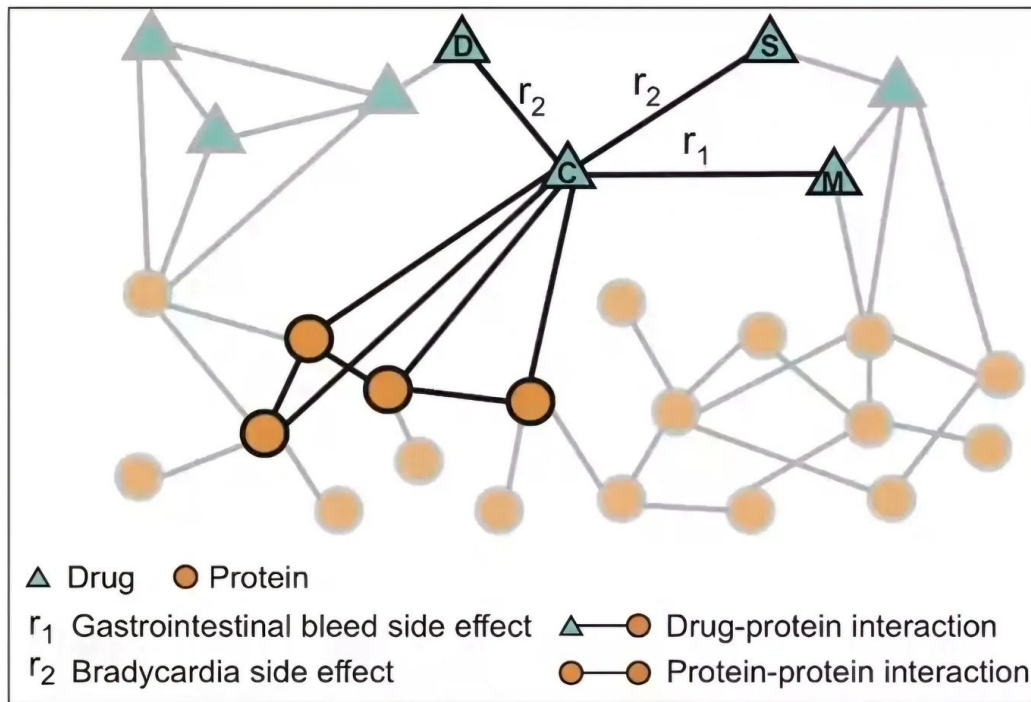


Recommendation system

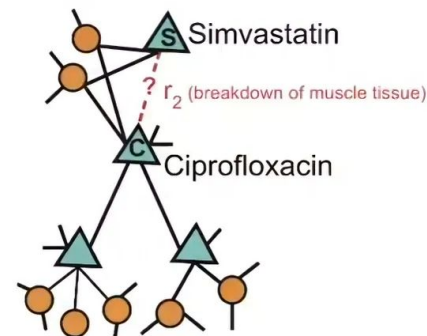
- Многие компании используют GNNs для более точных рекомендаций
 - Pinterest
 - LinkedIn
 - Instagram
- Например, изображения (вершины) в Pinterest кодируются на основе взаимодействия пользователей и визуального контента
- Похожие вершины получают близкие представления (эмбединги)
- В итоге качество рекомендаций становится выше, чем после анализа только изображений



Взаимодействия лекарств



Query: How likely will Simvastatin and Ciprofloxacin, when taken together, break down muscle tissue?



Предсказание побочных эффектов

ИТМО

Rank	Drug i	Drug j	Side effect r	Evidence
1	Pyrimethamine	Aliskiren	Sarcoma	Stage et al. 2015
2	Tigecycline	Bimatoprost	Autonomic neuropathy	
3	Omeprazole	Dacarbazine	Telangiectases	
4	Tolcapone	Pyrimethamine	Breast disorder	Bicker et al. 2017
5	Minoxidil	Paricalcitol	Cluster headache	
6	Omeprazole	Amoxicillin	Renal tubular acidosis	Russo et al. 2016
7	Anagrelide	Azelaic acid	Cerebral thrombosis	
8	Atorvastatin	Amlodipine	Muscle inflammation	Banakh et al. 2017
9	Aliskiren	Tioconazole	Breast inflammation	Parving et al. 2012
10	Estradiol	Nadolol	Endometriosis	

Zitnik et al., Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics 2018

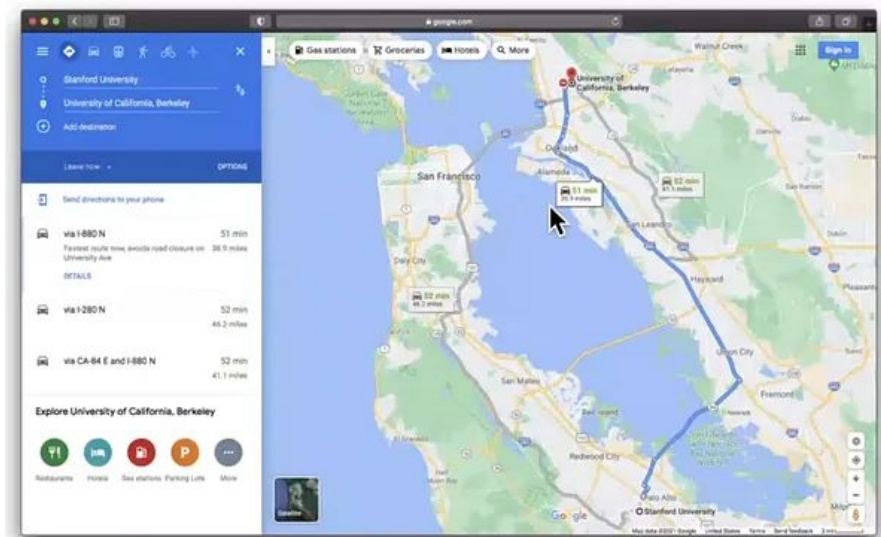


IITMO

Graph-level tasks

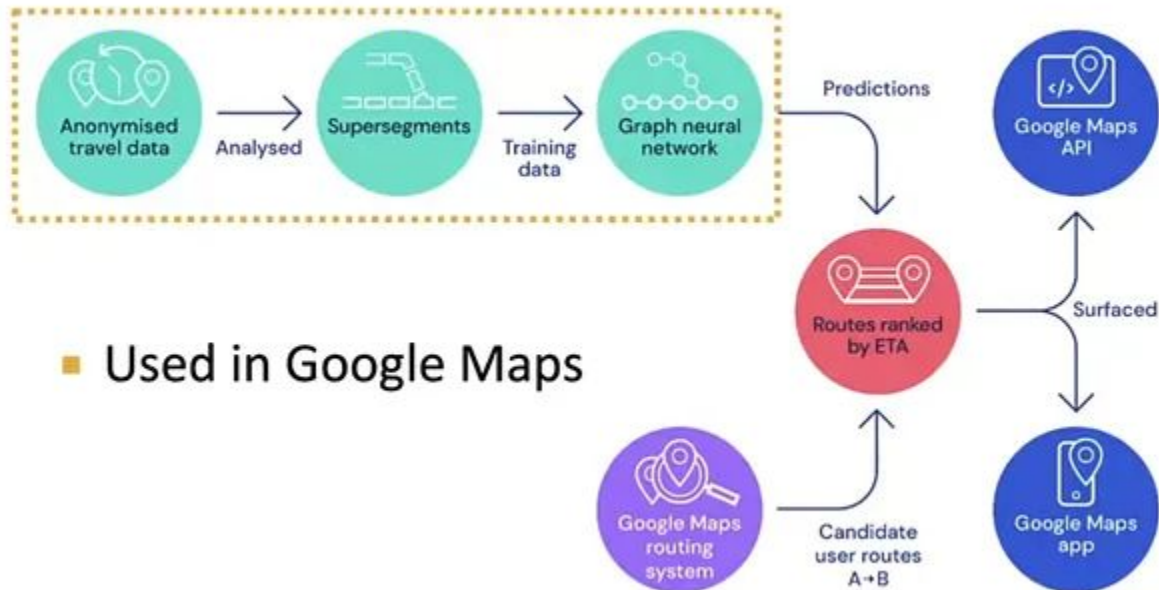
Sub-graph level

- Вершины - сегменты дорог
- Время поездки предсказывается с помощью GNN



Traffic prediction

Predict via Graph Neural Networks



■ Used in Google Maps

THE MODEL ARCHITECTURE FOR DETERMINING OPTIMAL ROUTES AND THEIR TRAVEL TIME.

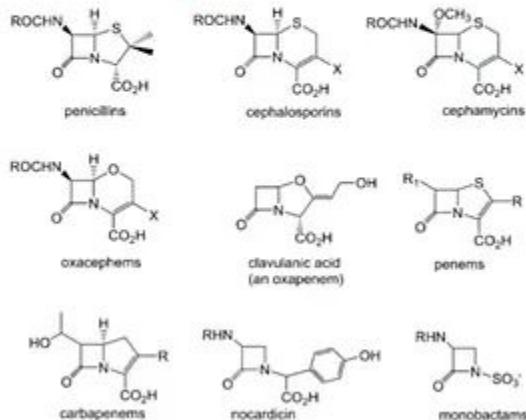
Image credit: [DeepMind](#)

Graph-level task

- **Antibiotics are small molecular graphs**

- **Nodes:** Atoms

- **Edges:** Chemical bonds



Konaklieva, Monika I. "Molecular targets of β -lactam-based antimicrobials: beyond the usual suspects." *Antibiotics* 3.2 (2014): 128-142.

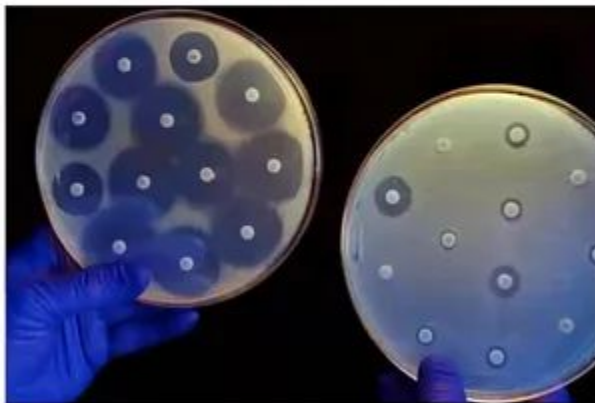
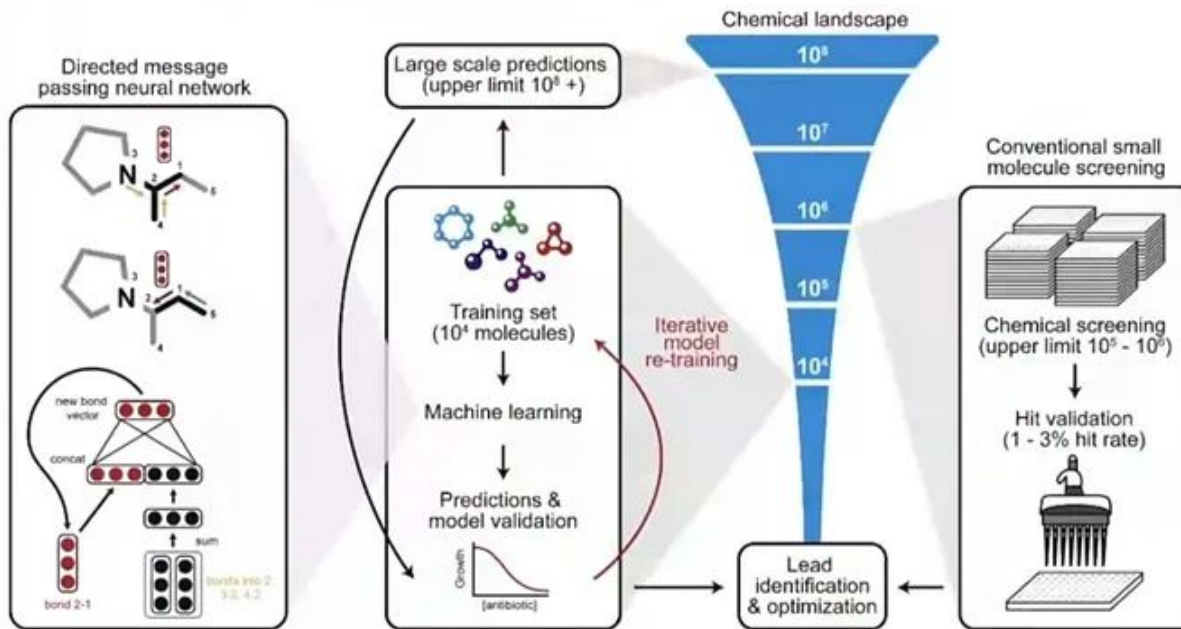


Image credit: [CNN](#)

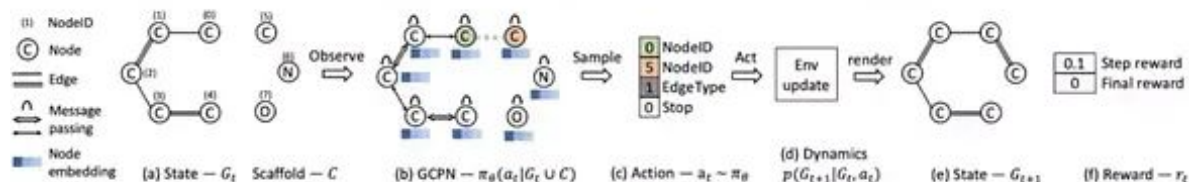
Antibiotic discovery

- Задача - предсказать нужные свойства графов (молекул)

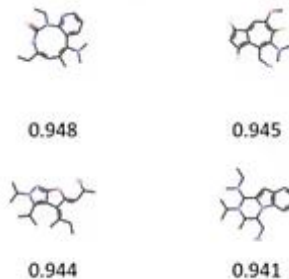


Генерация новых молекул

Graph generation: Generating novel molecules



Use case 1: Generate novel molecules with high drug likeness



Use case 2: Optimize existing molecules to have desirable properties



Симуляция изменений графа

A graph evolution task:

- **Goal:** Predict how a graph will evolve over

