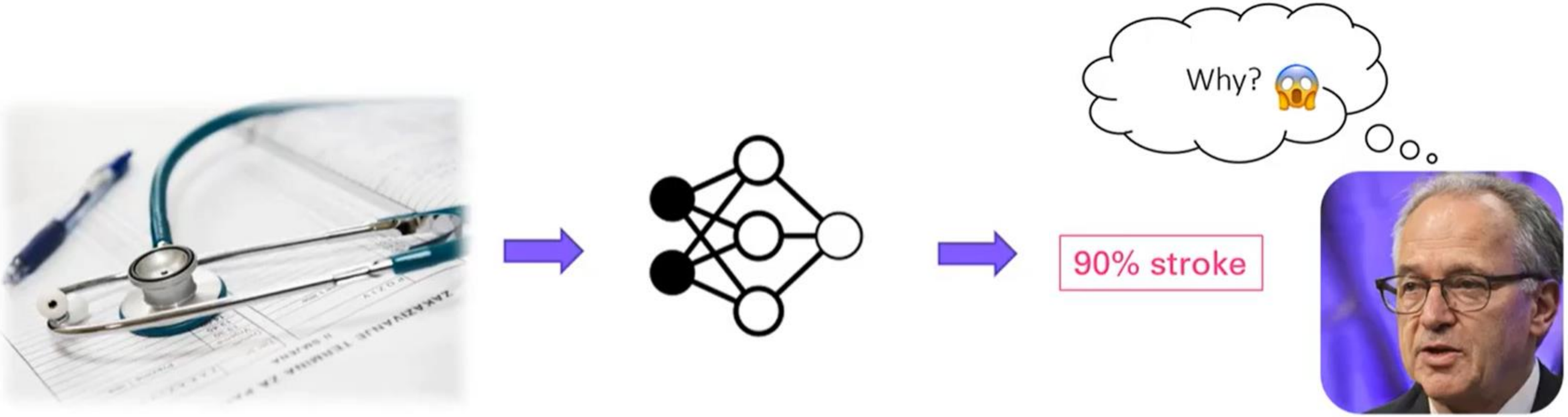
Interpretable AI & GNN Explainability

David Gu

SiDNN – 22.03.2022

Why care about Interpretability?



Stroke dataset

Black Box ML model

John

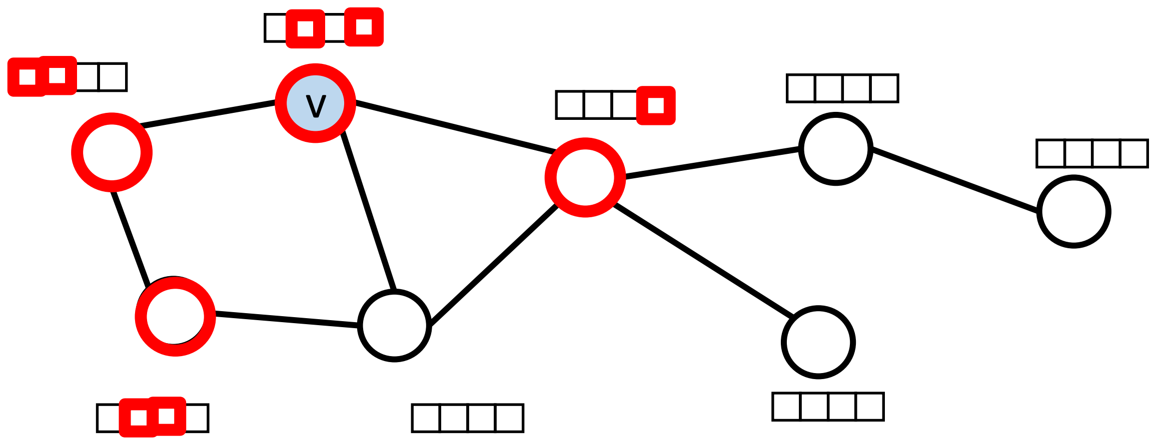
It‘s not always about predictive performance!

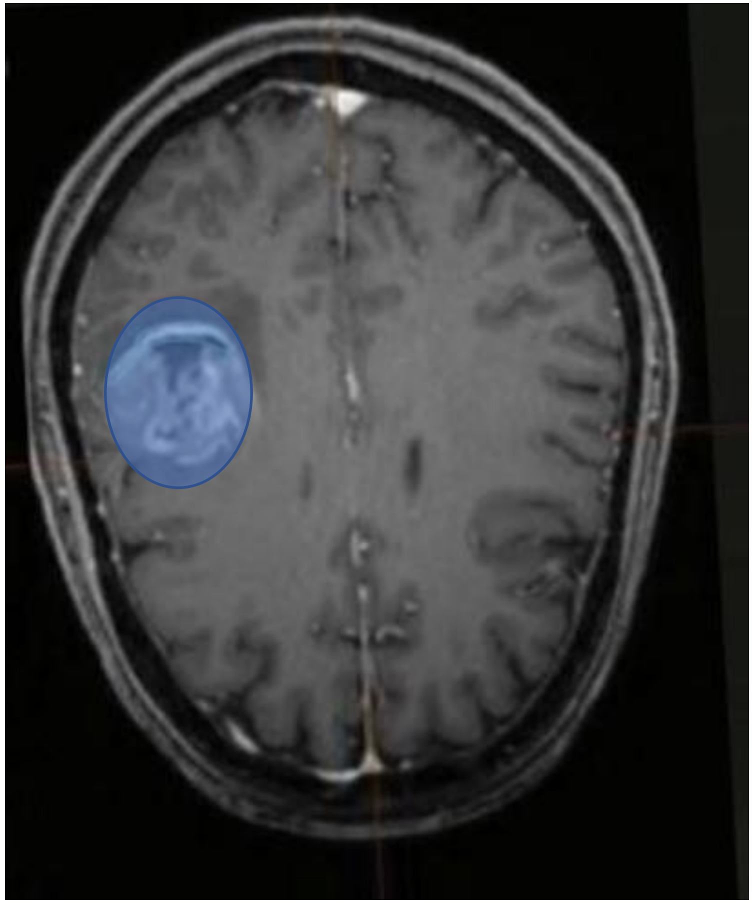
What we want:

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1

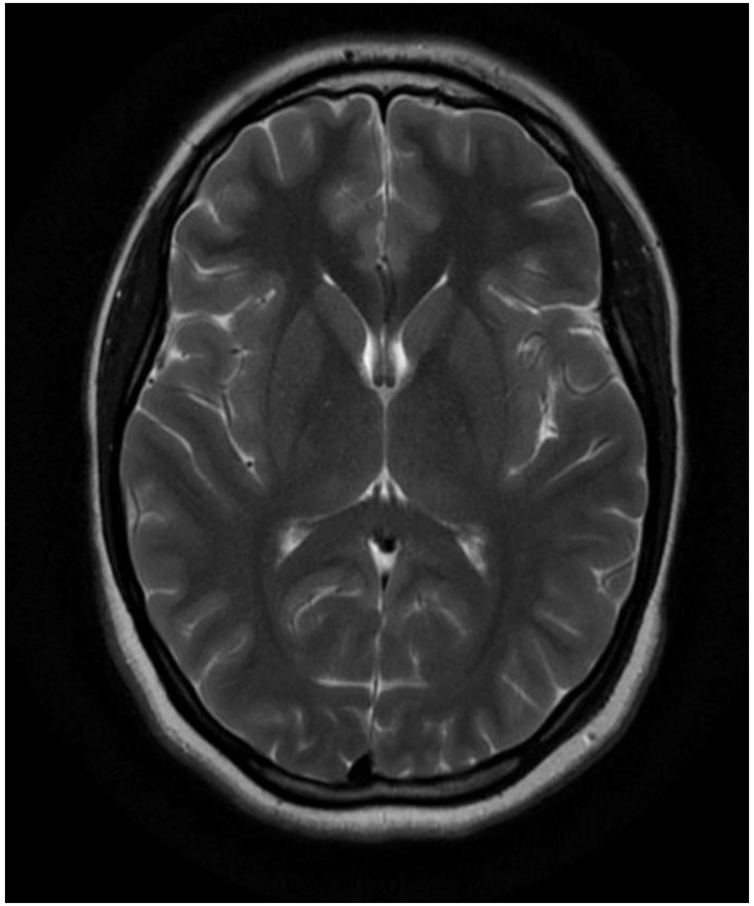
John‘s medical data

Today: Explaining = identifying important features!





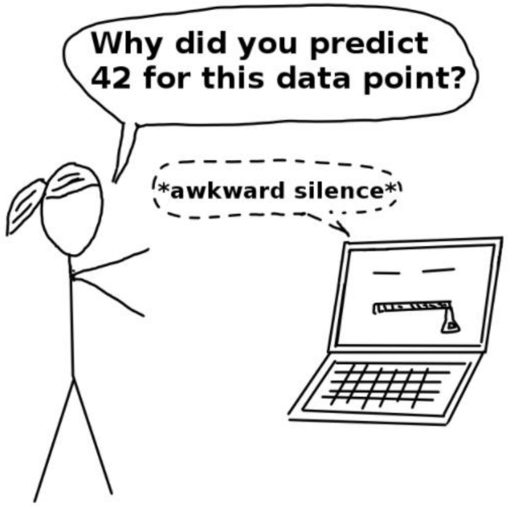
Brain with tumor



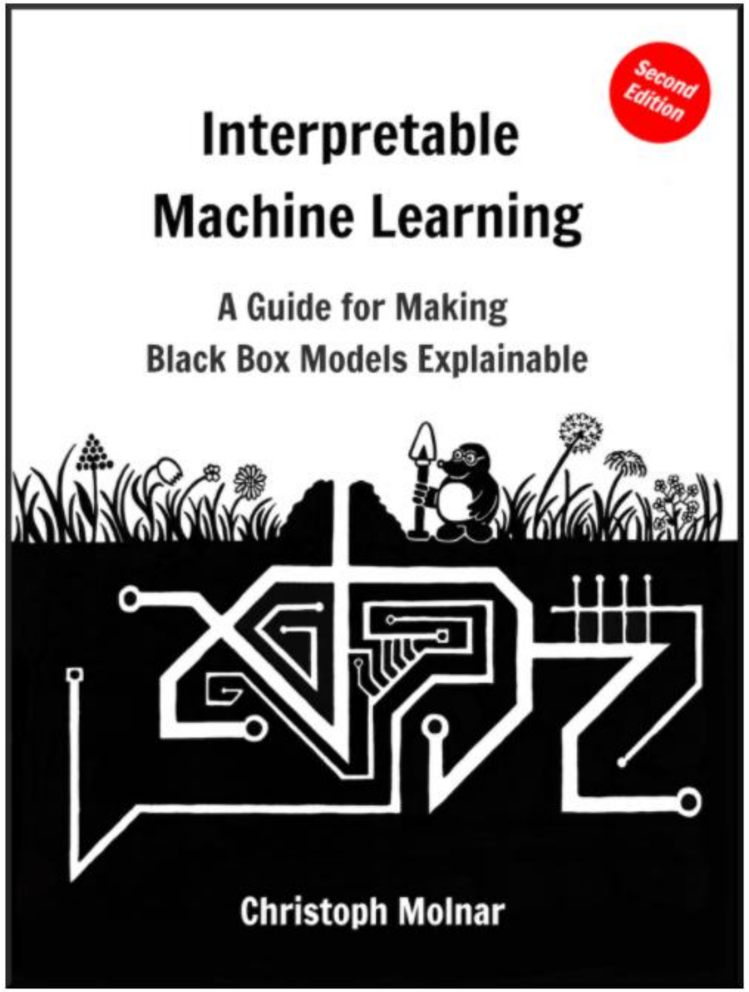
Brain without tumor

Why care about interpretability?

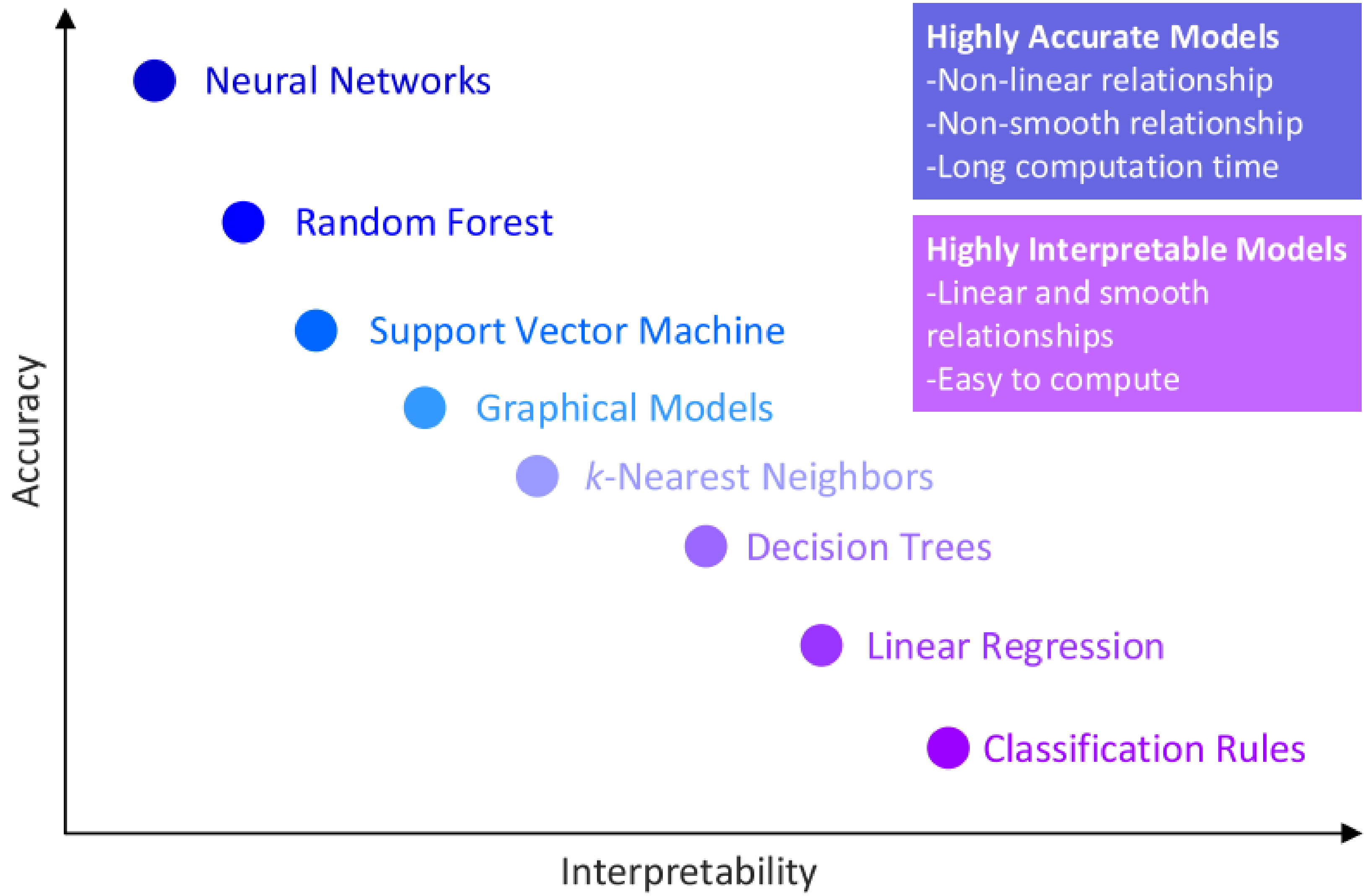
- Requirement by the end-user
- Model debugging
- Safety & Trust



Interpretable ML: Basics



https://christophm.github.io/interpretable-ml-book/



Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Interpretable ML: Basic Types

- Intrinsic interpretability vs. Post-hoc explanation methods
- Global vs. Local Interpretability
- Model-specific vs. Model-agnostic

Ex. 1: Permutation Feature Importance

	Date	Team	Opponent	Goal Scored	Ball Possession %	Attempts	On-Target	...	Man of the Match
0	14-06-2018	Russia	Saudi Arabia	5	40	13	7	...	Yes
1	14-06-2018	Saudi Arabia	Russia	0	60	6	0	...	No
2	15-06-2018	Egypt	Uruguay	0	43	8	3	...	No
3	15-06-2018	Uruguay	Egypt	1	57	14	4	...	Yes
4	15-06-2018	Morocco	Iran	0	64	13	3	...	No

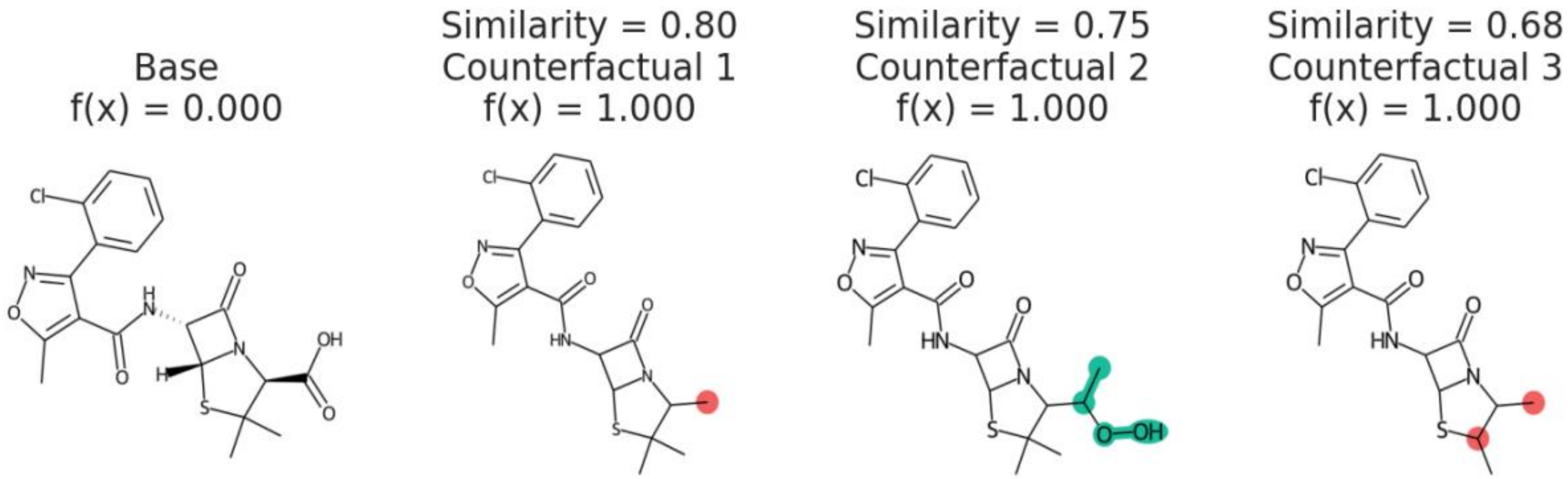
→ randomly permute!

Weight	Feature
0.1750 ± 0.0848	Goal Scored
0.0500 ± 0.0637	Distance Covered (Kms)
0.0437 ± 0.0637	Yellow Card
0.0187 ± 0.0500	Off-Target
0.0187 ± 0.0637	Free Kicks
0.0187 ± 0.0637	Fouls Committed
0.0125 ± 0.0637	Pass Accuracy %
0.0125 ± 0.0306	Blocked
0.0063 ± 0.0612	Saves
0.0063 ± 0.0250	Ball Possession %
0 ± 0.0000	Red
0 ± 0.0000	Yellow & Red
0.0000 ± 0.0559	On-Target
-0.0063 ± 0.0729	Offsides
-0.0063 ± 0.0919	Corners
-0.0063 ± 0.0250	Goals in PSO
-0.0187 ± 0.0306	Attempts
-0.0500 ± 0.0637	Passes

Ex. 2: Counterfactual Explanation

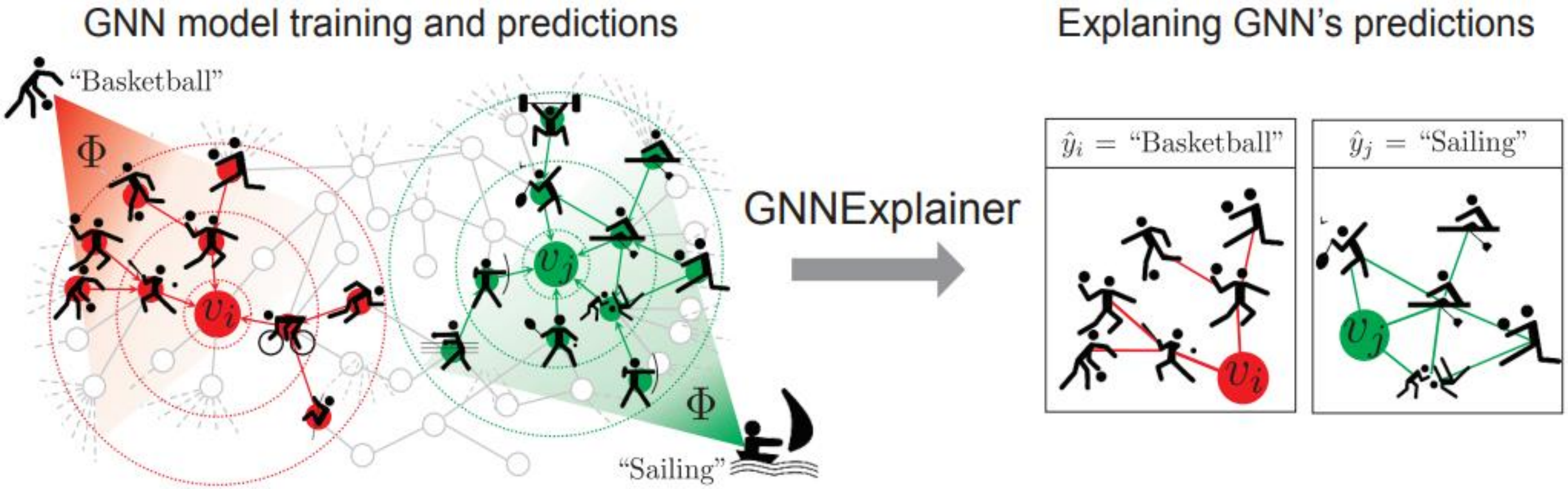
- „If X hadn‘t occured, Y hadn‘t occured.“
- Ex.: „If I hadn‘t partied all night, I wouldn‘t be hungover.“

Ex.: Graph classification task (Blood-Brain Barrier Permeation Prediction)



CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks

GNNExplainer



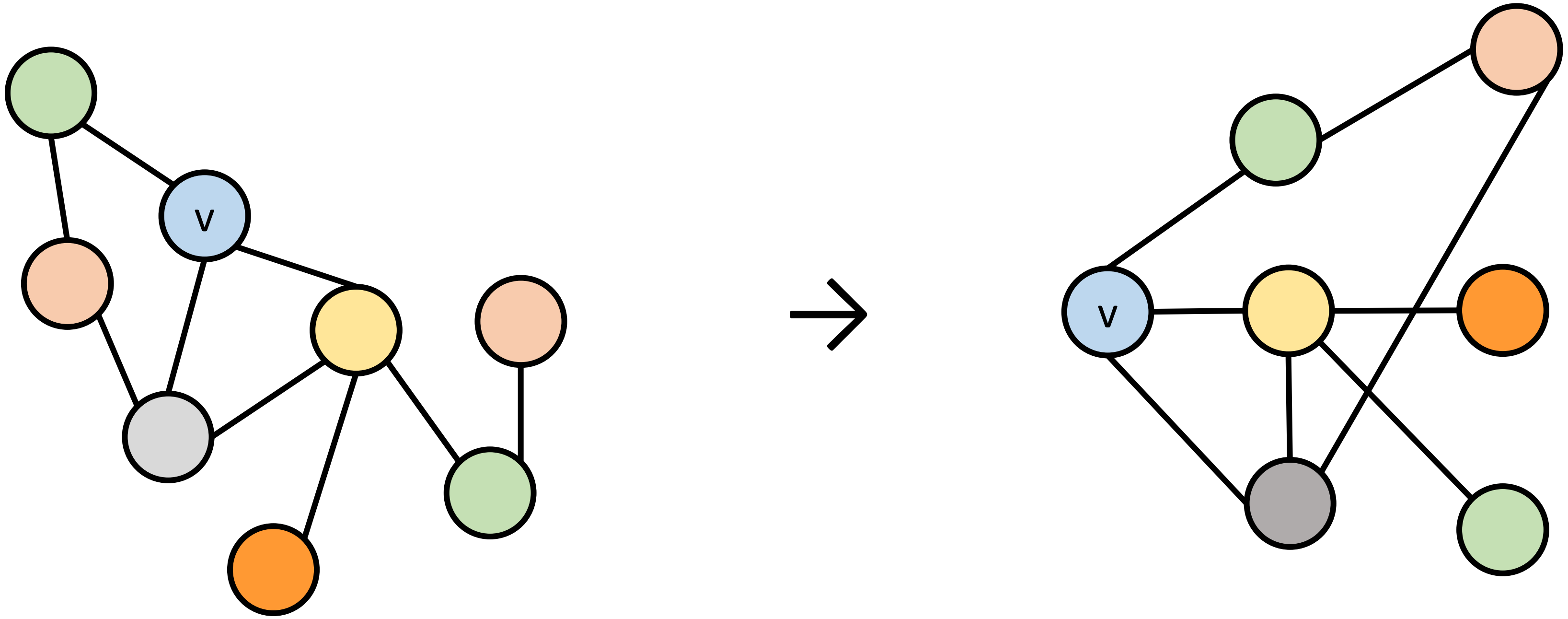
GNNExplainer: Generating Explanations for Graph Neural Networks

GNNExplainer

- Assume for now: node classification!

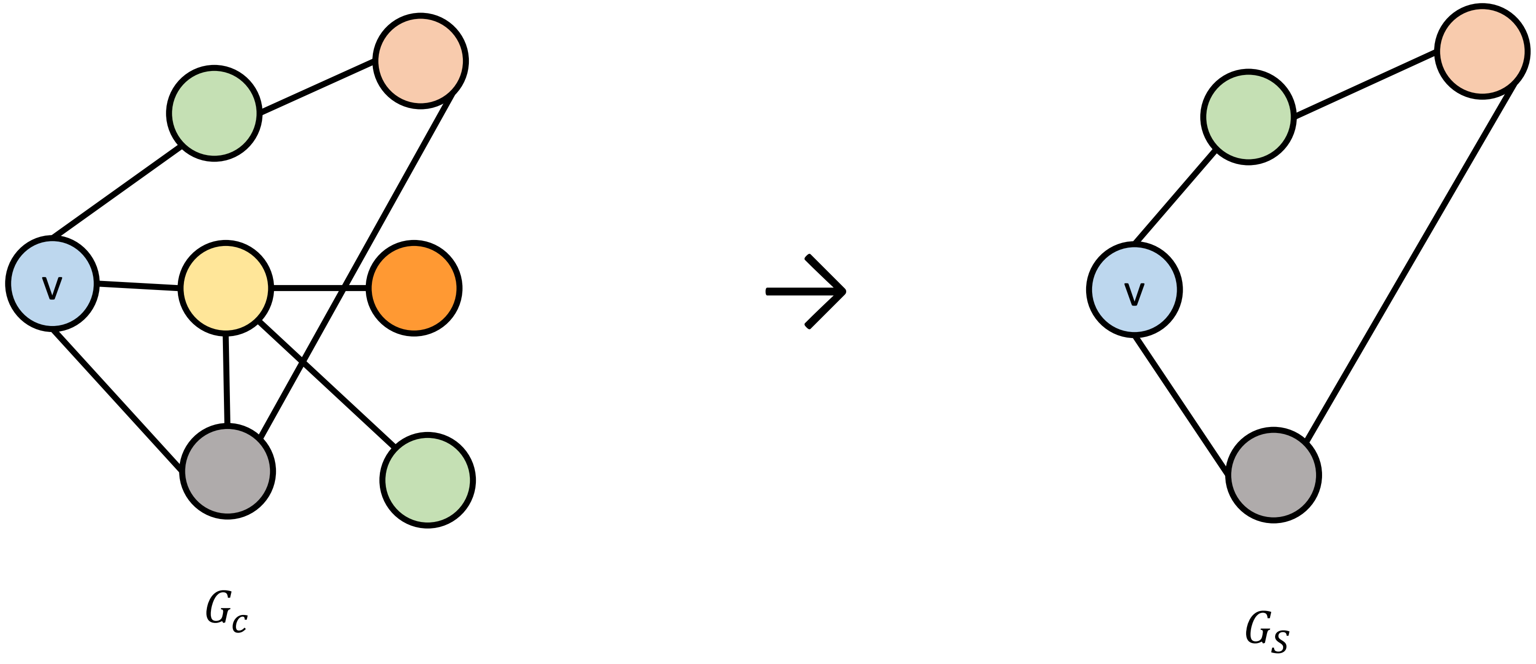
- GOAL: Identify **small** subgraph and associated features that **are important** for the GNN‘s prediction $\hat{y}$!

Computation graph:



$G_c(v)$

Intuition: Remove subset of nodes…



..if the prediction of the GNN changes, then the removed nodes are a good counterfactual explanation!

Mathematical Formalization

- GOAL: Choose subgraph G_S s.t. the mutual information between the prediction of the GNN using G_C and G_S and features X_S is maximized!

$$\max_{G_S} MI\left(Y, (G_S, X_S)\right) = H(Y) - H(Y|G = G_S, X = X_S)$$

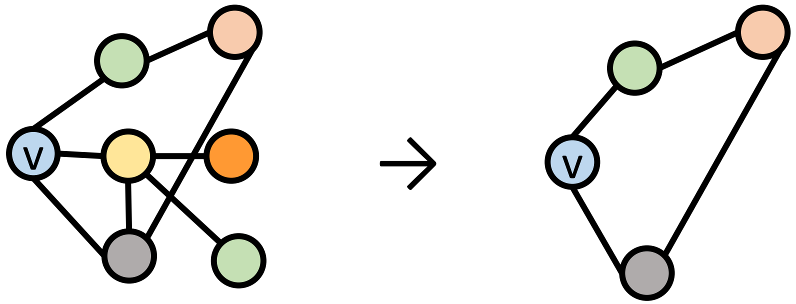
or equivalently minimize

$$H(Y|G{=}G_S, X{=}X_S) = -\mathbb{E}_{Y|G_S,X_S}\left[\log P_\Phi(Y|G{=}G_S, X{=}X_S)\right]$$

- Challenge: Exponentially many subsets G_S!

Continuous relaxation

- Idea: For tractability, learn mask M on the adjacency matrix of G_C



G_C G_S

	-	1	1	1	0	0	0
	1	-	0	0	1	0	0
	1	0	-	0	0	1	1
	1	0	0	-	1	0	0
	0	0	0	1	-	0	0
	0	0	1	0	0	-	0
	0	0	1	0	0	0	-

A_C

	-	7.2	0.3	5.2	0.5	1.2	0.9
	7.2	-	0.1	0.6	4.3	0.8	1.1
	0.3	0.1	-	0.7	0.6	0.0	0.1
	5.2	0.6	0.7	-	8.1	0.9	0.6
	0.5	4.3	0.6	8.1	-	0.2	0.8
	1.2	0.8	0.0	0.9	0.2	-	1.0
	0.9	1.1	0.1	0.6	0.8	1.0	-

M

Mathematical Formalization

- Learn mask M on the adjacency matrix of G_C that minimizes

$$\min_{M} - \sum_{c=1}^{C} \mathbb{1}[y=c] \log P_{\Phi}(Y=y|G=A_c \odot \sigma(M), X=X_c)$$

earlier: G_S

- Optimize objective via gradient descent!

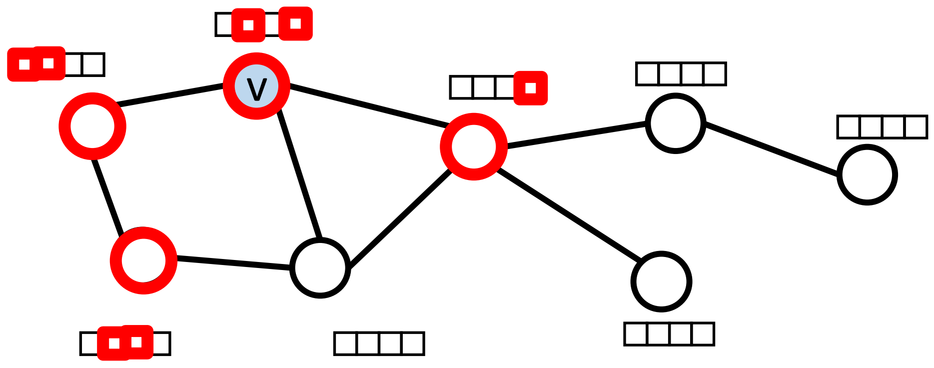
Feature selection

So far:

$$\min_M - \sum_{c=1}^{C} \mathbb{1}[y = c] \log P_\Phi(Y = y | G = A_c \odot \sigma(M), X = X_c)$$

What about this?

Apply same idea to learn optimal subset of the features via mask X_S^F!



Feature selection

Optimize jointly via gradient descent:

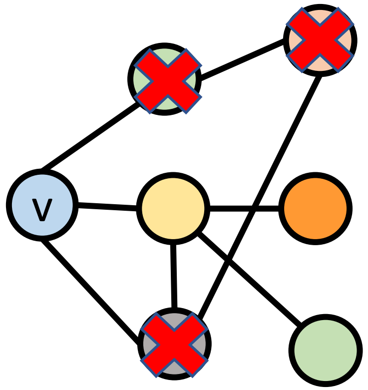
$$\max_{G_S, F} MI\left(Y, (G_S, F)\right) = H(Y) - H(Y|G = G_S, X = X_S^F)$$

Q: What is missing in this objective?

(Hint: How is this objective trivially maximized?)

- Regularization:
  - Mask size: Penalize <u>high explanation size</u> by adding sum of all mask parameters
  - Entropy of the parameters: Explanation should be discriminative

- Constraint:
  - Output should be a connected subgraph!

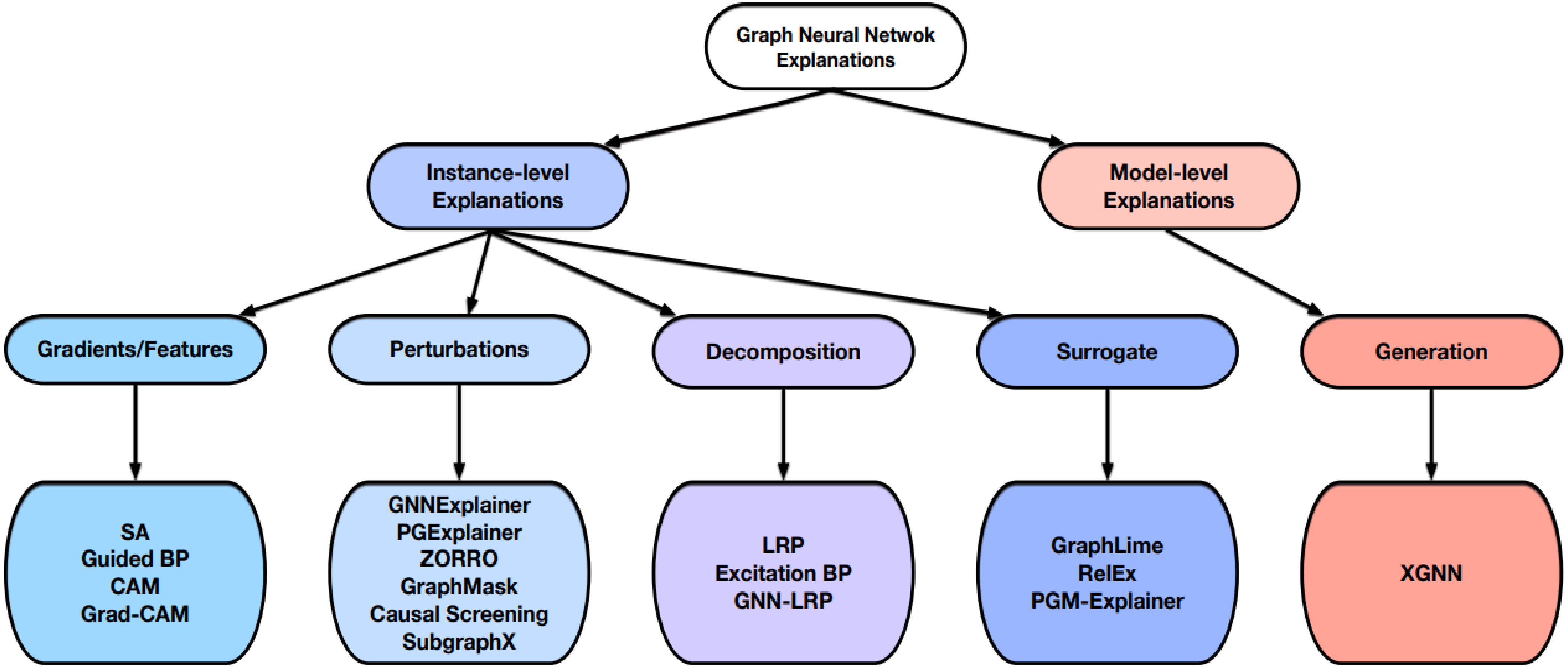


G_S

Extensions

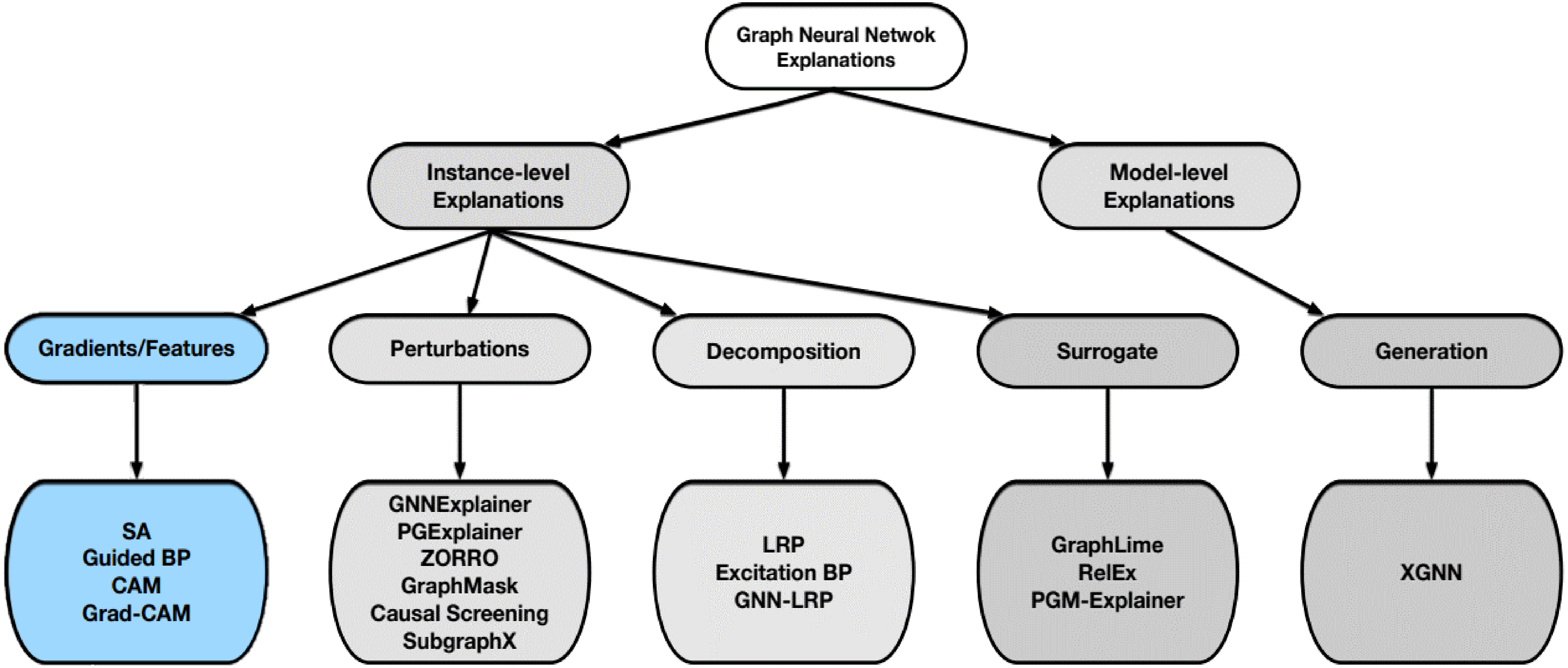
- Link prediction: Learn two masks explaining both endpoints of the link
- Multi-instance explanation: aggregate explanations of nodes to a class c to get a „typical explanation“

Taxonomy of GNN Explainability

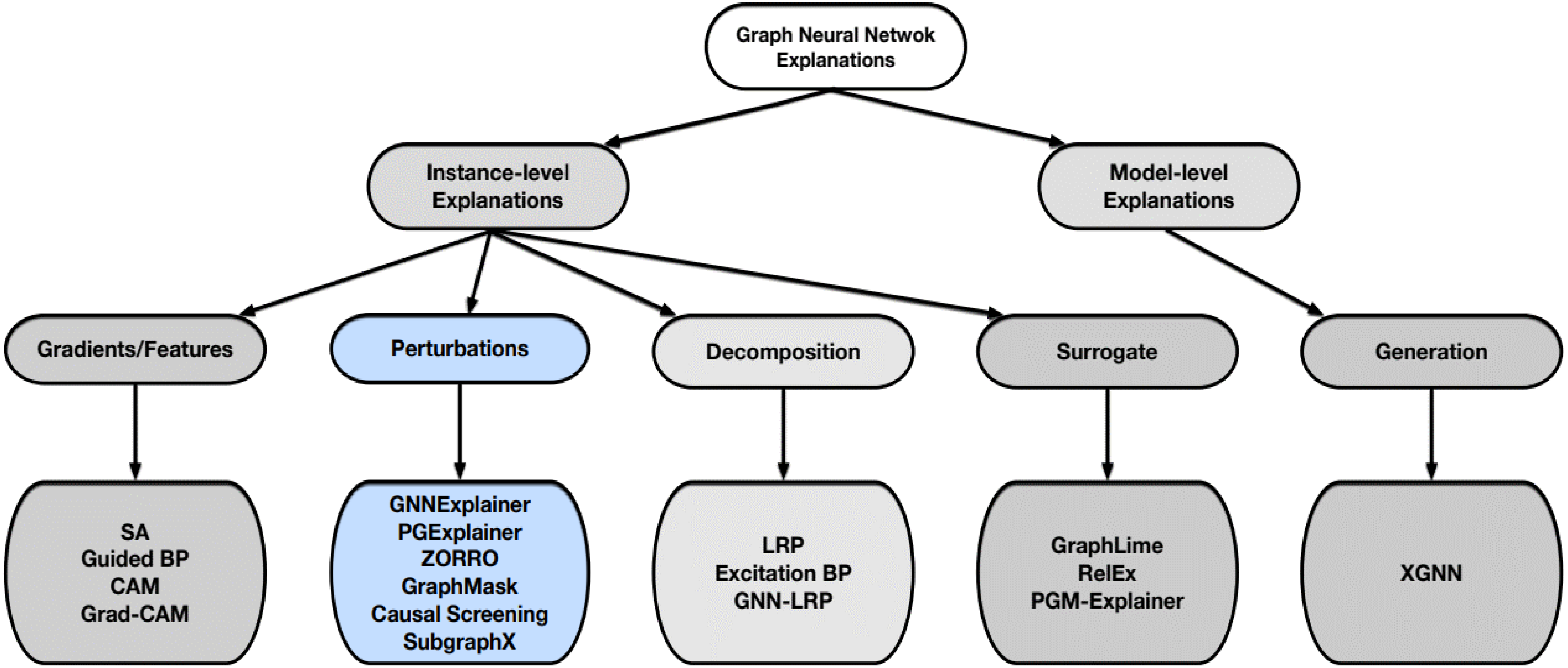


Explainability in Graph Neural Networks: A Taxonomic Survey

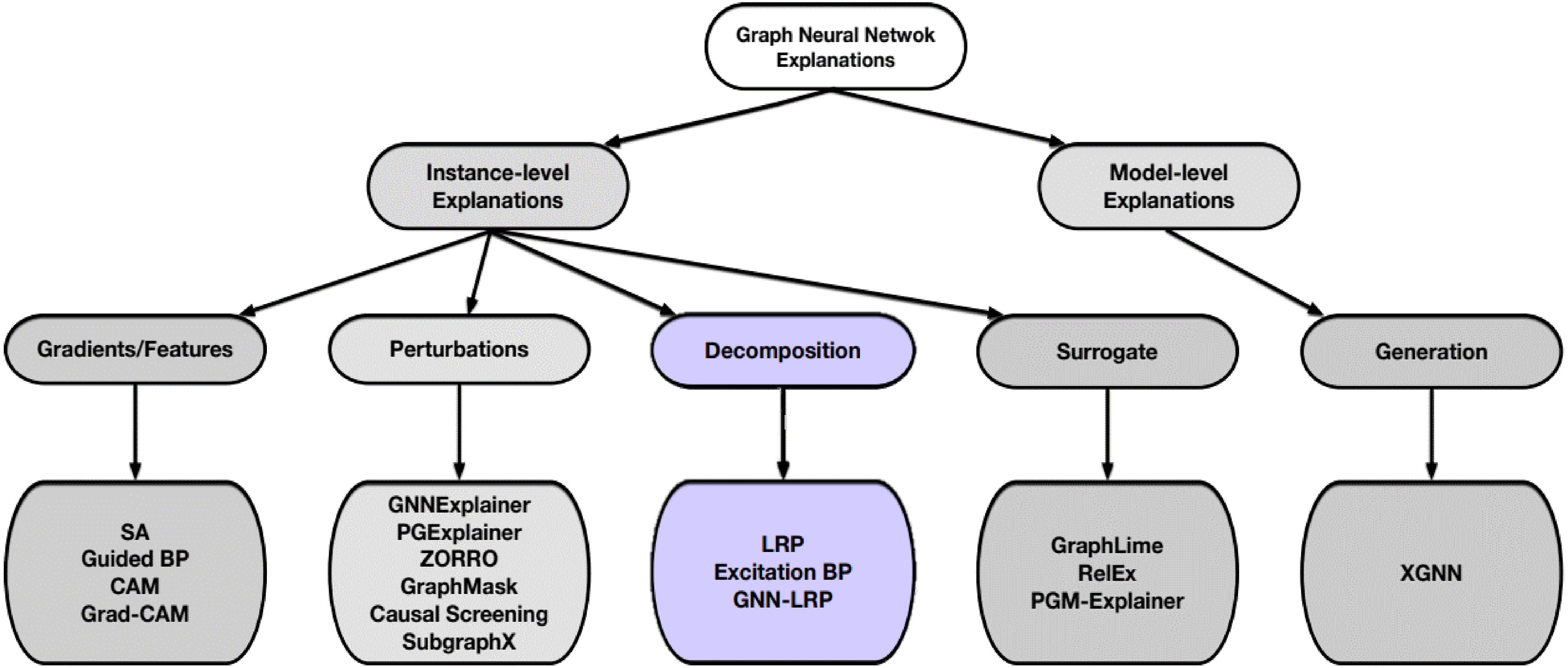
Taxonomy of GNN Explainability



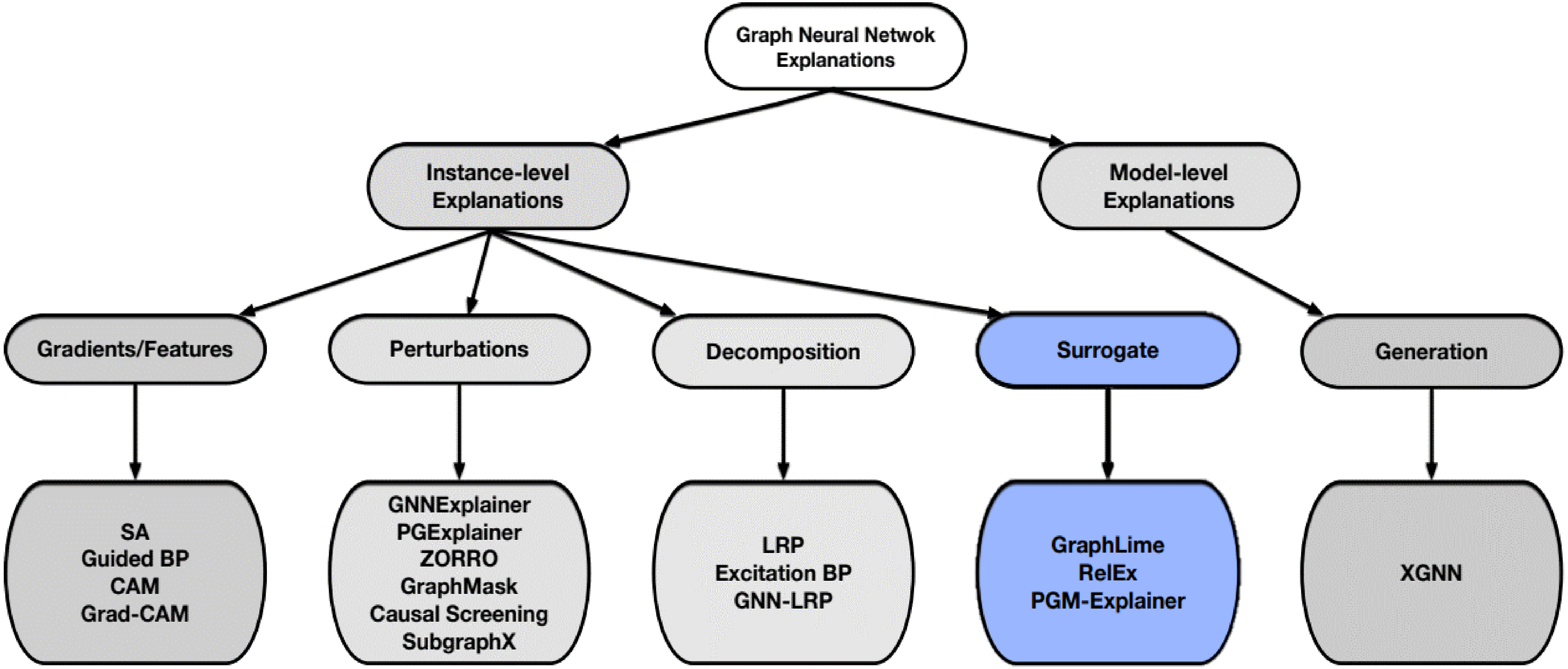
Taxonomy of GNN Explainability



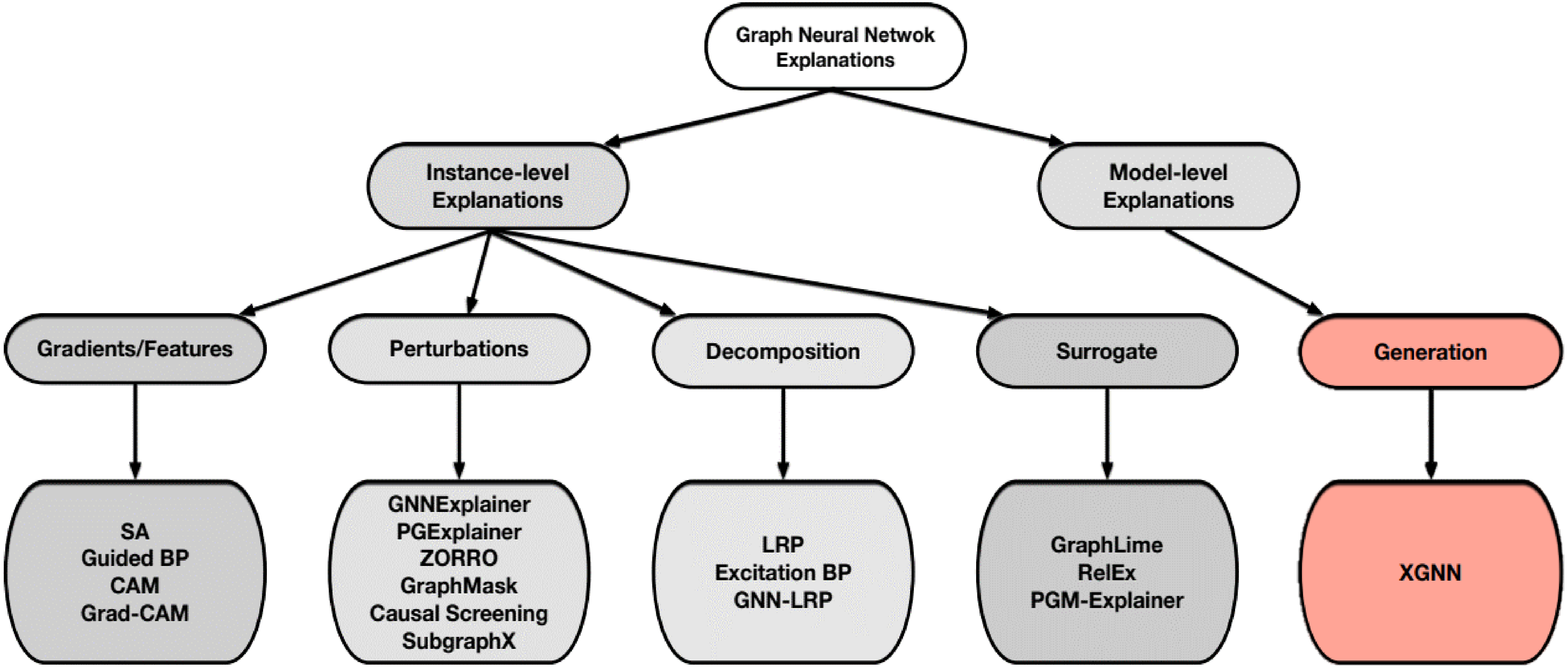
Taxonomy of GNN Explainability



Taxonomy of GNN Explainability



Taxonomy of GNN Explainability



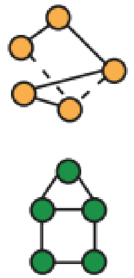
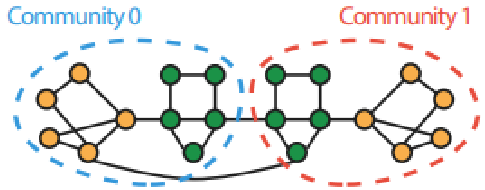
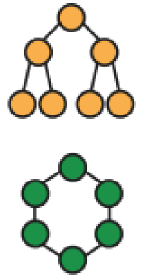
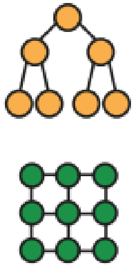
References

- Motivation:
  - Deepfindr: https://www.youtube.com/watch?v=NvDM2j8Jgvk
  - Stroke dataset: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
  - Brain MRI dataset: https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection
  - Self-driving car meme: https://m.facebook.com/TrolleyProblemMemes/photos/a.250373635311569.1073741827.250353181980281/353949958287269?locale=ar_AR&_rdr
  - Awkward silence meme: from book below
- Interpretable ML Book: https://christophm.github.io/interpretable-ml-book/
- Chart: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

References

- FIFA World Cup dataset: https://www.kaggle.com/mathan/fifa-2018-match-statistics
- Molecule example/CF-GNNExplainer: https://arxiv.org/abs/2102.03322
- GNNExplainer: https://arxiv.org/abs/1903.03894
- GNN Explainability Taxonomy: https://arxiv.org/pdf/2012.15445.pdf

Backup slide: GNNExplainer results

	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid
Base **Motif**				
Node Features	None	$\mathcal{N}(\mu_l, \sigma_l)$ where l = community ID	None	None
Explanation content	Graph structure	Graph structure Node feature information	Graph structure	Graph structure
Explanation accuracy				
Att	0.815	0.739	0.824	0.612
Grad	0.882	0.750	0.905	0.667
GNNExplainer	**0.925**	**0.836**	**0.948**	**0.875**