University of Pittsburgh

BIOINF 2071

Machine Learning on Graphs

February 3, 2021
Sanya Bathla Taneja
sbt12@pitt.edu

Machine Learning (ML)

The field of machine learning studies the design of computer programs (agents) capable of learning from past experience or adapting to changes in the environment.

Biomedical examples:

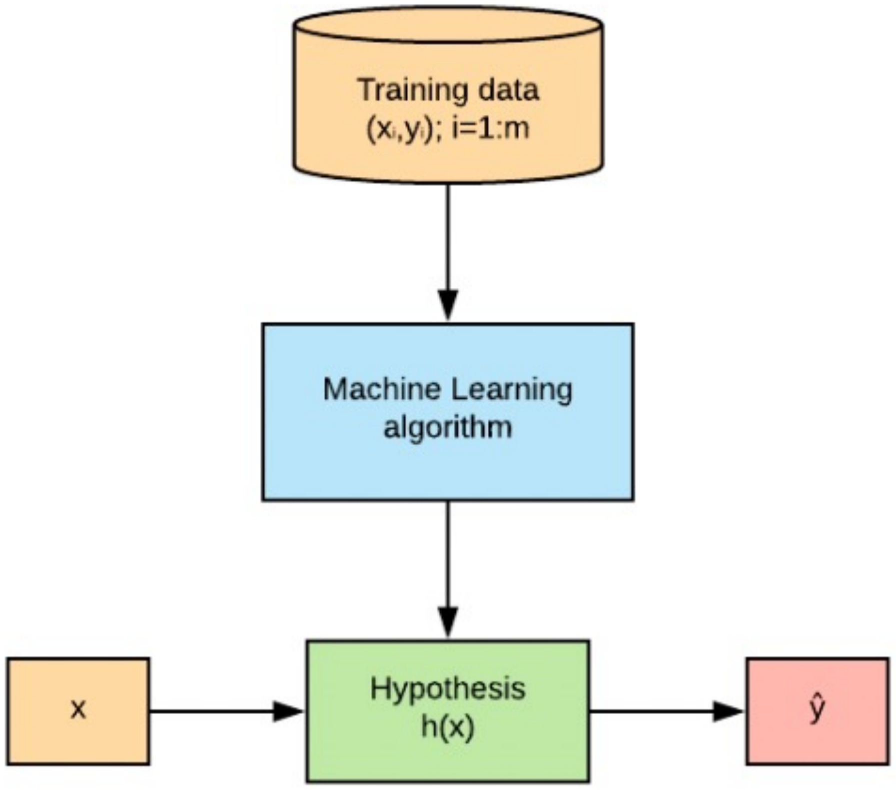
- Diagnose and treat illness with structured EHR data (CDS)
- Cancer detection with images/scans
- Pathology and radiology assistants
- Predicting gene expressions
- Drug design development
- Drug repurposing

Machine Learning (ML)

The field of machine learning studies the design of computer programs (agents) capable of learning from past experience or adapting to changes in the environment.

Biomedical examples:

- Diagnose and treat illness with structured EHR data (CDS)
- Cancer detection with images/scans
- Pathology and radiology assistants
- Predicting gene expressions
- Drug design development
- Drug repurposing



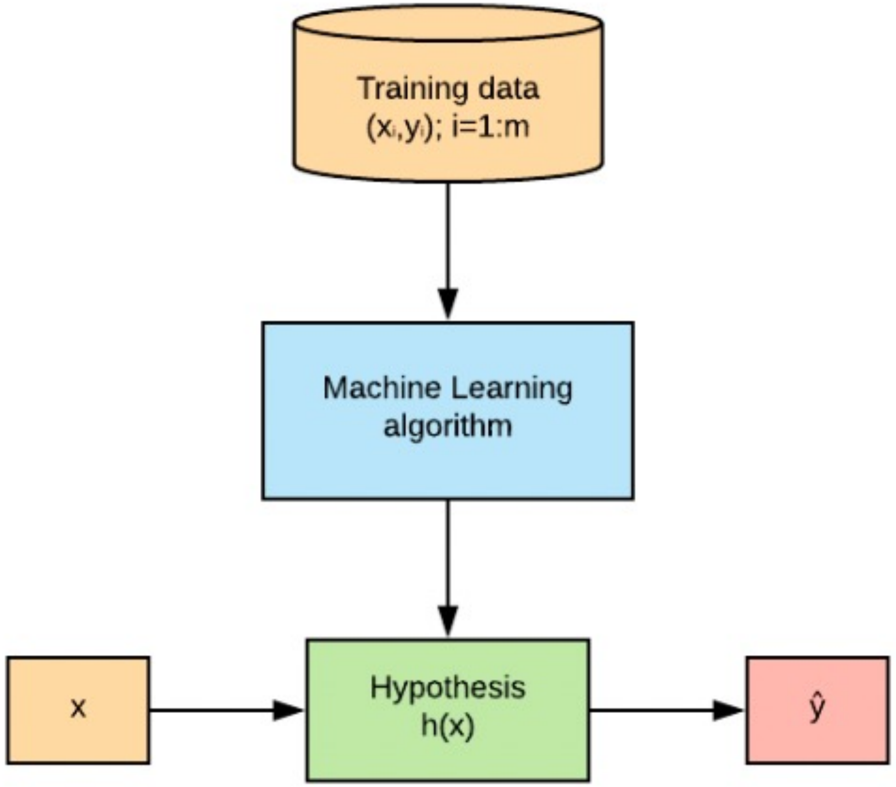
Machine Learning (ML)

Training Data → Features of model

	Age	Gender	Blood Pressure
Patient 1			
Patient 2			
…			

Patient characteristics such as –

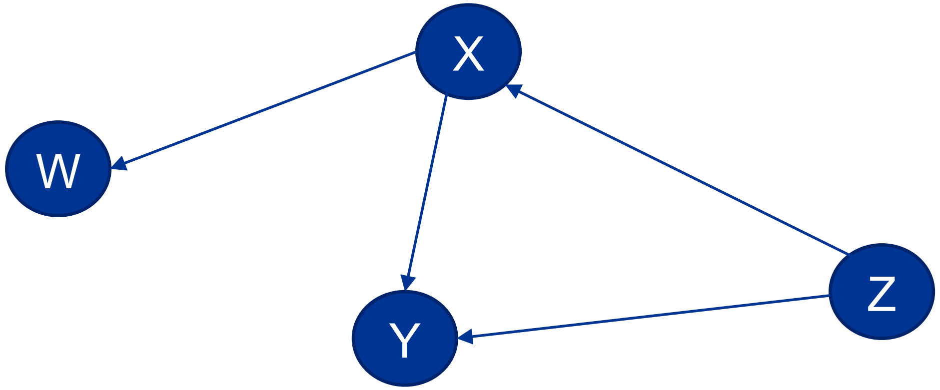
- Demographics
- Vitals
- Diagnoses
- Medical history
- Lab results
- Medication orders
- Family history
- CT scans
- Radiology reports



ML on graphs

Make predictions or discover new patterns using **graph-structured data** as feature information.

Graph structured data:



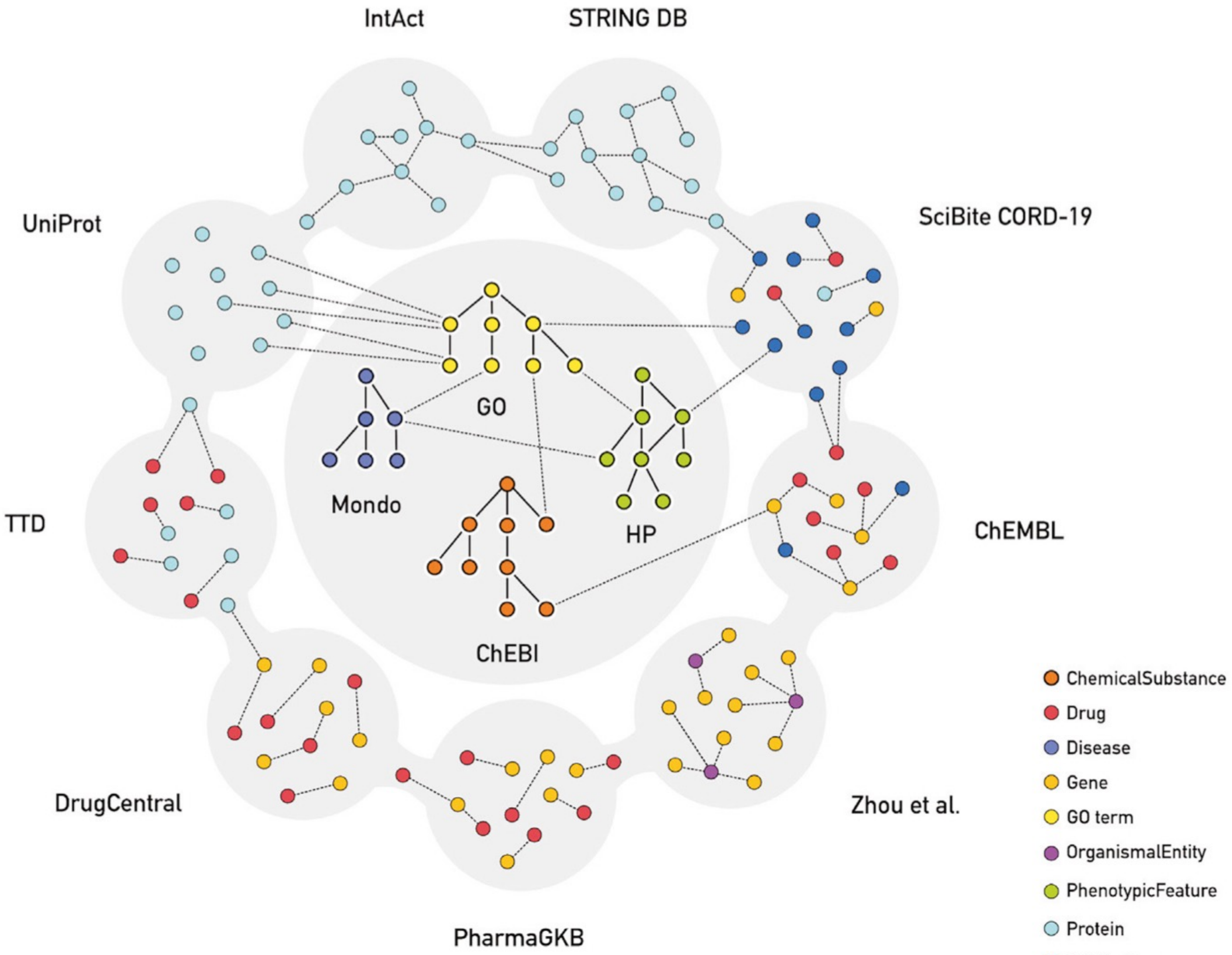
	W	X	Y	Z
W	0	0	0	0
X	1	0	1	0
Y	0	0	0	0
Z	0	1	1	0

Adjacency matrix

Nodes: W, X, Y, Z (*diseases, genes, molecules, enzymes …*)
Edges: relationships between the nodes (*causes, interacts with, has gene, participates in …*)

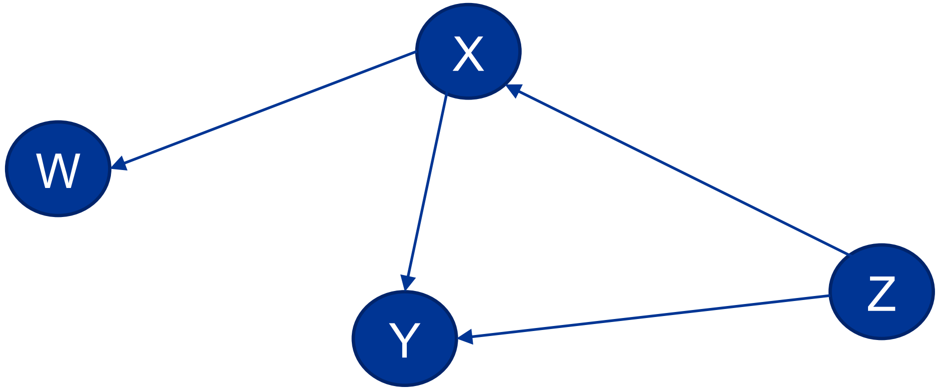
Question: Can ontologies be represented as graphs?

Kg-covid-19



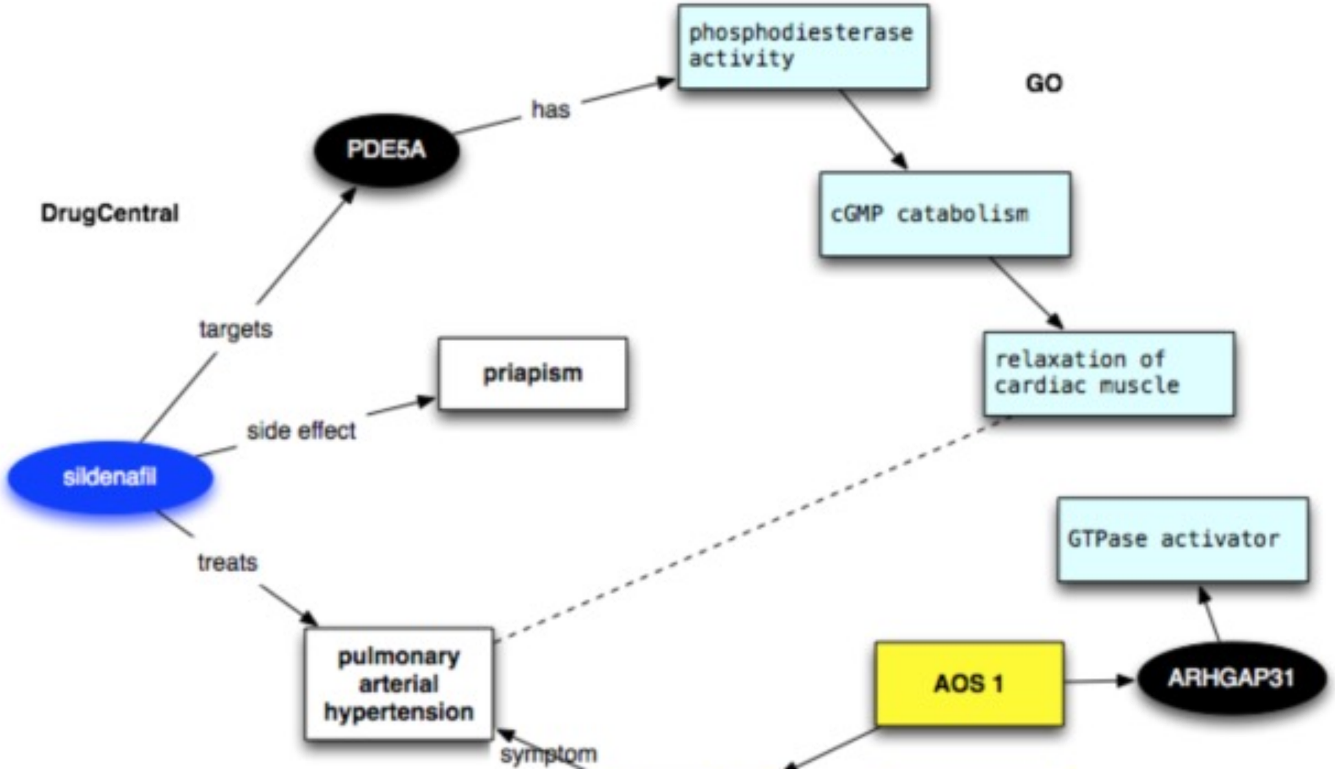
https://github.com/Knowledge-Graph-Hub/kg-covid-19

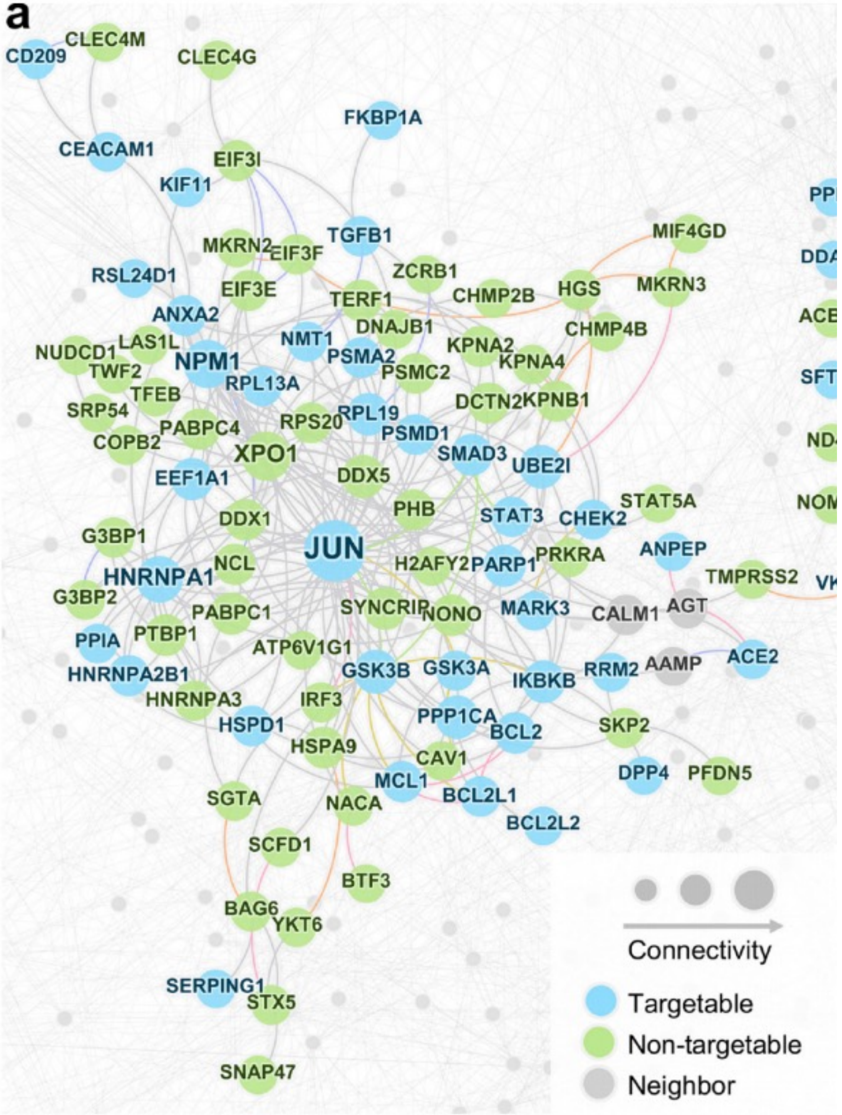
Homogenous vs heterogeneous graphs



Unique node “type”
Unique edge “type”

[Disease] <related_to> [Disease]





SARS-CoV-2 interaction graph

Why ML on graphs?

Make predictions or discover new patterns using **graph-structured data** as feature information.

- predict the role of a person in a collaboration network
- recommend new friends to a user in a social network
- predict new therapeutic applications of existing drug molecules (represented as graphs)
- **Link prediction:** find missing links in biological interaction graphs
- **Node classification:** classify the role of a protein in a biological interaction graph

https://cs.mcgill.ca/~wlh/comp766/files/graphs-against-covid.pdf

Why ML on graphs?

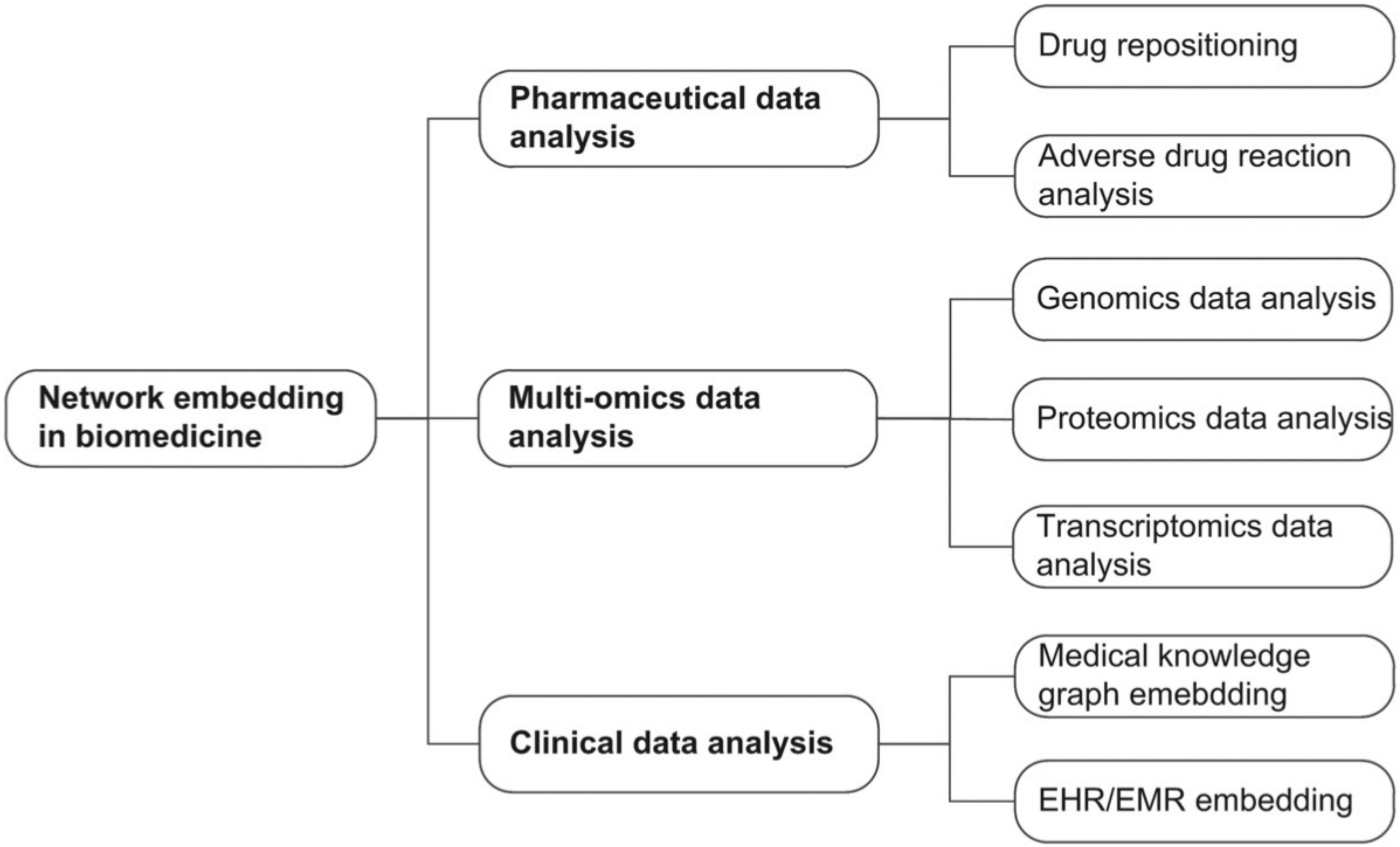


Figure 3. Illustration of applications of network embedding in biomedical data science.

Su C, Tong J, Zhu Y, Cui P, Wang F. Network embedding in biomedical data science. Briefings in bioinformatics. 2020 Jan;21(1):182-97.

ML on graphs

Apply machine learning algorithms (logistic regression, naïve Bayes, neural networks) with graph data.

What information do we want from the graph for machine learning?

- Position of node (local or global) in the graph
- Local neighborhood of a node (nodes+edges)
- Similarity between nodes (such as number of common edges)

Problems with encoding graphs:

- High dimensional representation (millions of nodes and edges)
- Statistical or kernel functions of graph data:
    - Time-consuming
    - Expensive (computation power)
    - Hand picked structural information

Graph Representation Learning (GRL)

Learn representations that encode structural information about the graph

Previous work

- treated encoding as a pre-processing step
- using hand-engineered statistics to extract structural information.

GRL

- treat encoding as machine learning task itself
- using a data-driven approach to learn **embeddings** that encode graph structure

Goal: Downstream tasks such as

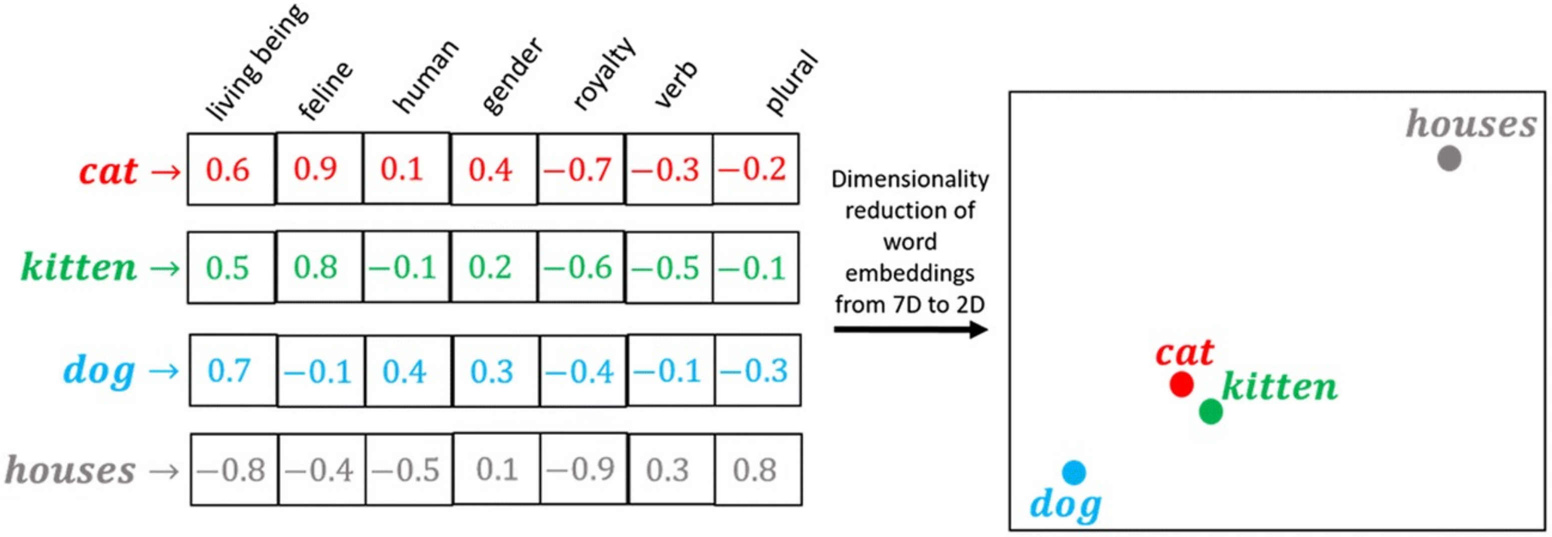
- Link prediction
- Graph completion
- Node classification

Embeddings on graphs

An **embedding** is a relatively low-dimensional space into which you can translate high-dimensional vectors.

Word embeddings (word2vec)

N-dimensional vectors



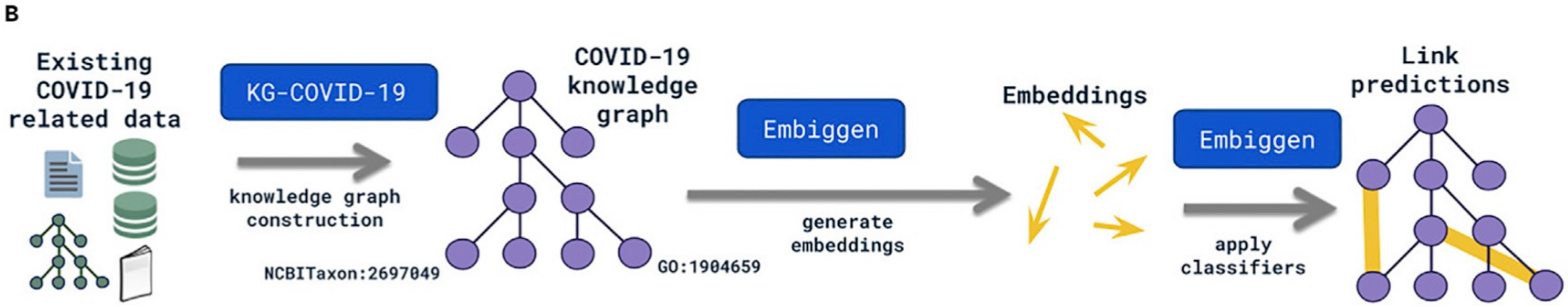
Node or graph or network embeddings

- Nodes that are in the same neighborhood in the original graph should be close in the embedding space
- “Local neighborhood” of node

Link prediction example

ML on KG-COVID-19 to perform link prediction in order to identify links that correspond to actionable knowledge:

- links between drugs and the COVID-19 disease
- links between drugs and SARS-CoV-2 protein targets
- links between drugs and host proteins that are involved in COVID-19 disease processes



Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Shefchek KA, Good BM, Balhoff JP, Fontana T, Blau H. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Patterns. 2020 Nov 9;2(1):100155.

Biomedical GRL

OXFORD
Briefings in Bioinformatics, 21(1), 2020, 182–197
doi: 10.1093/bib/bby117
Advance Access Publication Date: 10 December 2018
Review article

Network embedding in biomedical data science

Chang Su, Jie Tong, Yongjun Zhu, Peng Cui and Fei Wang

Corresponding author: Fei Wang, Division of Health Informatics, Department of Healthcare Policy and Research, Weill Cornell Medicine, Cornell University, 425 East 61 Street, New York, NY 10065, USA. Tel.: +1-646-962-9405; E-mail: few2001@med.cornell.edu

Crichton *et al. BMC Bioinformatics* (2018) 19:176
https://doi.org/10.1186/s12859-018-2163-9
BMC Bioinformatics
RESEARCH ARTICLE Open Access

Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches

Gamal Crichton*, Yufan Guo, Sampo Pyysalo and Anna Korhonen

To Embed or Not: Network Embedding as a Paradigm in Computational Biology

Walter Nelson[1,2], Marinka Zitnik[3], Bo Wang[3,4,5], Jure Leskovec[3,6], Anna Goldenberg[1,5,7] and Roded Sharan[8]**

[1] Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada, [2] Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada, [3] Department of Computer Science, Stanford University, Stanford, CA, United States, [4] Peter Munk Cardiac Center, University Health Network, Toronto, ON, Canada, [5] Vector Institute, Toronto, ON, Canada, [6] Chan Zuckerberg Biohub, San Francisco, CA, United States, [7] Department of Computer Science, University of Toronto, Toronto, ON, Canada, [8] School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Neuro-symbolic representation learning on biological knowledge graphs

Mona Alshahrani[1], Mohammad Asif Khan[1], Omar Maddouri[1,2], Akira R. Kinjo[3], Núria Queralt-Rosinach[4] and Robert Hoehndorf[1,]*

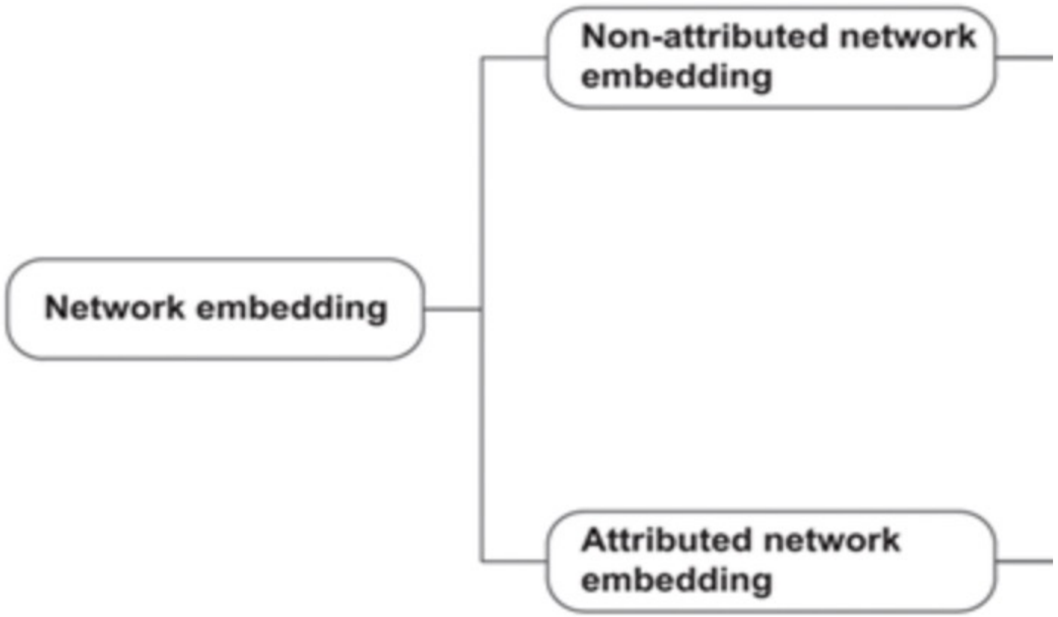
[1]Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia, [2]Life Sciences Division, College of Science & Engineering, Hamad Bin Khalifa University, HBKU, Doha, Qatar, [3]Institute for Protein Research, Osaka University 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan and [4]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, 92037 USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on December 13, 2016; revised on March 30, 2017; editorial decision on April 18, 2017; accepted on April 18, 2017

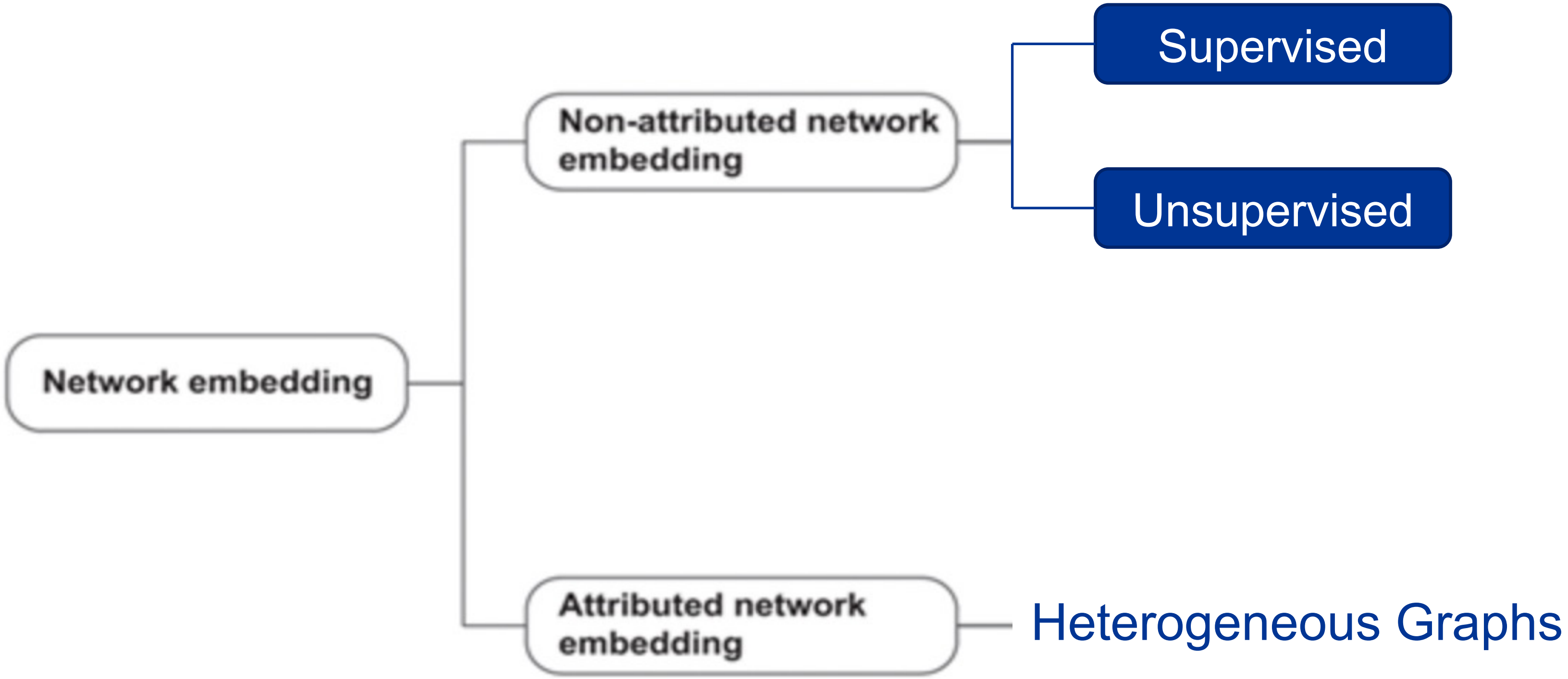
Embeddings on graphs



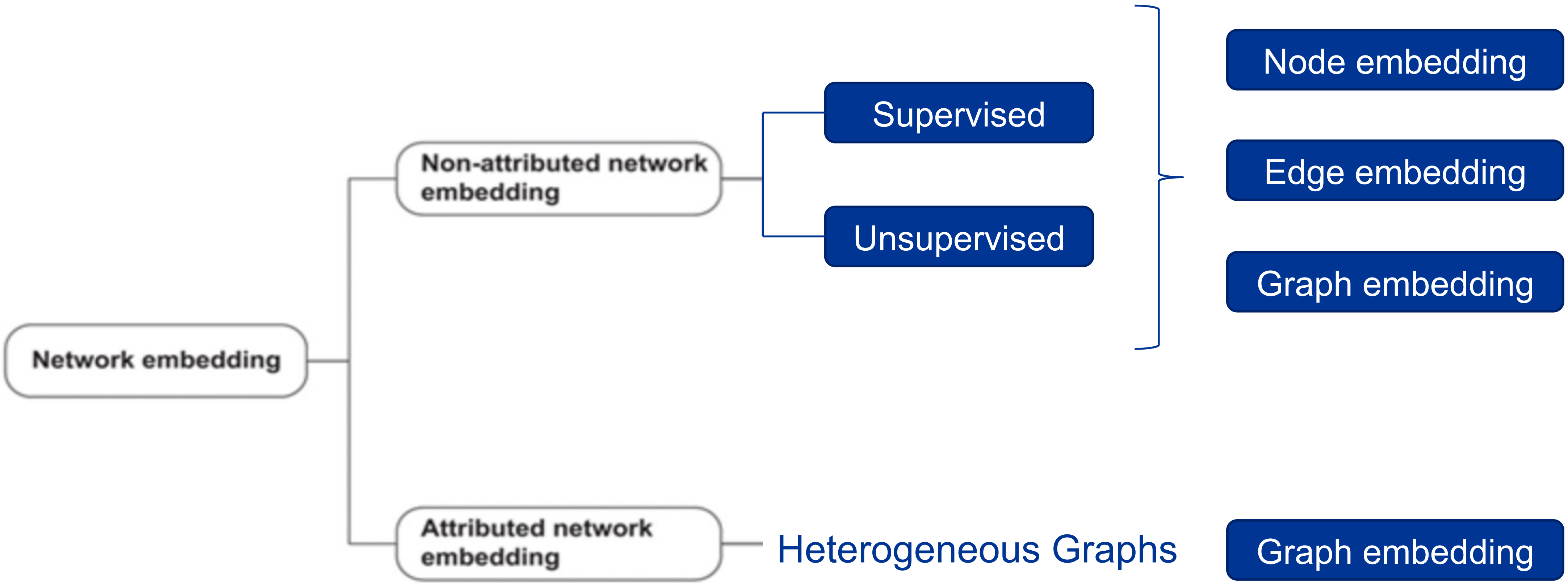
Homogeneous Graphs

Heterogeneous Graphs

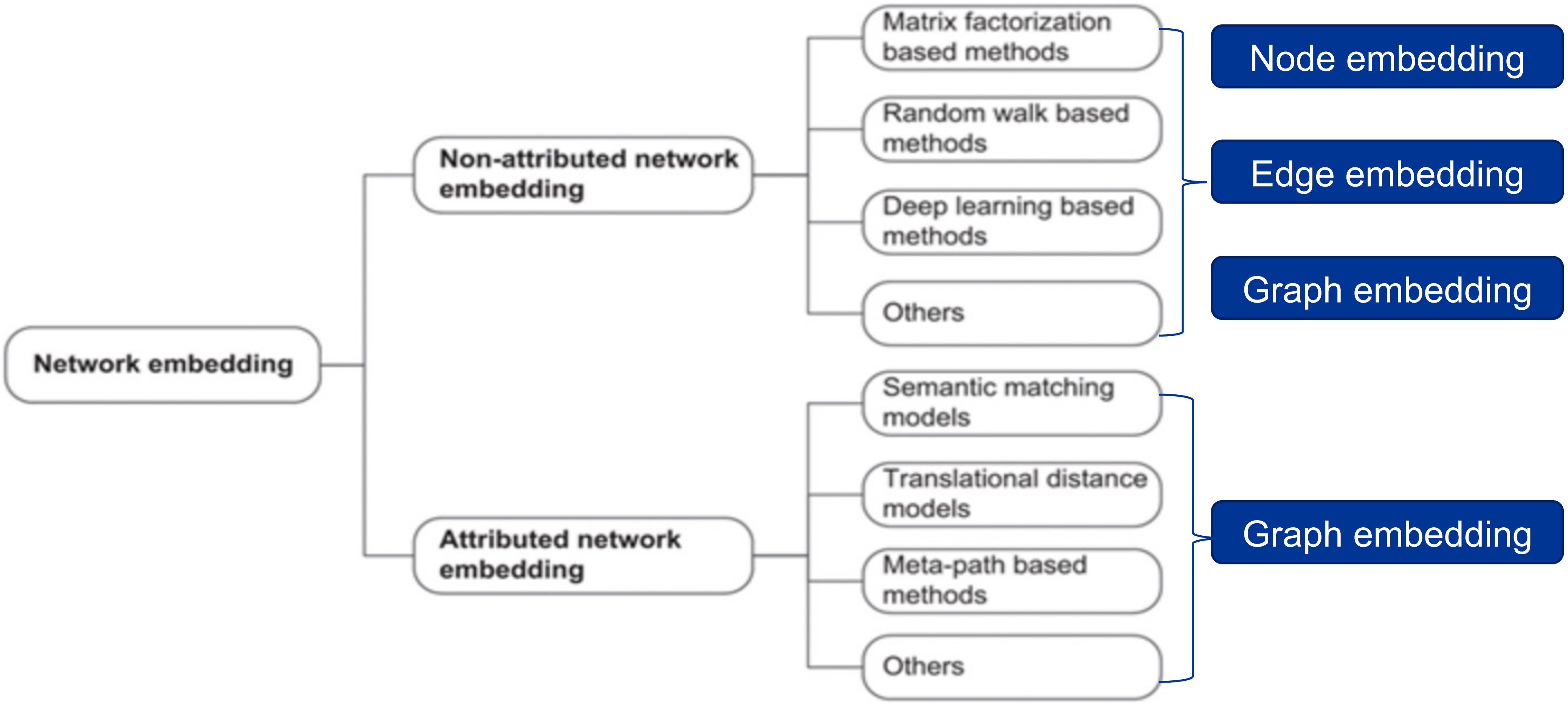
Embeddings on graphs



Embeddings on graphs

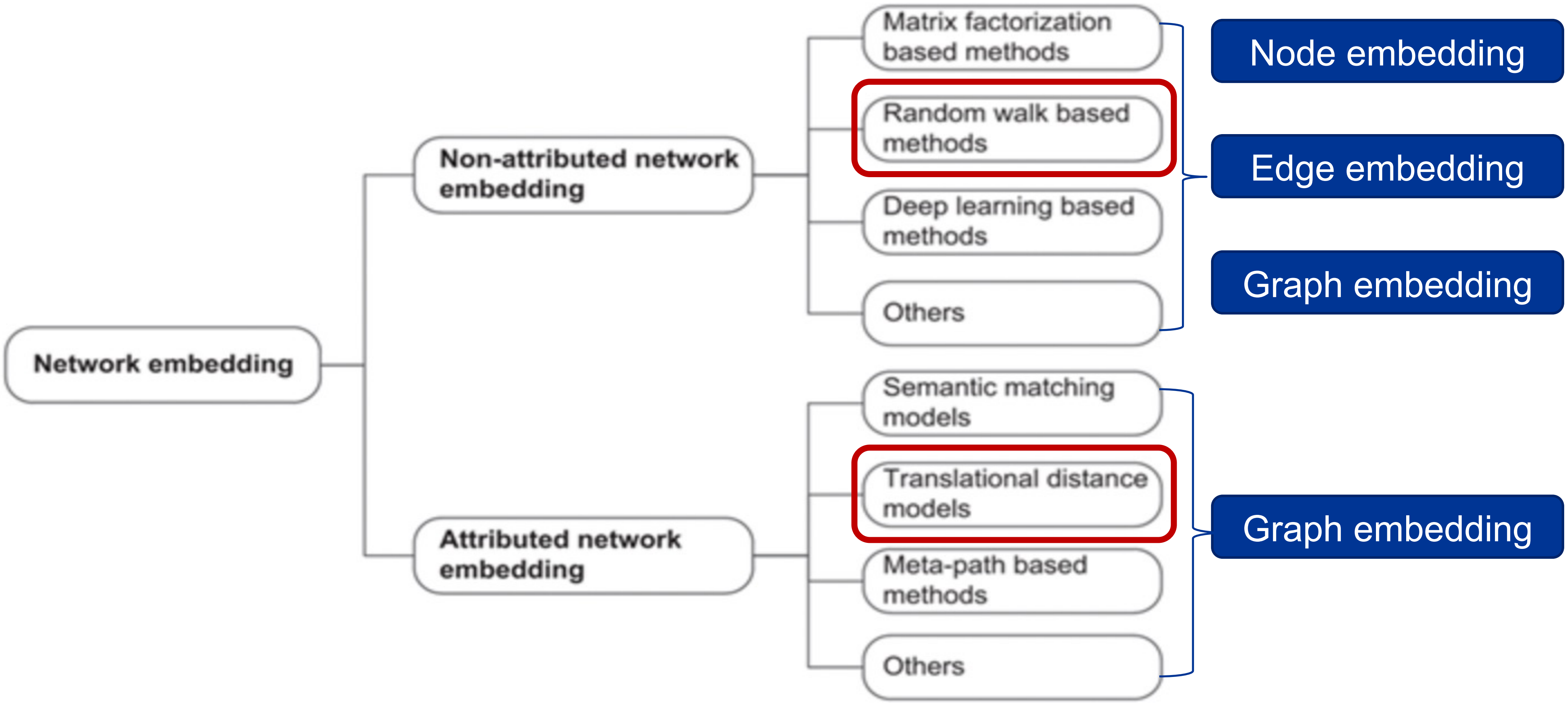


Embeddings on graphs



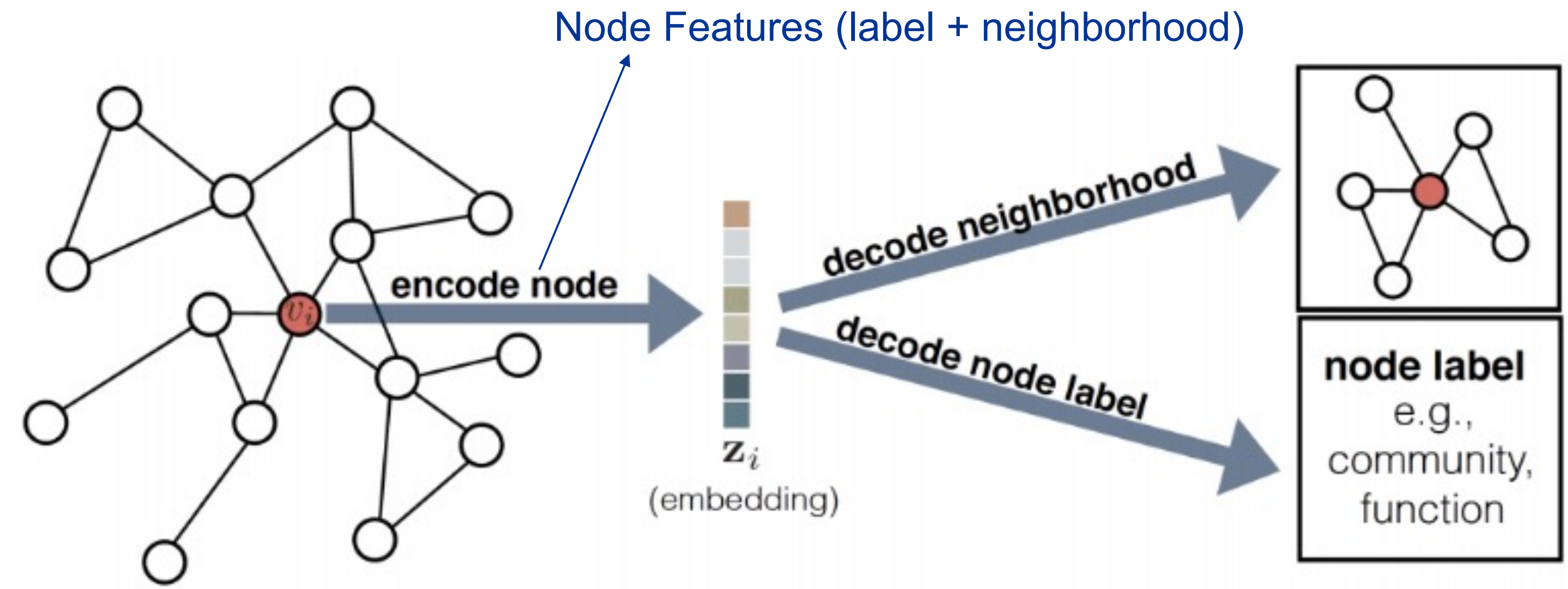
Su C, Tong J, Zhu Y, Cui P, Wang F. Network embedding in biomedical data science. Briefings in bioinformatics. 2020 Jan;21(1):182-97.

Embeddings on graphs

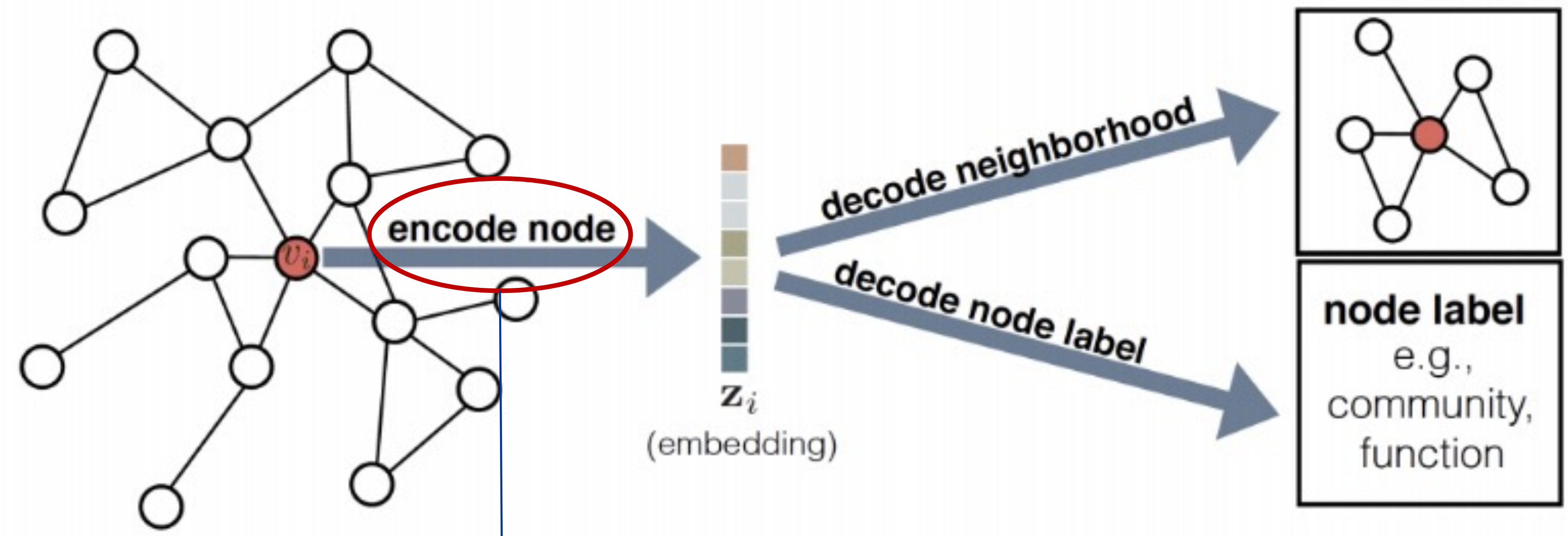


Su C, Tong J, Zhu Y, Cui P, Wang F. Network embedding in biomedical data science. Briefings in bioinformatics. 2020 Jan;21(1):182-97.

Node embeddings (Non-attributed)



Node embeddings (Non-attributed)



Shallow embeddings

Auto-encoding methods

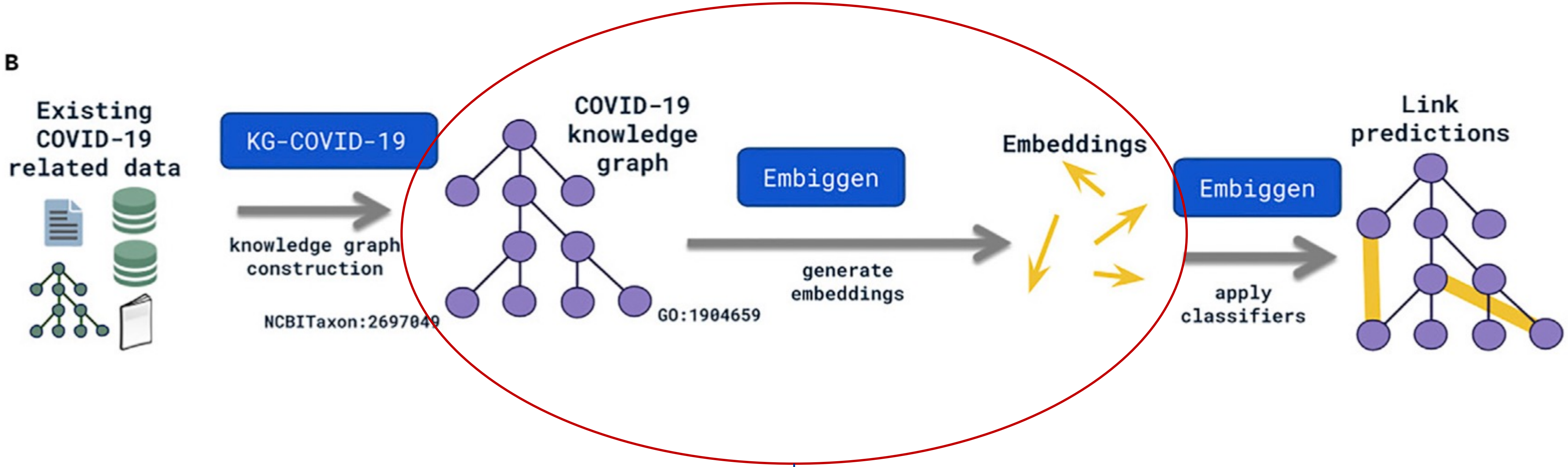
Graph Regularization

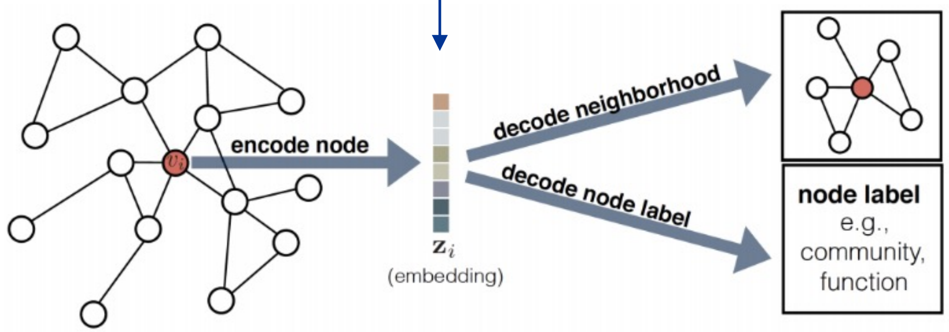
Graph Neural Networks

Random walk-based methods

Hamilton WL, Ying R, Leskovec J. Representation learning on graphs: Methods and applications. arXiv preprint arXiv:1709.05584. 2017 Sep 17.

Node embeddings (Non-attributed)

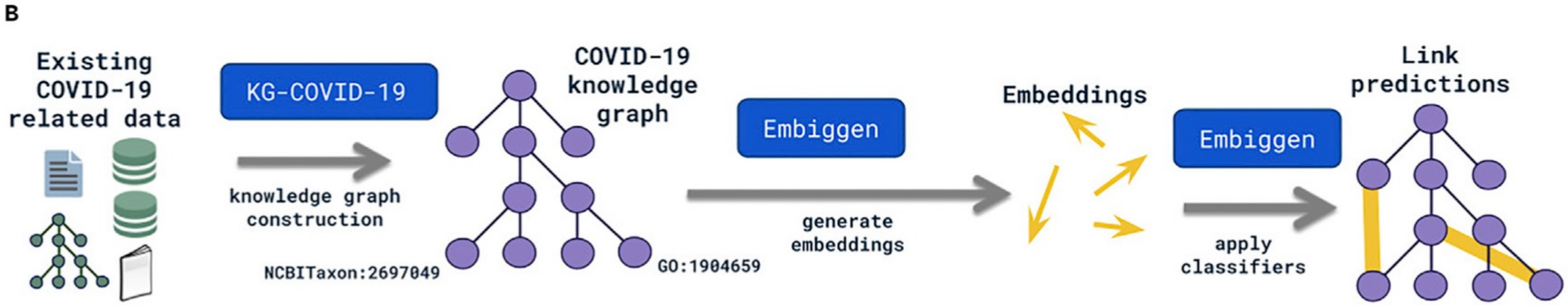




ASSUMPTION VIOLATED!!
Which one?

Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Shefchek KA, Good BM, Balhoff JP, Fontana T, Blau H. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Patterns. 2020 Nov 9;2(1):100155.

Node embeddings (Non-attributed)

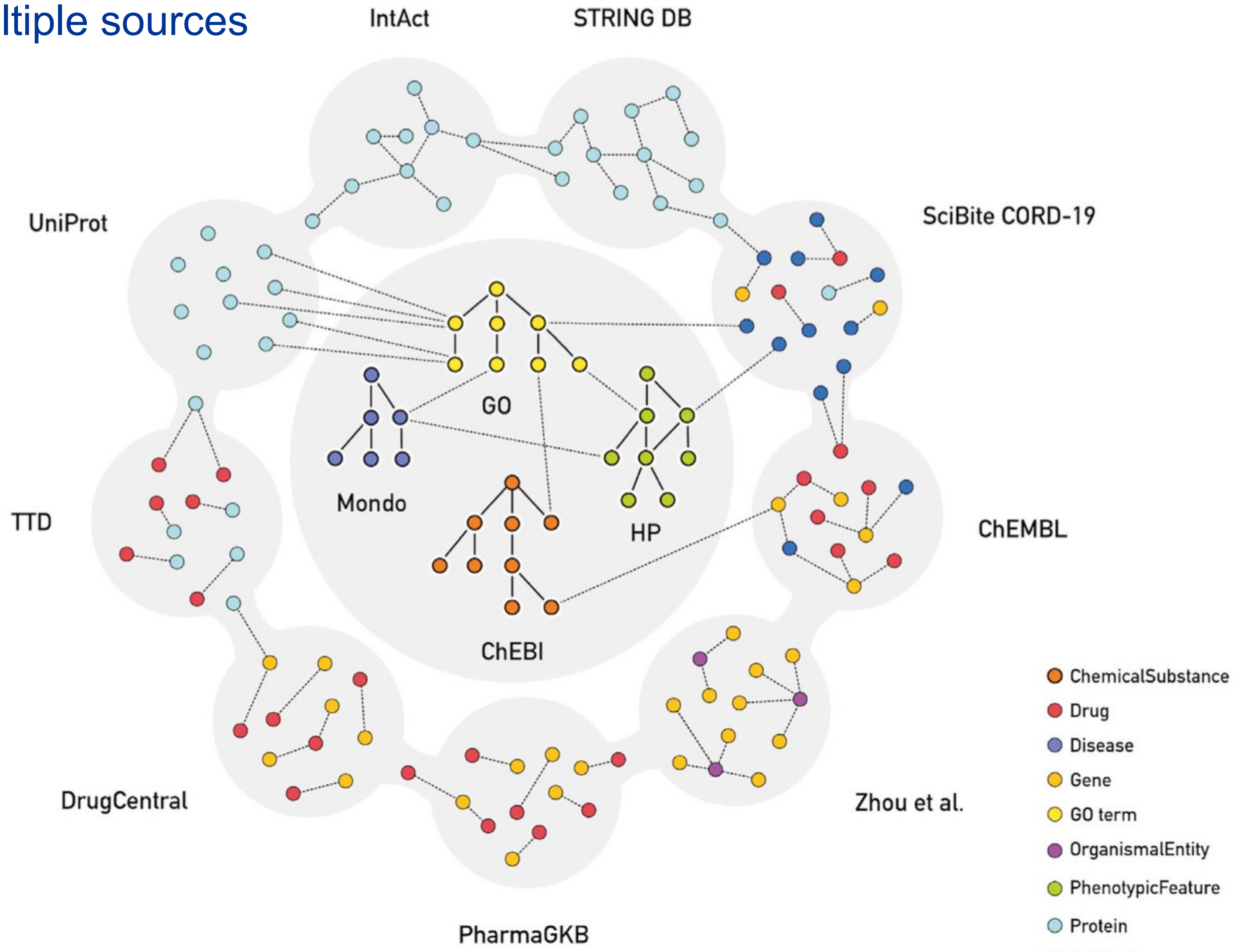


ASSUMPTION VIOLATED!!
-- Non-attributed embeddings built for homogeneous graphs

KG-COVID-19 created from multiple sources – ontologies, databases (DrugBank), literature (PubMed)

Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Shefchek KA, Good BM, Balhoff JP, Fontana T, Blau H. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Patterns. 2020 Nov 9;2(1):100155.

KG – created from multiple sources



https://github.com/Knowledge-Graph-Hub/kg-covid-19

Node embeddings (Non-attributed)

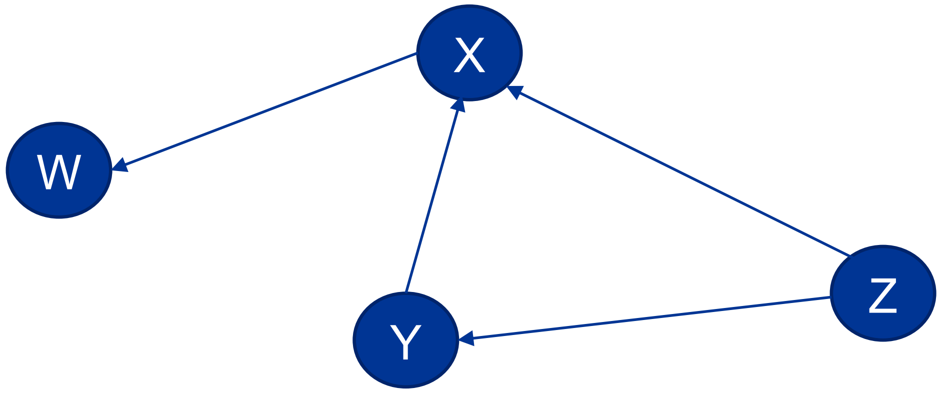
Node2Vec: Random walk-based algorithms for node embeddings

- Explore neighborhood of each node with random walks
- Uses the word2vec skip-gram model for word embeddings after generating random walks
- Skip-gram word2vec: https://ronxin.github.io/wevi/
- Subgraphs from random walk equivalent to sentences in word2vec corpus
- Hyperparameters:
    - Number of walks
    - Walk length
    - Window size (same as word2vec)
    - Dimensionality
    - P and Q – random walk parameters

Node2Vec

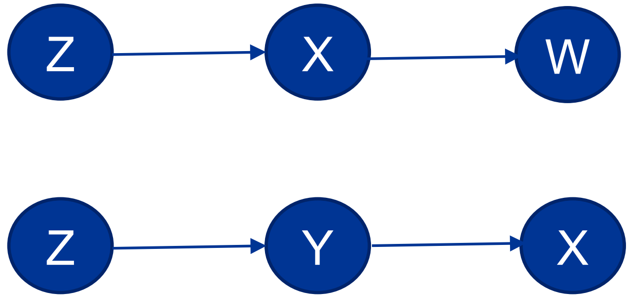
Node2Vec: Random walk-based algorithms for node embeddings

- Explore neighborhood of each node with random walks



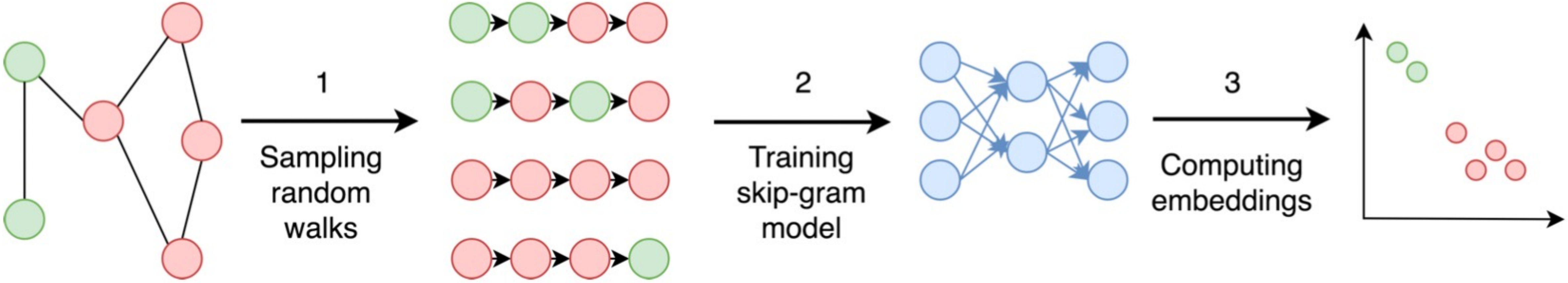
Adjacency Matrix

	W	X	Y	Z
W	0	0	0	0
X	1	0	0	0
Y	0	1	0	0
Z	0	1	1	0



Random walk subgraphs from random walk length = 3

Node2Vec



Random walk subgraphs
= sentences

One hot encoding of sentences
– input to neural network

Number of features
= dimensionality of
embedding vector

https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007

Graph embeddings (Non-attributed)

- Encode both nodes and edges – more complex

- Fully supervised approach where attributes of nodes in graph are always known

- Example: Embed graphs of different molecules to predict their therapeutic properties

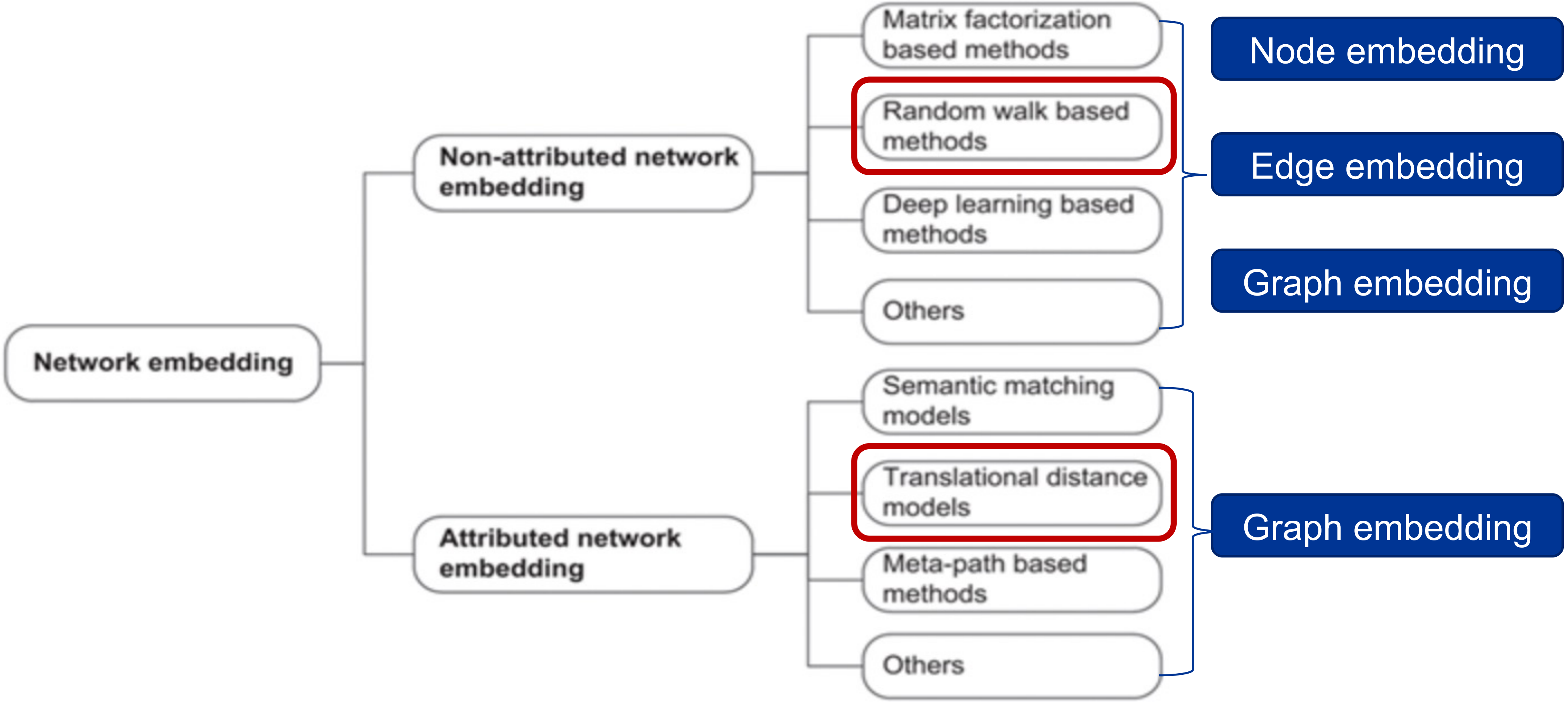
- Approach
  - Equate subgraphs with **sets of node embeddings**
  - Generate node embeddings (with node2vec) and **aggregate** for each subgraph (example – one molecule)
  - Aggregation may be summation, clustering, combing node and edge embeddings

Open problems in GRL

- Scalability – billions of nodes/edges
- Innovation in decoders – pairwise similarity is most common
- Modeling dynamic, temporal graphs
- Beyond graph classification – generating candidate subgraphs from embeddings
- Interpretability
- Heterogeneous graphs – node embeddings get more complicated with multi-modal data or even different node/edge types

Su C, Tong J, Zhu Y, Cui P, Wang F. Network embedding in biomedical data science. Briefings in bioinformatics. 2020 Jan;21(1):182-97.

Embeddings on graphs



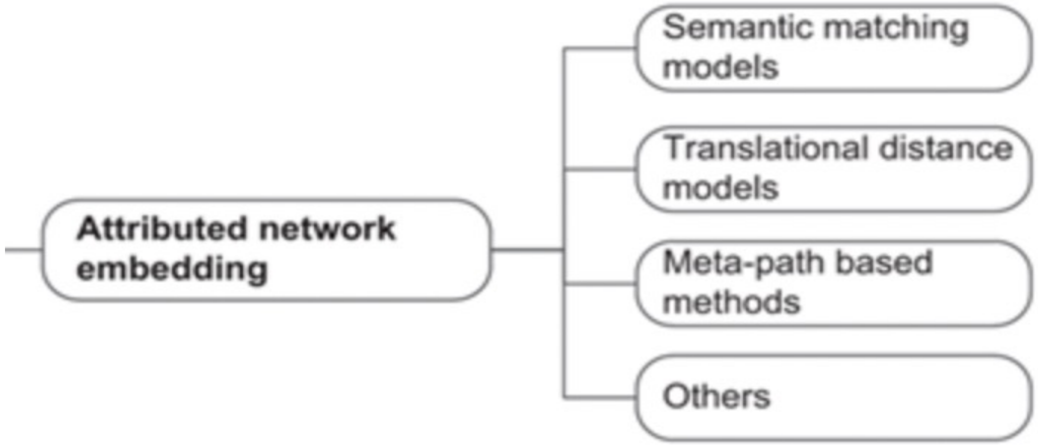
Su C, Tong J, Zhu Y, Cui P, Wang F. Network embedding in biomedical data science. Briefings in bioinformatics. 2020 Jan;21(1):182-97.

KG embeddings (Attributed)

Why?

- Embedding of **heterogeneous, large-scale knowledge graphs (KGs)**
- Multiple node types *(diseases, genes, chemicals)*
- Multiple edge types *(relation ontology – causes, interacts with, participates in)*
- Real world applications – homogeneous graphs are rare
- Applications are similar – link prediction, node classification, graph completion, hypothesis generation

Goal: embed components of KG to continuous vector space and preserve the inherent structure



Translational Models (Attributed)

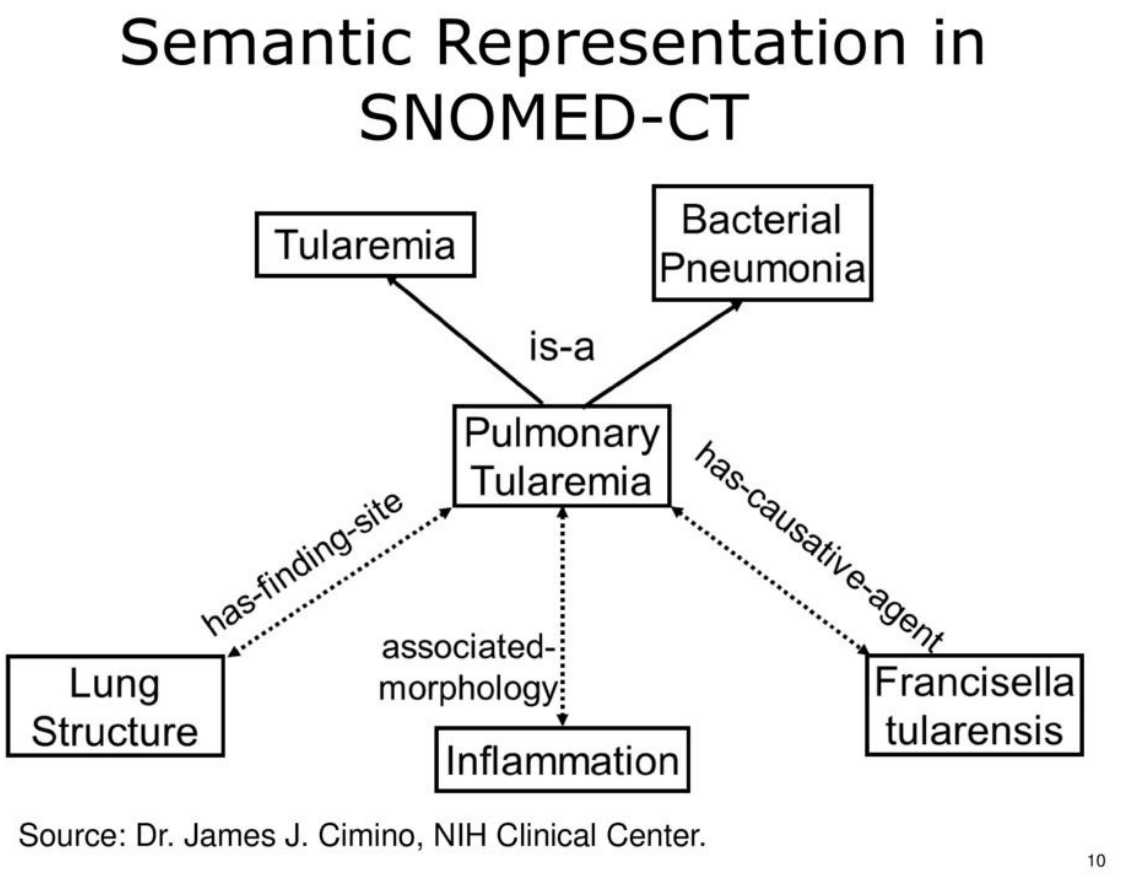
- Embed knowledge graph to continuous vector space while preserving properties of the original graph

- (Head, Relation, Tail) triples translated to the embedded space

Head ~ Subject
Relation ~ Predicate
Tail ~ Object

Translational Models (Attributed)

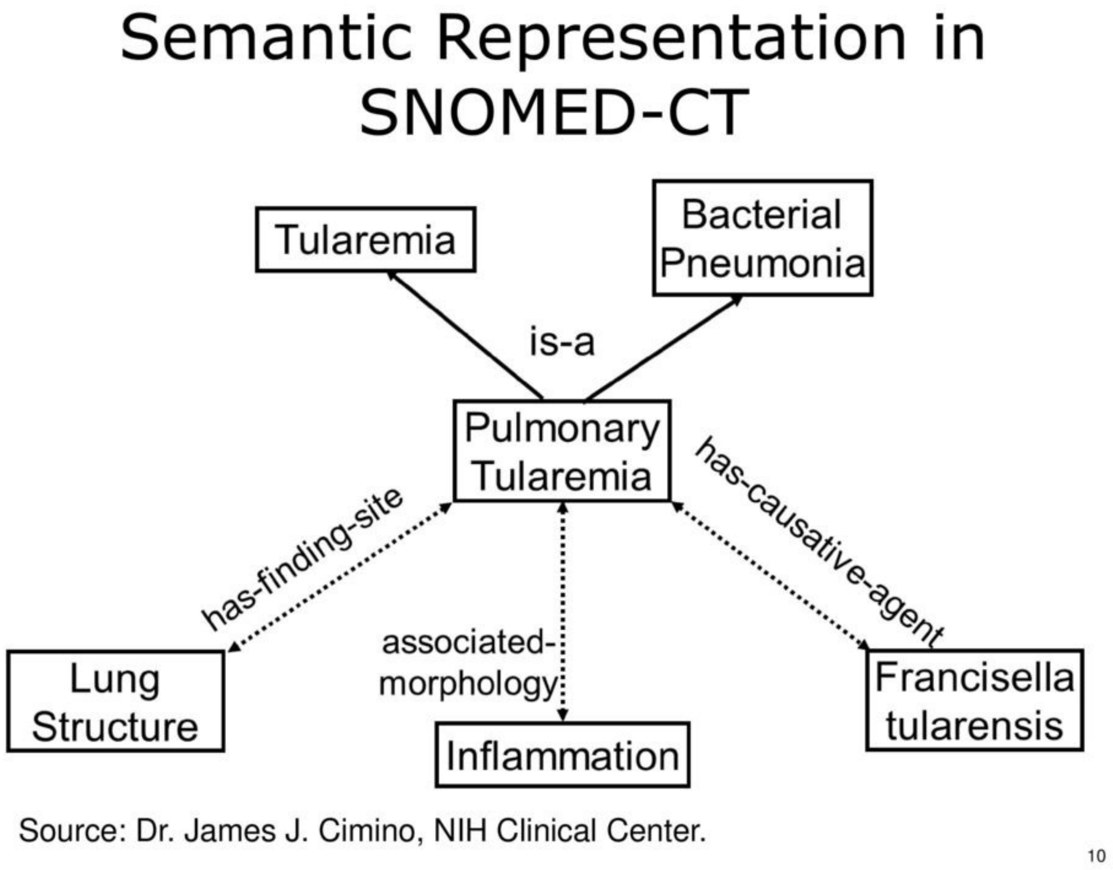
Head ~ Subject
Relation ~ Predicate
Tail ~ Object



Triples?

Translational Models (Attributed)

Head ~ Subject
Relation ~ Predicate
Tail ~ Object



Triples?

- Pulmonary Tularemia *<is-a>* Tularemia
- Pulmonary Tularemia *<is-a>* Bacterial Pneumonia
- Pulmonary Tularemia *<associated-morphology>* Inflammation

Translational Models (Attributed)

(Head, Relation, Tail) triples translated to the embedded space

Head ~ Subject
Relation ~ Predicate
Tail ~ Object

Triple
(h, r, t) → **Embedding/Vector**
h = (…)
r = (…)
t = (…)

- Head/Tail are vectors and the relation r is an operation in the embedding space (eg. Linear translation, projection) – represented as vectors **h**, **r** and **t**

- Representations of entities and relations are obtained by minimizing a global loss function involving all entities and relations.

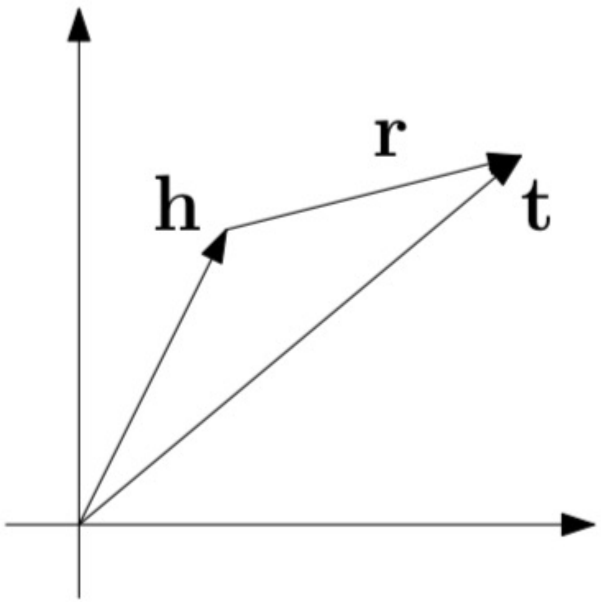
TransE (Attributed)

TransE performs linear transformation, and the scoring function is negative distance between:

Distance based embedding optimization - score function

$$f_r(h,t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2$$

$$L(h,r,t) = max(0, f_{pos} - f_{neg} + margin)$$



(a) TransE

TransH (Attributed)

Goal: Represent **relation** as **translating operation** on **hyperplane**

- Hyperplane: subspace of dimension (n-1)
- Relation: relation vector **r** – represented as 2 vectors on the hyperplane
  - Norm vector (wr)
  - Translation vector (dr)

Both norm and translation vectors are relation-specific

For a golden triplet (h, r, t) – taken from the knowledge graph - the *projections* of h and t on the hyperplane are expected to be connected by the translation vector dr with low error.

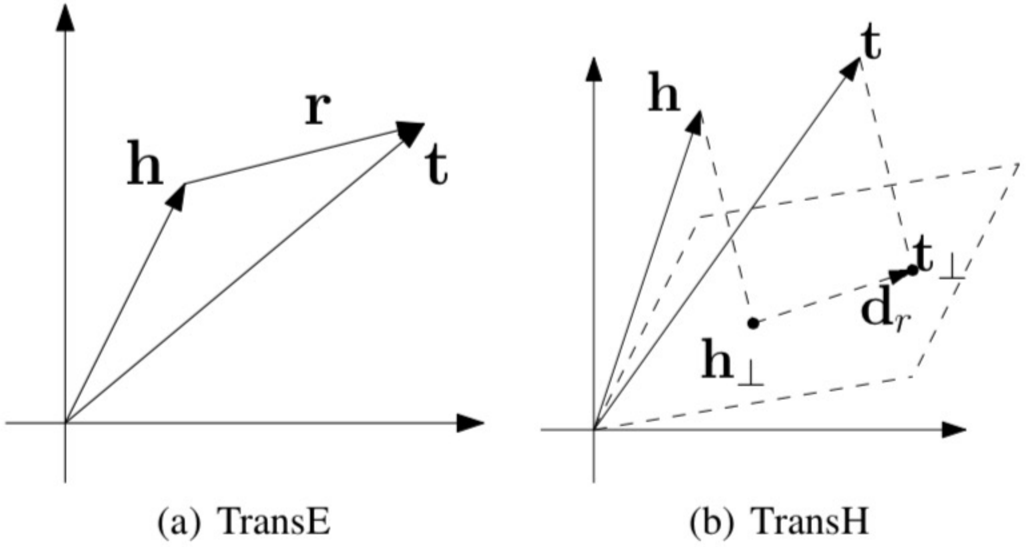


Figure 1: Simple illustration of TransE and TransH.

TransE vs TransH (Attributed)

Example:
1. (empire state building, location, NYC)
2. (ghostbusters, location, NYC)

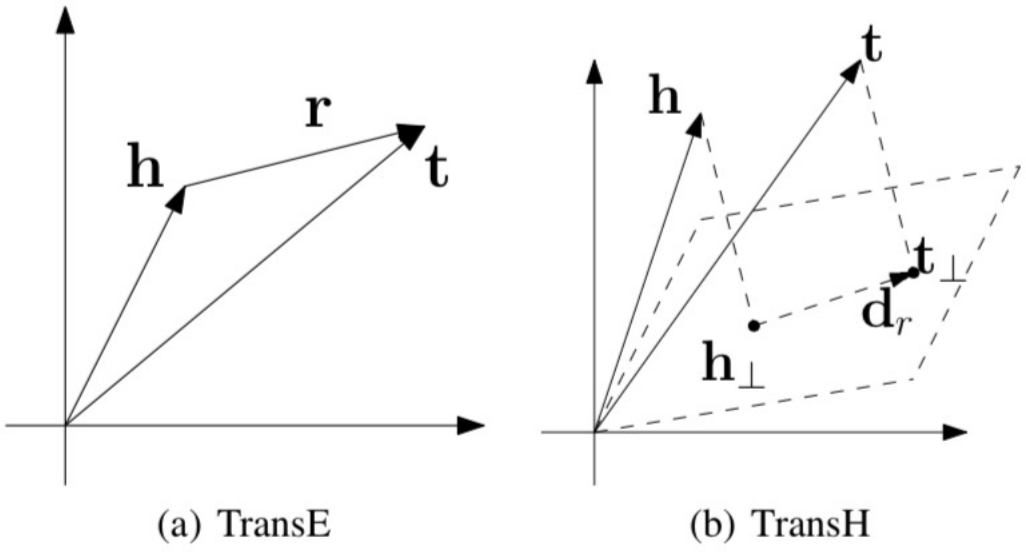


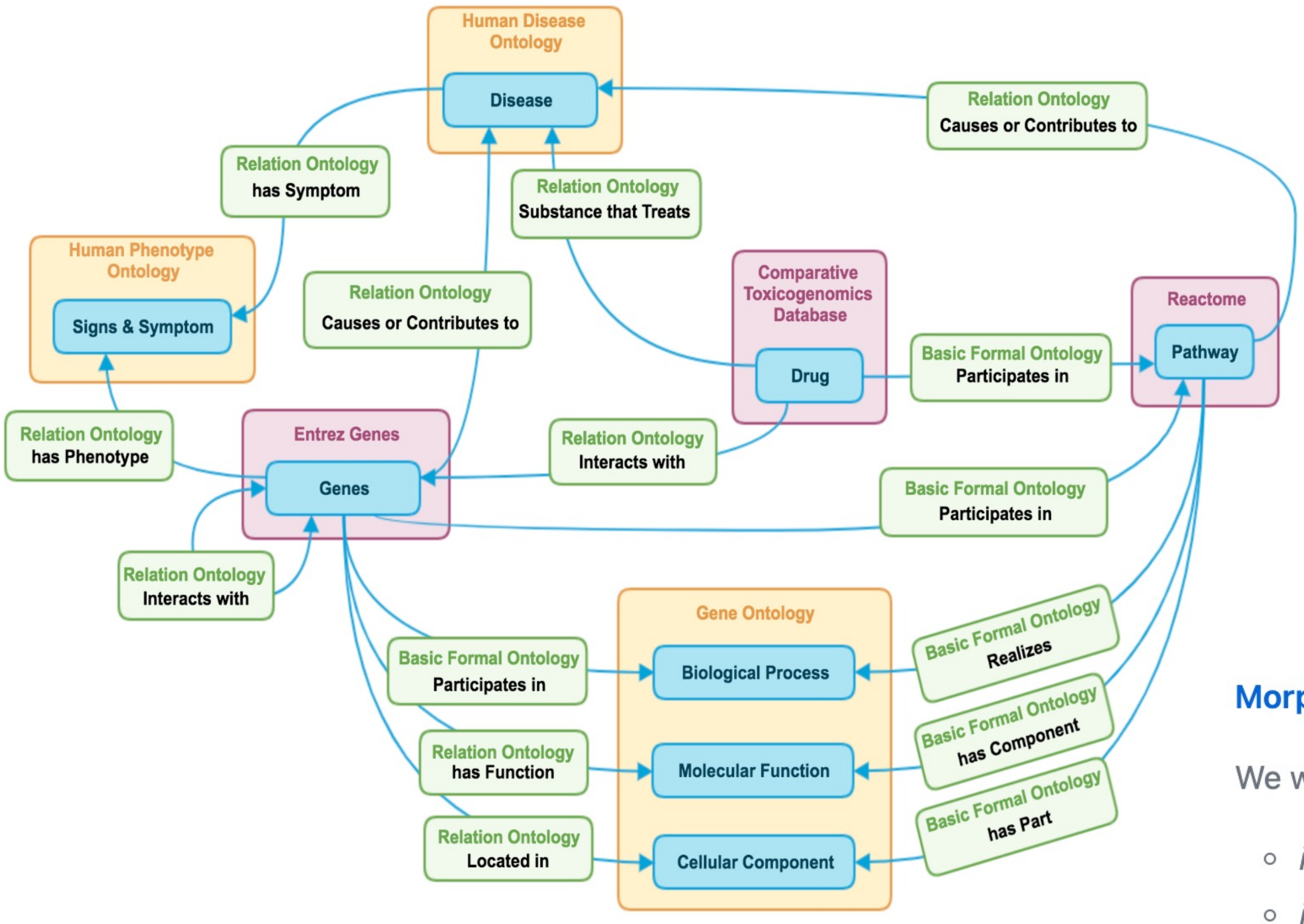
Figure 1: Simple illustration of TransE and TransH.

TransE
=> empire state building and ghostbusters close in semantic space (vectors) *but* have no or little similarity
- Entities are represented the same way in any relation

TransH
- Embeddings of 'Empire state building' and 'Ghostbusters' will be similar for a given relation 'location', however they might be far away from each other relative to other relations.

Also have TransR and TransD – further reading

KG embeddings



Morphine --> *is substance that treats* --> **Migraine**

We would need to create two axioms:

- *isSubstanceThatTreats(Morphine,* x_1*)*
- *instanceOf(*x_1*, Migraine)*

https://github.com/callahantiff/PheKnowLator/wiki/v1.0.0#molecular-mechanism-embeddings

Translational models (Attributed)

- Capable of handling heterogeneous, hierarchical data
- Not as intuitive as node embeddings
- Less widely used than node embeddings – node2vec, DeepWalk, LINE, PTE

Tools/Libraries

- Node2vec (in Python)
- DeepWalk (Python)
- StellarGraph library – all non-attributed algorithms for different tasks (Python)
- Embiggen (by MONARCH initiative)
- Tensorflow-TransX – translational models (Python)
- Scikit-kge
- NetworkX (Python graph library)

Other resources/links/bibliography

Su C, Tong J, Zhu Y, Cui P, Wang F. Network embedding in biomedical data science. Briefings in bioinformatics. 2020 Jan;21(1):182-97.

Crichton G, Guo Y, Pyysalo S, Korhonen A. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. BMC bioinformatics. 2018 Dec 1;19(1):176.

Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A, Sharan R. To embed or not: network embedding as a paradigm in computational biology. Frontiers in genetics. 2019;10:381.

Tripodi IJ, Callahan TJ, Westfall JT, Meitzer NS, Dowell RD, Hunter LE. Applying knowledge-driven mechanistic inference to toxicogenomics. Toxicology in Vitro. 2020 May 6:104877.

Hamilton WL, Ying R, Leskovec J. Representation learning on graphs: Methods and applications. arXiv preprint arXiv:1709.05584. 2017 Sep 17.

Chami I, Abu-El-Haija S, Perozzi B, Ré C, Murphy K. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. arXiv preprint arXiv:2005.03675. 2020 May 7. https://arxiv.org/pdf/2005.03675.pdf

Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Computational and structural biotechnology journal. 2020;18:1414.

Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. InAaai 2014 Jul 27 (Vol. 14, No. 2014, pp. 1112-1119).

Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. InNeural Information Processing Systems (NIPS) 2013 Dec 5 (pp. 1-9).

- https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007
- https://github.com/aditya-grover/node2vec
- https://stellargraph.readthedocs.io/en/stable/index.html
- https://cs.mcgill.ca/~wlh/comp766/files/graphs-against-covid.pdf