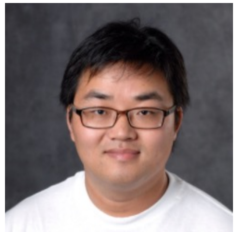
Introduction



Jiliang Tang

Privacy

Xiaorui Liu

Safety & Robustness

Yaxin Li

Explainability

Wenqi Fan

Non-discrimination & Fairness

Environmental Well-being

Haochen Liu

Accountability & Auditability

Dimension Interactions

Future Directions

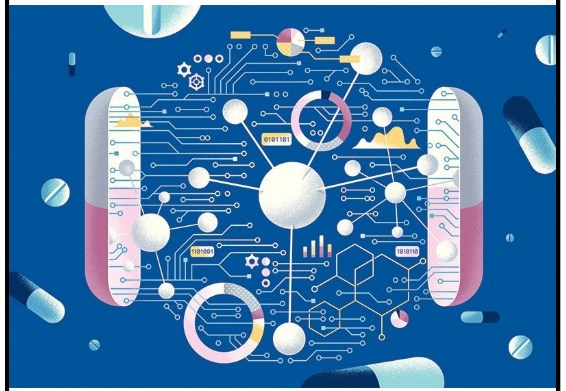
Yiqi Wang

AI in Critical Systems

Transportation

Finance

Security



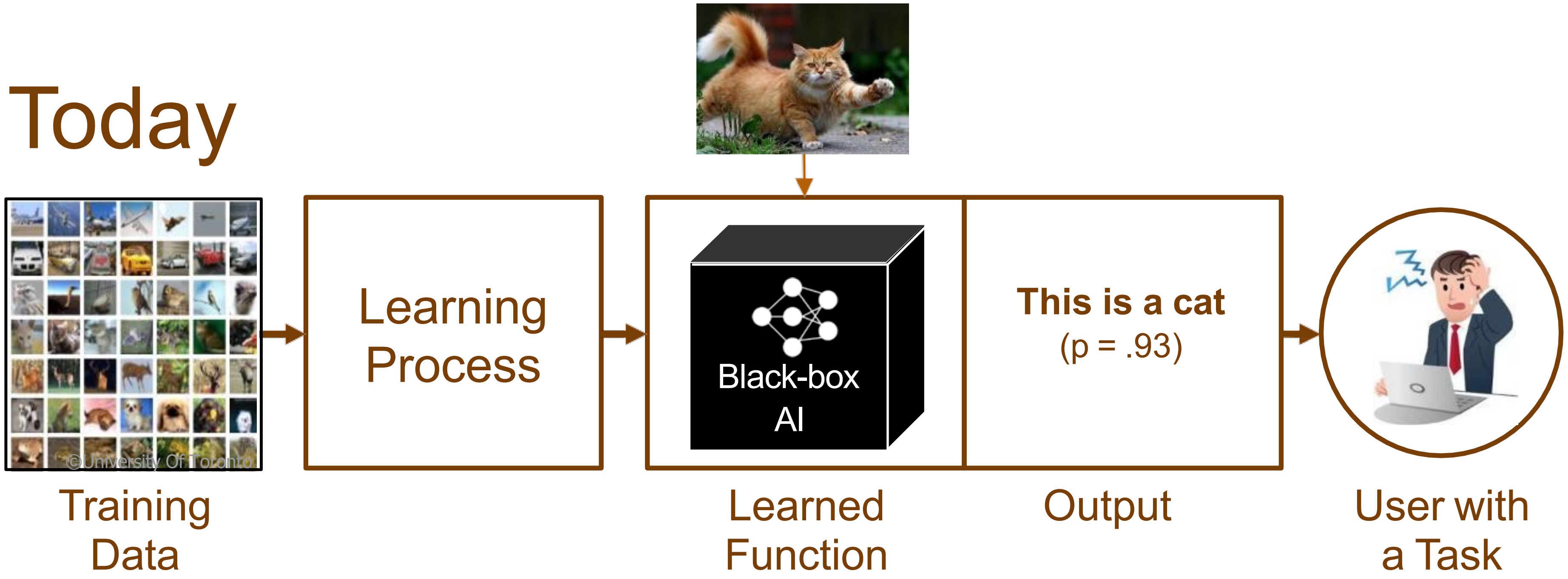
Medicine

Military

Legal

How an AI model works?

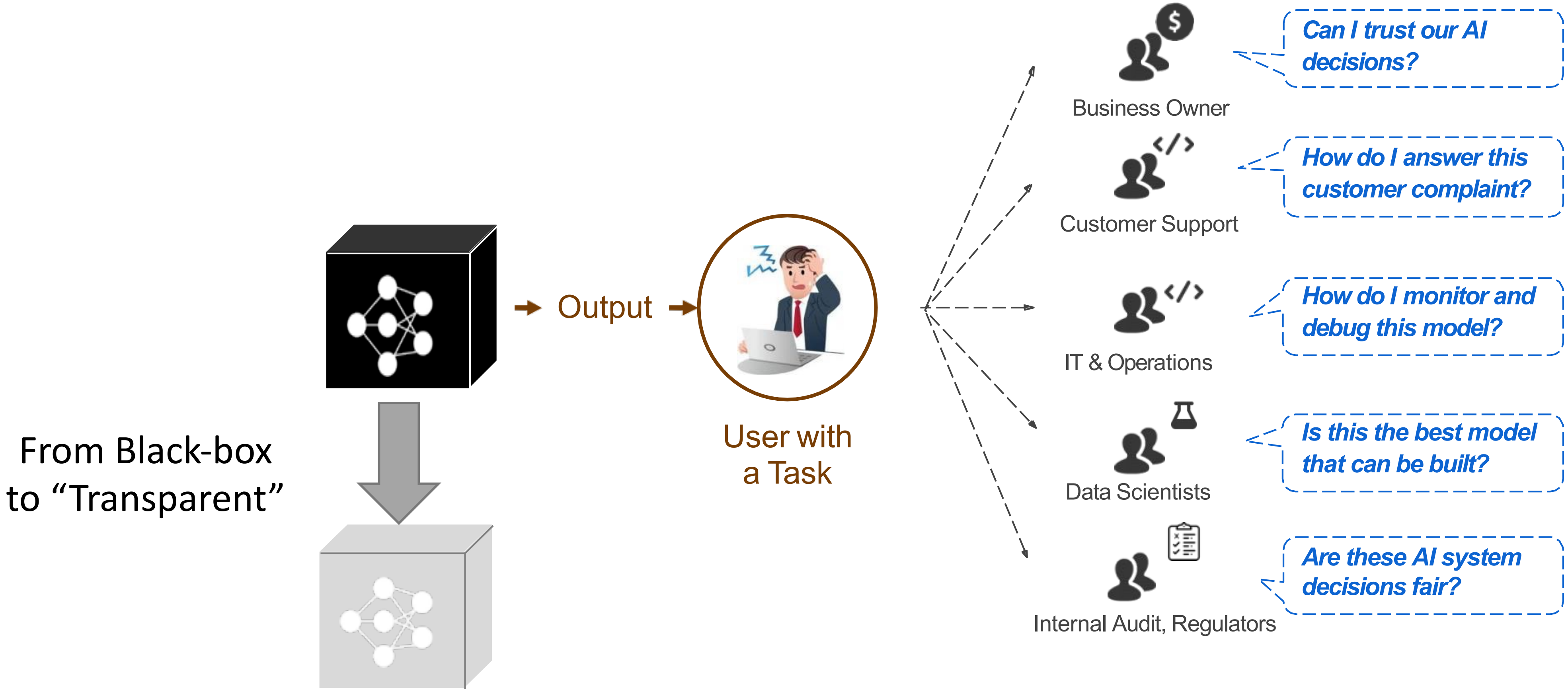
Today



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Gunning, David, et al. “Explainable Artificial Intelligence Research at DARPA”, 2019.

Black-box AI creates confusion and doubt

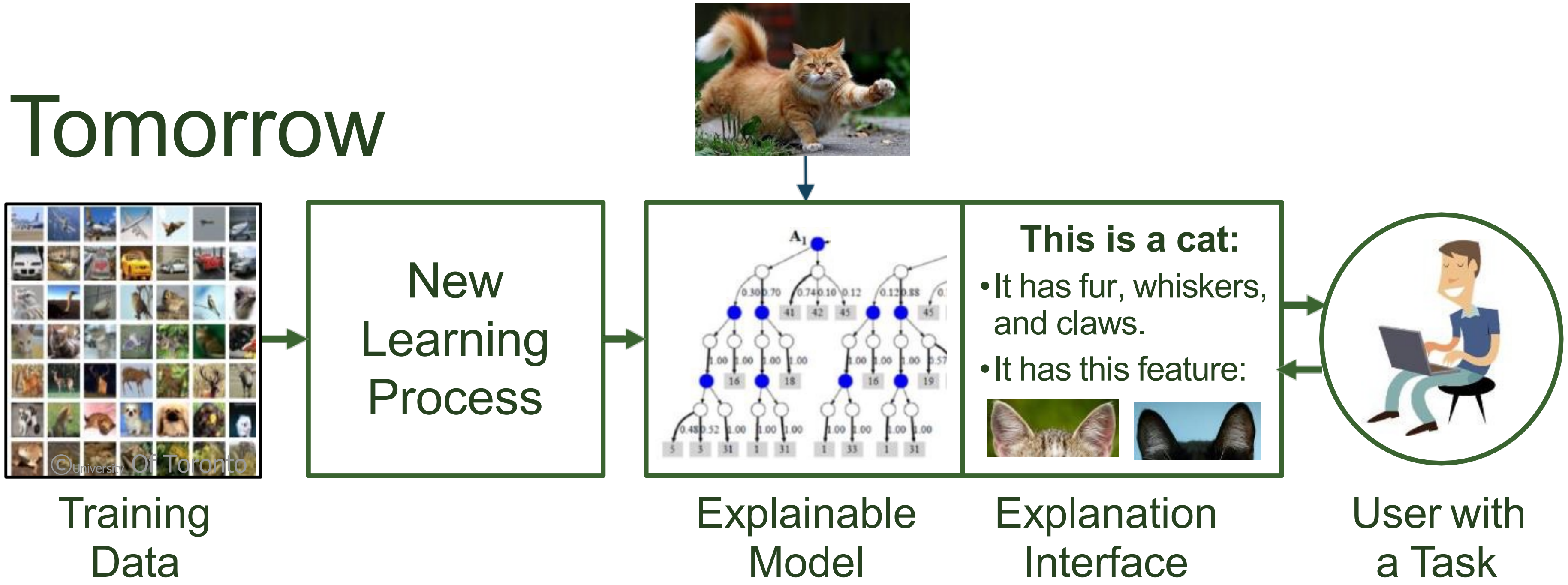


The Need for Explainable AI

Lecue, Freddy, et al. "Explainable ai: Foundations, industrial applications, practical challenges, and lessons learned.", 2020.

Explainable AI

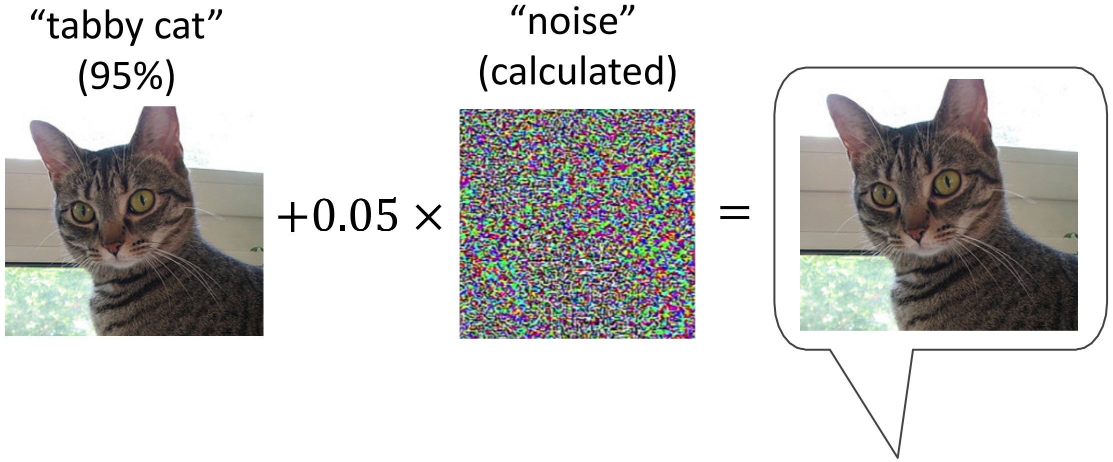
Tomorrow

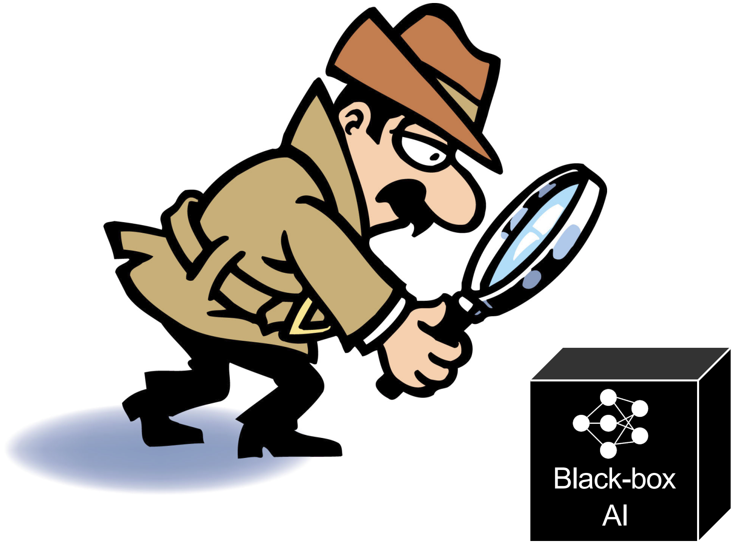


- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Gunning, David, et al. "Explainable Artificial Intelligence Research at DARPA", 2019.

Why Explainability: Debug (Mis-)Predictions





Top label: **“strawberry” (99%)**

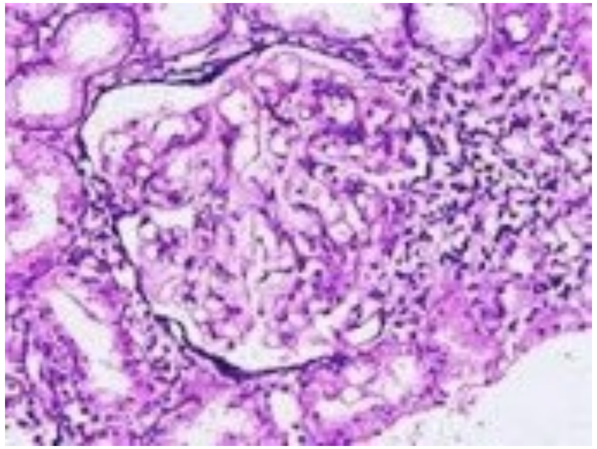
Why did the network label this image as “**strawberry**”?

Why Explainability: Verify the AI System

Wrong decisions can be costly and dangerous.

"Autonomous car crashes, because it wrongly recognizes …"

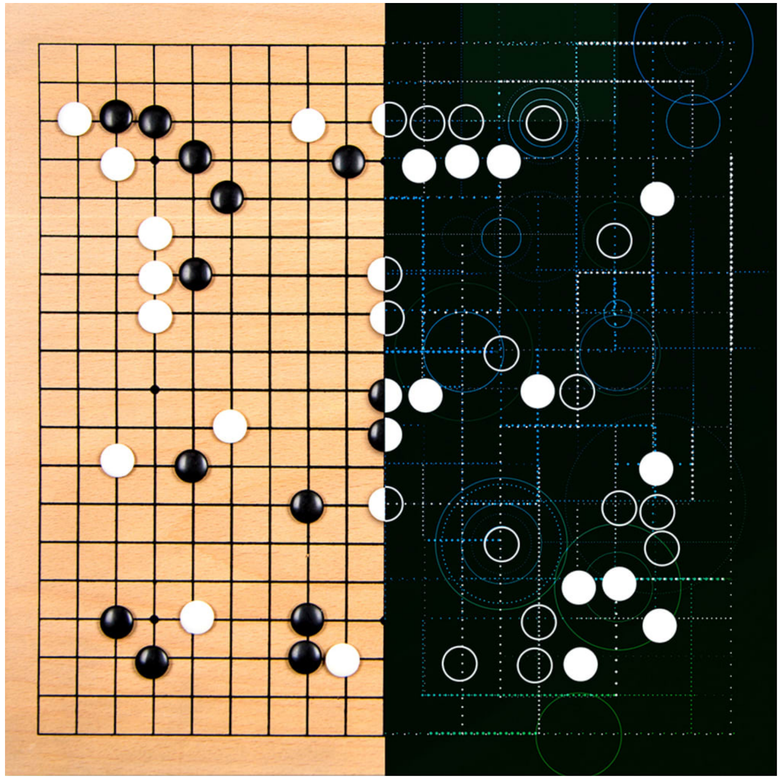
"AI medical diagnosis system misclassifies patient's disease …"



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… so beautiful." -- Fan Hui vs. AlphaGo



Outline

CONCEPTS AND TAXONOMY

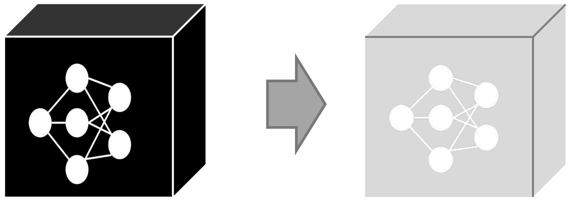
TECHNIQUES FOR EXPLAINABILITY IN AI (XAI)

APPLICATIONS IN REAL SYSTEMS

SURVEYS AND TOOLS

What is Explainable AI (XAI)?

- The degree to which a human can understand the cause of a decision.
  - **Interpretable** AI: intrinsically transparent and interpretable, rather than black-box/opaque models, such as decision trees and linear regression.
  - **Explainable** AI: additional (post hoc) explanation techniques, but still black-box and opaque, such as DNN.



From Black-box to “Transparent”

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences.", 2019.
Gilpin, Leilani H., et al. "Explaining explanations: An overview of interpretability of machine learning.", 2018.

Taxonomy

- **Model usage**: model-intrinsic and model-agnostic
  - Only restrict to a specific architecture of an AI model or not
- Differences in the **methodology**: gradient-based and perturbation-based
  - Employ the partial derivatives on inputs or change input data
- **Scope** of explanation: local and global
  - Provide an explanation only for a specific instance or for the whole model
- **Counterfactual** explanations
  - "If X had not occurred, Y would not have occurred."

Outline

CONCEPTS AND TAXONOMY

TECHNIQUES FOR EXPLAINABILITY IN AI

(XAI)

APPLICATIONS IN REAL SYSTEMS

SURVEYS AND TOOLS

Model usage

- **Only restrict to a specific architecture of an AI model or not**

- Model-intrinsic Explanations
  - Transparent or white-box explanation (model-specific)
- Model-agnostic Explanations
  - Interpret already well-trained models
  - Post-hoc or black-box explainability methods

Model usage: Model-intrinsic Explanations

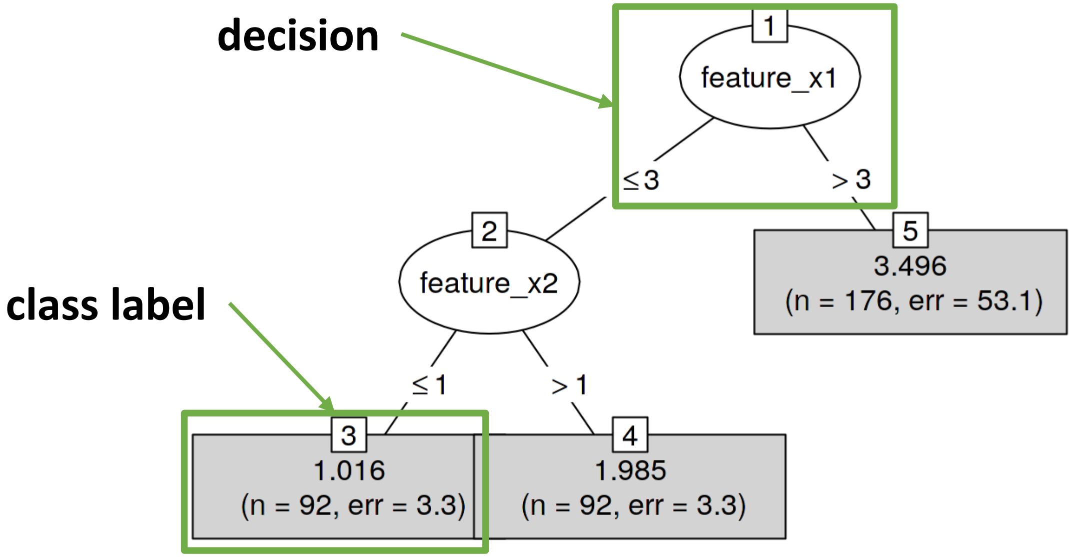
❑Transparent, or white-box explanation (model-specific)

- linear/logistic regression, decision trees, rule-based models, etc.

$$\hat{y} = \mathbf{w}^T\mathbf{x} + b = w_1x_1 + \ldots + \boxed{w_d}x_d + b$$

feature weight

linear regression model



Decision tree

Model usage: Model-agnostic Explanations

- Interpret already well-trained models
  - Post-hoc or black-box explainability methods

- Local Interpretable Model-Agnostic Explanations (LIME)
  - Approximating the black-box model by an interpretable one (such as linear model) learned on perturbations of the original instance.

$$\text{explanation}(x) = \arg\min_{\boxed{g \in G}} L(f, g, \pi_x) + \boxed{\Omega(g)}$$

Interpretable model
(linear models/decision tree, etc)

Model complexity

Ribeiro, Marco Tulio, et al. "" Why should i trust you?" Explaining the predictions of any classifier.", 2016.

Model usage: Model-agnostic Explanations

LIME:

Original Image

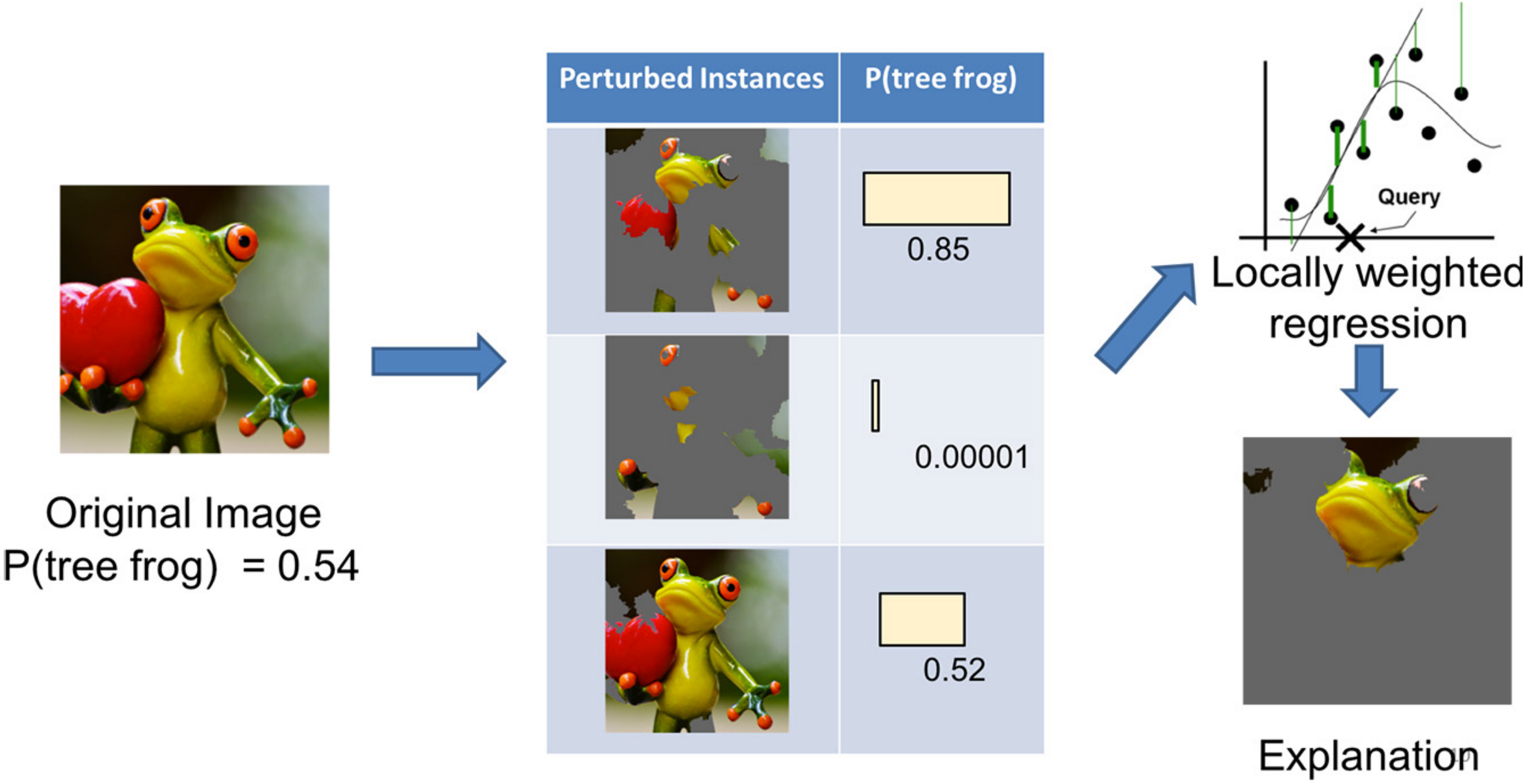
Interpretable Components

Transforming an image into interpretable components

Ribeiro, Marco Tulio, et al. "" Why should i trust you?" Explaining the predictions of any classifier.", 2016.

Model usage: Model-agnostic Explanations

LIME:



Ribeiro, Marco Tulio, et al. "" Why should i trust you?" Explaining the predictions of any classifier.", 2016.

Differences in the methodology

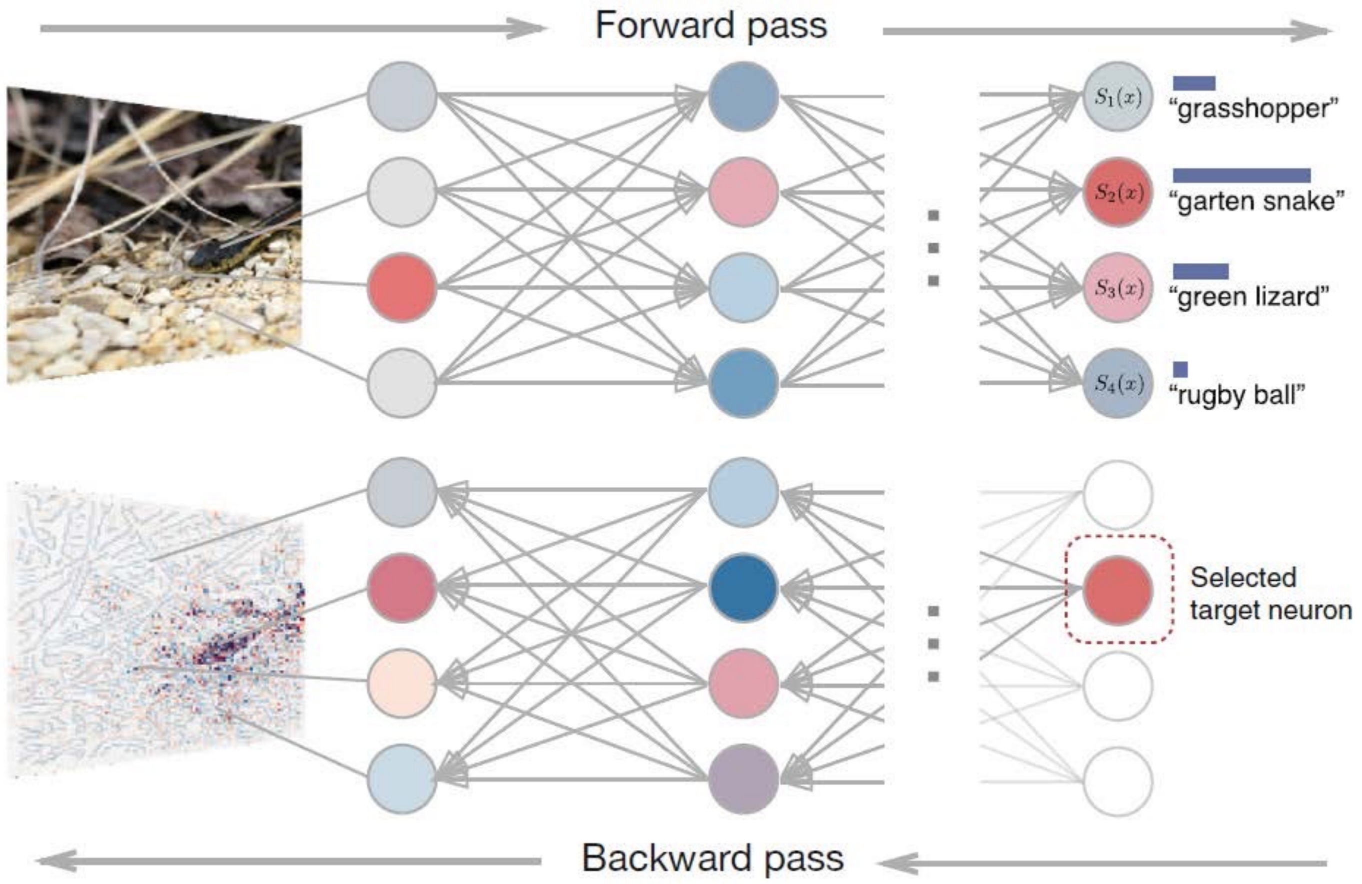
- ❑ Employ the partial derivatives on inputs or change input data

- ❑ Gradient-based Explanations
  - Combine network activations and gradients

- ❑ Perturbation-based Explanations
  - Change the input and observe the effect on the output

Methodology: Gradient-based Explanations

- Forward pass and back-propagation
  - Class activation mapping (CAM), Grad-GAM

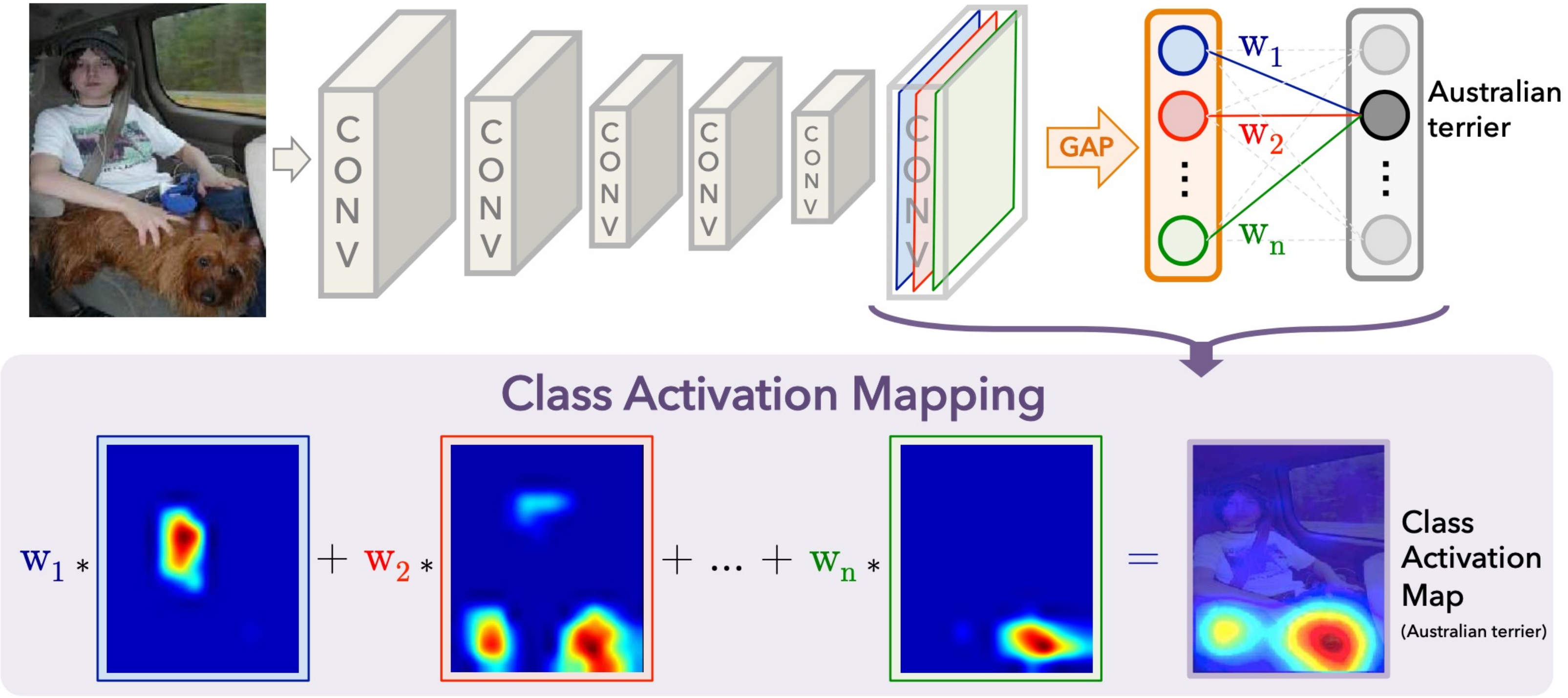


Zhou, Bolei, et al. "Learning deep features for discriminative localization.", 2016.

Methodology: Gradient-based Explanations

- ❑ Forward pass and back-propagation
  - Class activation mapping (CAM), Grad-GAM

GAP: Global Average Pooling



Zhou, Bolei, et al. "Learning deep features for discriminative localization.", 2016.

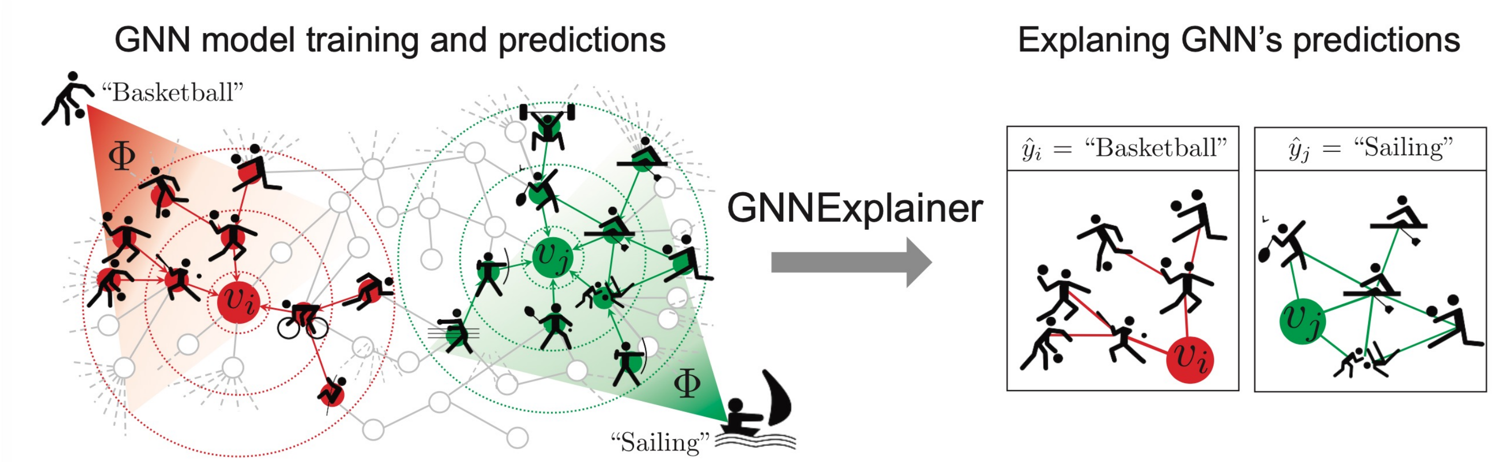
Methodology: Gradient-based Explanations

- Forward pass and back-propagation
  - Class activation mapping (CAM), Grad-GAM

Zhou, Bolei, et al. "Learning deep features for discriminative localization.", 2016.
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization.", 2017.

Methodology: Perturbation-based Explanations

- Change the input and observe the effect on the output
  - GNNExplainer on Graphs
    - A **small subgraph** of the input graph that are most influential for target prediction



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

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

Computation graph

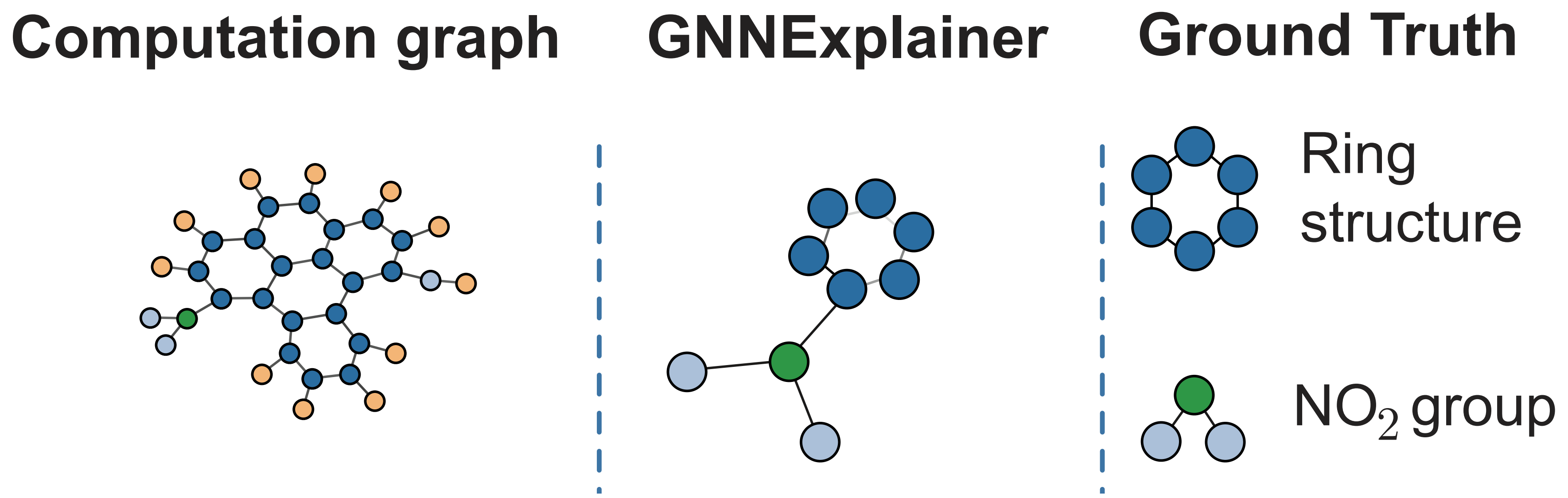
(Soft) Mask matrix

Ying, Rex, et al. "Gnnexplainer: Generating explanations for graph neural networks.", 2019

Methodology: Perturbation-based Explanations

- ❑Change the input and observe the effect on the output
  - GNNExplainer on Graphs
    - A **small subgraph** of the input graph that are most influential for target prediction

Molecular (atoms: hydrogen/carbon and bonds)



Ying, Rex, et al. "Gnnexplainer: Generating explanations for graph neural networks.", 2019

Scope of Explanation

- Provide an explanation only for a specific instance or for the whole model

- Local Explanations
  - Explain a specific instance

- Global Explanations
  - Explain the whole model or a class

Scope: Local Explanations

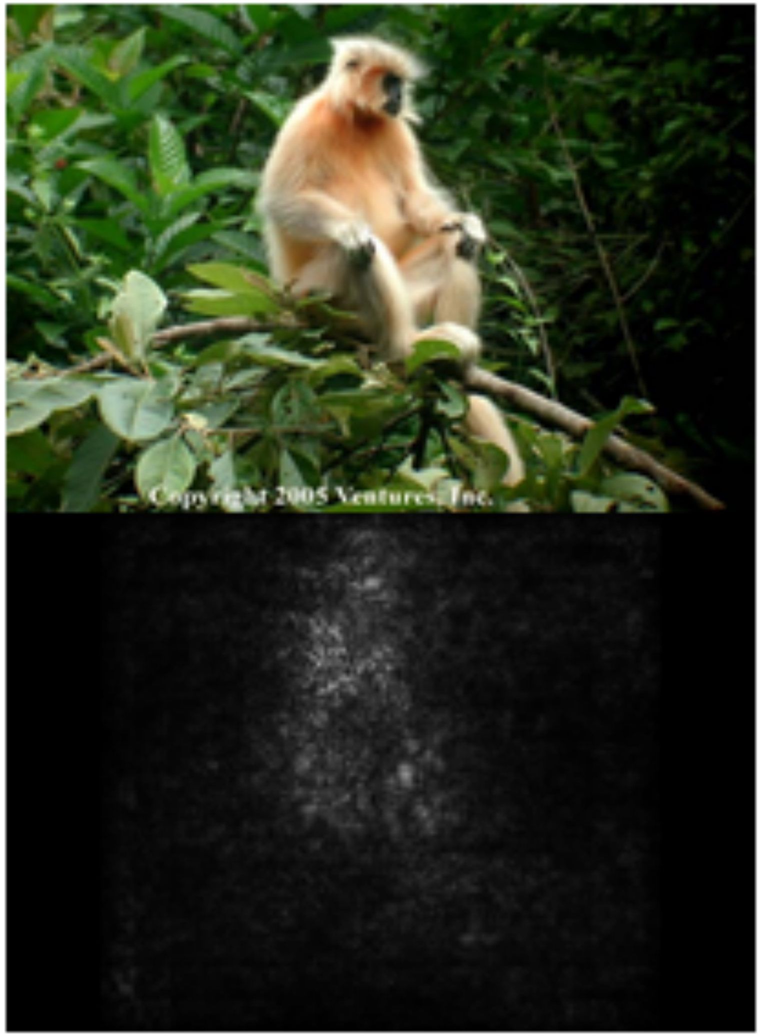
❑Explain a specific instance

- Image-Specific Saliency Map

$$SaliencyMap = gradient = \frac{\partial\, class\ score}{\partial\, input\ image}$$

$$I^* = argmax_I S_y(I) - R(I)$$

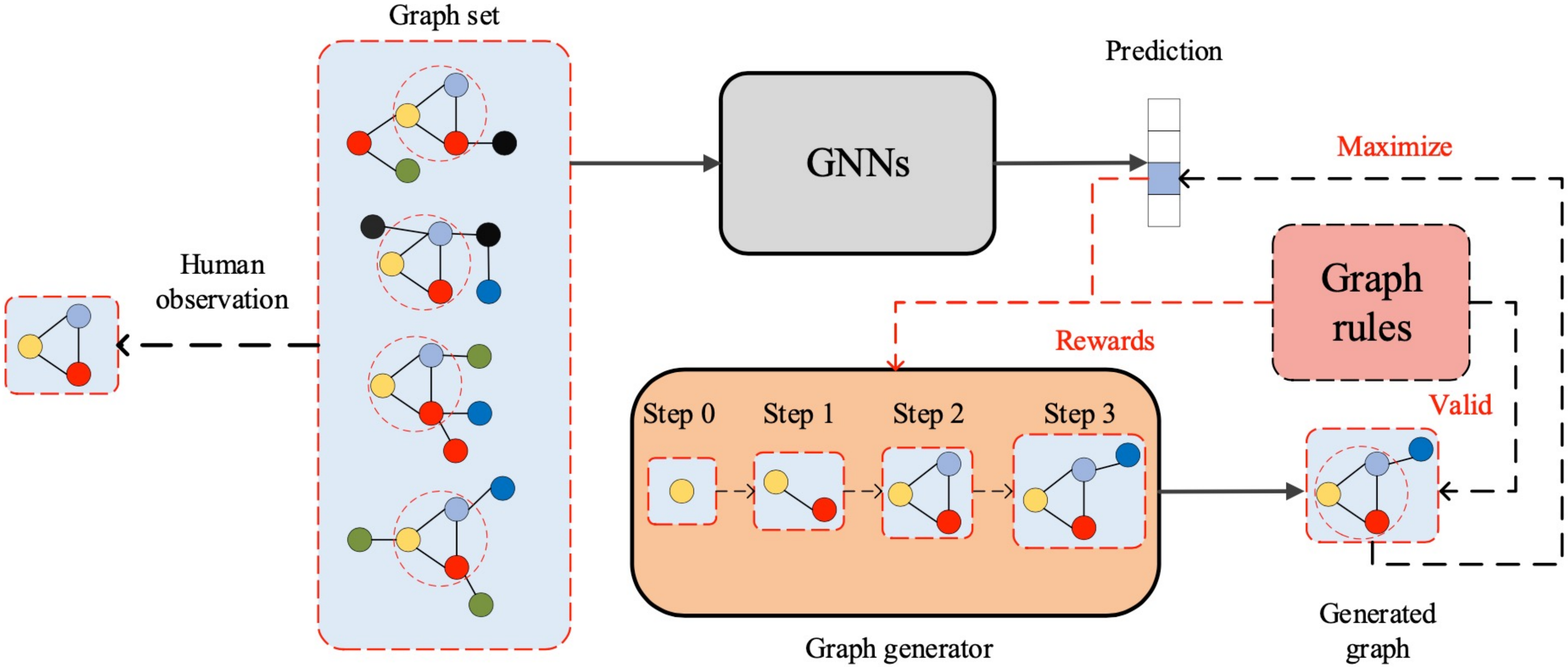
"Why is a given image classified as a **monkey?"**



Karen Simonyan, et al. "Deep inside convolutional networks: Visualising image classification models and saliency maps.", 2013

Scope: Global Explanations

- Explain the whole model or a class
  - XGNN: Model/Global-level Explanations on Graphs
    - Explain what **graph patterns** lead to a certain prediction (e.g., motifs)

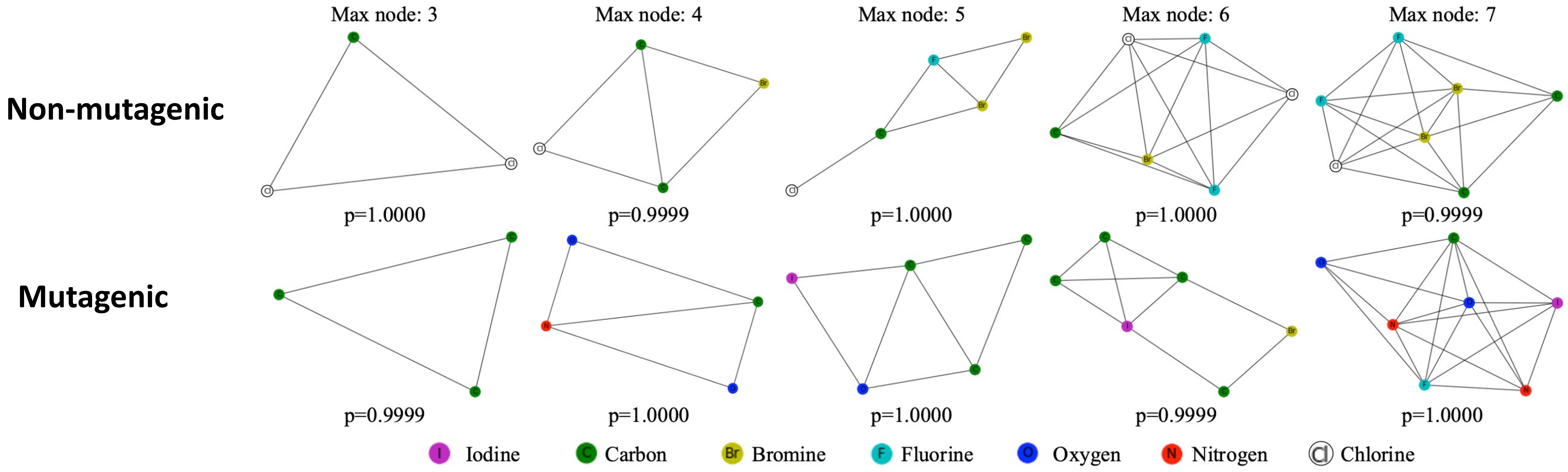


Yuan, Hao, et al. "XGNN: Towards Model-Level Explanations of Graph Neural Networks.", 2020

Scope: Global Explanations

XGNN: Model/Global-level Explanations on Graphs

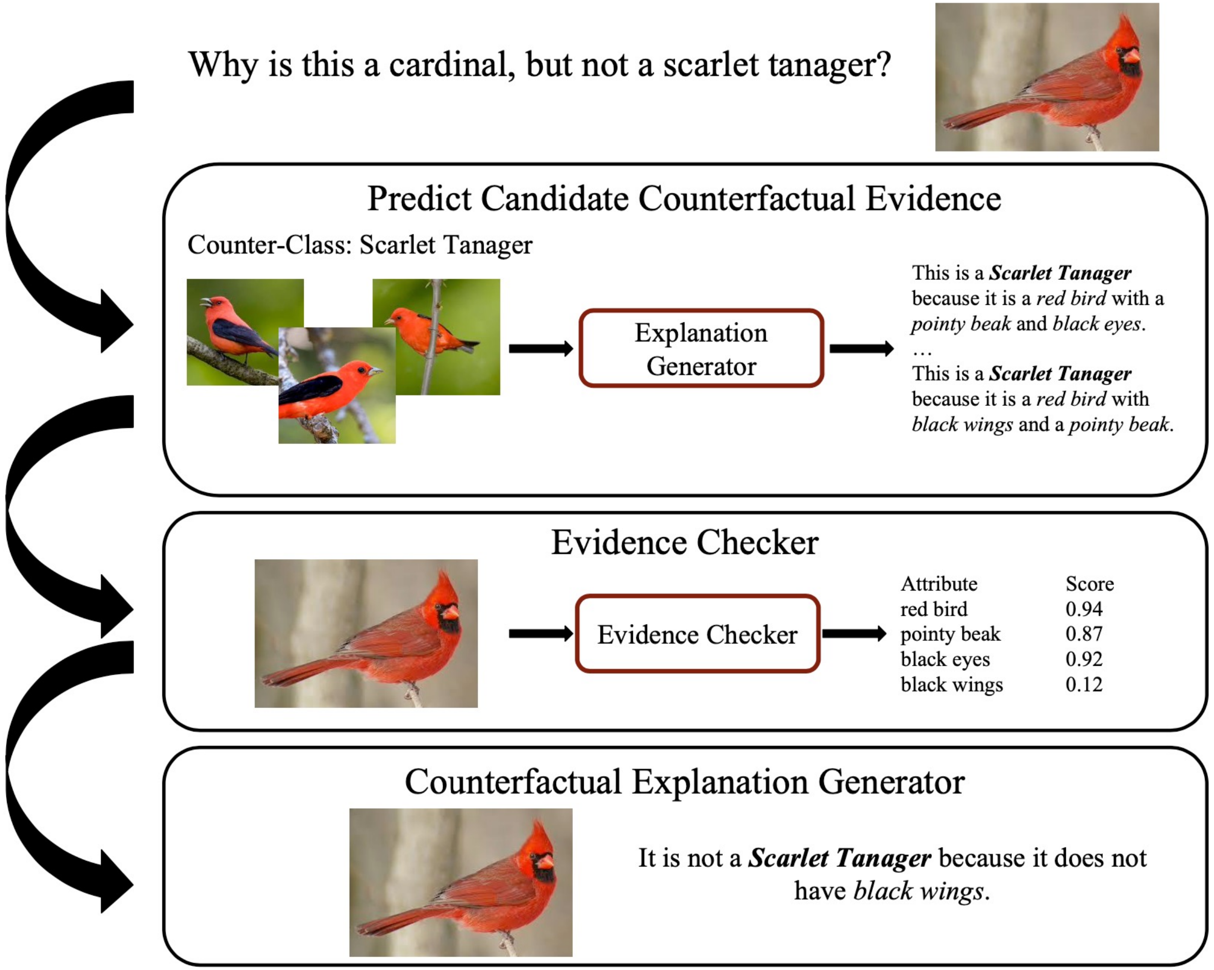
MUTAG (molecular: atoms/bonds)



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

Counterfactual Explanations

- ❑ Causal situation: "If X had not occurred, Y would not have occurred".



Hendricks, Lisa Anne, et al. "Generating Counterfactual Explanations with Natural Language.", 2018

Outline

CONCEPTS AND TAXONOMY

TECHNIQUES FOR EXPLAINABILITY IN AI

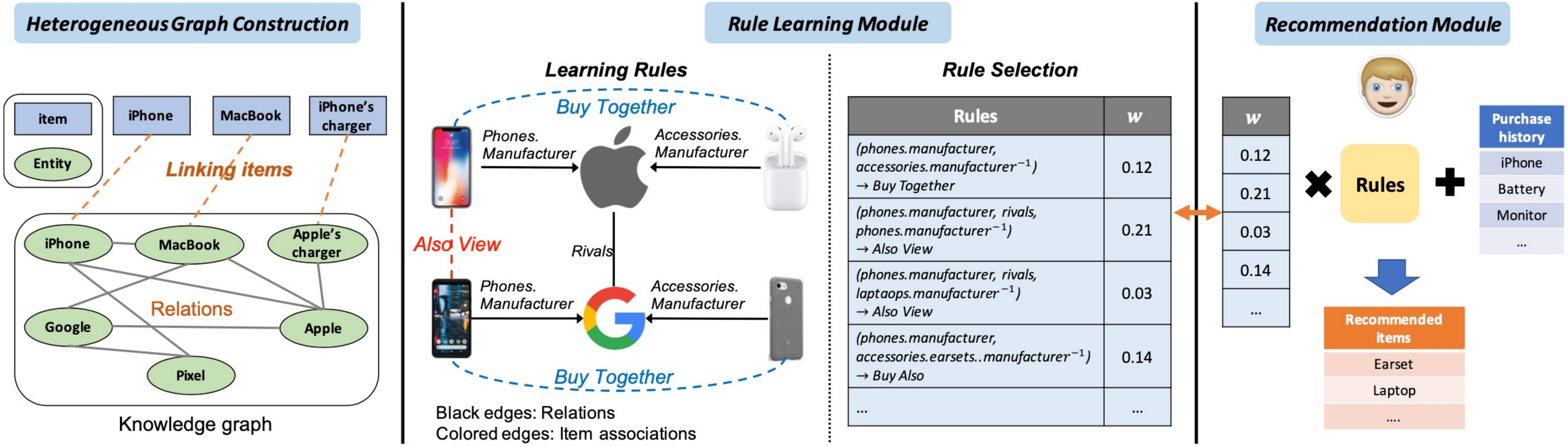
(XAI)

APPLICATIONS IN REAL SYSTEMS

SURVEYS AND TOOLS

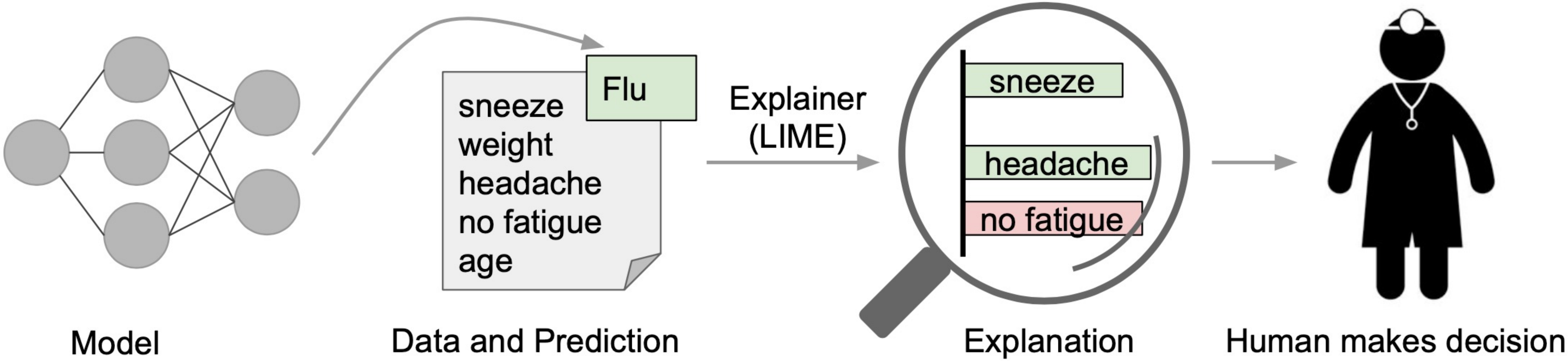
Recommender Systems

Explanations: Frequently Buy together, Also view, Buy after view, and Also buy, etc.



Ma, Weizhi, et al. " Jointly Learning Explainable Rules for Recommendation with Knowledge Graph.", 2019

Natural Language Processing (NLP)



Ribeiro, Marco Tulio, et al. "" Why should i trust you?" Explaining the predictions of any classifier.", 2016.

Outline

CONCEPTS AND TAXONOMY

TECHNIQUES FOR EXPLAINABILITY IN AI

(XAI)

APPLICATIONS IN REAL SYSTEMS

SURVEYS AND TOOLS

Surveys

- Doshi-Velez, Finale, et al. "Towards a rigorous science of interpretable machine learning.", 2017.
- Guidotti, Riccardo, et al. "A survey of methods for explaining black box models.", 2018.
- Du, Mengnan, et al. "Techniques for interpretable machine learning.", 2019.
- Belle, Vaishak, et al. "Principles and practice of explainable machine learning.", 2020
- Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences.", 2019
- Molnar, Christoph. "Interpretable machine learning.", 2020
- Yuan, Hao, et al."Explainability in Graph Neural Networks: A Taxonomic Survey.", 2020
- Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.",, 2020
- Linardatos, Pantelis, et al. "Explainable ai: A review of machine learning interpretability methods.", 2021
- …

Liu, Haochen, et al. "Trustworthy AI: A Computational Perspective." , 2021.

Tools

AIX360	• https://aix360.mybluemix.net
InterpretML	• https://github.com/interpretml/interpret
DeepExplain	• https://github.com/marcoancona/DeepExplain
DIG for graph deep learning research	• https://github.com/divelab/DIG

Liu, Haochen, et al. "Trustworthy AI: A Computational Perspective." , 2021.

Future Directions

- Security of explainable AI
- Evaluation methodologies
- Knowledge to target model: from white-box to black-box