Explaining the Explainers in Graph Neural Networks: a Comparative Study

Antonio Longa[1,2], Steave Azzolin[1], Gabriele Santin[2], Giulia Cencetti[2], Pietro Lio[3], Bruno Lepri[2], Andrea Passerini[1]

SML[1] Lab, University of Trento, Italy
MobS[2] Lab, Fondazione Bruno Kessler,Trento, Italy
Cambridge[3] University, Cambridge, UK

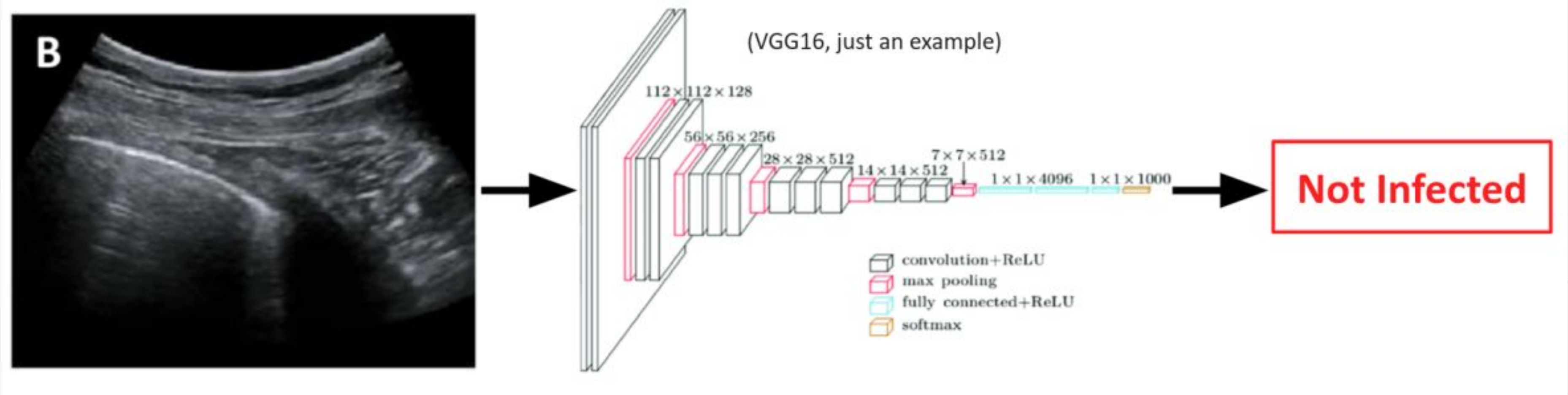
TABLE OF CONTENTS

01 Explainability

You need to be checked for COVID-19. The doctor takes a scan of your lungs and uses a state-of-the-art deep neural network to automatically compute a diagnosis. The model thinks that you are not infected.

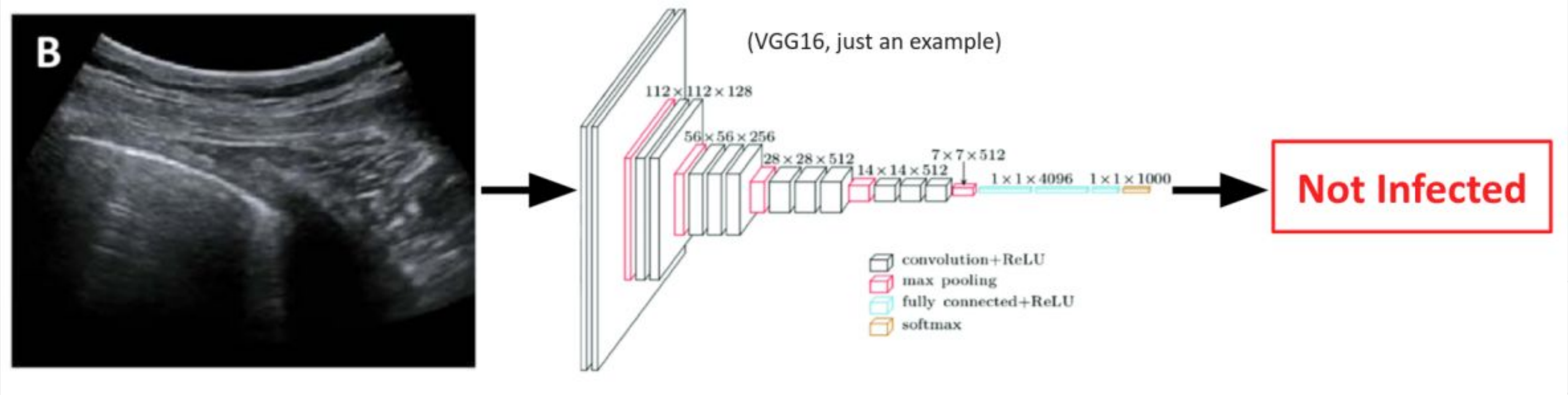
01 Explainability

You need to be checked for COVID-19. The doctor takes a scan of your lungs and uses a state-of-the-art deep neural network to automatically compute a diagnosis. The model thinks that you are not infected.



01 Explainability

You need to be checked for COVID-19. The doctor takes a scan of your lungs and uses a state-of-the-art deep neural network to automatically compute a diagnosis. The model thinks that you are not infected.



Question: Would you trust the model's prediction?

Thanks to Stefano Teso for slides

01 Explainability

People are finding more and more ways of integrating machine learning models into applications.

01 Explainability

People are finding more and more ways of integrating machine learning models into applications.

- Medical Diagnosis
- Crime (e.g., predicting recidivism in convicts)
- Credit Scoring (e.g., approving loan requests)
- Surveillance (e.g., face recognition, profiling)
- Hiring (e.g., ranking/filtering candidates)
- …

Thanks to Stefano Teso for slides

01 Explainability

People are finding more and more ways of integrating machine learning models into applications.

- Medical Diagnosis
- Crime (e.g., predicting recidivism in convicts)
- Credit Scoring (e.g., approving loan requests)
- Surveillance (e.g., face recognition, profiling)
- Hiring (e.g., ranking/filtering candidates)
- …

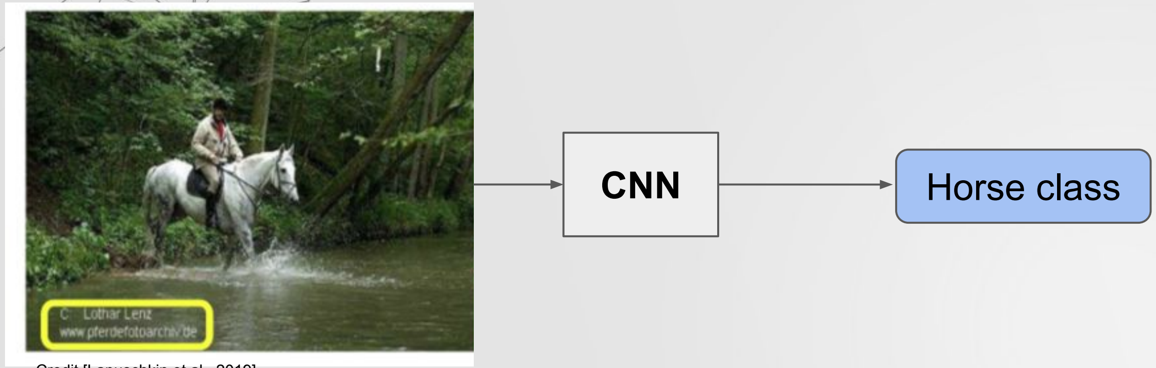
Right of explanation:

Example: you apply for a 50, 000 euro loan.
Unfortunately, your bank rejects your application.
You have a right to know why it was rejected: was it your credit history or your age/gender/ethnicity?

See https://en.wikipedia.org/wiki/Right_to_explanation

01 Explainability

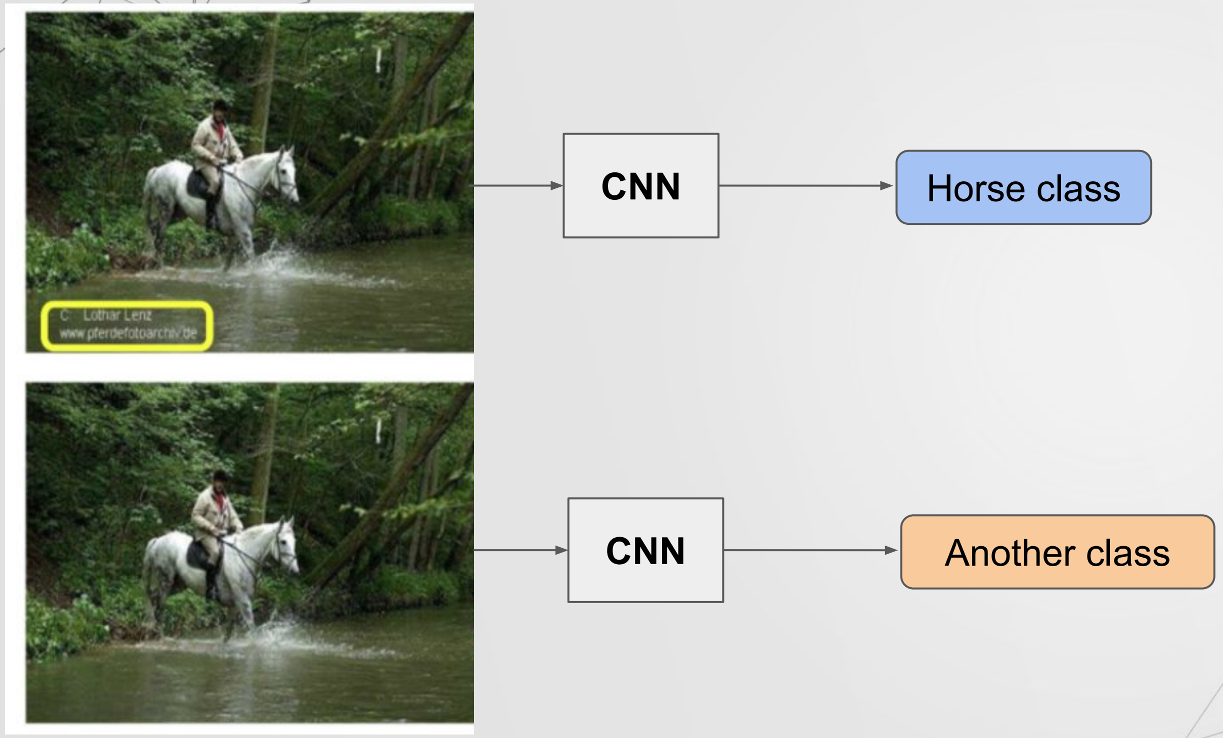
Horse-picture from Pascal VOC data set



Credit [Lapuschkin et al., 2019]

01 Explainability

Horse-picture from Pascal VOC data set

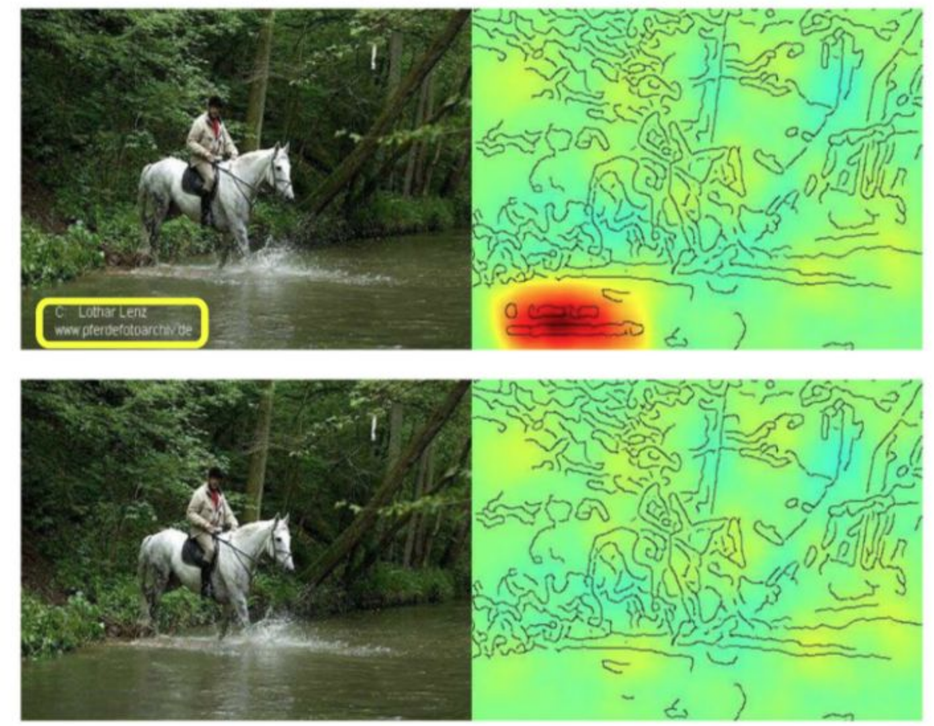


Credit [Lapuschkin et al., 2019]

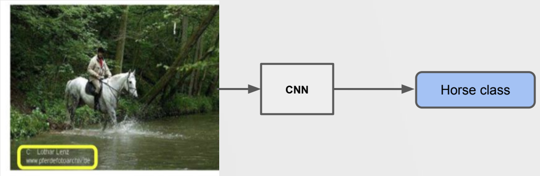
Thanks to Stefano Teso for slides

01 Explainability

Horse-picture from Pascal VOC data set



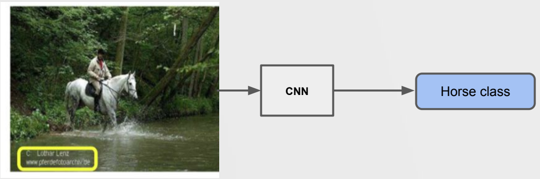
Credit [Lapuschkin et al., 2019]

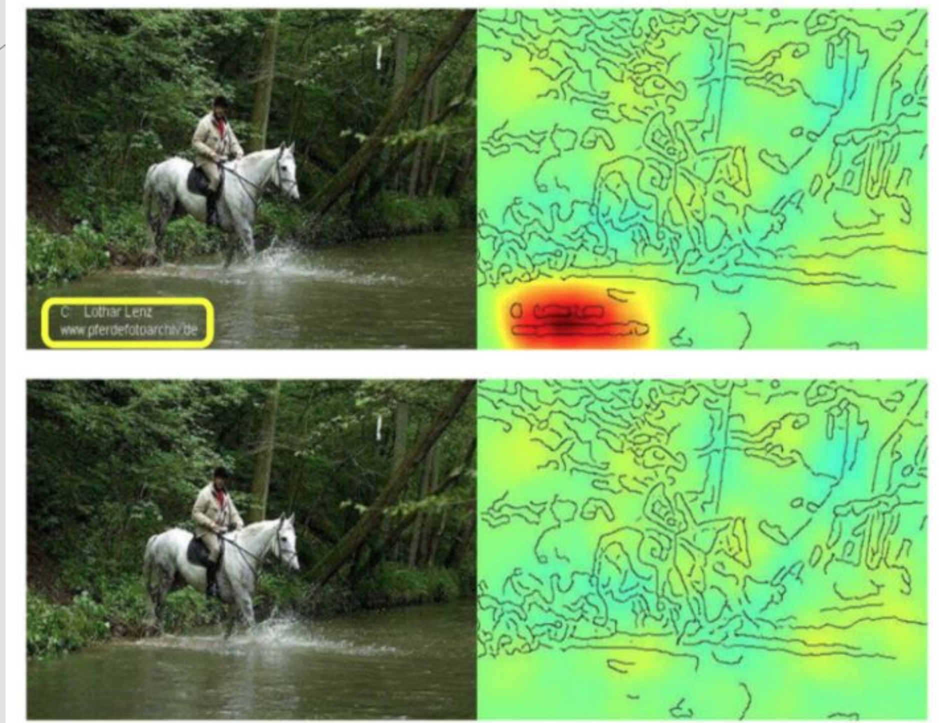


Thanks to Stefano Teso for slides

01 Explainability

Horse-picture from Pascal VOC data set





Credit [Lapuschkin et al., 2019]

Correlation between the presence of a watermark when an horse is present.

01 Explainability

Explanations are studied in epistemology & philosophy of science. There are many **incompatible** but **complementary** schools of thought:

Table 1: Philosophical Theories of Explanation

	Theory	**Explananda** *(things to be explained)*	**Explanantia** *(things doing the explaining)*
Logical	Deductive-Nomological	Observed phenomenon or pattern of phenomena	Laws of nature, empirical observations, and deductive syllogistic pattern of reasoning
	Unification	Observed phenomenon or pattern of phenomena	Logical argument class
Causal	Transmission	Observed output of causal process	Observed or inferred trace of causal process
	Interventionist	Variables representing output of causal process	Variables representing input of causal process and invariant pattern of counterfactual dependence between variables
Functional	Pragmatic	Answers to why-questions	True propositions defined by their relevance relation to the explanandum they explain and the contrast class against which the demand for explanation is made
	Psychological	Observed phenomenon or pattern of phenomena	True propositions defined by their relation to the user's knowledge base and to the explanandum

01 Explainability

Take-away:

- We need explainability!

01 Explainability

Take-away:

- We need explainability!
- No unique definition of explanation, even in philosophy

01 Explainability

Take-away:

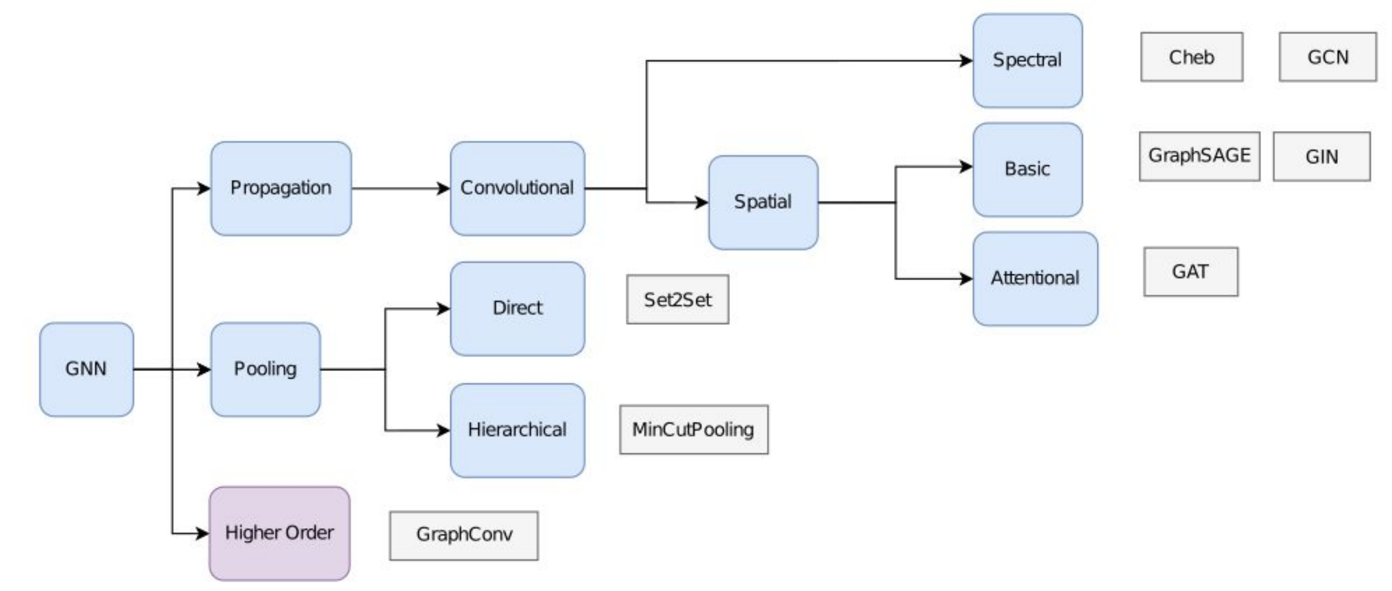
- We need explainability!
- No unique definition of explanation, even in philosophy
- Explaining machine learning models is still an open research question

02 Graph Neural Networks

GNN are well know to you.

02 Graph Neural Networks

Which network do we test?



Zhou et al. Graph neural networks: A review of methods and applications

An overview of the adopted GNN architectures structured in a taxonomy as defined by Zhou et al.

Blue boxes → Zhou et al.
Pink box → Our extension

02 Graph Neural Networks

Dataset	Architecture	GNN	Fully conn.	HyperParams	LR	Epochs	Train Acc	Test Acc
Grid	Gcn	30-30-30	10-2	-	0.001	1500	0.994	0.998
	GraphSage	30-30-30	10-2	-	0.01	3000	X	X
	Gat	30-30-30	10-2	heads = 1	0.01	3000	X	X
	Gin	30-30	30-2	-	0.001	1000	1.0	1.0
	Cheb	30-30	30-2	-	0.001	1000	1.0	1.0
	MinCutPool	32-32-32	32-2	-	0.001	700	0.92	0.93
	Set2Set	30-30-30	10-2	-	0.001	1500	0.97	0.97
	GraphConv	30-30	30-2	-	0.001	500	1.0	1.0
Grid-House	Gcn	60-60-60-60	60-10-2	-	0.001	7000	0.97	0.97
	GraphSage	60-60-60-60	60-10-2	-	0.01	3000	X	X
	Gat	60-60-60-60	60-10-2	heads = 3	0.01	3000	X	X
	Gin	30-30	30-2	-	0.001	1000	0.99	1.0
	Cheb	30-30-30	30-2	-	0.001	1000	1.0	0.98
	MinCutPool	32-32-32	32-2	-	0.001	700	0.95	0.95
	Set2Set	60-60-60-60	60-10-2	-	0.001	1500	0.97	0.97
	GraphConv	30-30	30-2	-	0.001	500	1.0	1.0
Stars	Gcn	70-70-70	30-3	-	0.005	1000	0.99	1.0
	GraphSage	30-30-30	30-3	-	0.01	3000	X	X
	Gat	30-30-30	10-3	heads = 1	0.01	3000	X	X
	Gin	40-40	30-3	-	0.001	3000	0.99	1.0
	Cheb	30-30	30-3	-	0.001	1000	0.99	0.99
	MinCutPool	32-32-32	32-3	-	0.001	400	0.99	0.99
	Set2Set	70-70-70	30-3	-	0.001	1500	0.99	0.99
	GraphConv	30-30	30-3	-	0.001	500	0.99	0.99

Mean agg

Mean agg

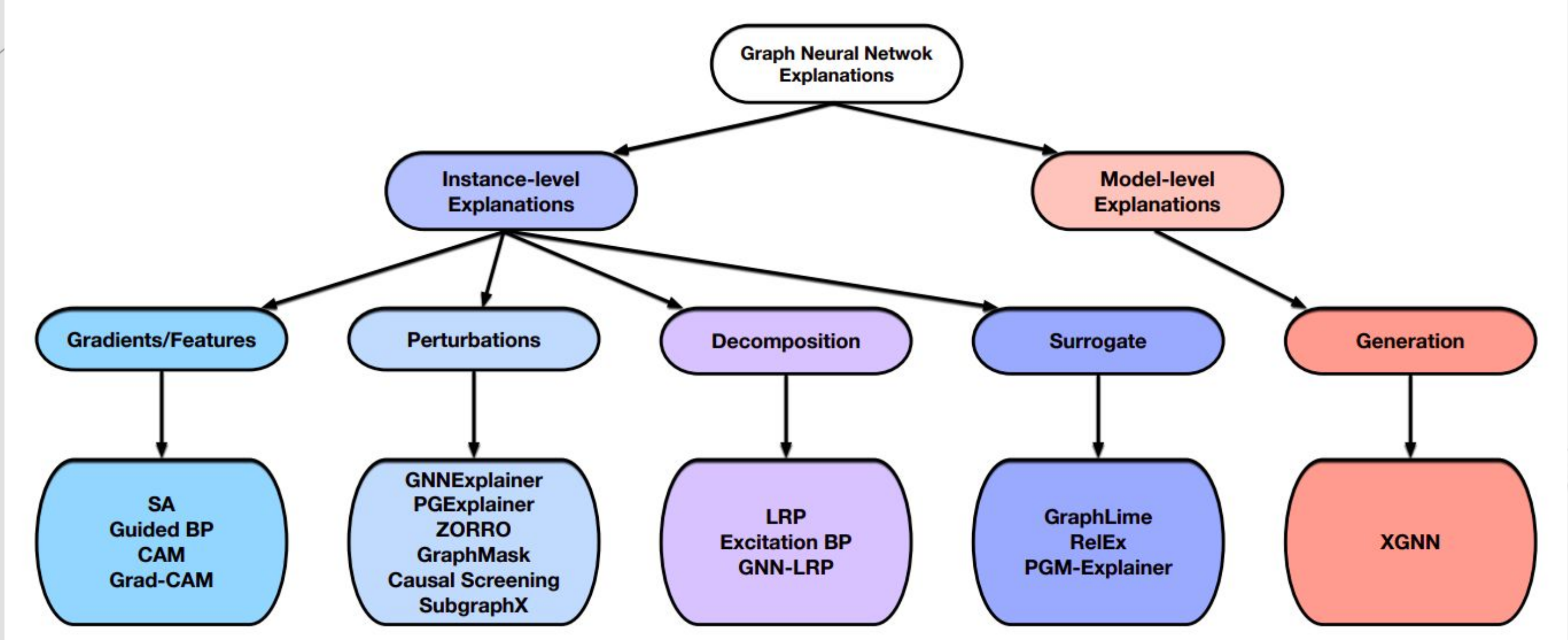
Sum agg

03 GNN explainers

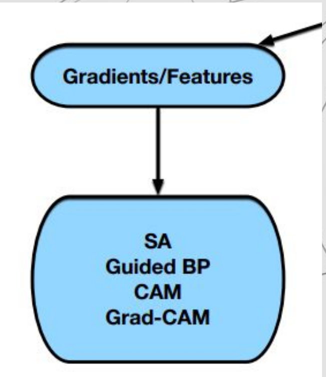
Many GNN explainers have been proposed.

03 GNN explainers

Many GNN explainers have been proposed.
We use Yuan taxonomy



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

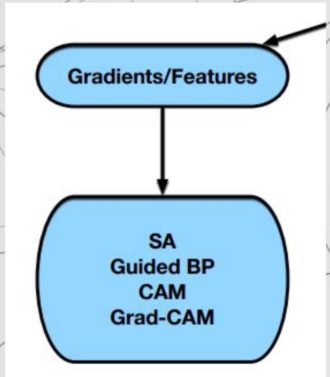


Yuan et al. Explainability in Graph Neural Networks: A Taxonomic Survey

03 GNN explainers

Gradient/Feature based:

- They uses gradients to explain the GNN.
- Widely used in image and text.
- Use the gradients as the approximations of input importance.



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

03 GNN explainers

Gradient/Feature based:

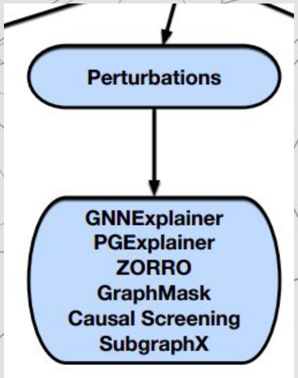
- They uses gradients to explain the GNN.
- Widely used in image and text.
- Use the gradients as the approximations of input importance.

What we use:

- GradExplNode [1] → Node importance mask
- GuidedBP [2] → Node importance mask
- IGNode [3] → Node importance mask
- CAM [4] → Node importance mask
- GradCAM [4] → Node importance mask
- GradExplEdge [1] → Edge importance mask
- IGEdge [3] → Edge importance mask

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.
[2] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014.
[3] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International conference on machine learning, pages 3319–3328. PMLR, 2017.
[4] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10772–10781, 2019.

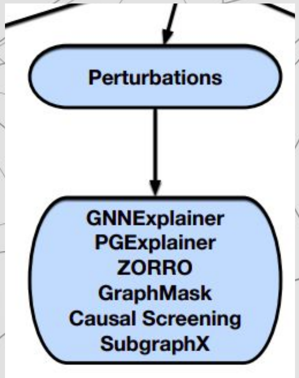
03 GNN explainers



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

Perturbation based:

- Study the output variations with respect to different input perturbations
- Widely used in image and text.
- Key idea → perturb important input information should impact the prediction



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

03 GNN explainers

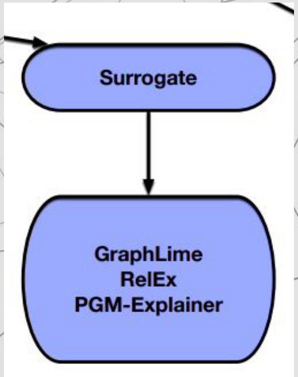
Perturbation based:

- Study the output variations with respect to different input perturbations
- Widely used in image and text.
- Key idea → perturb important input information should impact the prediction

What we use:

- GNNexplainer [5] → Edge importance mask
- PGExplainer [6] → Edge importance mask

[5] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. Advances in neural information processing systems, 32, 2019.
[6] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. Advances in neural information processing systems, 33:19620–19631, 2020.



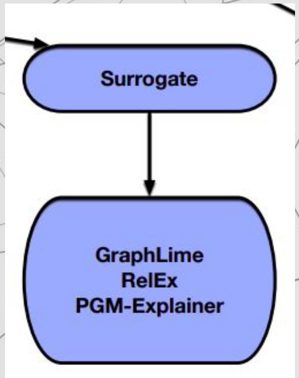
Yuan et al. Explainability in Graph Neural Networks: A Taxonomic Survey

03 GNN explainers

Perturbation based:

- Use a surrogate interpretable model to approximate the prediction.

03 GNN explainers



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

Perturbation based:

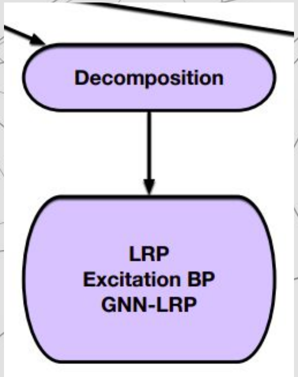
- Use a surrogate interpretable model to approximate the prediction.

What we use:

- PGM-Explainer [7] → Node importance mask

[5] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. Advances in neural information processing systems, 32, 2019.
[6] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. Advances in neural information processing systems, 33:19620–19631, 2020.

03 GNN explainers

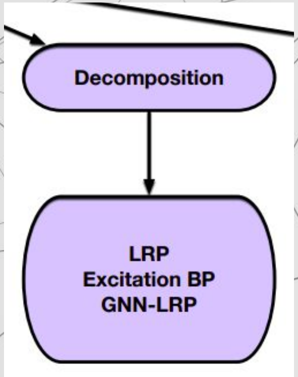


Yuan et al. Explainability in Graph Neural Networks: A Taxonomic Survey

Decomposition based:

- Decompose the original model prediction into several terms.
- Study the importance of those terms wrt the input feature

03 GNN explainers



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

Decomposition based:

- Decompose the original model prediction into several terms.
- Study the importance of those terms wrt the input feature

Model-level- based:

- XGNN [8]
- GLGExplainer [9]

[8] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. Xgnn: Towards model-level explanations of graph neural networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 430–438, 2020.
[9] Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Liò, and Andrea Passerini. Global explainability of gnns via logic combination of learned concepts, 2022

04 Benchmark datasets

Graph Classification:

- Grid
- Grid-House
- Stars
- House-Color

04 Benchmark datasets

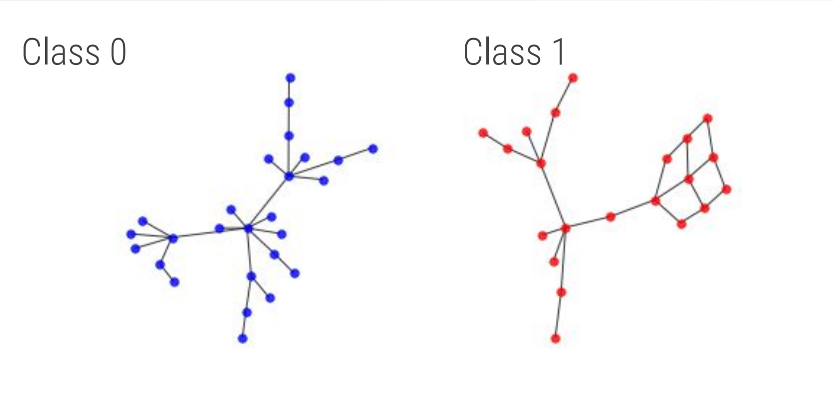
Grid:

- Binary graph classification
- Classes:
  - 0 → BA random graph
  - 1 → BA random graph + 3x3 grid graph

04 Benchmark datasets

Grid:

- Binary graph classification
- Classes:
  - 0 ⟶ BA random graph
  - 1 ⟶ BA random graph + 3x3 grid graph



04 Benchmark datasets

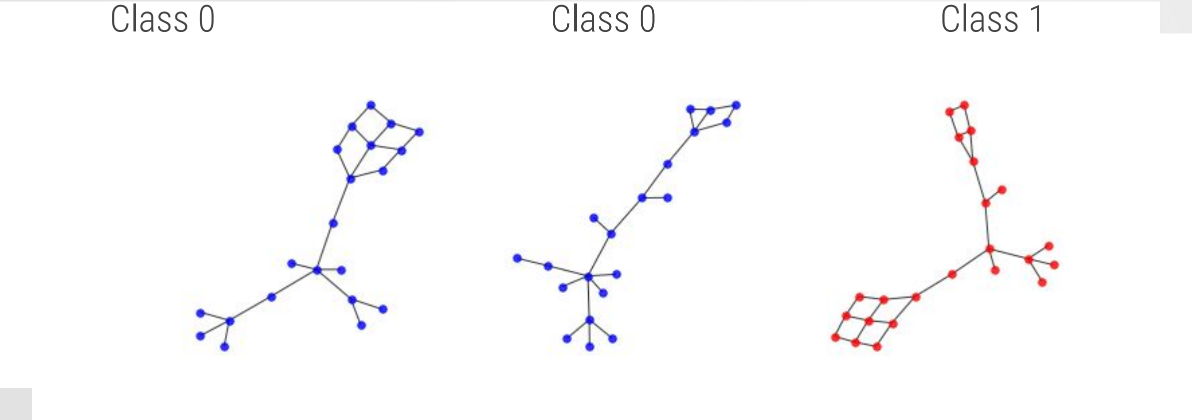
Grid house:

- Binary graph classification
- Classes:
  - 0 → BA random graph + 3x3 grid graph **OR** 5 node house graph
  - 1 → BA random graph + 3x3 grid graph **AND** 5 node house graph

04 Benchmark datasets

Grid house:

- Binary graph classification
- Classes:
  - 0 → BA random graph + 3x3 grid graph **OR** 5 node house graph
  - 1 → BA random graph + 3x3 grid graph **AND** 5 node house graph



04 Benchmark datasets

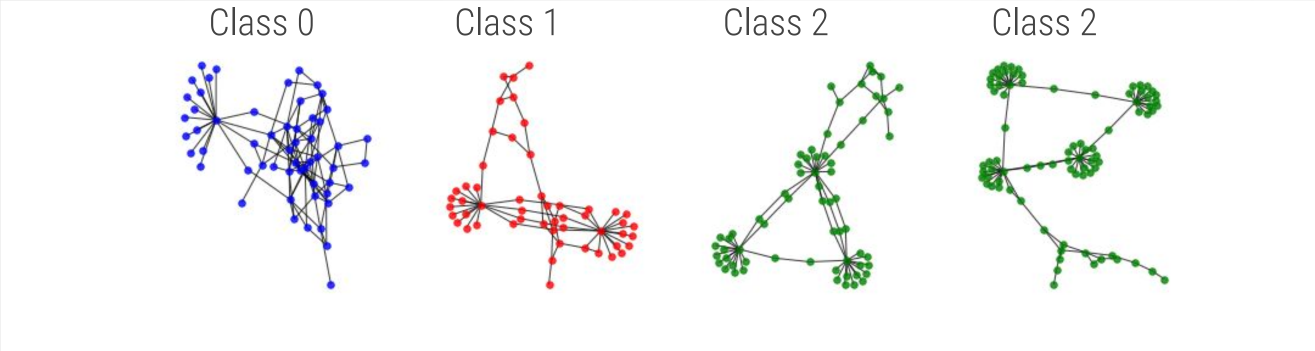
Stars:

- 3 class graph classification
- Classes:
  - 0 → ER random graph + 1 star
  - 1 → ER random graph + 2 stars
  - 2 → ER random graph + 3 stars **OR** 4 stars

04 Benchmark datasets

Stars:

- 3 class graph classification
- Classes:
  - 0 ⟶ ER random graph + 1 star
  - 1 ⟶ ER random graph + 2 stars
  - 2 ⟶ ER random graph + 3 stars **OR** 4 stars



Evaluation

Plausibility

Explanation

Binary GT

$$P = \text{AucROC}(G_{\text{exp}}, \overline{G}_{\text{exp}}),$$

Evaluation

Plausibility

Explanation

$$P = \mathrm{AucROC}(G_{\mathrm{exp}}, \overline{G}_{\mathrm{exp}}),$$

Binary GT

Fidelity

$$F_{f1} = 2\frac{(1 - F_{suf}) \cdot F_{com}}{(1 - F_{suf}) + F_{com}}.$$

Evaluation

Plausibility

Explanation

$$P = \mathrm{AucROC}(G_{\mathrm{exp}}, \overline{G}_{\mathrm{exp}}),$$

Binary GT

Fidelity

$$F_{f1} = 2\frac{(1 - F_{suf}) \cdot F_{com}}{(1 - F_{suf}) + F_{com}}.$$

$$F_{suf} = \frac{1}{N_t - 1}\sum_{k=1}^{N_t - 1}\left(g(G) - g(G_{\mathrm{exp}}(t_k))\right),$$

$$F_{com} = \frac{1}{N_t - 1}\sum_{k=1}^{N_t - 1}\left(g(G) - g(G \setminus G_{\mathrm{exp}}(t_k))\right),$$

So far…

- 8 GNN architectures
- 10 Explainers
- 3 Dataset (6 in the paper)

What we can do?

So far...

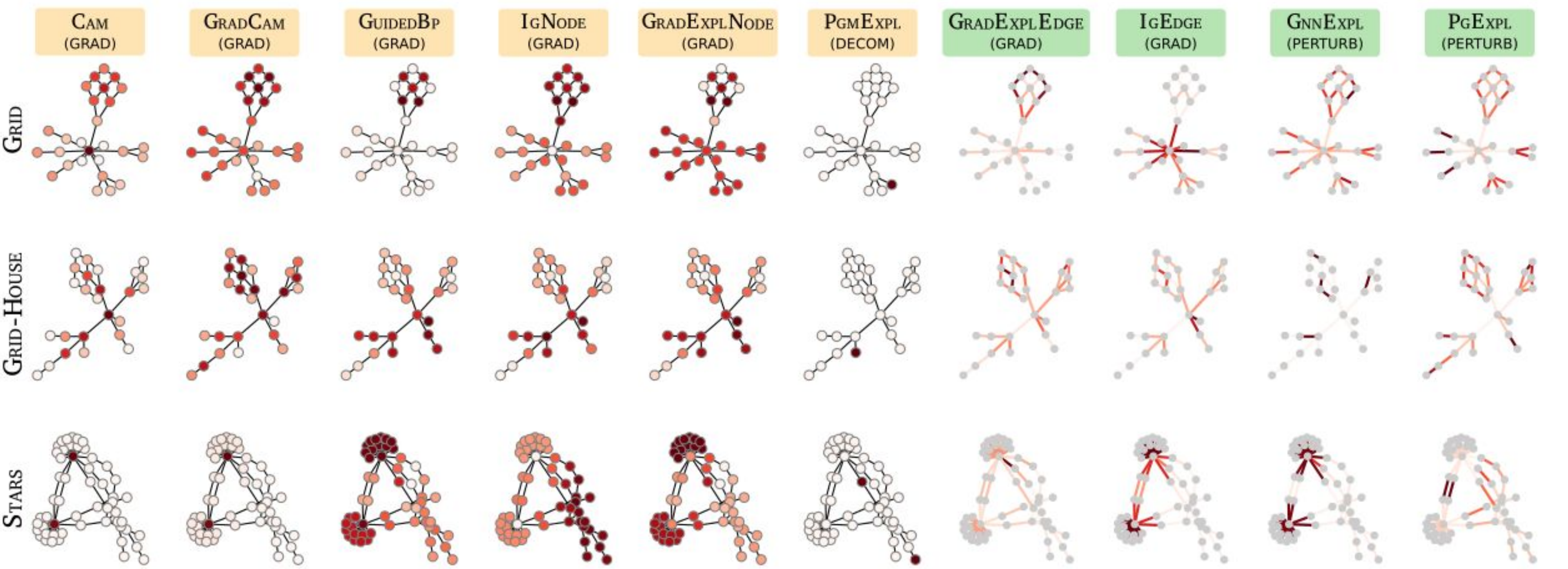
- 8 GNN architectures
- 10 Explainers
- 3 Dataset (6 in the paper)

What we can do?

8x10x6x(1000 graphs) = 480 000 (explanations)

Do not waste time (**and energy)**! If you need, they are available here:
https://github.com/AntonioLonga/GraphXAI/tree/main/Explanations

Look at them!


CAM (GRAD)
GRADCAM (GRAD)
GUIDEDBP (GRAD)
IGNODE (GRAD)
GRADEXPLNODE (GRAD)
PGMEXPL (DECOM)
GRADEXPLEDGE (GRAD)
IGEDGE (GRAD)
GNNEXPL (PERTURB)
PGEXPL (PERTURB)
GRID
GRID-HOUSE
STARS

05 Research Questions

- RQ1: How does the architecture affect the explanations?
- RQ2: How do explainers affect the explanations?
- RQ3: How do different types of data affect the explanations?

05 Research Questions

RQ1: How does the architecture affect the explanations?

- RQ1.1: Which is the architecture that has the best explanation?
- RQ1.2: Which is the easiest architecture to explain?
- RQ1.3: Which is the hardest architecture to explain?

	Plausibility			
	All	Grid	Grid-House	Stars
RQ1.1	**GraphConv**	GraphConv	Cheb	Set2Set
RQ1.2	**Gcn**	Cheb	Gcn	Set2Set
RQ1.3	**Gin**	Gin	Gin	MinCutPool

	Fidelity			
	All	Grid	Grid-House	Stars
RQ1.1	**GraphConv**	Cheb	Set2Set	GraphConv
RQ1.2	**Gcn**	Gcn	MinCutPool	GraphConv
RQ1.3	**Gin**	Gin	Gin	MinCutPool

05 Research Questions

RQ2: How do explainers affect the explanations?

- RQ2.1: Which is the explainer that explains in the best way?
- RQ2.2: Which is the explainer that explains the maximum number of architectures?
- RQ2.3: Which is the category of explainers that provides the best explanations? (Grad, Pert, Dec)
- RQ2.4: Which is the best mask type between node and edge?

Plausibility

	All	Grid	Grid-House	Stars
RQ2.1	**GradExplEdge**	IgEdge	PgExpl	IgEdge
RQ2.2	**GradExplEdge**	GradExplEdge	PgExpl	GradExplEdge
RQ2.3	**Pert**	Pert	Pert	Grad
RQ2.4	**Edge**	Edge	Edge	Edge

Fidelity

	All	Grid	Grid-House	Stars
RQ2.1	**IgEdge**	PgExpl	IgEdge	GradExplEdge
RQ2.2	**IgEdge**	IgEdge	IgEdge	GnnExpl
RQ2.3	**Pert**	Pert	Pert	Pert
RQ2.4	**Edge**	Edge	Edge	Edge

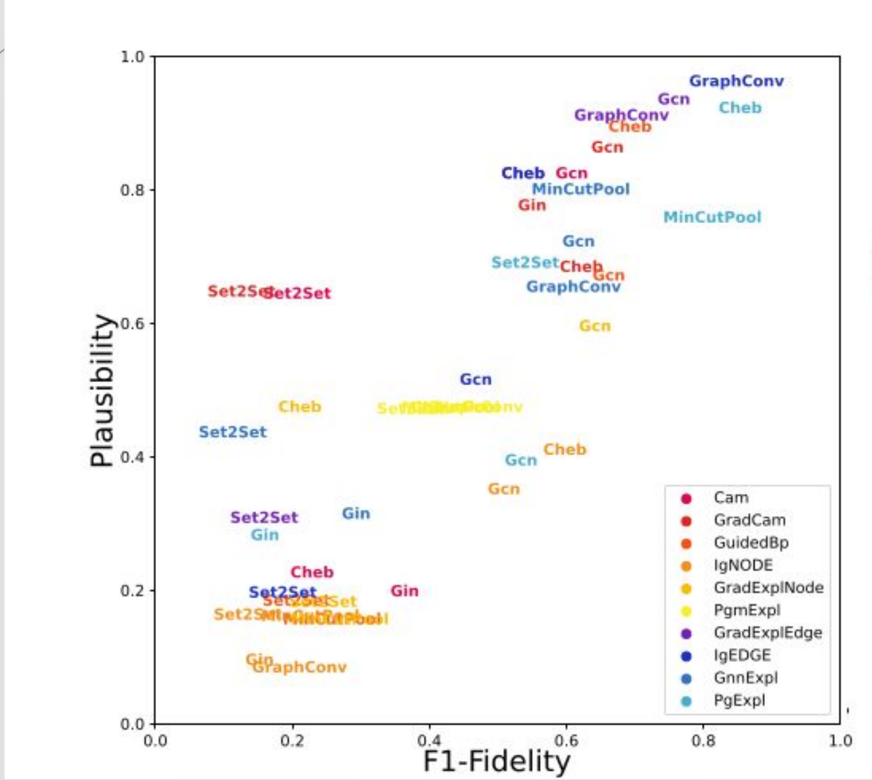
05 Research Questions

RQ3: How do different types of data affect the explanations?

05 Research Questions

RQ3: How do different types of data affect the explanations?

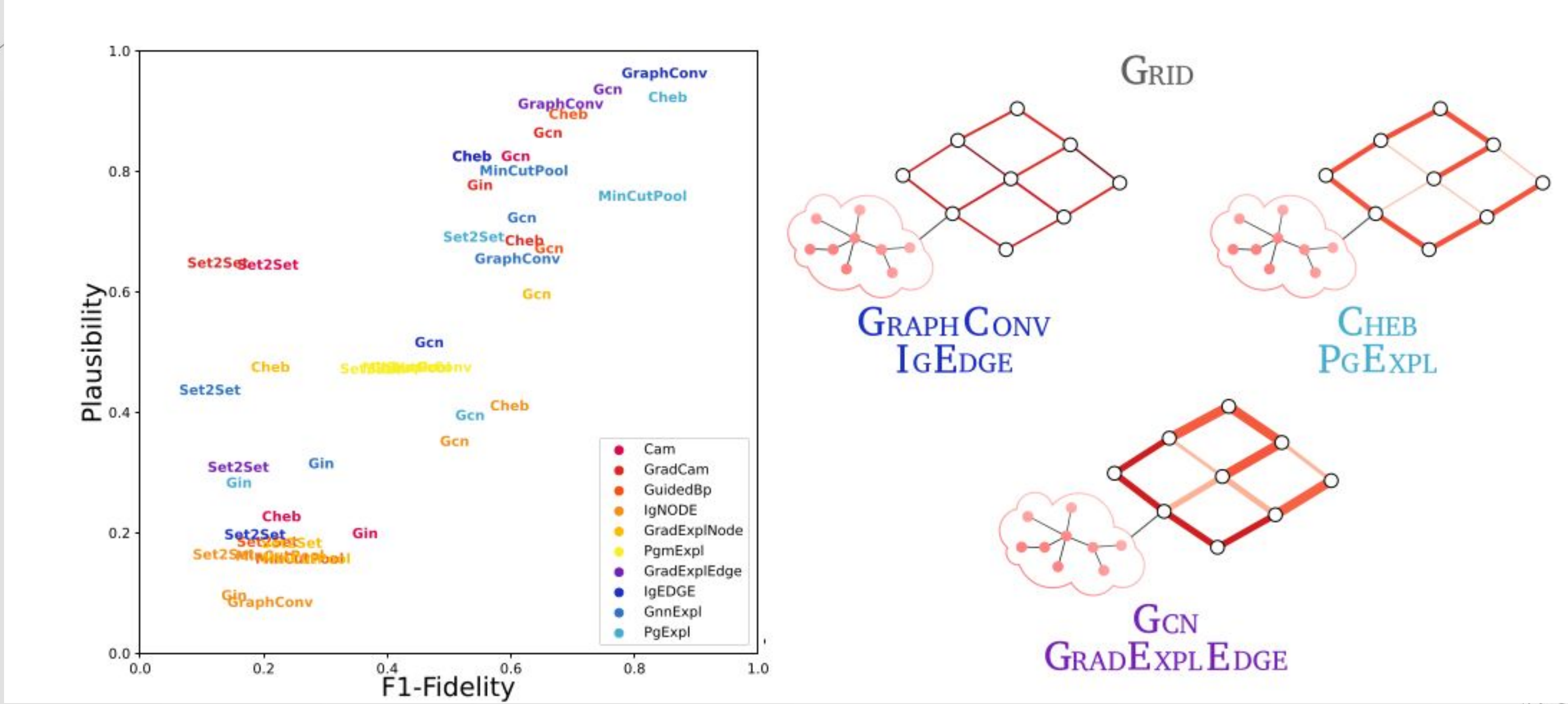
GRID



05 Research Questions

RQ3: How do different types of data affect the explanations?

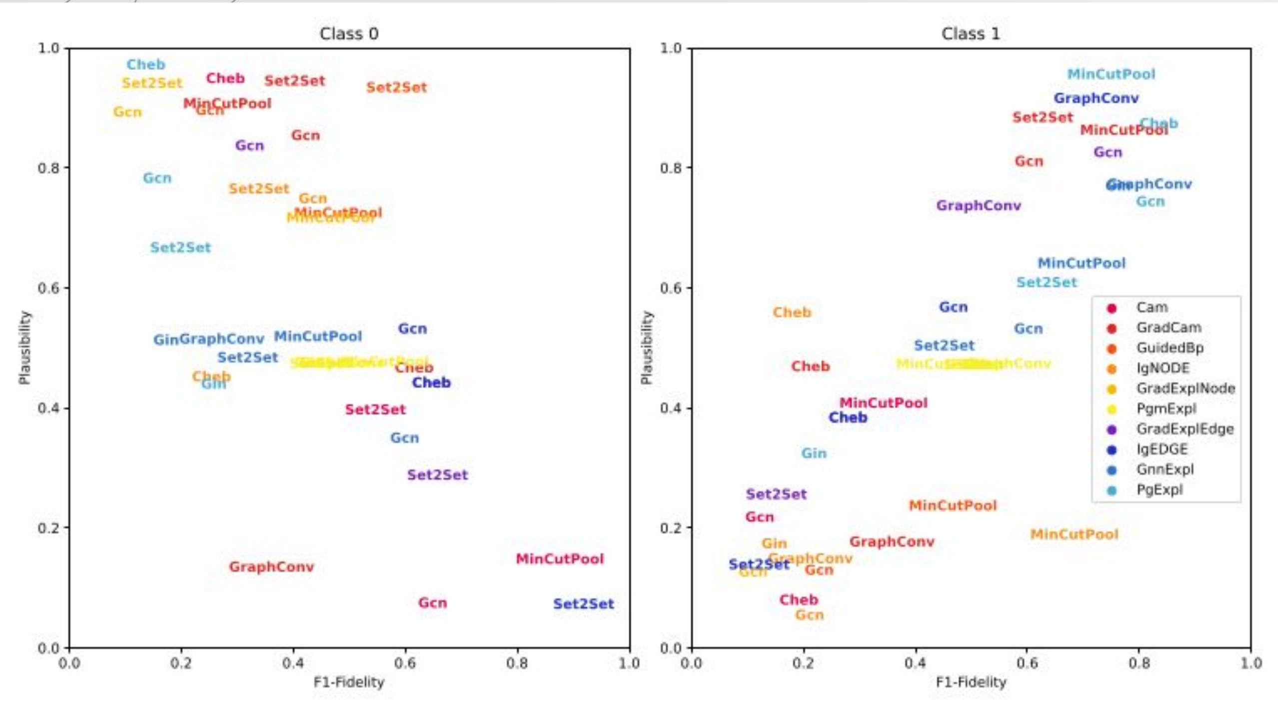
GRID



05 Research Questions

RQ3: How do different types of data affect the explanations?

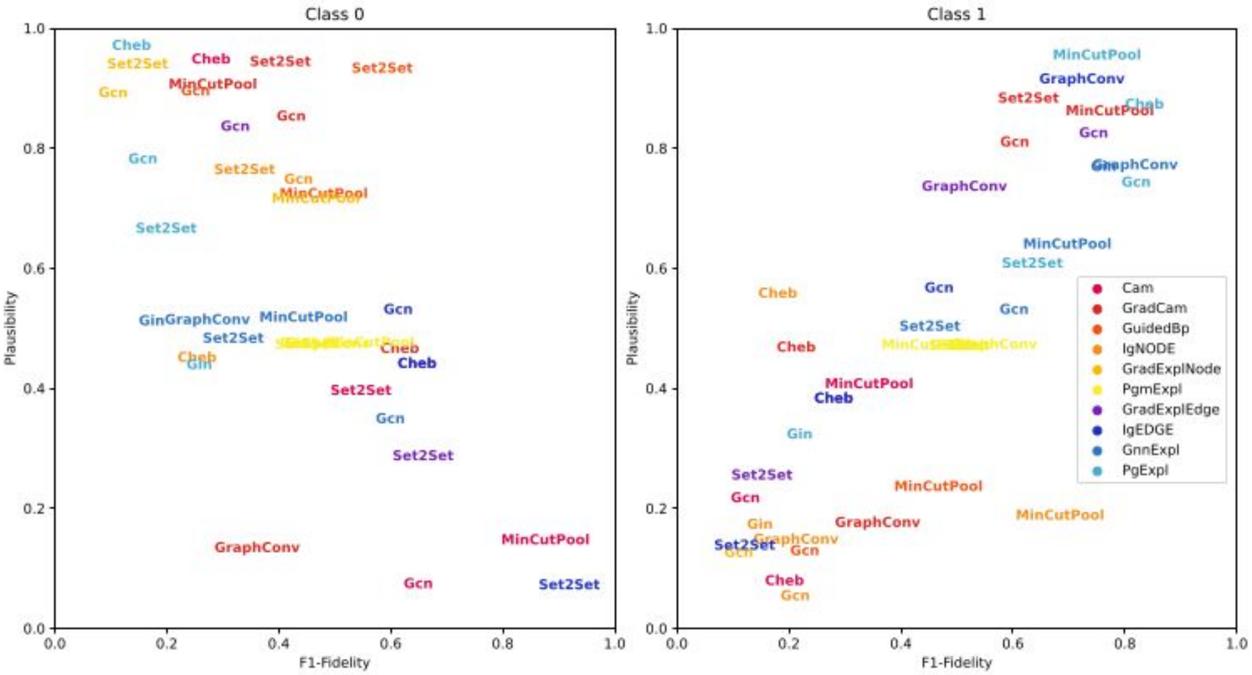
GRID HOUSE

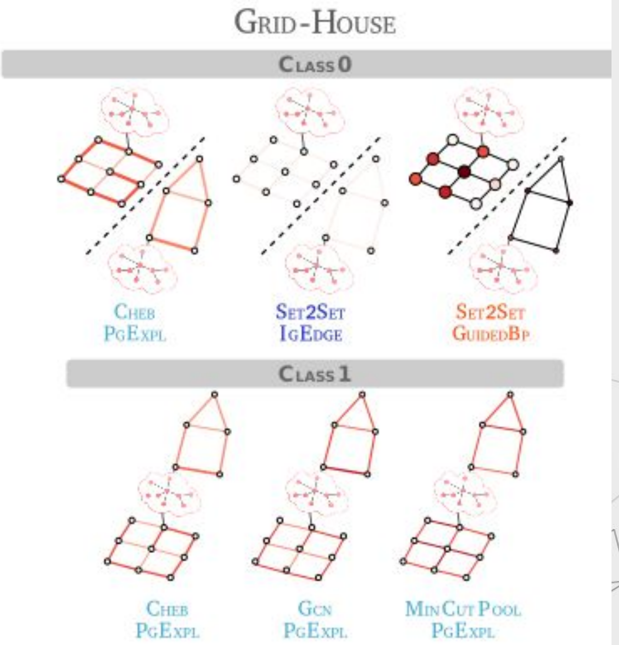


05 Research Questions

RQ3: How do different types of data affect the explanations?

GRID HOUSE

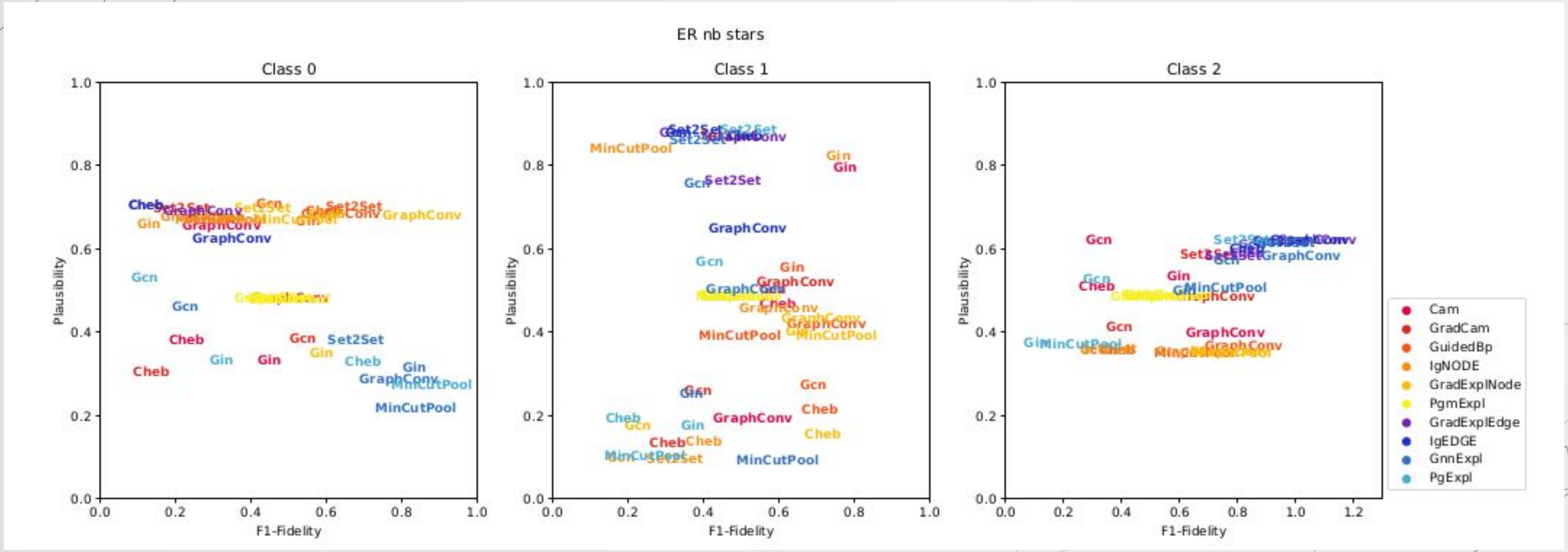




05 Research Questions

RQ3: How do different types of data affect the explanations?

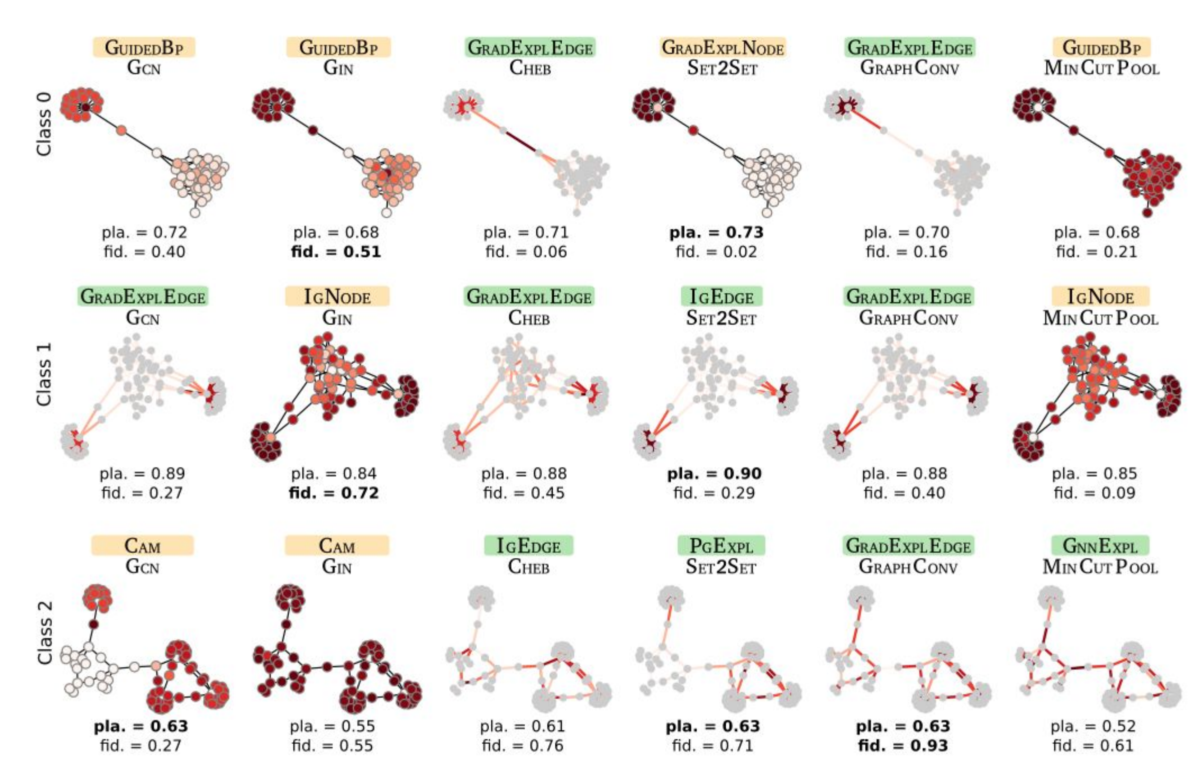
STARS



05 Research Questions

RQ3: How do different types of data affect the explanations?

STARS

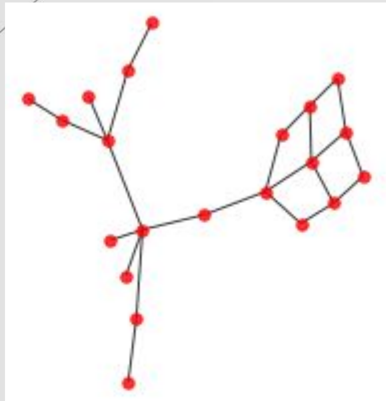


06 Conclusion & future directions

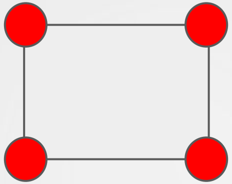
1) Human bias when defining Ground Truth.
    a) In GRID network do we need the entire grid?

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
    a) In GRID network do we need the entire grid?



A square is enough



06 Conclusion & future directions

1) Human bias when defining Ground Truth.
    a) In GRID network do we need the entire grid?
2) We could use only the fidelity…

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE

Graph Classification

STARS

Node Classification

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE

Graph Classification

STARS

1) Identify stars
2) Count them
3) Classify according to the frequence of stars

Node Classification

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE

Graph Classification

Node Classification

STARS

1) Identify stars
2) Count them
3) Classify according to the frequence of stars

Node embedding → **SUM** → Graph embedding

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE

Graph Classification

STARS

1) Identify stars
2) Count them
3) Classify according to the frequence of stars

Node embedding → **SUM** → Graph embedding

SUM ↓ Affects Fidelity

Node Classification

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE

Graph Classification

STARS

1) Identify stars
2) Count them
3) Classify according to the frequence of stars

Node embedding → **SUM** → Graph embedding

↓

Affects Fidelity

Node Classification

1) Comprehensiveness → difficult to define

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE

Graph Classification

STARS

1) Identify stars
2) Count them
3) Classify according to the frequence of stars

Node embedding → **SUM** → Graph embedding

SUM ↓ Affects Fidelity

Node Classification

1) Comprehensiveness → difficult to define

Only sufficiency

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE

Graph Classification

STARS

1) Identify stars
2) Count them
3) Classify according to the frequence of stars

Node embedding → **SUM** → Graph embedding

SUM ↓ Affects Fidelity

Node Classification

1) Comprehensiveness → difficult to define

Only sufficiency

↓

The entire graph has the perfect score!!!

06 Conclusion & future directions

1) Human bias when defining Ground Truth.
   a) In GRID network do we need the entire grid?
2) We could use only the fidelity…
   a) NOPE
3) Overall it seems that:
   a) Node Classification → Gradient based.
   b) Graph Classification → Edge mask based on Gradient or Perturbation.

Thanks!
Do you have any questions?

Antonio Longa
alonga@fbk.eu
https://antoniolonga.github.io/
https://twitter.com/AntonioLonga94