EXPLAINABILITY IN GRAPH NEURAL NETWORKS

DBDM, ENSL, March 2021

Links

- Explainable IA:
  - https://sites.google.com/view/www20-explainable-ai-tutorial (WWW'2020 tutorial)
  - https://xaitutorial2020.github.io/ (AAAI'2020 tutorial)
- GNN:
  - https://web.stanford.edu/class/cs224w/slides/08-GNN.pdf
  - https://github.com/snap-stanford/cs224w-notes/tree/master/machine-learning-with-networks
- GNN and Explainability:
  - Yuan, H., Yu, H., Gui, S., & Ji, S. (2020). Explainability in Graph Neural Networks: A Taxonomic Survey. *arXiv preprint arXiv:2012.15445*.

DNN: A revolution in ML & AI

- DNN have achieved promising performance in many research task:
  - Computer vision
    - S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 221–231, 2013.
    - K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
  - NLP
    - J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pretraining of deep bidirectional transformers for language understanding,” in NAACL-HLT (1), 2019.
    - A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 59986008.
  - Graph data analysis
    - T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” arXiv preprint arXiv:1609.02907, 2016.
    - K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?” in International Conference on Learning Representations, 2019. [Online]. Available: https://openreview.net/forum?id=ryGs6iA5Km

Development of DL methods for real-world applications in interdisciplinary domains

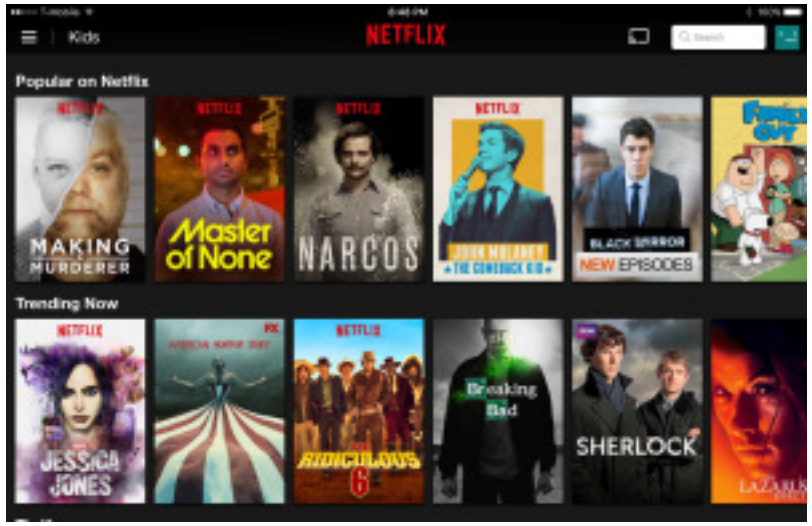
- Finance, Biology, Agriculture, Neuroscience, Astronomy, Defense, Sport analytics, Recommender systems, …
- Most deep models are developed without interpretability: **Black boxes**.
- Without reasoning the underlying mechanisms behind the predictions, deep models **cannot be fully trusted**, which prevents their use in critical applications pertaining to **fairness**, **privacy**, and **safety**.
- To **safely** and **trustfully** deploy deep models, it is necessary to provide **both accurate predictions** and **human-intelligible explanations**, *especially for users in interdisciplinary domains*.
- **The need of developing explanation techniques to explain deep neural networks.**

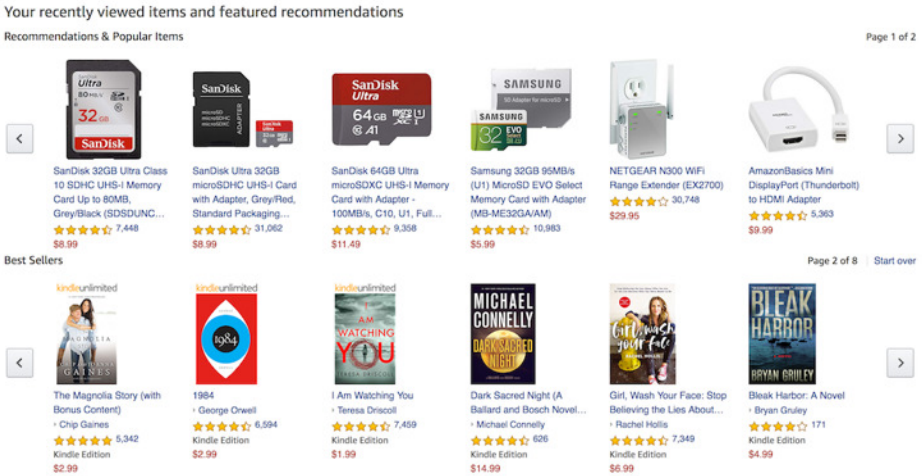
What is « Explainable AI»?

- **Explainable AI** explores and investigates methods to produce or complement **AI models** to make **accessible** and **interpretable** the internal logic and the outcome of the algorithms, making such process **understandable** by humans.
- **Explicability**, understood as incorporating both **intelligibility** (“*how does it work?*” for non experts, e.g., patients or business customers, and for experts, e.g., product designers or engineers) and **accountability** (“*who is responsible for*”).
- 5 core principles for ethical AI:
  - **beneficence, non maleficence**, **autonomy**, and **justice**
  - a new principle is needed in addition: **explicability**

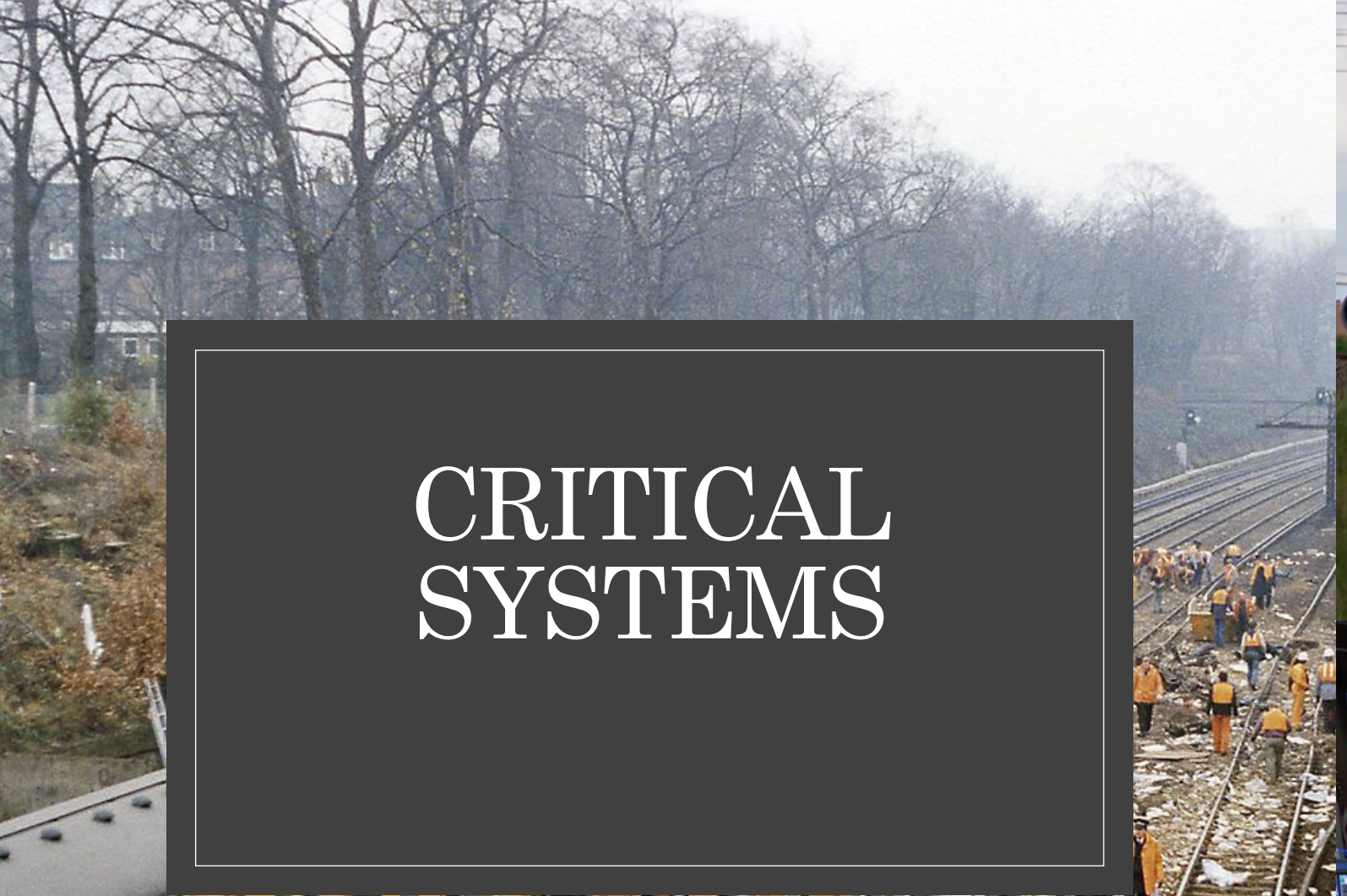
MOTIVATING EXAMPLES

Business to Customer AI





CRITICAL SYSTEMS



Not only …

- Criminal Justice
  - People wrongly denied
  - Unfair Police dispatch
  - Recidism prediction


How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin
May 23, 2016


STATEMENT OF CONCERN ABOUT PREDICTIVE POLICING BY ACLU AND 16 CIVIL RIGHTS PRIVACY, RACIAL JUSTICE, AND TECHNOLOGY ORGANIZATIONS


The New York Times

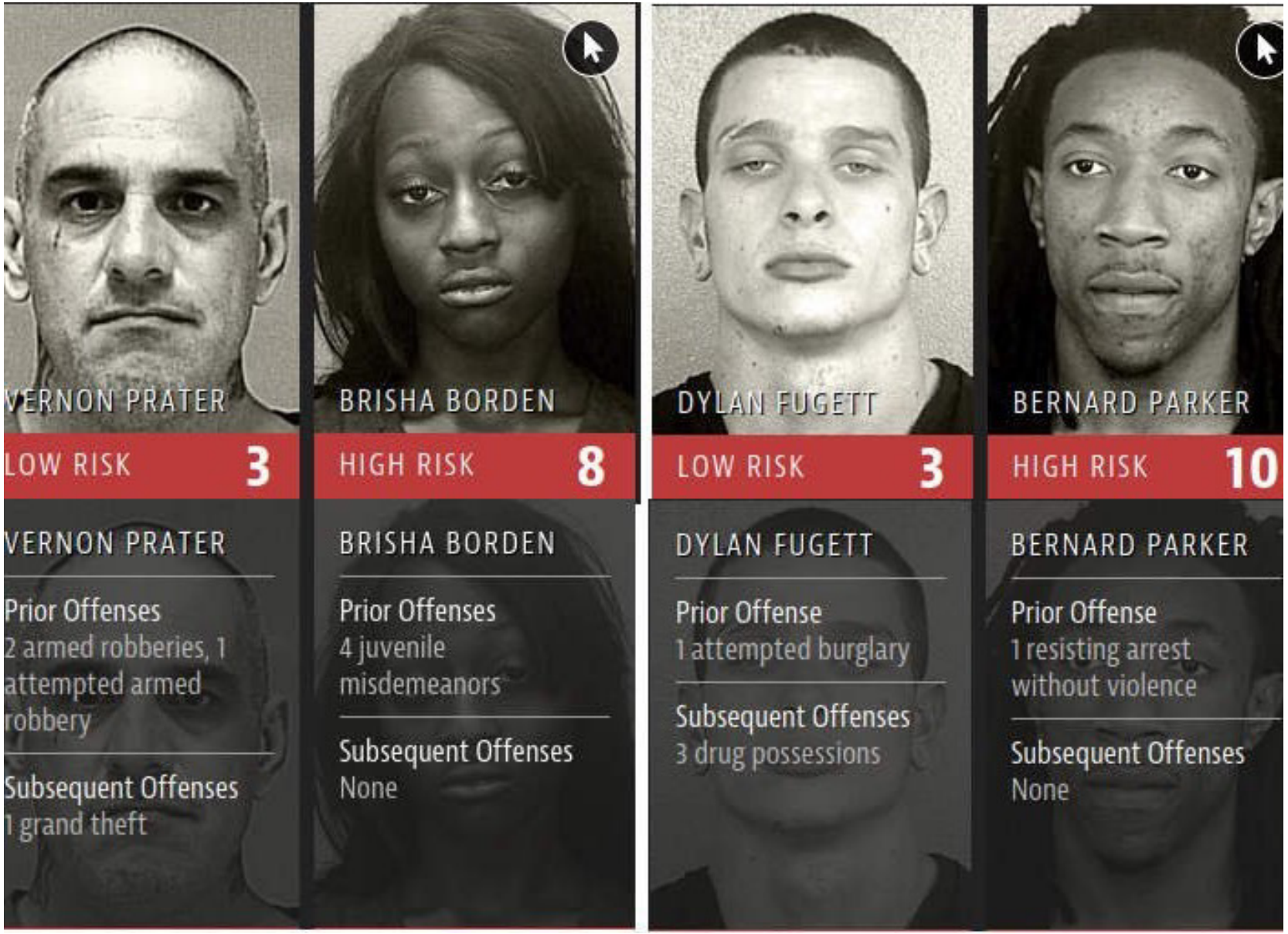
Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017



COMPAS recidivism black bias

- Compass has become very **unreliable.**
- only 20% of people considered at risk of recidivism ended up committing a new crime.
- Researchers at Dartmouth College conducted an experiment that proved that the predictions **Compas provided were no better than those made by people with no legal training.**

Finance: credit scoring, insurance quotes

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

Artificial intelligence could revolutionise the industry but may also allow clients to calculate if they need protection

Save

Oliver Ralph MAY 16, 2017 24

Healthcare

- Applying ML methods in medical care is **problematic**
- AI as **3rd party actor** in Physician/Patient relationship
- Responsibility, Confidentiality ?
- Learning must be done with available data
  - Can not randomize care given to patients !
- Must validate models before use.

Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*. 2020;295-336. doi:10.1016/B978-0-12-818438-7.00012-5

Pesapane, F., Volonté, C., Codari, M. *et al.* Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* **9,** 745–753 (2018).

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *Journal of Clinical Neuroscience*, *64*, 277-282.

Black-box AI creates business risk for Industry

Bloomberg Businessweek

Apple Card's Gender-Bias Claims Look Familiar to Old-School Banks

Updated on November 12, 2019, 4:23 AM

MIT News

Study finds gender and skin-type bias in commercial AI systems

Feb 12, 2018

BBC NEWS

Tay: Microsoft issues apology over racist chatbot fiasco

Sep 22, 2017

Missouri S&T News and Research

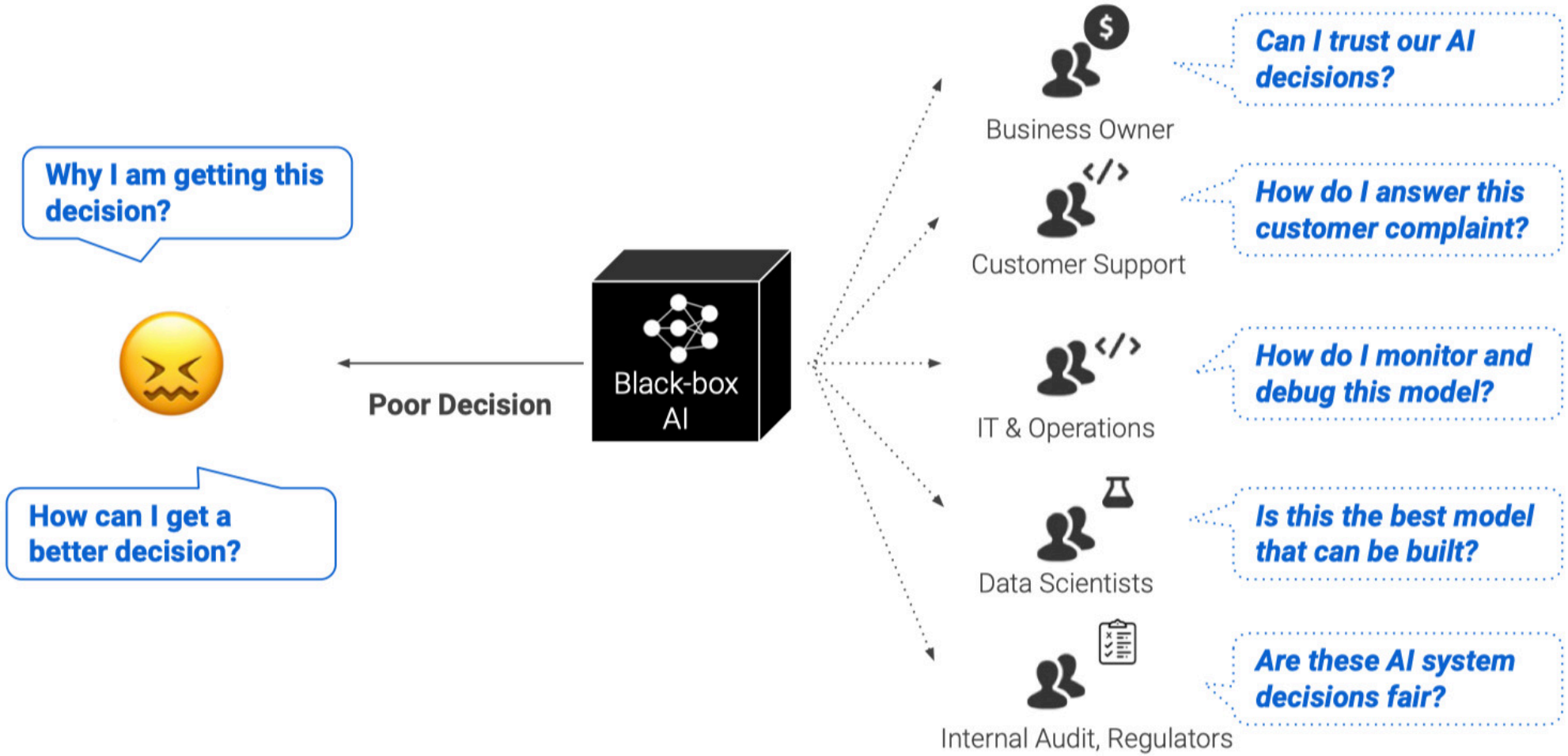
After Uber, Tesla incidents, can artificial intelligence be trusted?

Apr 10, 2018

Guilty! AI Is Found to Perpetuate Biases in Jailing

1 day ago

Black-box AI creates confusion and doubt



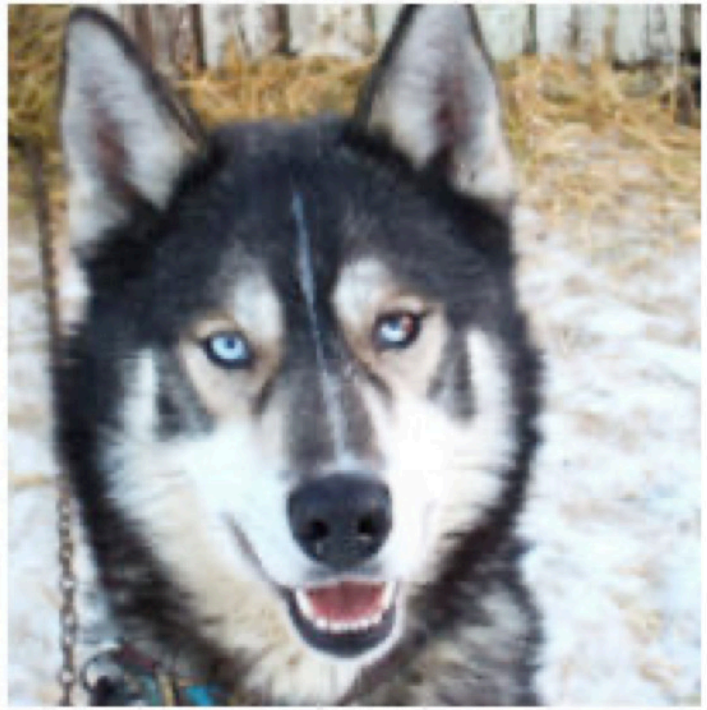
EXPLANATION - FROM A MODEL PERSPECTIVE

Why Explainability: Debug (Mis-)Predictions

Top label: **“clog”**

Why did the network label this image as **“clog”**?

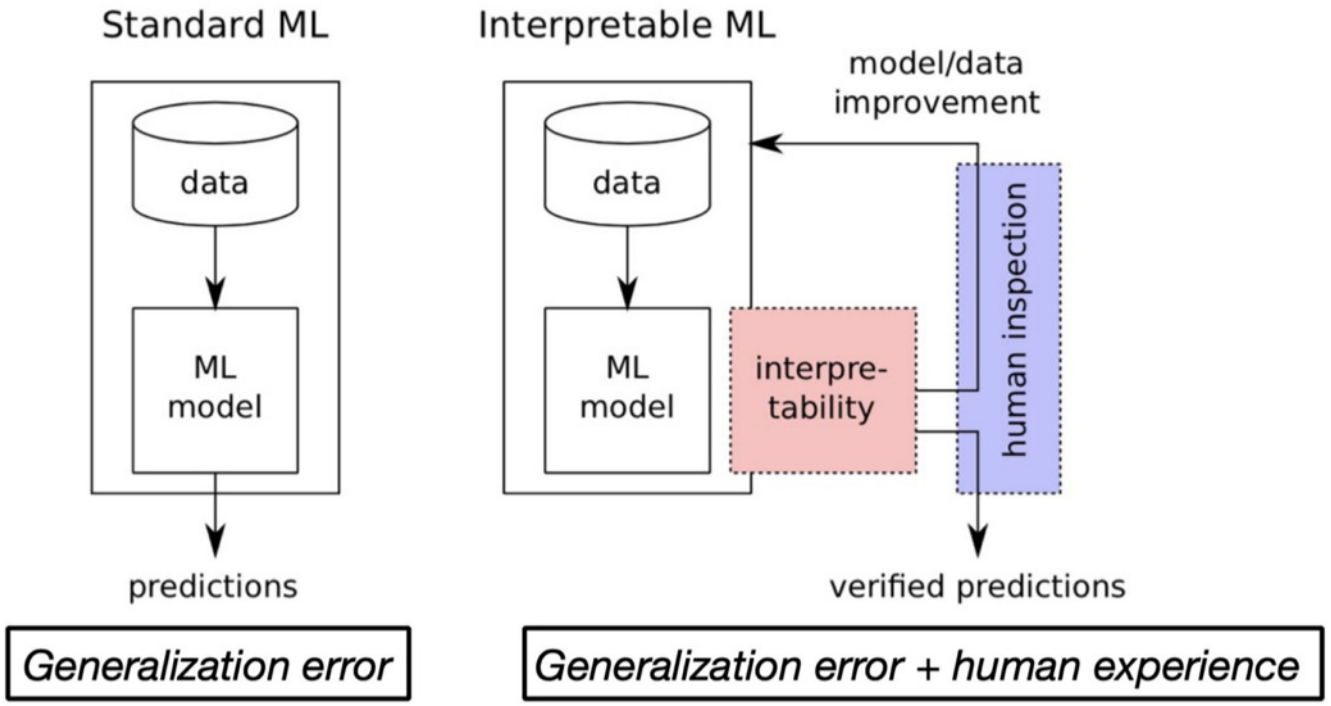
The background bias



(a) Husky classified as wolf

(b) Explanation

Why Explainability: Improve ML Model



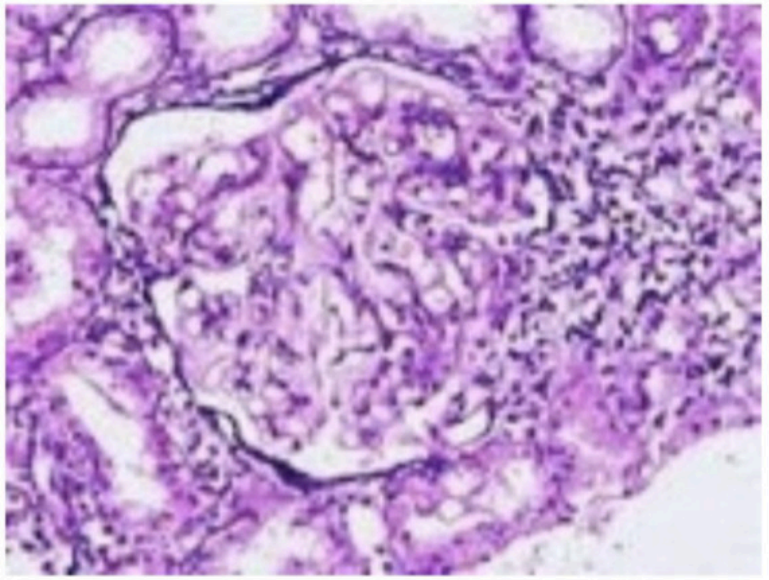
Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Verify the ML Model / System

Wrong decisions can be costly
and dangerous

"Autonomous car crashes, because it wrongly recognizes …"

"AI medical diagnosis system misclassifies patient's disease …"

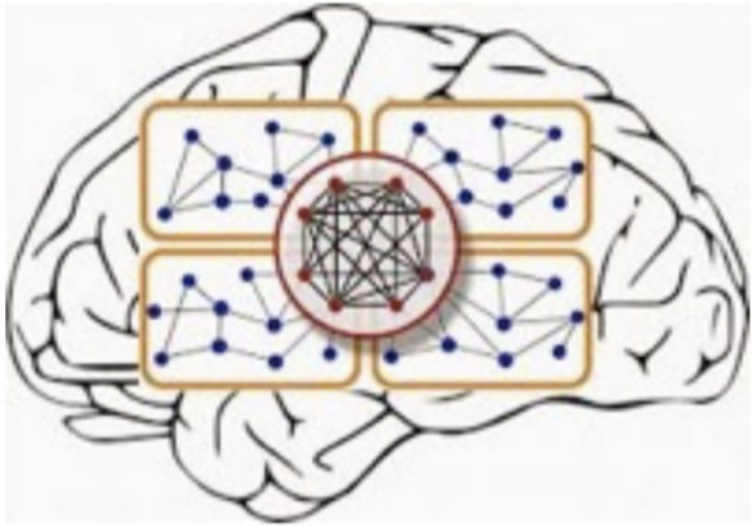


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Learn New Insights

"It's not a human move. I've never seen a human play this move." (Fan Hui)

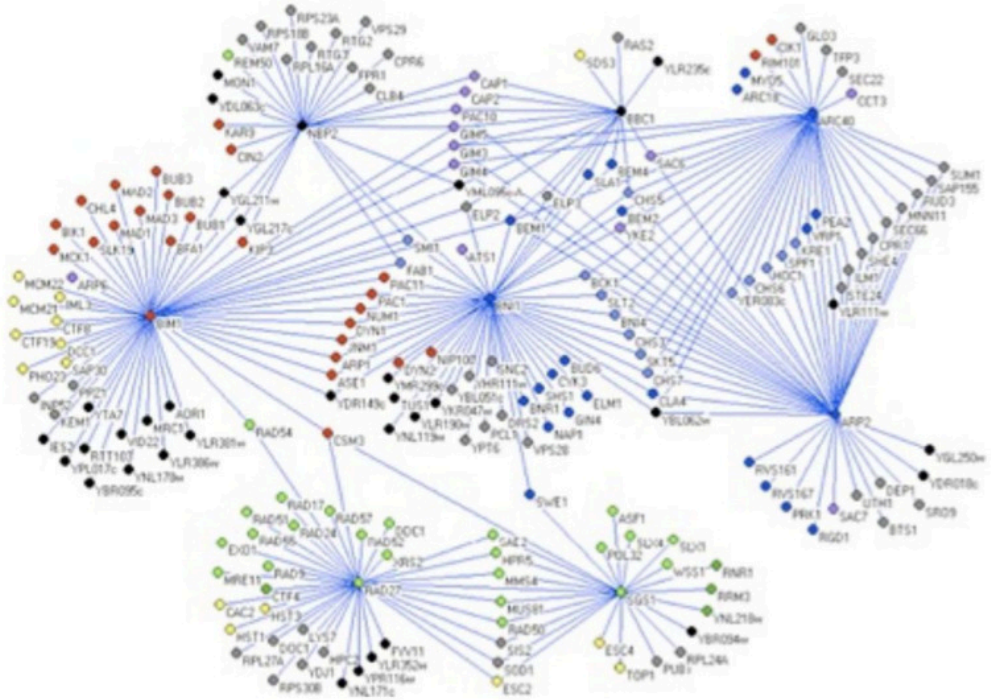
Old promise:
"Learn about the human brain."

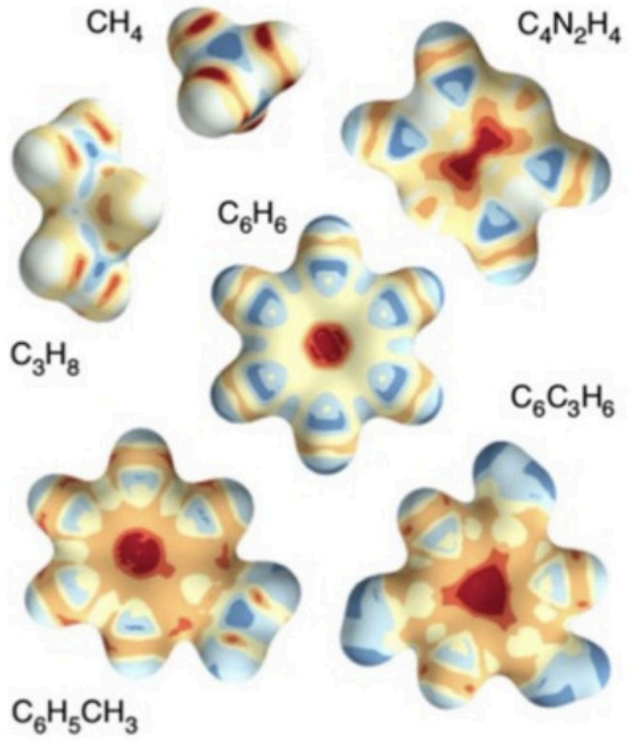


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Learn Insights in the Sciences

Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites …)

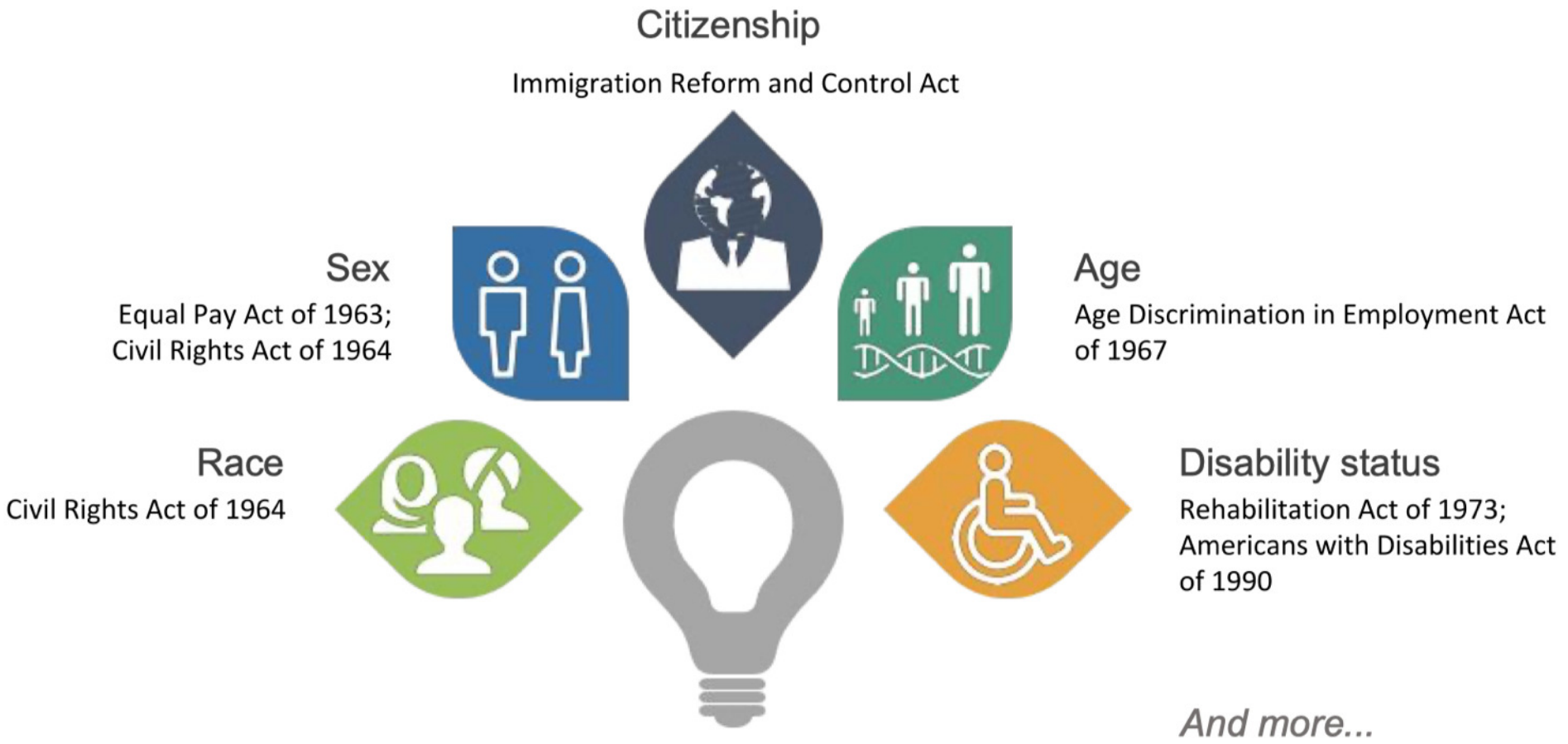




Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

EXPLANATION - FROM A REGULATORY PERSPECTIVE

Why Explainability: Laws against Discrimination



SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

FAIRNESS

PRIVACY

EXPLAINABILITY

TRANSPARENCY

GDPR concerns about lack of explainability in AI

"

Companies should commit to ensuring systems that could fall under GDPR, including AI, will be compliant. The threat of ***sizeable fines of €20 million or 4% of global turnover*** *provides a sharp incentive.*

Article 22 of GDPR empowers individuals with the ***right to demand an explanation of how an AI system made a decision that affects them.***

"

- European Commision

VP, European Commision

Article 22
EU GDPR
"Automated individual decision-making, including profiling"

=> Recital: 71, 72
=> administrative fine: Art. 83 (5) lit b
=> Dossier: Automated Decision In Individual Cases, Profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
=> Article: 4

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
=> Dossier: Legitimate Interests (Data Subject), Opening Clause

(c) is based on the data subject's explicit consent.
=> Dossier: Consent

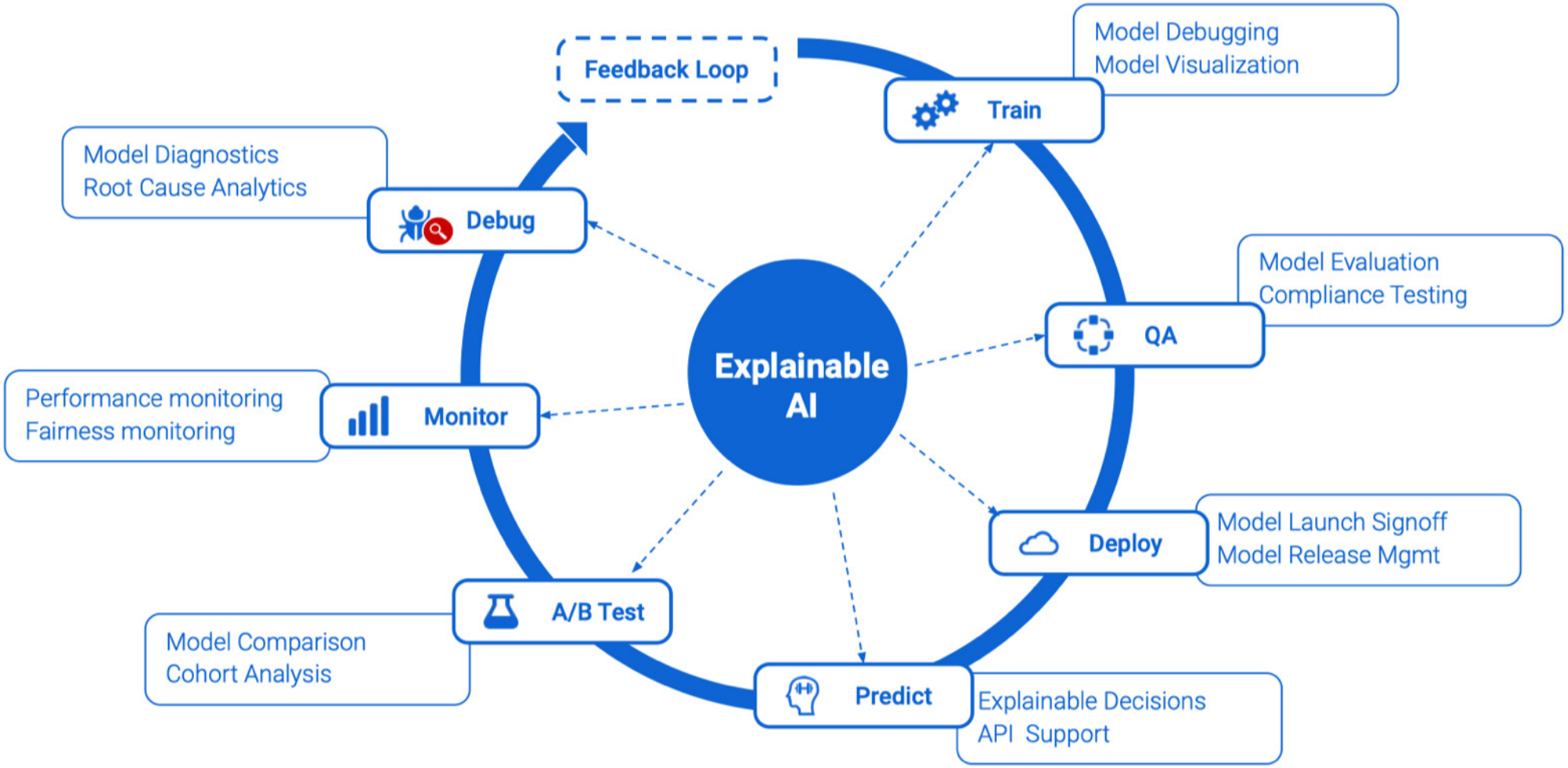
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller **shall implement suitable measures** to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
=> Recital: 70
=> Dossier: Legitimate Interests (Data Subject), Obligation

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.
=> Dossier: Legitimate Interests (Data Subject)

Why Explainability: Growing Global AI Regulation

- GDPR: Article 22 empowers individuals with the right to demand an explanation of how an automated system made a decision that affects them.
- Algorithmic Accountability Act 2019: Requires companies to provide an assessment of the risks posed by the automated decision system to the privacy or security and the risks that contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers
- California Consumer Privacy Act: Requires companies to rethink their approach to capturing, storing, and sharing personal data to align with the new requirements by January 1, 2020.
- Washington Bill 1655: Establishes guidelines for the use of automated decision systems to protect consumers, improve transparency, and create more market predictability.
- Massachusetts Bill H.2701: Establishes a commission on automated decision-making, transparency, fairness, and individual rights.
- Illinois House Bill 3415: States predictive data analytics determining creditworthiness or hiring decisions may not include information that correlates with the applicant race or zip code.

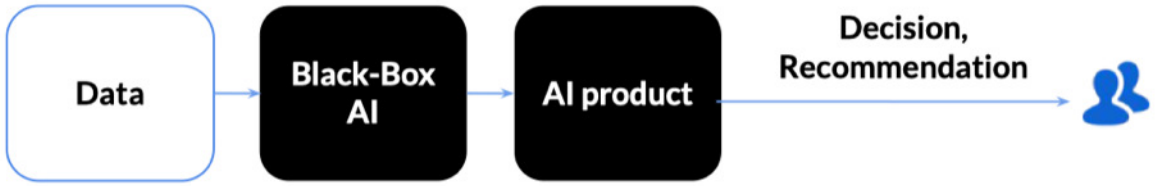
“Explainability by Design” for AI products



EXPLANATION - IN A NUTSHELL

What is Explainable AI?

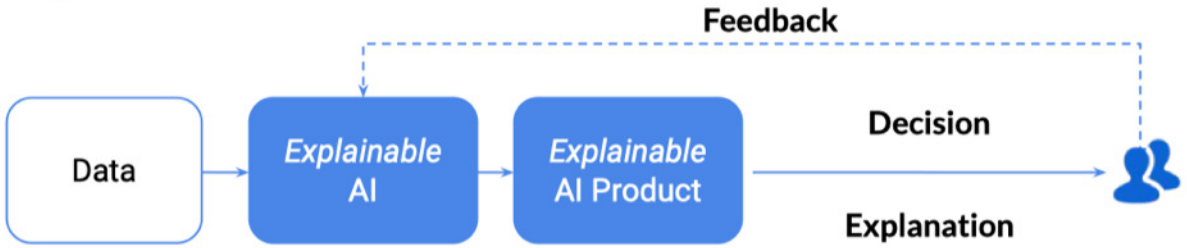
Black Box AI



Confusion with Today's AI Black Box

- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

Explainable AI

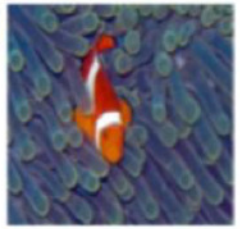


Clear & Transparent Predictions

- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Example of an End-to-End XAI System

1

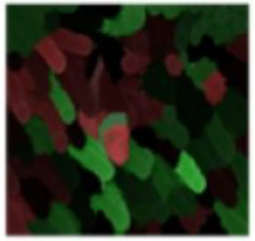


ML Classifier

C: I predict FISH

2 H: Why?

C: See below:



Green regions argue for FISH, while RED pushes towards DOG. There's more green.

3 H: (Hmm. Seems like it might be just recognizing anemone texture!) Which training examples are most influential to the prediction?

C: These ones:

4 H: What happens if the background anemones are removed? E.g.,

C: I still predict FISH, because of these green superpixels:

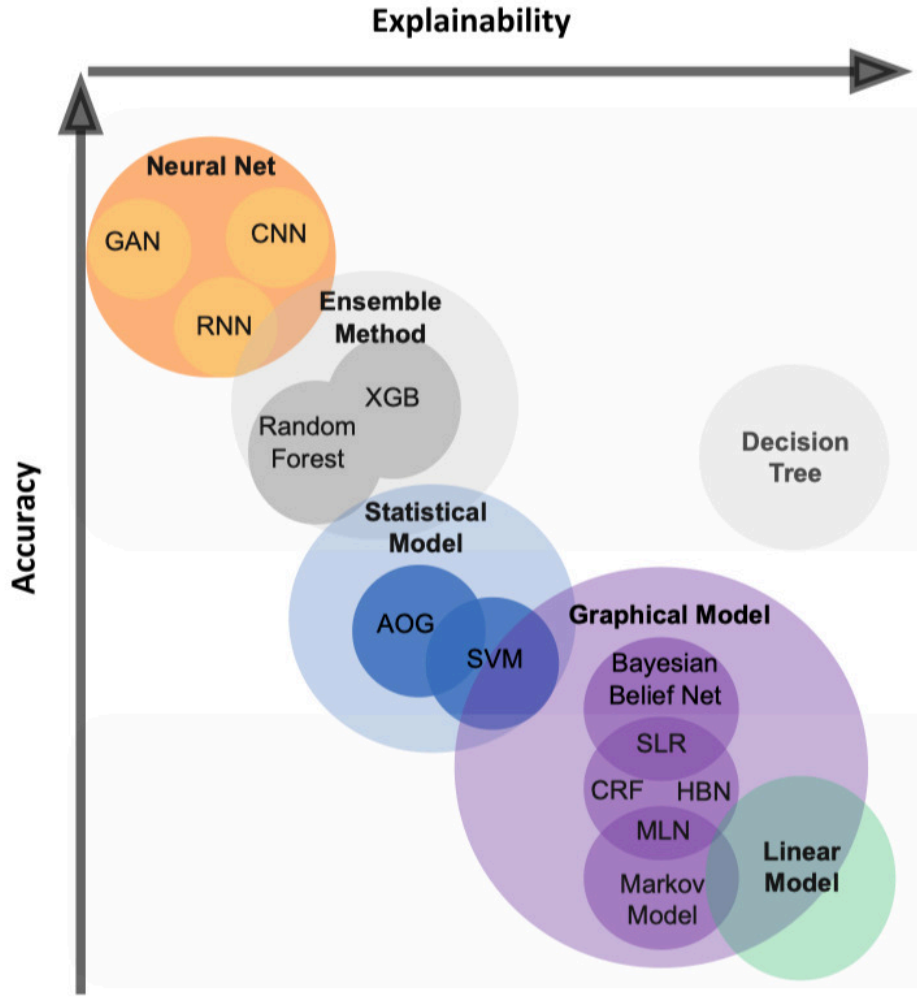
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

Weld, D., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of ACM (2018).

How to Explain? Accuracy vs. Explainability

Learning

- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - **Correlation**
  - **No causation**



Interpretability

- Non-Linear functions
- Polynomial functions
- Quasi-Linear functions

EXPLANATION AND GNN

Many explanation techniques for images and text

- Input-dependent explanations:
  - Studying the important score for input features
  - Studying the gradient or weights to analyse the sensitivity between input features and the predictions
  - Occlusion of input features
- Input independent explanations:
  - Studying the input patterns that maximize the predicted score of a certain class.

M. Du, N. Liu, and X. Hu, “Techniques for interpretable machine learning,” Communications of the ACM, vol. 63, no. 1, pp. 68–77, 2019.

C. Molnar, Interpretable Machine Learning, 2019, https:// christophm.github.io/interpretable-ml-book/.

Much less for graphs …

- GNNs have become increasingly popular since many real-world data are represented as graphs, such as social networks, chemical molecules, and financial data.
- Several graph-related tasks are widely studied :
  - node classification
  - graph classification
  - link prediction
- Many advanced GNN operations are proposed to improve the performance:
  - graph convolution,
  - graph attention,
  - graph pooling.
- However, compared with image and text domains, the explainability of graph models are less explored, which is critical for understanding deep graph neural networks

The challenges

Explaining deep graph models is an important but challenging task:

- 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.
- Graphs contain important topology information and are represented as **feature** matrices and **adjacency** matrices:
  - To explain feature importance, we may directly extend the explanation methods for image data to graph data
    - However, **the adjacency matrices** represent the topology information and only contain **discrete** values.
    - **Existing methods cannot be directly applied.**
    - For example, input optimization methods are popular to explain the general behaviors of image classifiers. It treats the input as trainable variables and optimizes the input via back-propagation to obtain abstract images to explain the model. However, the discrete adjacency matrices cannot be optimized in the same manner.
    - In addition, several methods learn soft masks to capture important image regions. However, **applying soft masks to the adjacency matrices will destroy the discreteness property**.

The challenges (2)

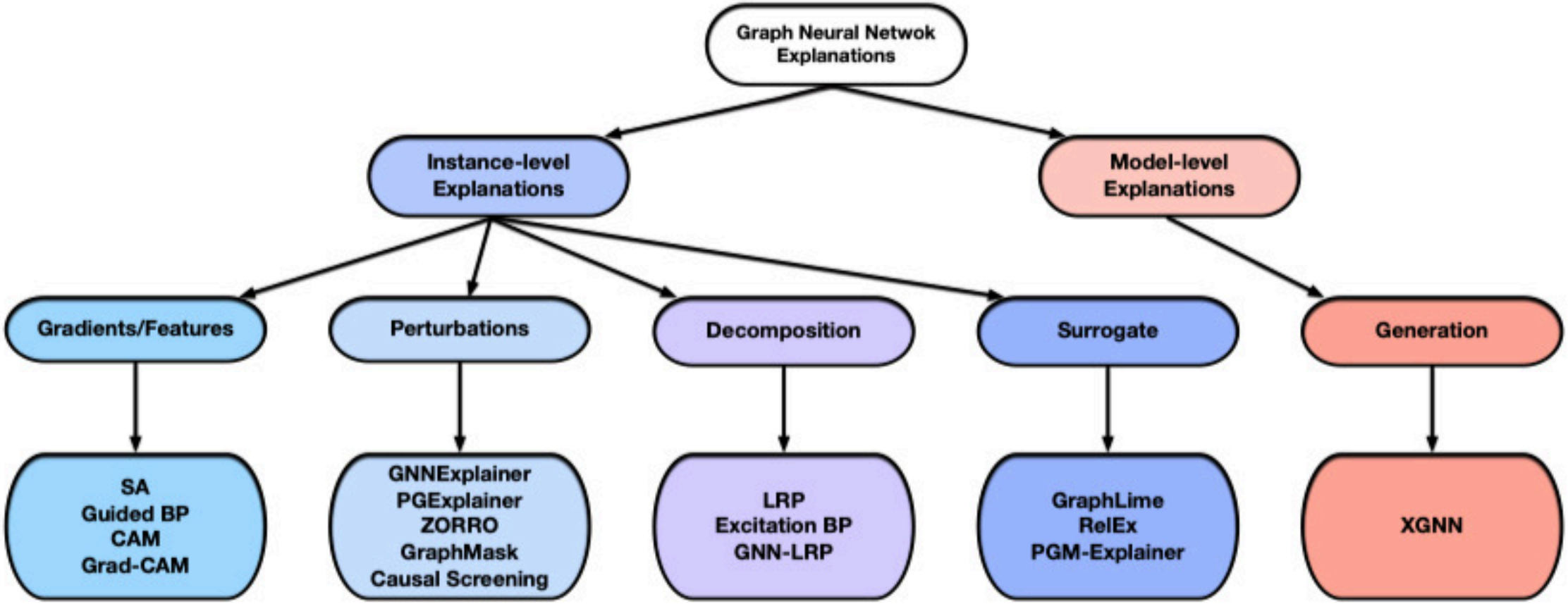
- For images and texts, we study the **importance** of **each pixel** or **word**.
- **It is more important to study the structural information for graph data:**
  - the **nodes** in graphs may be **unlabeled** and the **labels** of the whole graphs are determined by **graph structures**.
  - **Studying each node may be meaningless** since those unlabeled nodes contain no semantic meaning.
  - For graphs in biochemistry, neurobiology, ecology, and engineering, **graph substructures are highly related to their functionalities**.
    - Ex: **network motifs are the building blocks** of many complex networks.
  - Then **such structural information should not be ignored in the explanation tasks**.
    - However, existing methods from image domains **cannot provide explanations regarding the structures**.
  - For node classification tasks:
    - the **prediction** of each node is **determined. by different message walks from its neighbors**
    - Investigating such message walks is **meaningful** but challenging.
    - **None of the existing methods in the image domain can consider such walk information**, which needs further explorations.
  - graph data are **less intuitive** than images and texts. To understand deep models, **domain knowledge for the datasets is necessary**.
    - it is challenging for humans to understand the meaning of graphs.
    - In interdisciplinary areas such as chemistry and biology, there are many **unsolved mysteries and the domain knowledge is still lacking**.
    - **non-trivial to obtain human-understandable explanations** for graph models.
    - **need** of **standard datasets** and **evaluation metrics** for explanation tasks

Overview

- Explanation methods focus on different aspects of the graph models and provide different views to understand these models.
- They generally answer a few questions:
  - which input edges are more important?
  - which input nodes are more important?
  - which node features are more important?
  - what graph patterns will maximize the prediction of a certain class?

Taxonomy of methods

Based on what types of explanations are provided, different techniques are categorized into two main classes: instance-level methods and model-level methods.



INSTANCE LEVEL EXPLANATIONS

Gradients/Features-Based Methods

- Employing gradients or features to explain the deep models is the **most straightforward solution**, which is widely used in image and text tasks.
- **Key idea**: use **the gradients or hidden feature map** values as the **approximations of input importance**.
  - **Gradients-based methods** compute the gradients of target prediction with respect to input features by **back-propagation**.
  - **Features-based methods** map the **hidden features** to the **input space** via interpolation to measure importance scores.
  - Larger gradients or feature values indicate higher importance.
- **Methods**:
  - SA, Guided BP, CAM and Grad-CAM.
  - The key difference among these methods lies in the **procedure of gradient backpropagation** and how different hidden feature maps are **combined**.

- F. Baldassarre and H. Azizpour, "Explainability techniques for graph convolutional networks," in International Conference on Machine Learning (ICML) Workshops, 2019 Workshop on Learning and Reasoning with Graph-Structured Representations, 2019.
- P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 772–10 781.

Perturbation-Based Methods

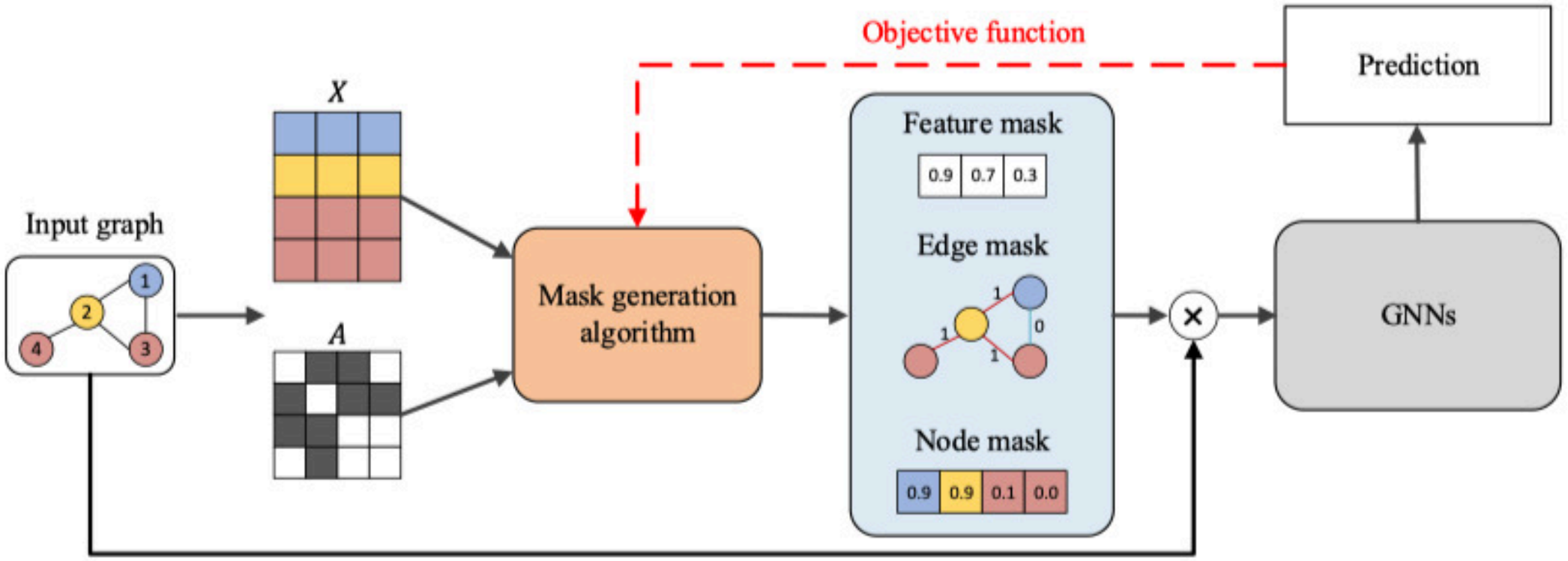


Fig. 2. The general pipeline of the perturbation-based methods. They employ different mask generation algorithms to obtain different types of masks. Note that the mask can correspond to nodes, edges, or node features. In this example, we show a soft mask for node features, a discrete mask for edges, and an approximated discrete mask for nodes. Then the mask is combined with the input graph to capture important input information. Finally, the trained GNNs evaluate whether the new prediction is similar to the original prediction and can provide guidance for improving the mask generation algorithms.

Methods

- **GNNExplainer** learns **soft masks** for edges and node features to explain the predictions via **mask optimization**. The soft masks are randomly initialized and treated as trainable variables.
- **PGExplainer** learns **approximated discrete masks** for edges to explain the predictions. It trains a parameterized mask predictor to predict edge masks.
- **GraphMask** is a **post-hoc method** for explaining the edge importance in each GNN layer. Similar to the PGExplainer, it trains a classifier to predict whether an edge can be dropped without affecting the original predictions.

- Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, “Gnnexplainer: Generating explanations for graph neural networks,” in Advances in neural information processing systems, 2019, pp. 92449255.
- D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, “Parameterized explainer for graph neural network,” in Advances in neural information processing systems, 2020.
- M. S. Schlichtkrull, N. De Cao, and I. Titov, “Interpreting graph neural networks for nlp with differentiable edge masking,” arXiv preprint arXiv:2010.00577, 2020.

Surrogate Methods

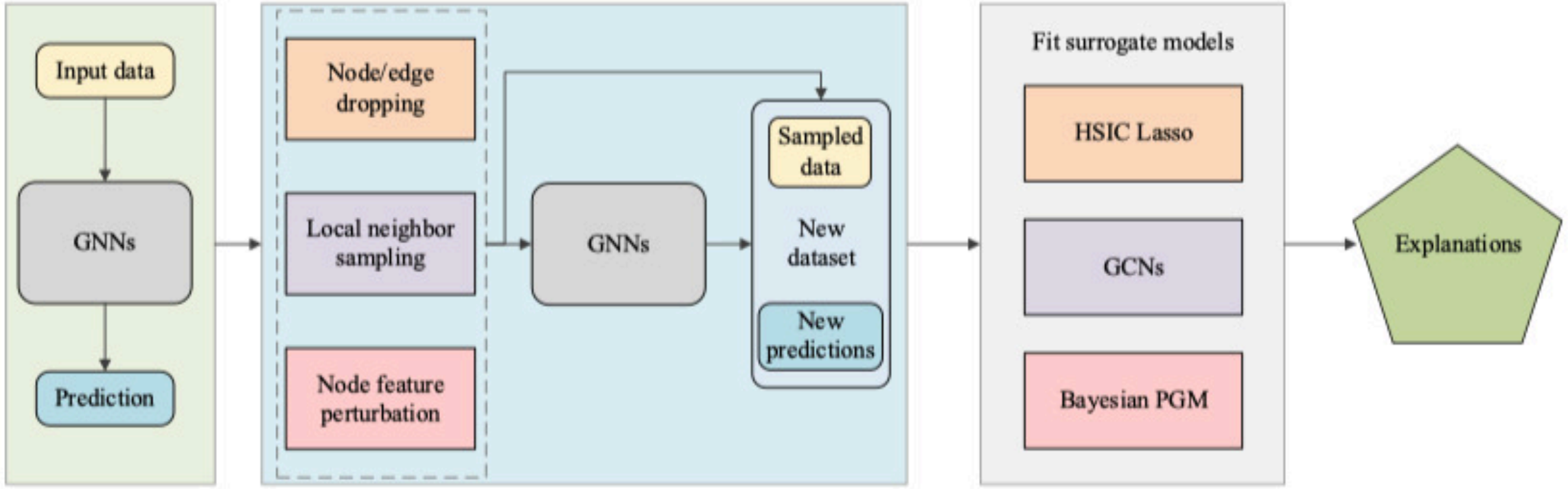


Fig. 3. The general pipeline of the surrogate methods. Given an input graph and its prediction, they first sample a local dataset to represent the relationships around the target data. Then different surrogate methods are applied to fit the local dataset. Note that surrogate models are generally simple and interpretable ML models. Finally, the explanations from the surrogate model can be regarded as the explanations of the original prediction.

Methods

- **GraphLime** extends the LIME algorithm to deep graph models and studies the importance of different node features for **node classification** tasks.
  - Given a target node in the input graph, GraphLime considers its N-hop neighboring nodes and their predictions as its local dataset where a reasonable choice of N is the number of layers in the trained GNNs.
  - Then a nonlinear surrogate model, Hilbert-Schmidt Independence Criterion (HSIC) Lasso [65], is employed to fit the local dataset.
  - HSIC Lasso is a kernel based feature selection algorithm.
  - Finally, based on the weights of different features in HSIC Lasso, it can select important features to explain the HSIC Lasso predictions. Those selected features are regarded as the explanations of the original GNN prediction.
  - GraphLime can only provide explanations for node features **but ignore graph structures**, such as nodes and edges, which are more important for graph data.
- Other methods (e.g., PGM-Explainer) use different strategy to sample the neigborhood and different surrogate models.

Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, and Y. Chang, “Graphlime: Local interpretable model explanations for graph neural networks,” arXiv preprint arXiv:2001.06216, 2020.

M. N. Vu and M. T. Thai, “Pgm-explainer: Probabilistic graphical model explanations for graph neural networks,” in Advances in neural information processing systems, 2020.

MODEL LEVEL EXPLANATIONS

Model level explanations

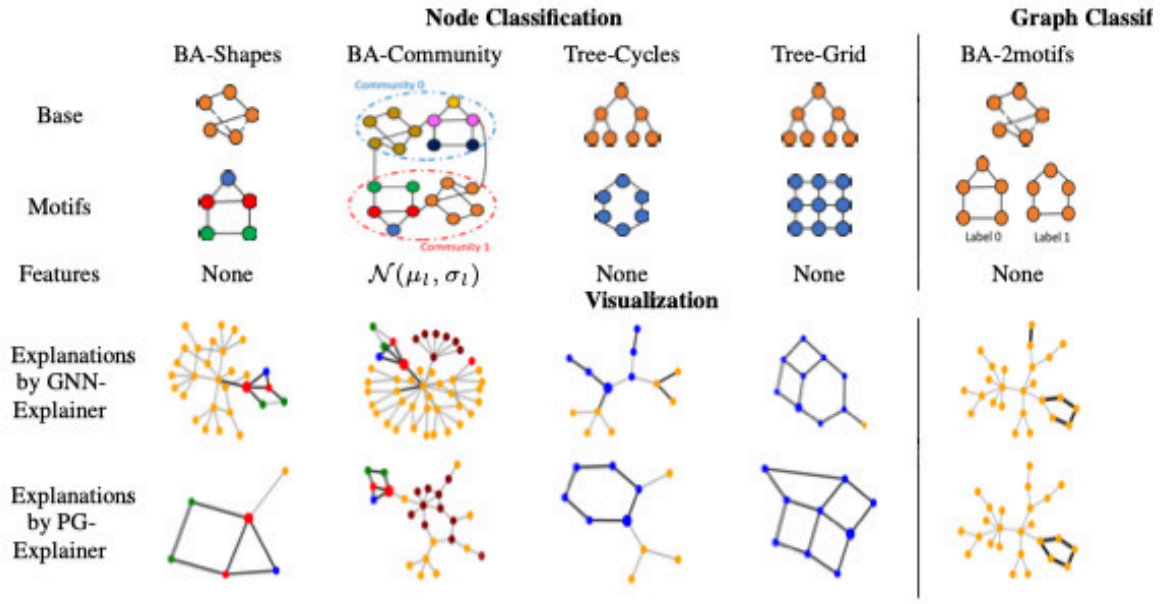
- Model-level methods aim at providing **the general insights and high-level understanding** to explain deep graph models.
- Study what input graph patterns can lead to a certain GNN behavior, such as maximizing a target prediction.
  - Input optimization is a **popular** direction to obtain model-level explanations for **image** classifiers.
  - it **cannot be directly applied to graph models** due to the **discrete** graph topology information
- **XGNN** proposes to explain GNNs via graph generation
  - Trains a graph generator so that the generated graphs can **maximize** a **target graph prediction**.
  - **Generated graphs** are regarded as the **explanations** for the target prediction and are expected to contain **discriminative graph patterns**.
  - The graph generation is formulated as a **reinforcement learning problem**
    - For each step, the generator **predicts how to add an edge** to the current graph.
    - Then the generated graphs are fed into the trained GNNs to obtain feedback to train the generator via policy gradient.
    - Several graph rules are incorporated to encourage the explanations to be both valid and human-intelligible.
- H. Yuan, J. Tang, X. Hu, and S. Ji, “XGNN: Towards model-level explanations of graph neural networks,” ser. KDD ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 430–438. [Online]. Available: https://doi.org/10.1145/3394486.3403085

EVALUATION

What is a good explanation?

- **Good** explanations should **faithfully** explain the **behaviors** of GNN models
- **Evaluating** the explanation results is **non trivial** due to **the lack of ground truths.**
- Need for :
  - **Datasets** with **ground truths**
  - **Evaluation metrics**

Synthetic datasets



Explanation AUC

GRAD	0.882	0.750	0.905	0.612	0.717
ATT	0.815	0.739	0.824	0.667	0.674
Gradient	-	-	-	-	0.773
GNNExplainer	0.925	0.836	0.948	0.875	0.742
PGExplainer	**0.963**±0.011	**0.945**±0.019	**0.987**±0.007	**0.907**±0.014	**0.926**±0.021
Improve	4.1%	13.0%	4.1%	3.7%	24.7%

Inference Time (ms)

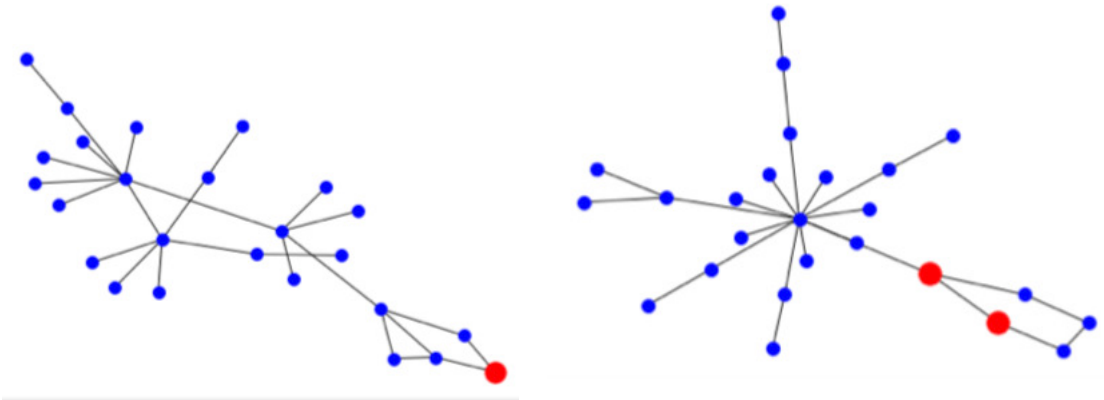
GNNExplainer	650.60	696.61	690.13	713.40	934.72
PGExplainer	10.92	24.07	6.36	6.72	80.13
Speed-up	59x	29x	108x	106x	12x

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In NeurIPS 2020

Explanation evaluation seen as a **classification pb**: a **good explanation must uncover the ground truth.**
$\Rightarrow$ Use of AUC, Precision, Recall, ...

Concerns:

- Too strong hypothesis (what we like to have but not what the model actually capture!)
- only contain **simple relation, not enough for comprehensive evaluation**



Real-world datasets

- **Sentiment graph data:**
  - From text sentiment analysis data (SST2, SST5, Twitter) to a graph that each node represents a word while the edges reflect the relationships between different words.
    - Easy to understand, yet **not enough for comprehensive evaluation**
- **Molecule data:**
  - Molecular datasets are also widely used in explanation tasks, such as MUTAG, BBBP, and Tox21.
  - Each graph in such datasets corresponds to a molecule where nodes represent atoms and edges are the chemical bonds.
  - The labels of molecular graphs are generally determined by the chemical functionalities or properties of the molecules.
  - Employing such datasets for explanation tasks requires **domain knowledge**, such as what **chemical groups are discriminative for their functionalities**.
  - In MUTAG, different graphs are labeled based on their mutagenic effects on a bacterium.
  - Known that **carbon rings and NO 2 chemical** groups may lead to mutagenic effects.
  - ***Then we can study whether the explanations can identify such patterns for the corresponding class.***
  - ***Is the domain knowledge exhaustive ? No !***

Metrics

The **Fidelity** metric studies **the prediction change by removing important** nodes/edges/node features

$$Fidelity^{acc} = \frac{1}{N}\sum_{i=1}^{N}(\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{1-m_i} = y_i)),$$

Fidelity^acc metric studies the change of prediction accuracy (i.e., the model prediction changes).

$$Fidelity^{prob} = \frac{1}{N}\sum_{i=1}^{N}(f(\mathcal{G}_i)_{y_i} - f(\mathcal{G}_i^{1-m_i})_{y_i}),$$

Fidelity^prop focuses on the predicted probability

Infidelity studies prediction change by keeping important input features and removing unimportant features.

$$Infidelity^{acc} = \frac{1}{N}\sum_{i=1}^{N}(\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{m_i} = y_i)),$$

Important features should contain **discriminative information** so that they should lead to similar predictions as the original predictions even unimportant features are removed

$$Infidelity^{prob} = \frac{1}{N}\sum_{i=1}^{N}(f(\mathcal{G}_i)_{y_i} - f(\mathcal{G}_i^{m_i})_{y_i}),$$

Metrics

Good explanations should be sparse, which means they should capture the most important input features and ignore the irrelevant ones.

The metric Sparsity measures such a property.

$$Sparsity = \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{|m_i|}{|M_i|}),$$

where $|m_i|$ denotes the number of important input features (nodes/edges/node features) identified in m_i and $|M_i|$ means the total number of features in Gi.

END

Conclusion

- Importance of providing explanations
- Taxonomy of methods
- Difficulty to assess good explanation
- Many challenges still opened
  - GNN introspection
  - From a pattern mining perspective: define new pattern languages to “open the black box”
  - …

Exam (DM part)

- **Pattern mining:**
  - Frequent pattern mining (Apriori)
  - Be able to perform an extraction with Monotone/Antimonotone/Convertible constraints (with Depthfirst enumeration if relevant)
  - Output space sampling (possible but only in an open question without too many calculations)
- **Clustering:**
  - Be able to perform a kmeans or hierarchical clustering
- Possibility to have **open question:** (some problem with some generalization of what we studied)
  - E.g., sequence mining …
  - In that case, every new concept will be defined.
- Documents **allowed iff allowed** for DB part (To be checked).
- **Good luck !**