Introduction

Jiaxuan You

Assistant Professor at UIUC CDS

CS512: Data Mining Principles, 2025 Fall

https://ulab-uiuc.github.io/CS512/

Introduction

Course Logistics

Course Logistics

- **Instructor:** Jiaxuan You
  - Lectures: Wed/Fri 2:00 PM - 03:15 PM, DCL Room 1320
  - Office hours: Wed 3:30PM – 5:00PM (Siebel 2122)
- **Teaching Assistants:** Haofei Yu (50% TA), Pengrui Han (currently 25% TA)
  - TA OH: Haofei (Tue/Thur 12:45PM - 1:45PM), Pengrui (TBD)
  - Lecture slides will be before each lecture
- **Prerequisites:** Mostly self-contained
  - General knowledge on data mining, deep learning, Python
- **Course website:** https://ulab-uiuc.github.io/CS512/
  - **Communications:** Join Slack, for announcements, teammate findings, discussions
  - **Canvas:** Homework submissions, grading, major annoucements
- Structure of lectures: 60-70 minutes of a lecture. Feel free to ask questions
- **Textbook:** Research papers (classic and recent)

Workload and Grading

- Final grade will be composed of (subject to minor adjustments):
  - Assignments: 20%
    - 4 coding assignments using Google Colab - each 5%
  - Paper reading and analysis: 15%
    - Deep dive into research papers and their related works -> Analysis
  - Idea brainstorm and discussion: 20%
    - Research idea proposals, in person discussion sessions (5%)
  - Course project report: 25% [Group of 3]
    - Project Proposal, submitted version, final version after review and response
  - Review and response: 10%
    - Participate in OpenReview discussions as authors and reviewers
  - Research Presentations: 10% [Group of 3]
    - Recorded presentations for efficiency. Bonus 1% for final in-person presentation and discussion

Course Schedule

Week	Date	Knowledge learning	Research training	Events	Deadlines
1	Aug 27 Wed	Introduction	Paper reading & analysis		
	Aug 29 Fri	Graph learning basics	Paper reading & analysis	Writing task 1, out	
2	Sept 03 Wed	Embedding-based graph learning	Paper reading & analysis		
	Sept 05 Fri	Graph neural networks: perspective	Paper reading & analysis	Writing task 2, out	
3	Sept 10 Wed	Graph neural networks: model I	Paper reading & analysis		
	Sept 12 Fri	Graph neural networks: model II	Paper reading & analysis	Idea proposal, out	Writing task 1 due
4	Sept 17 Wed	Paper reading discussions	Ideate & discussion		
	Sept 19 Fri	Graph neural networks: objective	Ideate & discussion	HW 1, out	Writing task 2 due
5	Sept 24 Wed	Graph neural networks: pipeline	Ideate & discussion		
	Sept 26 Fri	Graph neural networks: alternatives and add-ons	Ideate & discussion	Project proposal, out	Idea proposal due
6	Oct 01 Wed	GNN implementation: PyG & GraphGym	Ideate & discussion		
	Oct 03 Fri	Project idea discussions	Ideate & discussion	HW 2, out	HW 1 due
7	Oct 08 Wed	Beyond simple graphs: heterogeneous graphs	Prototype implementation		
	Oct 10 Fri	Beyond simple graphs: knowledge graphs	Prototype implementation	Project submission, out	Project Proposal due
8	Oct 15 Wed	Beyond simple graphs: knowledge graph reasoning	Prototype implementation		
	Oct 17 Fri	Beyond prediction: graph generative models	Prototype implementation	HW 3, out	HW 2 due
9	Oct 22 Wed	Beyond message passing: expressive GNNs	Prototype implementation		
	Oct 24 Fri	Beyond small graphs: scale GNNs to large graphs	Prototype implementation	HW 4, out	
10	Oct 29 Wed	Beyond small GNNs: graph foundation models	Paper Writing		
	Oct 31 Fri	Beyond sparse graphs: graph transformers	Paper Writing		HW 3 due
11	Nov 05 Wed	GNN applications: recommender systems	Paper Writing		
	Nov 07 Fri	GNN applications: graph mining	Paper Writing		HW 4 due
12	Nov 12 Wed	GNN applications: science	Review & Response		
	Nov 14 Fri	Graph & Multiagent: ResearchTown, MultiAgentBench	Review & Response		
13	Nov 19 Wed	Graph & LLM Ecosystem: Graphrouter+Grapheval	Review & Response		
	Nov 21 Fri	Graph & World Model	Review & Response	Review & response task, out	Submission due
14	Dec 03 Wed	No class	Remote Presentation		
	Dec 05 Fri	No class	Remote Presentation		Review & response due
15	Dec 10 Wed	Conclusion			Final project due

LLM Policy: We Embrace "ChatGPT Moment"

- LLMs fundamentally changed AI & beyond
- In this course, **we embrace LLMs**
  - **Assignments:** feel free to use LLMs to help, but you *must include your prompt,* ***or you may subject to Honor Code violation***
  - **Ideation/Projects:** feel free to use LLM as copilot, but we have designs such that you should think critically with LLM
- But we cherish **non-LLM-able** things
  - **Research training:** asking good questions is more important than answering questions

Course Objective

Takeaways from this course:

- **Knowledge** about **data mining via graph deep learning**
  - Core knowledge: Insights, Coding, Math
  - Latest knowledge: Recent research papers
- **Training** for AI research
  - Experience the **full lifecycle as an AI researcher**
    - Read, Ideate, Discuss, Code, Write, Review, Present
  - **Learn from peers** from discussions, collaborations, and sharing

Goal: a journey that cannot be directly replaced by LLM

Class Attendance

- Expected and highly encouraged!
  - Regularly attend the class meetings
  - Questions are highly welcomed
  - Will grant 1-2% extra grade to active participants
  - Attendance to the 2 discussion sessions count toward final grades
  - Will grant 1% to in-person presentation and discussions in the final session

Assignment Policy

- The homework is due at **11:59 PM CT on the due date**.
- Canvas: https://canvas.illinois.edu/ for collecting homework
  - Contact the TAs if you are having technical difficulties in submitting the assignment **before the deadline**. **2 total late days** will be permitted, after that **1% per late day will be deducted** to final grade (until any submission get zero grade)
- **Academic Integrity:**
  - Feel free to discuss with others, but you must submit your homework **independently**. Aiding others would have the same consequence as the direct violation. Academic standards (e.g., Plagiarism) apply to course projects and other writeups.

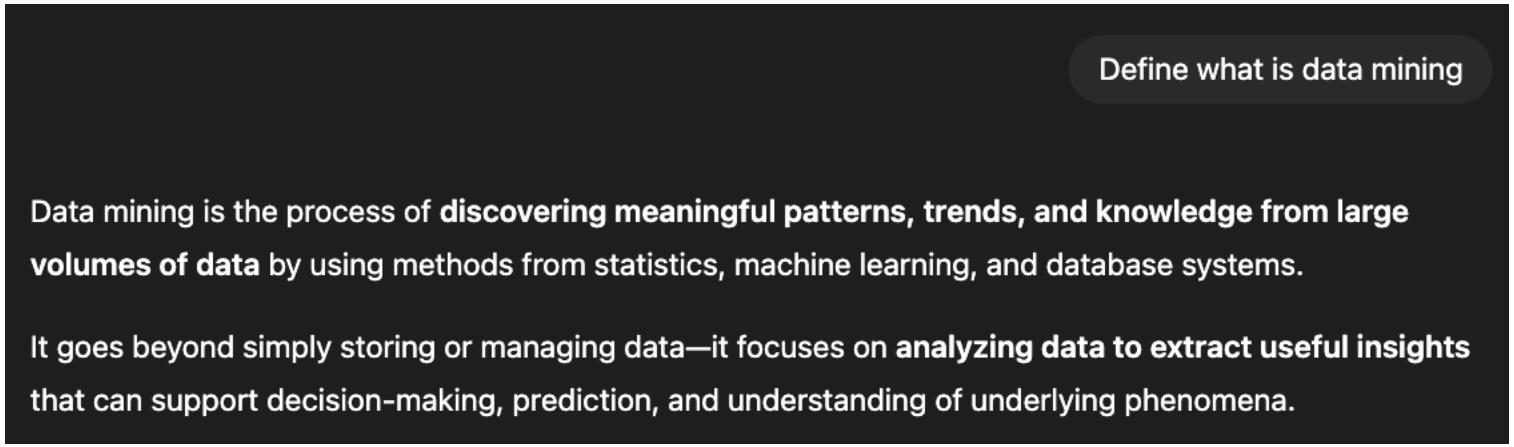
Introduction

What is Data Mining?

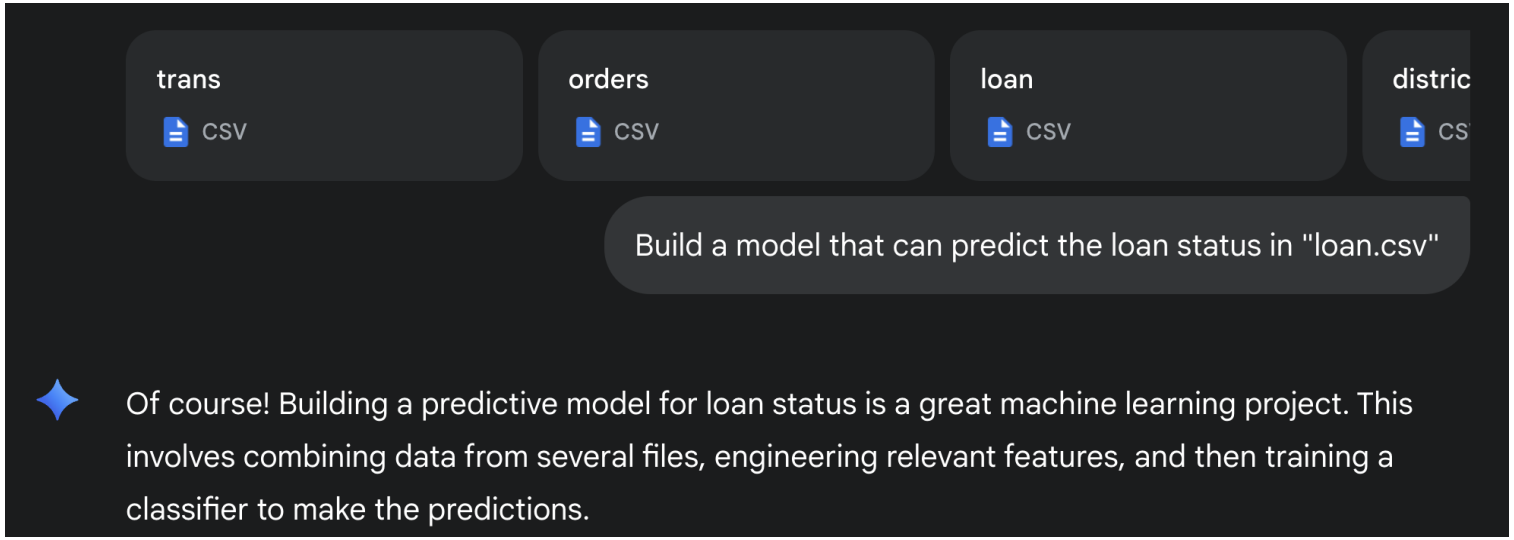
Popular Definitions of Data Mining

- From **Wikipedia:**
  - Data mining is the process of **extracting and finding patterns** in **massive datasets**
- From **Prior CS512 (Prof. Tong):**
  - Data mining (knowledge discovery from data): Extraction of **interesting** (non-trivial, implicit, previously unknown and potentially useful) **patterns or knowledge** from **huge amount** of data
- **A very broad scope:**
  - **Classic:** Frequent pattern mining, Clustering, Classification, Regression
  - **Recent:** Deep learning, Generative models, LLMs, AI agent
  - **This course: Graph-based data mining**

Why not just Use LLMs for Data Mining?



- What we would expect LLM could do in 2023



- Today: It's possible to prompt LLM for basic data mining https://g.co/gemini/share/0c433dfb735d

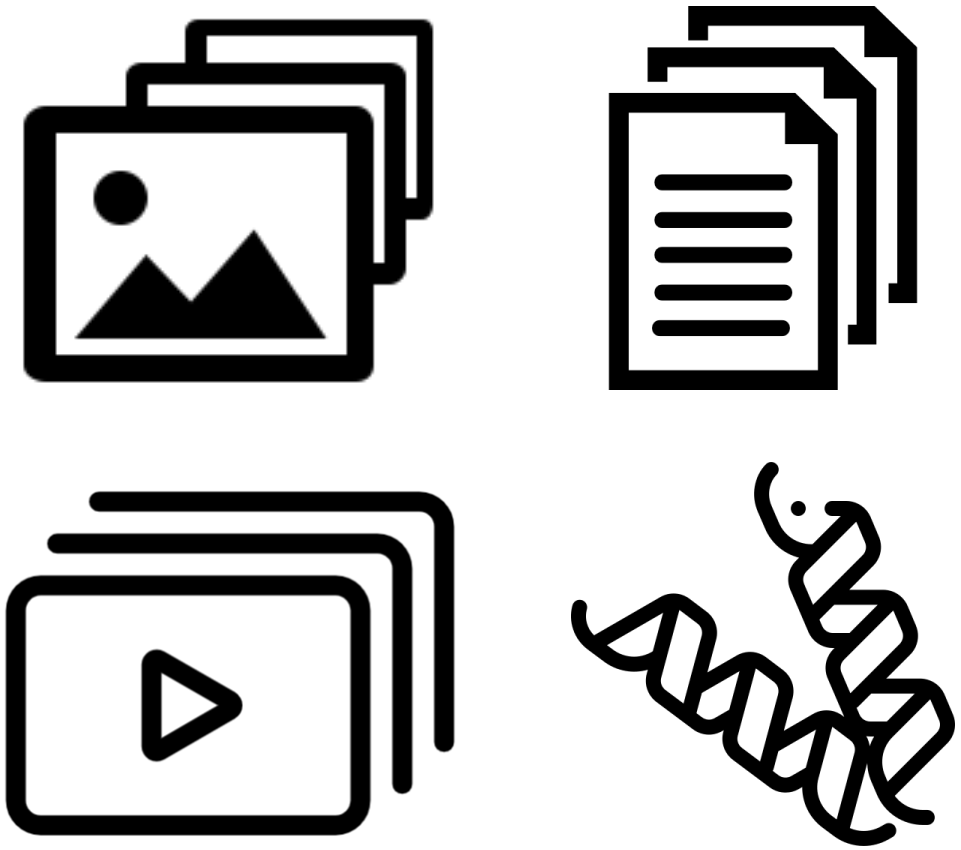
Why not just Use LLMs for Data Mining?

- **However, LLMs are not perfect**
- Technical Limitations of LLMs
  - **High cost**: Classic data mining methods can be far cheaper and efficient
  - **Availability:** SOTA LLMs are not open-sourced
  - **Applicability:** LLMs have limited capability to specific/emerging domains (e.g., graph deep learning to be covered in this course)
- Human Rationale
  - **LLM Copilot:** We, humans, need to understand the concepts and insights
  - **Creativity:** True innovation comes from human understanding of data
  - **Career preparation:** Develop our unique and valuable skills

Introduction

Why Graphs?

Gap



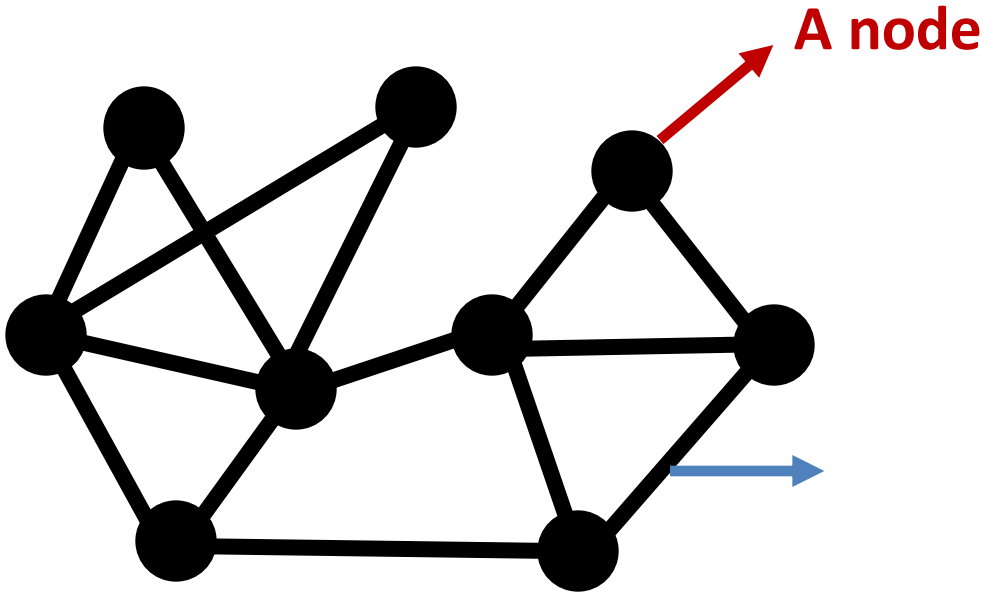
Interconnected world

Modern ML

How to Represent **Interconnected Data?**

Interconnected world

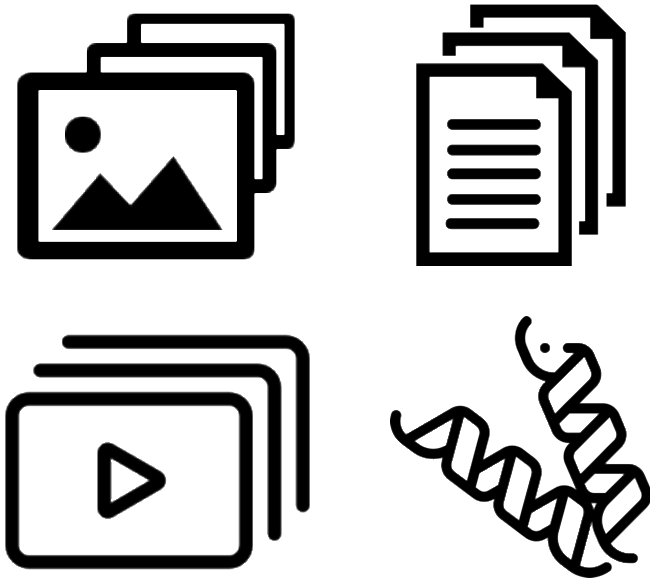
Represent



Graph-structured data

Graph: The language for **describing entities with relations**

Gap



Interconnected world

Modern ML

Goal of Graph Deep Learning

Mine valuable insights from interconnected data

Graph: Ubiquitous across Disciplines

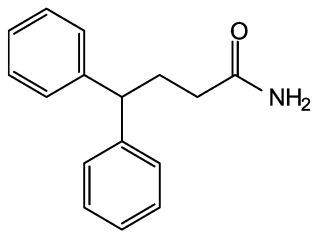


Image credit: MDPI

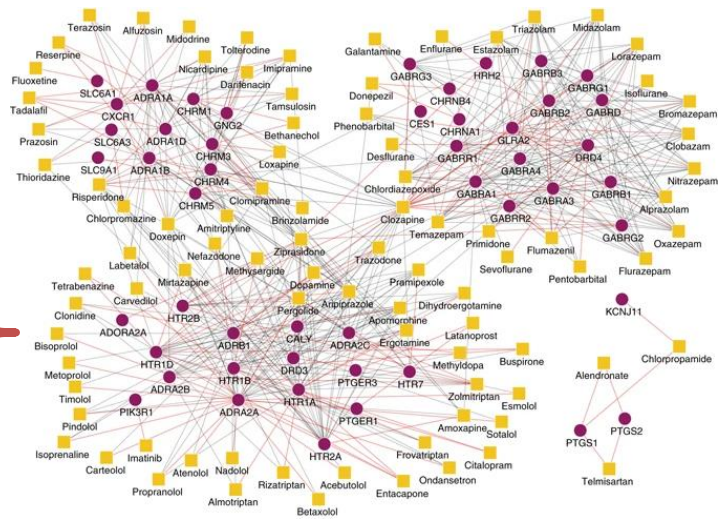


Image credit

Image credit: Medium

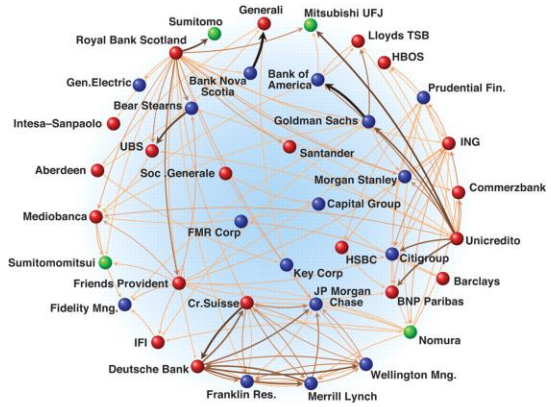


Image credit: Science

Molecule
Molecule design

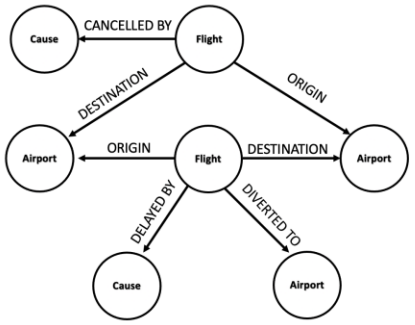
Protein interaction
Drug discovery

Social network
Recommender systems

Economic network
Policy making

- **Graphs:** ***flexible*** and ***expressive***
- **Graphs** can **bridge interdisciplinary data**

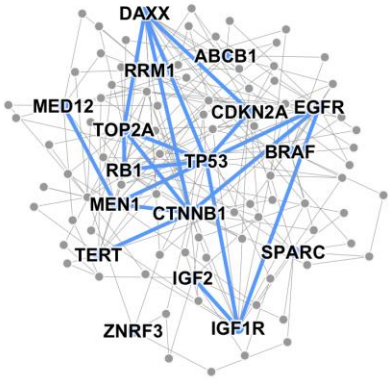
Many Types of Data are Graphs (1)



Event Graphs

Image credit: SalientNetworks

Computer Networks



Disease Pathways

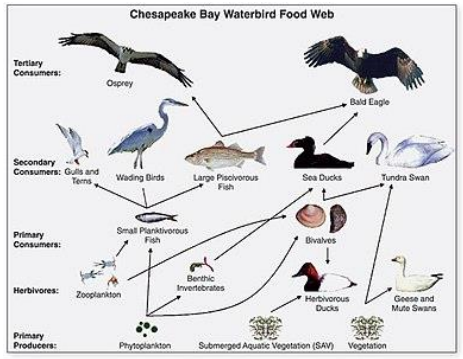


Image credit: Wikipedia

Food Webs

Image credit: Pinterest

Particle Networks

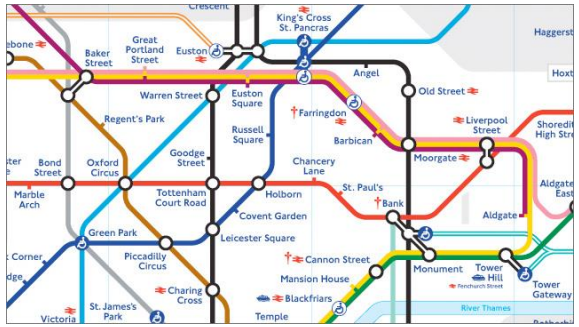


Image credit: visitlondon.com

Underground Networks

Many Types of Data are Graphs (2)

Image credit: Medium

Social Networks

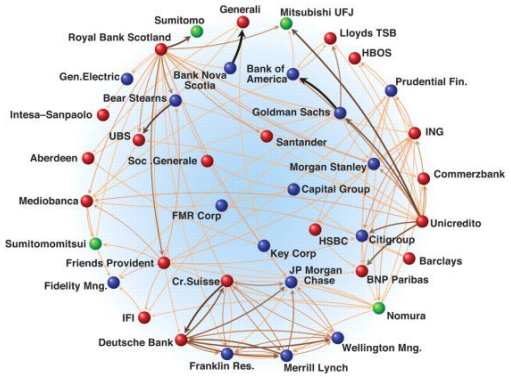


Image credit: Science

Economic Networks

Image credit: Lumen Learning

Communication Networks

Citation Networks

Image credit: Missoula Current News

Internet

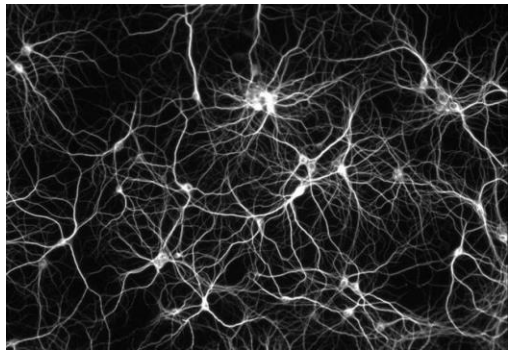


Image credit: The Conversation

Networks of Neurons

Many Types of Data are Graphs (3)

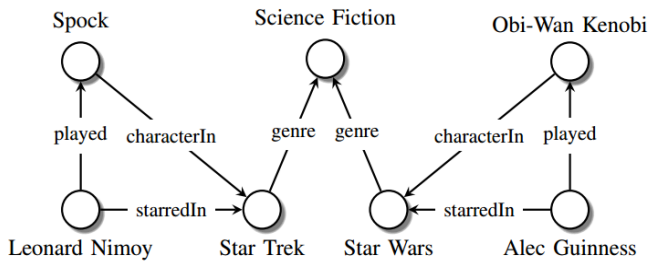


Image credit: Maximilian Nickel et al

Knowledge Graphs

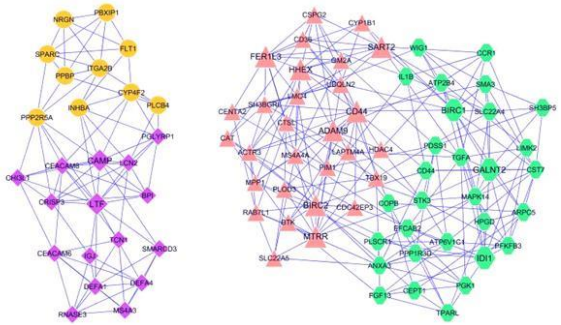


Image credit: ese.wustl.edu

Regulatory Networks

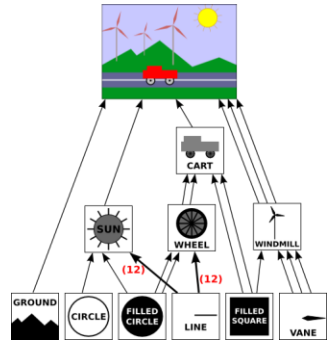


Image credit: math.hws.edu

Scene Graphs

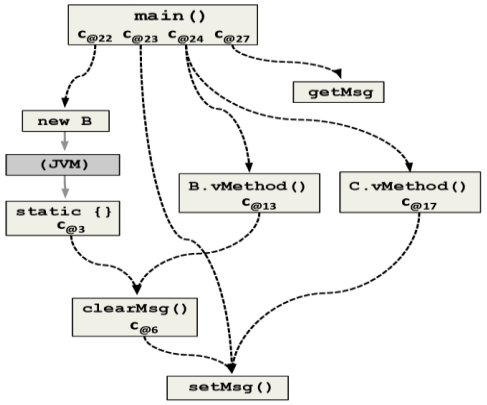


Image credit: ResearchGate

Code Graphs

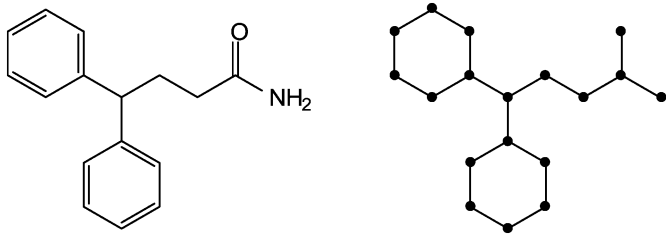


Image credit: MDPI

Molecules

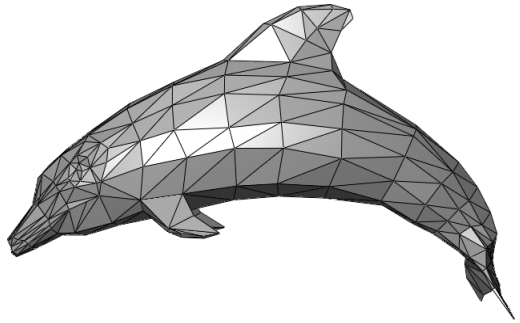


Image credit: Wikipedia

3D Shapes

Graph Machine Learning

Machine learning: mine insights from data

Observation:

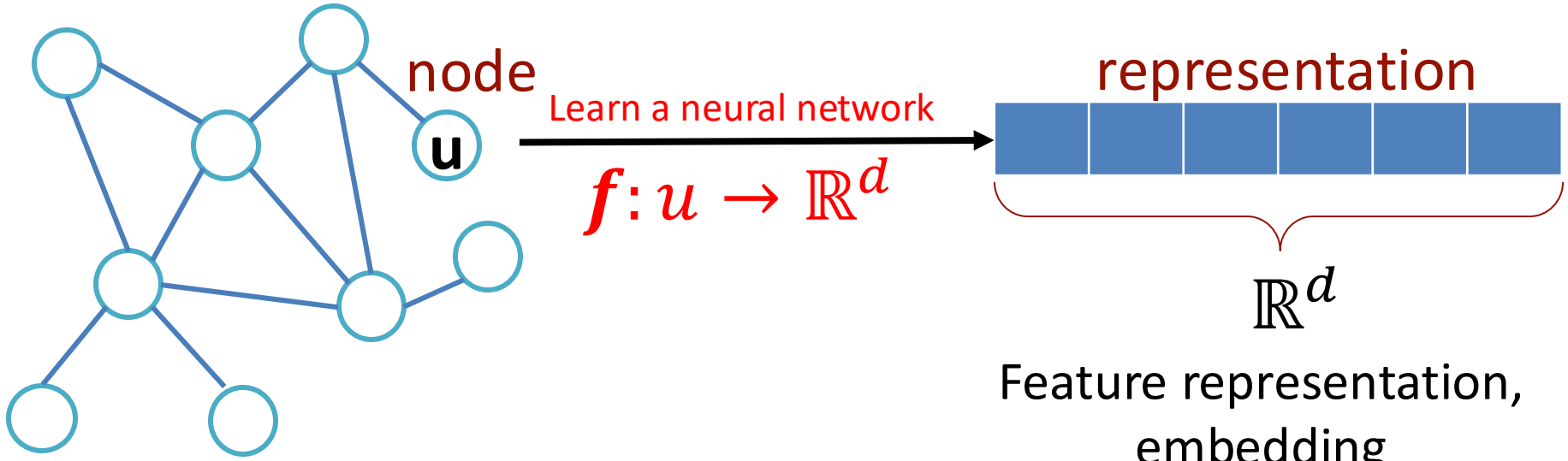
- Relational information – graphs – are ubiquitous
- Standard ML methods do not (explicitly) model relations

Graph machine learning:

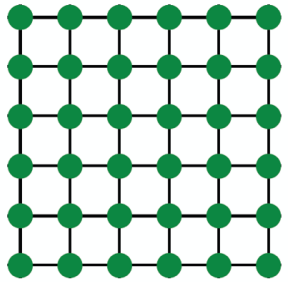
- How do we take advantage of **relational information** for better data mining?

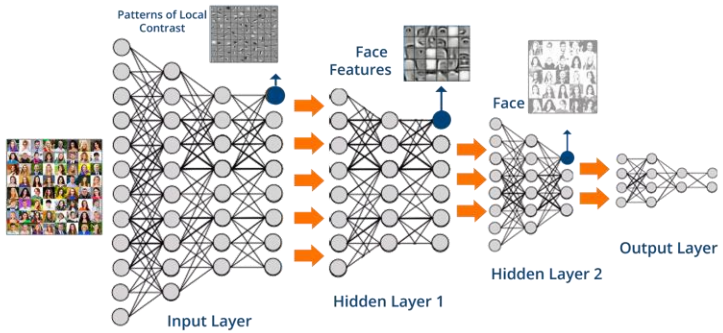
Graph Representation Learning

- Map nodes to d-dimensional embeddings to fit an objective function
  - E.g., Similar nodes in the graph are embedded close together



Modern Deep Learning Toolbox

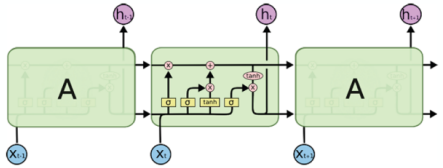




Images

Audio signals

Modern deep learning toolbox is designed for simple sequences & grids.



Text/Speech

Images

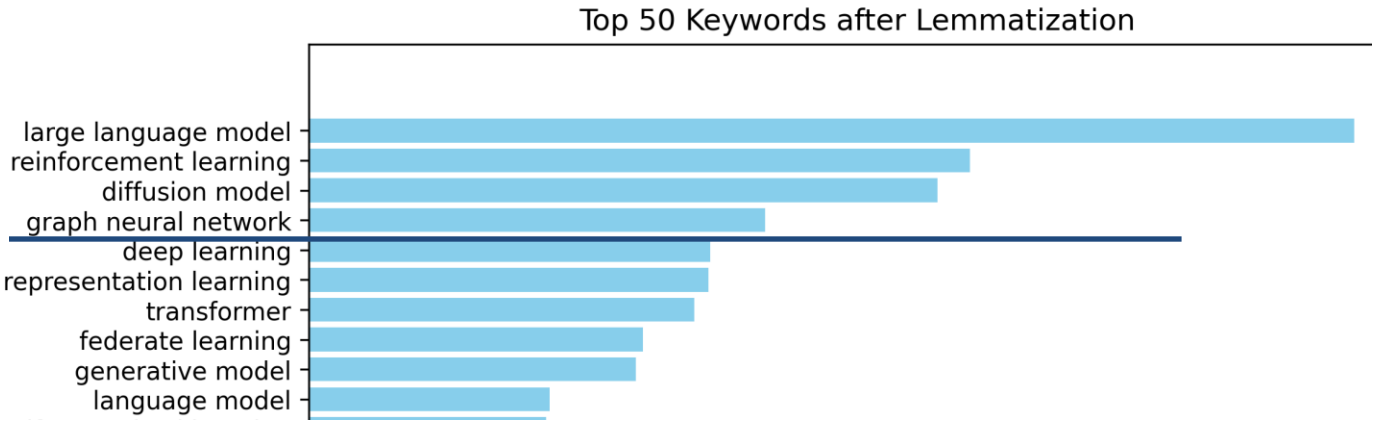
Modern ML -> Graph ML

- Not everything can be represented as a sequence or a grid.
- How can we develop neural networks that are much more broadly applicable, beyond sequences and images?

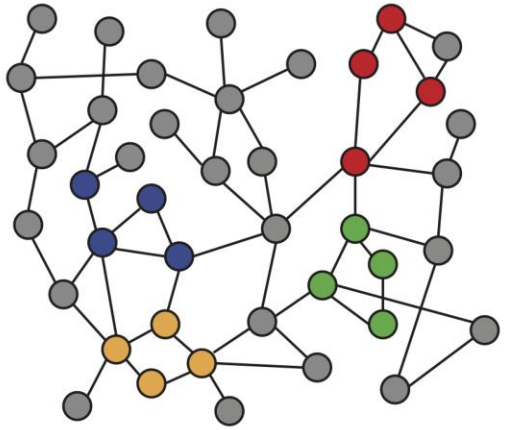
- **Graph neural network** is a new class of neural networks, representing a new frontier of deep learning

Graph - Hot Subfield in AI

- ICLR 2024 keyword visualization

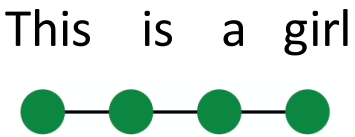


Machine Learning with Graphs is Hard

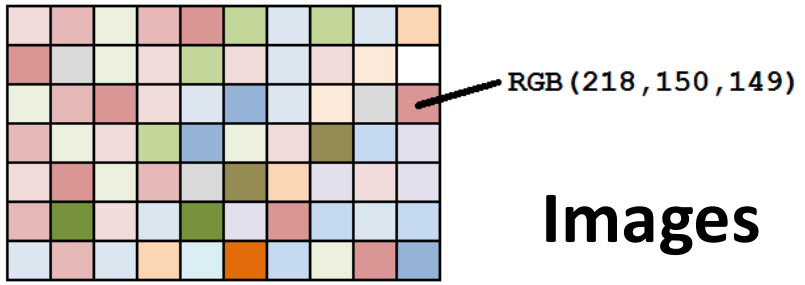


Graphs

vs.



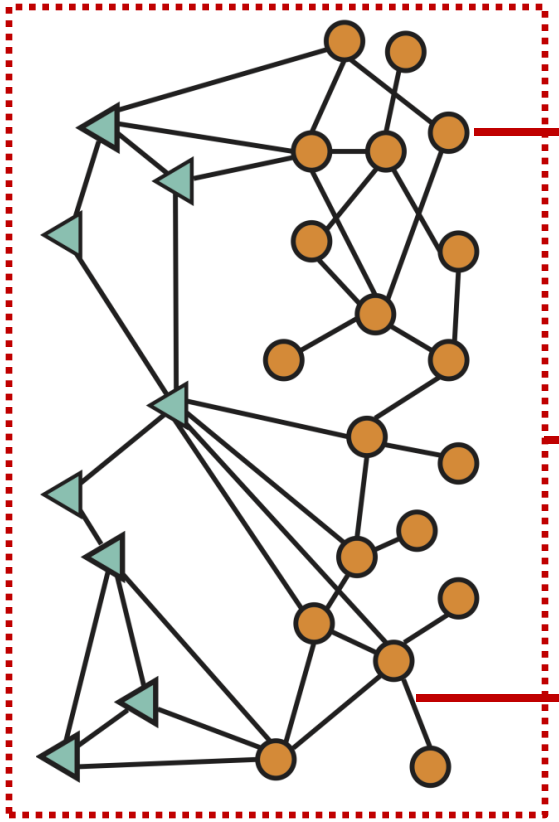
Text



Images

- **Arbitrary size** and topological **structure**
- Nodes have **no fixed ordering**

Graph Machine Learning **Tasks**



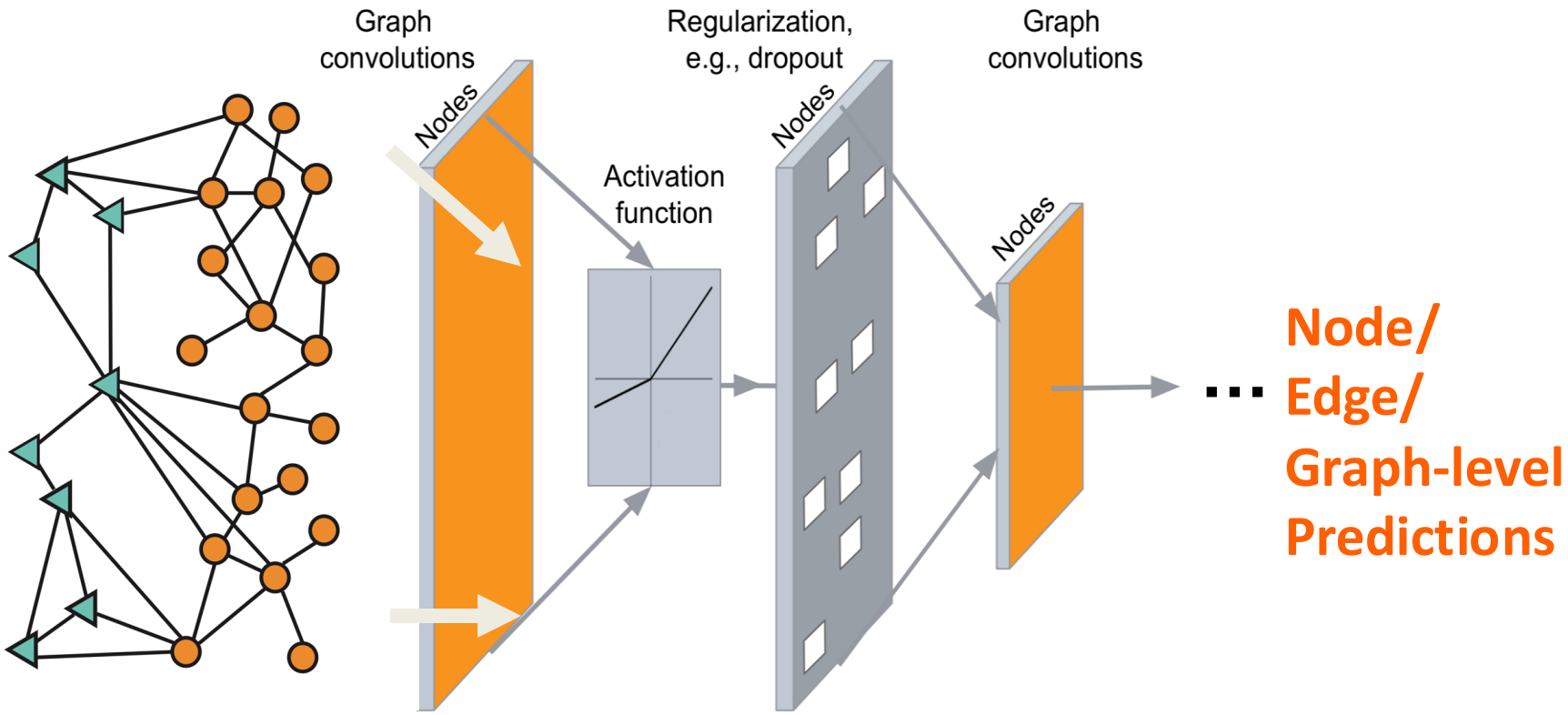
Node-level prediction
"Classify user by their type in a social network"

Graph-level prediction
"Predict which molecules are drug-like"

Edge-level prediction
"Recommend item nodes to user nodes"

Deep Learning with Graphs

Input:
Network



Node/ Edge/ Graph-level Predictions

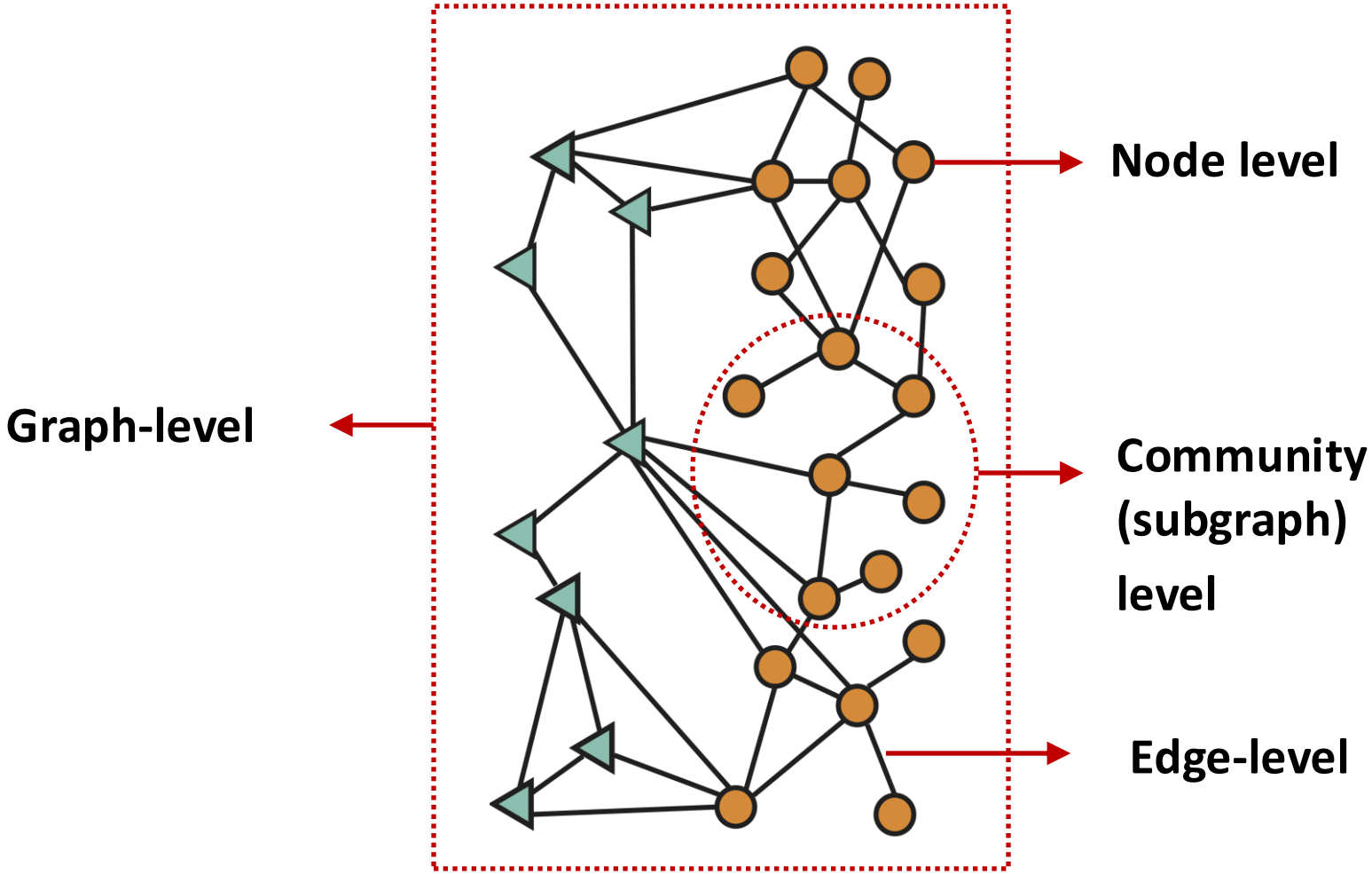
Reminder: Use Graphs Wisely

- Graph is a general language, but don't over-use it!
  - E.g., graphs subsumes lattices/sequences
- Suggested checklist
  - Will representing my data as graphs bring more information?
    - If your graph can be fully induced from existing data, think twice
      - E.g., construct K nearest neighbor graph from your embeddings, worth it?
  - Will representing my data as graphs lose information?
    - Graphs are unordered. Graph nodes/edges need features
      - E.g., molecule as graphs -> lose the distance information -> add as edge feature
  - Are there more efficient alternative representations?
    - Understand the trade-off between expressiveness & efficiency. How important is the relational info?
      - E.g., images as grids, text as sequences

Introduction

Applications of Graph ML

Different Types of Tasks



Classic Graph ML Tasks

- **Node classification**: Predict a property of a node
  - Example: Categorize online users / items
- **Link prediction**: Predict whether there are missing links between two nodes
  - Example: Knowledge graph completion
- **Graph classification**: Categorize different graphs
  - Example: Molecule property prediction
- **Clustering**: Detect if nodes form a community
  - Example: Social circle detection
- Other tasks:
  - **Graph generation**: Drug discovery
  - **Graph evolution**: Physical simulation

Classic Graph ML Tasks

- **Node classification**: Predict a property of a node
  - Example: Cat
- **Link prediction**
  - Example: Kn
- **Graph classific**
  - Example: Mo
- **Clustering**: Det
  - Example: Social circle detection
- Other tasks:
  - **Graph generation**: Drug discovery
  - **Graph evolution**: Physical simulation

veen two nodes

These Graph ML tasks lead to high-impact applications!

Example (1): Protein Folding

- A protein chain acquires its native 3D structure.

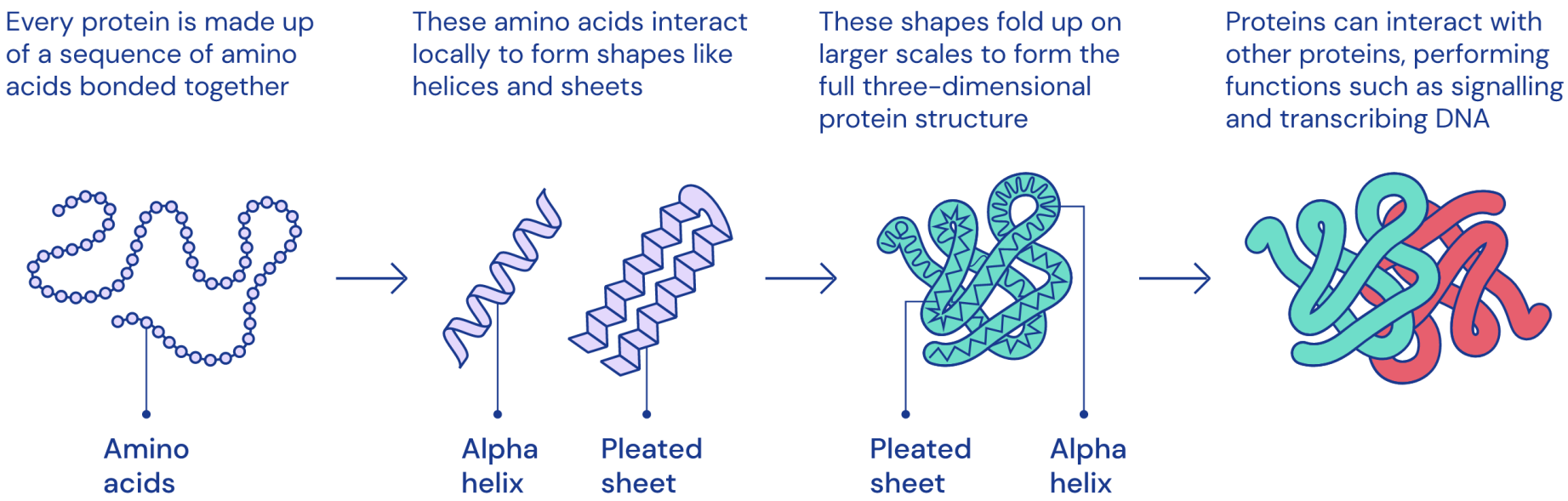


Image credit: DeepMind

Example (1): Protein Folding

- The Protein Folding Problem: Computationally predict a protein's 3D structure based solely on its amino acid sequence.

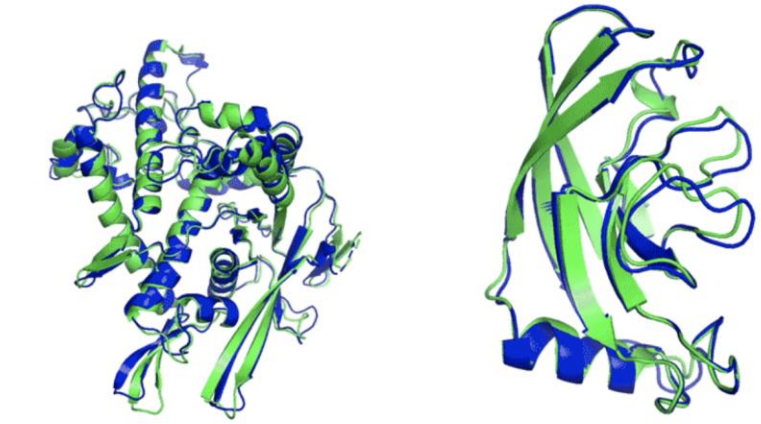
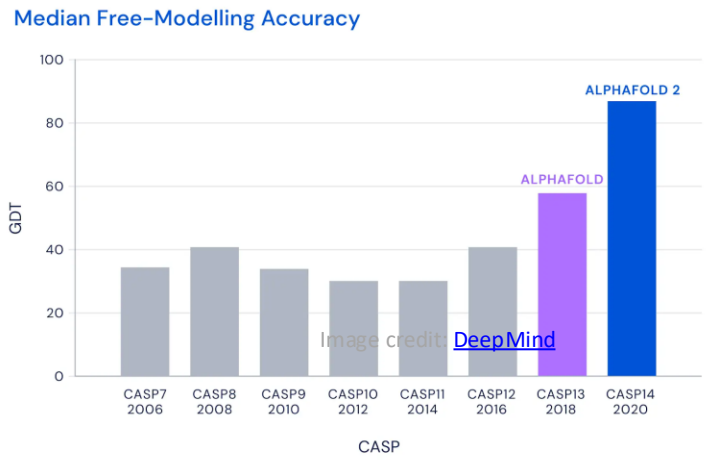


Image credit: DeepMind

AlphaFold: Impact



DeepMind's latest AI breakthrough can accurately predict the way proteins fold

AlphaFold's AI could change the world of biological science as we know it

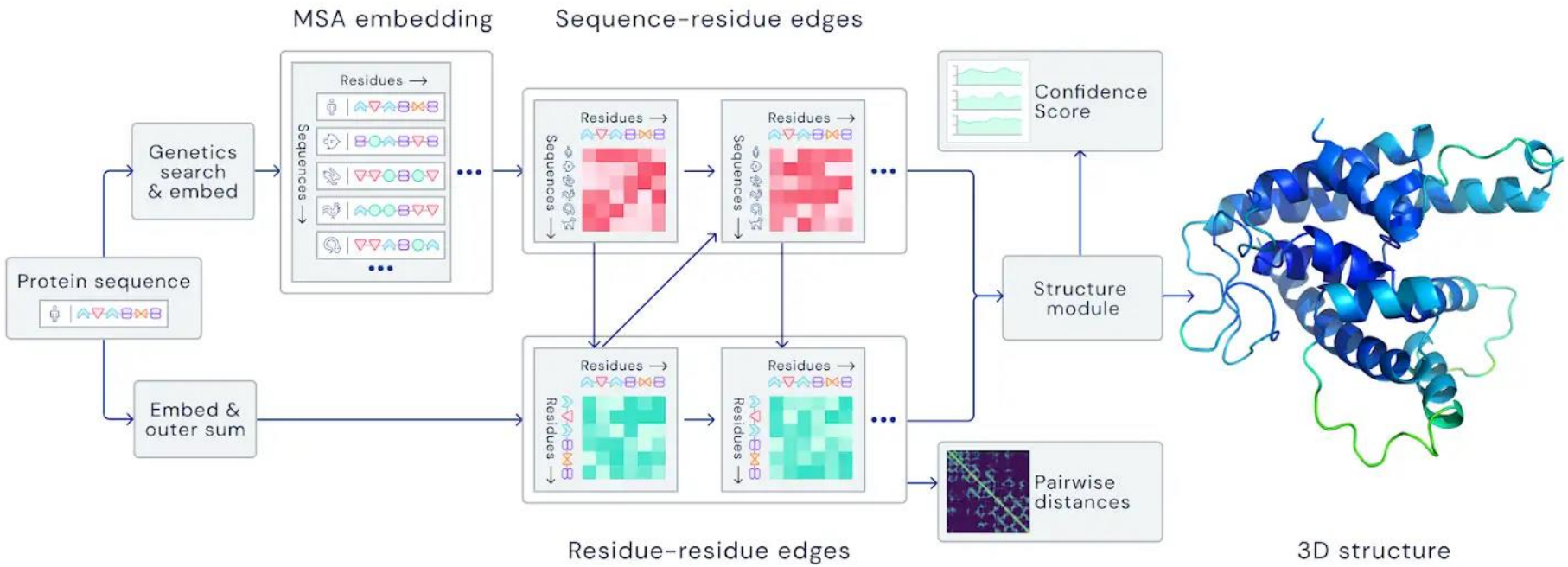
12-14-20

DeepMind's latest AI breakthrough could turbocharge drug discovery

Has Artificial Intelligence 'Solved' Biology's Protein-Folding Problem?

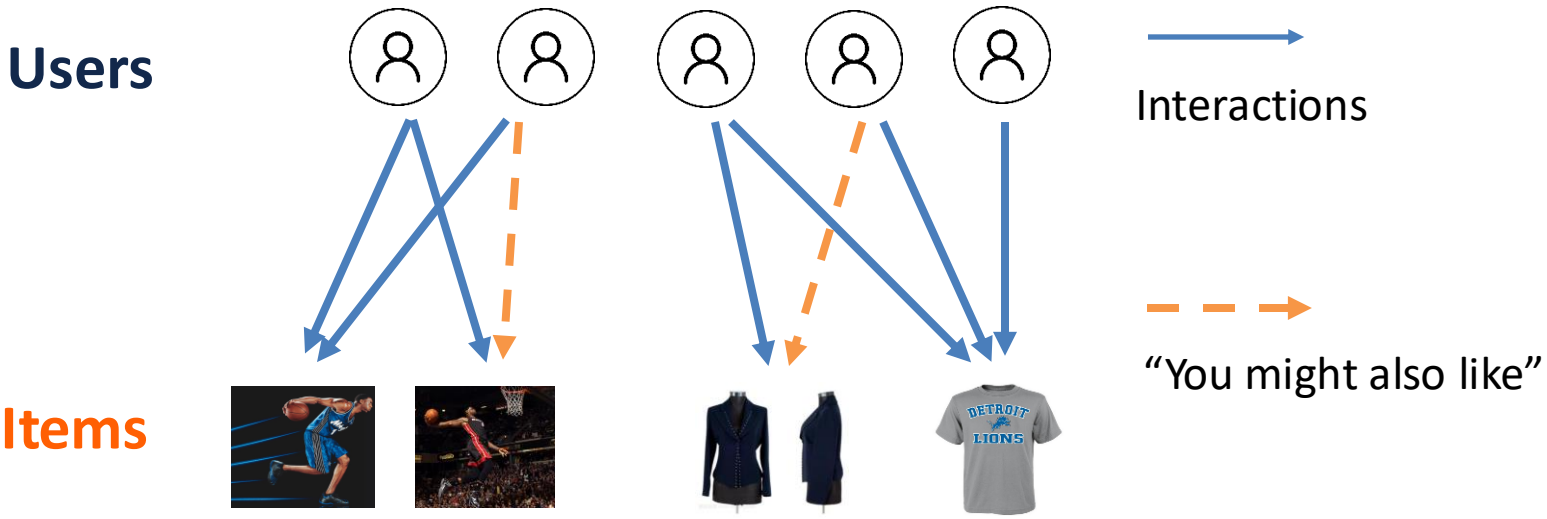
AlphaFold: Solving Protein Folding

- Key idea: “Spatial graph”
  - **Nodes**: Amino acids in a protein sequence
  - **Edges**: Proximity between amino acids (residues)



Example (2): Recommender Systems

- **Users interacts with items**
  - Watch movies, buy merchandise, listen to music
  - Nodes: Users and items
  - Edges: User-item interactions
- **Goal: Recommend items users might like**

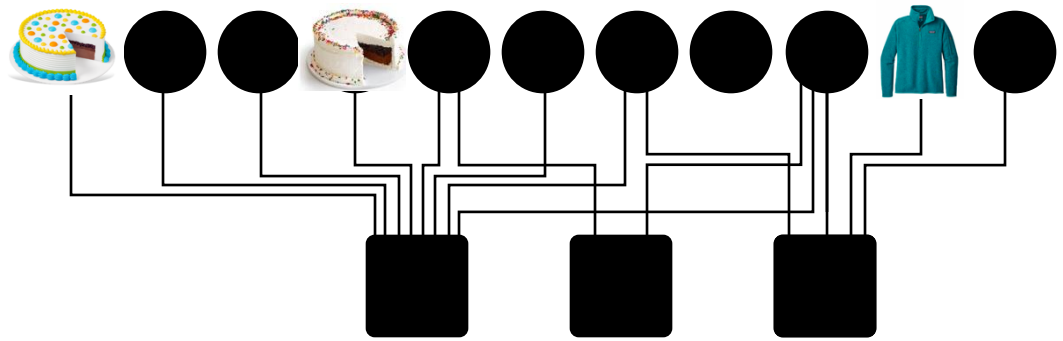


PinSage: Graph-based Recommender

- Task: Recommend related pins to users

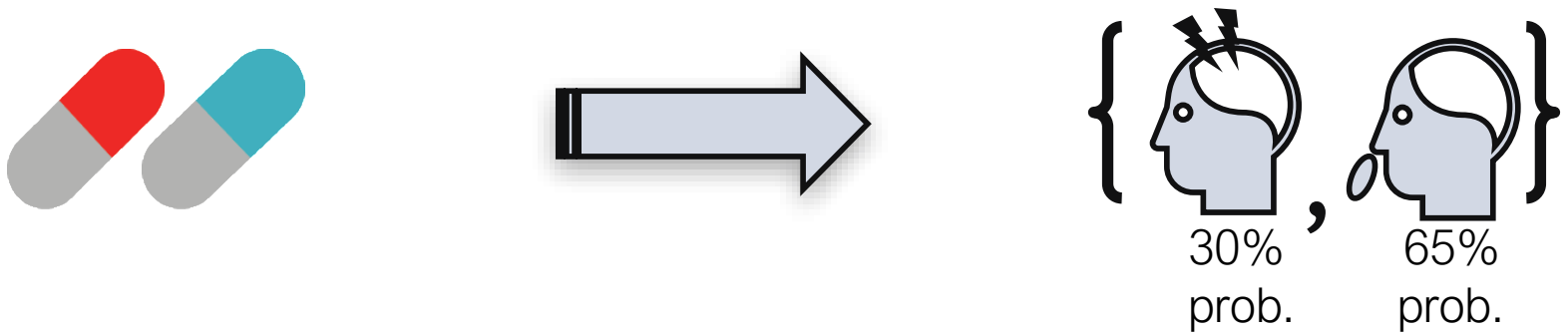
Task: Learn node embeddings z_i such that $d(z_{cake1}, z_{cake2}) < d(z_{cake1}, z_{sweater})$

Predict whether two nodes in a graph are related



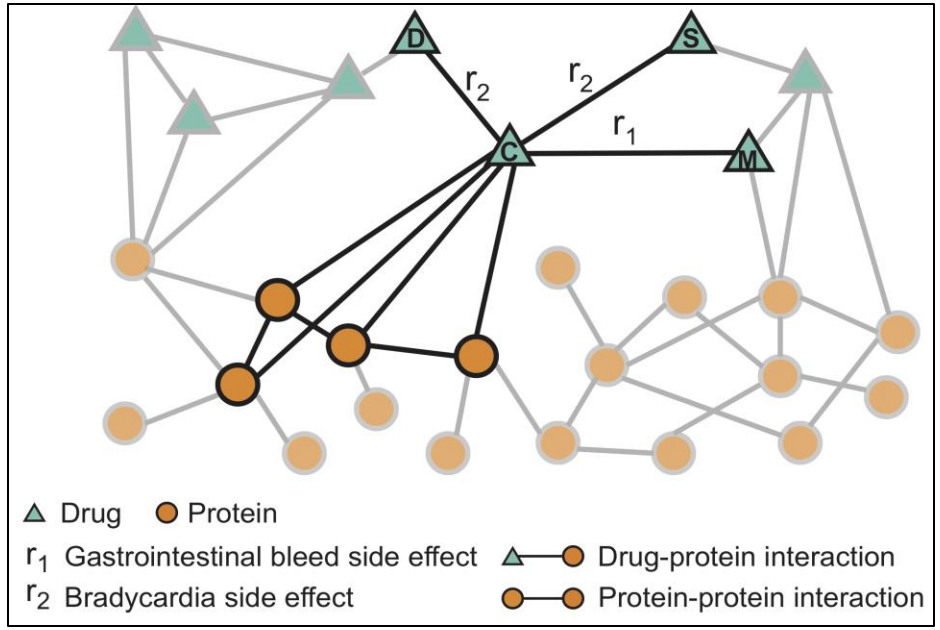
Example (3): Drug Side Effects

- Many patients take multiple drugs to treat complex or co-existing diseases:
  - 46% of people ages 70-79 take more than 5 drugs
  - Many patients take more than 20 drugs to treat heart disease, depression, insomnia, etc.
- **Task: Given a pair of drugs predict adverse side effects.**

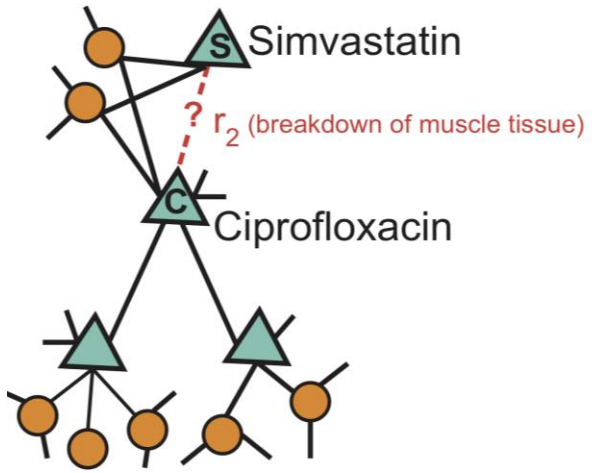


Biomedical Graph Link Prediction

- **Nodes**: Drugs & Proteins
- **Edges**: Interactions



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



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

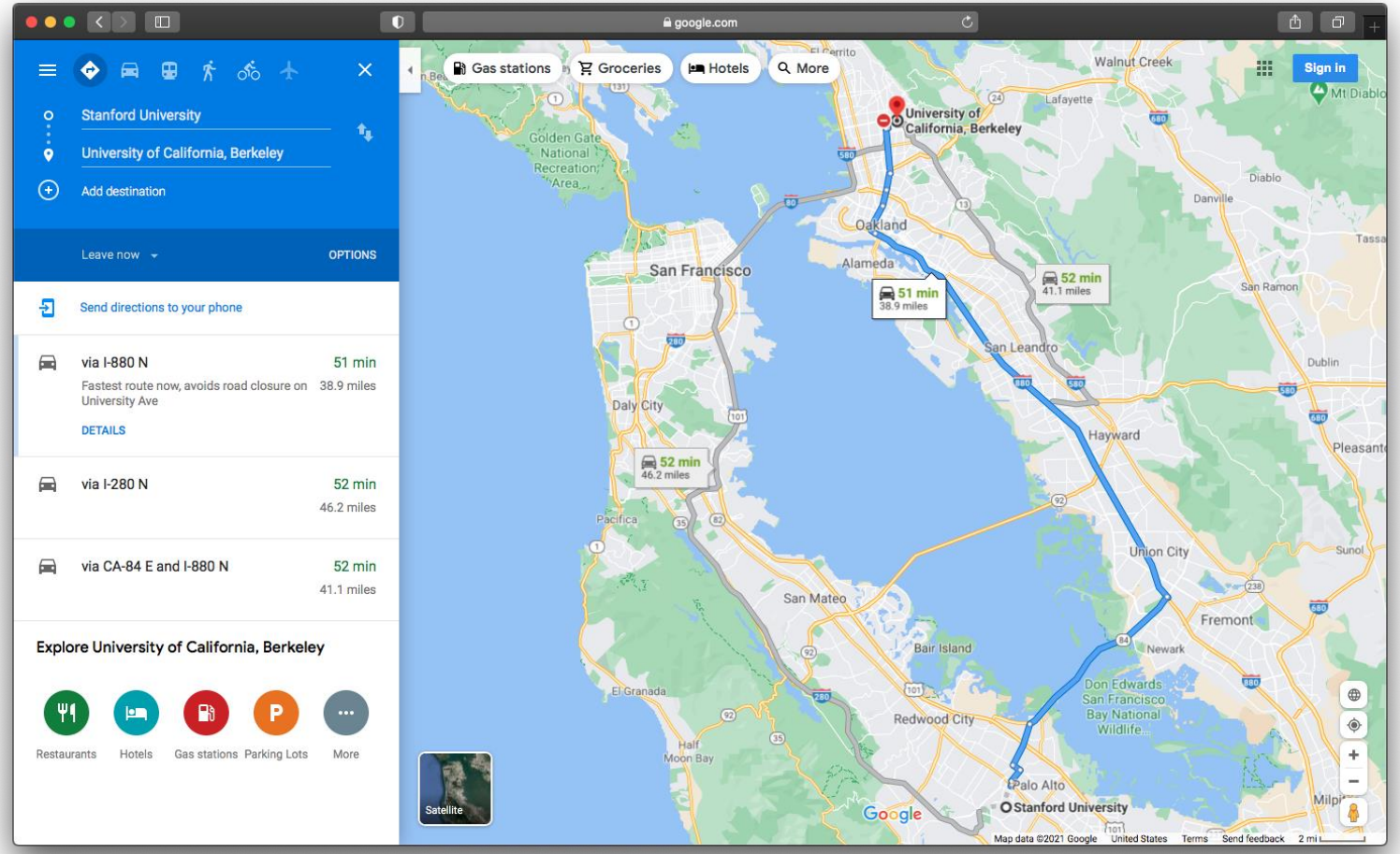
Results: *De novo* Predictions

Rank	Drug c	Drug d	Side effect r	Evidence found
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	

Case Report

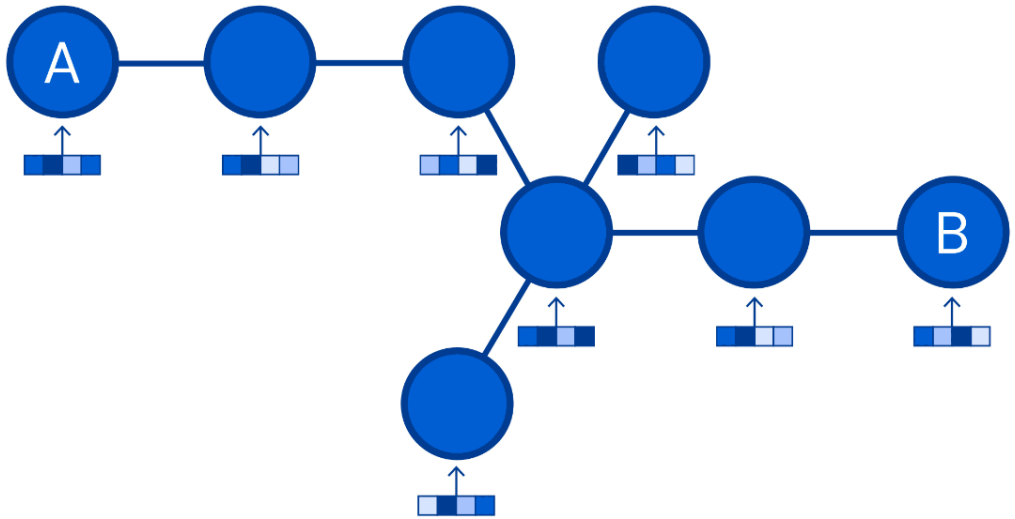
Severe Rhabdomyolysis due to Presumed Drug Interactions between Atorvastatin with Amlodipine and Ticagrelor

Example (4): Traffic Prediction



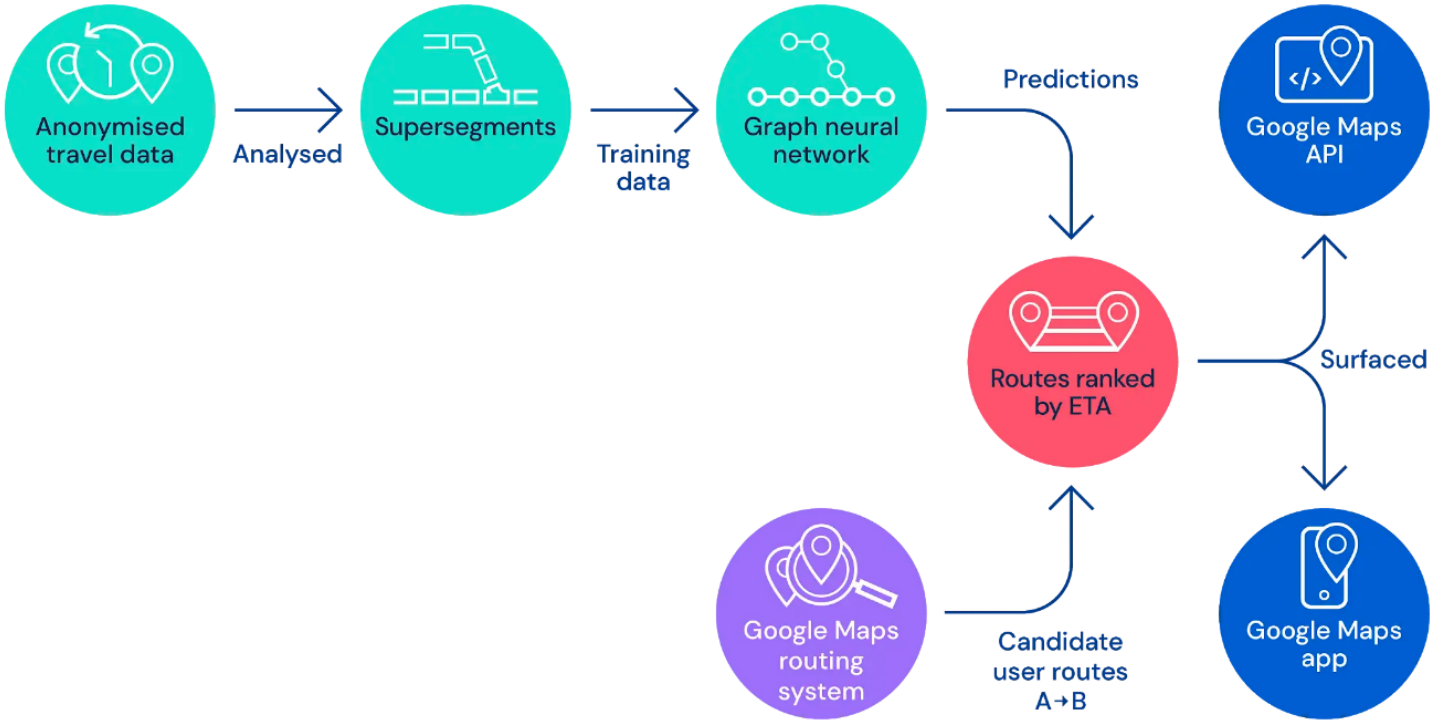
Road Network as a Graph

- **Nodes**: Road segments
- **Edges**: Connectivity between road segments
- **Prediction**: Time of Arrival (ETA)



Traffic Prediction via GNN

- Predicting Time of Arrival with Graph Neural Networks

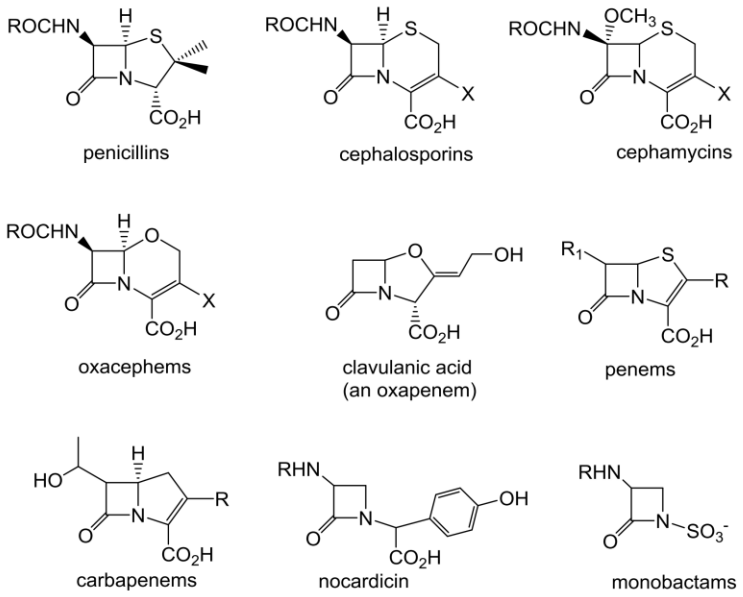


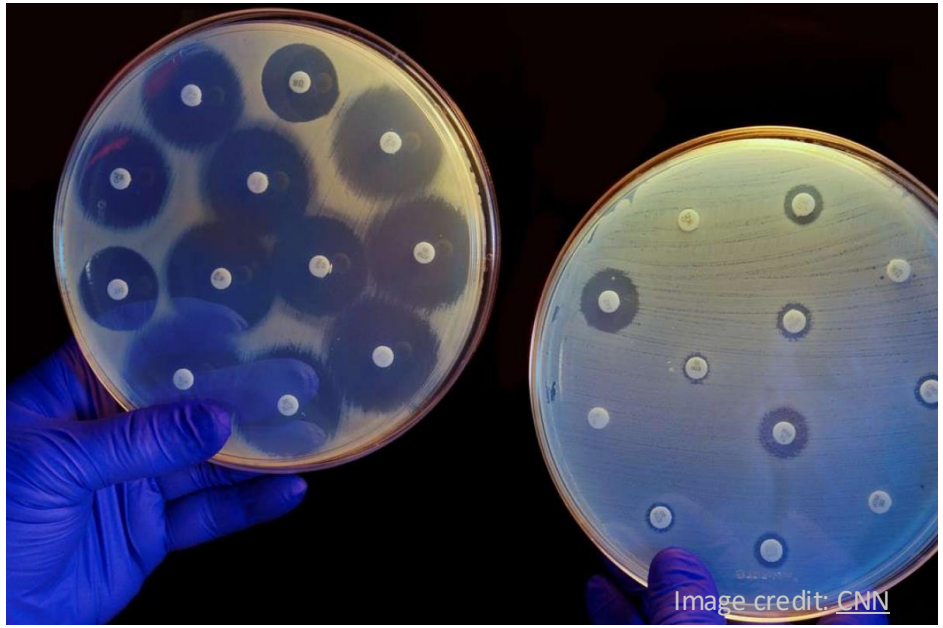
THE MODEL ARCHITECTURE FOR DETERMINING OPTIMAL ROUTES AND THEIR TRAVEL TIME.

Used in Google Maps

Example (5): Drug Discovery

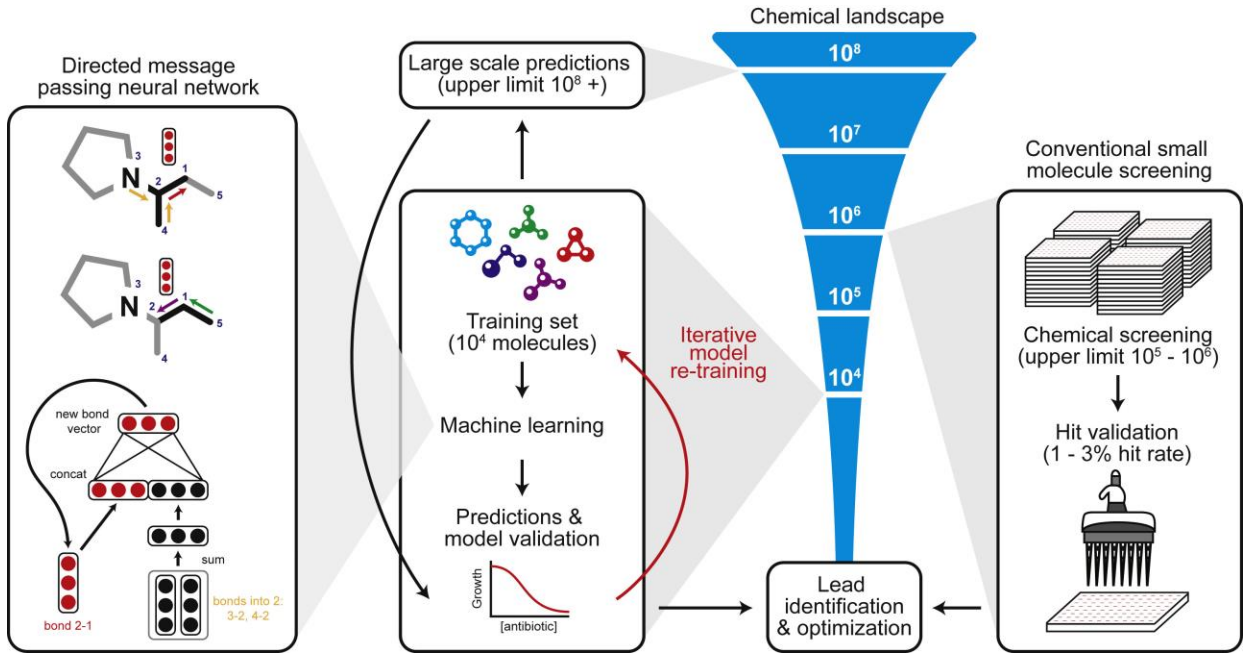
- Antibiotics are small molecular graphs
  - **Nodes**: Atoms
  - **Edges**: Chemical bonds





Deep Learning for Antibiotic Discovery

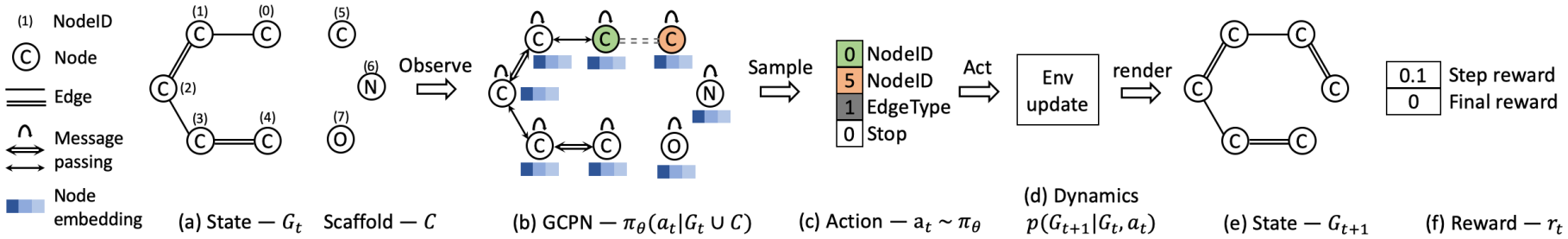
- A Graph Neural Network **graph classification model**
- Predict promising molecules from a pool of candidates



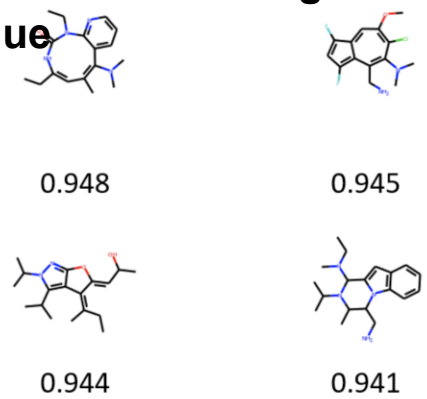
Stokes et al., A Deep Learning Approach to Antibiotic Discovery, Cell 2020

Molecule Generation / Optimization

- **Graph generation**: Generating novel molecules

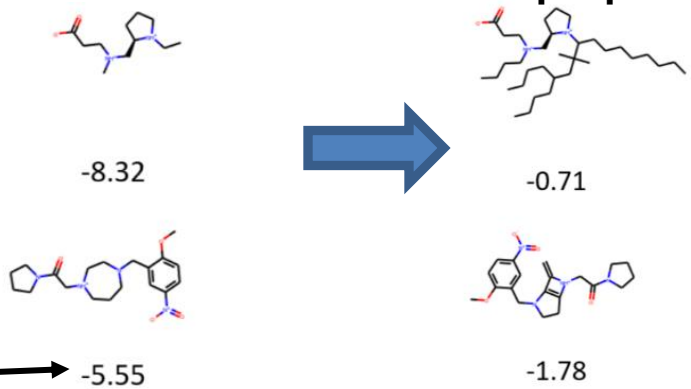


Use case 1: Generate novel molecules with high Drug likeness value



Drug likeness

Use case 2: Optimize existing molecules to have desirable properties



You et al., Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation, NeurIPS 2018

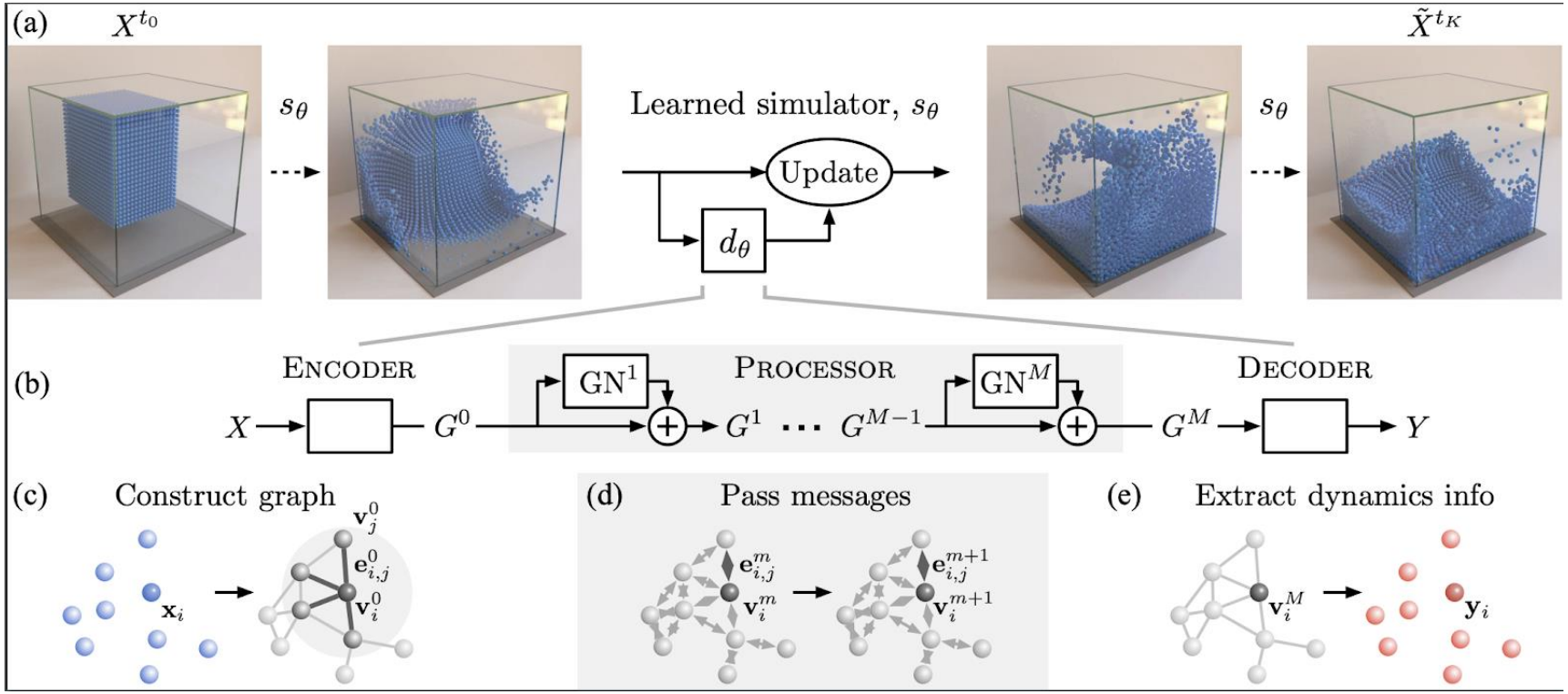
Example (6): Physics Simulation

- Physical simulation as a graph:
  - **Nodes**: Particles
  - **Edges**: Interaction between particles

Sanchez-Gonzalez et al., Learning to simulate complex physics with graph networks, ICML 2020

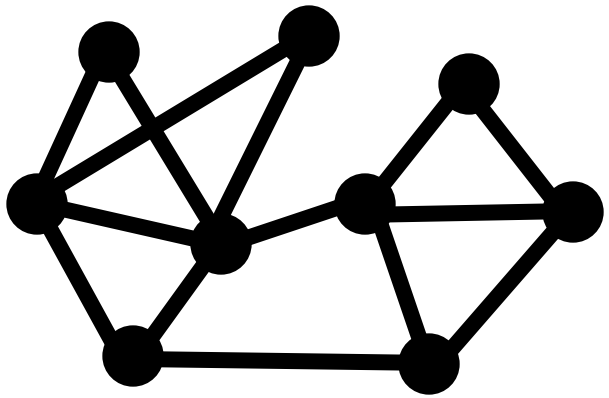
Simulation Learning Framework

- **A graph evolution task**:
  - **Goal**: Predict how a graph will evolve over time

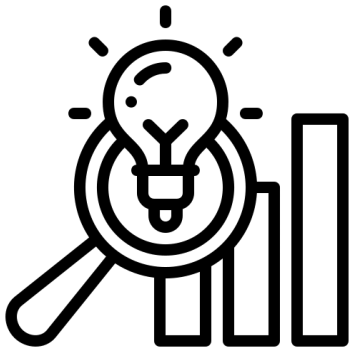


Summary

Represent



Extract



Interconnected world

Graph-structured data

Insights

Graph: A perspective to represent the world and extract insights