Explainability in Graph Machine Learning

Megha Khosla (TU Delft),

https://khosla.github.io
m.khosla@tudelft.nl

TU Delft

Outline

Why this tutorial?

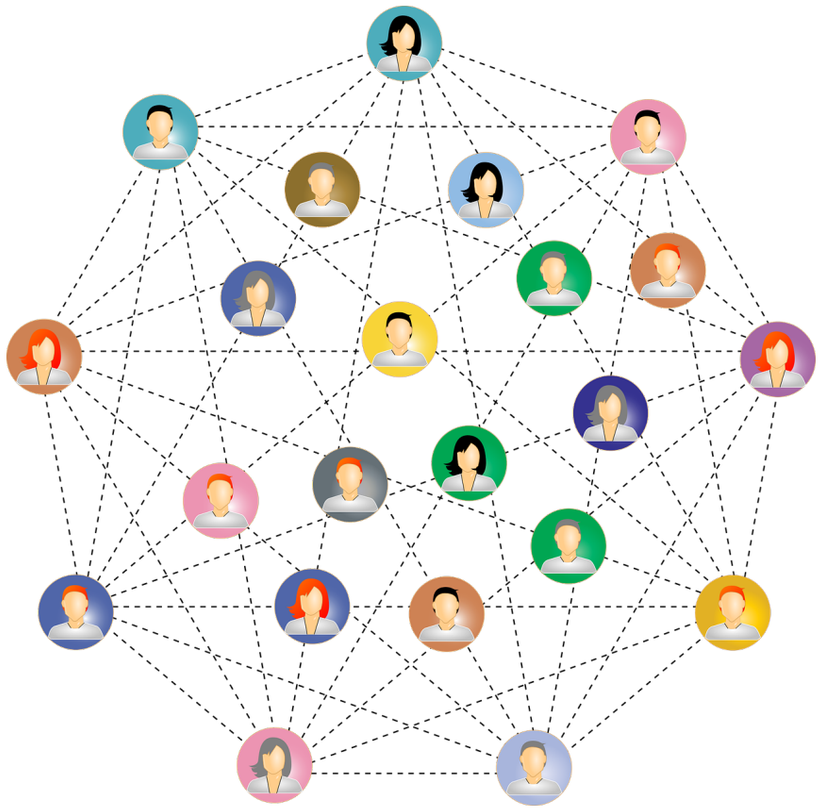
Explanations over graph data

Posthoc explainability in graphs : Instance-wise and Global explanations

Evaluation of post-hoc explanations

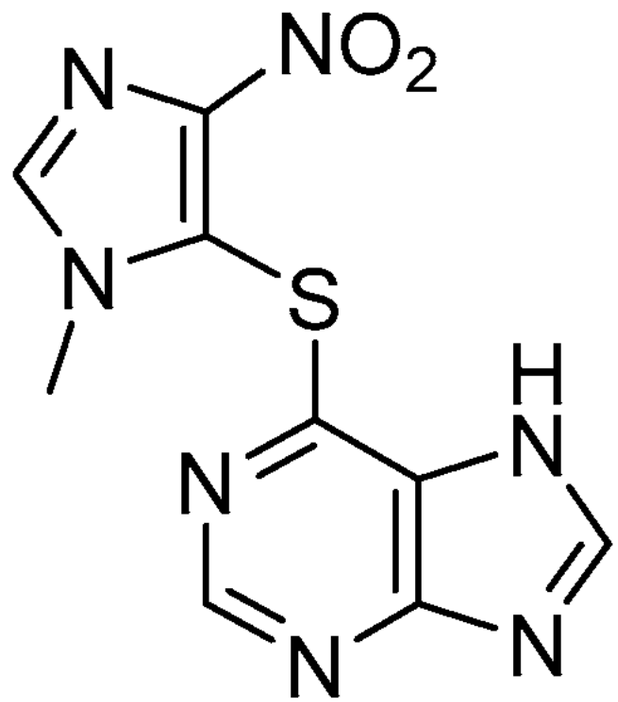
Hands-On Session

Graphs are everywhere



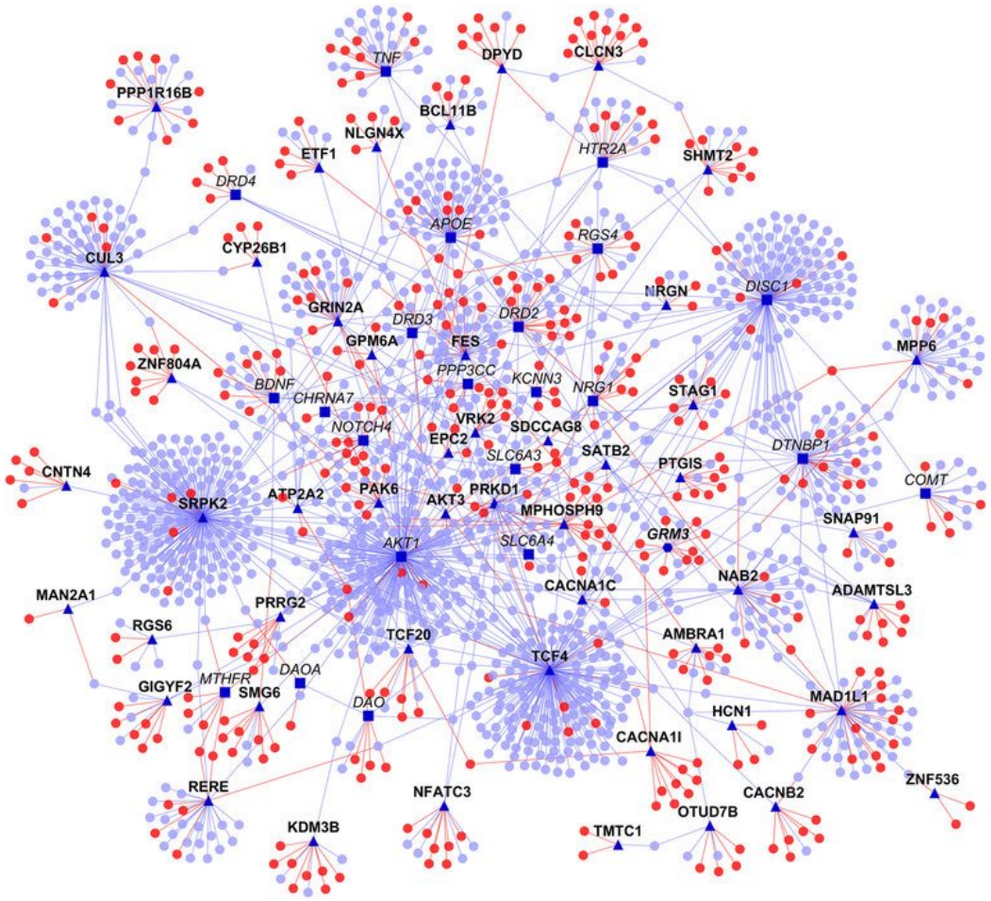
Social Networks

Image Source : Medium



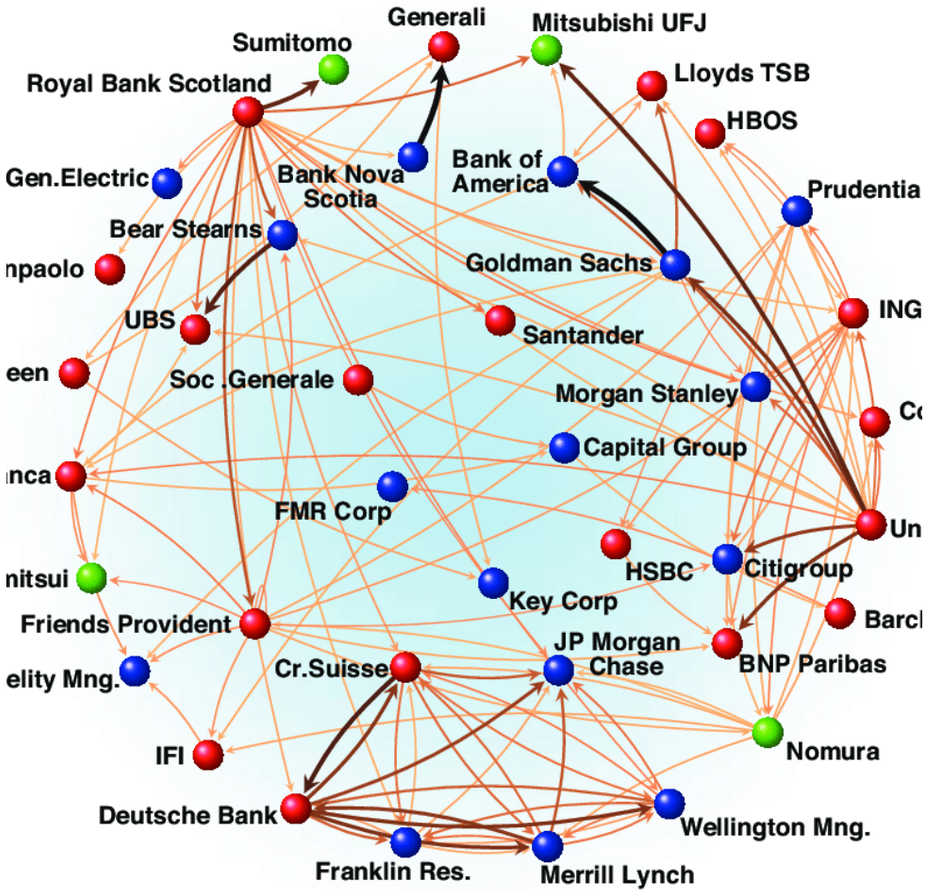
Drug Molecules

Graphs are everywhere



Protein interaction network

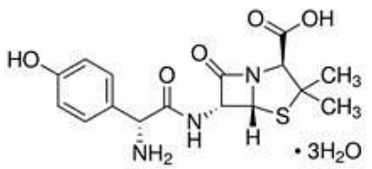
Image Source : wikipedia



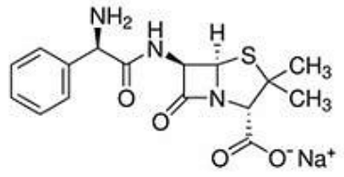
Financial network

Image Source : Schweitzer et al. 2009

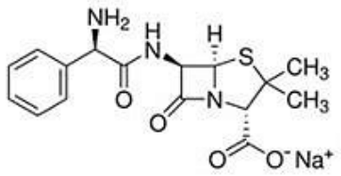
Success of Graph Machine Learning



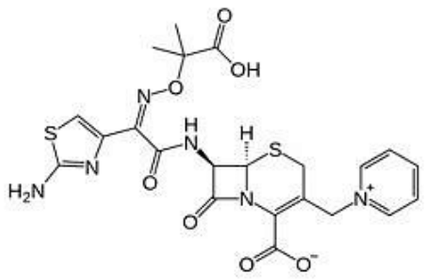
Amoxicillin



Ampicillin



Penicillin G

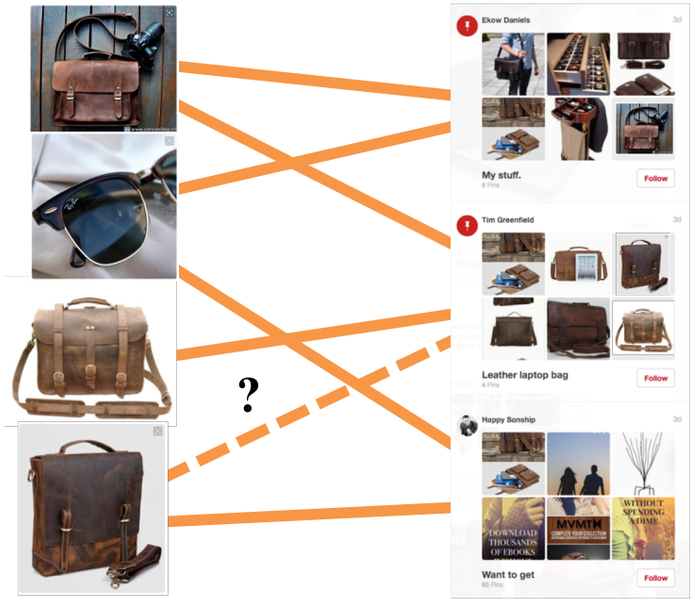


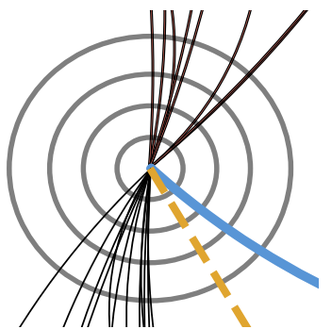
Ceftazidime

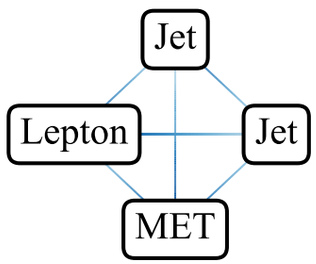
discover **novel antibiotics** (Stokes *et al.*, Cell'20)

Image Source : Coman et al. 2017

power **web-scale recommender systems** (Ying *et al.*, KDD'18; Pal *et al.*, KDD'20)





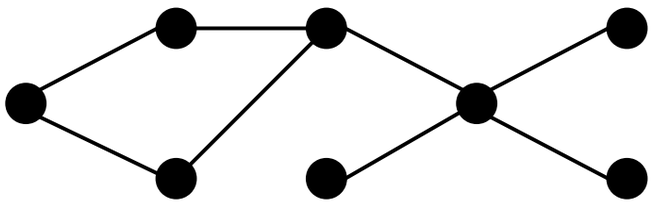


assist **particle physicists** (Shlomi *et al.*, Mach. Learn.: Sci. Technol'21)

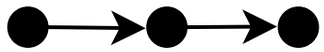
Different kinds of graphs

Some of the graph types

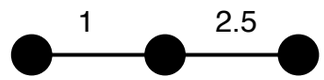
Undirected, unweighted, static



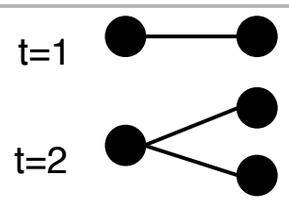
Directed



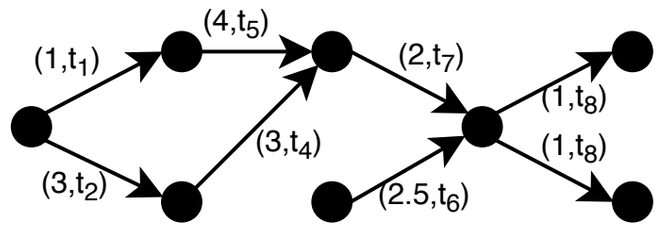
Weighted



Sliced



Directed, weighted, continuous



Feature types:

No features

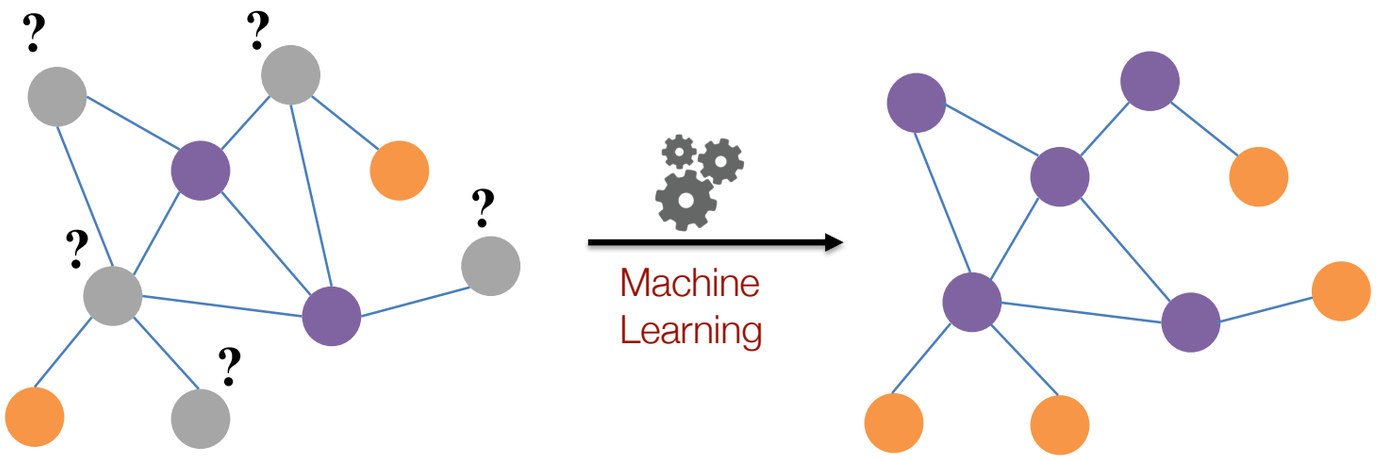
Node features

Edge features

Dense features, e.g. word embeddings

Sparse features

Typical ML Tasks on Graphs



Node classification

Link prediction

Graph classification

Community detection

Graph Machine Learning (GraphML)

Shallow Network Embedding Methods

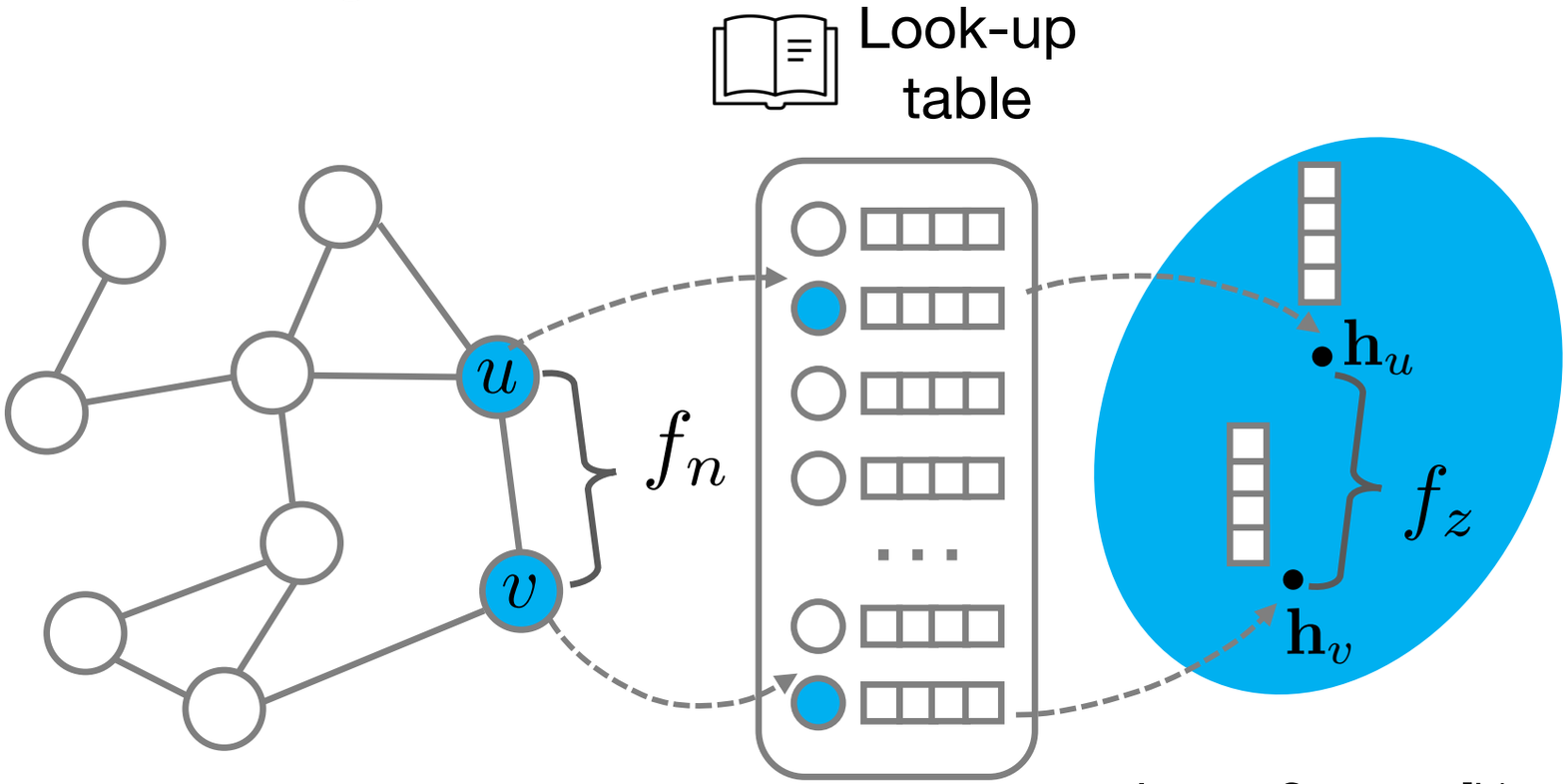


Image Source: [Li et al., 2022]

Graph Machine Learning (GraphML)

Shallow Network Embedding Methods

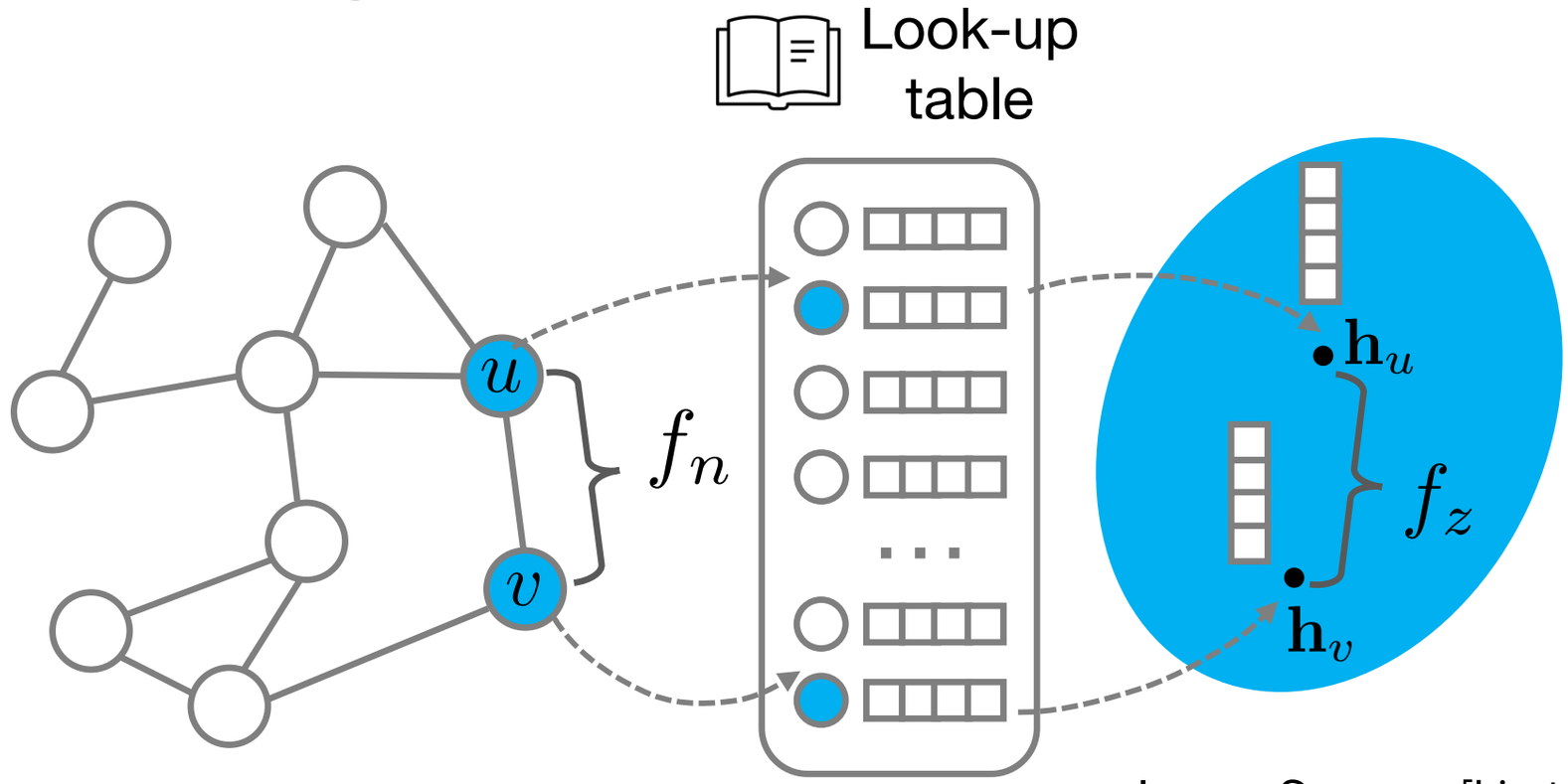


Image Source: [Li et al., 2022]

- Generate a look up table for node representations
- Similar nodes get embedded closer

Examples :

DeepWalk, Node2Vec, NERD, HOPE

Graph Neural Networks (GNNs)

$$\boldsymbol{z}_i^{(\ell)} = \text{AGGREGATE}\left(\left\{\boldsymbol{x}_i^{(\ell-1)}, \left\{\boldsymbol{x}_j^{(\ell-1)} \mid j \in \mathcal{N}_i\right\}\right\}\right)$$

$$\boldsymbol{x}_i^{(\ell)} = \text{TRANSFORM}\left(\boldsymbol{z}_i^{(\ell)}\right)$$

Examples :

GCN, GAT, GIN

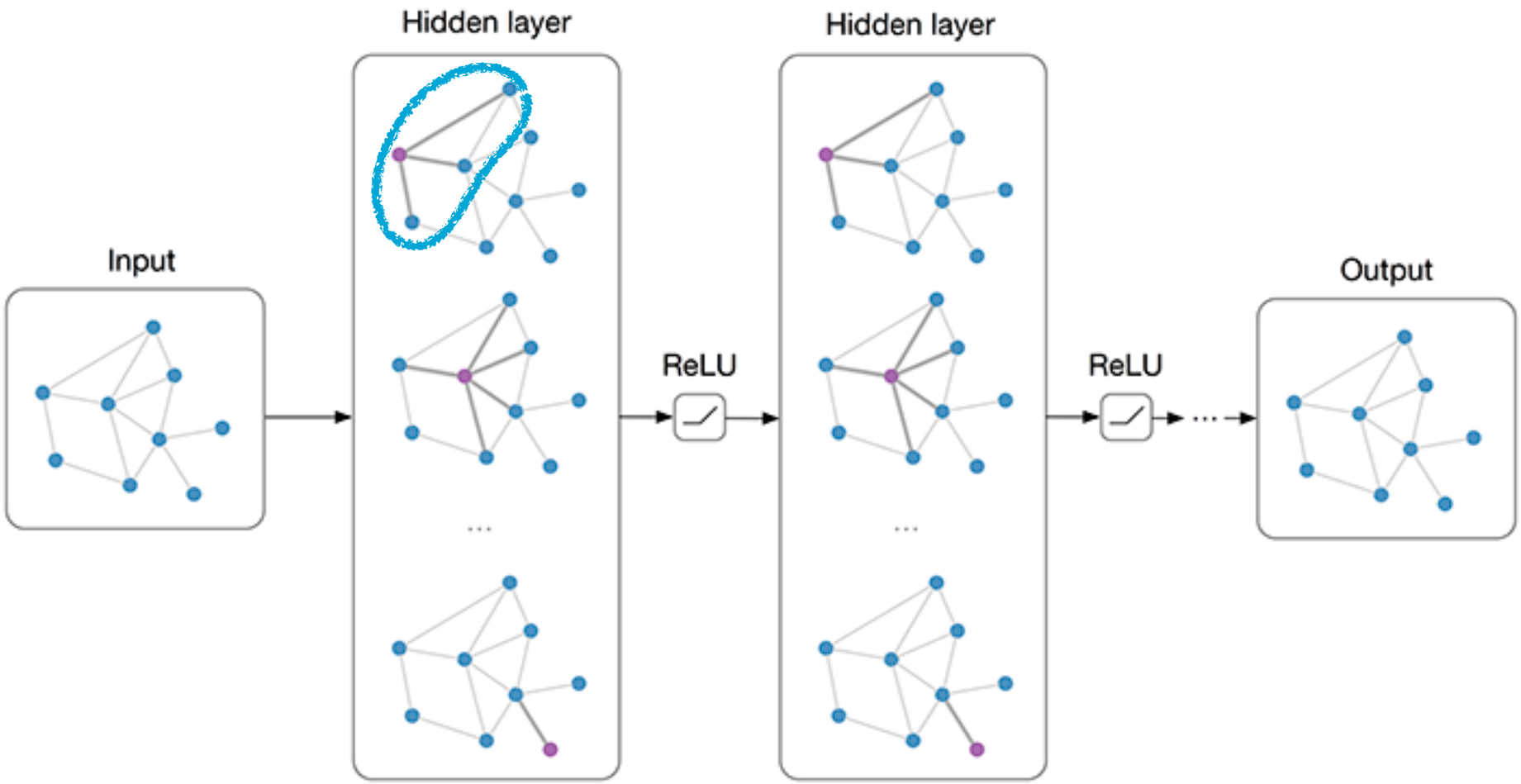
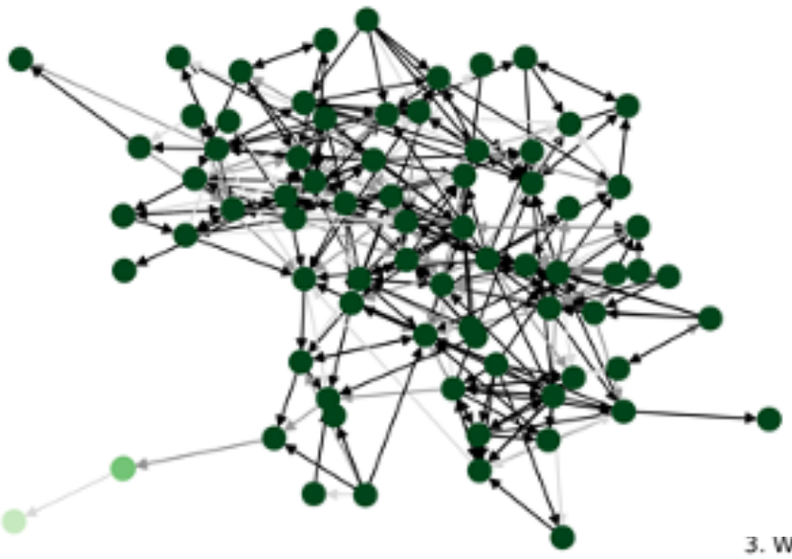


Image Source : https://tkipf.github.io/graph-convolutional-networks/

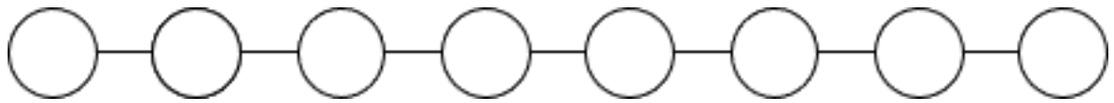
Our Focus : Explainability of GNNs

Why special techniques for Graphs?

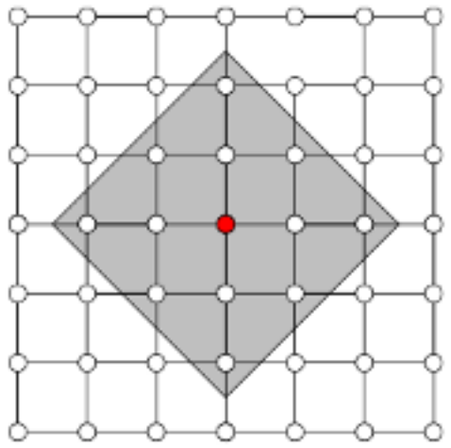


Graphs

vs.



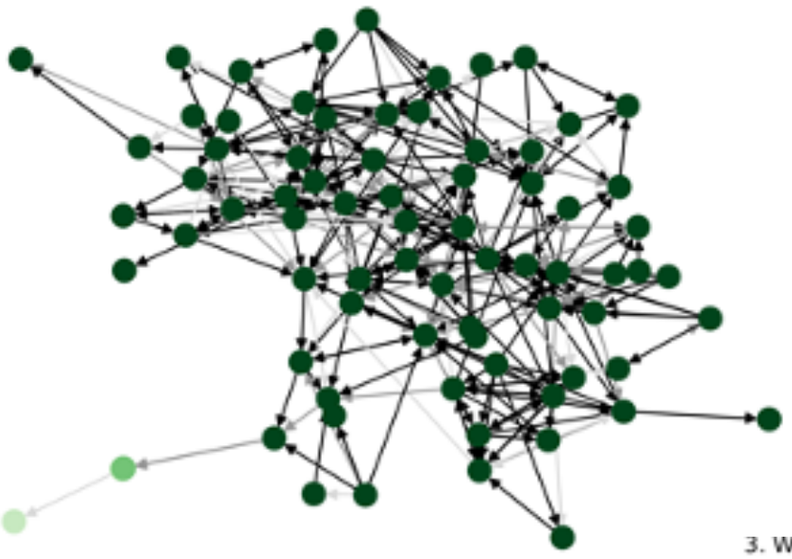
Text



Images

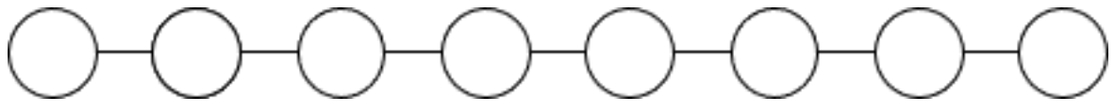
Our Focus : Explainability of GNNs

Why special techniques for Graphs?

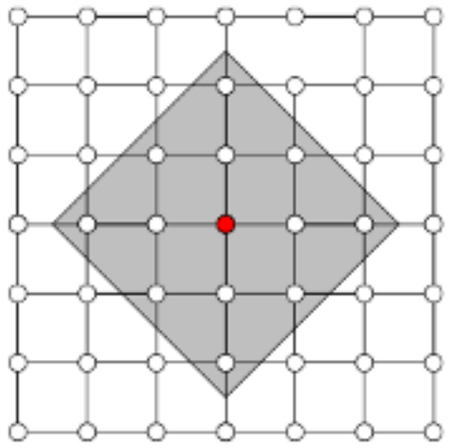


Graphs

vs.



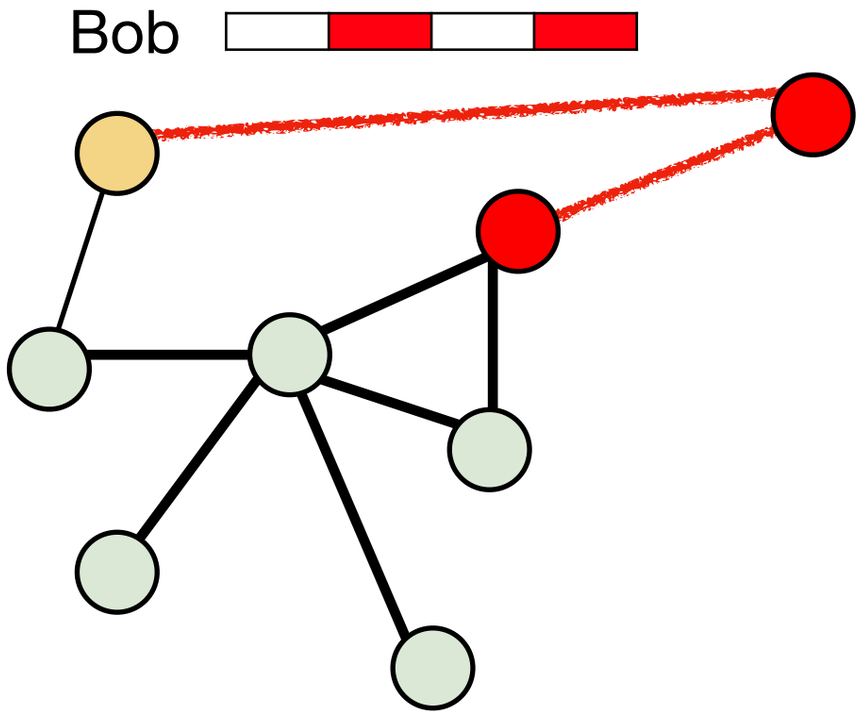
Text



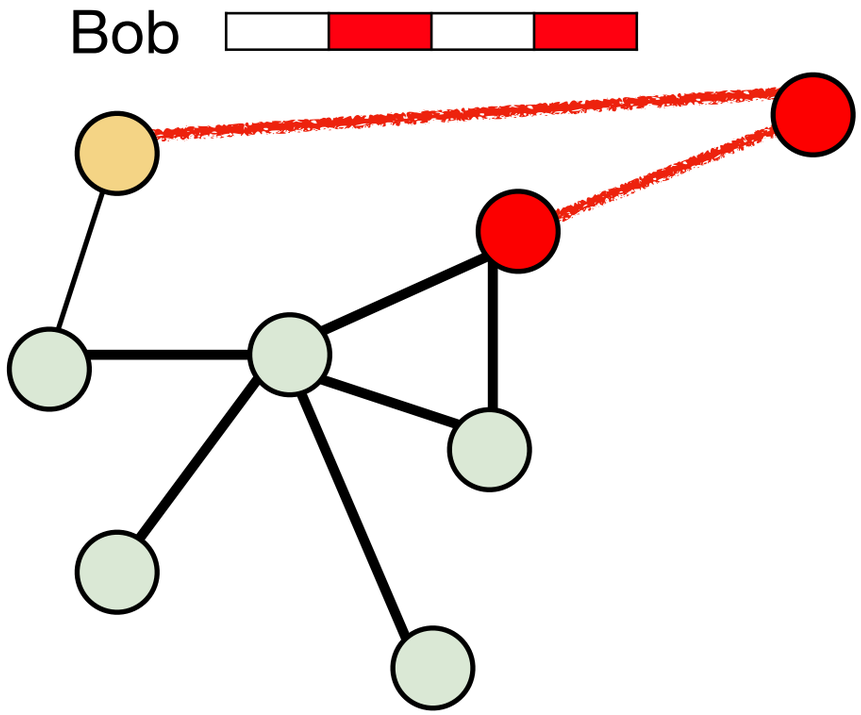
Images

Unlike images and texts, graphs are not grid-like data, which means there is no locality information and each node has different numbers of neighbors. Regions like in image are not even defined.

Challenges for Graphs



Challenges for Graphs



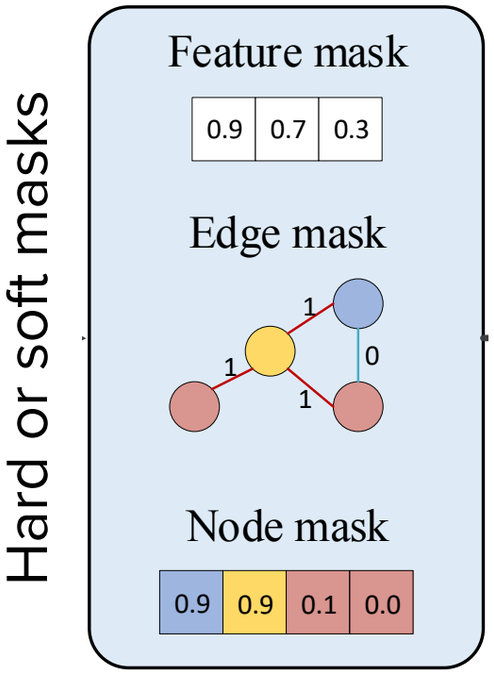
Decision has to be explained not only in terms of features but also graph structure. General explainability methods cannot be trivially applied for graphs.

Explanation for GNNs

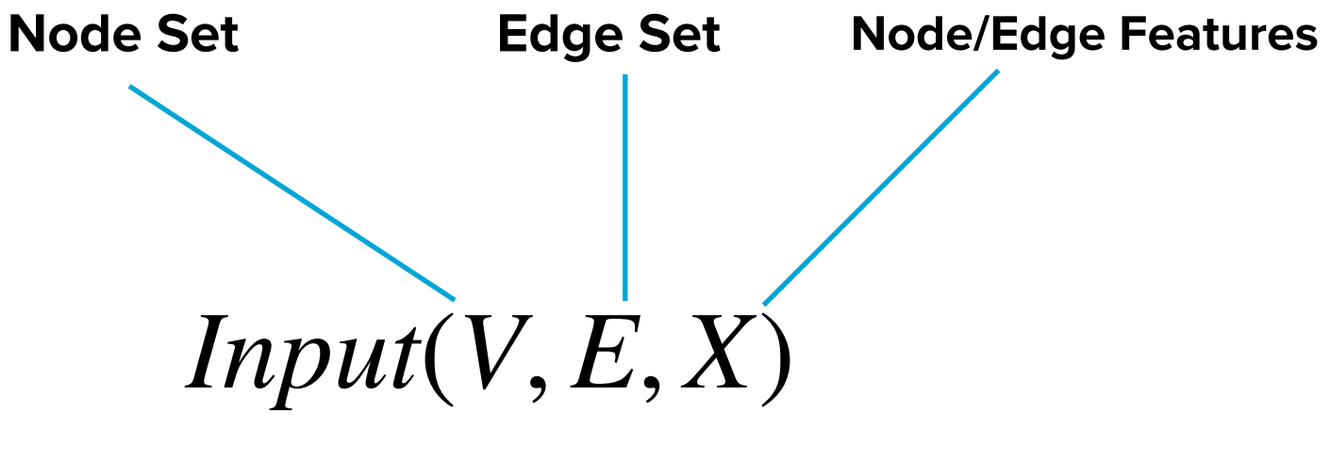
Node Set **Edge Set** **Node/Edge Features**

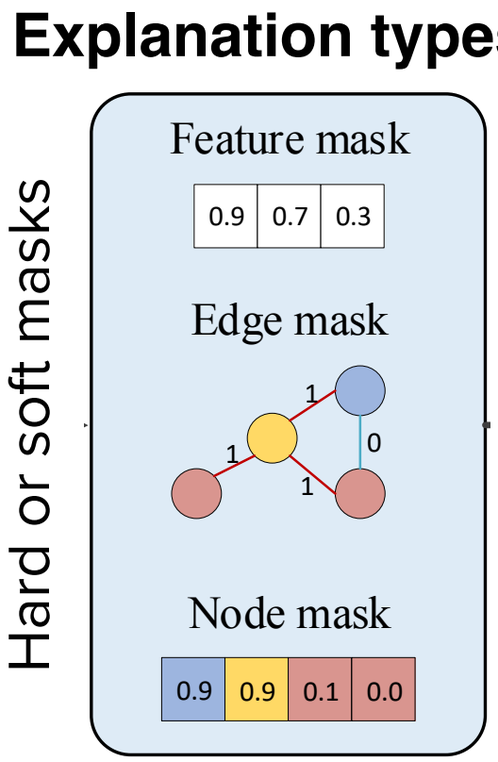
$$Input(V, E, X)$$

Explanation types



Explanation for GNNs





Explanation types:

Feature explanations in terms of most relevant features $X' \subset X$

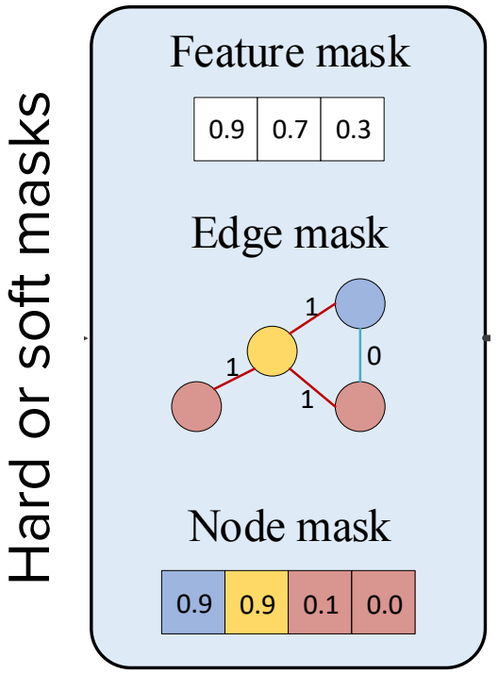
Structure explanations in terms of most relevant nodes $(V' \subset V)$ or edges $(E' \subset E)$

Explanation for GNNs

Node Set **Edge Set** **Node/Edge Features**

$$Input(V, E, X)$$

Explanation types



Explanation types:

Feature explanations in terms of most relevant features $X' \subset X$

Structure explanations in terms of most relevant nodes $(V' \subset V)$ or edges $(E' \subset E)$

We are interested in finding both feature and structure explanations which effectively capture interplay of structure and features in model's decision making.

Computational Graph for GNNs

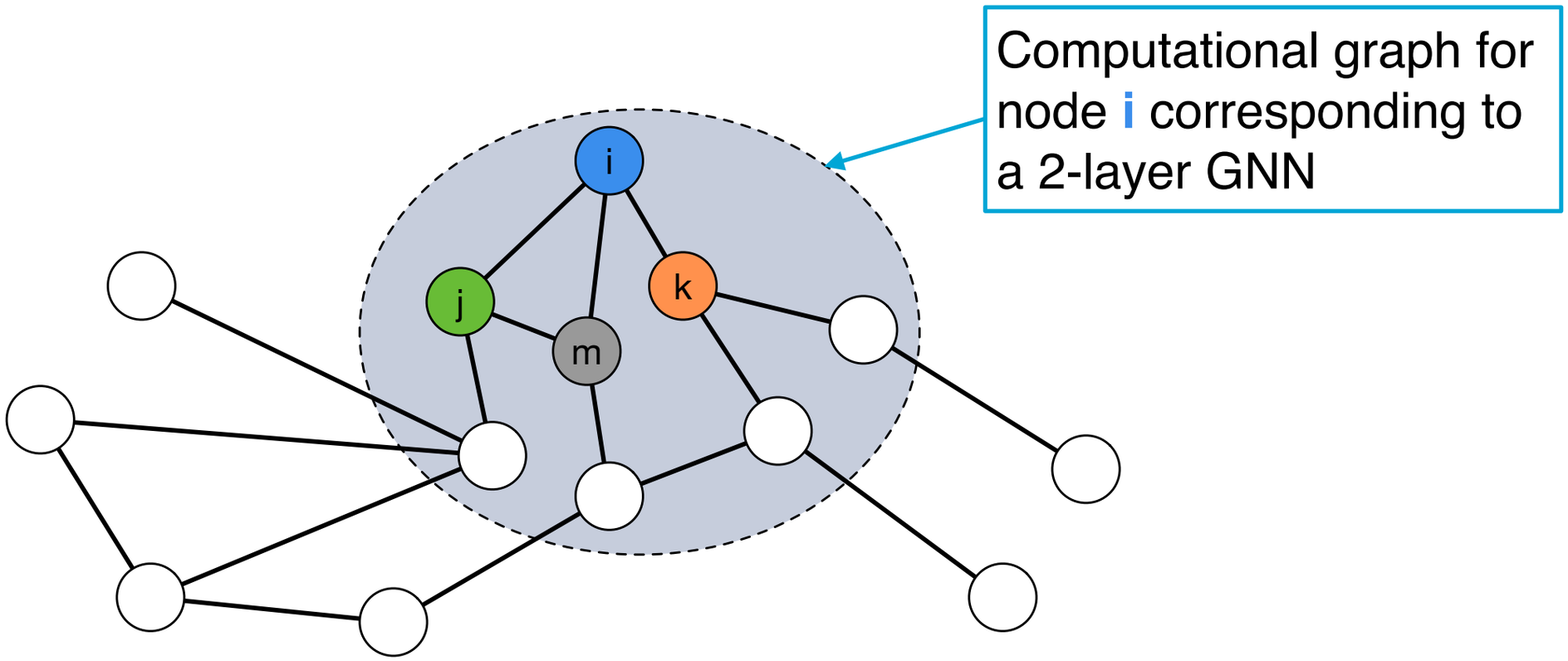


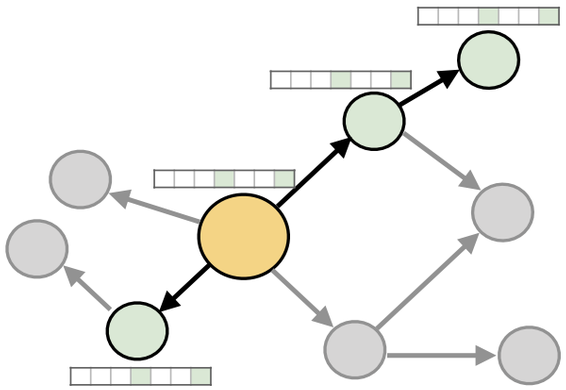
Image Source: [Lin et al., 2021]

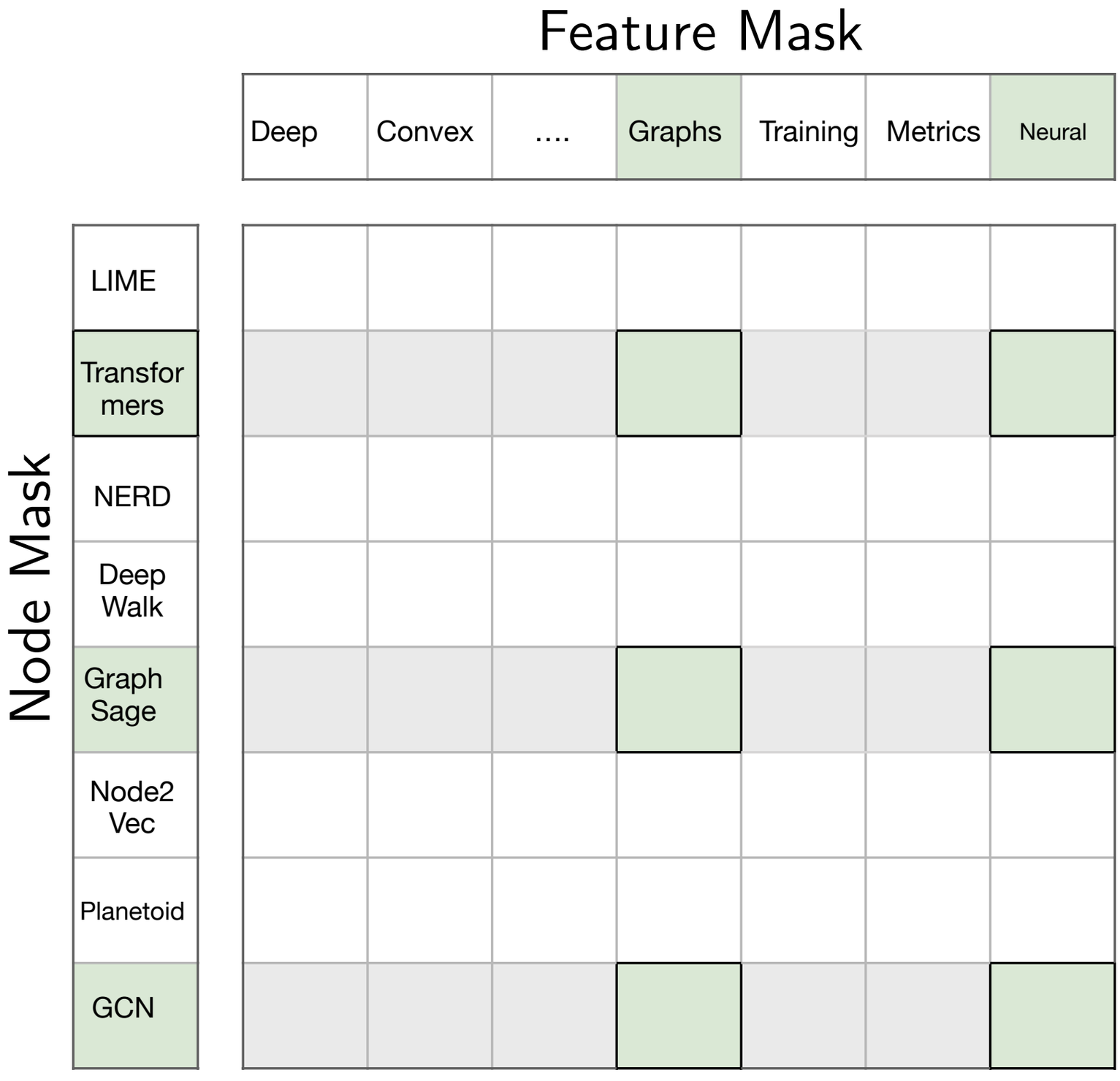
At inference time decision of a GNN on a particular node can be attributed to important nodes/edges and their features in its computational graph.

An example explanation

Computation graph for the node representing the paper *"Graph Attention Network"*

Prediction class label: **"GNN"**





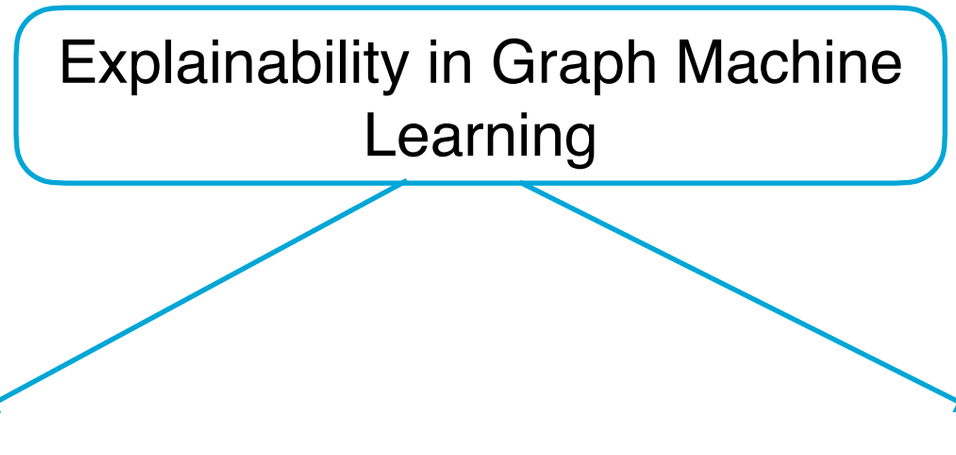
Example Explanation

The learnt model focusses for the prediction **"GNN"**

1) Features (Words in this example)
   - "Graphs"
   - "Neural"
2) Neighbourhood Nodes (Papers in this example):
   - Transformers (Vaswani et al. '17)
   - GraphSage (Hamilton et al. '17)
   - GCN (Kipf & Welling '16)

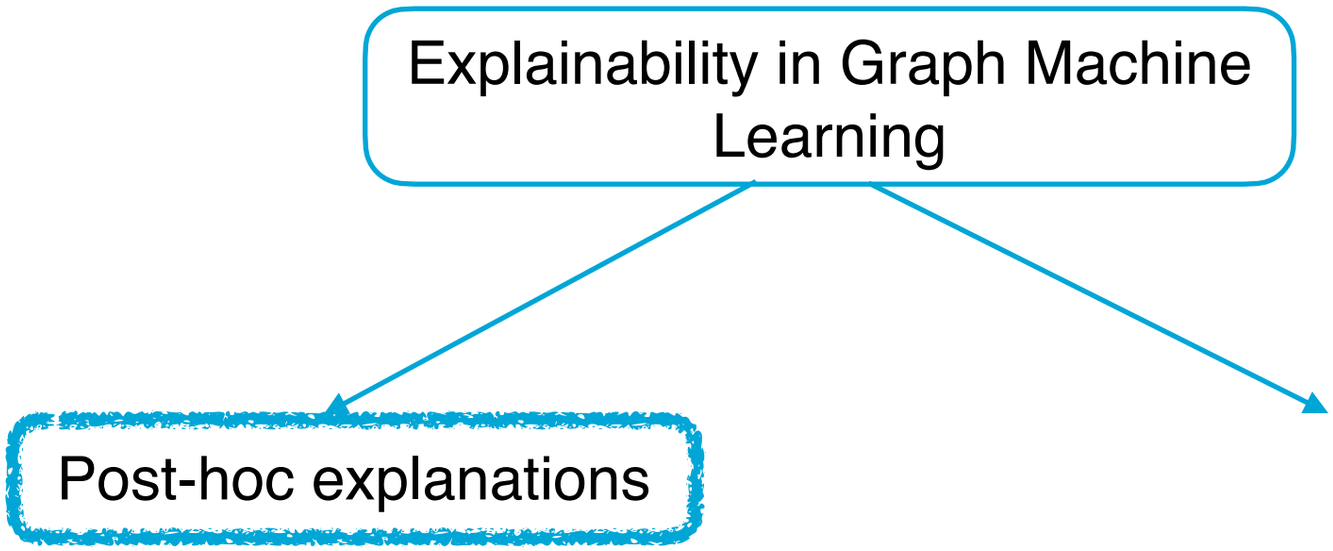
Explaining GraphML models

Explaining/interpreting decisions of models learnt over graph data



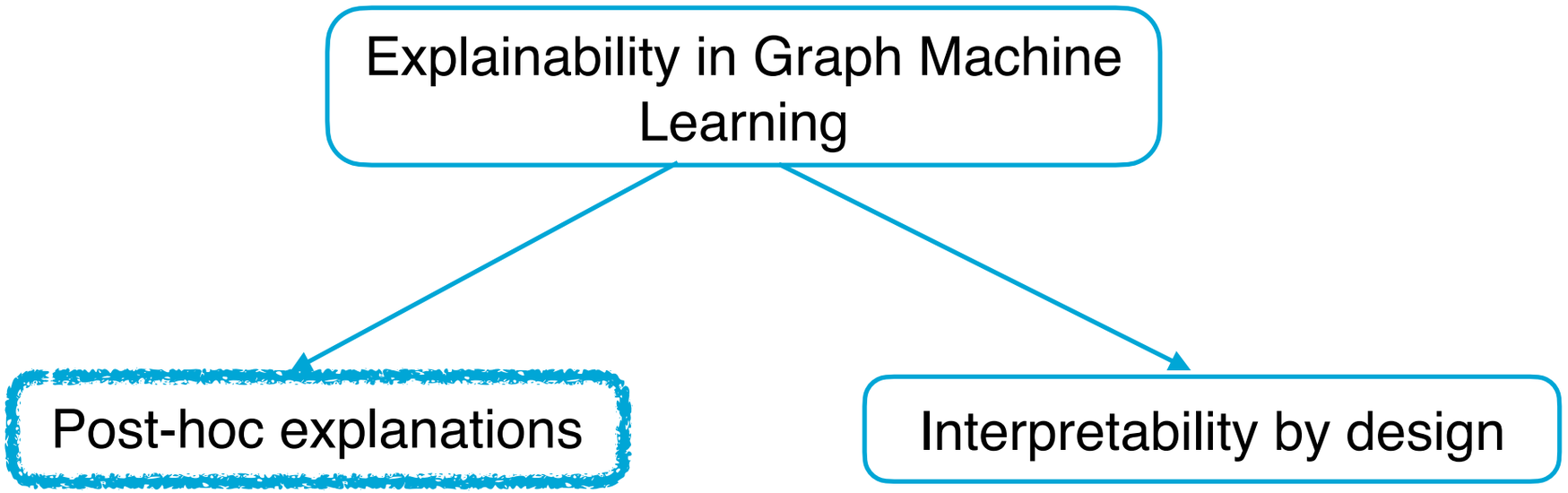
Explaining GraphML models

Explaining/interpreting decisions of models learnt over graph data



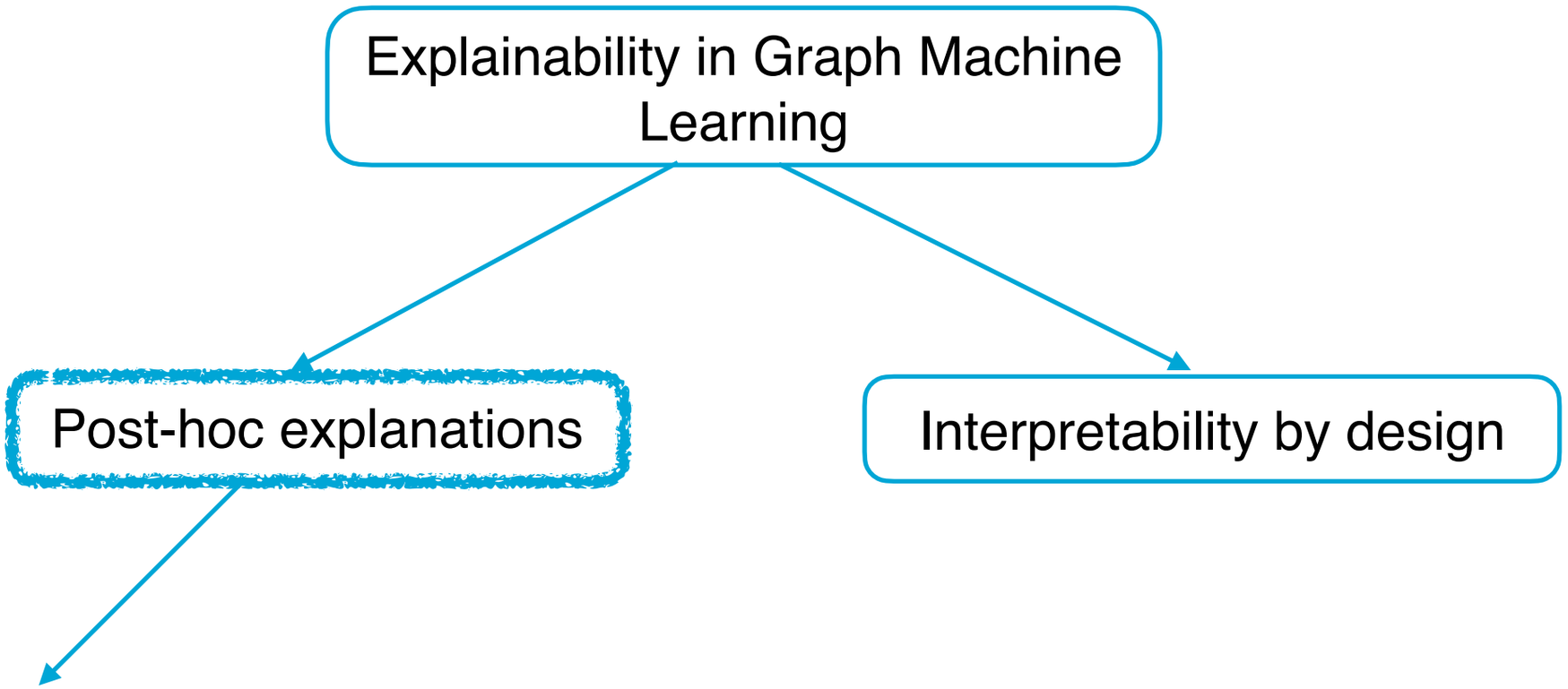
Explaining GraphML models

Explaining/interpreting decisions of models learnt over graph data



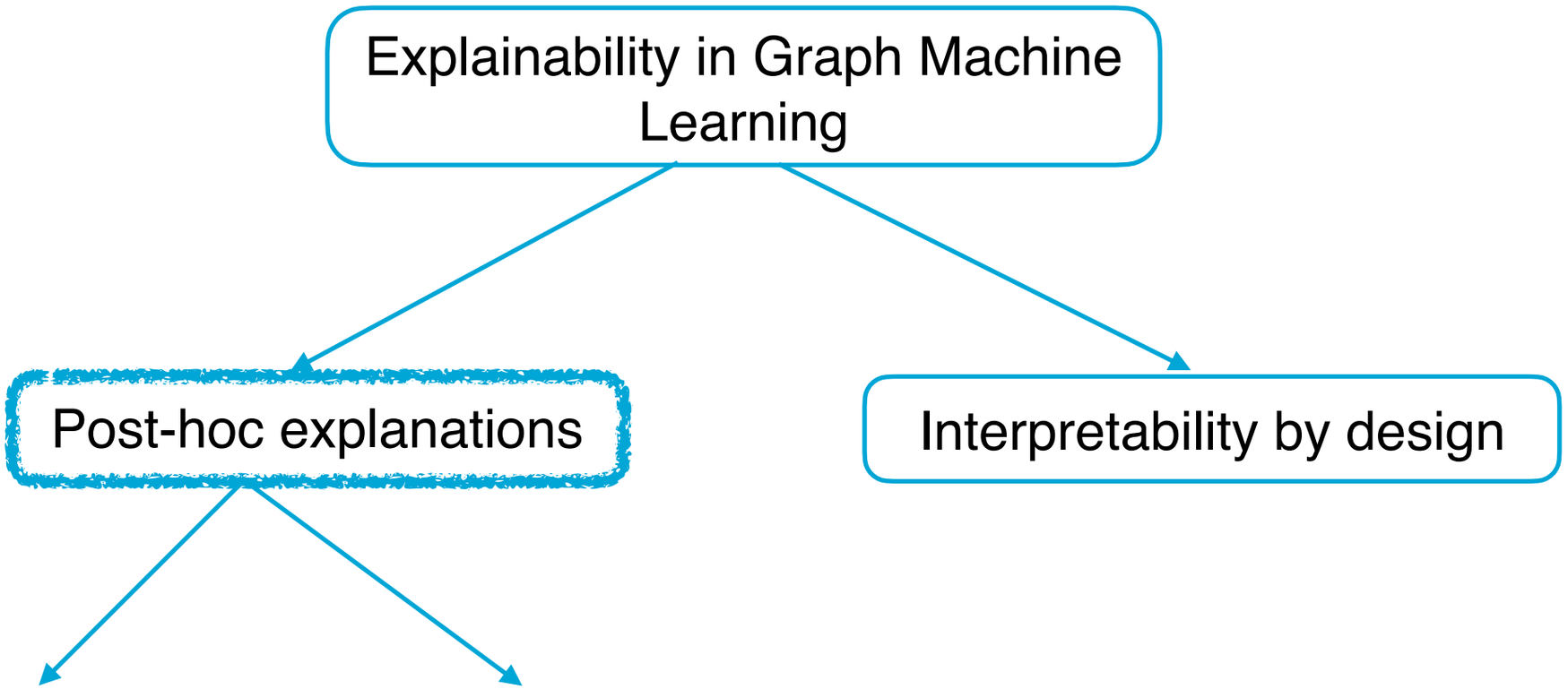
Explaining GraphML models

Explaining/interpreting decisions of models learnt over graph data



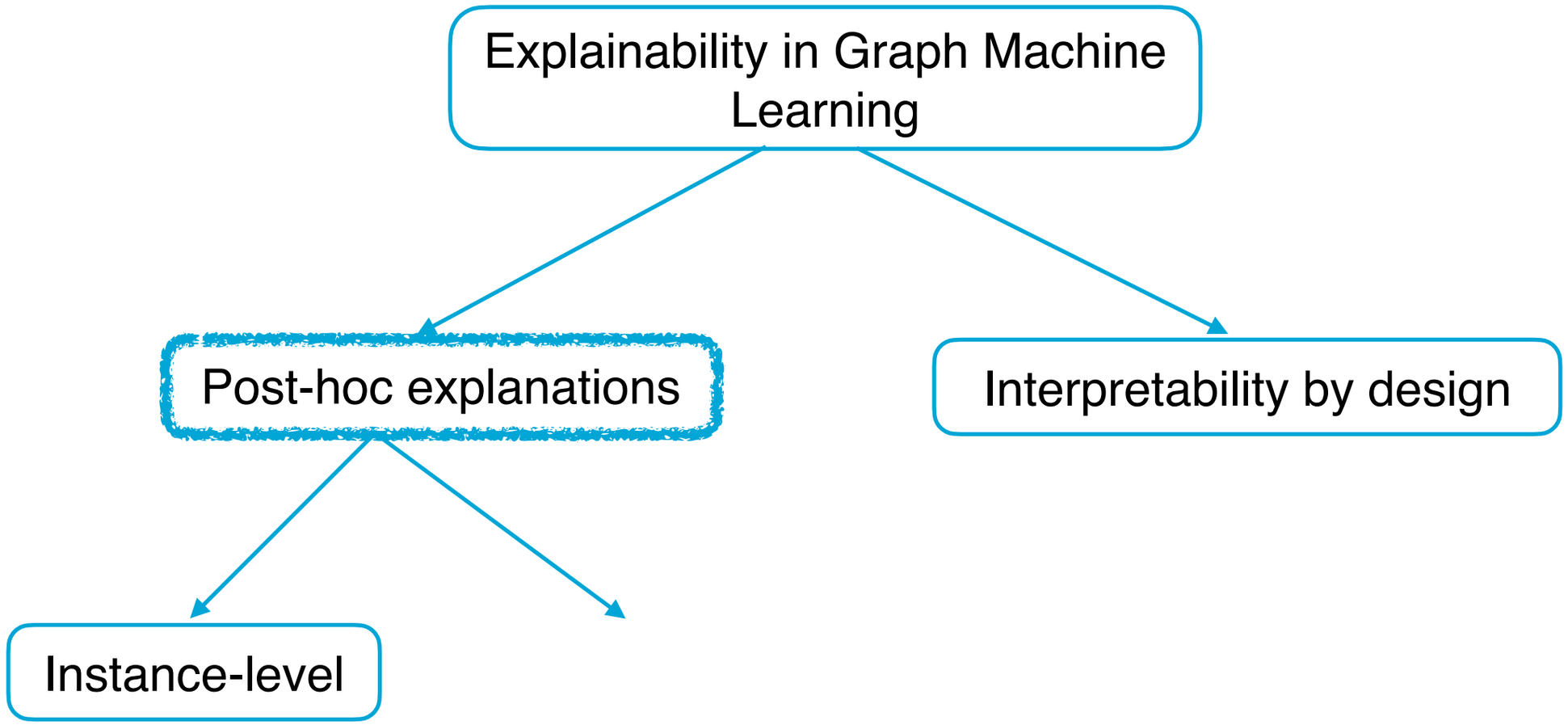
Explaining GraphML models

Explaining/interpreting decisions of models learnt over graph data



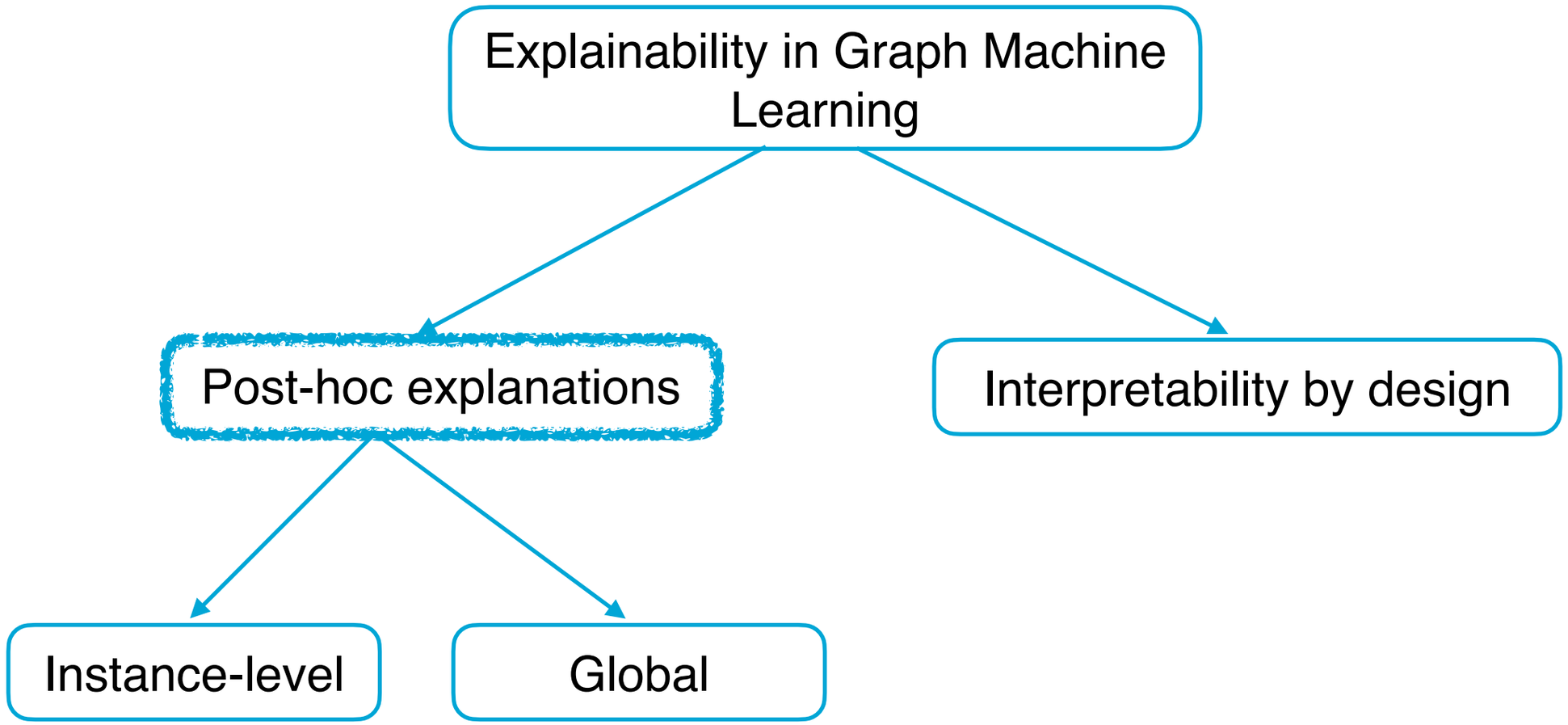
Explaining GraphML models

Explaining/interpreting decisions of models learnt over graph data



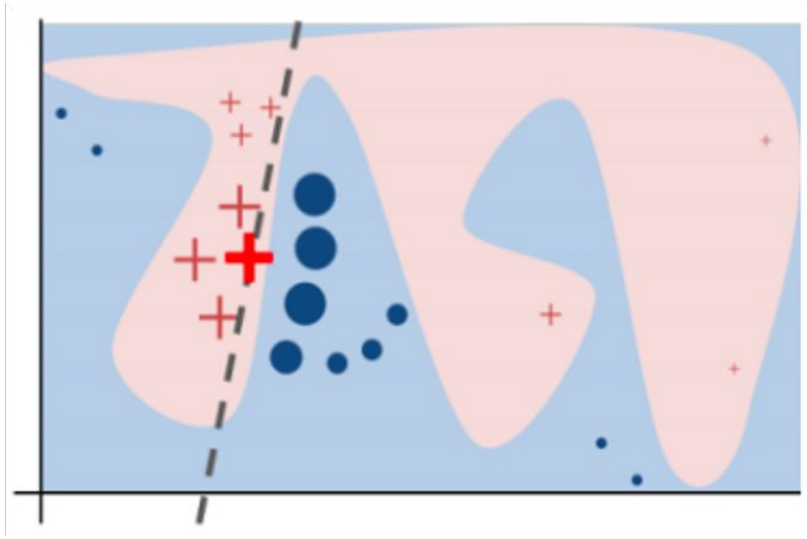
Explaining GraphML models

Explaining/interpreting decisions of models learnt over graph data



Post hoc Vs. Interpretability by design

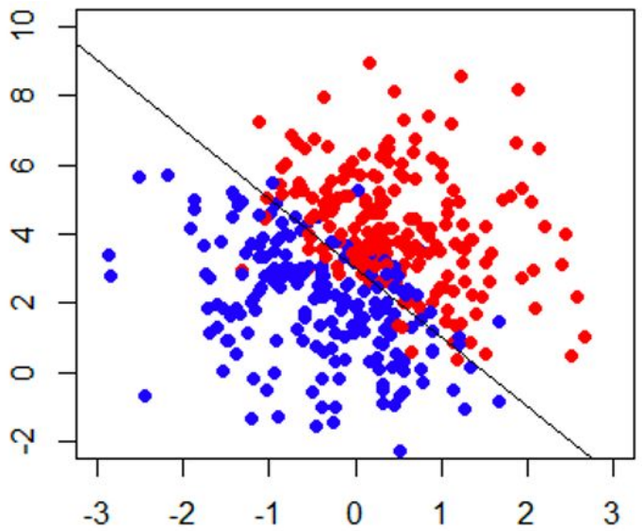
Post-hoc explanations



Explaining an already trained complex model
does not affect its performance

Explanations might not be faithful to the model

Inherently interpretable models



Simpler inherently interpretable models could incur loss of performance

Explanations are by design faithful to the model

This tutorial- Explainability of GNNs

Approaches for Post-hoc Explainability

Evaluation of Explanations

Hands on Session

Approaches for Post-hoc Explainability

Approaches for Post-hoc Explainability

Instance-wise or Local explanations

Explain individual predictions

Help to judge if individual predictions are right for the right reasons

Shed light on local biases

Global explanations

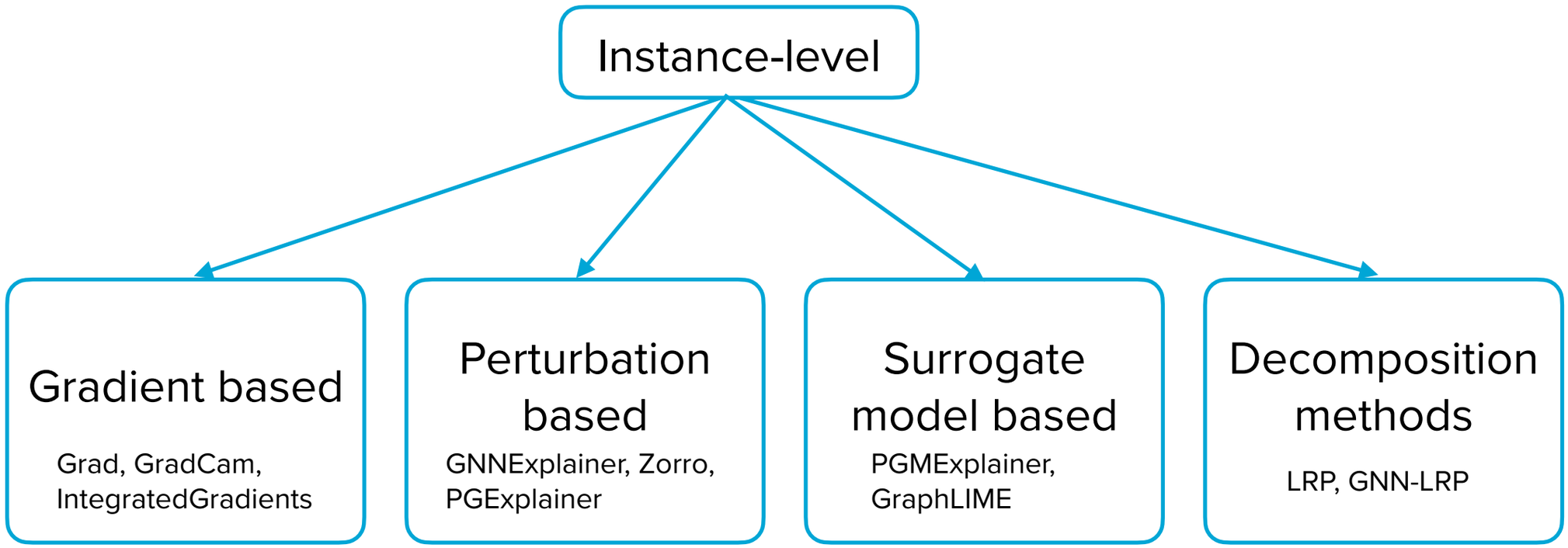
Ideally should explain complete model behaviour

Help to judge if the model at a higher level is suitable for deployment

Shed light on big picture biases affecting larger subgroups

Instance-level Explanations

Explain individual predictions using local structure of the given instance



Gradient based Explainers

Gradient based Explainers

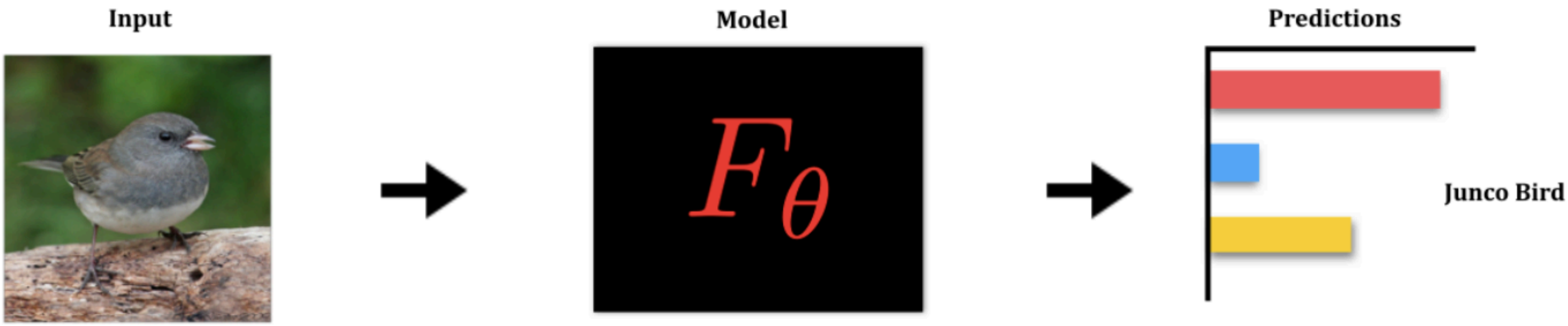


Image Source : https://explainml-tutorial.github.io

$$F : \mathbb{R}^n \rightarrow \mathbb{R}^c$$

Model F with n-dimensional input and c classes.

The class specific logic is given by

$$F_i : \mathbb{R}^n \rightarrow \mathbb{R}$$

Gradient based Explainers

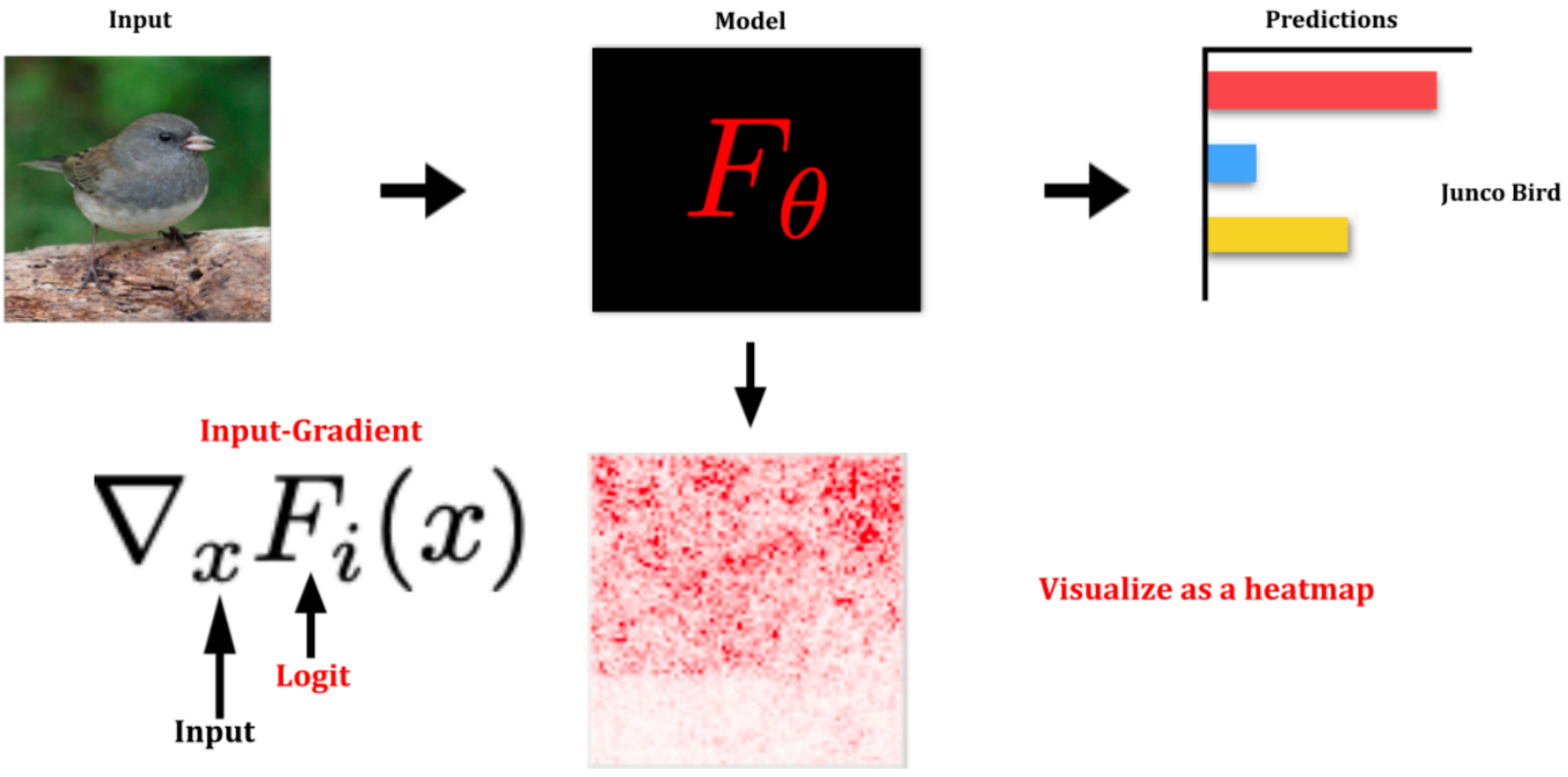


Image Source : https://explainml-tutorial.github.io

Gradient based Explainers

Gradient (Grad)

$$\nabla_x F_i(x)$$

GradInput

$$\nabla_x F_i(x) \odot x$$

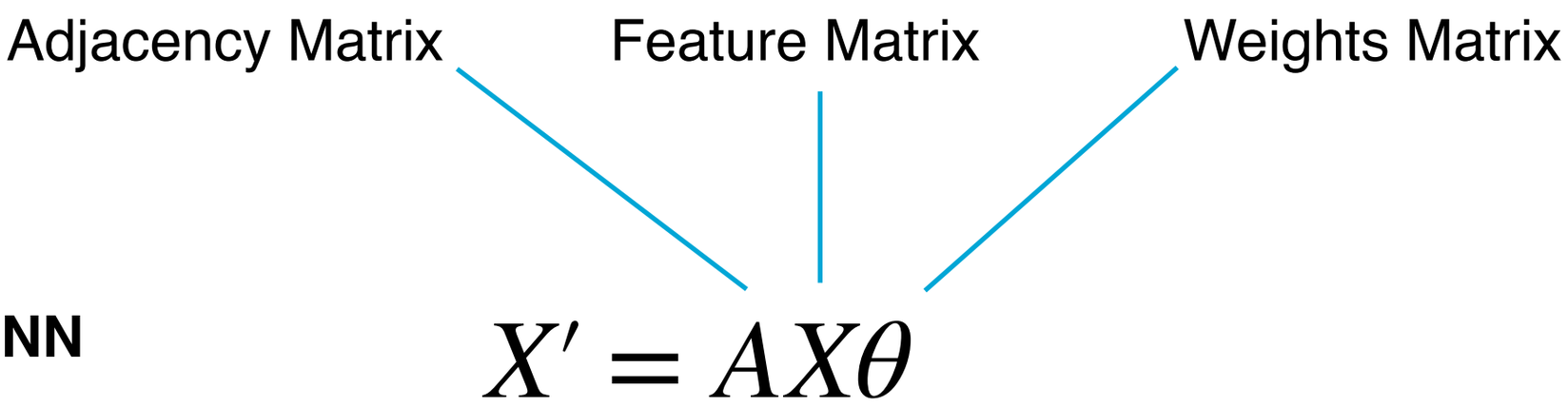
SmoothGrad

$$\frac{1}{N}\sum_{i=1}^{N} \nabla_{x+\varepsilon} F_i(x+\varepsilon)$$

IntegratedGrad

$$(x-\tilde{x}) \times \int_{\alpha=0}^{1} \frac{\delta F(\tilde{x}+\alpha \times (x-\tilde{x}))}{\delta x}$$

Gradient based Explainers for GNNs



Simplified GNN

$$X' = AX\theta$$

Node importance $\frac{\delta}{\delta x_i} X'$ (column-wise gradient)

Feature importance $\frac{\delta}{\delta x_j} X'$ (row-wise gradient)

Edge importance $\frac{\delta}{\delta a_{ij}} X'$

Perturbation based Explainers

Perturbation based Explainers

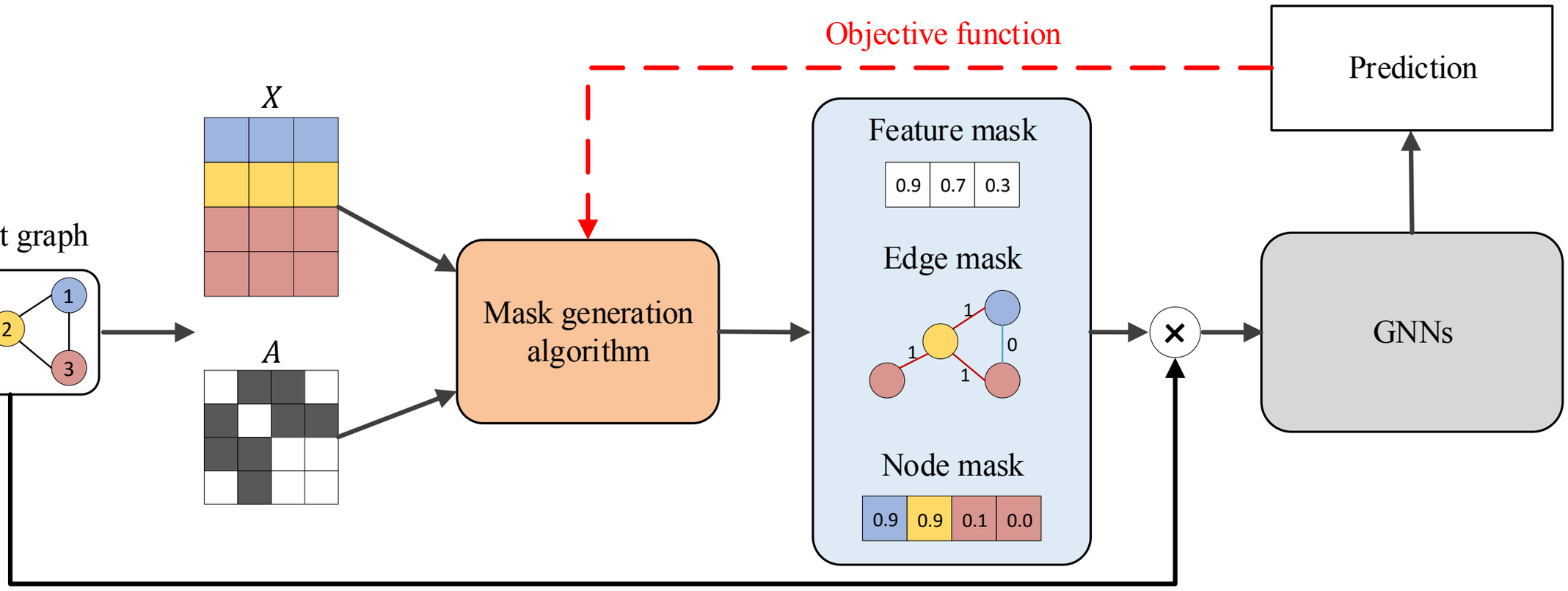


Image Source : Yuan et al. 2021

Key Idea: Given a trained GNN, generate a mask over the input features/ nodes/edges such that the masked input leads to the similar prediction as the original one produced using the whole input.

Example 1 : GNNExplainer

[Ying et al. NeurIPS 2019]
https://arxiv.org/abs/1903.03894

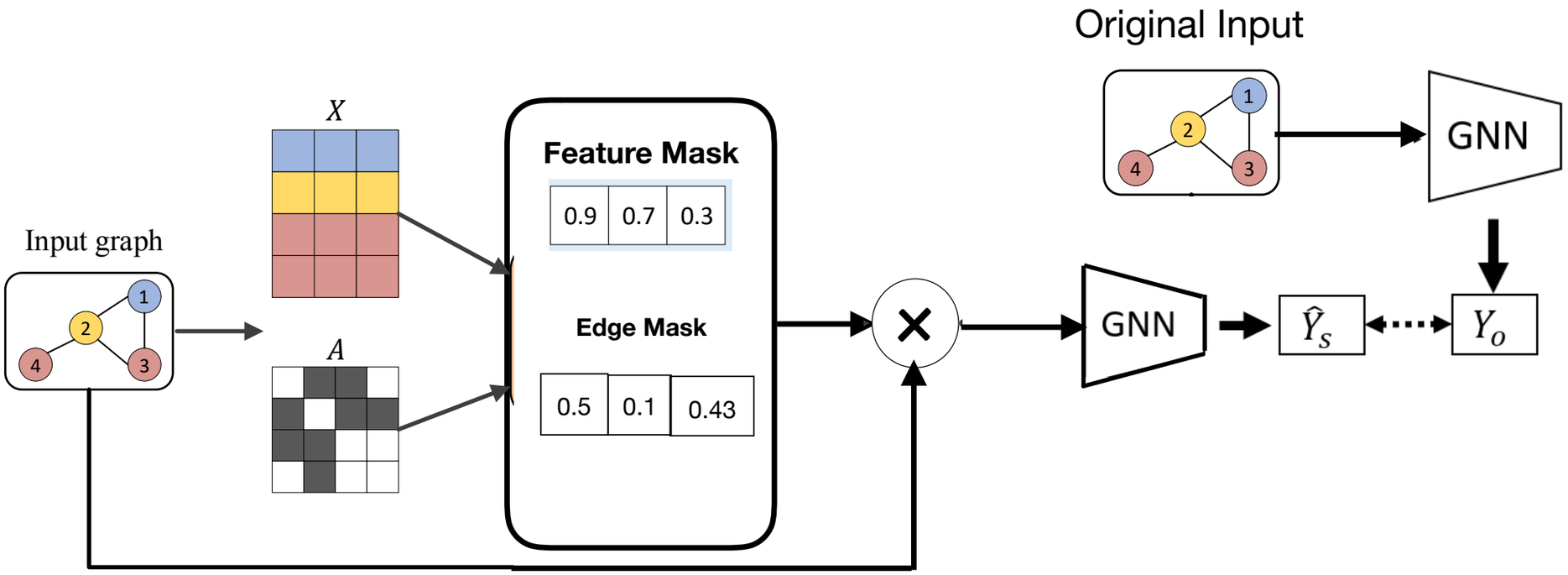
Key Idea

Find an explanation such that the mutual information between the explanation and the original prediction is maximized

Explanation Type

Continuous importance scores over features and edges
(Soft edge and feature masks)

Example 1 : GNNExplainer



Learn an edge and feature mask such that the log probability for original predicted class is maximised.

$$\min_{M_e, M_f} - \sum_{c=1}^{C} \delta_{y=c} \log P_{\Phi} \left(Y_0 = y \mid G = A_{comp} \odot \sigma(M_e), X = X_{comp} \odot M_f \right)$$

Example 2: Zorro

[Funke et al. TKDE 2022]
https://arxiv.org/abs/2105.08621

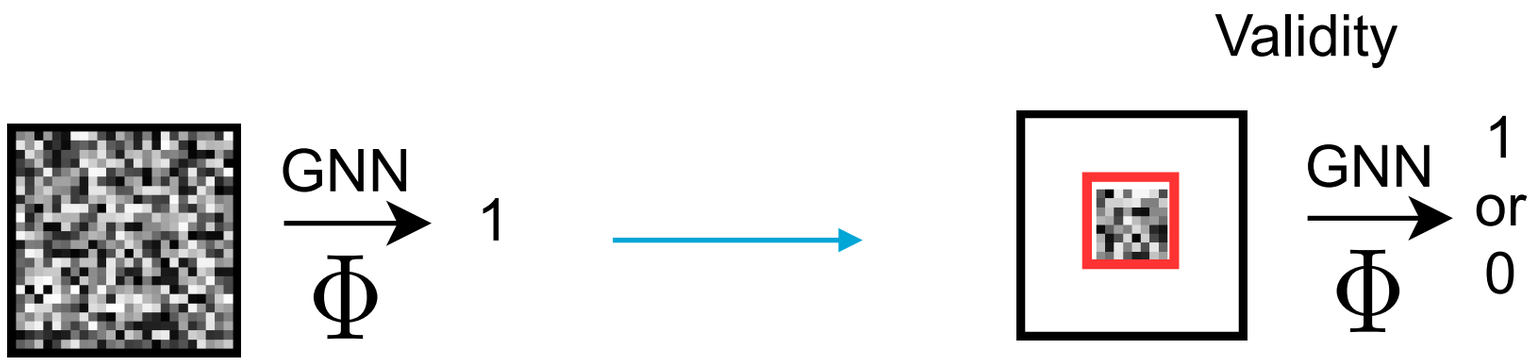
Key Idea

Compute a **valid**, **sparse** and **stable** explanation such that the prediction using the explanation is in **expectation** close to the original prediction

Explanation Type

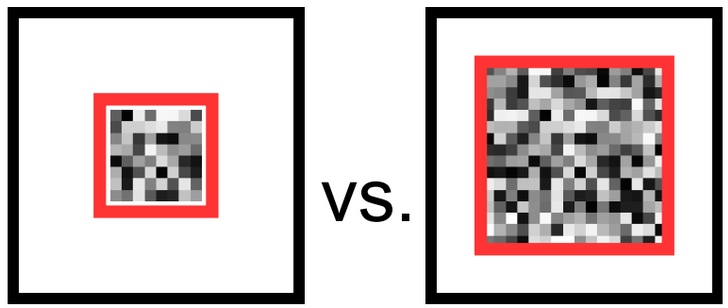
Binary importance scores over features and nodes
(Hard node and feature masks)

Valid Explanation



A **subset** of the input such that the prediction while just using the input stays the same as the original prediction is a **valid** explanation

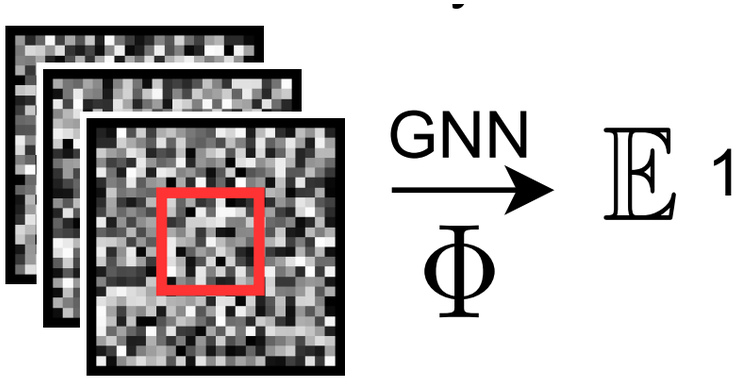
Sparsity



The chosen subset (explanation) should be sparse

Stability

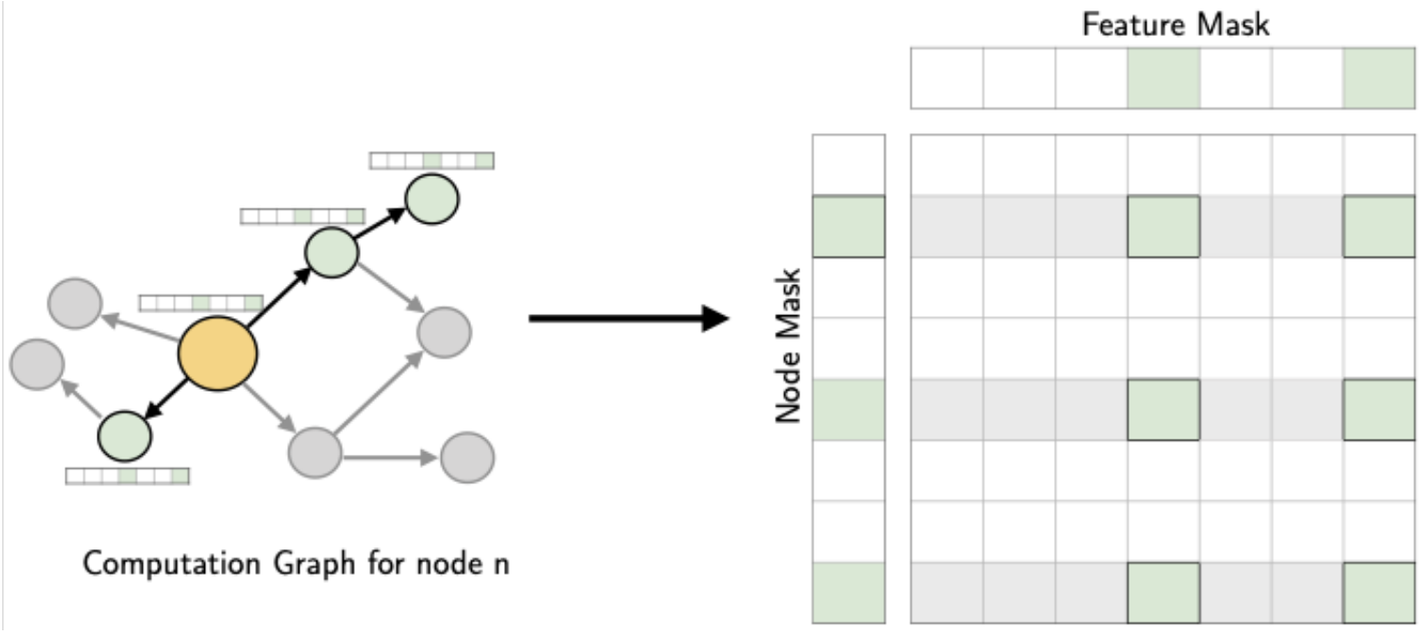
What happens to the not selected part of the input?



- Set the not selected part by some noisy values.
- Check the expected prediction over multiple such perturbations.

A stable explanation is one which achieves in expectation a close prediction to that of the original prediction

Constructing a perturbed input



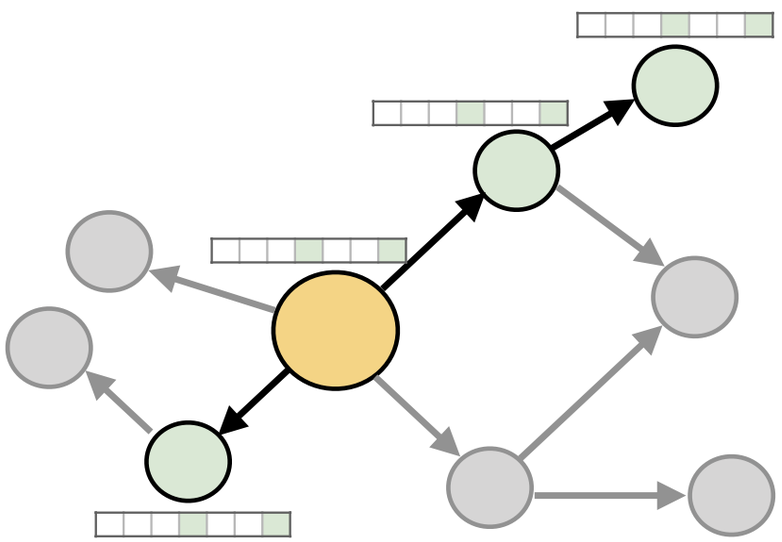
Selected nodes and features are marked green

Construct a perturbed input by setting selected features of selected nodes (the **green** cells) to their true values and others to random noisy values

Mathematically if M(S) corresponds to product of feature and node masks

$$Y_{\mathcal{S}} = X \odot M(\mathcal{S}) + Z \odot (\mathbf{1} - M(\mathcal{S})), Z \sim \mathcal{N}$$

RDT-Fidelity of an explanation



Computation Graph for node n

$$\mathcal{F}(\mathcal{S}) = \mathbb{E}_{Y_{\mathcal{S}}|Z\sim\mathcal{N}}\left[\mathbf{1}_{f(X)=f(Y_{\mathcal{S}})}\right]$$

with

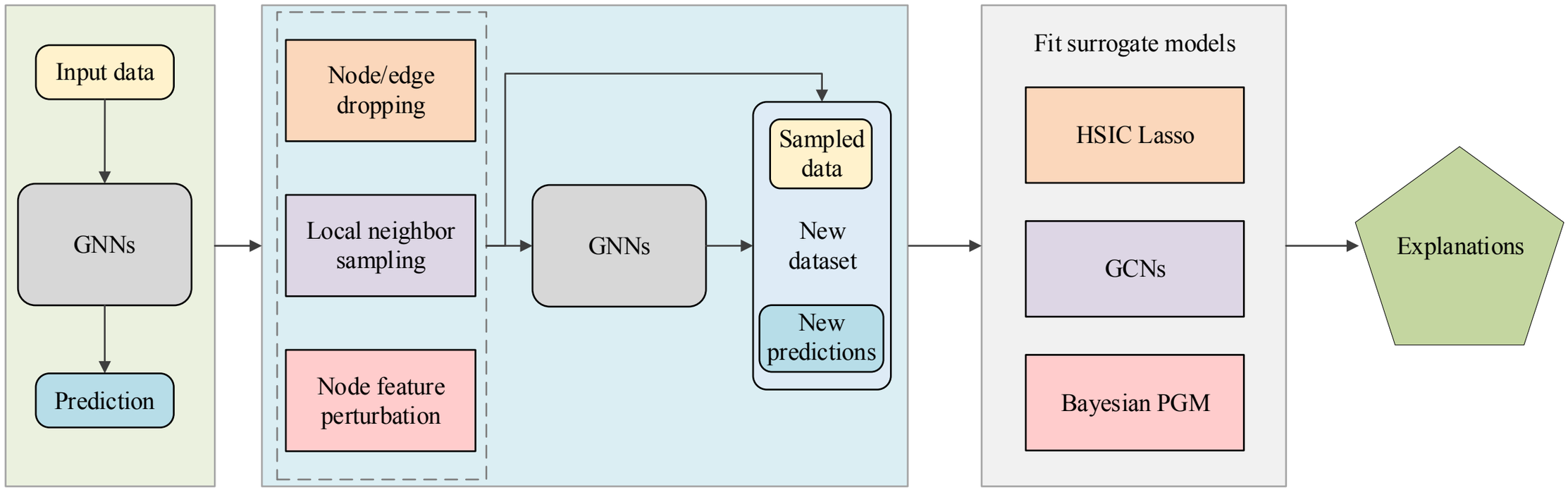
$$Y_{\mathcal{S}} = X \odot M(\mathcal{S}) + Z \odot (\mathbf{1} - M(\mathcal{S})), Z \sim \mathcal{N}$$

Zorro

Find the sparsest explanation such that its RDT-fidelity is maximised.

Surrogate Model Based

Surrogate Model based



Key Idea: Given an input graph and its prediction, sample a local dataset to represent the relationships around the target data. Surrogate methods which are usually interpretable by design are applied to fit the sampled local dataset.

Example: PGMExplainer

[Wu and Thai, NeurIPS 2020]
https://arxiv.org/abs/2010.05788

Key Idea

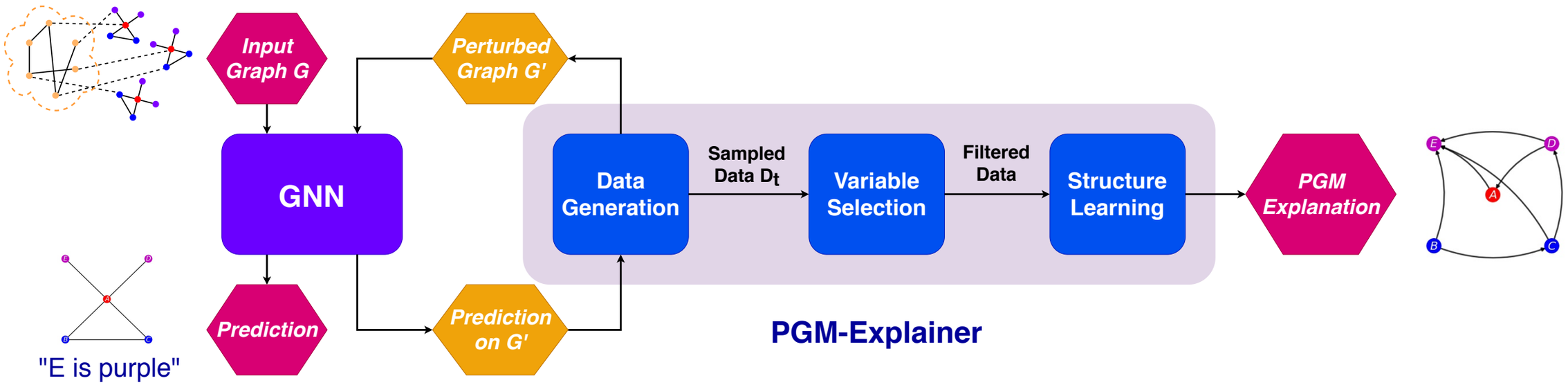
Fits an interpretable model to a local perturbed dataset

Explanation Type

Hard node masks
(Top nodes are output based on node importances)

Example : PGMExplainer

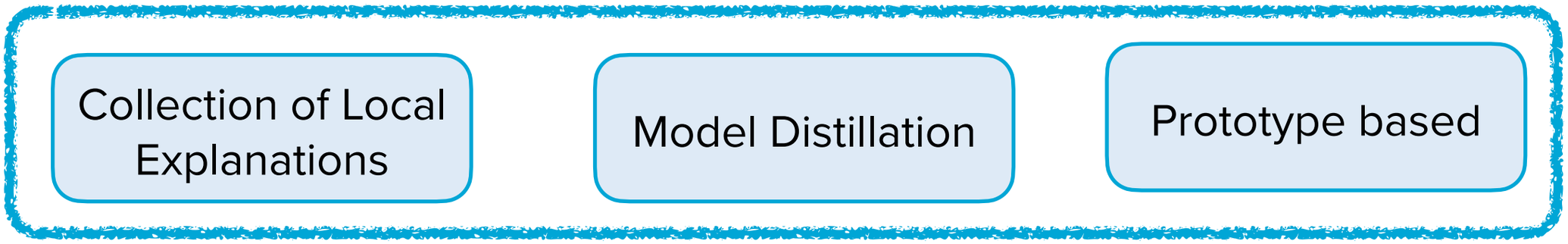
[Wu and Thai, NeurIPS 2020]



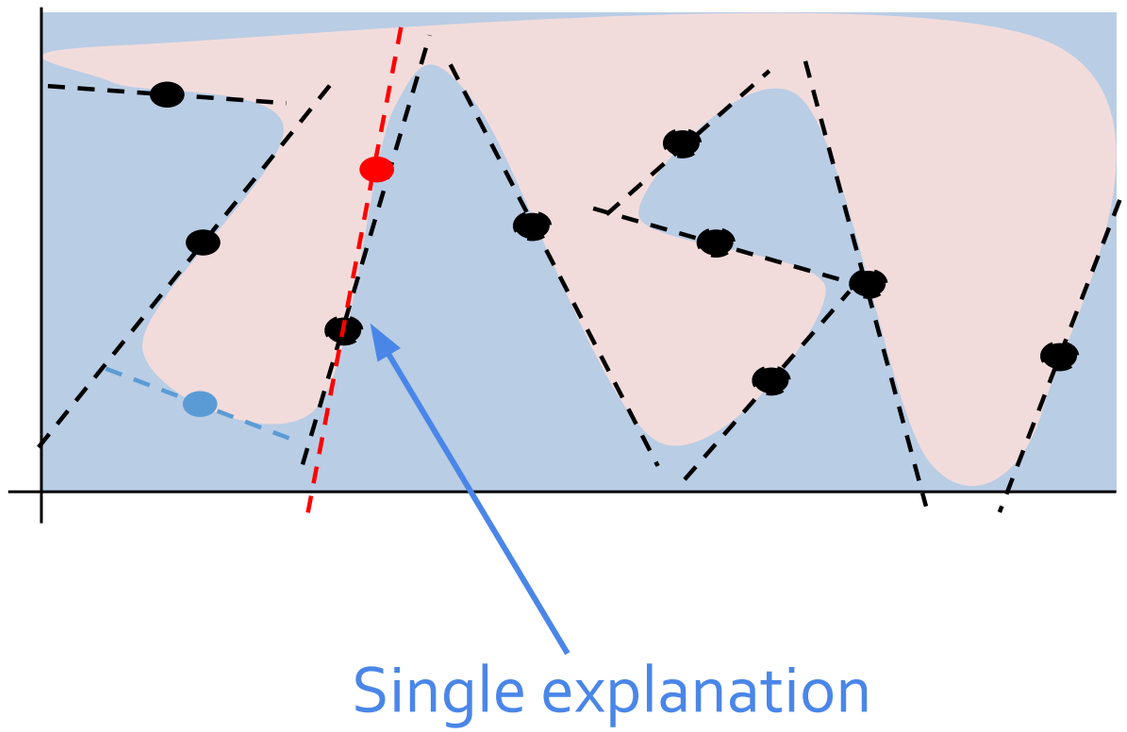
Generate a local dataset by randomly perturbating features of nodes in the computational graph and corresponding predictions.

Fit a Bayesian model on the perturbed data to obtain **node importances**

Global Explanations



Collection of Local Explanations

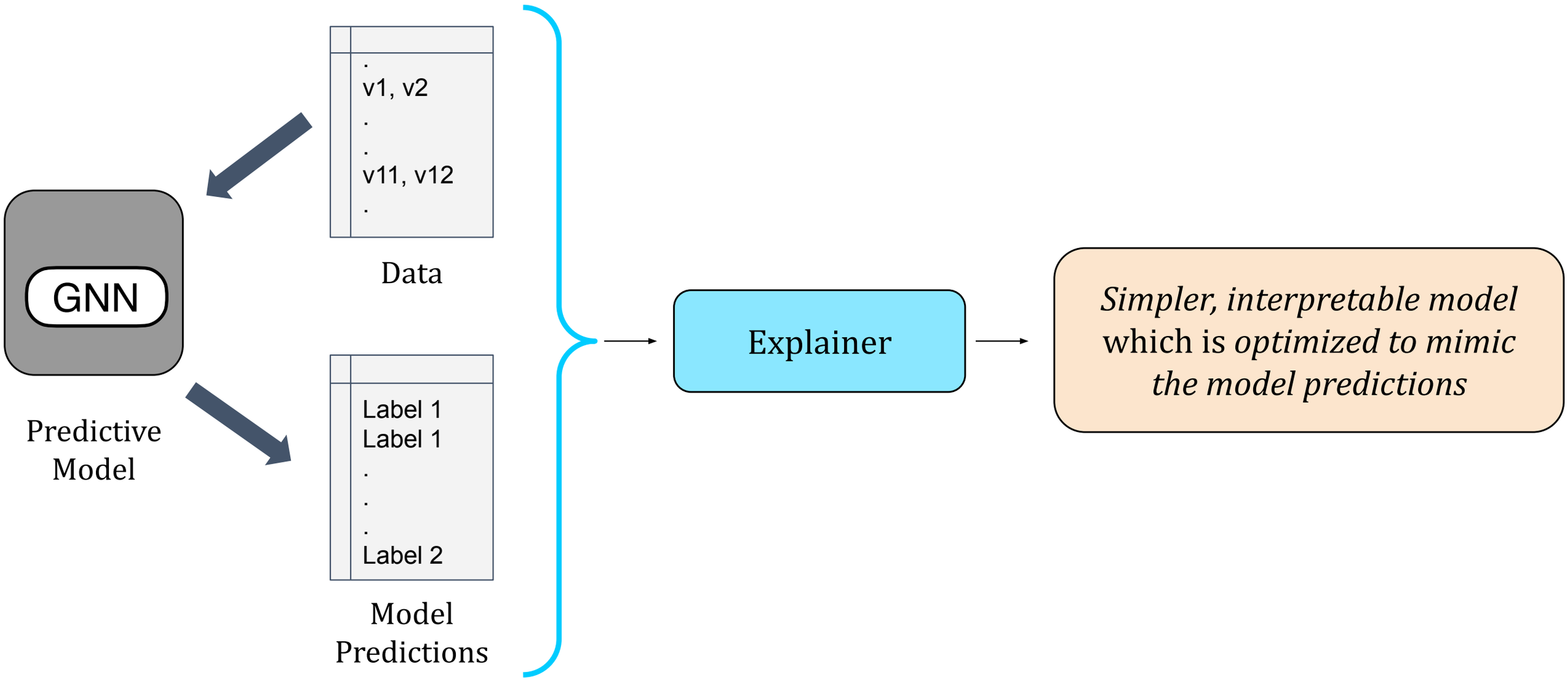


- Generate a local explanation for each instance
- Pick a subset of k explanations to constitute a global explanation

So far no explainer for graph data follow this strategy

Image credit : ExplainML tutorial AAAI 21

Model Distillation

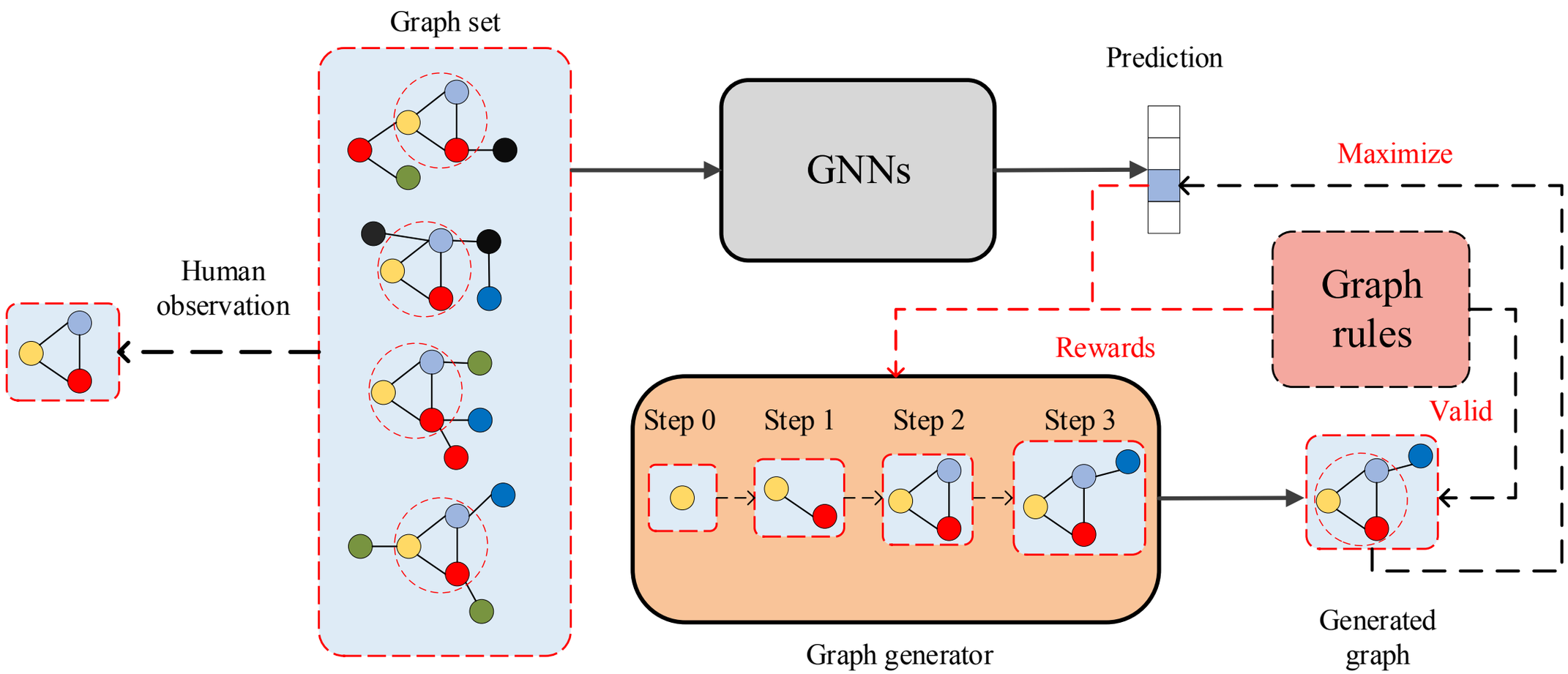


Viable strategy but so far no explainer for graph data follow this strategy

Image credit : ExplainML tutorial AAAI 21

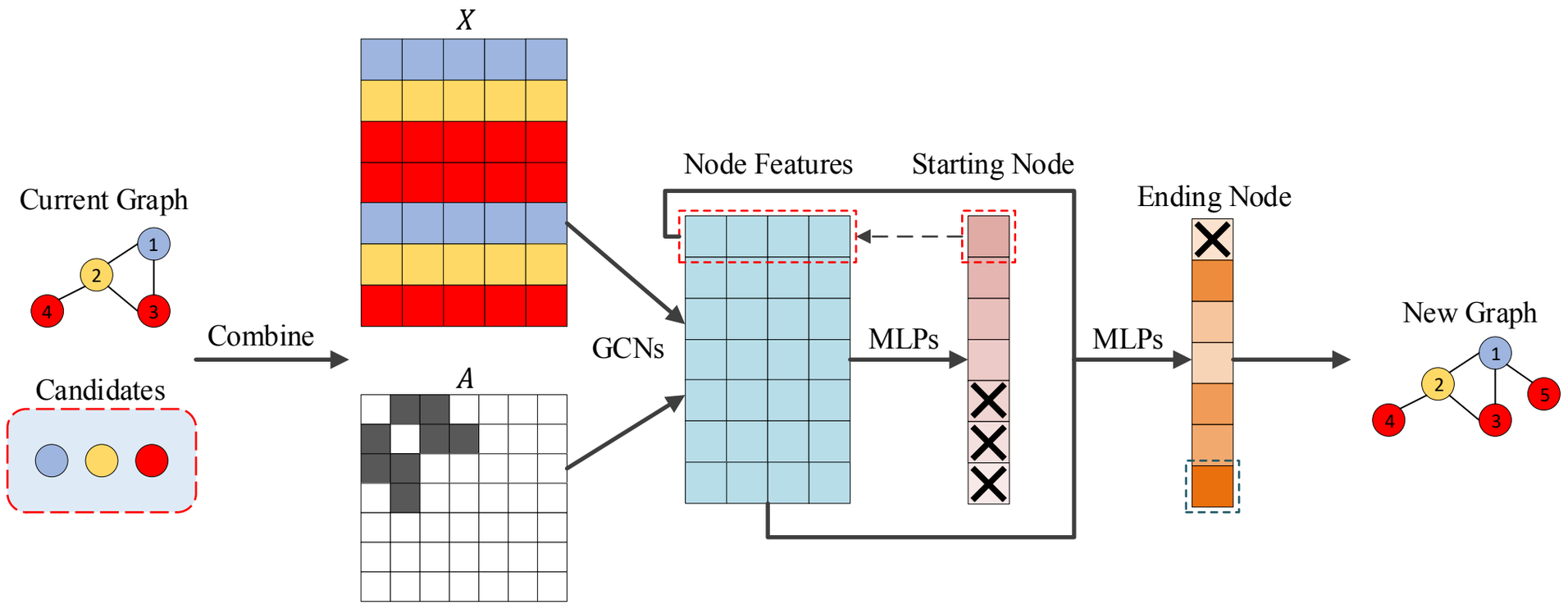
Global Explanations by XGNN

[Yuan et al., 2020]
https://arxiv.org/abs/2006.02587



Key Idea: Generate graph, $G^\star$ which maximise the prediction probability of a particular class: $G^\star = \underset{G}{\operatorname{argmax}}\, \mathbb{P}(f(G) = c)$

Generation of Explanation Graph

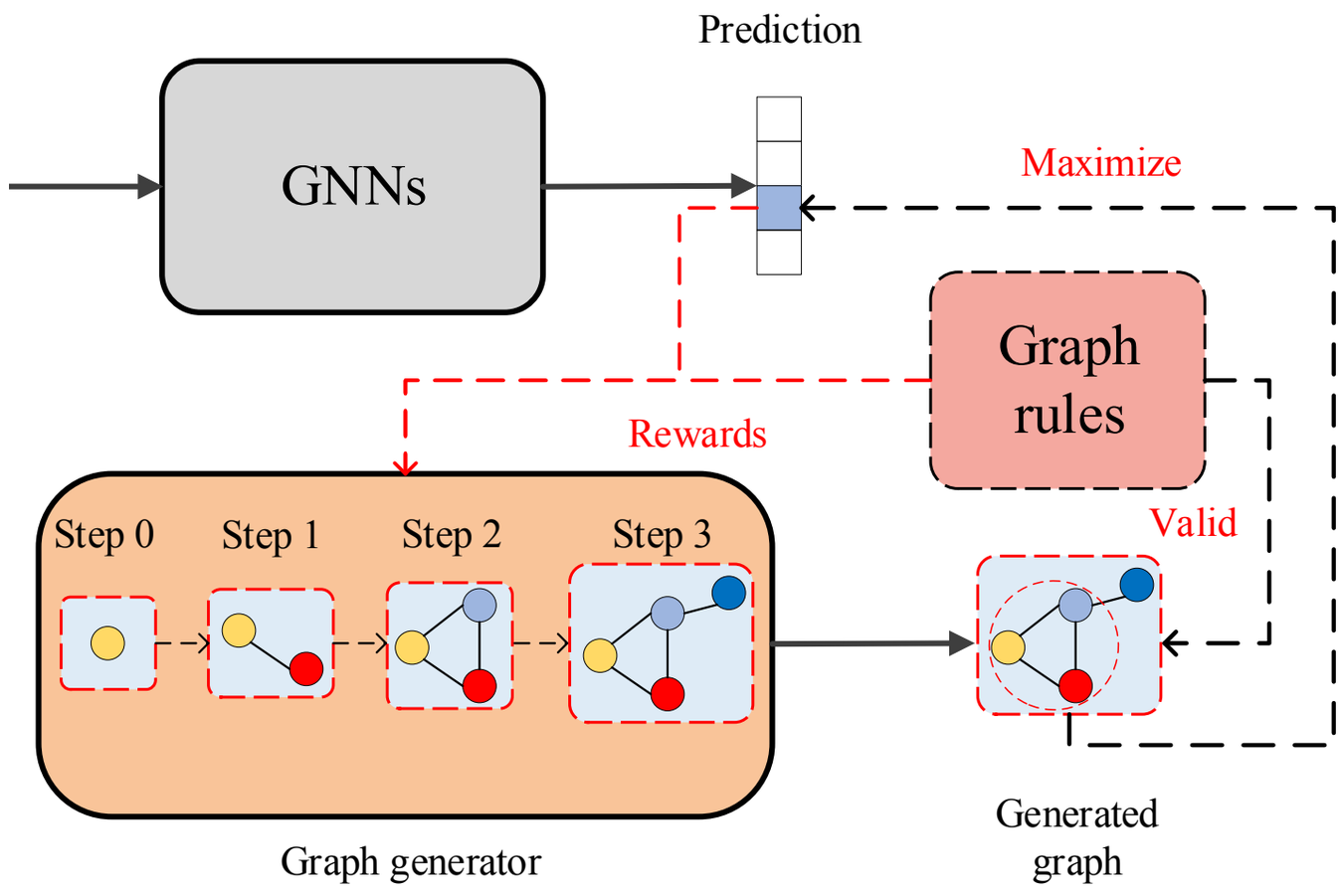


Starting with a candidate node, predict the start and end node for the edge to be added

Image Source : Yuan et al., 2020

Training of Graph Generation

- Considering the Reinforcement Learning setting each intermediate generated graph corresponds to a state.
- Action corresponds to selection of start and end node of the edge to added
- Training by Policy Gradient
- Reward computed using prediction probability and graph rules



Evaluation of Post-hoc Explanations

What is a good explanation?

How to measure the goodness of an explanation?

Evaluating Post-Hoc Explanations

Functionally-grounded evaluation

Faithfullness

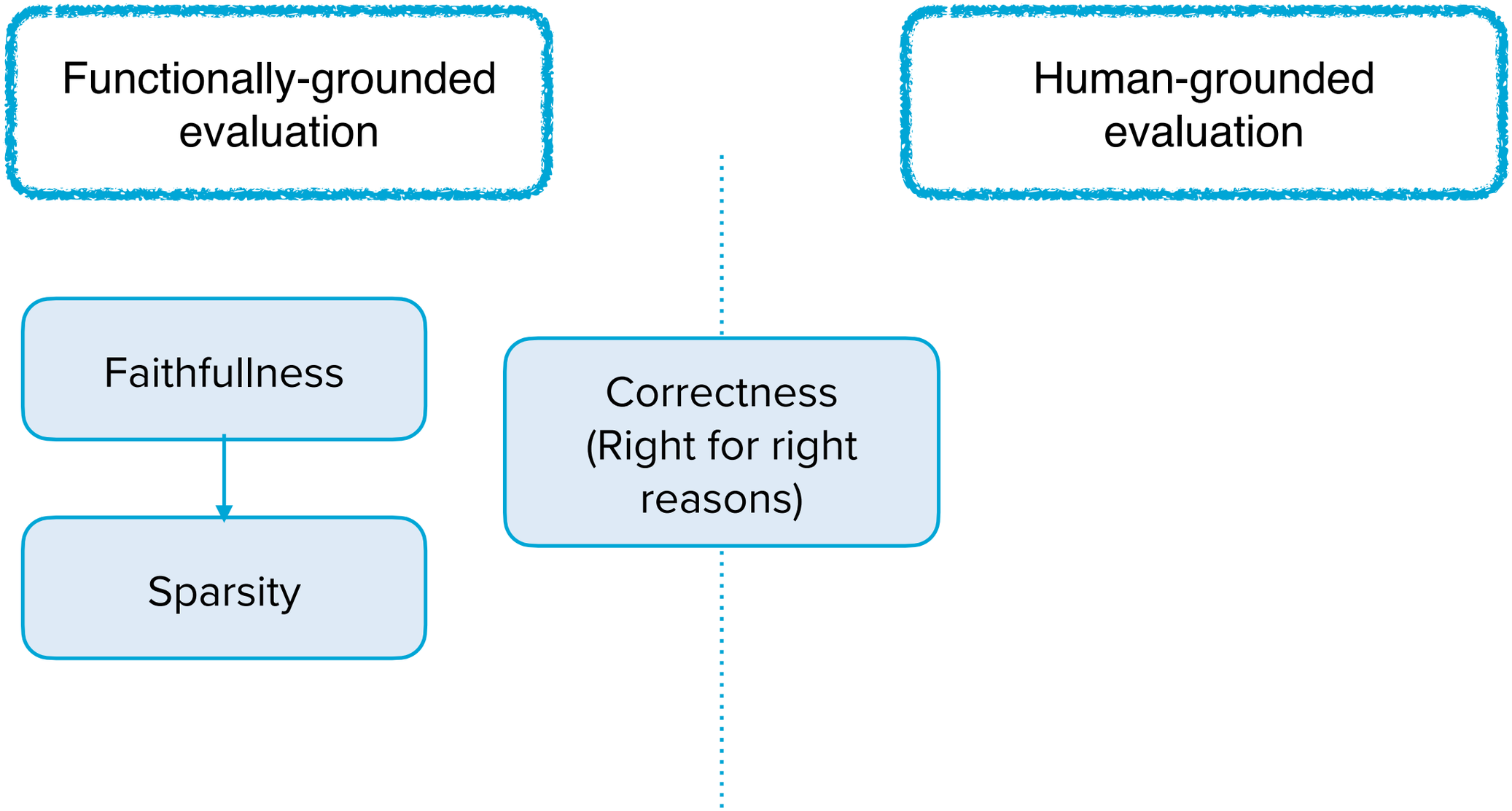
Sparsity

Human-grounded evaluation

[BAGEL Benchmark, Rathee et al. 2022]

https://github.com/Mandeep-Rathee/Bagel-benchmark

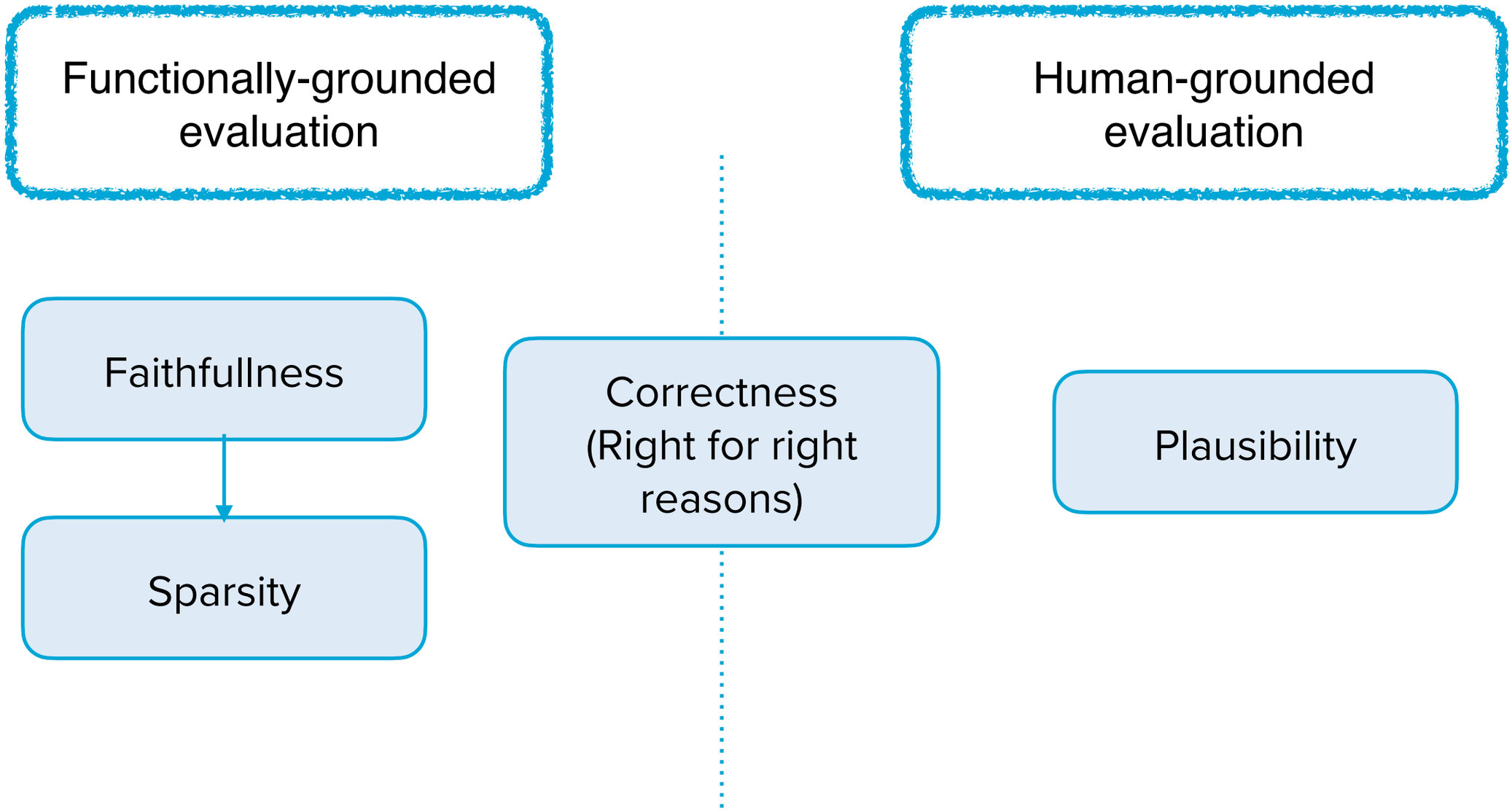
Evaluating Post-Hoc Explanations



[BAGEL Benchmark, Rathee et al. 2022]

https://github.com/Mandeep-Rathee/Bagel-benchmark

Evaluating Post-Hoc Explanations



[BAGEL Benchmark, Rathee et al. 2022]

https://github.com/Mandeep-Rathee/Bagel-benchmark

Faithfullness

Take 1: Check *sufficiency* and *comprehensiveness* of the explanation

Faithfullness

Take 1: Check *sufficiency* and *comprehensiveness* of the explanation

Sufficiency

Keep the most important features/nodes/edges and check if they alone can predict the original decision.

Faithfullness

Take 1: Check *sufficiency* and *comprehensiveness* of the explanation

Sufficiency

Keep the most important features/nodes/edges and check if they alone can predict the original decision.

Comprehensiveness

Remove the features/nodes/edges not in the explanation and check if the original prediction changes.

Faithfullness

How to compute sufficiency and comprehensiveness for soft masks?

Faithfullness

How to compute sufficiency and comprehensiveness for soft masks?

What happens when you cannot remove features?

Faithfullness

How to compute sufficiency and comprehensiveness for soft masks?

What happens when you cannot remove features?

$$Y_{\mathcal{S}} = X \odot M(\mathcal{S}) + Z \odot (\mathbf{1} - M(\mathcal{S})), Z \sim \mathcal{N}$$

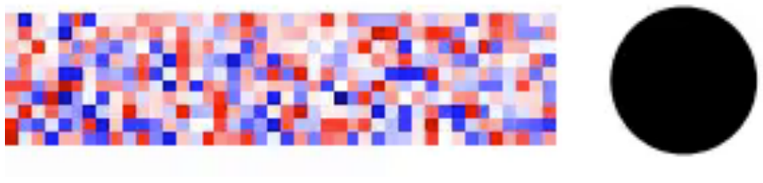
Faithfullness

How to compute sufficiency and comprehensiveness for soft masks?

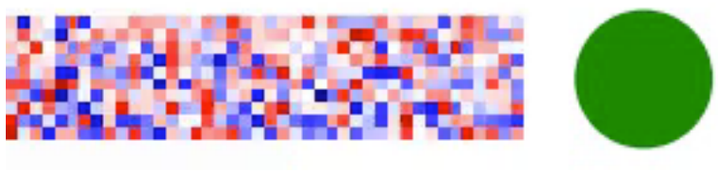
What happens when you cannot remove features?

$$Y_{\mathcal{S}} = X \odot M(\mathcal{S}) + Z \odot (\mathbf{1} - M(\mathcal{S})), Z \sim \mathcal{N}$$

Take 2: Use RDT-Fidelity to check if the explanation is predictive and stable



Low fidelity



High fidelity

Sparsity

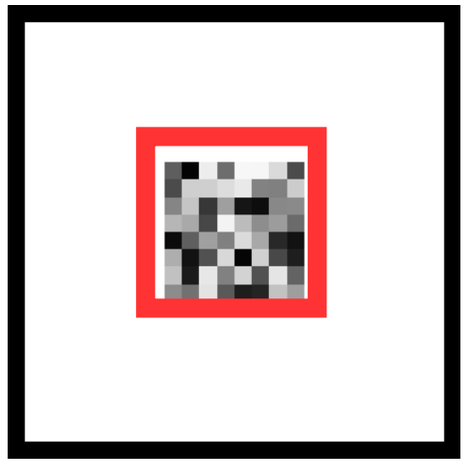
But the full input is also a faithful explanation

Are the explanations non-trivial?

Sparsity

But the full input is also a faithful explanation

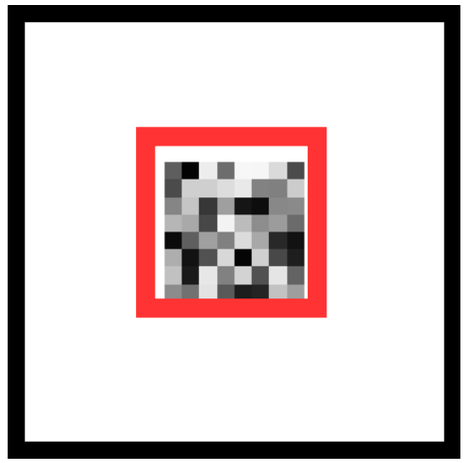
Are the explanations non-trivial?



Sparsity

But the full input is also a faithful explanation

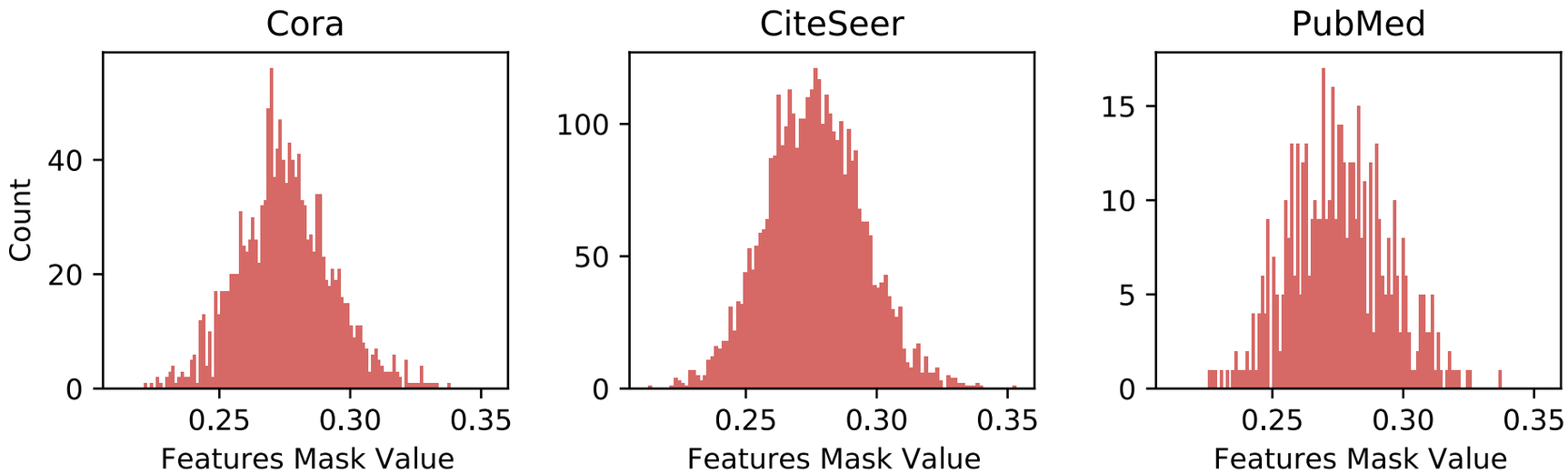
Are the explanations non-trivial?



Take 1: Sparsity for hard masks = Selection size / total

Sparsity

What about soft masks ?



A uniform distribution of normalised mask distribution implies complete input

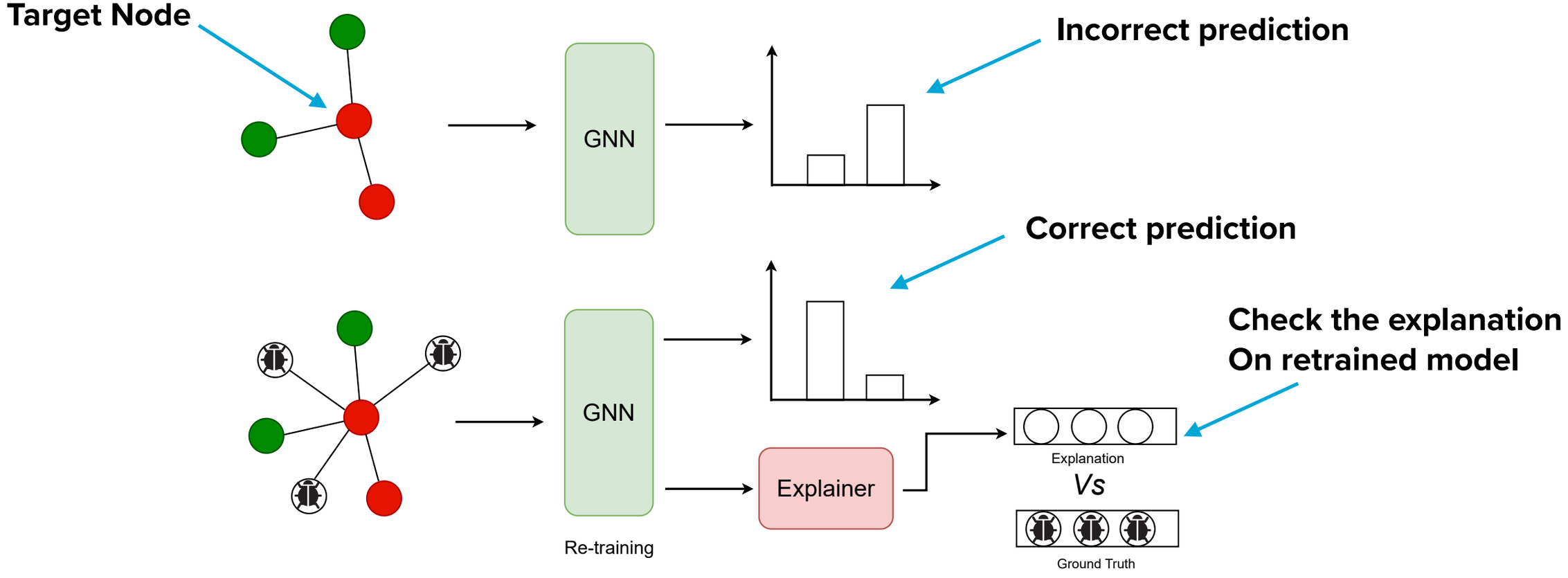
Take 2: Check Entropy of normalised distribution of masks

Lower the entropy sparser the explanation

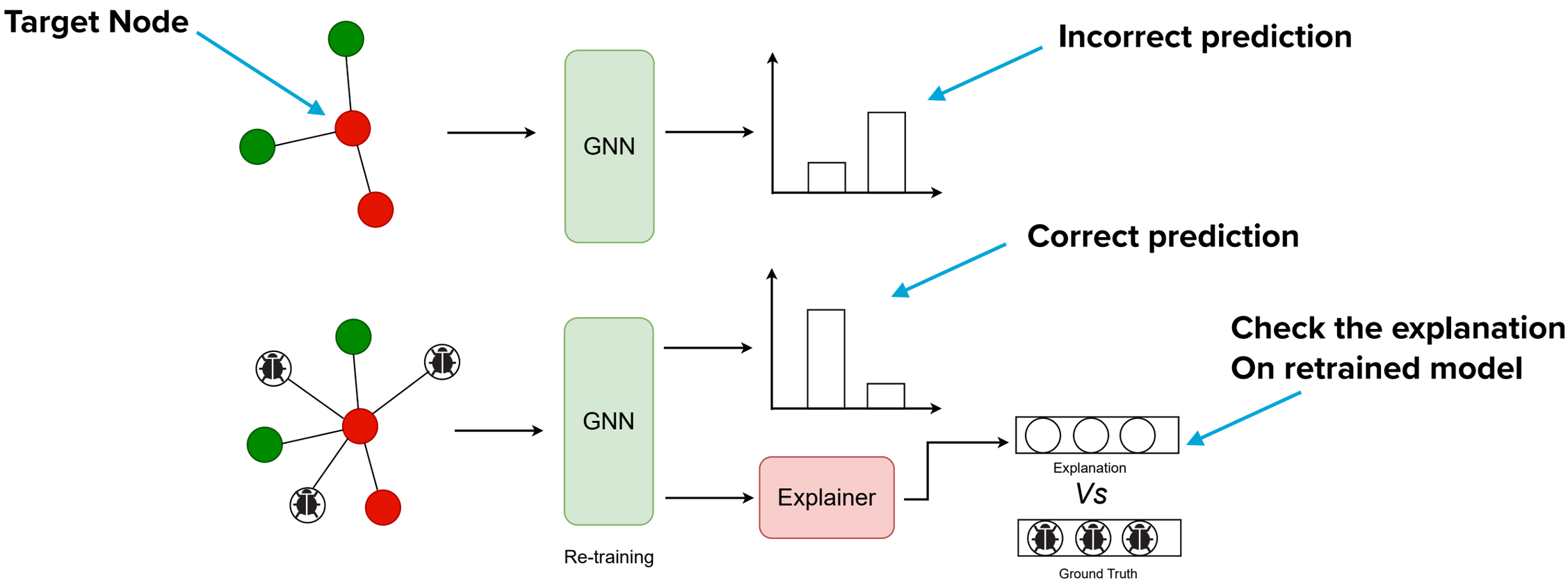
Correctness

Can the explanation model detect any injected correlations responsible for altering model's behavior ?

Introduce **correlations** in the training data which can change the decision on a node/graph. Then check if explanation discovers the added correlations.



Correctness



Drawbacks :

(i) Choosing correlations is tricky in the first place
(ii) Requires model retraining

Plausibility

How close are the explanations to human rationales ?

Human Rationales	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
GNNExp	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
Grad	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
CAM	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...

Compute agreement of explanation with human rationales

Metrics : F1 score for hard masks, AUPRC score for soft masks

Plausibility

Should be used in conjunction with a suitable faithfulness metric

First ensure that the explanation is in fact approximating model's decision

GCN: The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...

GAT: The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...

APPNP: The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...

Given the explainer is faithful to the model one can use plausibility to compare GNN models for the agreement of their decision making process with human rationales.

Other Evaluation schemes

Measuring agreement (explanation accuracy) with planted subgraph in a synthetic graph

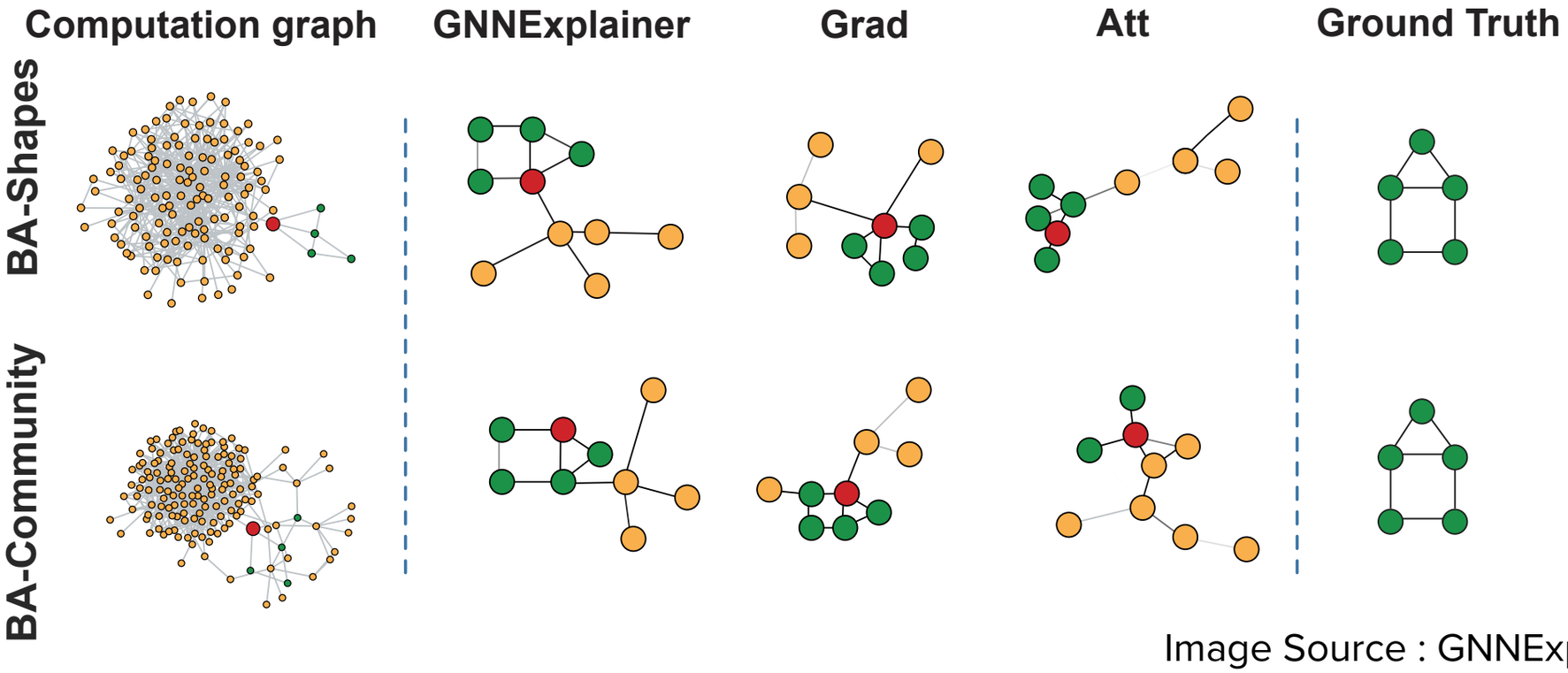


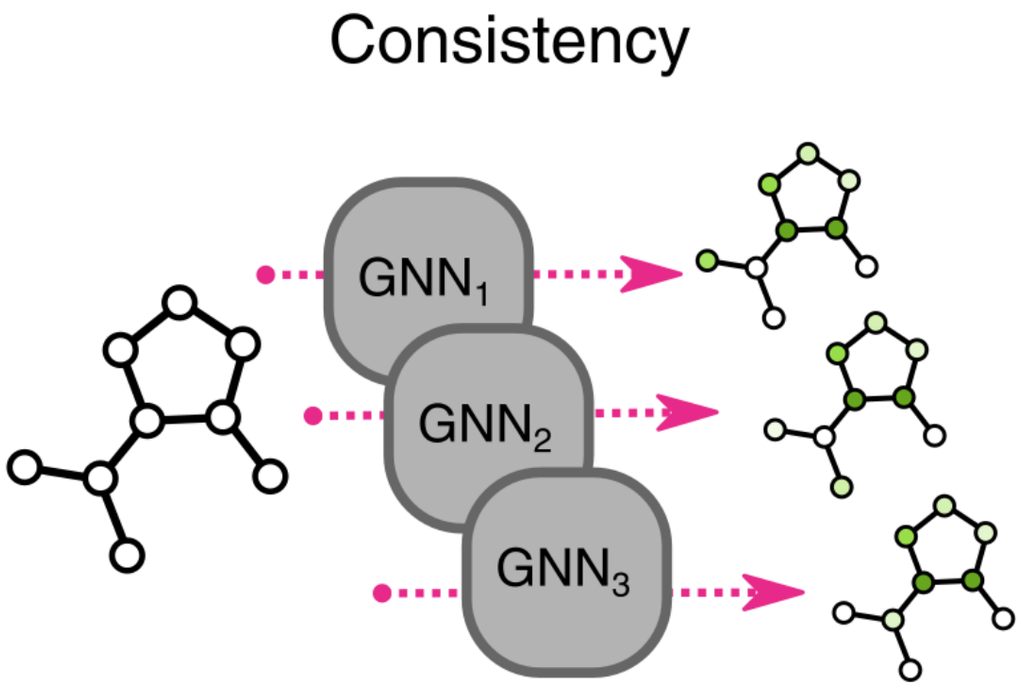
Image Source : GNNExplainer

Drawback : How to be sure if the model picked the planted subgraph?

Other Evaluation schemes

Measuring attribution (explanation) consistency across high performing models

[Sanchez-Lengeling et al. 2020]



Quantifies the variability in explanation accuracy using the top 10% of models through a hyperparameter scan over model architectures

Drawback : How to be sure if the model used the intended explanation?

A few parting words

Explaining GraphML models is inherently tricky because of the complex interplay of structure and features in the decision making process

Several graph specific approaches are proposed with no clear winner

Evaluation of is inherently tricky in general but trickier for graphs because of additional structural explanations

A possible direction to investigate is the threat of explainability to data privacy

https://arxiv.org/abs/2207.10896, https://arxiv.org/abs/2206.14724

Hands-on-Session