GraphSAGE: Deep Learning for Relational Data

Jure Leskovec

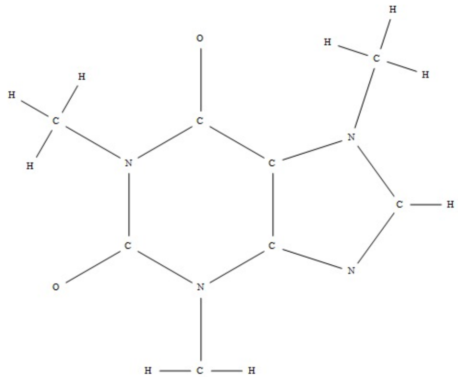
STANFORD INFOLAB

LELAND STANFORD JUNIOR UNIVERSITY ORGANIZED 1891

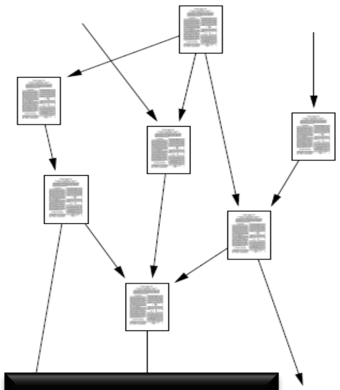
CHAN ZUCKERBERG BIOHUB

Includes joint work with **W. Hu, J. You**, **R. Ying. H. Ren**, M. Fey, Y. Dong, B. Liu, M. Catasta, M. Zitnik, P. Eksombatchai, W. Hamilton

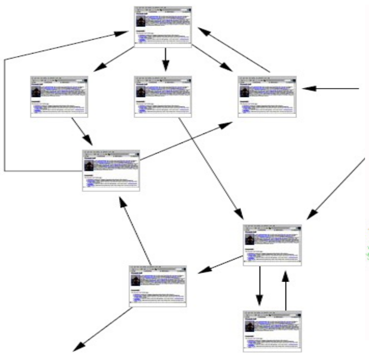
Networks around us!



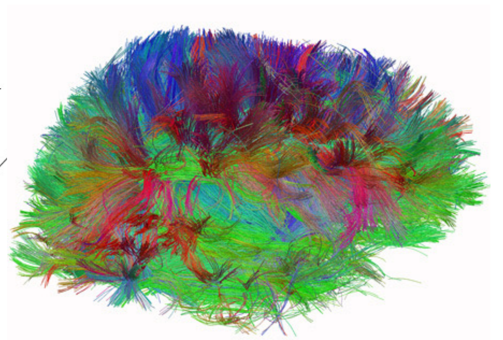
Molecules



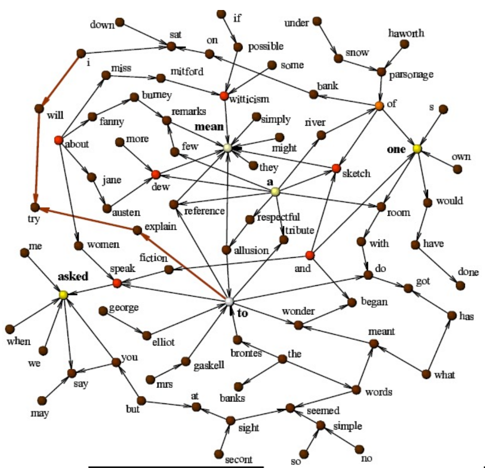
Knowledge



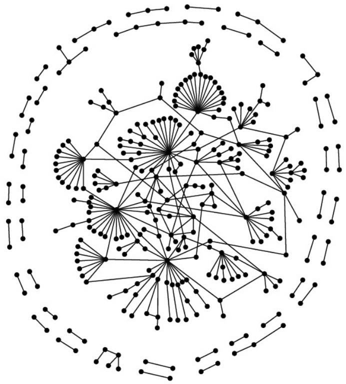
Information



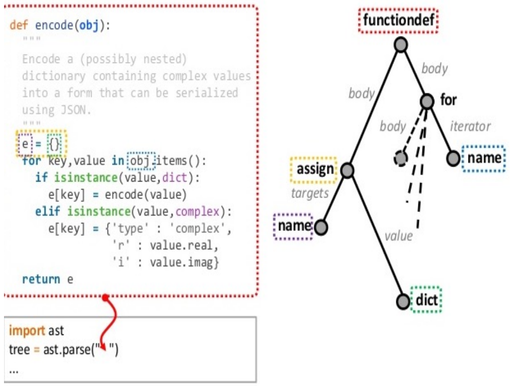
Brain/neurons



Genes



Communication



Software

Social

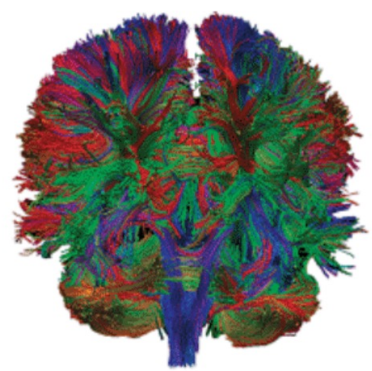
Applications of DL on Graphs

Computer graphics

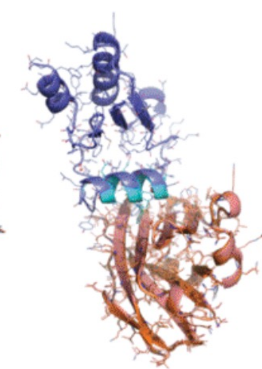
Virtual/augmented reality

Robotics

Autonomous driving

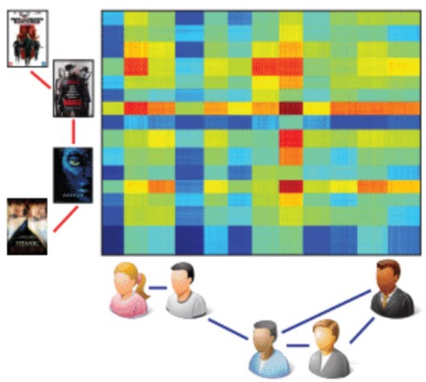


Medicine

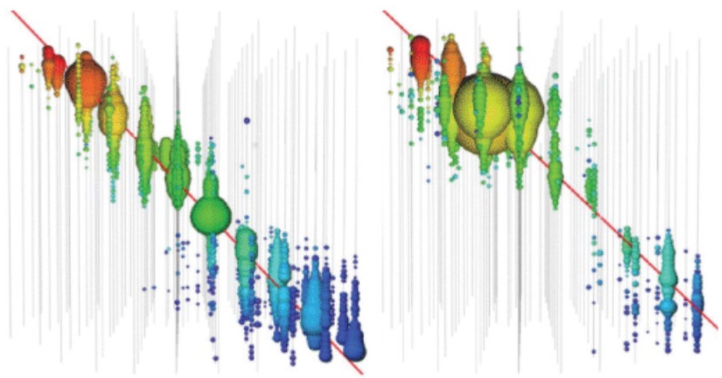


Drug design

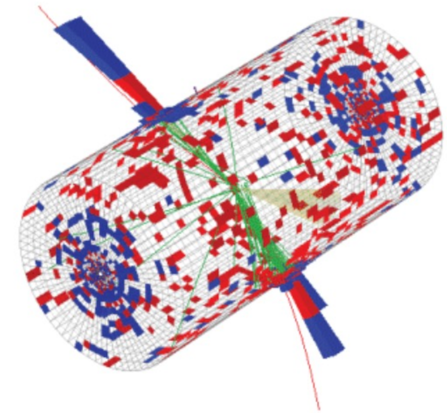
Applications of DL on Graphs



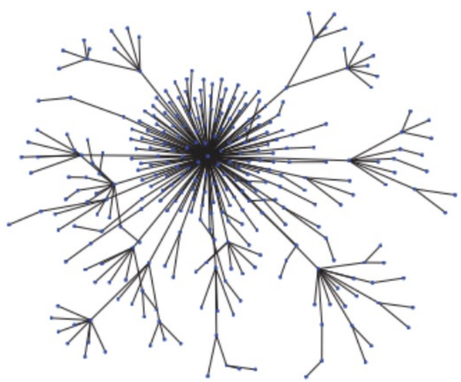
Recommender system



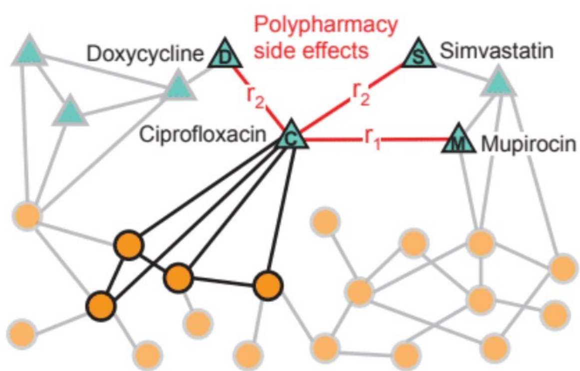
Neutrino detection



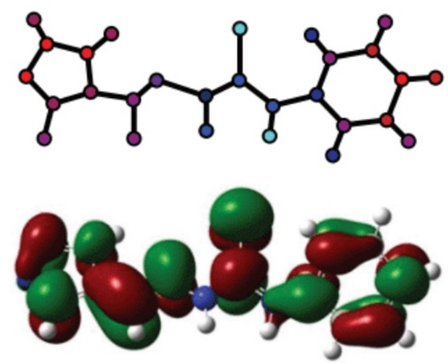
LHC



Fake news detection

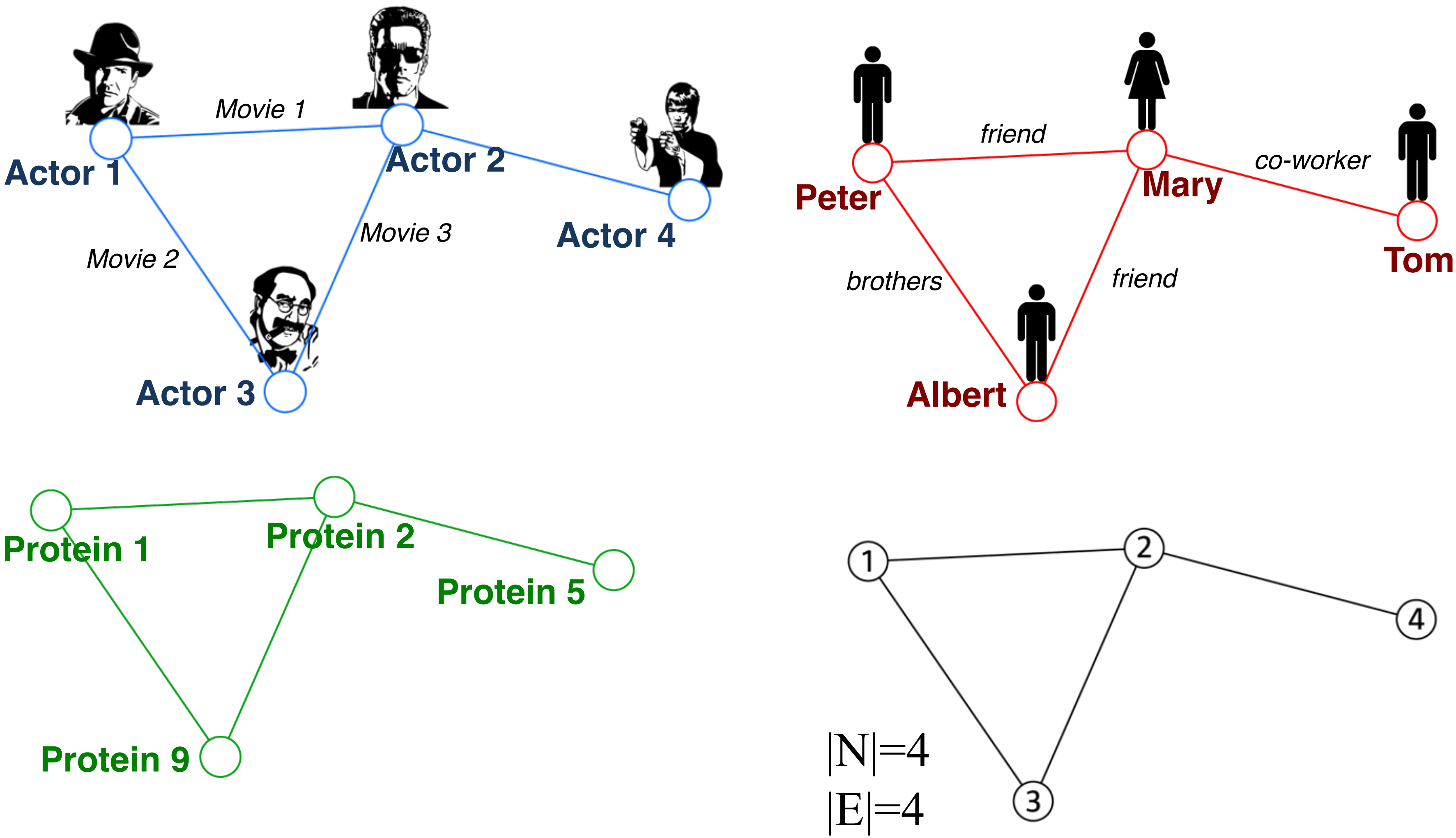


Drug repurposing



Chemistry

Graphs: Common Language



Our Collaborations

- We work with many external organizations
- Discuss and identify big problems
- Obtain and anonymize data, get consent/IRB

Fundamental research, results in public domain

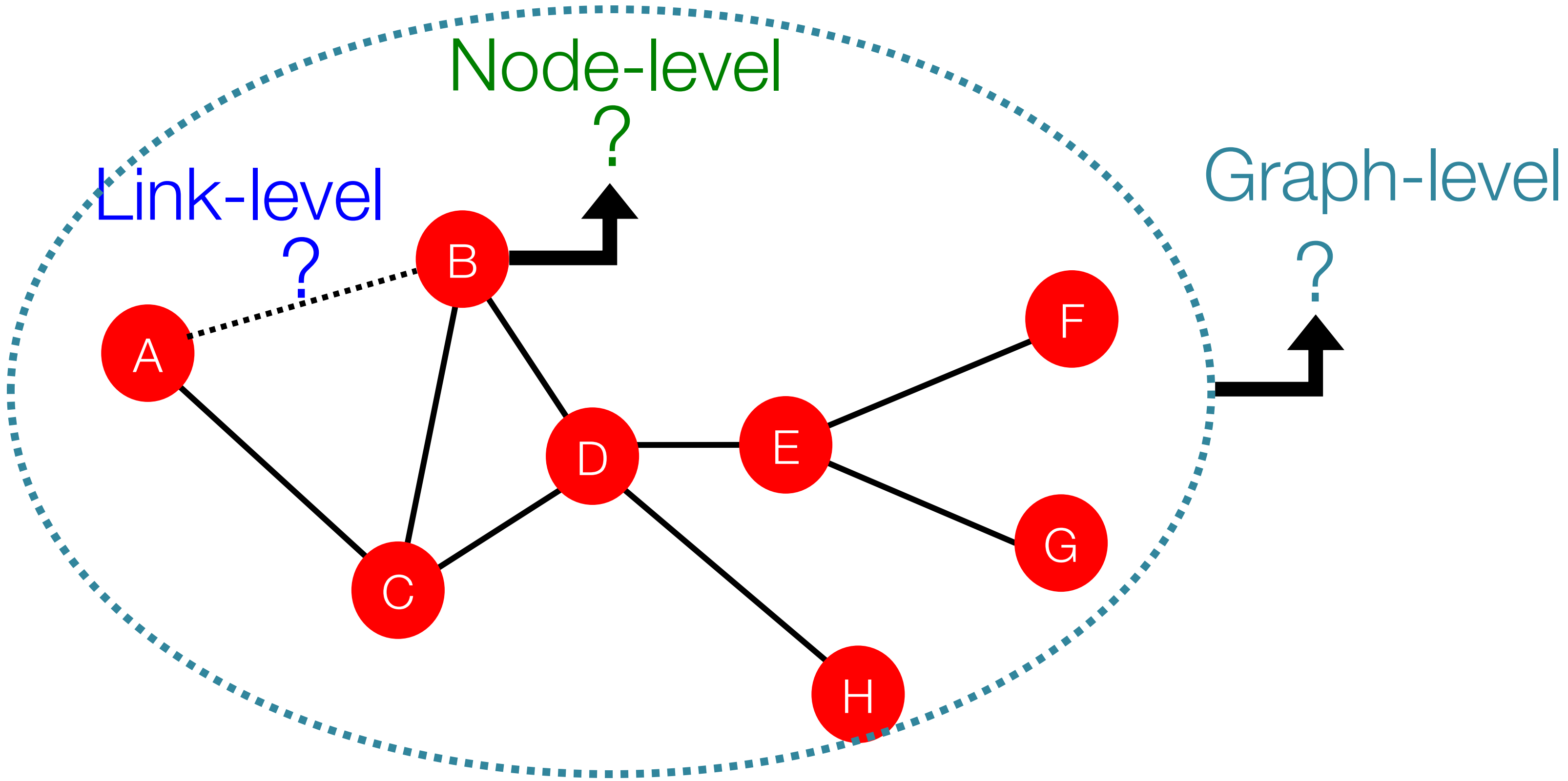
New Ways of Thinking

Working on real-world problems leads to new ways of thinking:

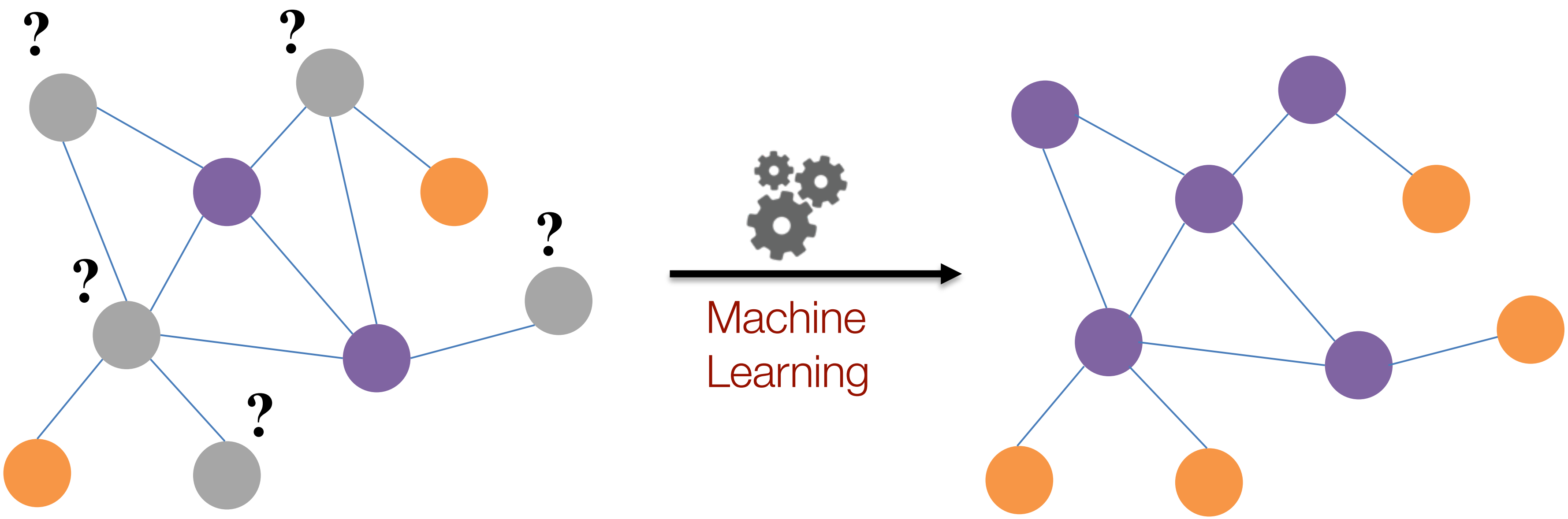
- Incremental algorithmic improvements turn out not to be so important
- More important is methodology and computational modeling of the domain

Leads to new research that would be impossible in isolation

Machine Learning Tasks

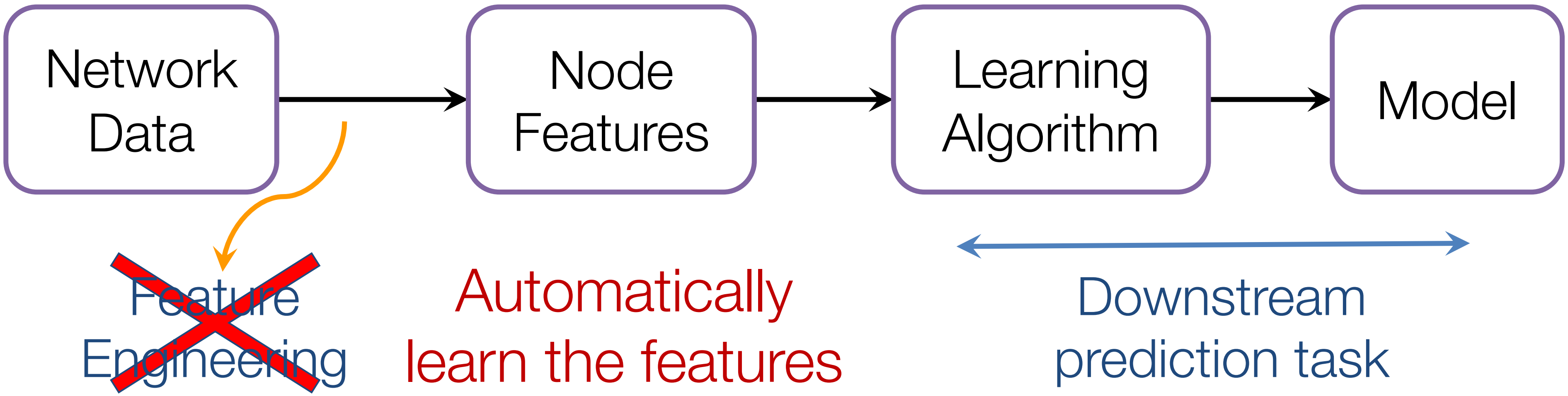


Example: Node Classification



- What users are going to churn?
- What is the disease of a patient?
- What are functions of proteins?

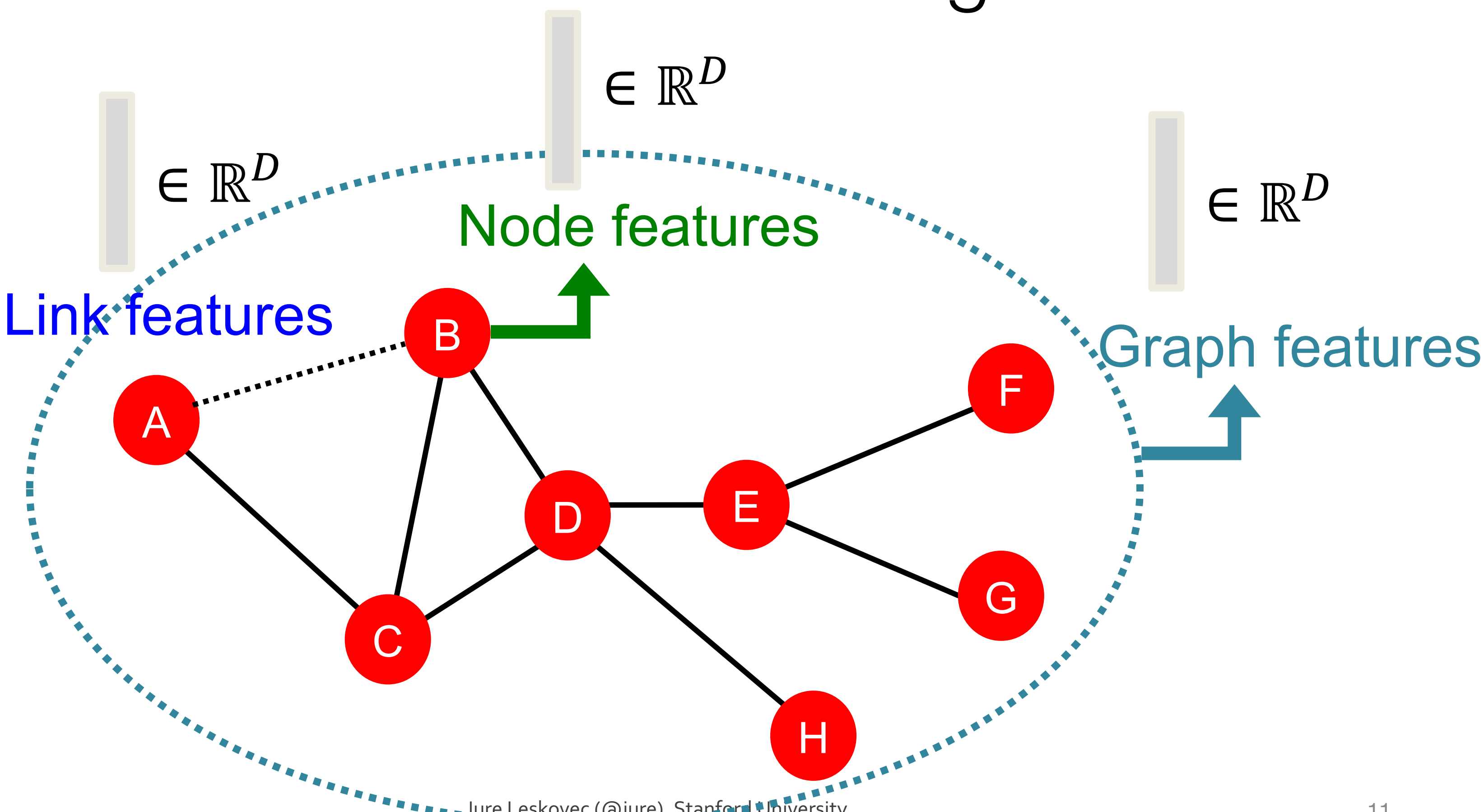
Machine Learning Lifecycle



(Supervised) Machine Learning Lifecycle: This feature, that feature.
Every single time!

Graph Feature Engineering

- Design features for nodes/links/graphs
- Obtain features for all training data



Two Pain Points: One

Data Scientist's pain point #1:

- Data scientists have to hand encode features to solve prediction problems.
- Hand encoding graph features is…

… complex and involves expensive queries

… error prone

… suboptimal

… labor intensive

Two Pain Points: Two

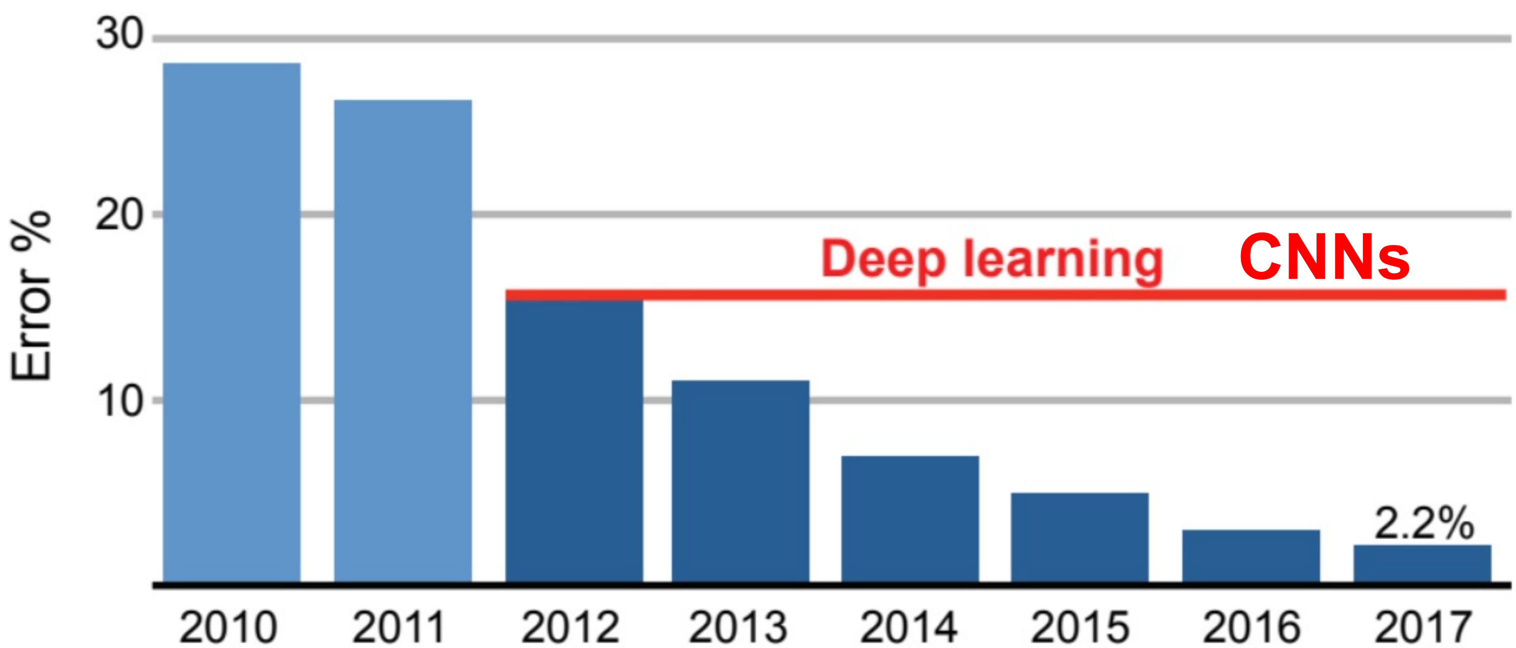
Data Scientist's pain point #2:

- Data is often incomplete.
  - Address Books, Follows, Interests, Protein Protein Interaction, Ancestry
- Entity information is incomplete.
- Predictions often entail completing the "missing information".
  - Relational structure is often not leveraged due to scalability issues.

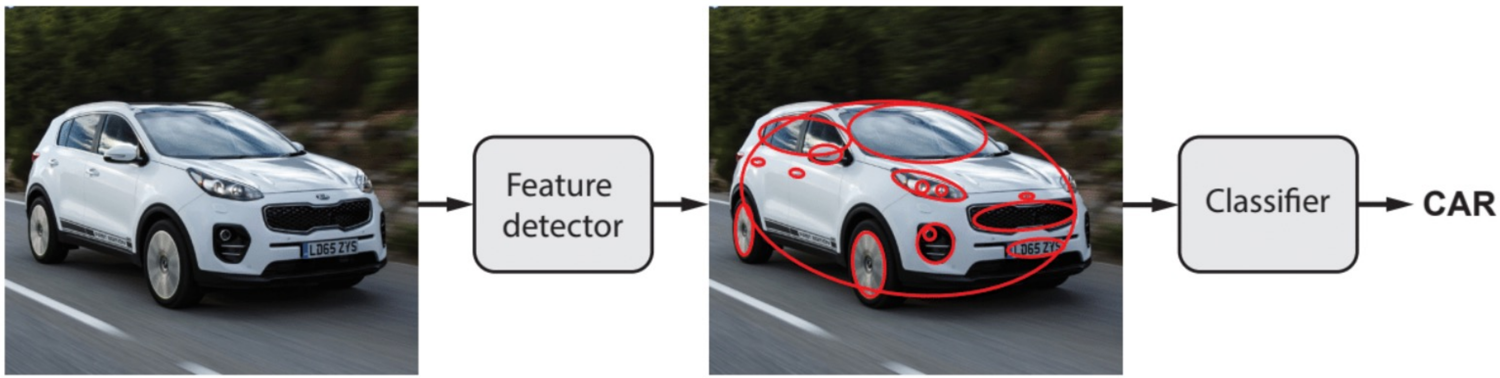
We are in the middle of a big revolution...

The Deep Learning Revolution

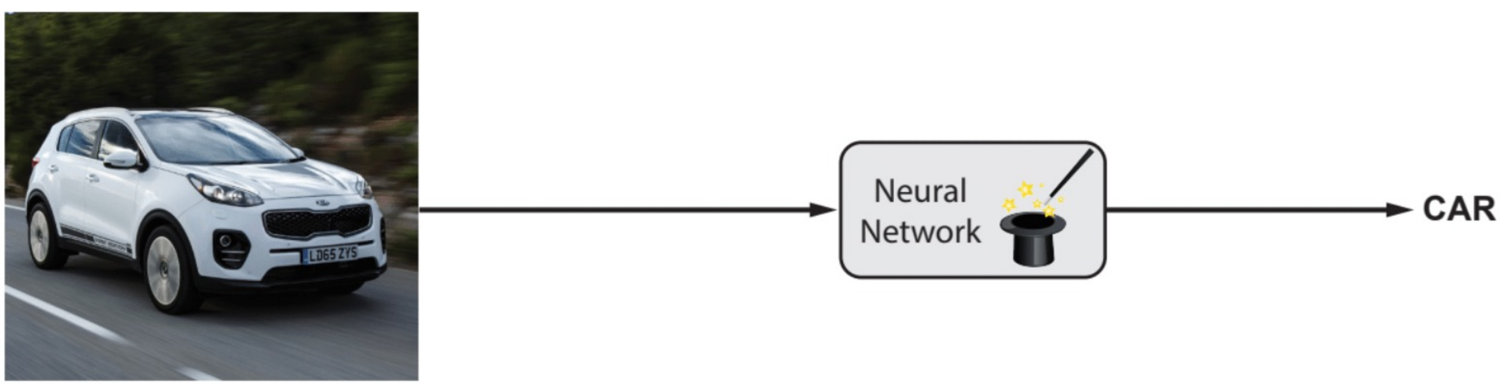
Breakthroughs in image recognition fueled by Convolutional Neural Networks.



Representation Learning



Classical computer vision: hand-crafted features (e.g. SIFT)
+ simple classifier (e.g. SVM)

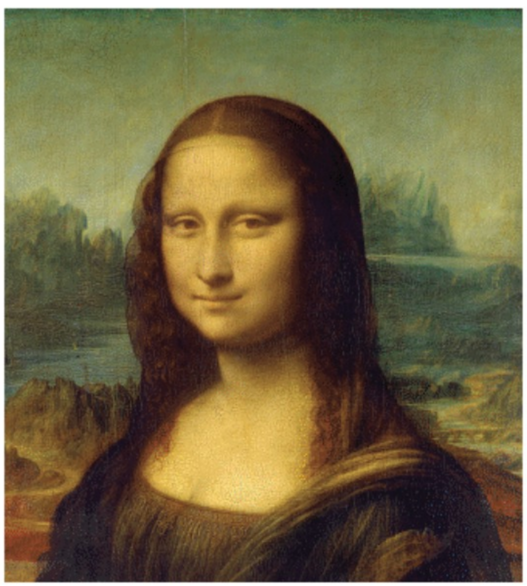


Modern computer vision: data-driven end-to-end systems

Doubt thou the stars are fire;
Doubt that the sun doth move;
Doubt truth to be a liar;
But never doubt I love...

Text

Audio signals



Images

But, modern deep learning toolbox is designed for sequences & grids

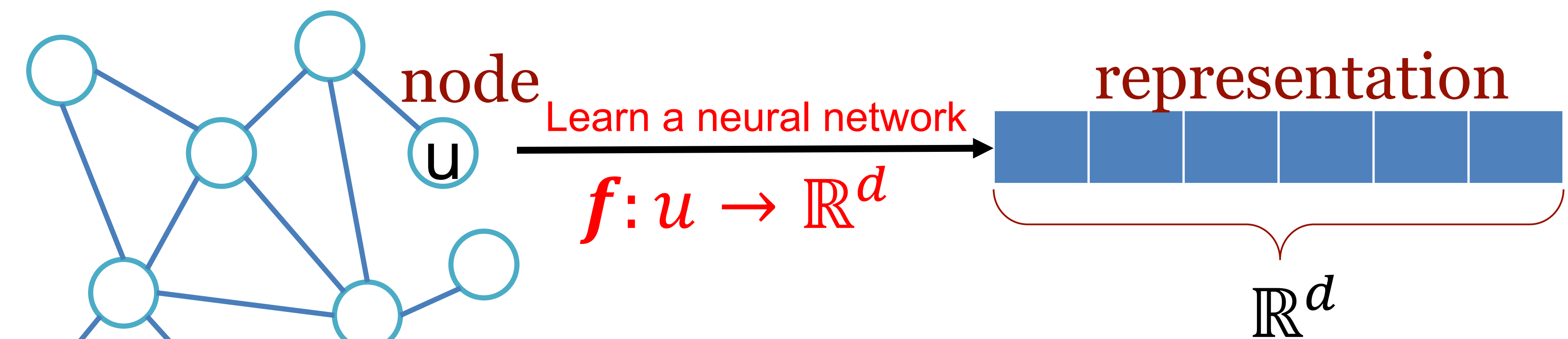
My Research

How can we develop neural networks that are much more broadly applicable?

Graphs are the new frontier of deep learning

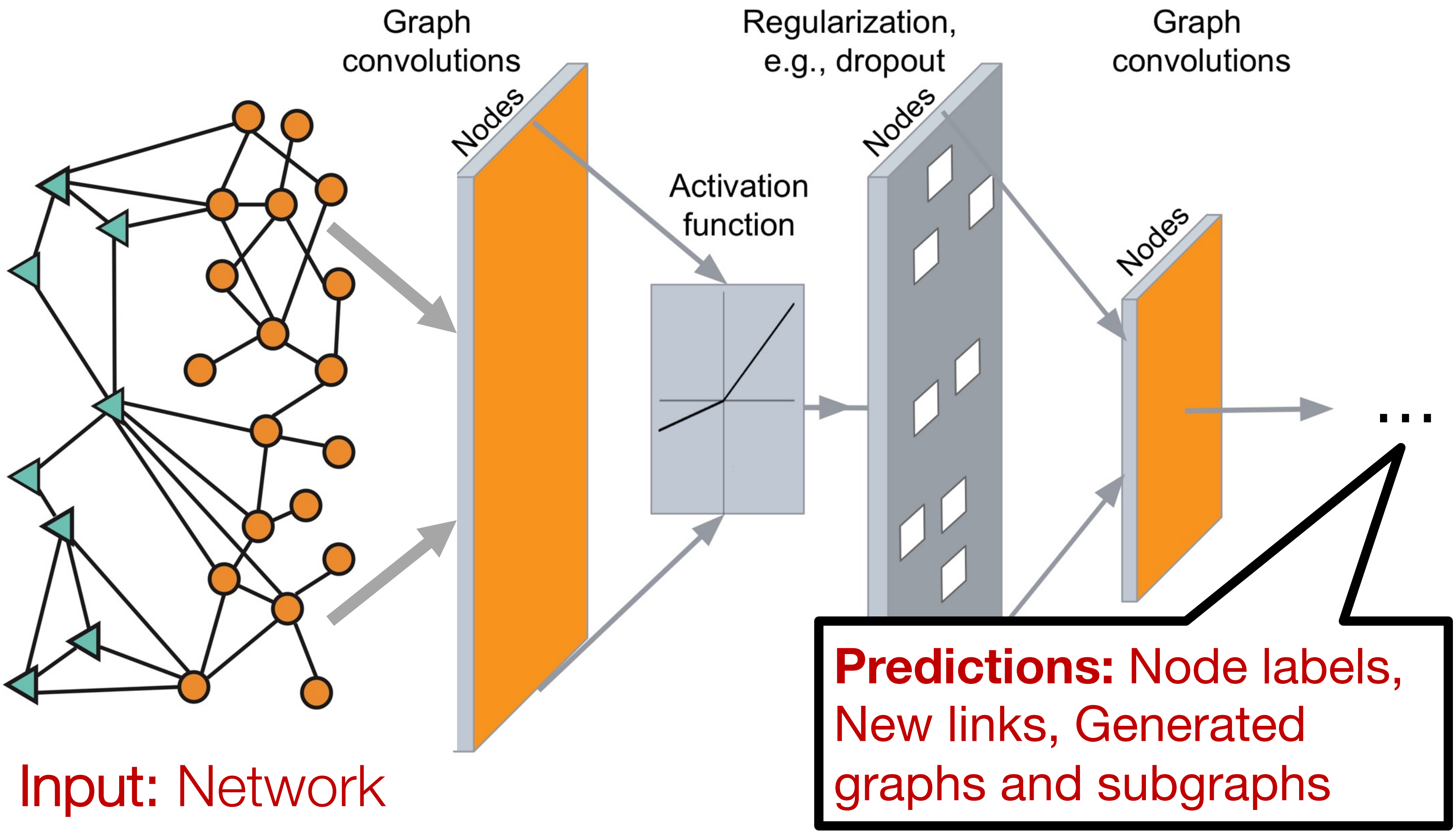
Goal: Representation Learning

Map nodes to d-dimensional **embeddings** such that similar nodes in the network are embedded close together



Feature representation, embedding

Deep Learning in Graphs

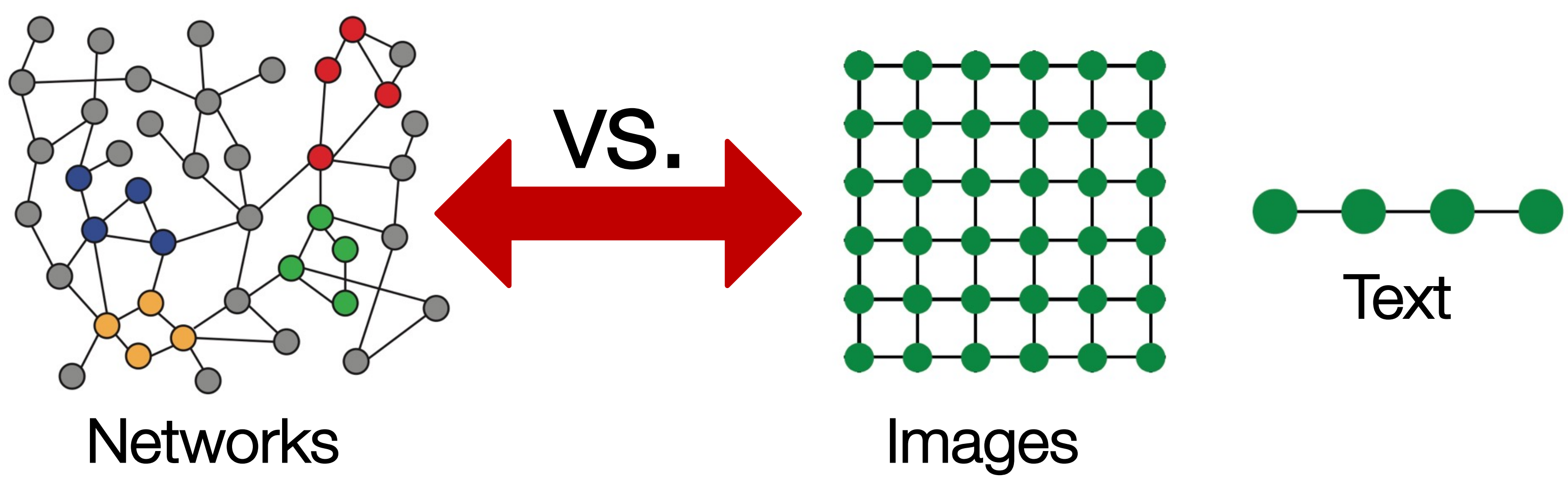


Input: Network

Why is it Hard?

Networks are complex!

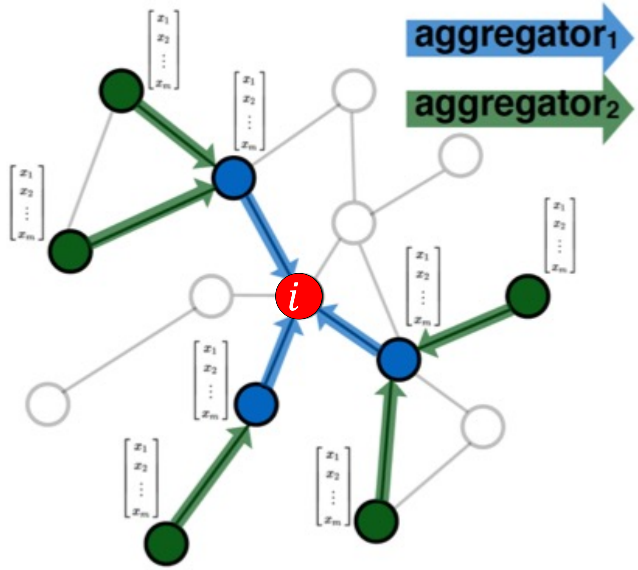
- Arbitrary size and complex topological structure (i.e., no spatial locality like grids)



- No fixed node ordering or reference point
- Often dynamic and have multimodal features

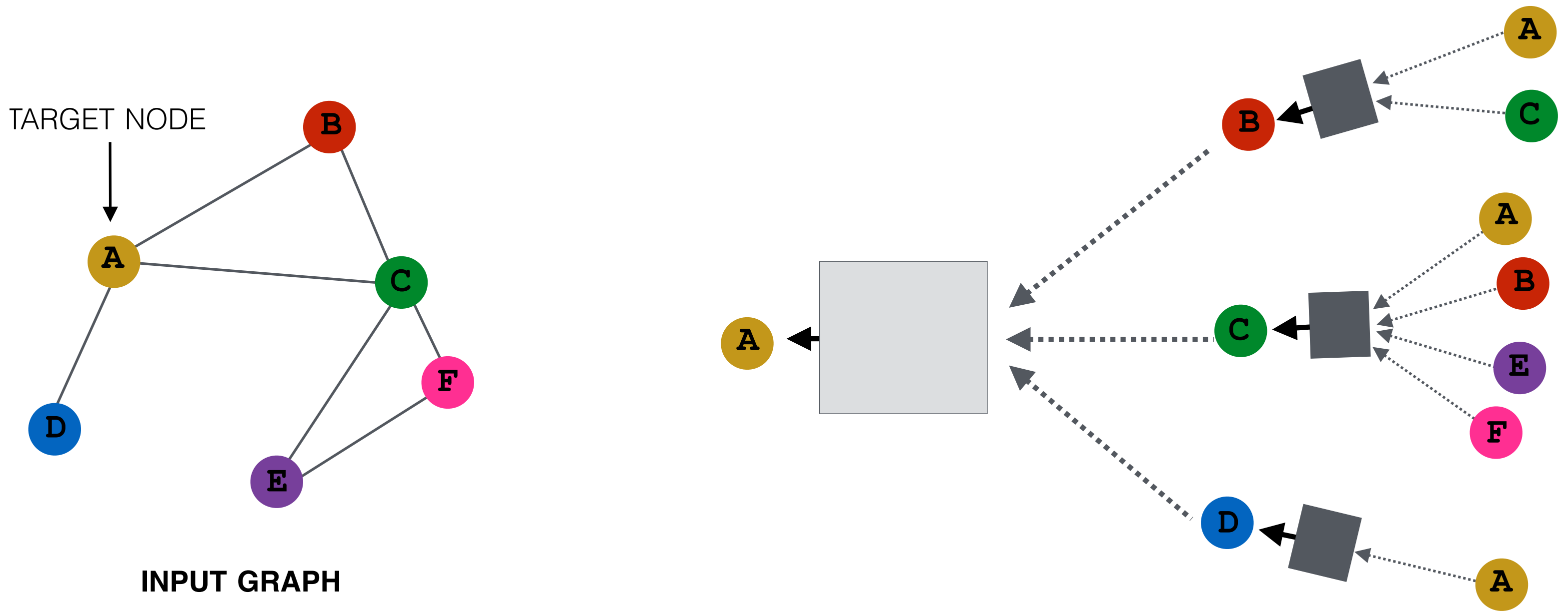
Networks as computation graphs

Key idea: Network is a computation graph



<u>Learn</u> how to propagate information across the network

GraphSAGE

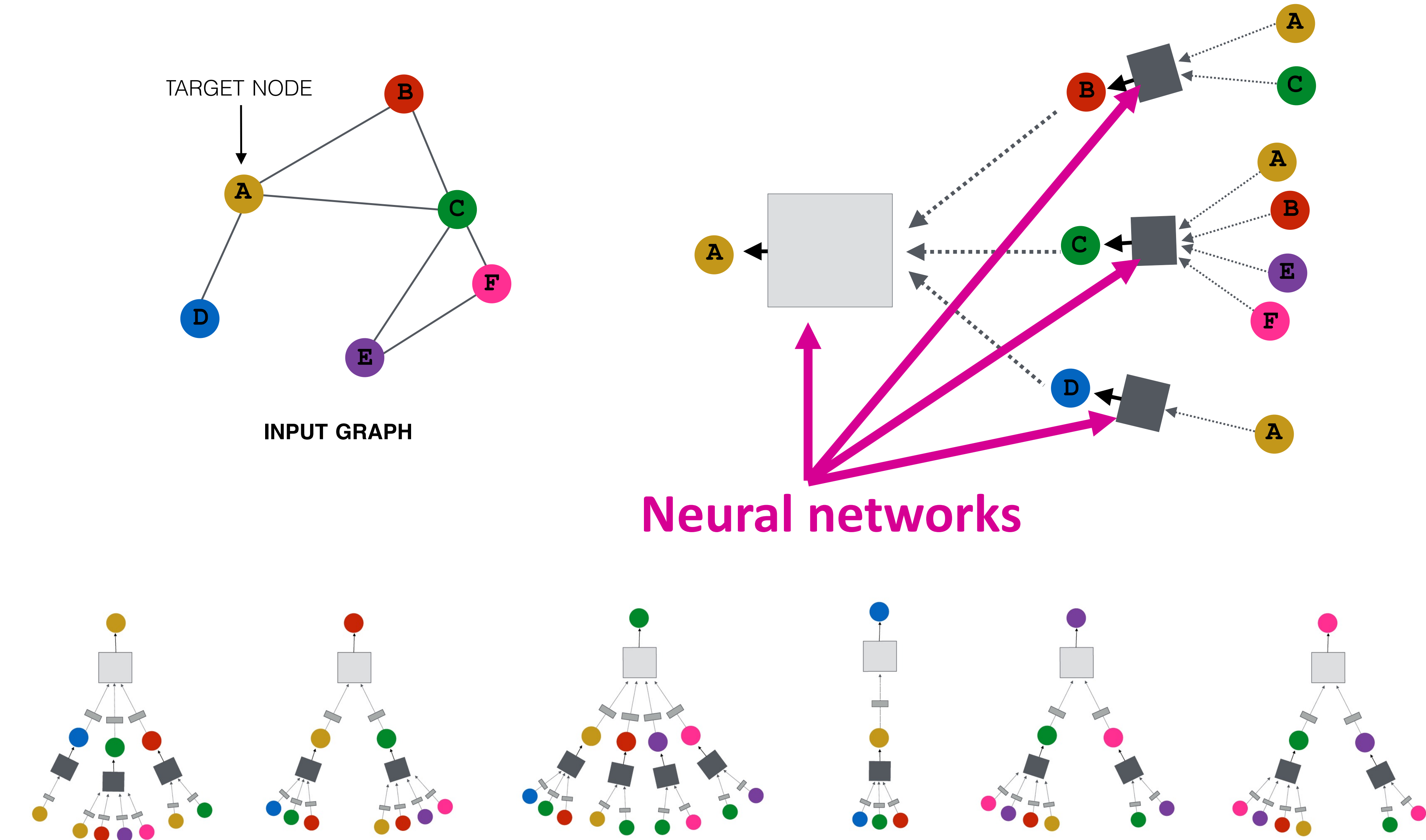


Each node defines a computation graph

- Each edge in this graph is a transformation/aggregation function

Scarselli et al. 2005. [The Graph Neural Network Model](). *IEEE Transactions on Neural Networks*.

GraphSAGE



Key Benefits of GraphSAGE

- No manual feature engineering needed
- End-to-end learning results in optimal features.
- Any graph machine learning task:
  - Node-level, link-level, entire graph-level prediction
- Scalable to billion node graphs!

What are some applications of GraphSAGE?

Computational Drug Discovery: Drug Side Effect Prediction

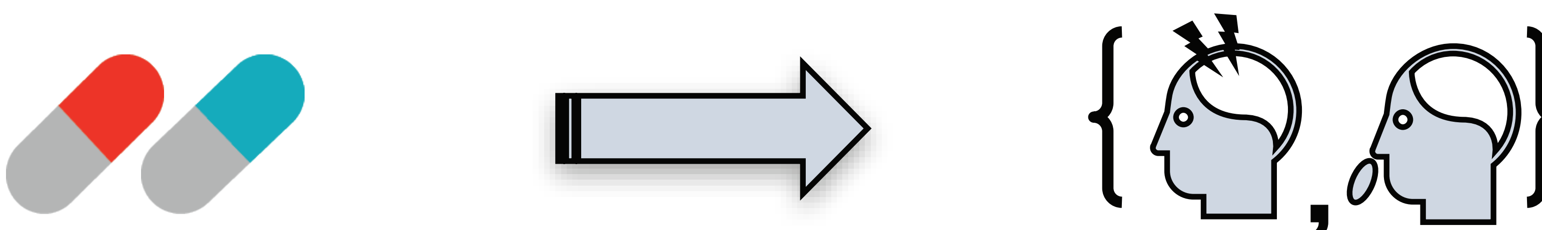
Modeling Polypharmacy Side Effects with Graph Convolutional Networks.
M. Zitnik, M. Agrawal, J. Leskovec. *Bioinformatics*, 2018.

Polypharmacy side effects

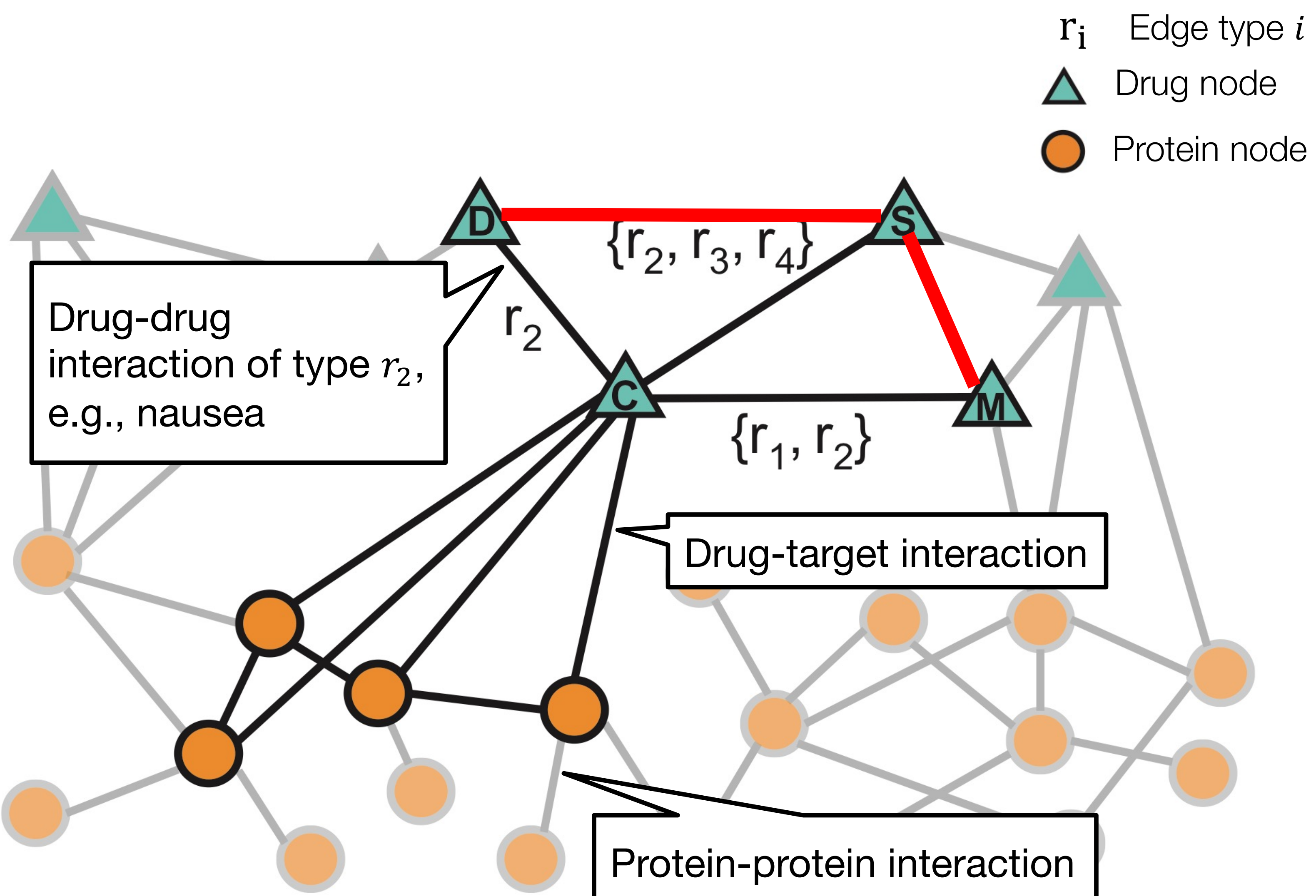
Many patients take multiple drugs to treat complex or co-existing diseases:

- 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.

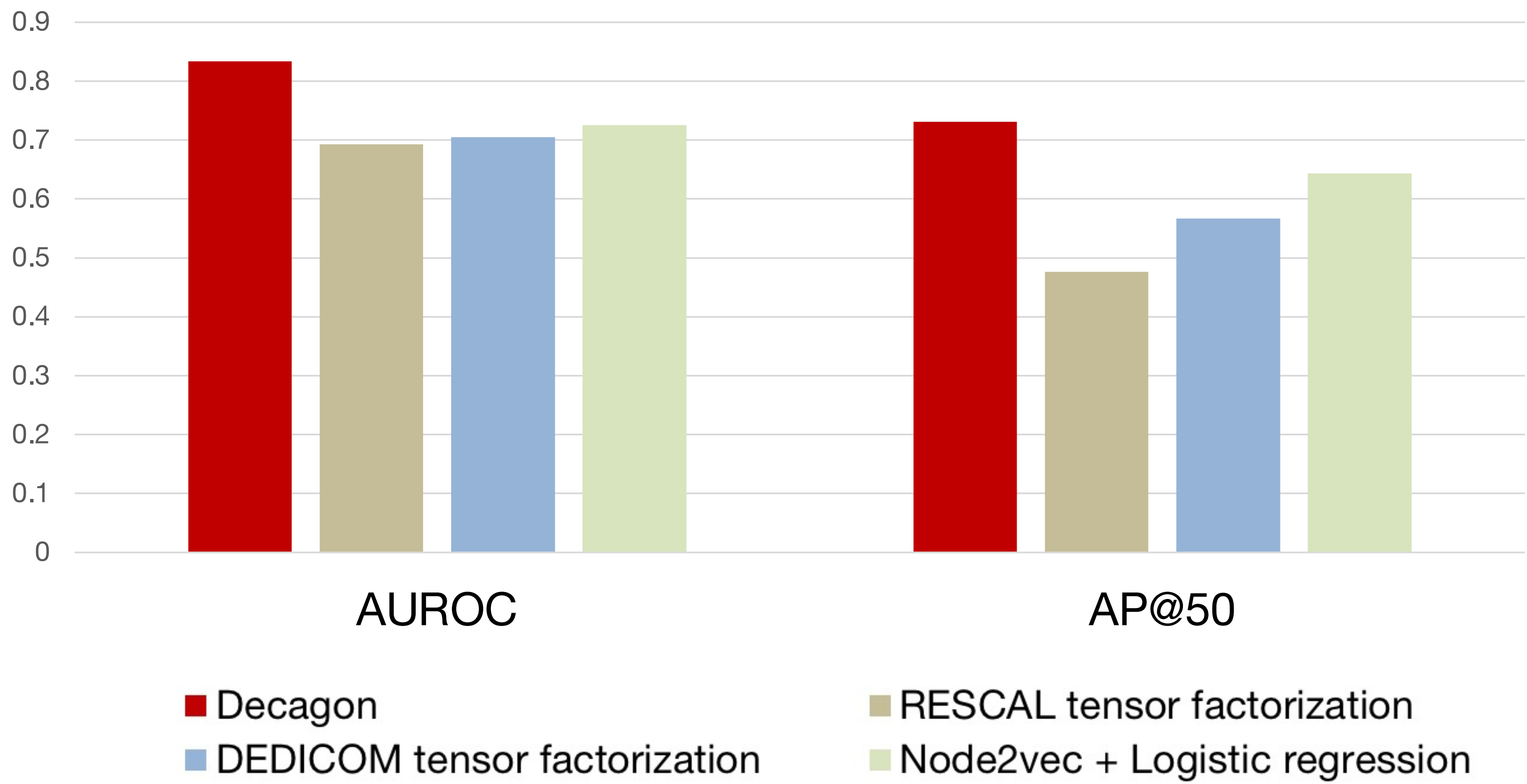
Task: Given a pair of drugs predict adverse side effects



Approach: Link Prediction



Results: Side Effect Prediction



36% average in AP@50 improvement over baselines

De novo Predictions

Rank	Drug c	Drug d	Side effect r
1	Pyrimethamine	Aliskiren	Sarcoma
2	Tigecycline	Bimatoprost	Autonomic neuropathy
3	Omeprazole	Dacarbazine	Telangiectases
4	Tolcapone	Pyrimethamine	Breast disorder
5	Minoxidil	Paricalcitol	Cluster headache
6	Omeprazole	Amoxicillin	Renal tubular acidosis
7	Anagrelide	Azelaic acid	Cerebral thrombosis
8	Atorvastatin	Amlodipine	Muscle inflammation
9	Aliskiren	Tioconazole	Breast inflammation
10	Estradiol	Nadolol	Endometriosis

De novo Predictions

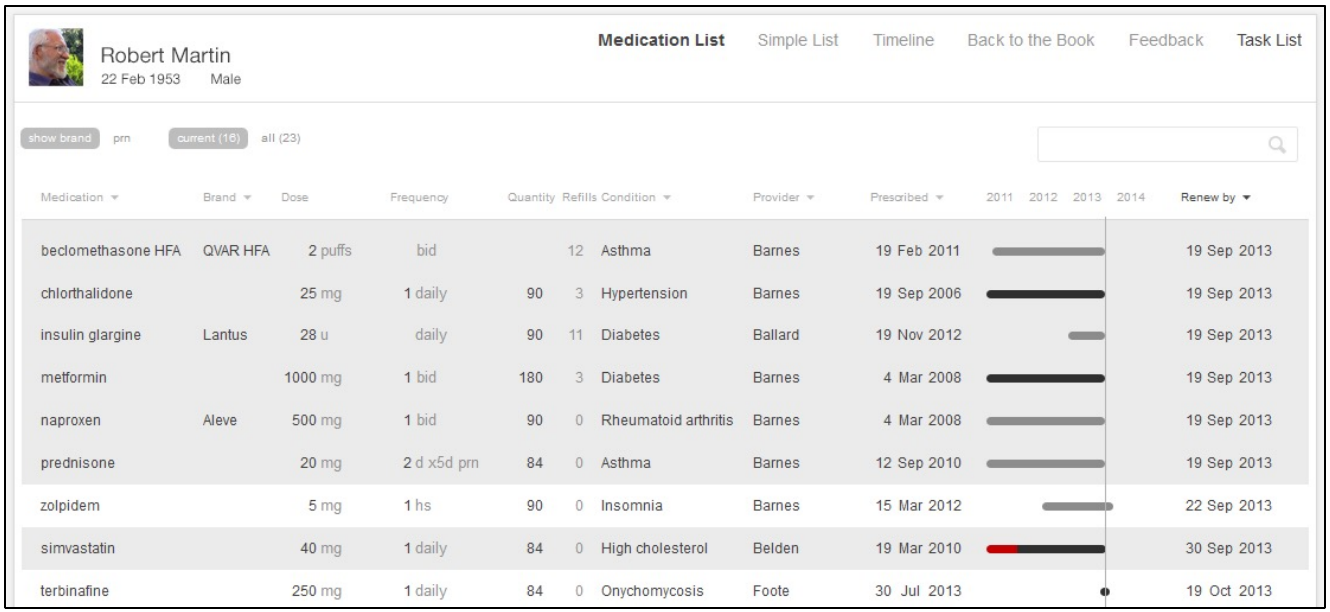
Rank	Drug c	Drug d	Side effect r	Evidence found
1	Pyrimethamine	Aliskiren	Sarcoma	Stage *et al.* 2015
2	Tigecycline	Bimatoprost	Autonomic neuropathy	
3	Omeprazole	Dacarbazine	Telangiectases	
4	Tolcapone	Pyrimethamine	Breast disorder	Bicker *et al.* 2017
5	Minoxidil	Paricalcitol	Cluster headache	
6	Omeprazole	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

Predictions in the Clinic

Clinical validation via drug-drug interaction markers, lab values, and


Robert Martin
22 Feb 1953 Male

Medication List | Simple List | Timeline | Back to the Book | Feedback | Task List

show brand | prn | current (16) | all (23)

Medication	Brand	Dose	Frequency	Quantity	Refills	Condition	Provider	Prescribed	2011 2012 2013 2014	Renew by
beclomethasone HFA	QVAR HFA	2 puffs	bid		12	Asthma	Barnes	19 Feb 2011		19 Sep 2013
chlorthalidone		25 mg	1 daily	90	3	Hypertension	Barnes	19 Sep 2006		19 Sep 2013
insulin glargine	Lantus	28 u	daily	90	11	Diabetes	Ballard	19 Nov 2012		19 Sep 2013
metformin		1000 mg	1 bid	180	3	Diabetes	Barnes	4 Mar 2008		19 Sep 2013
naproxen	Aleve	500 mg	1 bid	90	0	Rheumatoid arthritis	Barnes	4 Mar 2008		19 Sep 2013
prednisone		20 mg	2 d x5d prn	84	0	Asthma	Barnes	12 Sep 2010		19 Sep 2013
zolpidem		5 mg	1 hs	90	0	Insomnia	Barnes	15 Mar 2012		22 Sep 2013
simvastatin		40 mg	1 daily	84	0	High cholesterol	Belden	19 Mar 2010		30 Sep 2013
terbinafine		250 mg	1 daily	84	0	Onychomycosis	Foote	30 Jul 2013		19 Oct 2013

NEWTON-WELLESLEY HOSPITAL

MASSACHUSETTS GENERAL HOSPITAL

Stanford MEDICINE

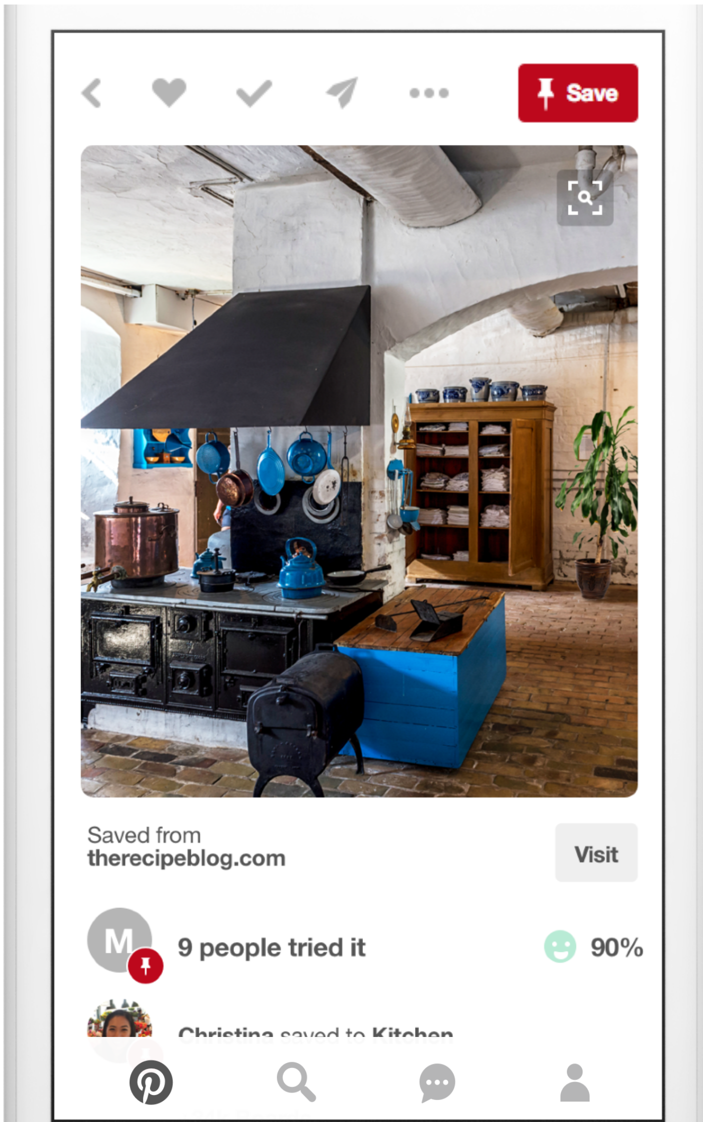
HARVARD MEDICAL SCHOOL

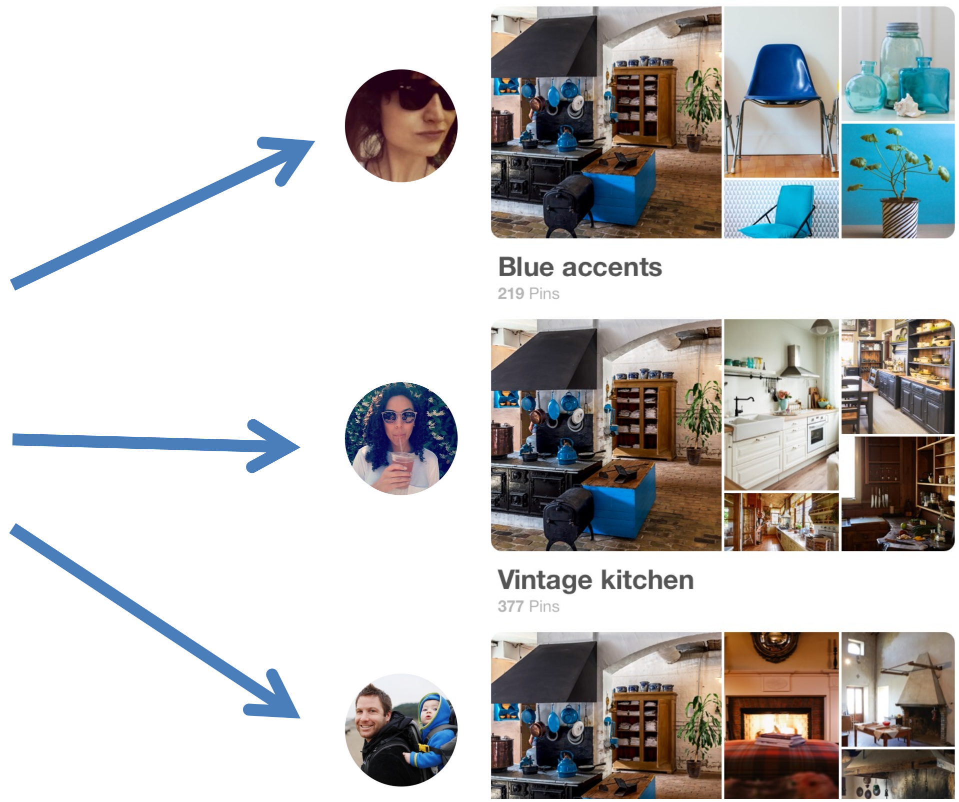
First method to **predict side effects of drug pairs,** even for drug combinations not yet used in patients

Massive Social Networks: Example of Pinterest

Graph Convolutional Neural Networks for Web-Scale Recommender Systems. R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec. *KDD*, 2018.

Pinterest



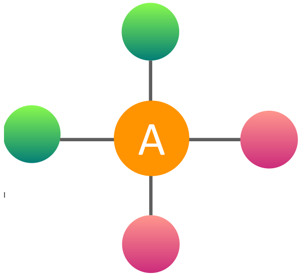


- 300M users
- 4+B pins, 2+B boards

Application: Pinterest

PinSage graph convolutional network:

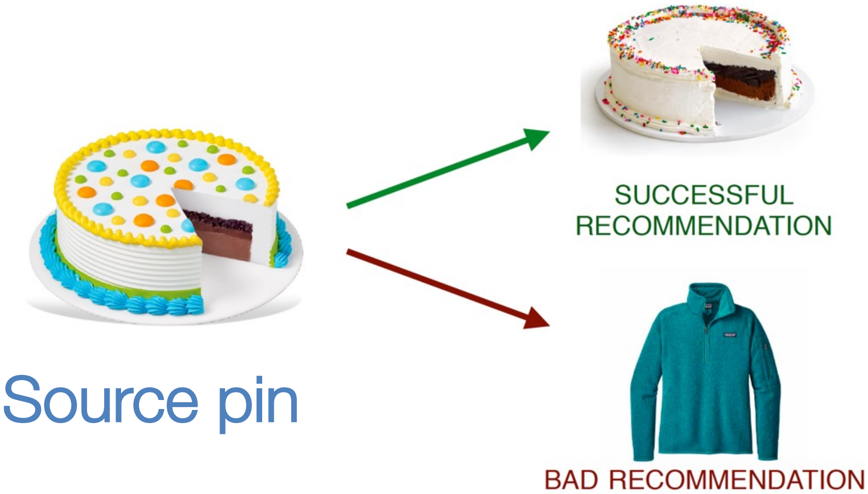
- Goal: Generate embeddings for nodes in a large-scale Pinterest graph containing billions of objects
- Key Idea: Borrow information from nearby nodes
  - E.g., bed rail Pin might look like a garden fence, but gates and beds are rarely adjacent in the graph



  - Pin embeddings are essential to various tasks like recommendation of Pins, classification, ranking
    - Services like “Related Pins”, “Search”, “Shopping”, “Ads”

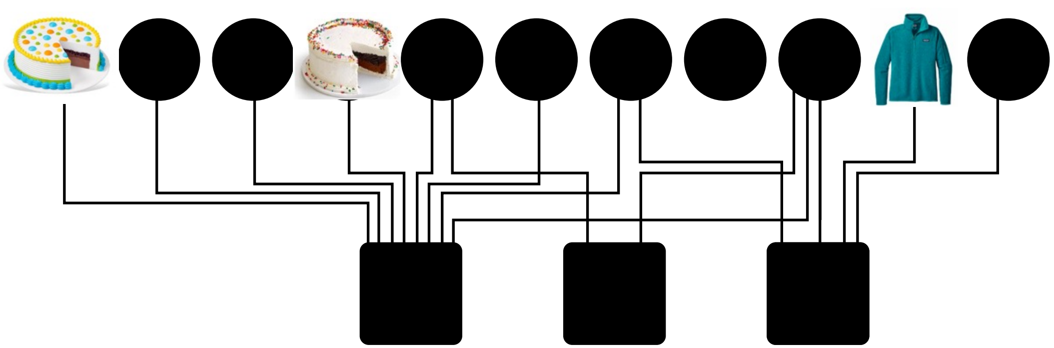
Pin Recommendation

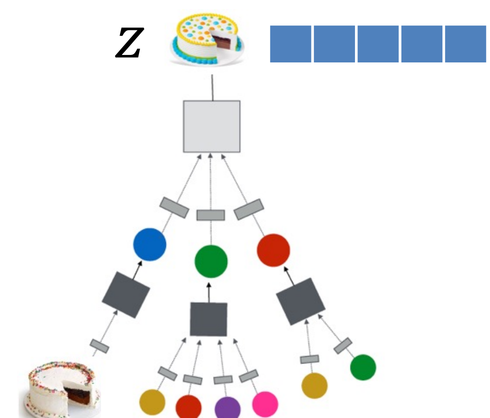
Task: Recommend related pins to users



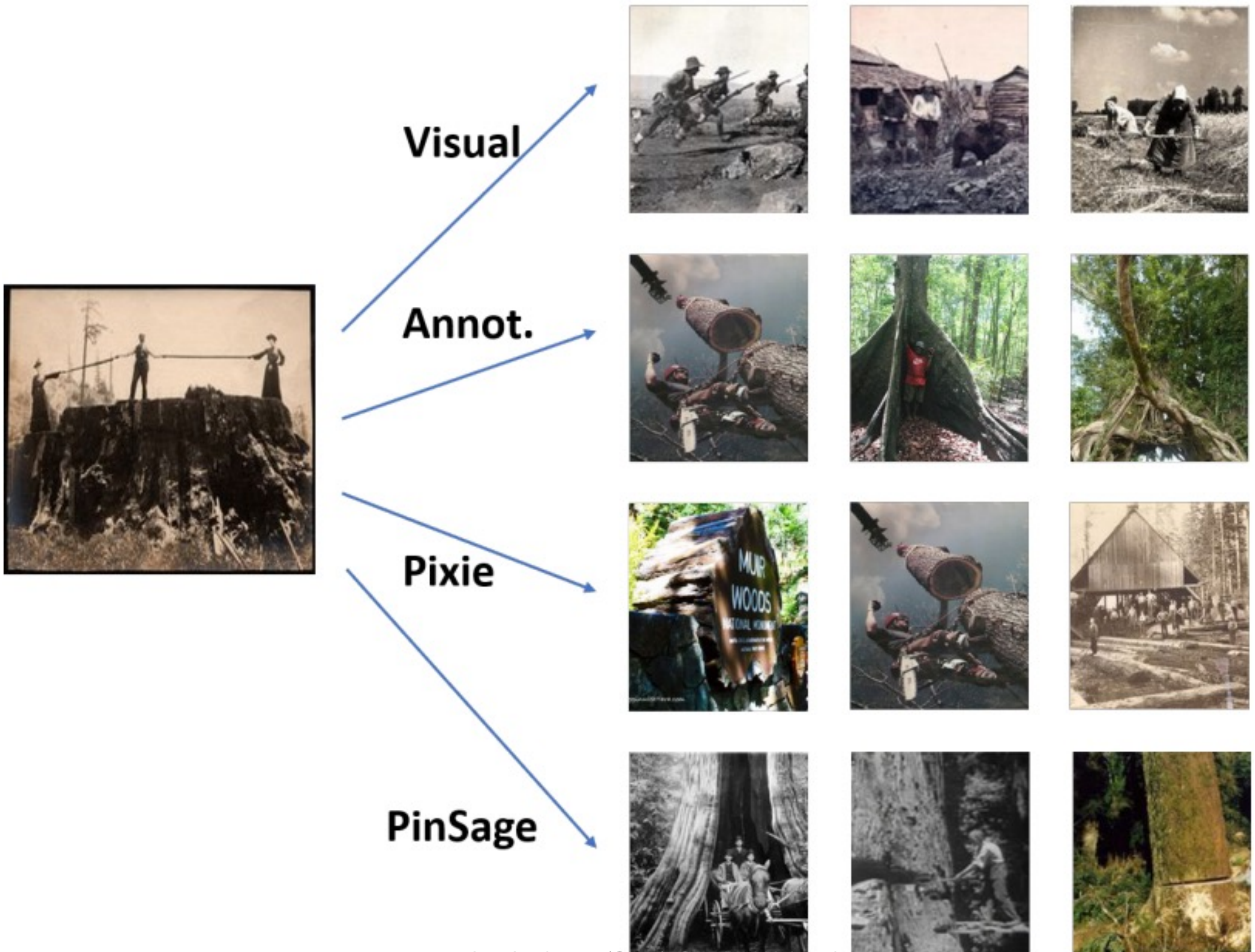
Task: Learn node embeddings z_i such that
$d(z_{cake1}, z_{cake2})$
$< d(z_{cake1}, z_{sweater})$

Predict whether two nodes in a graph are related



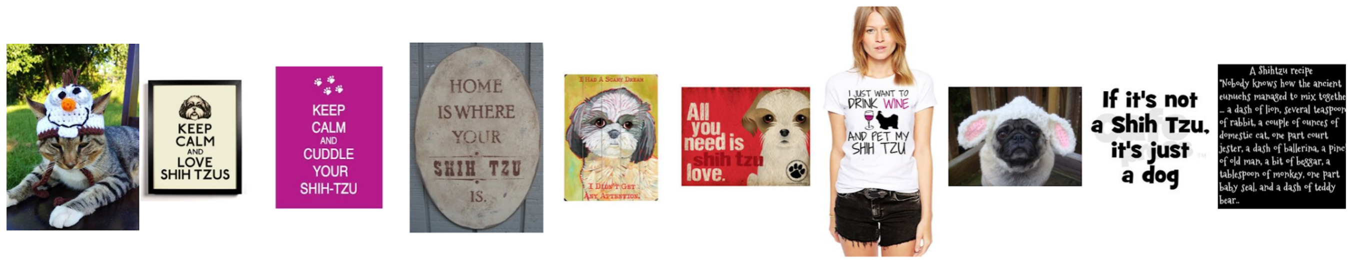


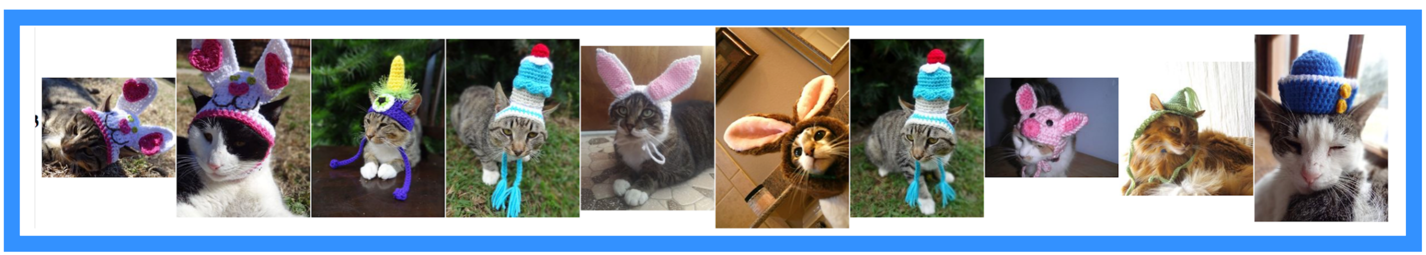
PinSAGE Example



Results

Query





PinSAGE

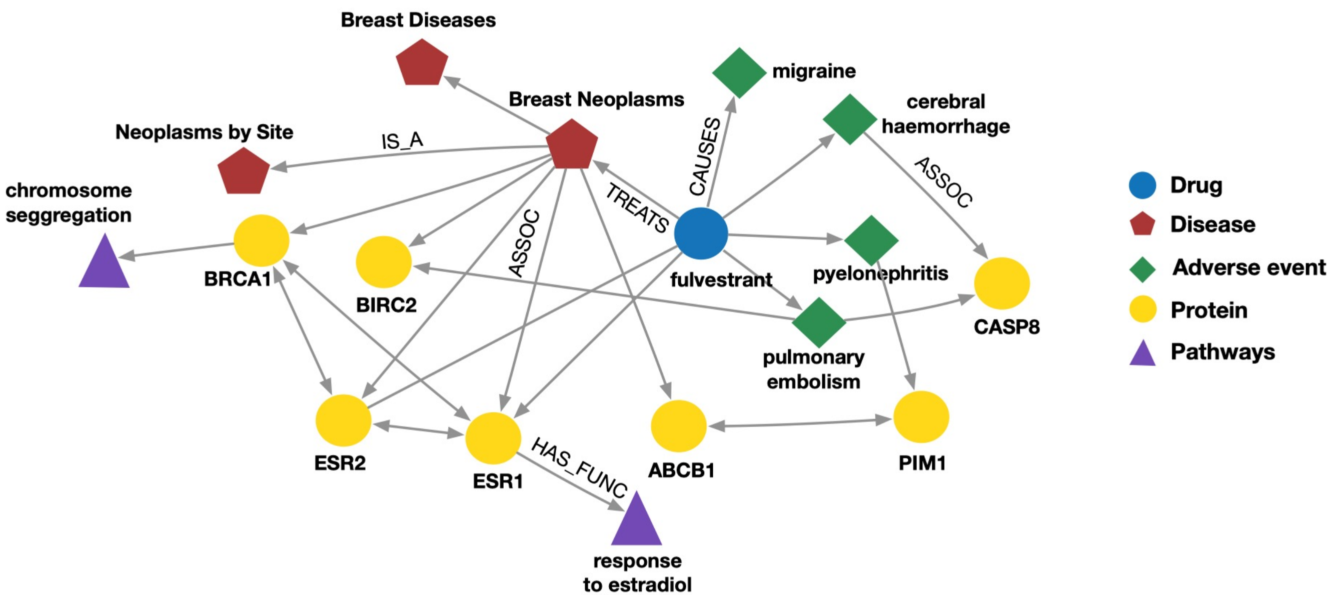
Reasoning in Incomplete Knowledge Graphs

Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. H. Ren, J. Leskovec. *Neural Information Processing Systems (NeurIPS)*, 2020.
Identification Of Disease Treatment Mechanisms Through The Multiscale Interactome. C. Ruiz, M. Zitnik, J Leskovec. *Nature Communications*, 2021.

Knowledge Graphs

Knowledge in a graph form:

- Captures entities, types, and relationships



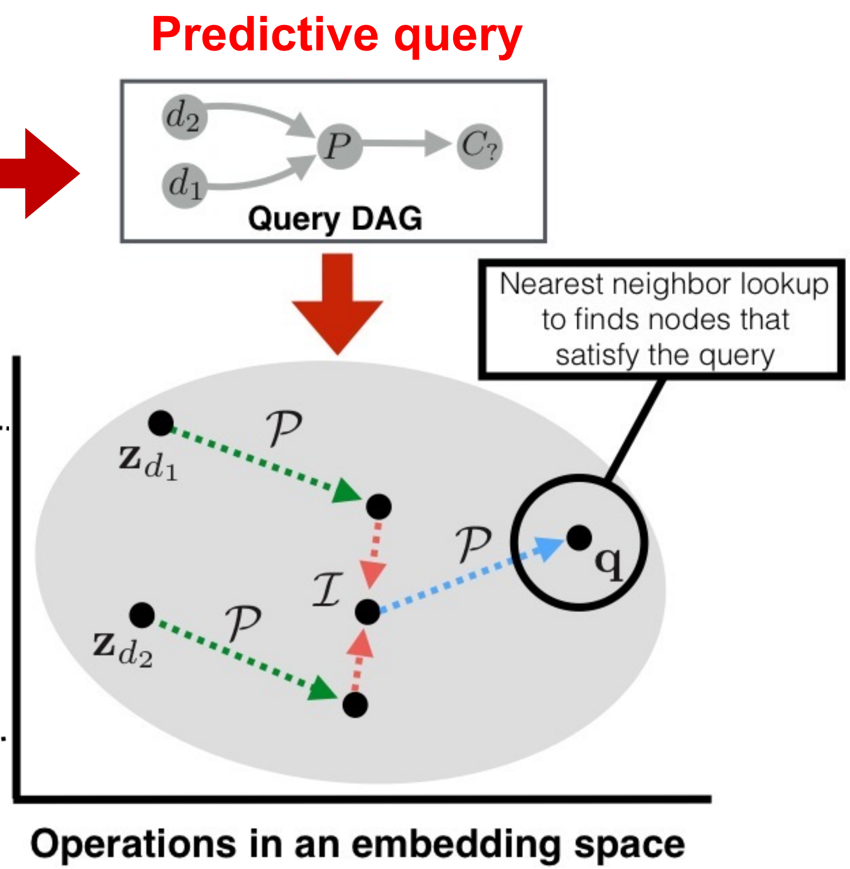
Node types: drug, disease, adverse event, protein, pathway

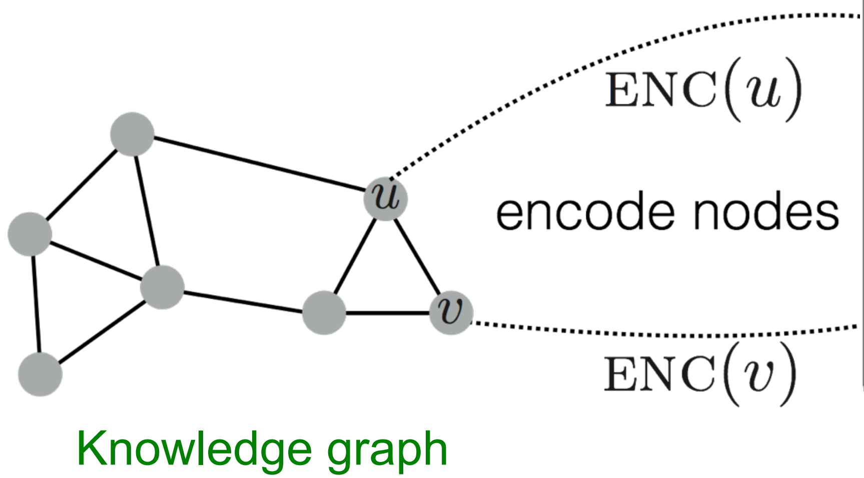
Relation types: has_func, causes, assoc, treats, is_a

Overview of Our Framework

Goal: Complex predictions in KGs

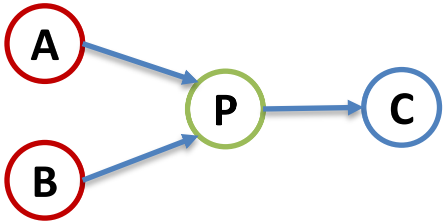
E.g.: “**Predict drugs C** likely *target* proteins P *associated* with diseases d_1 and d_2”.



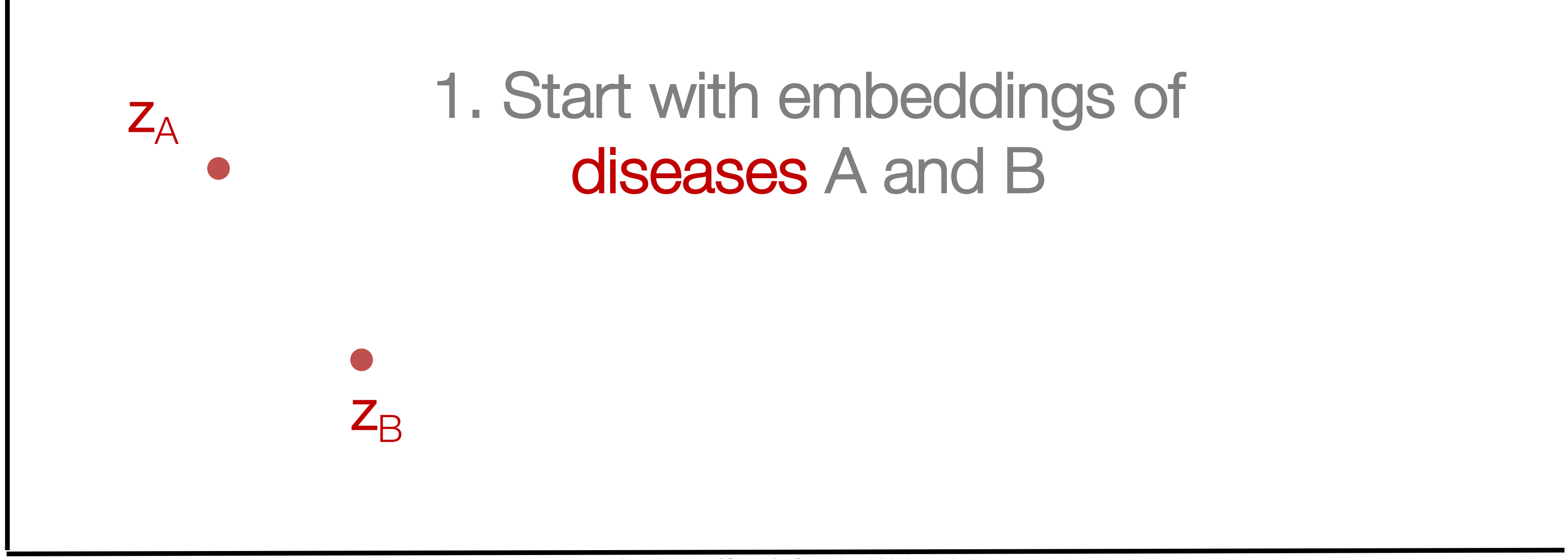


Example: Drug Discovery

Query: "**Predict drugs C** likely *target* proteins P *associated* with diseases A and B"

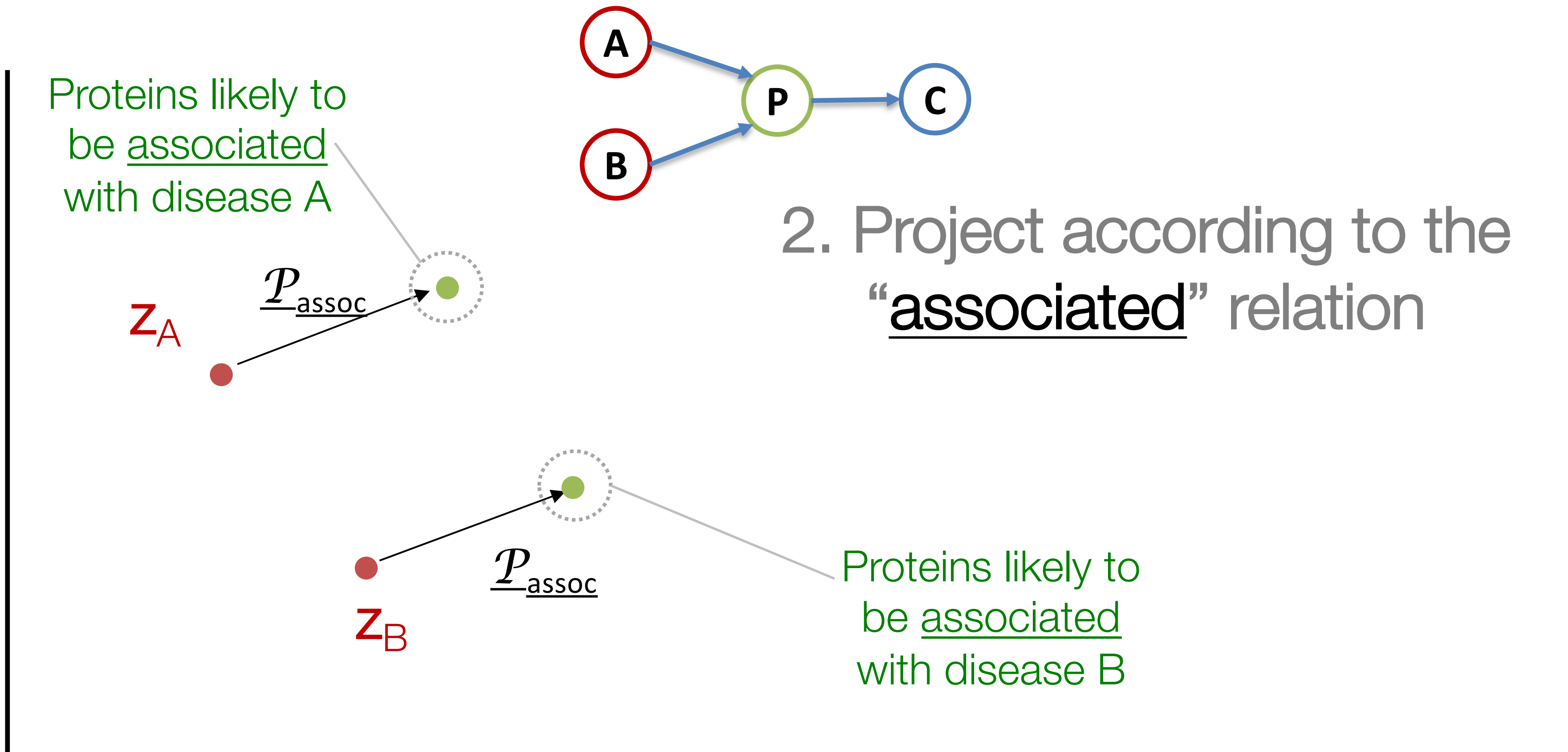


1. Start with embeddings of diseases A and B



Example: Drug Discovery

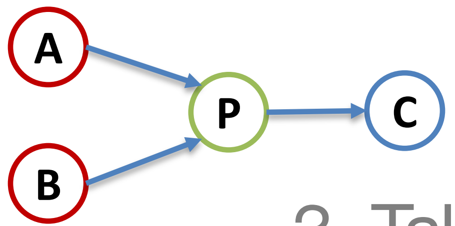
Query: "**Predict drugs C** likely *target* proteins P *associated* with diseases A and B"



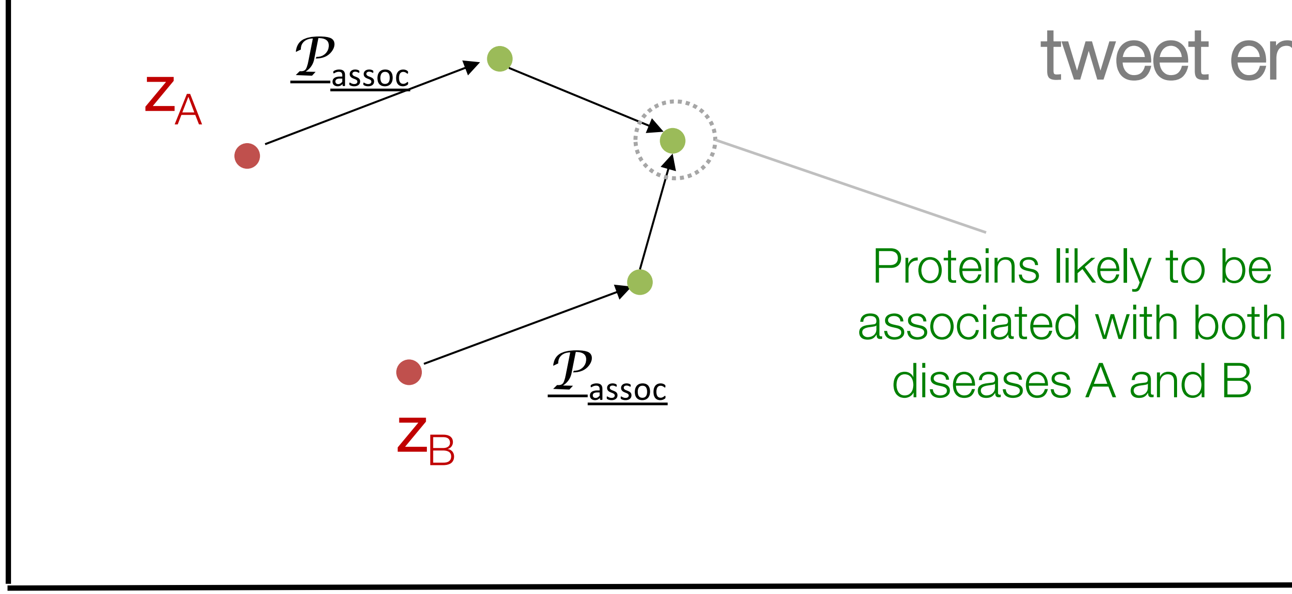
2. Project according to the "**associated**" relation

Example: Drug Discovery

Query: "**Predict drugs C** likely *target* proteins P *associated* with diseases A and B"

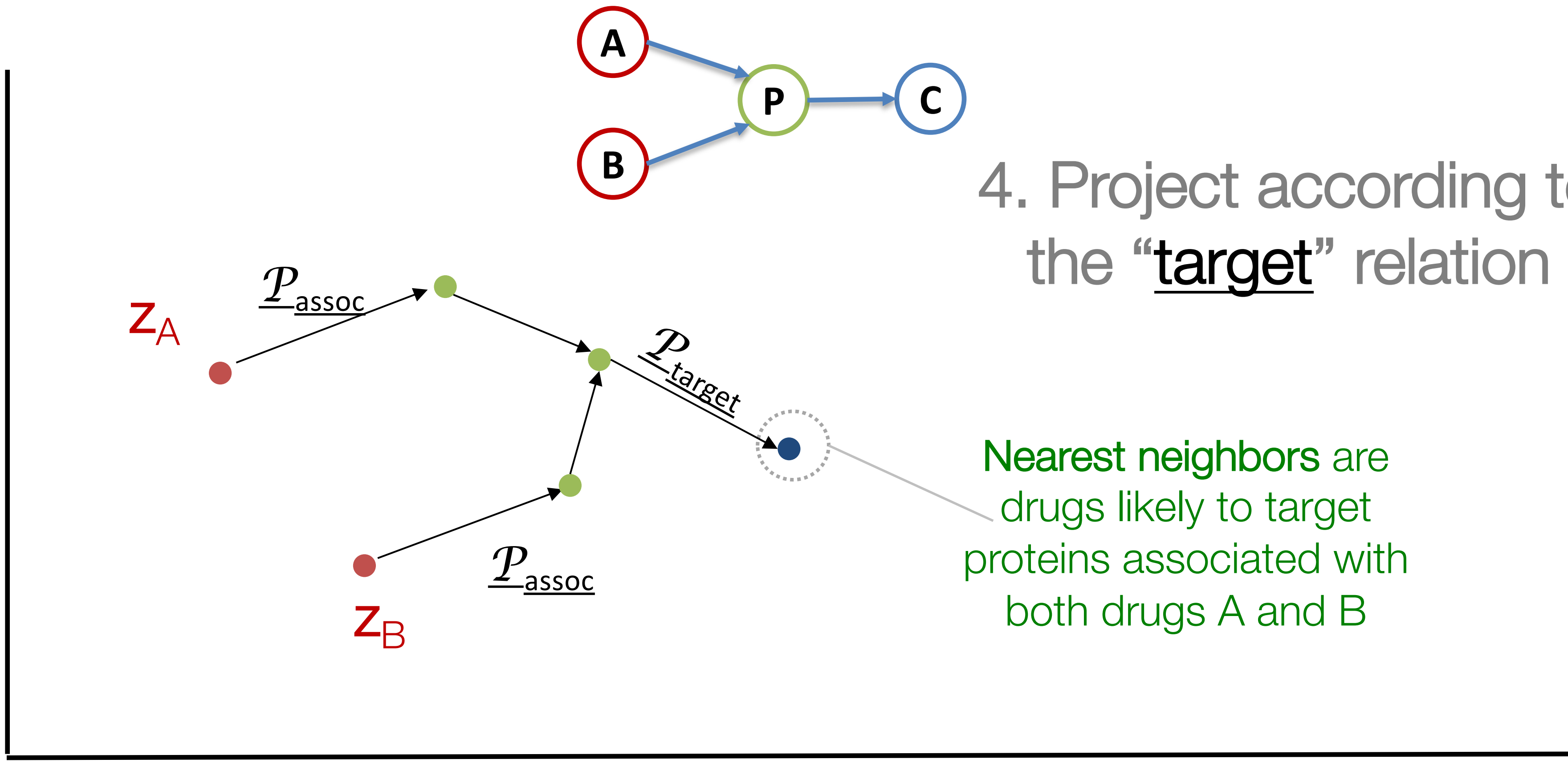


3. Take **intersection** of the tweet embeddings



Example: Drug Discovery

Query: "**Predict drugs C** likely *target* proteins P *associated* with diseases A and B"



4. Project according to the "**target**" relation

Nearest neighbors are drugs likely to target proteins associated with both drugs A and B

Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. H. Ren, J. Leskovec. *Neural Information Processing Systems (NeurIPS)*, 2020.
Identification Of Disease Treatment Mechanisms Through The Multiscale Interactome. C. Ruiz, M. Zitnik, J Leskovec. *Nature Communications*, 2021.

How can this technology be used for other problems?

We can now apply neural networks much more broadly

New frontiers beyond classic neural networks that learn on images and sequences

Many other applications:

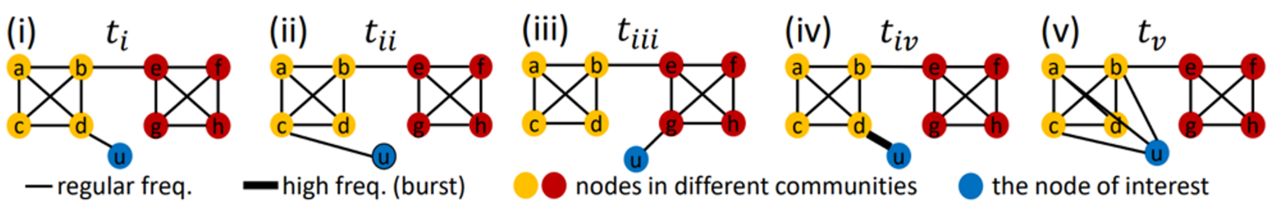
- Fraud and Anomaly Detection
- Graph generation
- Common sense reasoning

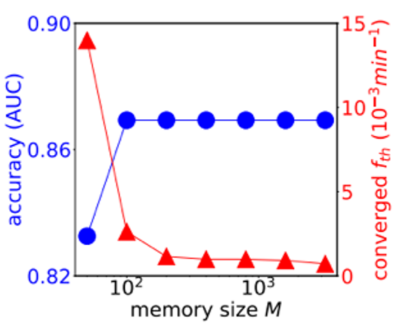
(1) Fraud & Intrusion Detection

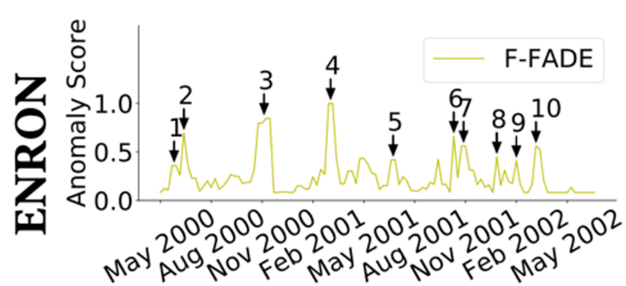
Fraud and intrusion detection in dynamic transaction graphs

Financial networks

Communication networks







Barracuda

CHASE

amazon

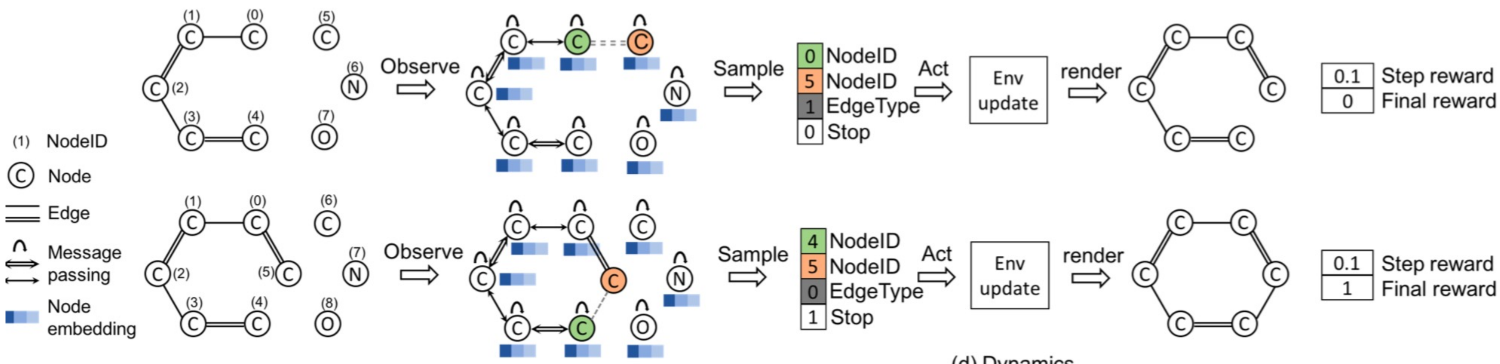
VISA

(2) Targeted Molecule Generation

Goal: Generate molecules that optimize a given property (Quant. energy, solubility)

Solution: Combination of

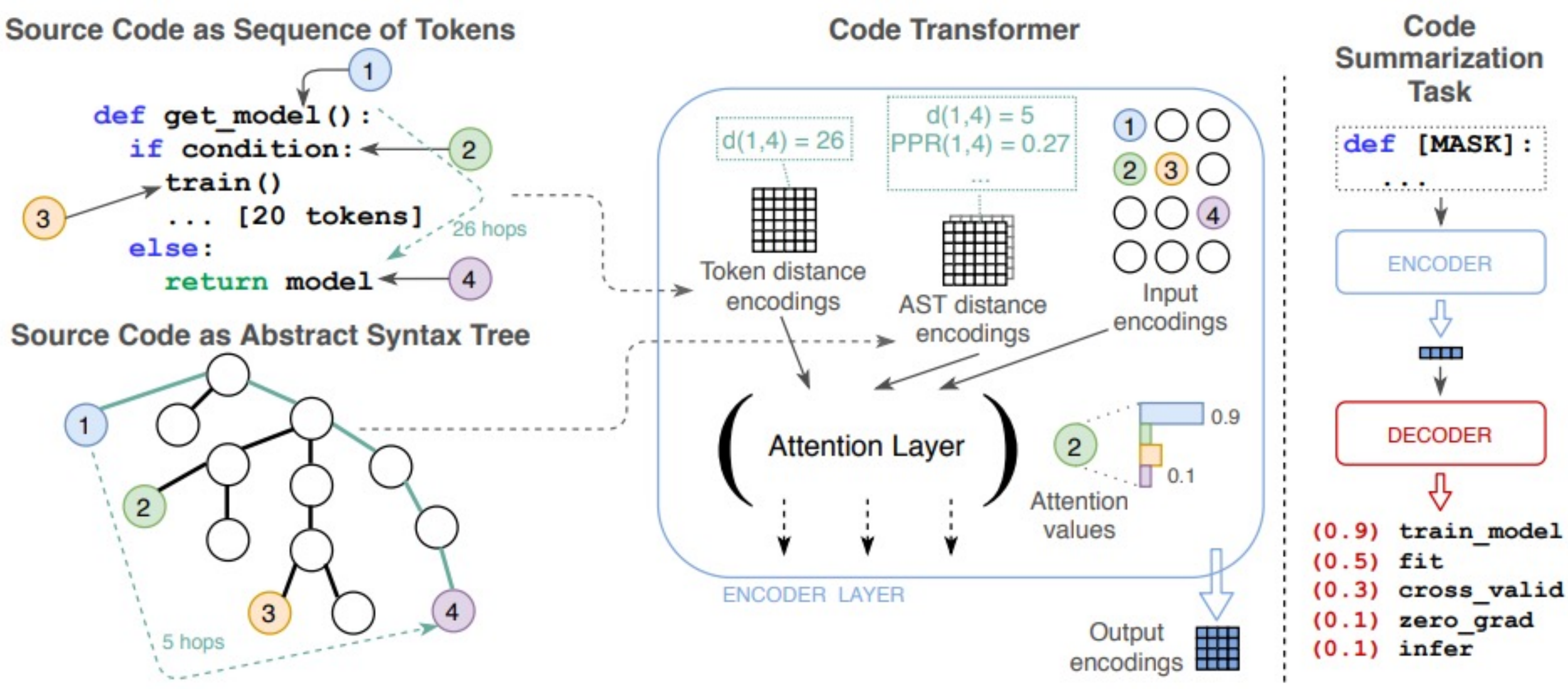
- Graph representation learning
- Adversarial training
- Reinforcement learning



(a) State — G_t Scaffold — C (b) GCPN — $\pi_\theta(a_t|G_t \cup C)$ (c) Action — $a_t \sim \pi_\theta$ (d) Dynamics $p(G_{t+1}|G_t, a_t)$ (e) State — G_{t+1} (f) Reward — r_t

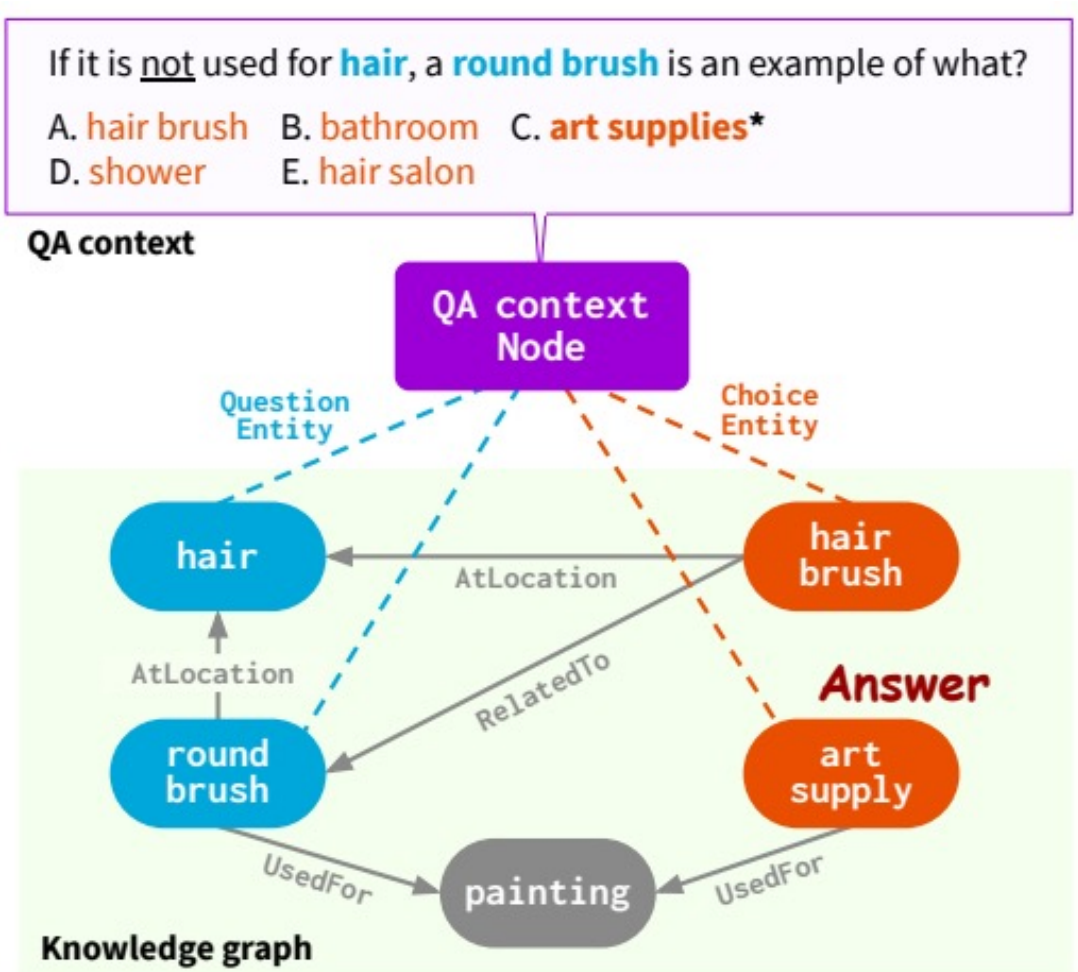
Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. J. You, B. Liu, R. Ying, V. Pande, J. Leskovec, NeurIPS 2018.

(3) Resoning with Programs



[Language-agnostic Representation Learning Of Source Code From Structure And Context](). D. Zugner, T. Kirschstein, M. Catasta, J. Leskovec, S. Gunnemann. *International Conference on Learning Representations (ICLR)*, 2021.

(4) Common Sense Reasoning



[QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering](). M. Yasunaga, H. Ren, A. Bosselut, P. Liang, J. Leskovec. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

Summary

GraphSAGE brings the power of deep learning to graphs!

- Fuses node features & graph info
  - State-of-the-art accuracy graph machine learning tasks
- Model size independent of graph size; can scale to billions of nodes
  - Largest embedding to date (3B nodes, 20B edges)
- Leads to significant performance gains

Conclusion

Results from the past 2-3 years have shown:

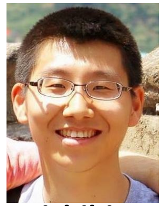
- Representation learning paradigm can be extended to graphs
- No feature engineering necessary
- Can effectively combine node attribute data with the network information
- State-of-the-art results in a number of domains/tasks
- Use end-to-end training instead of multi-stage approaches for better performance

PhD Students

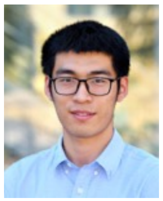

Alexandra Porter

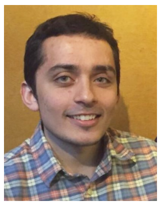

Camilo Ruiz

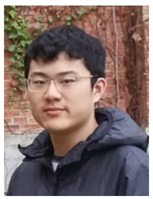

Serina Chang


Michihiro Yasunaga


Weihua Hu


Jiaxuan You


Yusuf Roohani


Hongyu Ren


Rex Ying

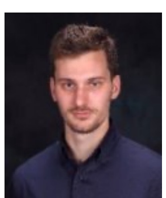
Post-Doctoral Fellows


Maria Brbic


Tailin Wu


Antoine Bosselut

Research Staff


Adrijan Bradaschia


Rok Sosic

Industry Partnerships

أرامكو السعودية saudi aramco

amazon, twitter, Pinterest, WIKIPEDIA The Free Encyclopedia, DELL, facebook, TOSHIBA, BOEING, CHASE, intel, NTT docomo, NVIDIA, JD.COM, HITACHI, Viaduct

Funding

NIH, IARPA, DARPA, NSF, Office of Naval Research ONR Science & Technology, MURI ARO, CHAN ZUCKERBERG INITIATIVE

Collaborators

Learn more at:

http://snap.stanford.edu

Contact us at:

jure@cs.stanford.edu

References

- Tutorial on Representation Learning on Networks at WWW 2018 http://snap.stanford.edu/proj/embeddings-www/
- Inductive Representation Learning on Large Graphs.
  W. Hamilton, R. Ying, J. Leskovec. NIPS 2017.
- Representation Learning on Graphs: Methods and Applications. W. Hamilton, R. Ying, J. Leskovec.
  IEEE Data Engineering Bulletin, 2017.
- Graph Convolutional Neural Networks for Web-Scale Recommender Systems. R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec. KDD, 2018.
- Modeling Polypharmacy Side Effects with Graph Convolutional Networks. M. Zitnik, M. Agrawal, J. Leskovec. Bioinformatics, 2018.
- Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. J. You, B. Liu, R. Ying, V. Pande, J. Leskovec, NeurIPS 2018.
- Embedding Logical Queries on Knowledge Graphs. W. Hamilton, P. Bajaj, M. Zitnik, D. Jurafsky, J. Leskovec. NeuIPS, 2018.
- How Powerful are Graph Neural Networks? K. Xu, W. Hu, J. Leskovec, S. Jegelka. ICLR 2019.
- Position-aware Graph Neural Networks. J. You, R. Ying, J. Leskovec. *ICML*, 2019.
- To Embed or Not: Network Embedding as a Paradigm in Computational Biology. W. Nelson, M. Zitnik, B. Wang, J. Leskovec, A. Goldenberg, R. Sharan. *Frontiers in Genetics, 10:381*, 2019.
- G2SAT: Learning to Generate SAT Formulas. J. You, H. Wu, C. Barrett, R. Ramanujan, J. Leskovec. *Neural Information Processing Systems (NeurIPS)*, 2019.
- Hyperbolic Graph Convolutional Neural Networks. I. Chami, R. Ying, C. Re, J. Leskovec. *Neural Information Processing Systems (NeurIPS)*, 2019.
- GNNExplainer: Generating Explanations for Graph Neural Networks. R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec. *Neural Information Processing Systems (NeurIPS)*, 2019.
- Strategies For Pre-training Graph Neural Networks. W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, J. Leskovec. *International Conference on Learning Representations (ICLR)*, 2020.
- Query2box: Reasoning Over Knowledge Graphs In Vector Space Using Box Embeddings. H. Ren, W. Hu, J. Leskovec. *International Conference on Learning Representations (ICLR)*, 2020.

- Code:
  - http://snap.stanford.edu/graphsage
  - http://snap.stanford.edu/decagon/
  - https://github.com/bowenliu16/rl_graph_generation
  - https://github.com/williamleif/graphqembed
  - https://github.com/snap-stanford/GraphRNN