GLOBAL EXPLAINABILITY OF GNNs VIA LOGIC COMBINATION OF LEARNED CONCEPTS

AIT Journal Club
03/02/2023
Steve Azzolin
steve.azzolin1@gmail.com

Global Explainability of GNNs via Logic Combination of Learned Concepts. S. Azzolin et al., 2022. ICLR2023

Outline

1. GNNs are Cool 😎 (?)

2. Why XAI?

3. XAI for GNNs

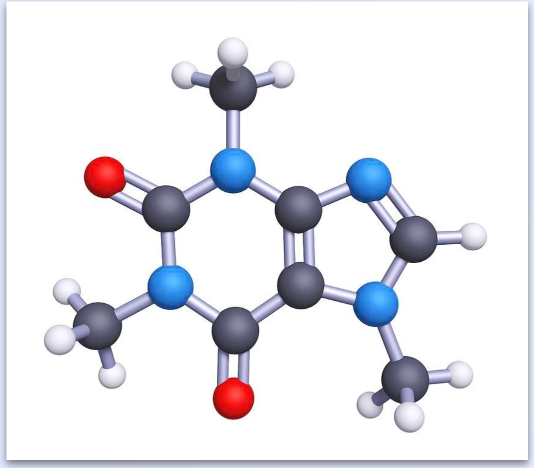
4. Proposed Solution

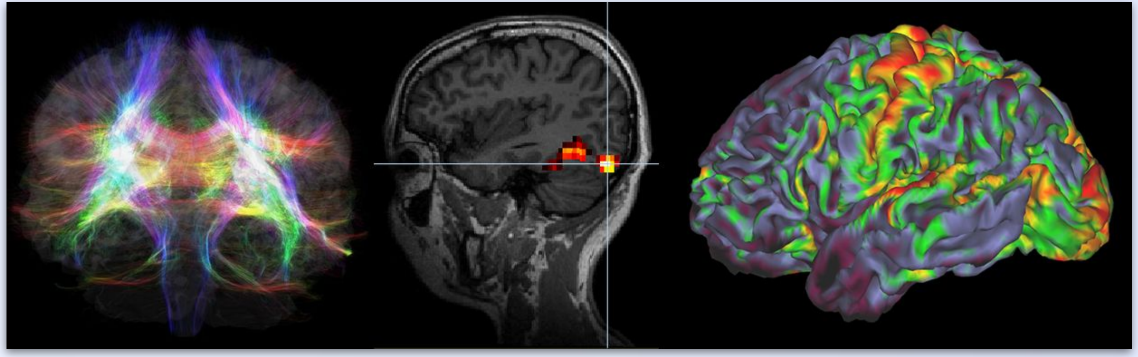
5. Results

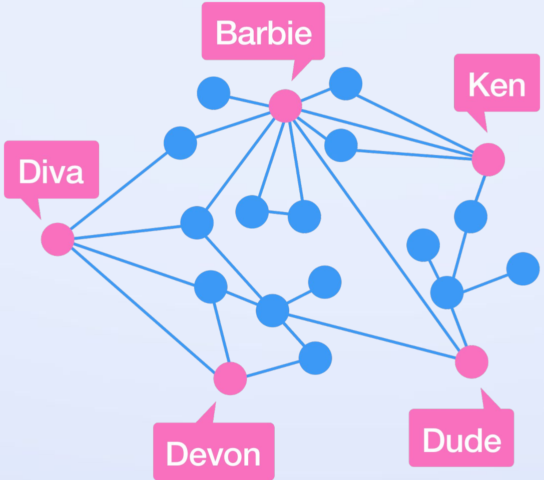
1.

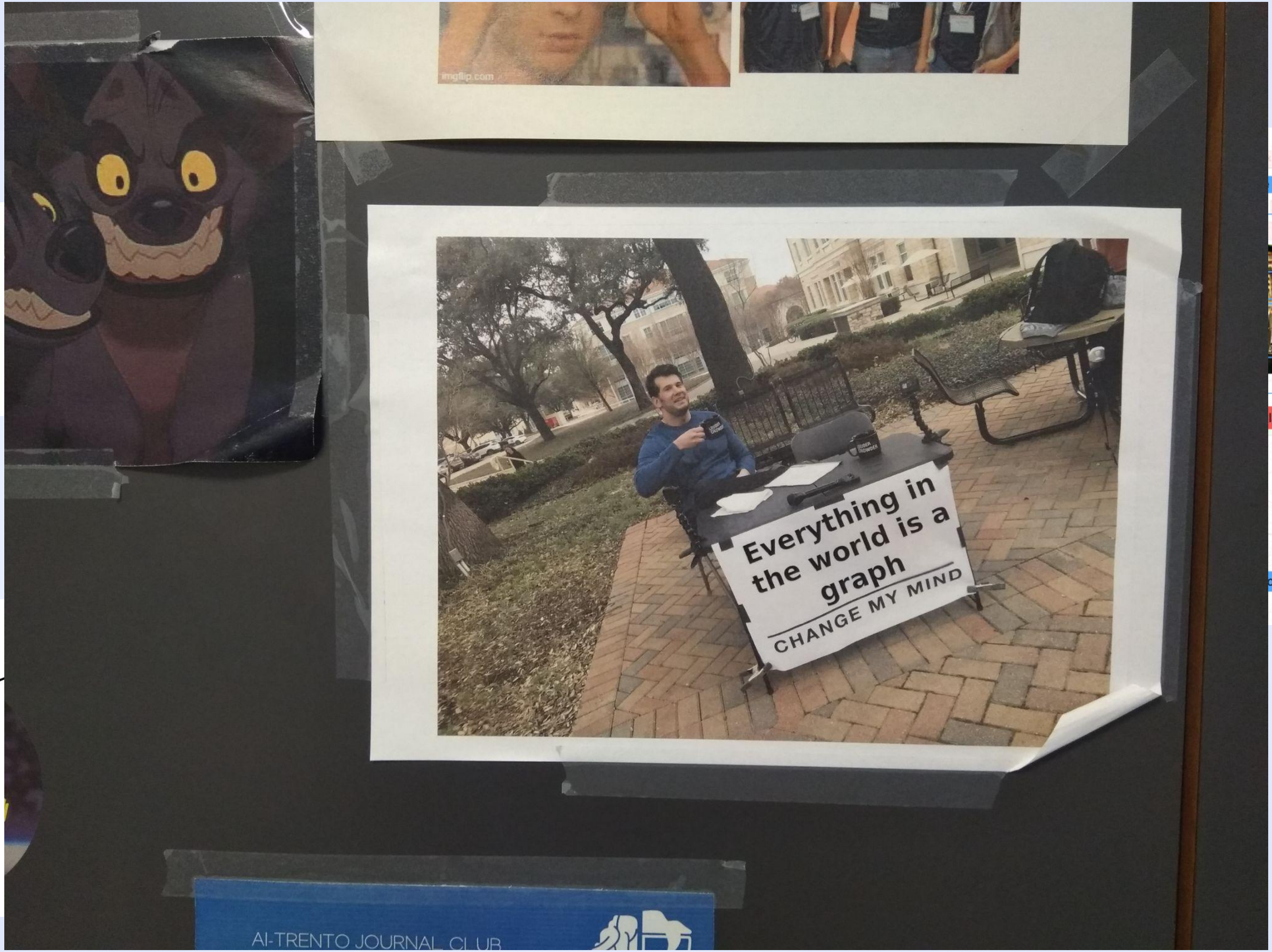
GNNs are Cool (?)

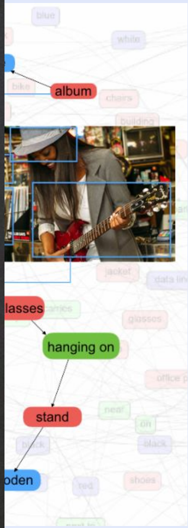
1. Why GNNs

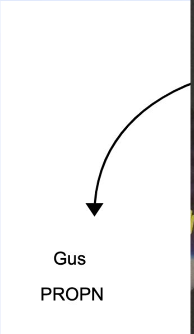




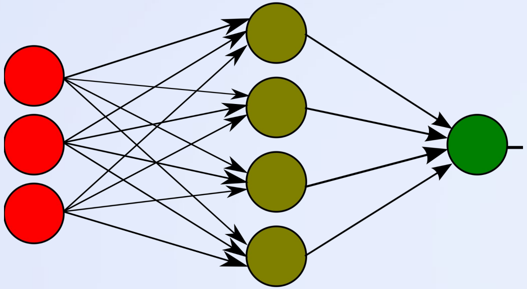



imgflip.com
Everything in the world is a graph
CHANGE MY MIND
AI-TRENTO JOURNAL CLUB

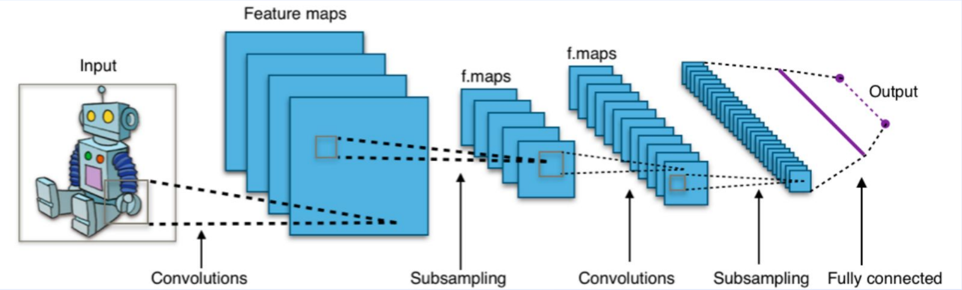

album
lasses
hanging on
stand
oden


Gus
PROPN

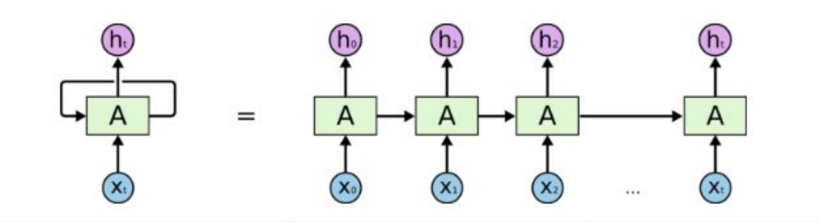
1. Why GNNs



Fully Connected Network



Convolutional Network



Recurrent Network

Transformers are Graph Neural Networks

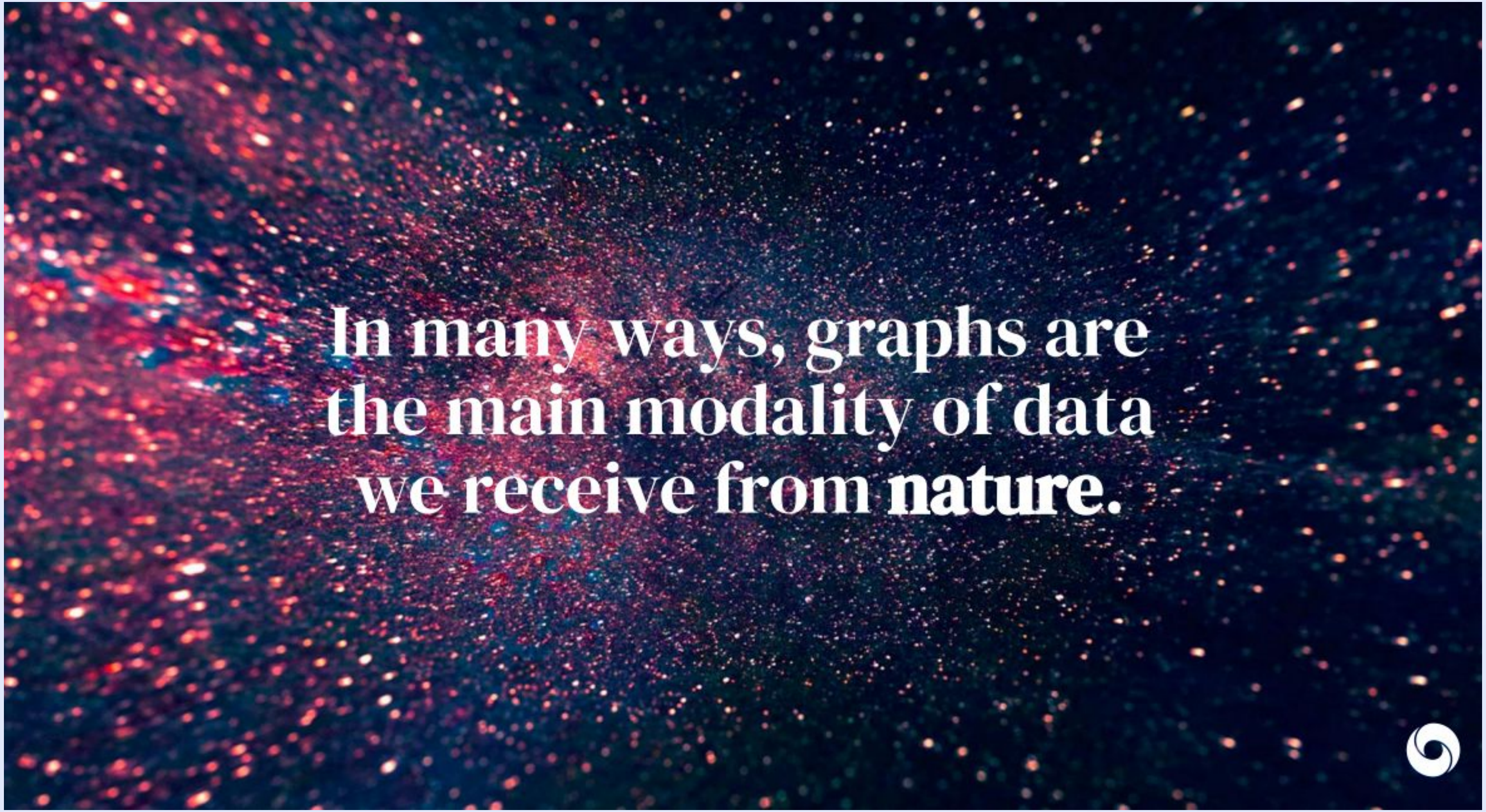
Exploring the connection between Transformer models such as GPT and BERT for Natural Language Processing, and Graph Neural Networks.

Chaitanya K. Joshi
Last updated on Jun 21, 2021 · 12 min read

Cite Project Project Slides The Gradient Towards Data Science Stanford CS224W Cambridge L45 Probabilistic ML textbook

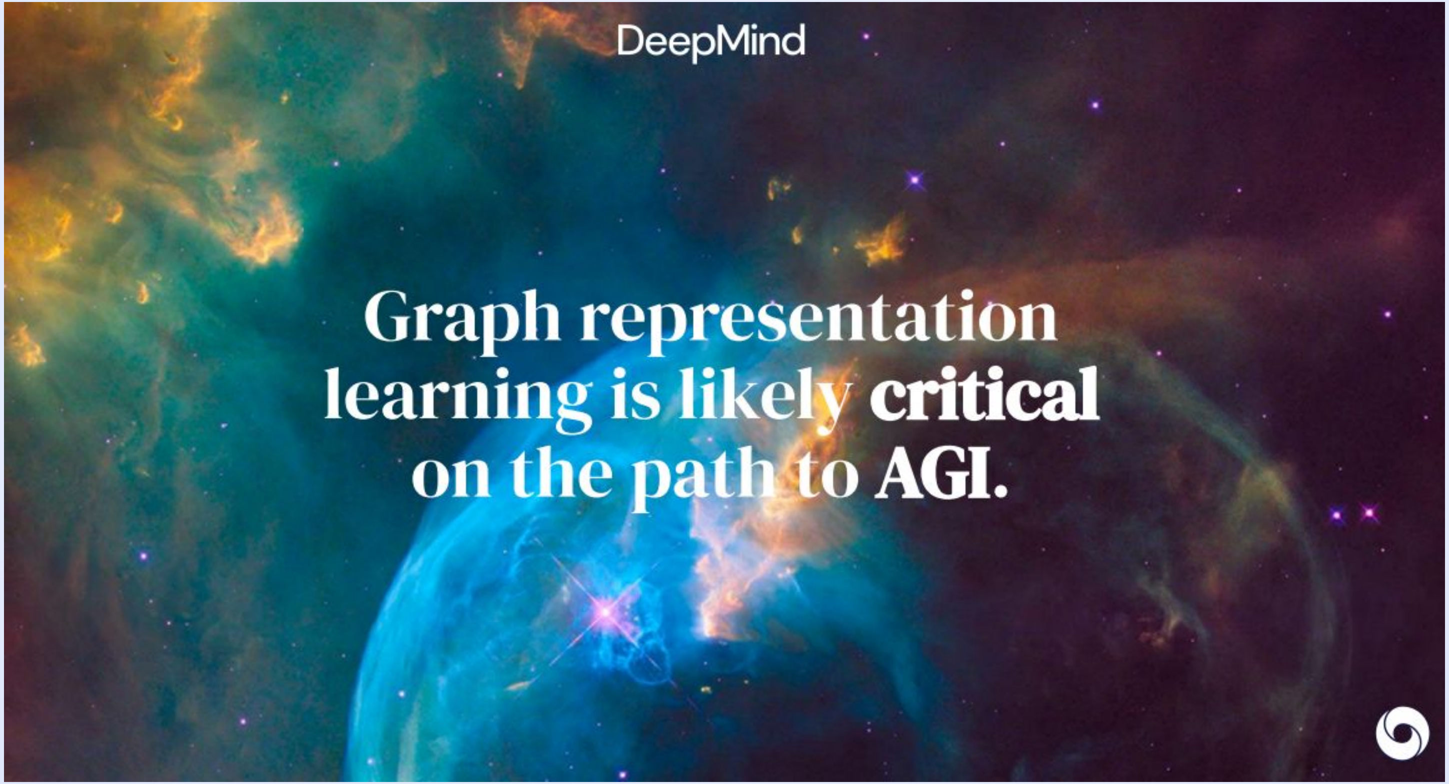
1. Why GNNs

1. Why GNNs



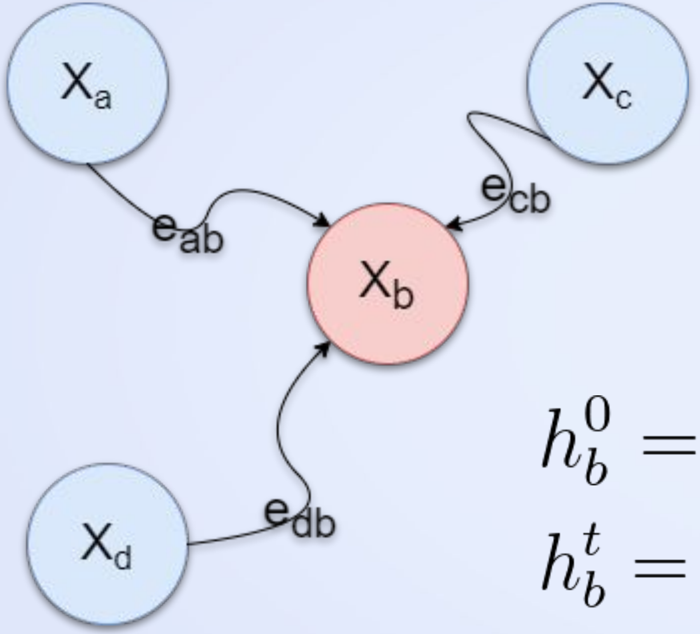
Slide from https://petar-v.com/talks/MLPL-GNN.pdf

1. Why GNNs



Slide from https://petar-v.com/talks/MLPL-GNN.pdf

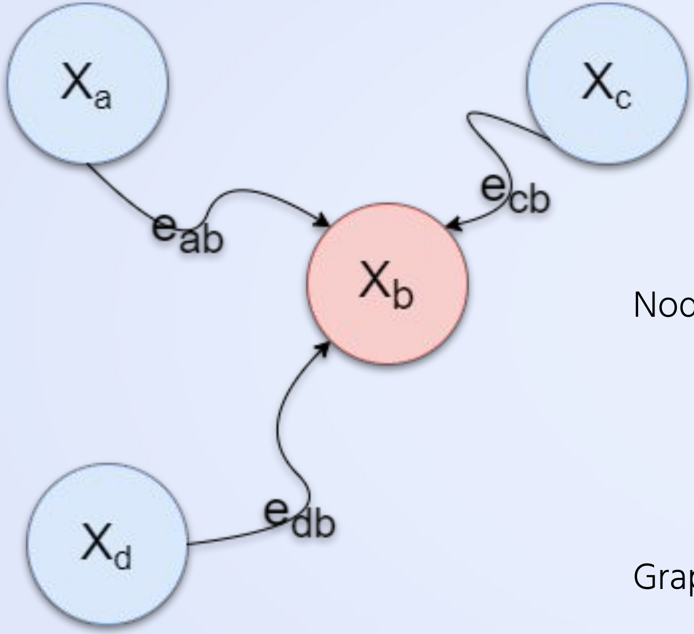
1. Why GNNs



$$h_b^0 = x_b$$

$$h_b^t = U^t(h_b^{t-1}, A^t(\{h_u^{t-1} : \forall u \in N_b\}))$$

1. Why GNNs



Node prediction:

$$\hat{y}_b = MLP(h_b^t)$$

Graph prediction:

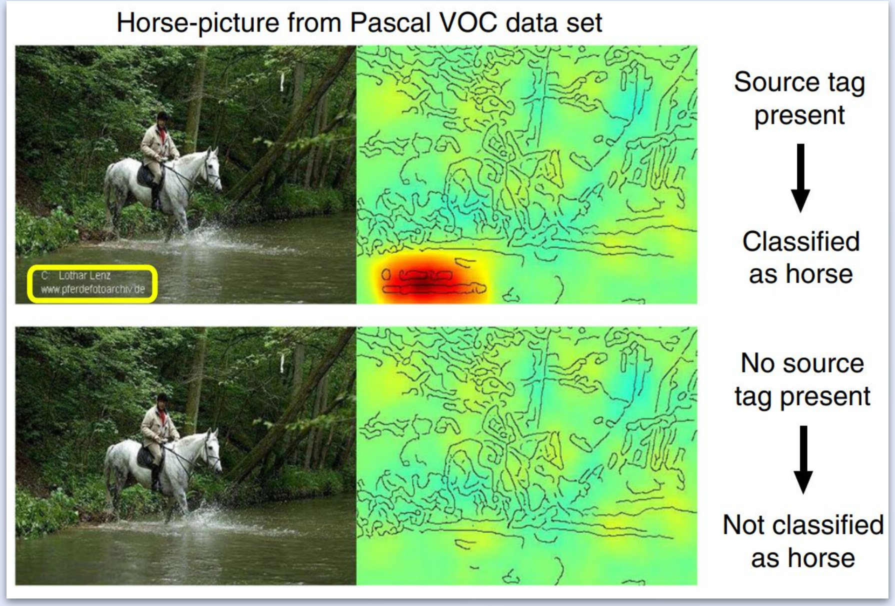
$$\hat{y} = MLP(P(\{h_u^t : u \in G\}))$$

2.

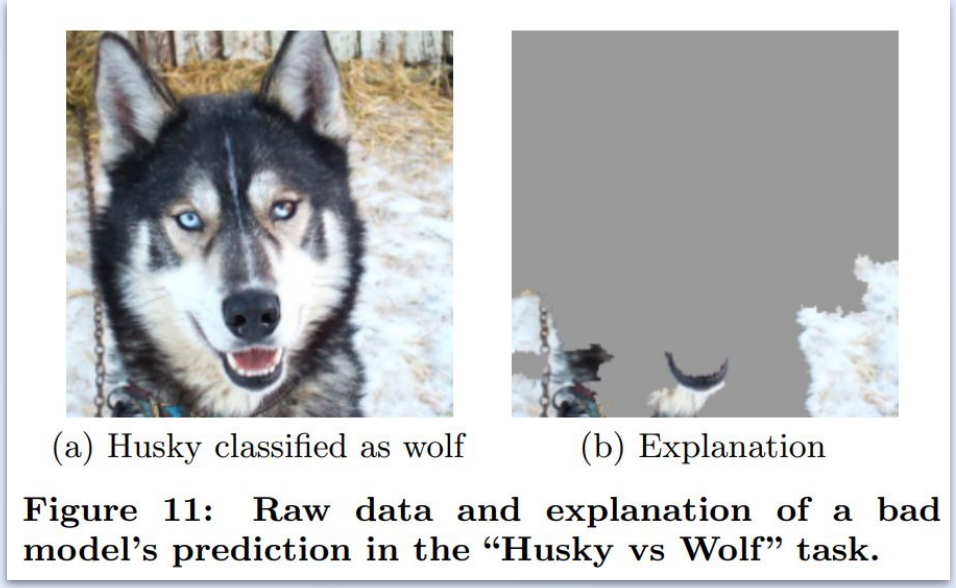
Why XAI?

2. Why XAI

Neural Networks achieve great performances in many tasks.
However, predictions are difficult to be interpreted (**black box**).



Unmasking Clever Hans predictors and assessing what machines really learn. S. Lapuschkin et al., 2019



"Why Should I Trust You?" Explaining the Predictions of Any Classifier . M. T. Ribeiro et al., 2016

2. Why XAI

Neural Networks achieve great performances in many tasks.
However, predictions are difficult to be interpreted (**black box**).

Ad-hoc methods are required to shed light over predictions:

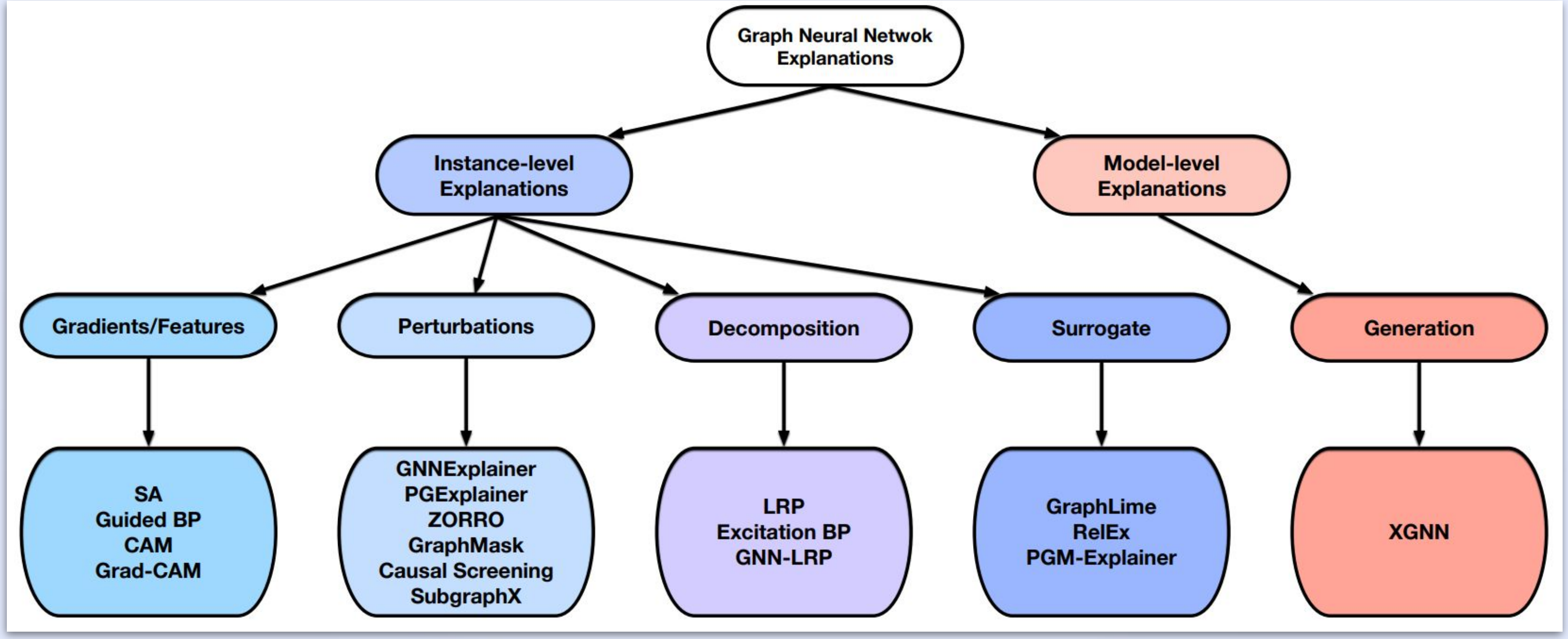
1. **CAM:** Is Object Localization for Free? - Weakly-Supervised Learning With Convolutional Neural Networks. M. Oquab et al., CVPR, 2015
2. **LIME:** “Why Should I Trust You?” Explaining the Predictions of Any Classifier . M. T. Ribeiro et al., ACM SIGKDD, 2016
3. **Integrated Gradients:** Axiomatic Attribution for Deep Networks. M. Sundararajan, ICML, 2017
4. …

3.

XAI for GNNs

3. XAI for GNNs

Also GNNs are black box.
As for non-graph architectures, methods have been proposed to shed light over predictions:

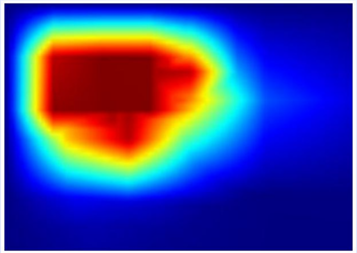


Explainability in Graph Neural Networks: A Taxonomic Survey. H. Yuan et al., 2022

3. XAI for GNNs

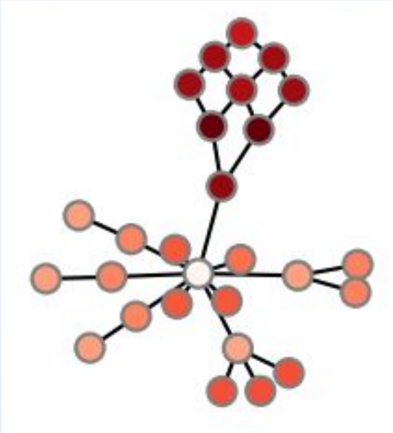
Also GNNs are black box.
As for non-graph architectures, methods have been proposed to shed light over predictions

IMAGES

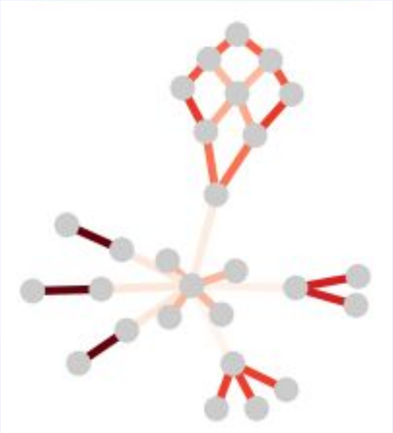


GRAPHS

Node attribution

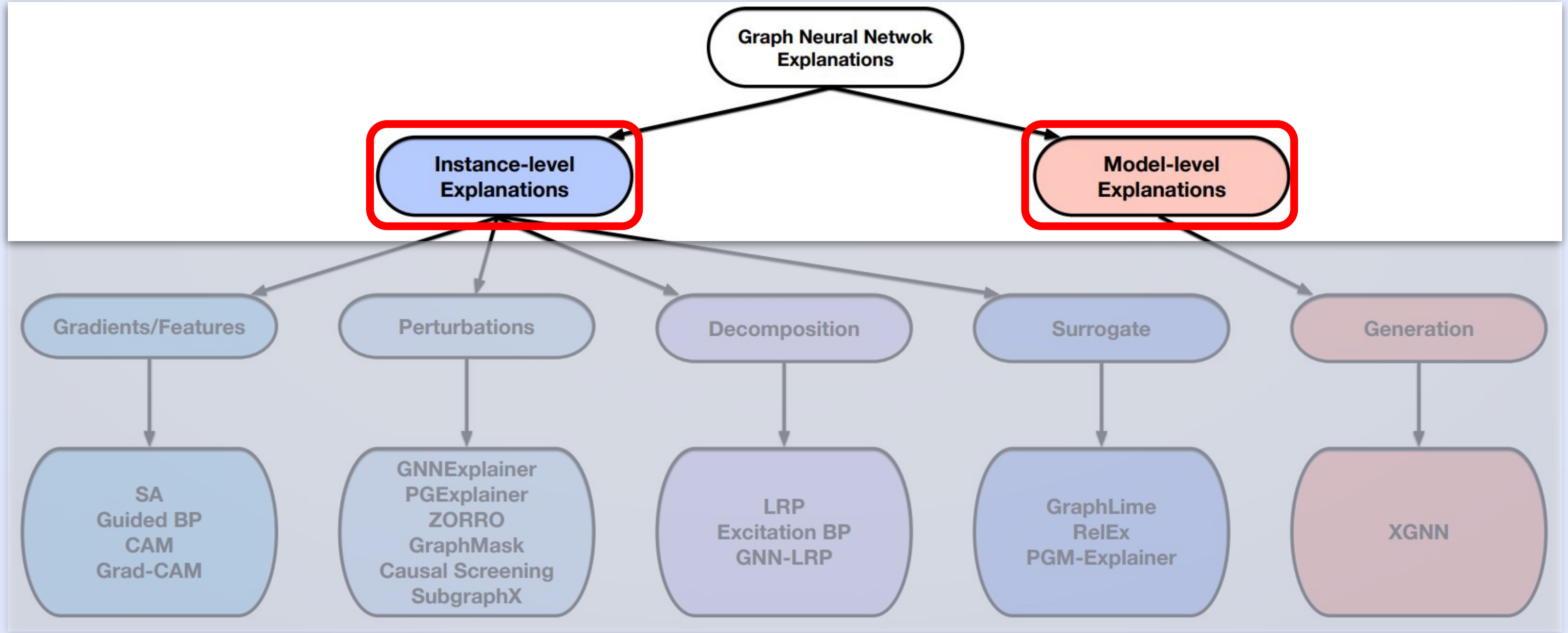


Edge attribution



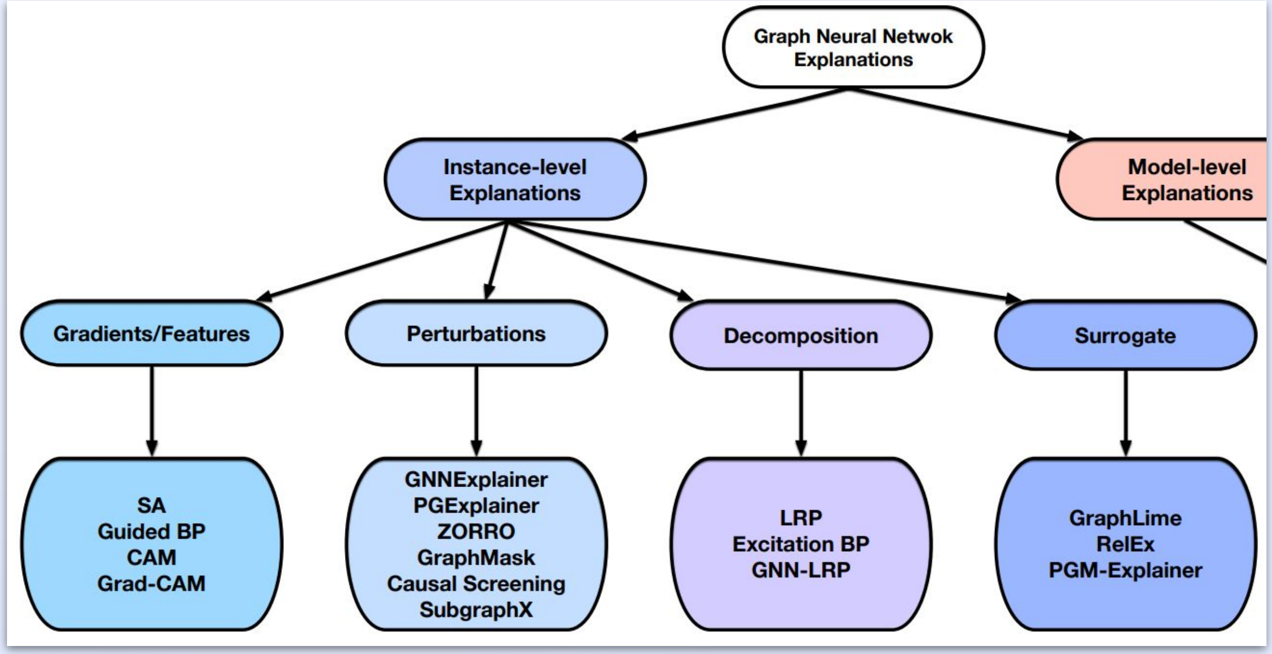
3. XAI for GNNs

Also GNNs are black box.
As for non-graph architectures, methods have been proposed to shed light over predictions:



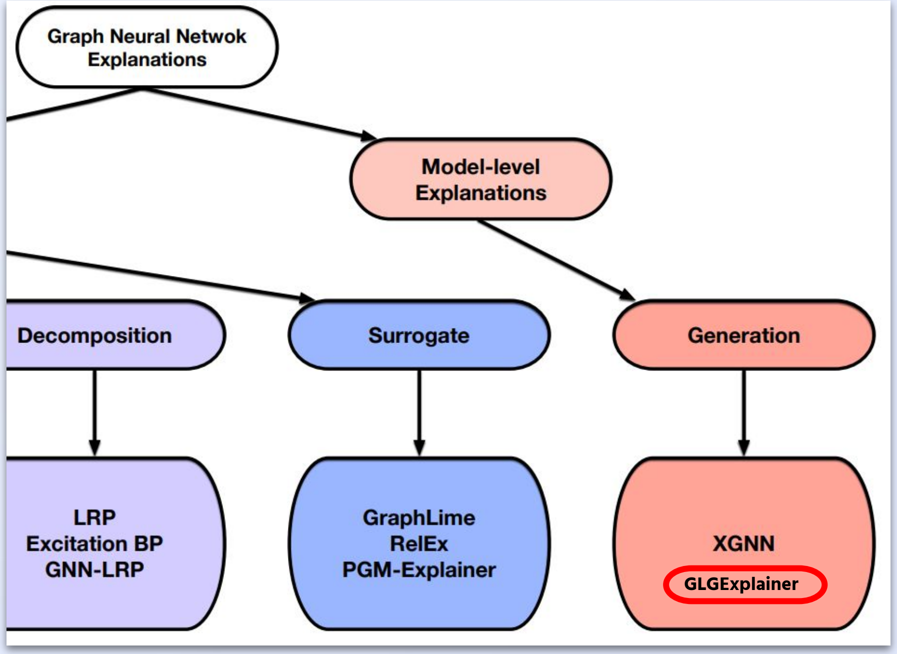
3. XAI for GNNs

Local (or Instance-level) **Explainers** highlight the input features most relevant for the prediction of the model to explain



3. XAI for GNNs

Global (or Model-level) **Explainers** capture the behaviour of the model as a whole, abstracting individual noisy local explanations



Why global explanations?

Global Explainers are seldom studied + mining local explanations is hard:

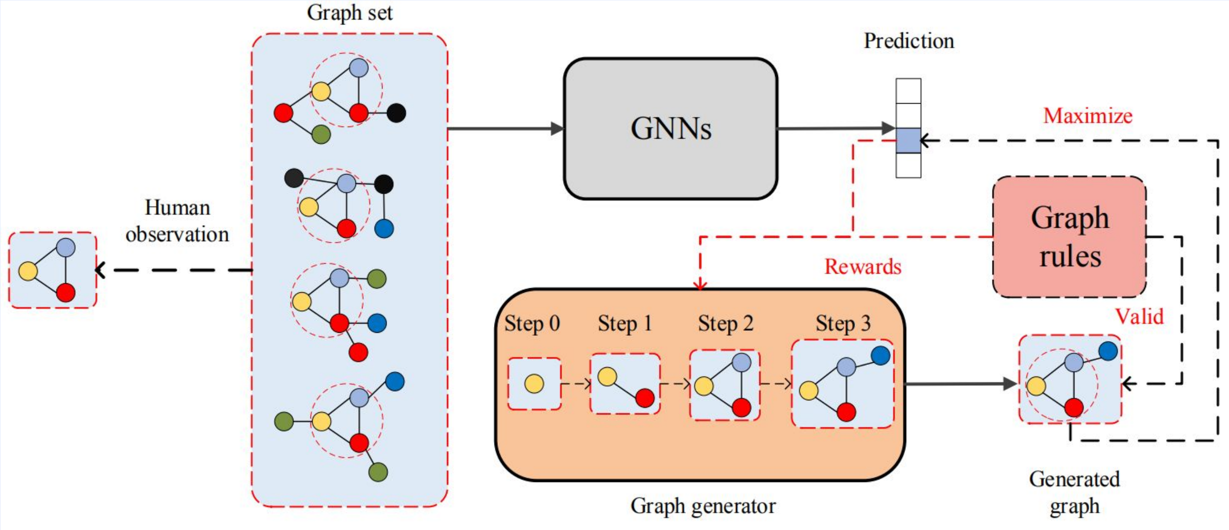
1) 1+ for every input sample
2) Often noisy
3) Difficult quality evaluation[1,2]

A summarized view is amenable to a prompt debugging

1. When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods. L. Faber et al., 2021
2. On Consistency in Graph Neural Network Interpretation. T. Zhao et al., 2022

3. XAI for GNNs

XGNN: Towards Model-Level Explanations of Graph Neural Networks



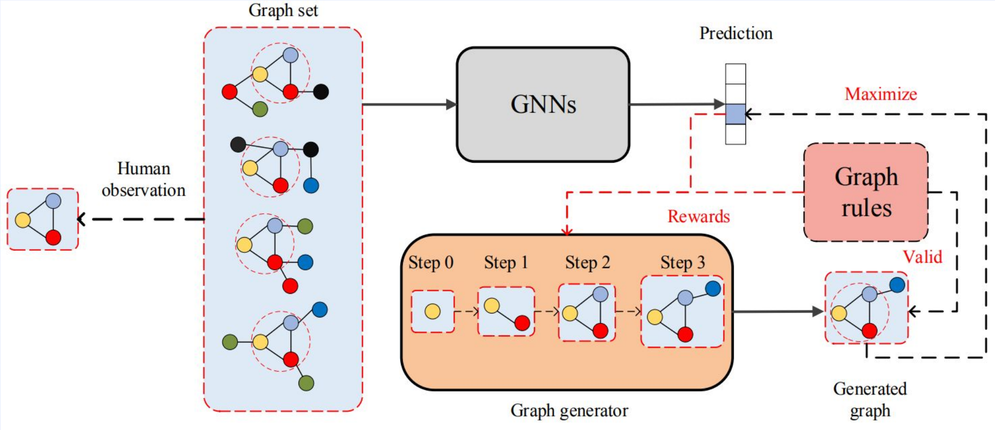
XGNN: Towards Model-Level Explanations of Graph Neural Networks. H. Yuan et al., 2020

3. XAI for GNNs

XGNN: Towards Model-Level Explanations of Graph Neural Networks

Open challenges:

1. Graph rules require strong domain knowledge
2. Explanations not faithful to the data domain



XGNN: Towards Model-Level Explanations of Graph Neural Networks. H. Yuan et al., 2020

4.

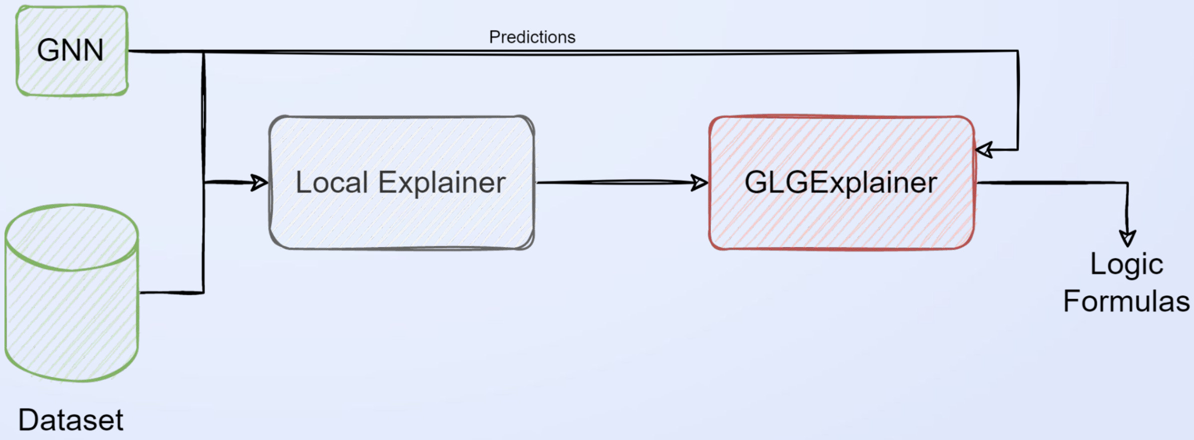
Proposed solution

GLGExplainer (Global Logic-based GNN Explainer)

4. GLGExplainer

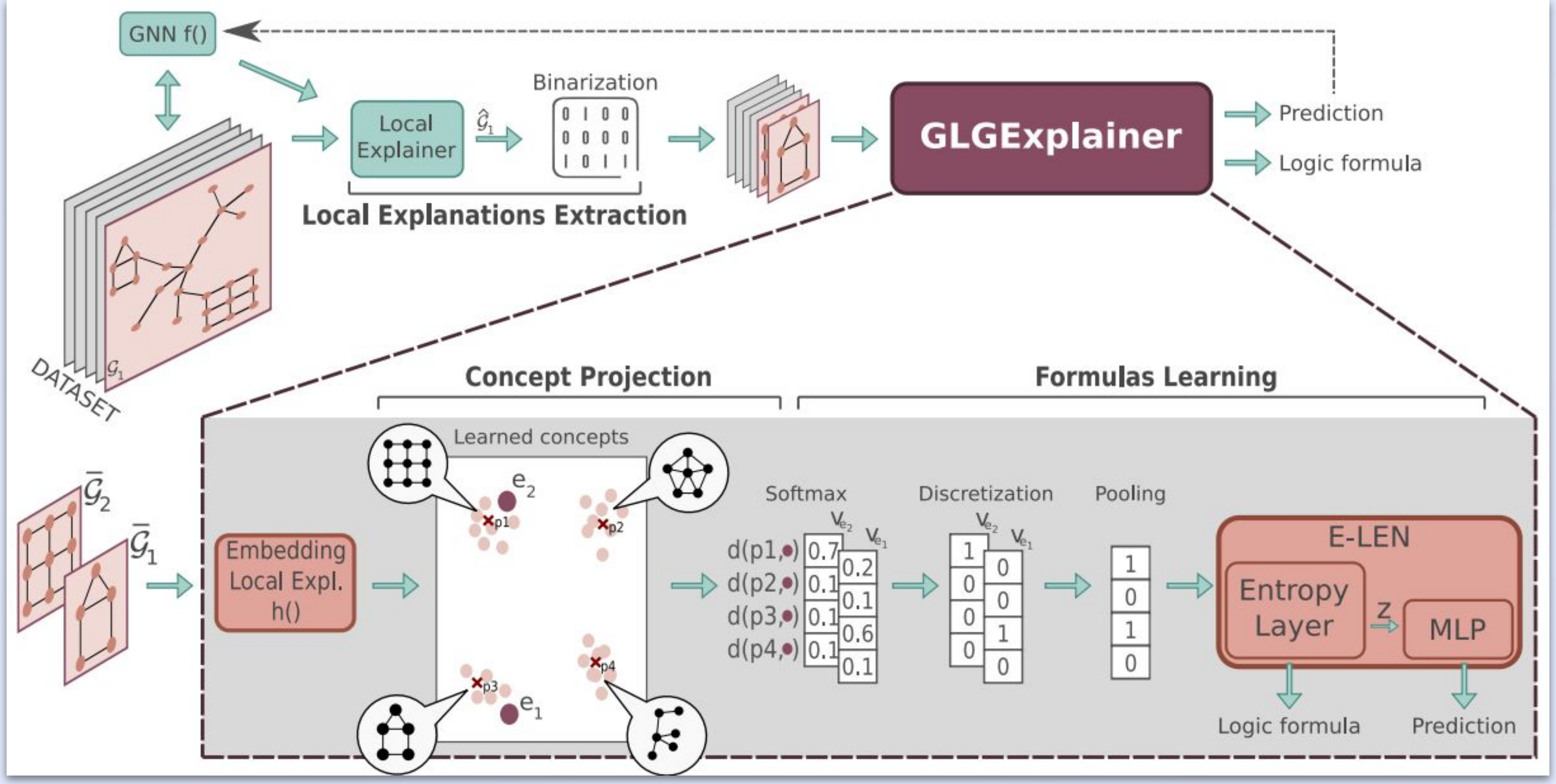
GLGExplainer in short**:**

1. Extract local explanations with a local explainer
2. Run GLGExplainer over those local explanations
3. Inspect the generated logic formulas summarizing the behaviour of the GNN in terms of human-understandable concepts



Global Explainability of GNNs via Logic Combination of Learned Concepts. S. Azzolin et al., 2022

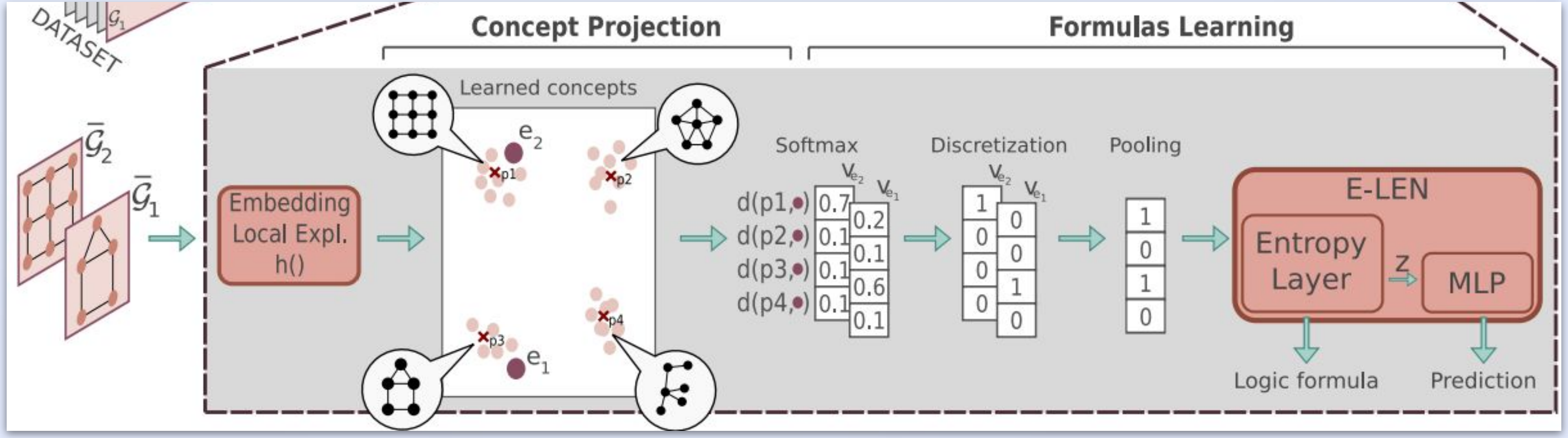
4. GLGExplainer



4. GLGExplainer

Formulas Learning:

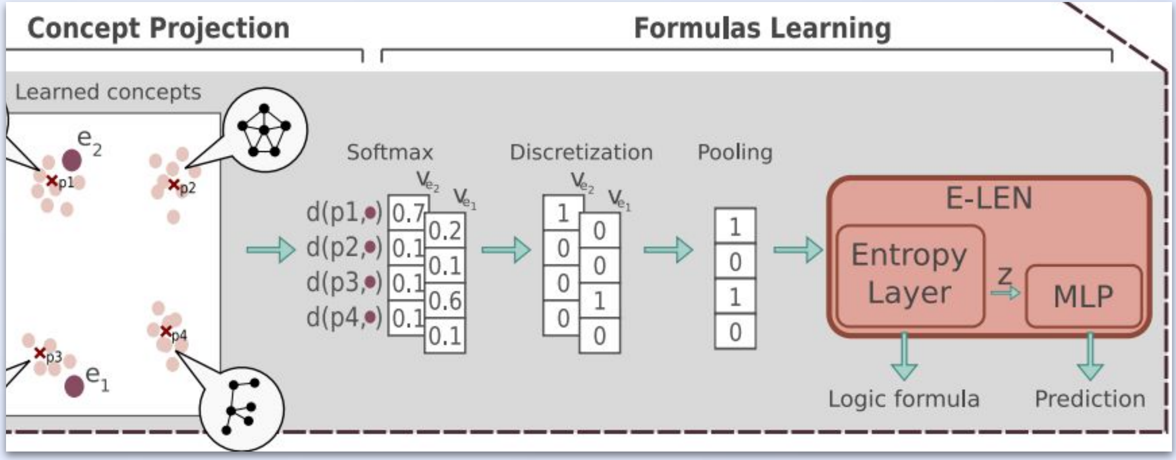
1. **Discretize** concept vectors
   a. promotes discreteness of formulas
   b. promotes formulas-MLP alignment
2. **Pooling** of concept activations of the same input sample
3. Feed the **E-LEN** with the pooled concept vector



4. GLGExplainer

E-LEN (Entropy-based Logic Explained Network):

1. Fully-connected layer *with steroids*
2. Applies entropy regularization for concept selection
3. Builds a Truth Table T for each output class that will be used to extract the final formulas



Entropy-based Logic Explanations of Neural Networks. P. Barbiero et al., 2022

4. GLGExplainer

GLGExplainer is **trained end-to-end** with, as losses:

1. CELoss between E-LEN predictions and GNN predictions (surrogate loss)
2. Distance loss to push every prototype to be close to at least one local explanation
3. Distance loss to push every local explanation to be close to at least one prototype

$$L_{R1} = \frac{1}{m} \sum_{j=1}^{m} \min_{\bar{\mathcal{G}} \in D} \|p_j - h(\bar{\mathcal{G}})\|^2$$

$$L_{R2} = \frac{1}{|D|} \sum_{\bar{\mathcal{G}} \in D} \min_{j \in [1,m]} \|p_j - h(\bar{\mathcal{G}})\|^2$$

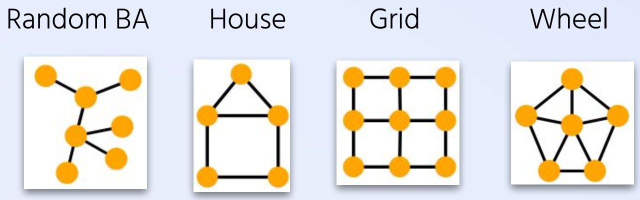
–> No supervision on the concepts, which emerge as prototypical representations of local explanations

Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. O. Li et al., 2018
This looks like that: Deep learning for interpretable image recognition. C. Chen et al., 2019

5. Results

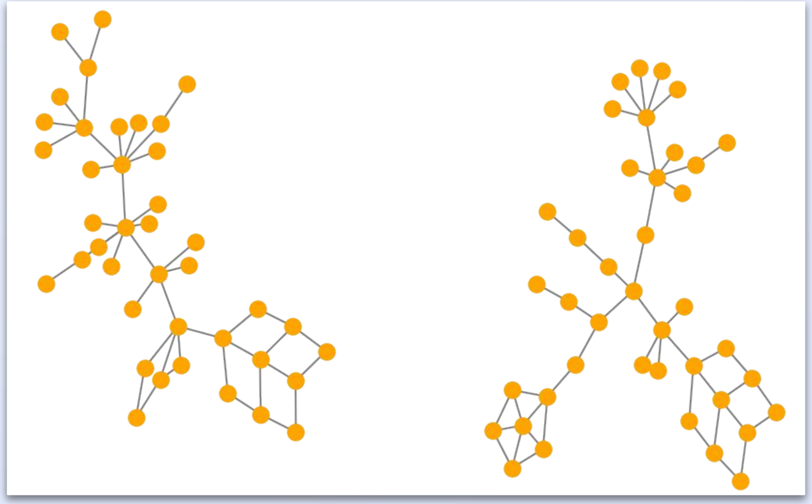
5. Results

1. BAMultiShapes Dataset



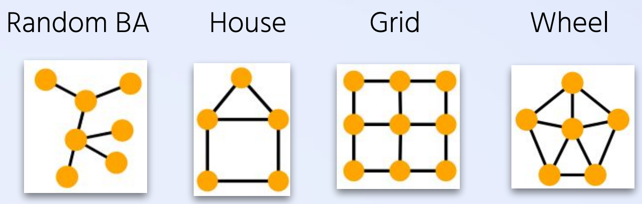
Class 1

- BA + House + Grid
- BA + House + Wheel
- BA + Grid + Wheel



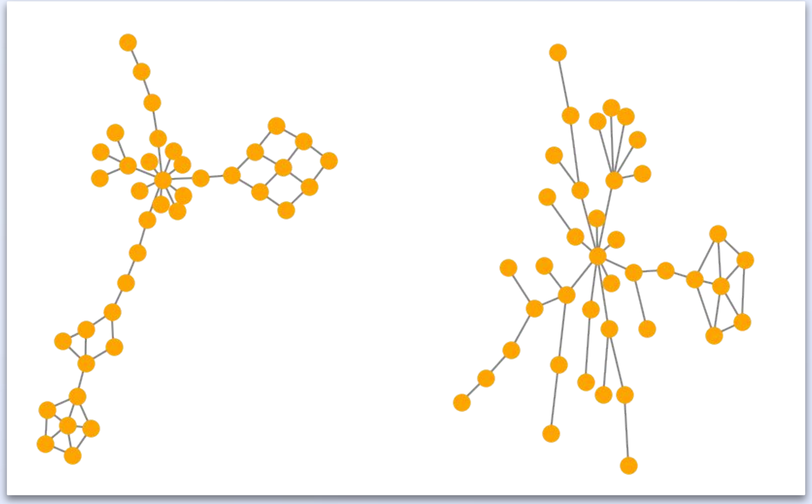
5. Results

1. BAMultiShapes Dataset



Class 0

- BA + House
- BA + Wheel
- BA + Grid
- BA + Grid + House + Wheel
- BA



5. Results

2. GNN to explain

- 3-layers GCN (20-20-20) with mean pooling
- Single FC layer for graph predictions

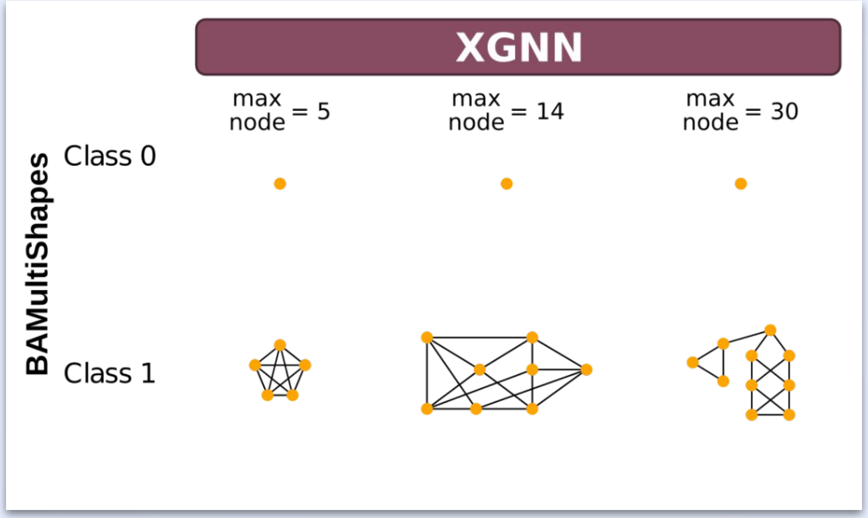
Split	BAMultiShapes
Train	0.94
Val	0.94
Test	0.99

	Class 0					Class 1		
Motifs	∅	H	G	W	All	H + G	H + W	G + W
Accuracy (%)	1.0	1.0	0.85	1.0	0.0	1.0	0.98	1.0

Semi-Supervised Classification with Graph Convolutional Networks. T. Kipf et al. 2022

5. Results

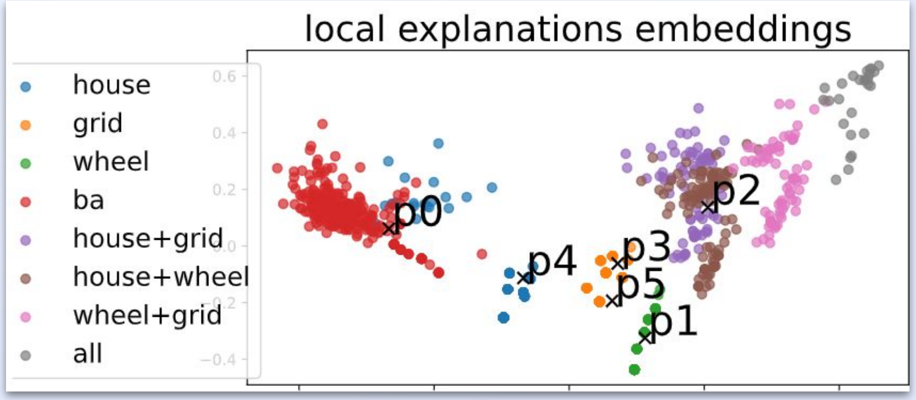
3. XGNN



5. Results

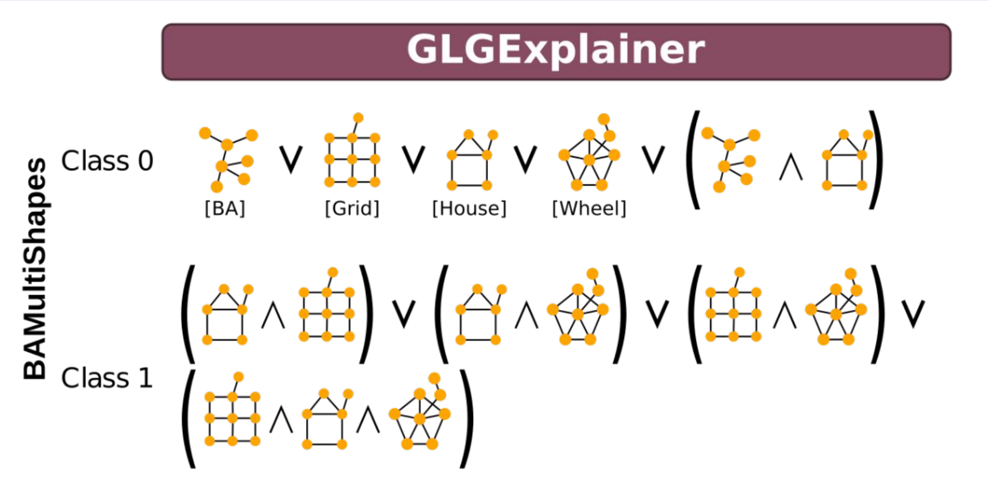
4. GLGExplainer

Dataset			Raw Formulas	Fidelity
BAMultiShapes	Class_0	$\iff$	$P_0 \vee P_3 \vee P_1 \vee P_4 \vee P_5$	0.98
	Class_1	$\iff$	$(P_4 \wedge P_3) \vee (P_5 \wedge P_4) \vee (P_3 \wedge P_1) \vee (P_5 \wedge P_1) \vee (P_4 \wedge P_1) \vee (P_4 \wedge P_2) \vee (P_1 \wedge P_2) \vee (P_3 \wedge P_2) \vee P_2$	



5. Results

4. GLGExplainer



Conclusions

Conclusions

Main contributions:

1. Global Explainer for GNNs which
   a. provides logic formulas
      i. more informative than previous SOTA
   b. faithful to the data domain

2. Unsupervised algorithm for concept discovery

E.O.F.

Global Explainability of GNNs via Logic Combination of Learned Concepts. S. Azzolin et al., 2022. ICLR2023