RICE ENGINEERING AND COMPUTING
Department of Computer Science

DATA LAB

Introduction to Interpretable Machine Learning

Xia "Ben" Hu
Associate Professor
Department of Computer Science, Rice University

Outline

1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

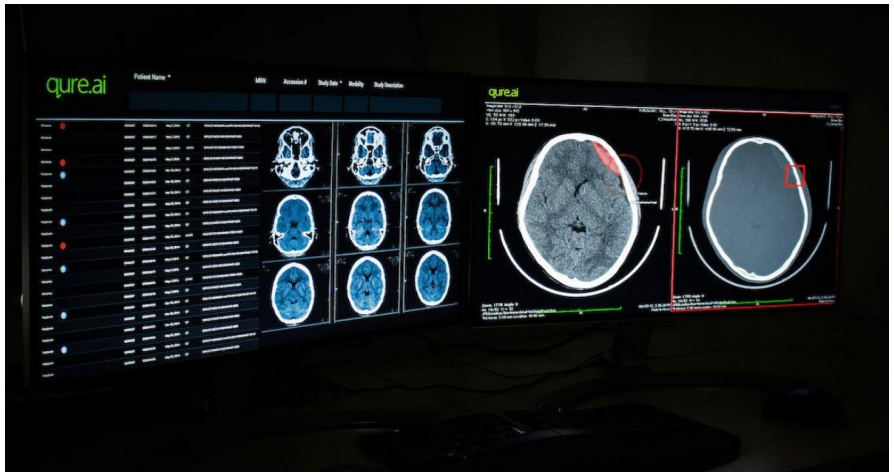
Outline

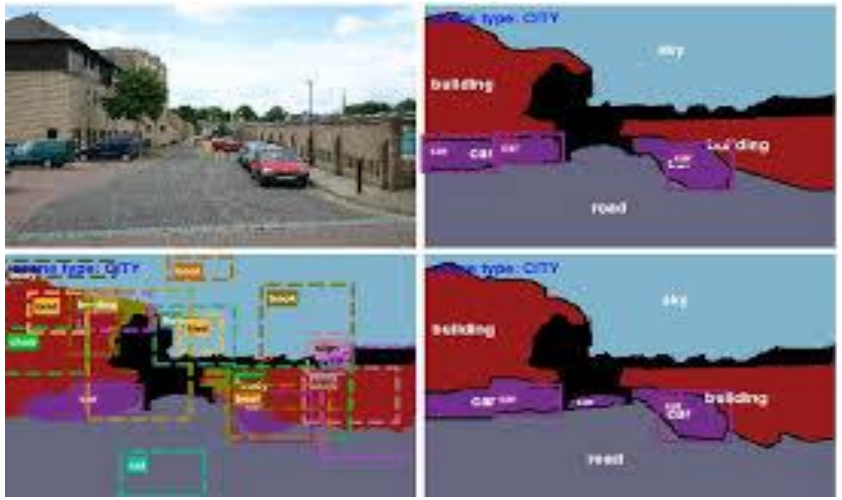
1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

Machine Learning is Everywhere

Playing Go

Medical Diagnosis





Scene Understanding

Voice Recognition

Machine Learning is Everywhere

What have been learned inside the models?

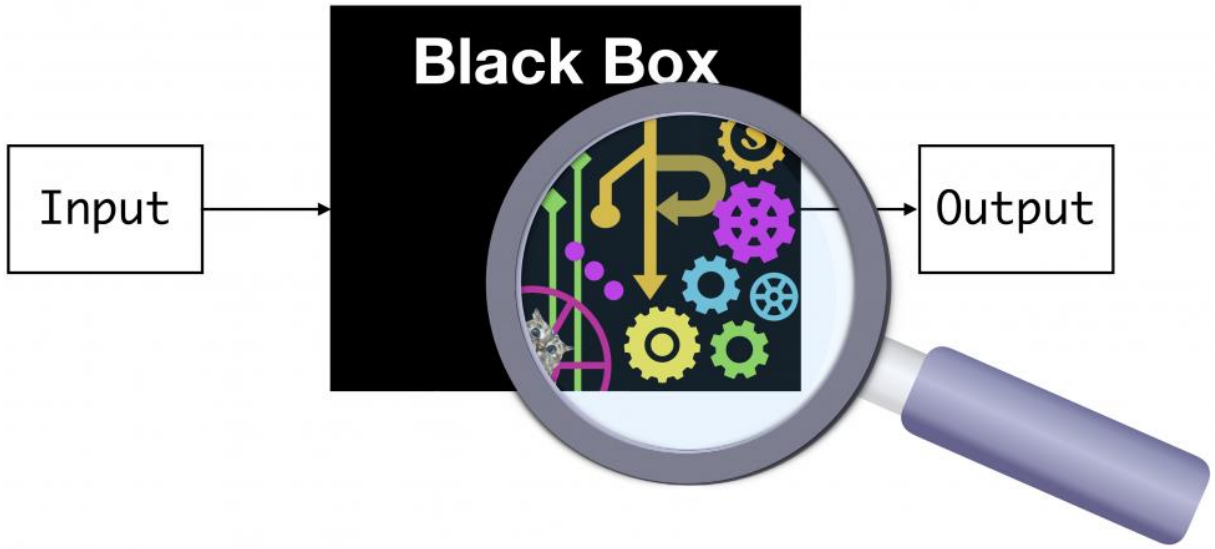
Interpretable Machine Learning

Safety of AI Models

Trust of AI Decision

Policy and Regularization

What is Interpretable Machine Learning



Interpretable Machine Learning is the ability to explain or to present the behavior of a black-box ML model in understandable terms to a human

Explaining a Black-box Using Deep Variational Information Bottleneck Approach

Interpretation Case

Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

Input X

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

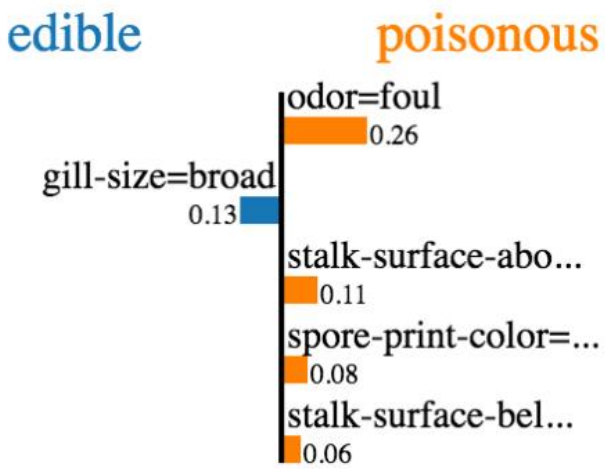
LASSO Model

Prediction probabilities

edible 0.00

poisonous 1.00

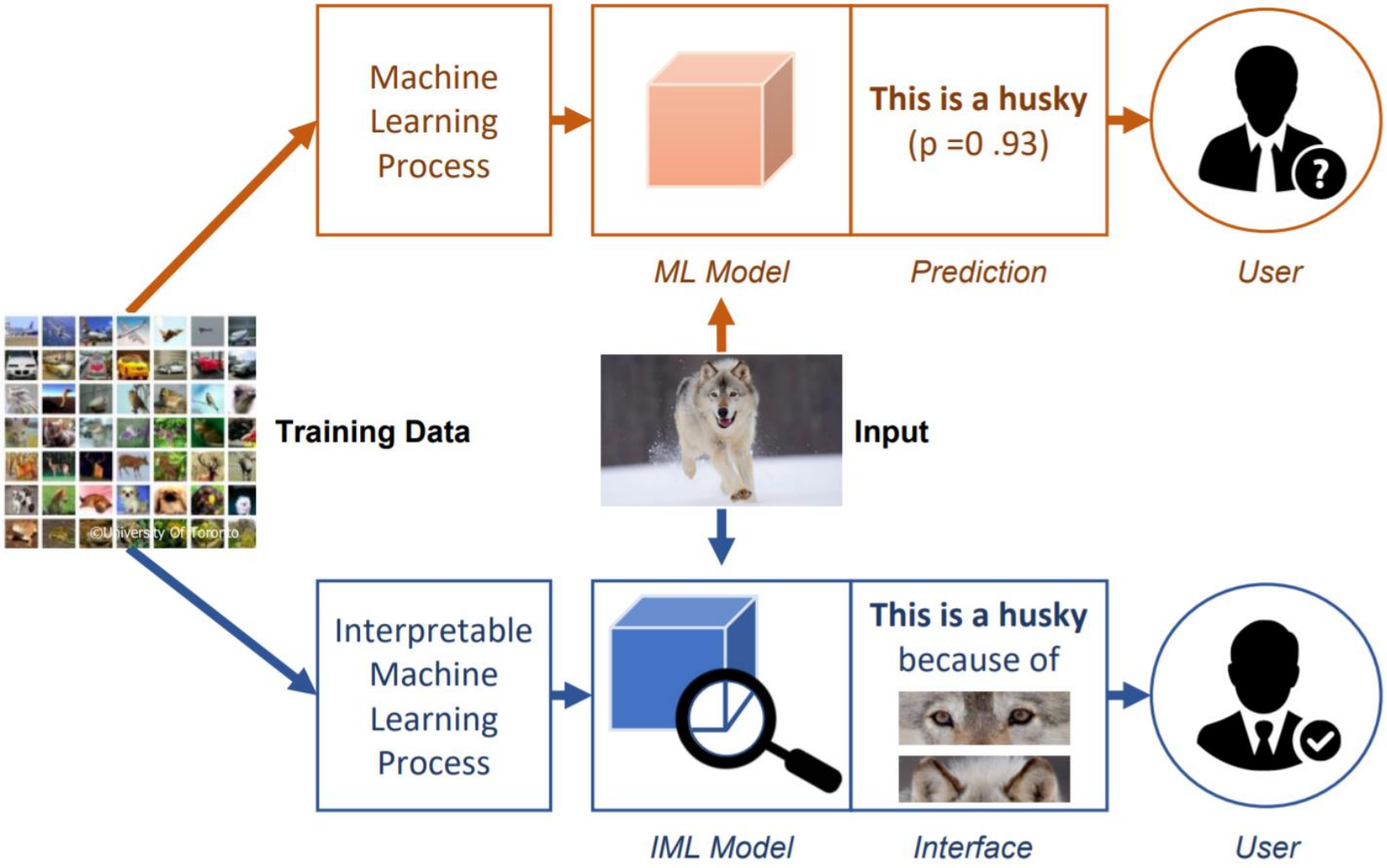
Output y



Interpretation to LASSO model:
Feature importance vector of
The linear weight β

https://github.com/marcotcr/lime

Pipeline

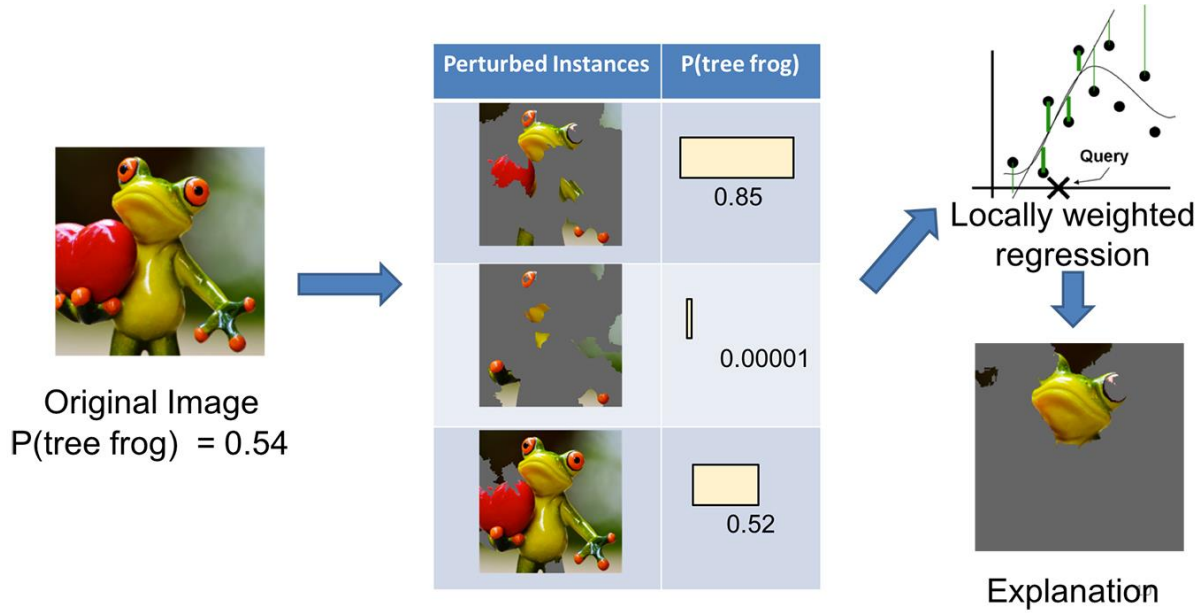


ML

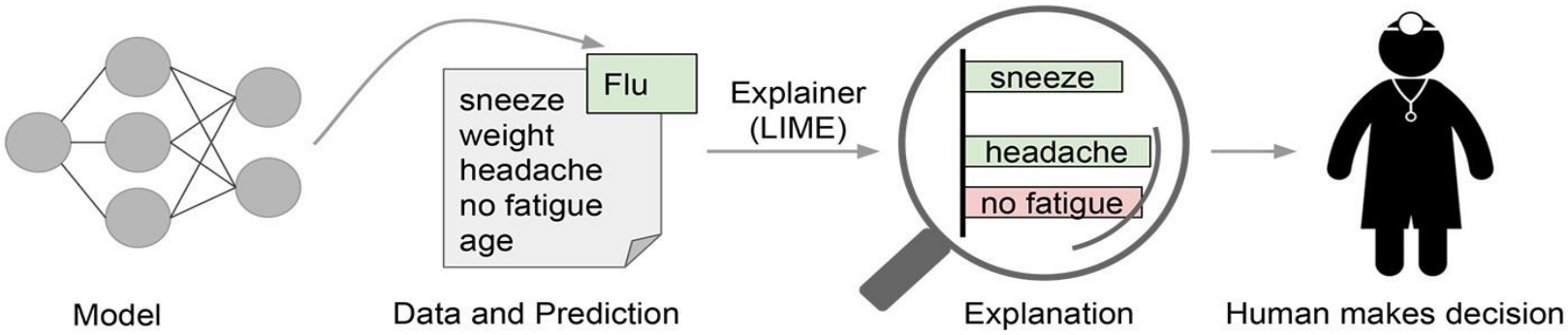
Interpretable ML

Examples

1 *Image Classification*



2 *Medical Diagnosis*

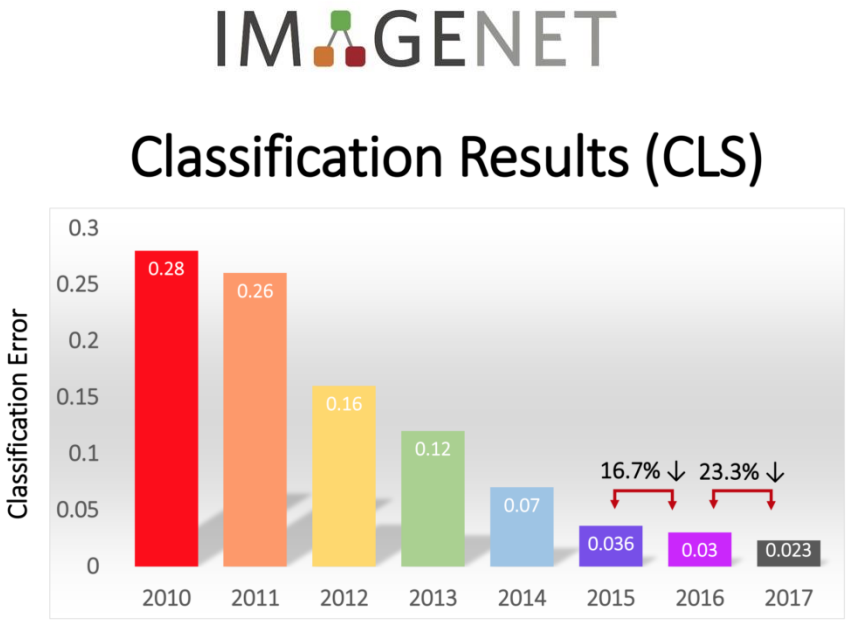


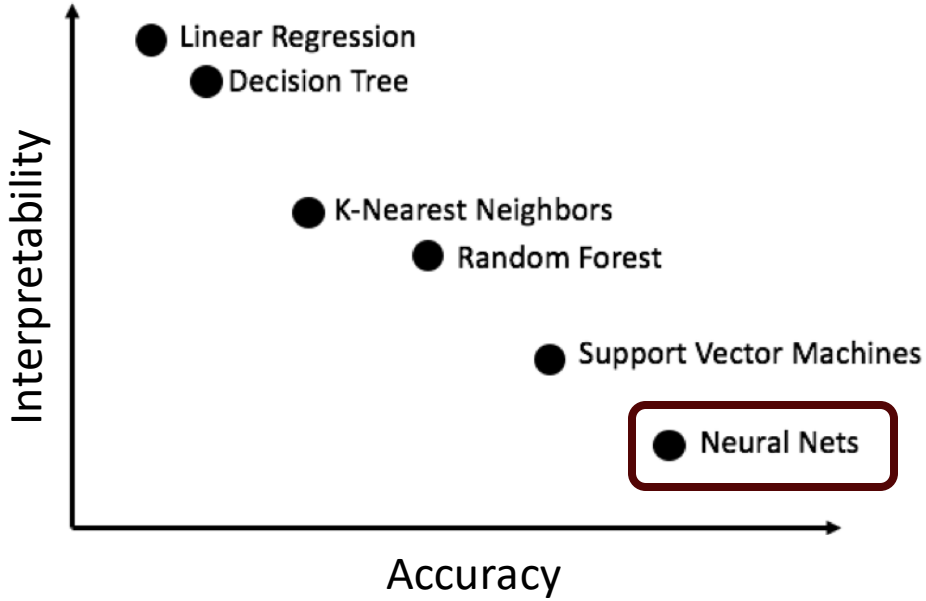
"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Outline

1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

Interpretable Deep Learning

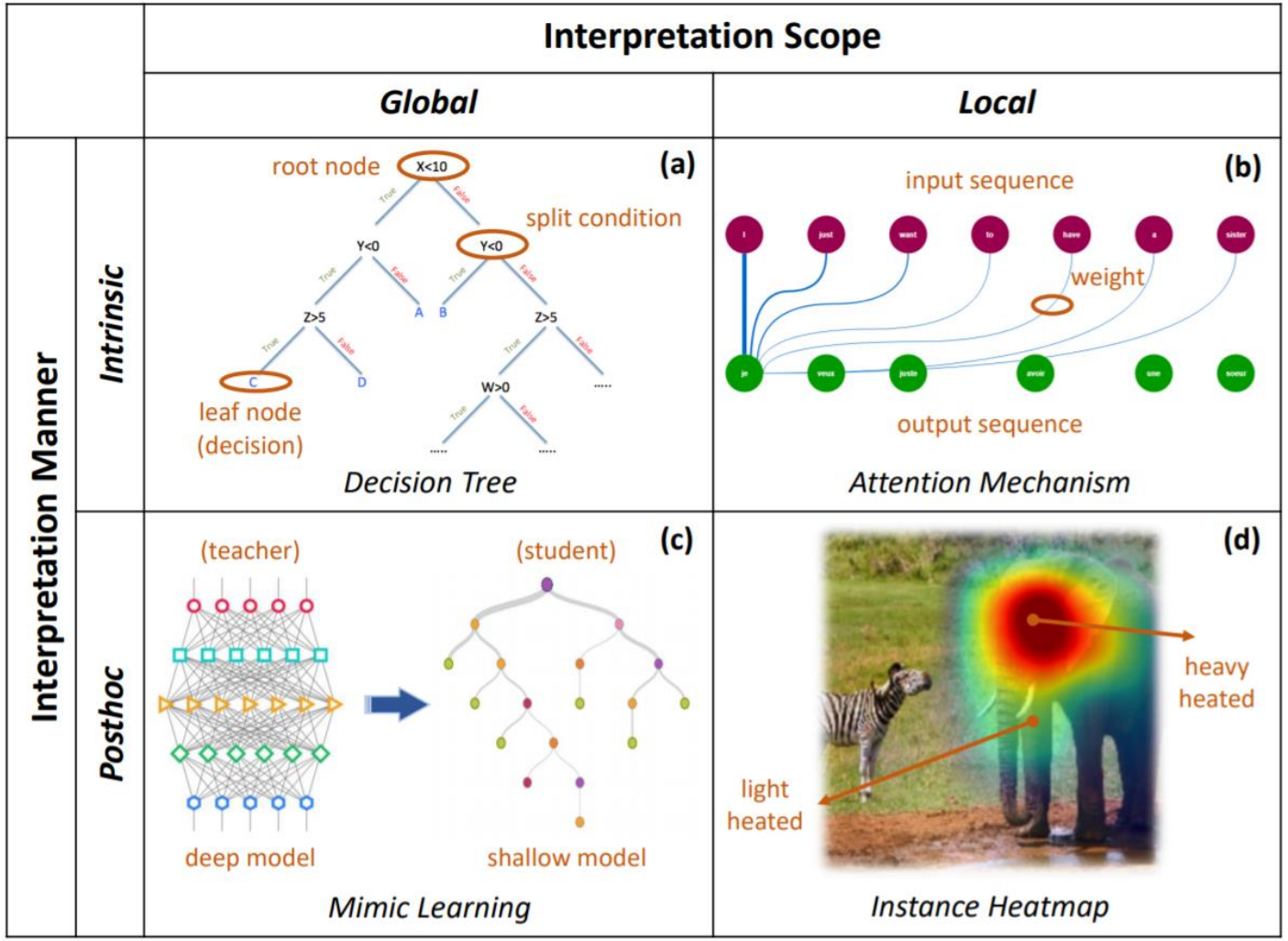




DNNs make lots of *progresses*

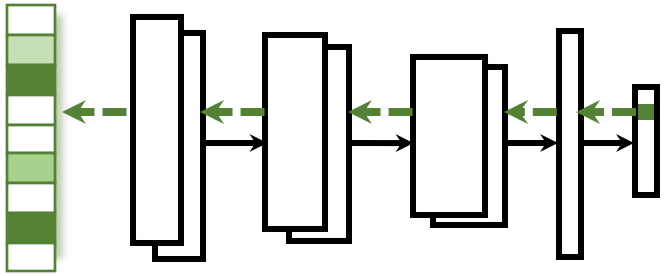
DNNs are regarded as *black boxes*

Categorization



- ✓ Intrinsic - Global
  - *decision tree*
  - *rule base*
- ✓ Intrinsic - Local
  - *Attentional model*
- ✓ Posthoc - Global
  - *Mimic learning*
- ✓ Posthoc - Local
  - *Heatmap*
  - *Influential sample*

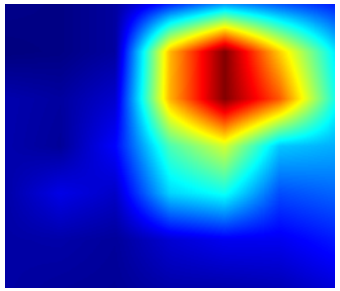
Post-hoc Local Explanation



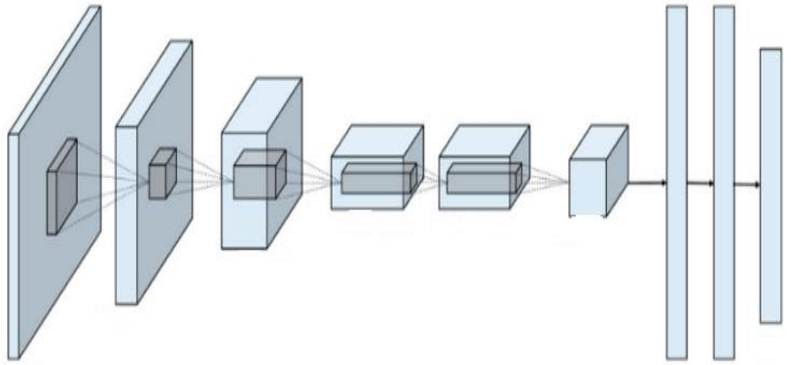
Post-hoc Interpretation

- *Given an input instance*
- *A pre-trained DNN*
- *Contribution score for each feature in input*

Input



Explanation Heatmap

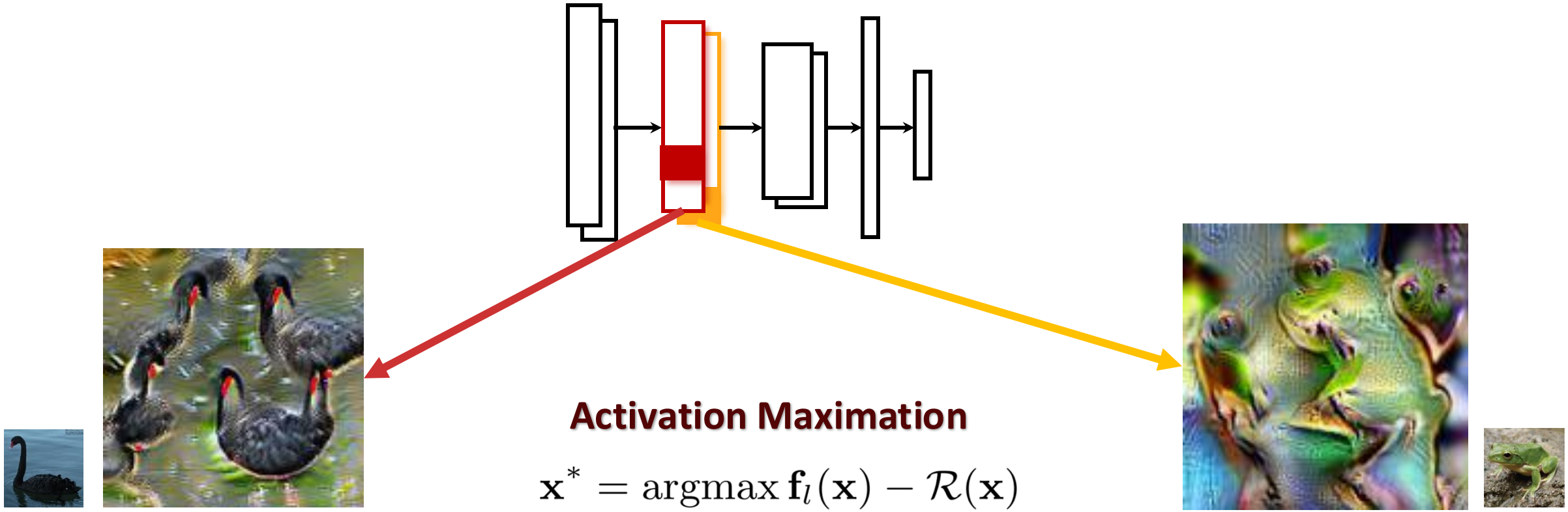


Output Probability

*Find **evidence** for prediction "elephant"*

Post-hoc Global Explanation

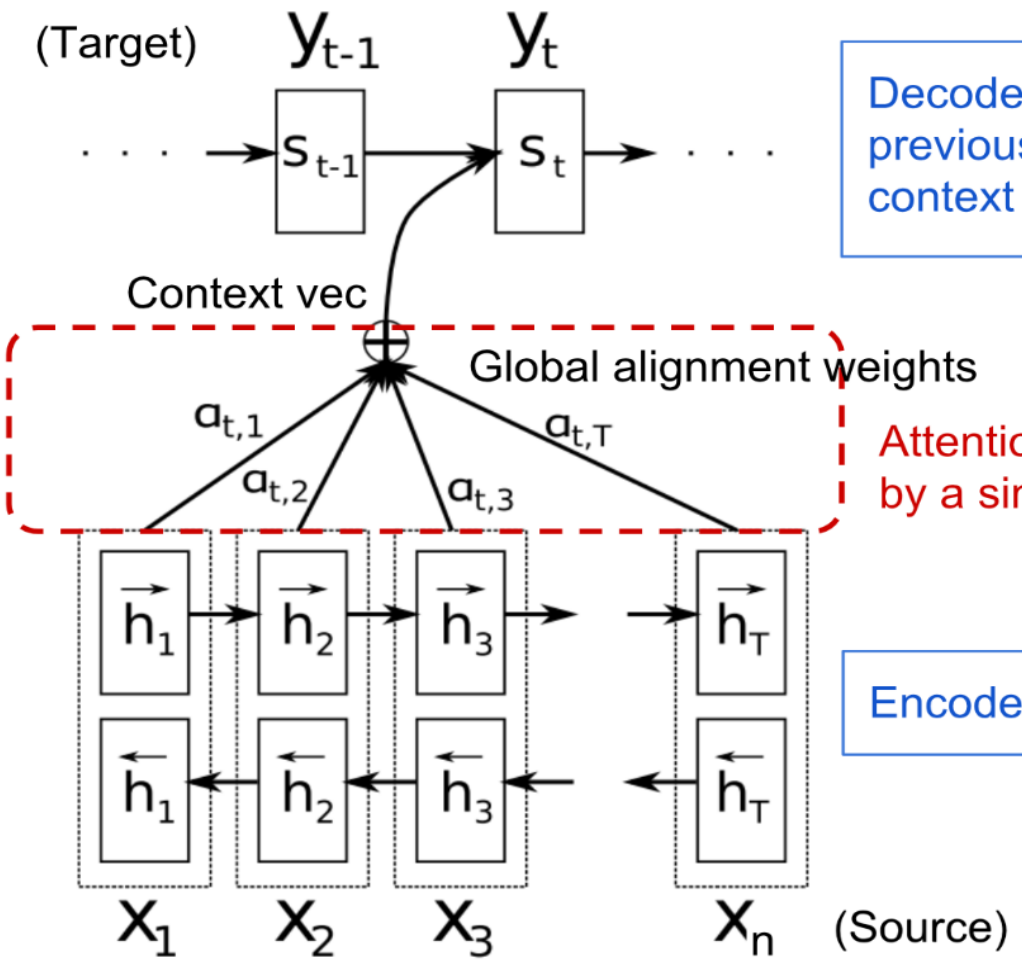
Give a global understanding about what knowledge has been captured by a DNN model



Activation Maximation

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmax}}\, \mathbf{f}_l(\mathbf{x}) - \mathcal{R}(\mathbf{x})$$

Intrinsic Attentional Model



Decoder: RNN with input from previous state + dynamic context vector.

Attention layer: parameterized by a simple feed-forward network

Additive Attention

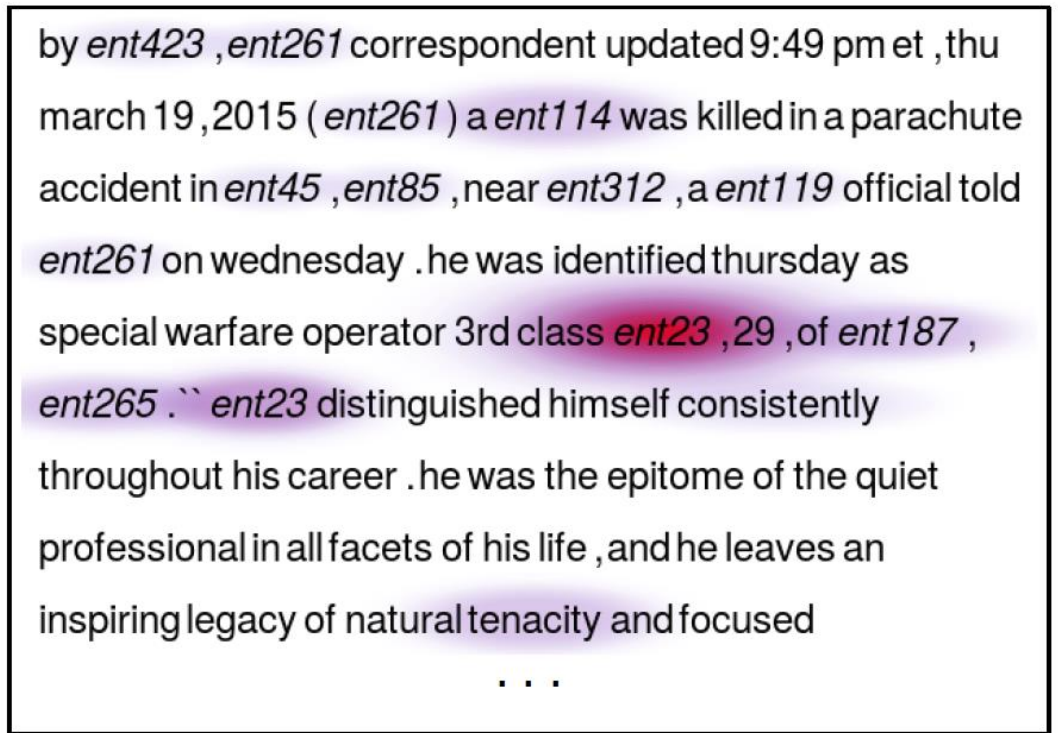
Encoder: bidirectional RNN

https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Intrinsic Interpretable Model (Local)

Design justifiable model architectures that can explain why a specific decision is made

Interpretation heatmap



Interpretation Visualization

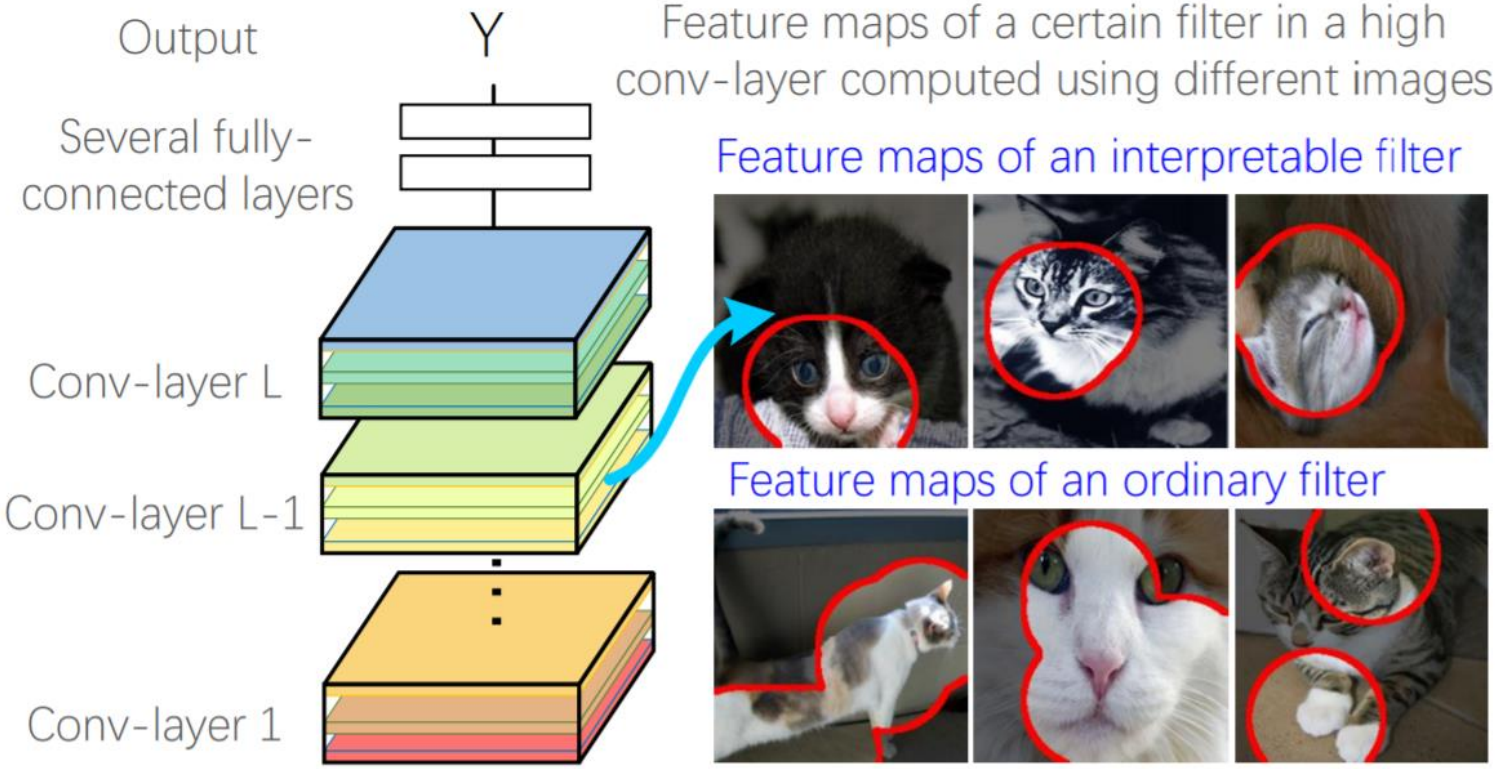
--- Contribution score for each feature in input

--- Deeper color indicates higher contribution

Teaching Machines to Read and Comprehend, NeurIPS

Intrinsic Interpretable Model (Global)

Globally interpretable models that offer a certain extent of working transparency



In interpretable CNN, each filter in high-layers represents a specific object part

Interpretable Convolutional Neural Networks, CVPR

Outline

1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

Evaluation Perspectives

Are the generated explanations ***faithful*** to the original model?

Fidelity

Ensure the explanations can ***faithfully reflect*** the model

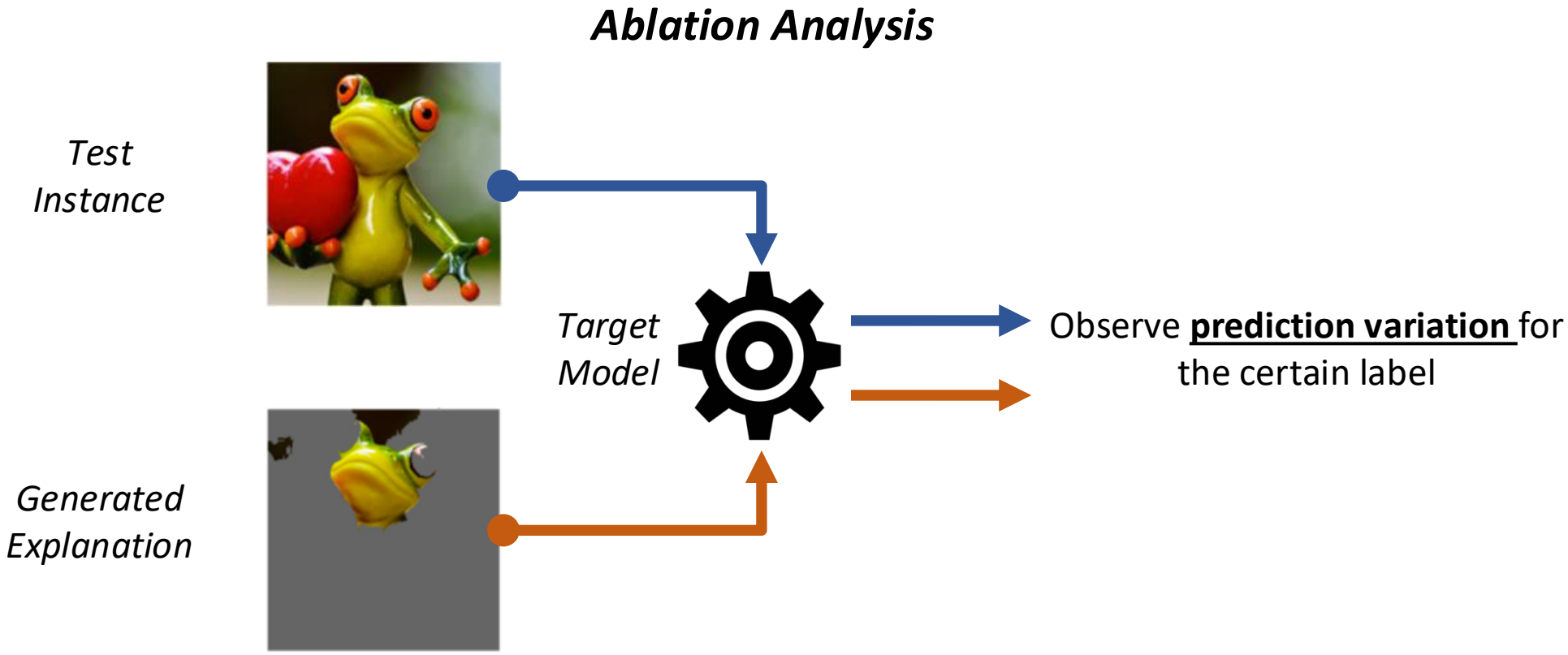
Are the generated explanations ***friendly*** to the human users?

Persuasibility

Ensure the explanations can be ***easily comprehended*** by humans

Philosophy of Fidelity Evaluation

Ablation Analysis



If the generated explanation is **faithful** to the target model, the **prediction variation** should be **small**.

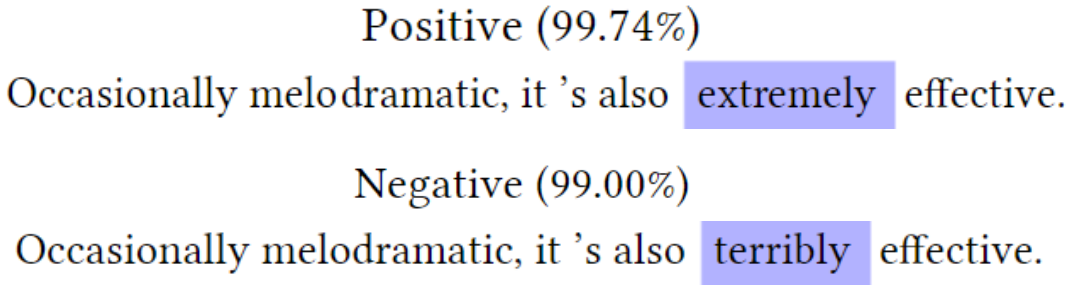
MT Ribeiro, et al. "Why should I trust you? Explaining the predictions of any classifier." KDD, 2016.

Fidelity Evaluation Cases

Image Feature

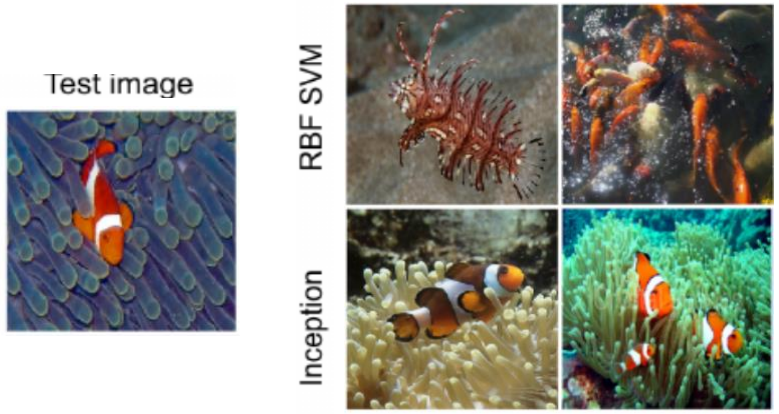
Fong, Ruth C., et al. “Interpretable explanations of black boxes by meaningful perturbation.” ICCV, 2017.

Text Feature



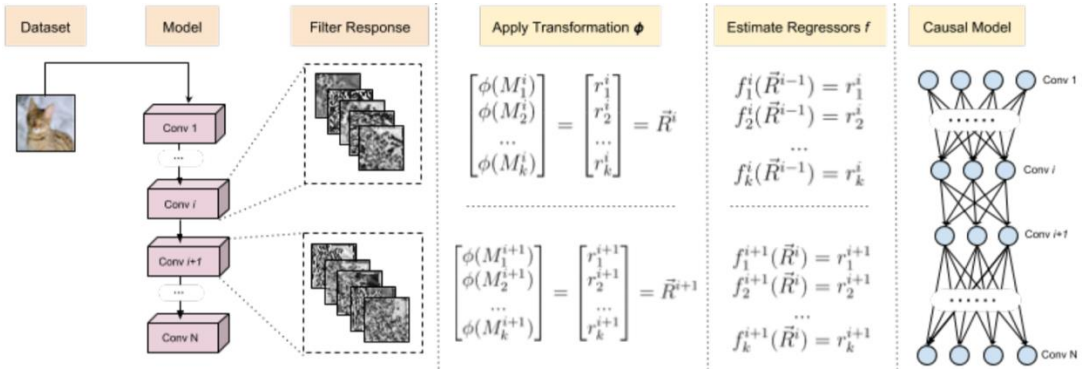
Du, Mengnan, et al. “On attribution of recurrent neural network predictions via additive decomposition.” The WebConf, 2019.

Training Data



Koh, Pang Wei, et al. "Understanding black-box predictions via influence functions." ICML, 2017.

Model Component

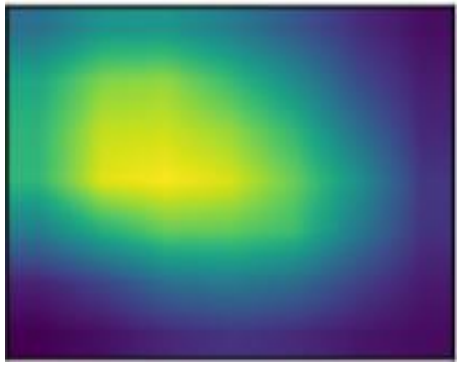


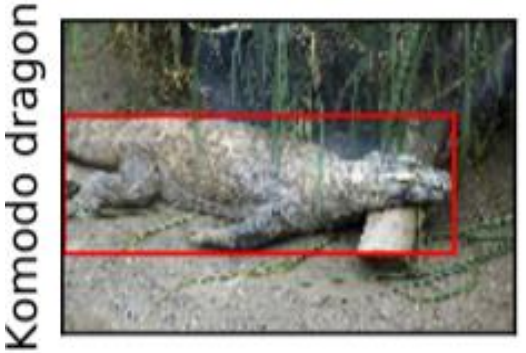
Narendra, Tanmayee, et al. "Explaining deep learning models using causal inference." arXiv, 2018.

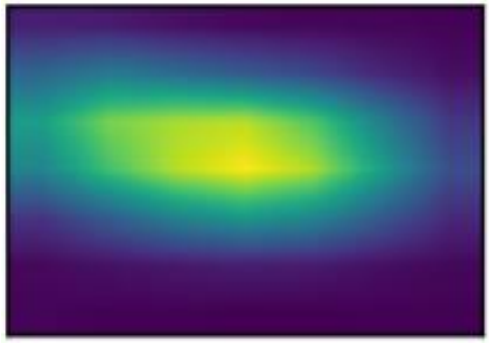
Persuasibility with Image Bounding

Evaluation with Bounding Box

"Interpretable explanations of black boxes by meaningful perturbation." ICCV, 2017.

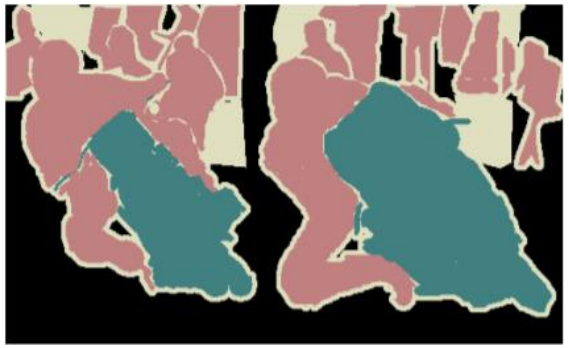






Evaluation with Semantic Segmentation

Long, Jonathan, et al. "Fully convolutional networks for semantic segmentation." CVPR, 2015.



Persuasibility with Text Rationale

Evaluation with Text Annotation

Task: movie review **Label**: negative

The movie is *so badly put together* that even the most casual viewer may notice the *miserable pacing and stray plot threads*.

Task: beer appearance **Label**: positive

A beautiful beer, coal black with a thin brown head. Extremely powerful flavors, but everything is muted by the intense alcohol . the alcohol is so strong.

"Learning credible deep neural networks with rationale regularization." ICDM, 2019.

Persuasibility with User Study

Evaluation with Human-Computer Interaction (HCI)

Mental Model ?

User Satisfaction ?

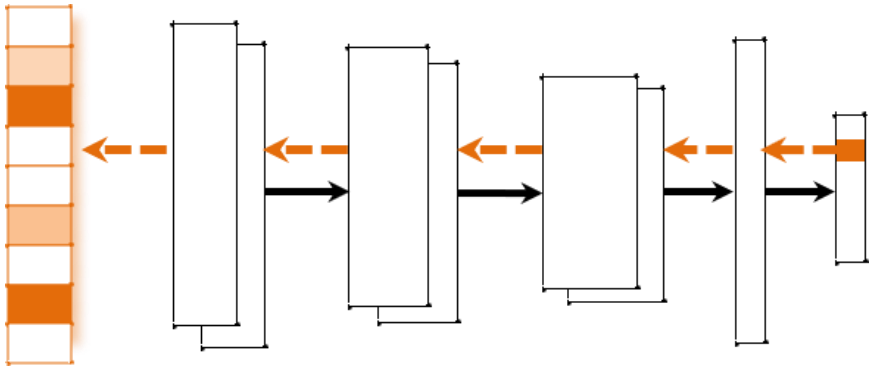
User Trust ?

Lage, Isaac, et al. "An evaluation of the human-interpretability of explanation." arXiv, 2019.

Outline

1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

Post-Hoc CNN Interpretation



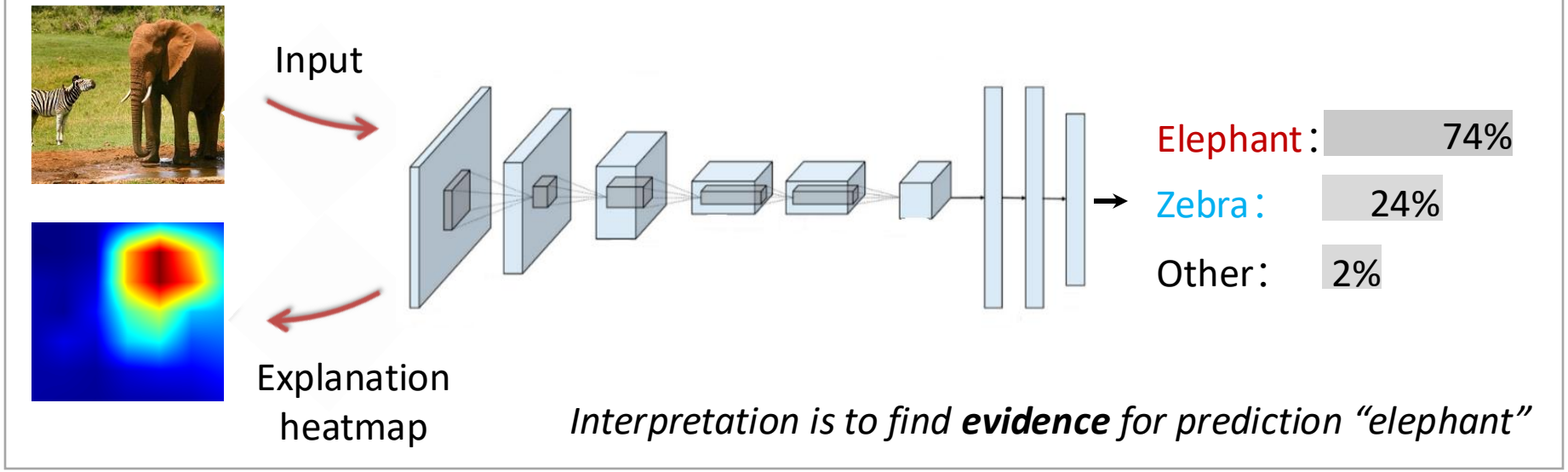
Key Factors

--- A pre-trained DNN and an input instance

--- The prediction of DNN

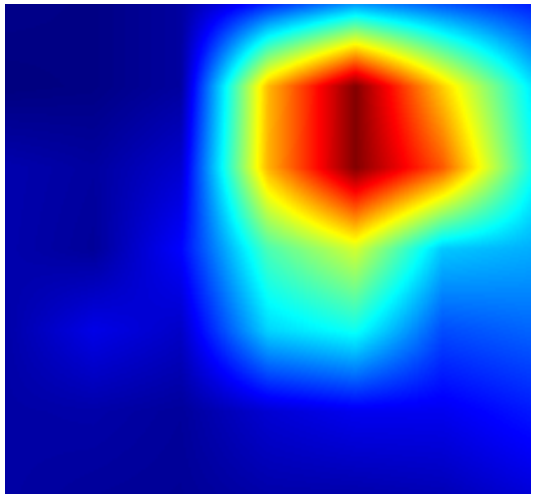
Post-hoc Interpretation

--- Contribution score for each feature in input

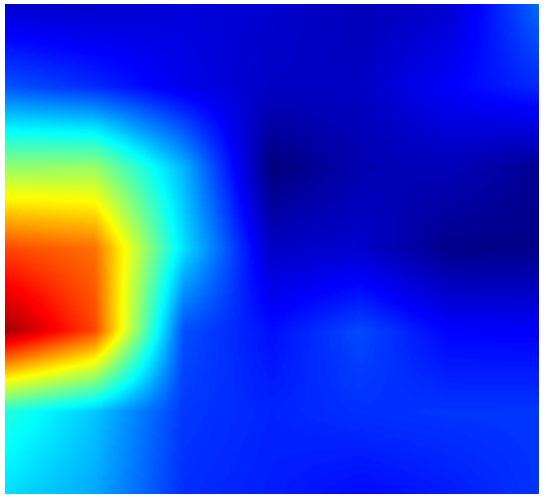


Motivation: Using ***deep representations*** in intermediate layers to derive interpretations

Challenges



Elephant



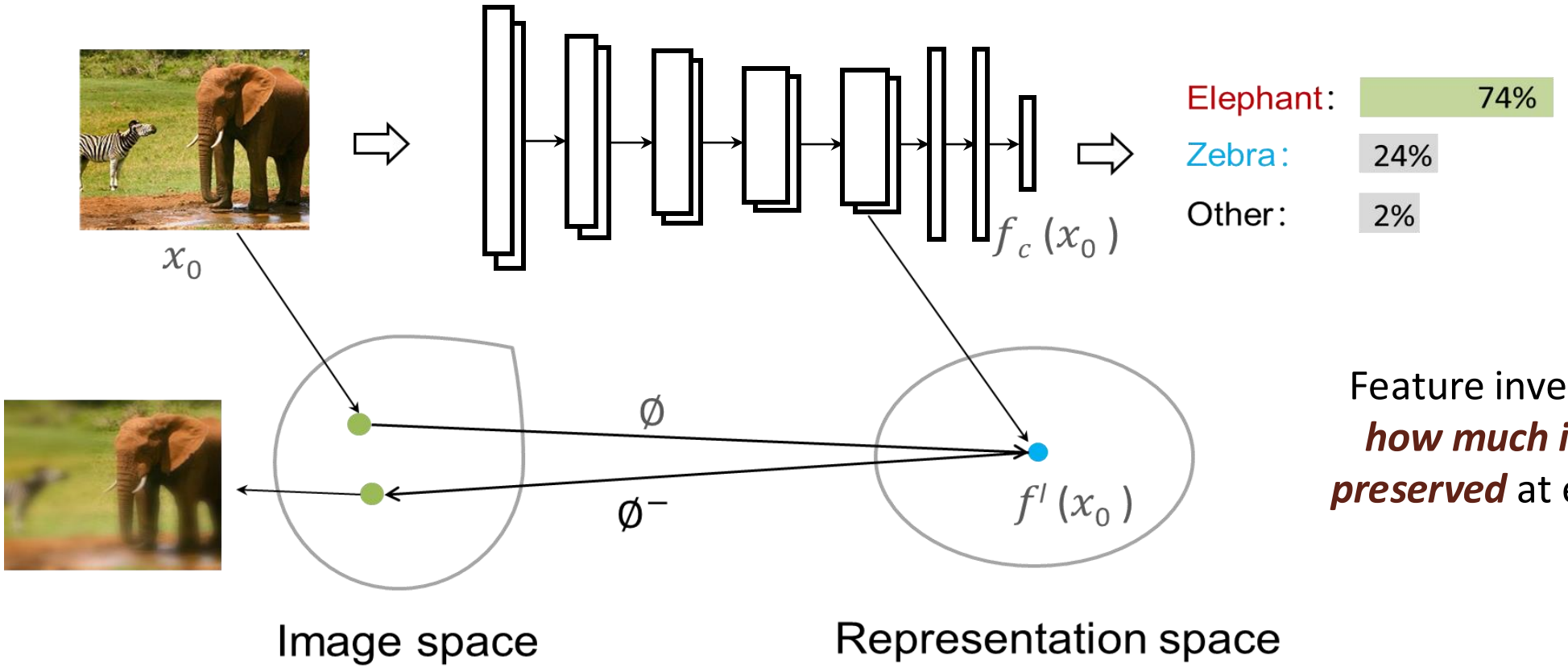
Zebra

1. How to guarantee that the ***interpretations are faithful*** to the decision-making process of the original CNN model?

2. How to generate ***class-discriminative interpretation***?

Representation Inversion

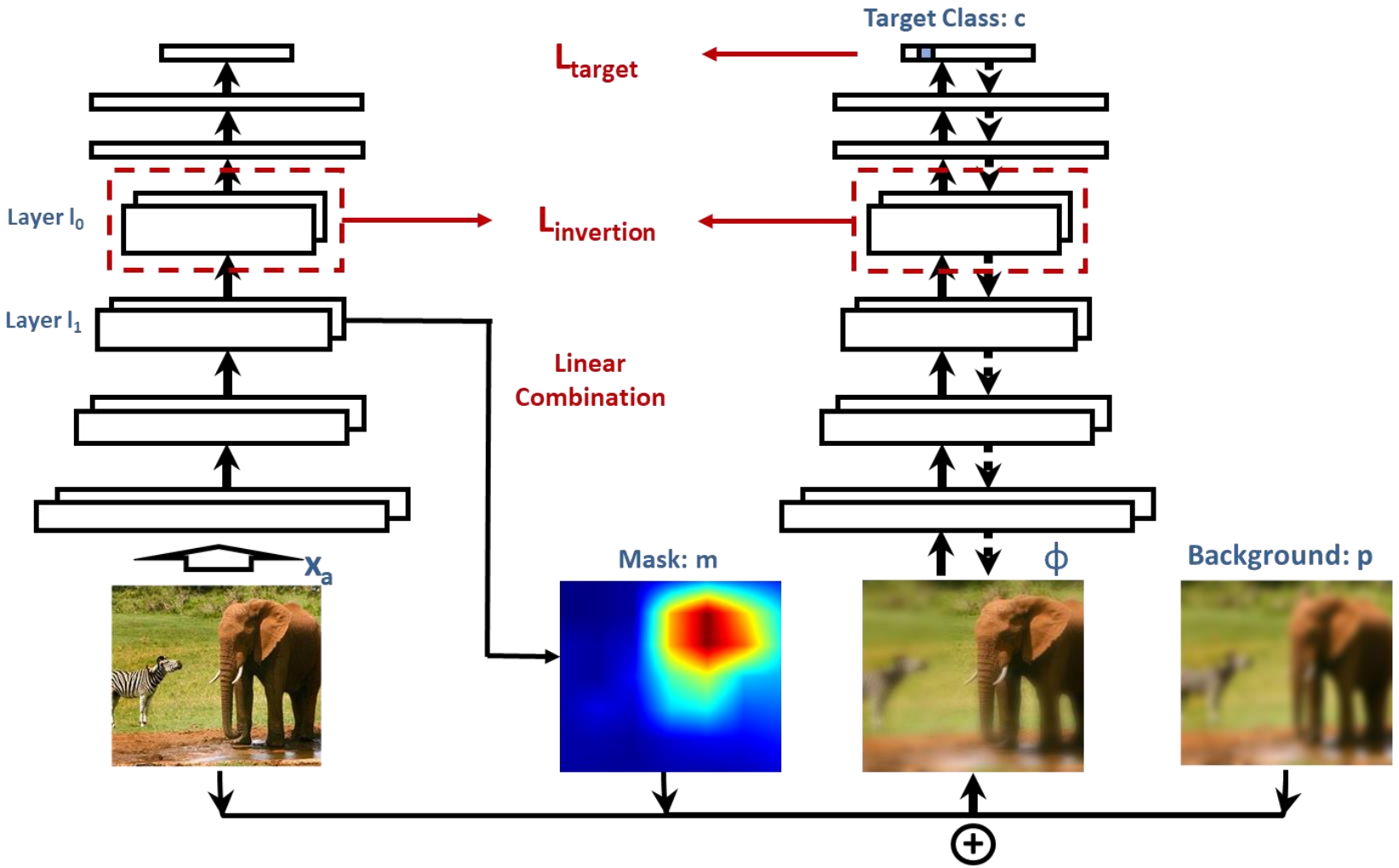
Sub-network Ø maps **input x_0** to a **representation $f^l(x_0)$**



Feature inversion to obtain ***how much information is preserved*** at each inner layer

"Understanding deep image representations by inverting them". CVPR, 2015.

The Proposed Method



- ***Guided feature inversion*** for preserving the object location in a mask;
- ***Model target neuron in output layer*** for getting class-discriminative interpretation;
- ***Regularization by inner layers*** for further reduced artifacts;

Guided Feature Inversion

Representation of the **inverted input**

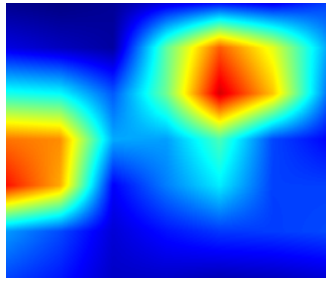
Representation of the **original input**

$$L_{\text{inversion}}(\mathbf{x}_a, \mathbf{m}) = \|\mathbf{f}^{l_0}(\Phi(\mathbf{x}_a, \mathbf{m})) - \mathbf{f}^{l_0}(\mathbf{x}_a)\|^2 + \alpha \cdot \frac{1}{d} \sum_{i=1}^{d} \mathbf{m}_i$$

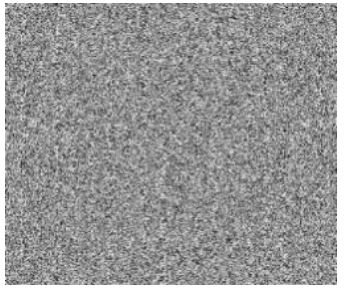
$$\Phi(\mathbf{x}_a, \mathbf{m}) = \mathbf{x}_a \odot \mathbf{m} + \mathbf{p} \odot (1 - \mathbf{m})$$

Inverted input

Original input

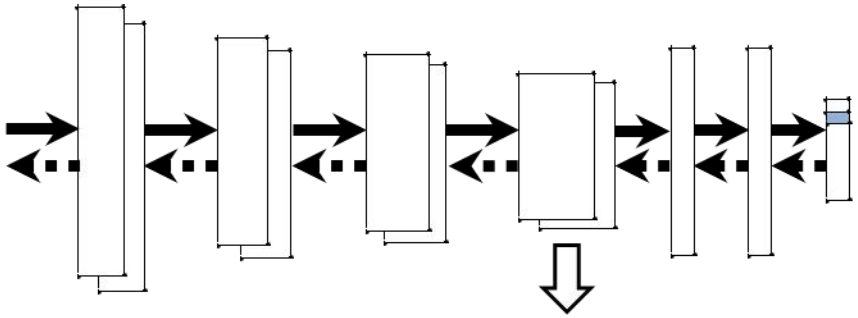


Weight matrix



Noise

Class-Discriminative Interpretation

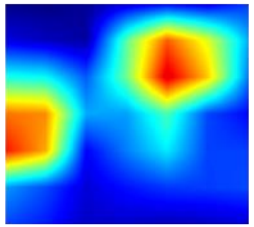


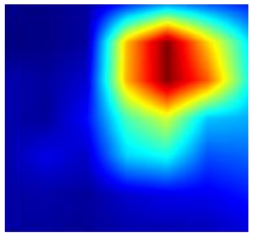
Elephant: 74%

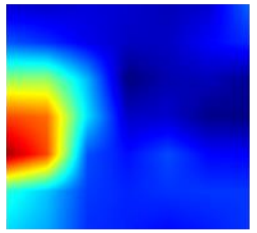
Zebra: 24%

Other: 2%

Interpretation heatmap with Guided feature inversion







Elephant

Zebra

Class-discriminative

$$L_{\text{target}}(\mathbf{x}_a, \mathbf{m}) = -\mathbf{f}_c^L(\Phi(\mathbf{x}_a, \mathbf{m})) + \lambda \mathbf{f}_c^L(\Phi_{bg}(\mathbf{x}_a, \mathbf{m})) + \beta \cdot \frac{1}{d} \sum_{i=1}^{d} \mathbf{m}_i$$

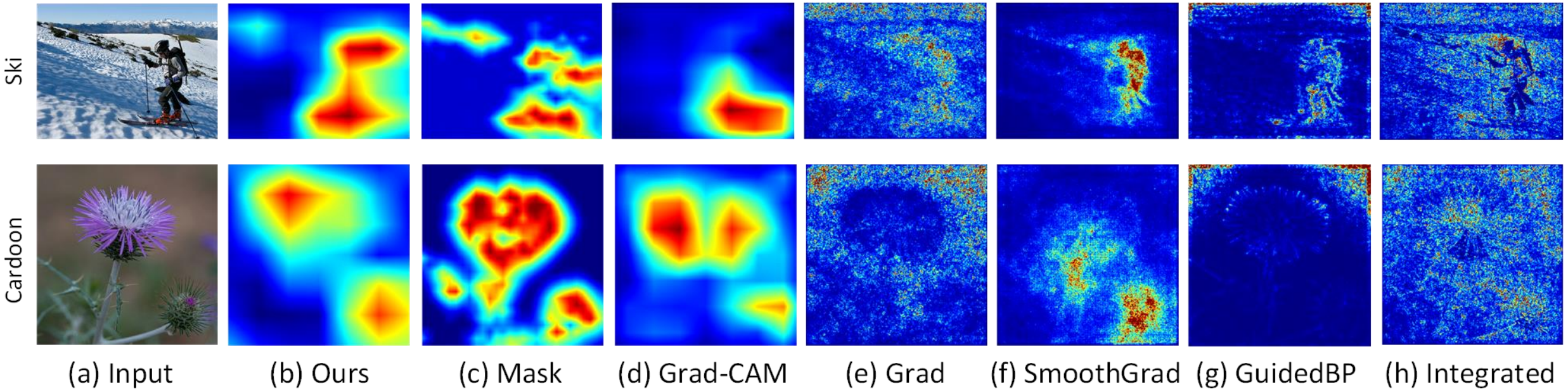
highlight ***suppress***

Output of the **target neuron**

Accurate Interpretation (1/2)

Question: Are the interpretations *accurate*, *class-discriminative* and not *affected by artifacts*?

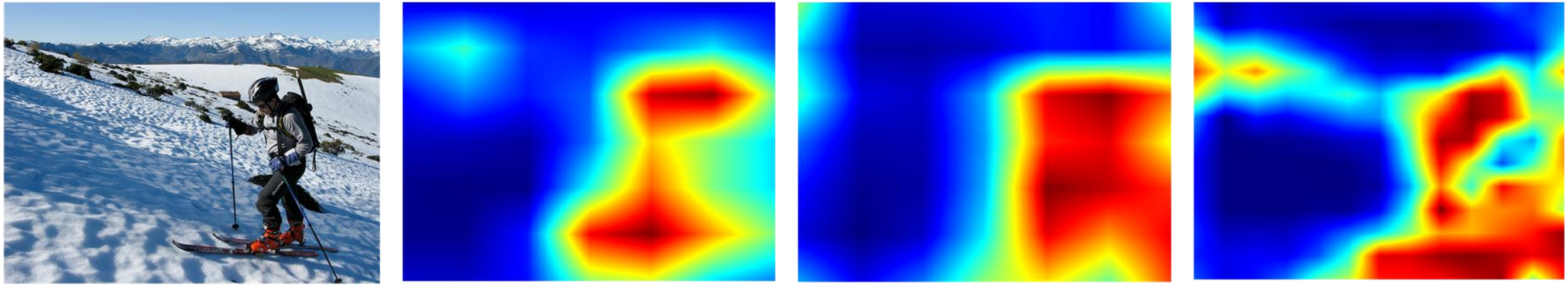
***Visualization comparison** with 6 state-of-the-art methods*



Our interpretation can **accurately identify the evidence for prediction**

Accurate Interpretation (2/2)

*Interpretation results for **three DNN architectures***

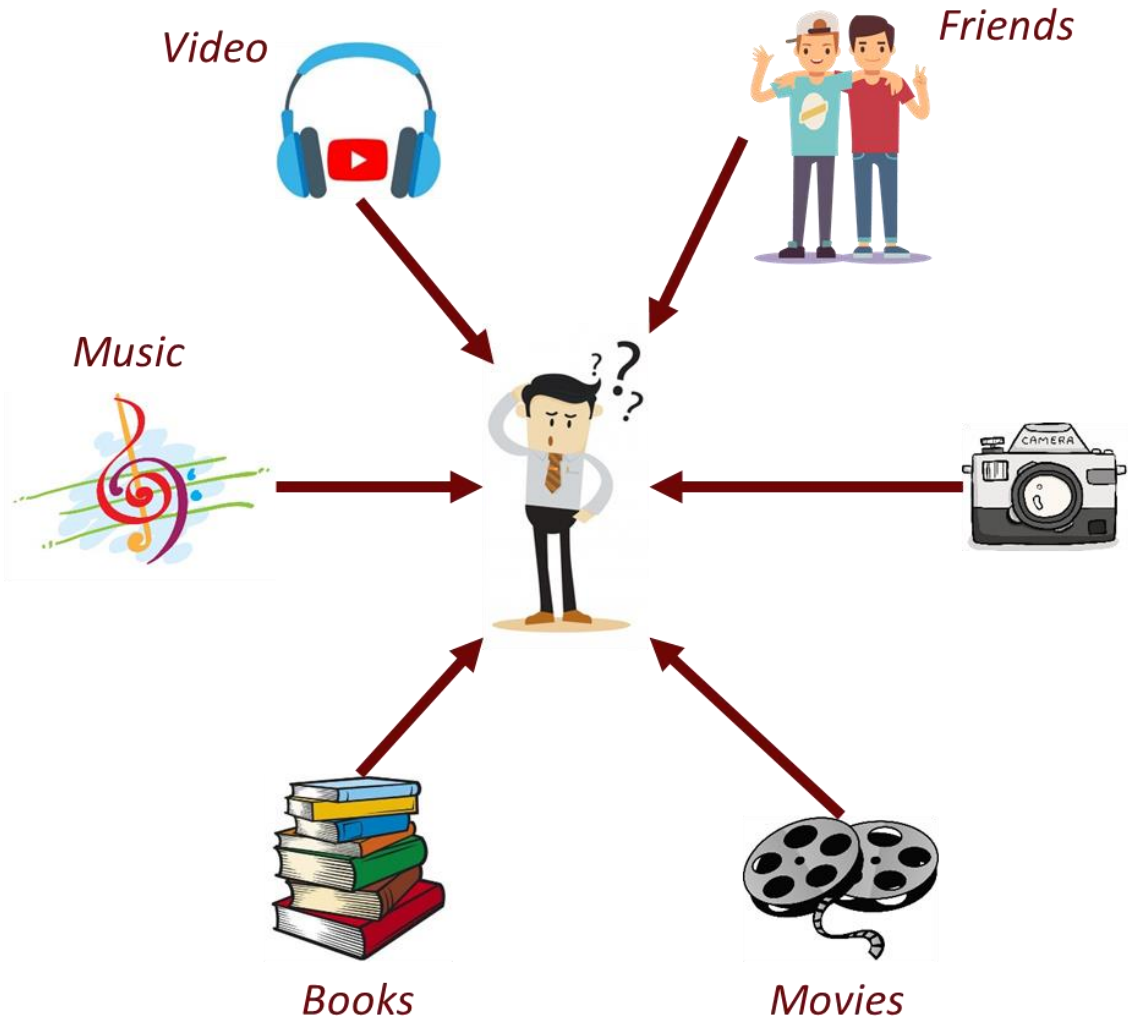


Interpretations help ***capture the pros and cons*** of different network architectures

Outline

1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

Why Interpretations for RecSys



Having deeper insights into RecSys may benefit from multiple ways:

For Customers ---

- *Identify personal **needs***
- *Facilitate **decisions***

For Vendors ---

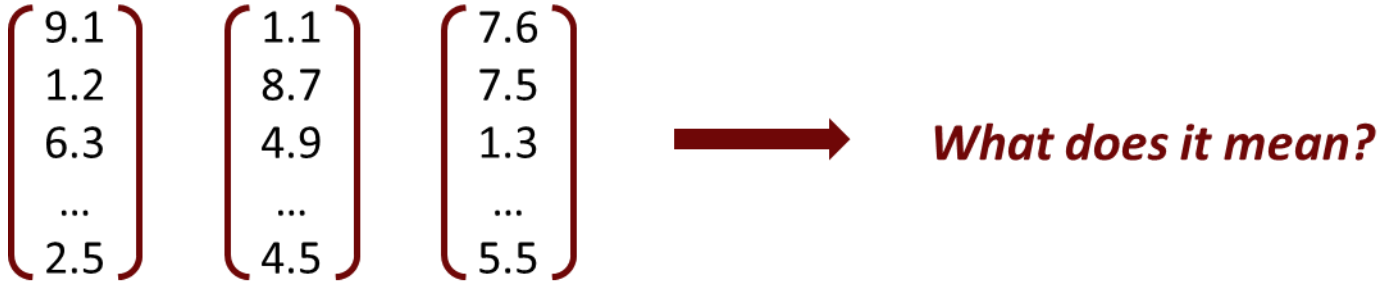
- *Make good **strategies***
- *Choose effective **target***

For Deployers ---

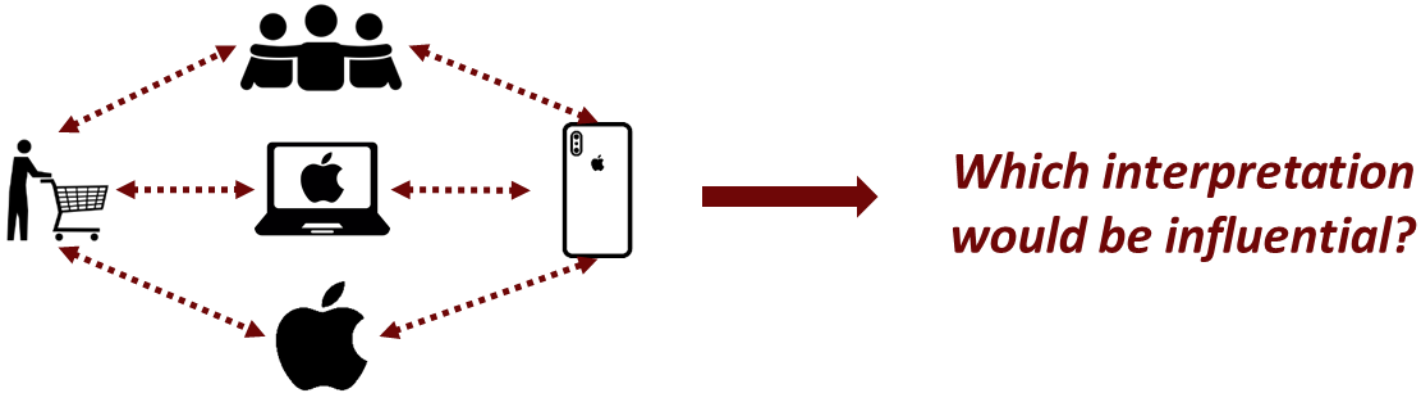
- ***Debug** the system*
- ***Refine** the system*

Challenges

1. The latent factors of users and items are the **uninterpreted vectors** to humans.

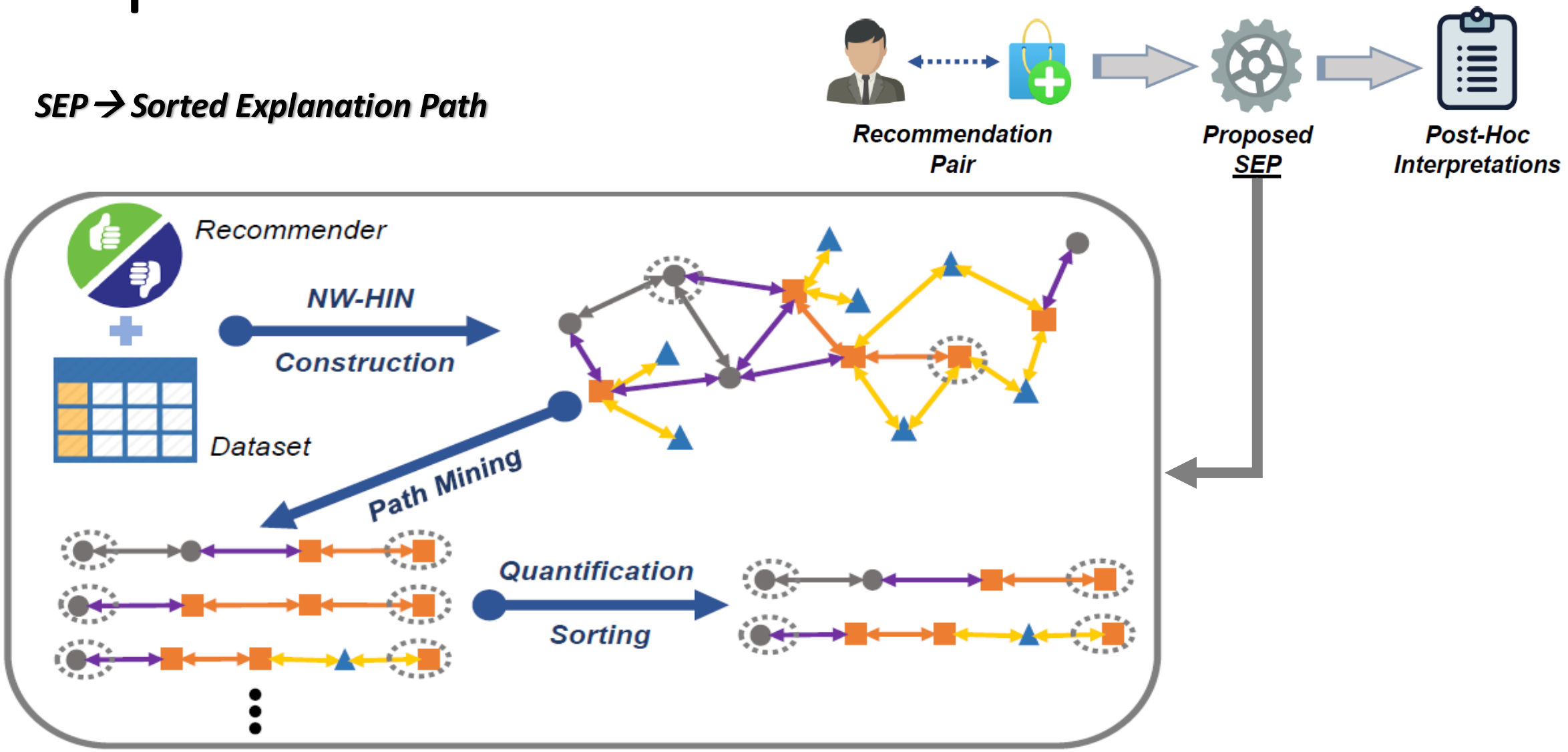


2. The possible interpretations can be diversified, and **appropriate selections** would be difficult.



Proposed Framework

SEP → Sorted Explanation Path



HIN Components

Our Constructed HIN Structure ---

Node Type: User Node

Item Node

Aspect Node

Link Type: User-User

Item-Item

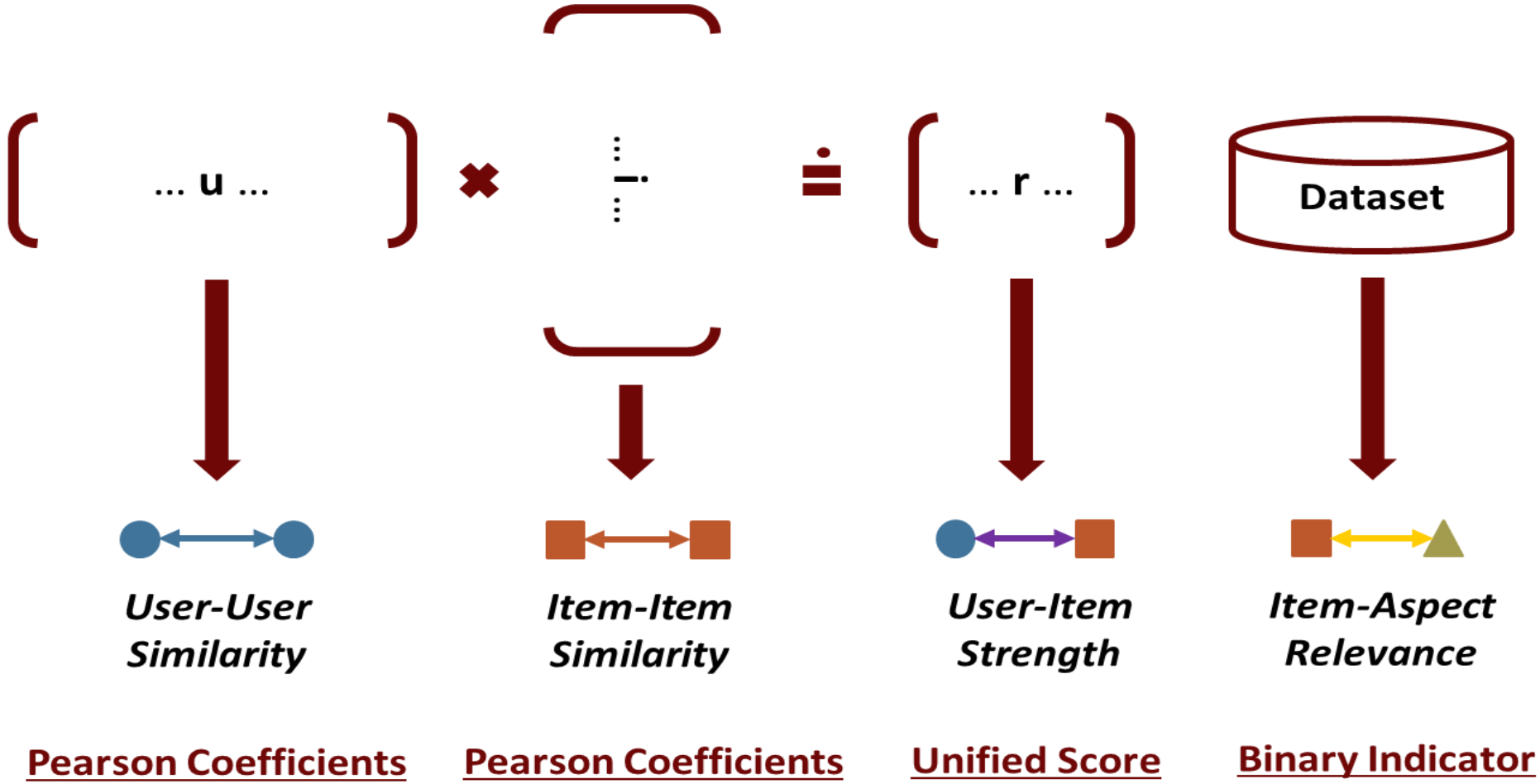
User-Item

Item-Aspect

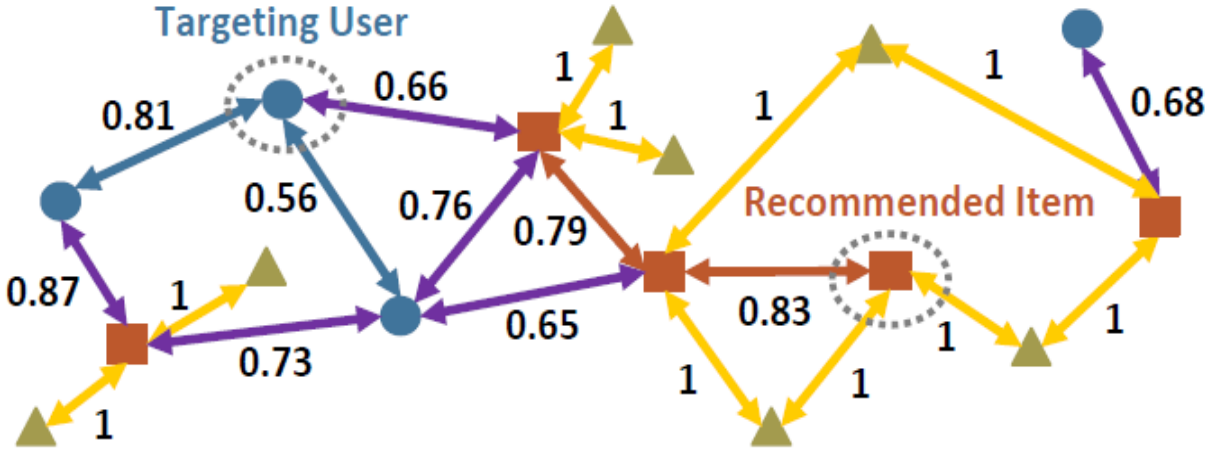
Network Schema:

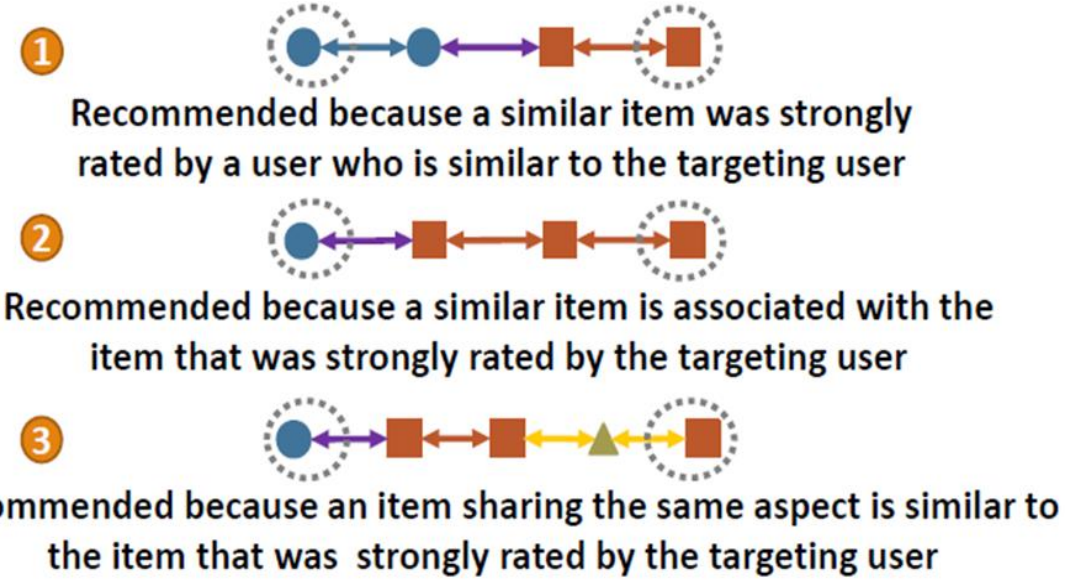
HIN Construction

Latent-Factor Recommender System ---



Explanation Path Mining



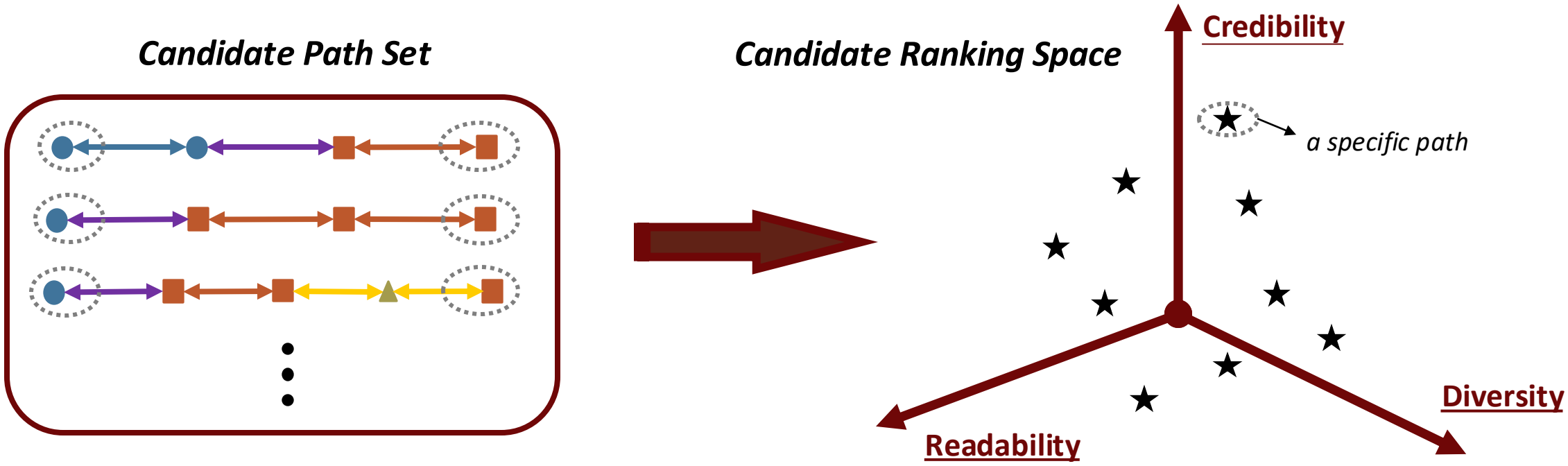


To keep the process effective and efficient, we conduct the mining based on a ***depth-first-search*** based algorithm with ***constraints on weight and length thresholds***

Path Quantification

For each explanation path k, we have →

$$\mathbf{k} = [Q^C(k), Q^R(k), Q^D(k)]^\intercal$$



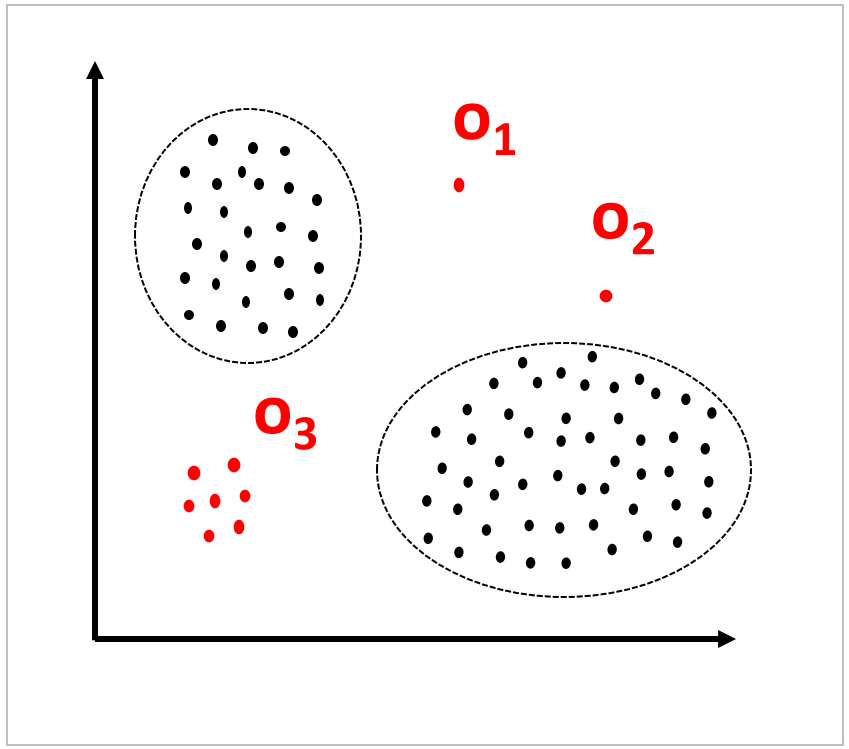
Summary

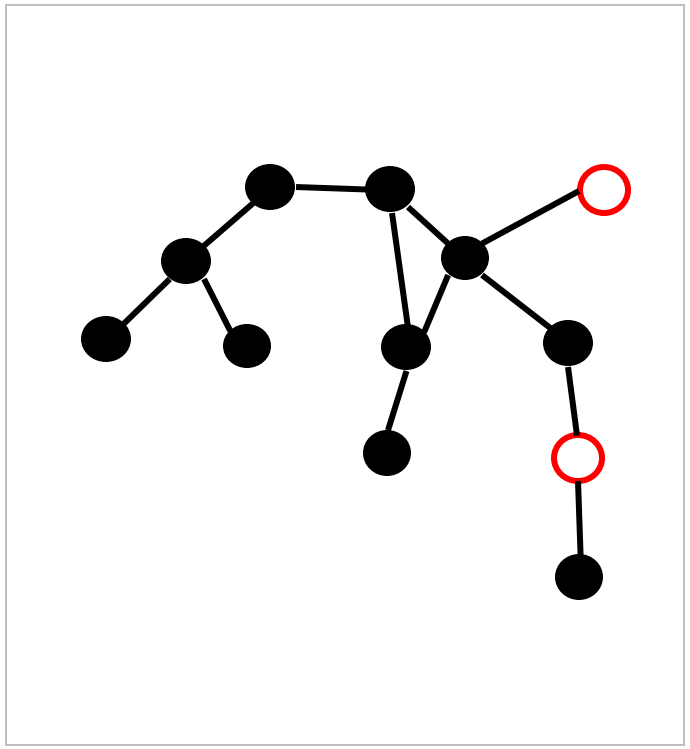
- We propose a **post-hoc interpretation method** called SEP for explaining the results of recommender systems.
- Three heuristic metrics (i.e., **Credibility**, **Readability** and **Diversity**) are designed to better quantify the quality of the generated explanations.
- An **effective ranking algorithm** is applied to sort the explanations, which helps to filter out these bad explanations.
- Empirical experiments show some **promising results** of SEP.

Outline

1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

What are Outliers?

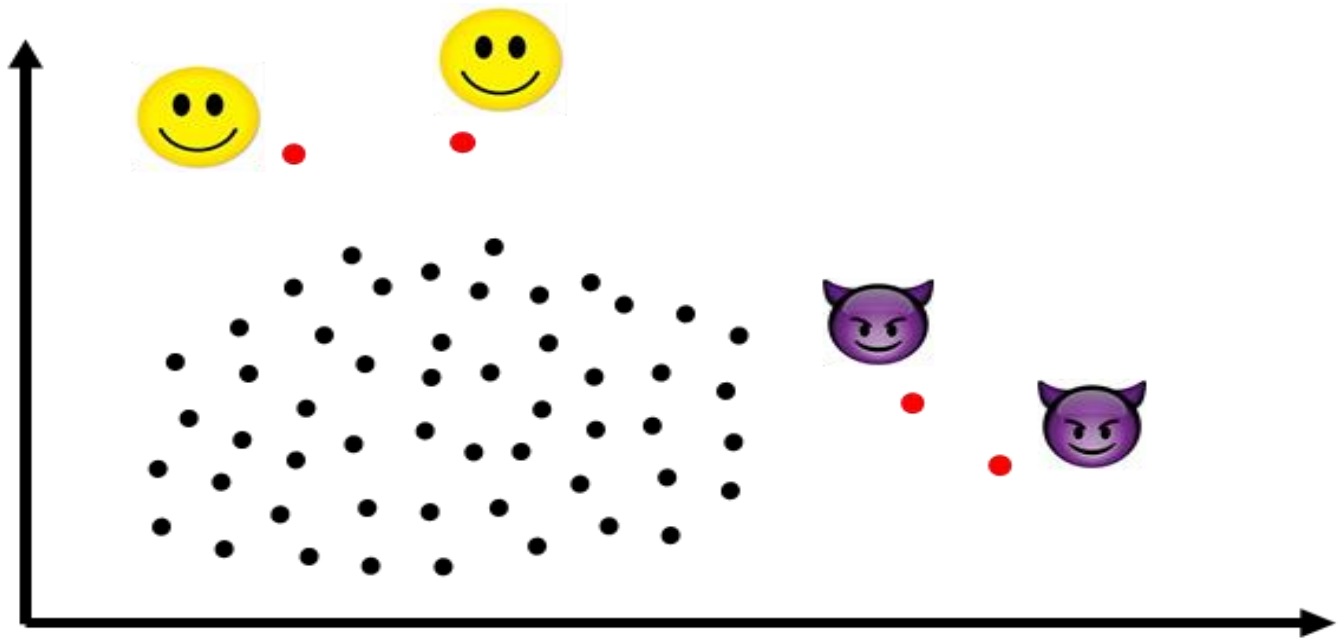




The noteworthy objects with patterns or behaviors that significantly deviate from the chosen background (or context).

Why is Interpretation Needed?

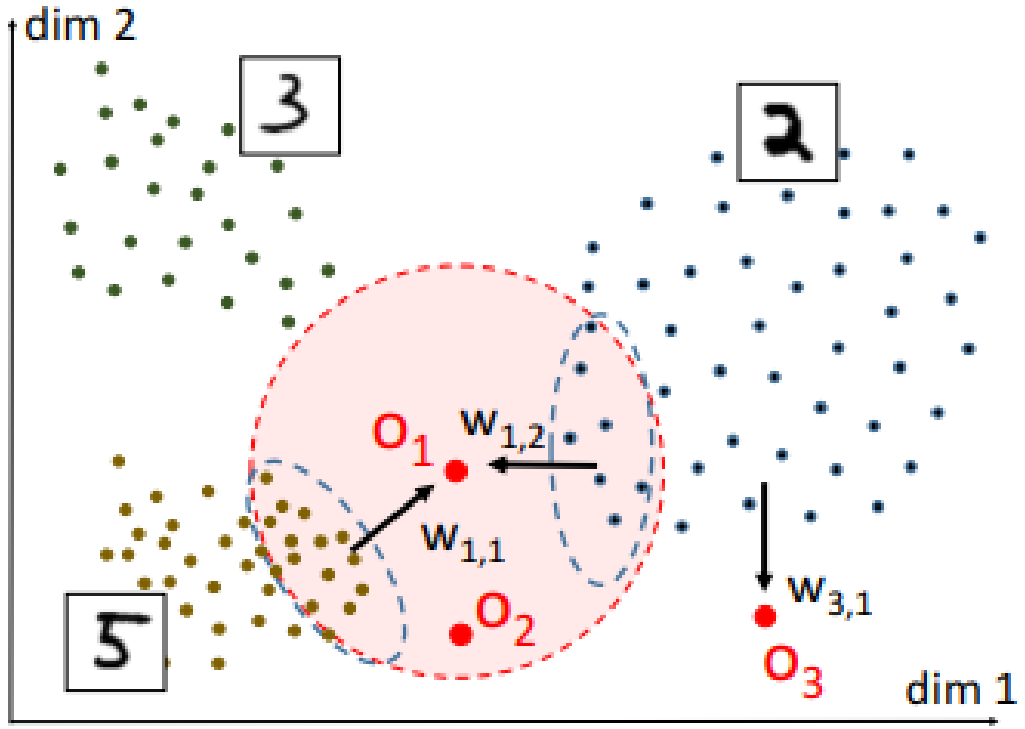
- Hard to tell whether the detected outliers are relevant to the application scenario;
- Existing metrics (e.g., ROC AUC) are unstable or limited in measuring performance;



Key Factors for Outlier Interpretation

- The definition of interpretation for outlier detection.
- The design of a model-agnostic interpretation framework.
- Identification of application-specific anomalies by utilizing interpretation with human prior knowledge.

Definition of Interpretation



A toy example explaining why context clustering is needed

Given a dataset $\mathbf{X} = \{\mathbf{x_n}\}$ and the detected outlier set $\mathcal{O}$, the interpretation for each outlier $\mathbf{o}_i \in \mathcal{O}$ is defined as:

$$\{\mathcal{A}_i, d(\mathbf{o}_i), \mathcal{C}_i = \{\mathcal{C}_{i,l} | l \in [1, L]\}\}$$

where

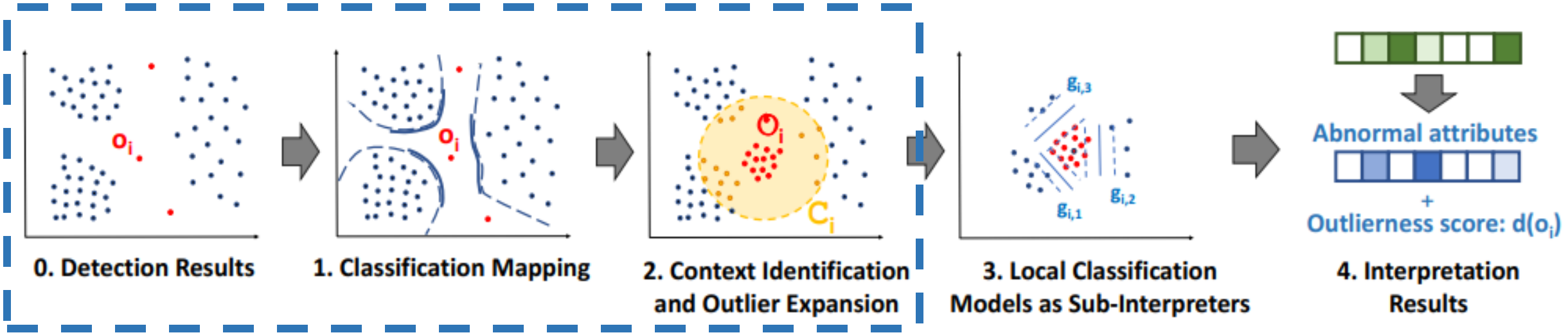
$\mathcal{C}_i$: the context (e.g., k-nearest normal neighbors) of the outlier;

$\mathcal{C}_{i,1}, \mathcal{C}_{i,2}, \ldots, \mathcal{C}_{i,L}$s identified from the context;

$\mathcal{A}_i$: the set of outlying attributes;

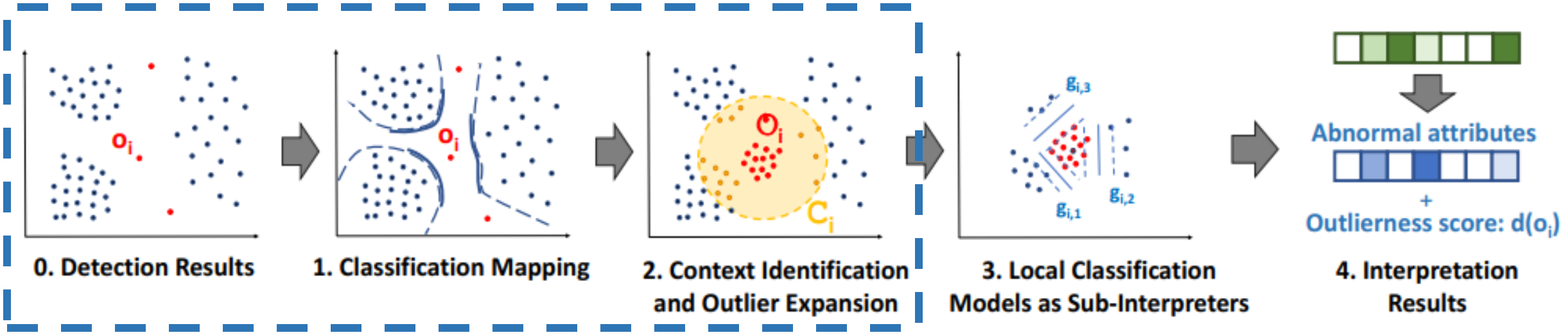
$d(\mathbf{o}_i) \in \mathbb{R}_{\geq 0}$tlierness score.

Proposed Framework (1/3)



- h : The given outlier detector.
- There could be an imaginary classification boundary, denoted by f , to separate outliers from normal instances.

Proposed Framework (2/3)



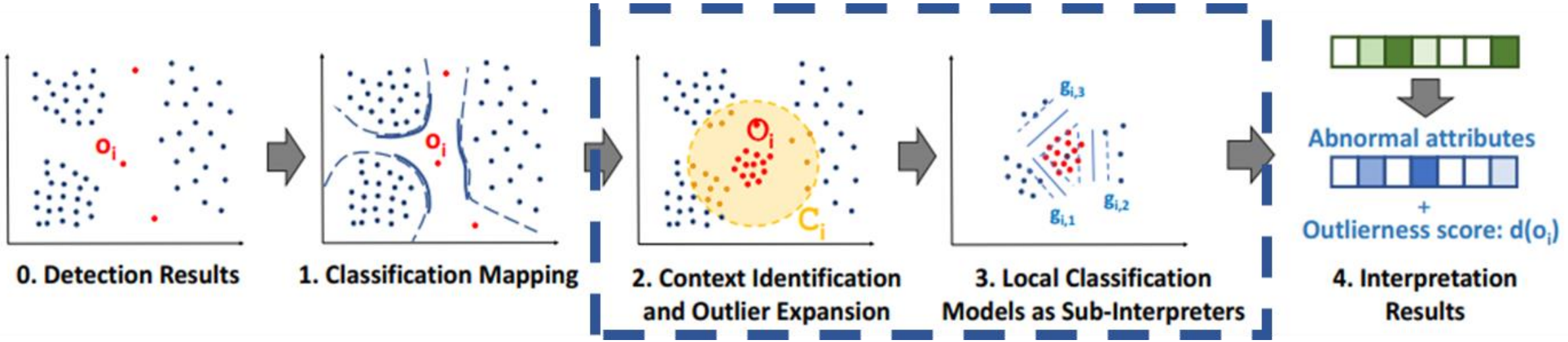
- We use f to interpret h :

$$\min_f \mathcal{L}(h, f; \mathcal{O}, \mathcal{X} - \mathcal{O}) \Rightarrow \min_f \sum_i \mathcal{L}(h, f; \mathbf{o}_i, \mathcal{C}_i) \quad \text{Decomposition}$$

$$\Rightarrow \sum_i \min_{g_i} \mathcal{L}(h, g_i; \mathbf{o}_i, \mathcal{C}_i) \quad g_i\text{: Local boundary}$$

$$\Rightarrow \sum_i \min_{g_i} \mathcal{L}(h, g_i; \mathcal{O}_i, \mathcal{C}_i) \quad \text{Synthetic sampling}$$

Proposed Framework (3/3)



Classification error between $\mathcal{C}_i$ and $\mathcal{O}_i$

$$P^{err}(\mathcal{O}_i, \mathcal{C}_i) = P(\mathcal{O}_i)\int_{\mathcal{C}_i} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x} + P(\mathcal{C}_i)\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_i)d\mathbf{x}$$

$$\approx \Big(\sum_{l\in[1,L]} P(\mathcal{O}_i)\int_{\mathcal{C}_{i,l}} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x}\Big) + \Big(\sum_{l\in[1,L]} P(\mathcal{C}_{i,l})\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_{i,l})d\mathbf{x}\Big)$$

$$= \sum_{l\in[1,L]} \Big(P(\mathcal{O}_i)\int_{\mathcal{C}_{i,l}} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x} + P(\mathcal{C}_{i,l})\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_{i,l})d\mathbf{x}\Big)$$

$$\approx \sum_{l\in[1,L]} P^{err}(\mathcal{O}_{i,l}, \mathcal{C}_{i,l}).$$

$g_{i,l}$

Local classification error between $\mathcal{C}_{i,l}$ and $\mathcal{O}_{i,l}$

Experiments - Settings

	SYN1	SYN2	WBC	Twitter	MNIST
N	405	405	458	11,000	42,000
M	15	15	9	16	150
$\lvert\mathcal{O}\rvert$	30	30	25	1,000	1,000

Table: Details of the datasets

Baseline Methods:

- **LIME**: A method for explaining supervised classification models.
- **IPS-BS**: An outlier interpretation method based on beam search and isolation path.
- **CAL**: An outlier interpretation method based on LASSO, without considering context clusters.

Experiments - Results

	COIN			CAL			IPS-BS			LIME		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
SYN1	**0.97**	**0.89**	**0.93**	0.89	0.81	0.84	0.87	0.44	0.58	0.82	0.79	0.80
SYN2	**0.99**	**0.90**	**0.94**	0.92	0.70	0.80	**1.00**	0.37	0.54	0.91	0.70	0.79
WBC	0.86	0.37	**0.52**	0.84	0.37	0.51	**0.90**	0.15	0.26	0.35	**0.39**	0.37
Twitter	**0.91**	0.33	0.48	0.75	0.34	0.47	0.72	0.29	0.41	0.60	**0.67**	**0.63**

Table: Performance of outlying attributes identification

- Noise *attrs* are appended as false positives to WBC and Twitter.
- SYN2 (multimodal distribution) is more complex than SYN1, which causes performance degradation for CAL and LIME.
- COIN is effective especially on complex datasets.

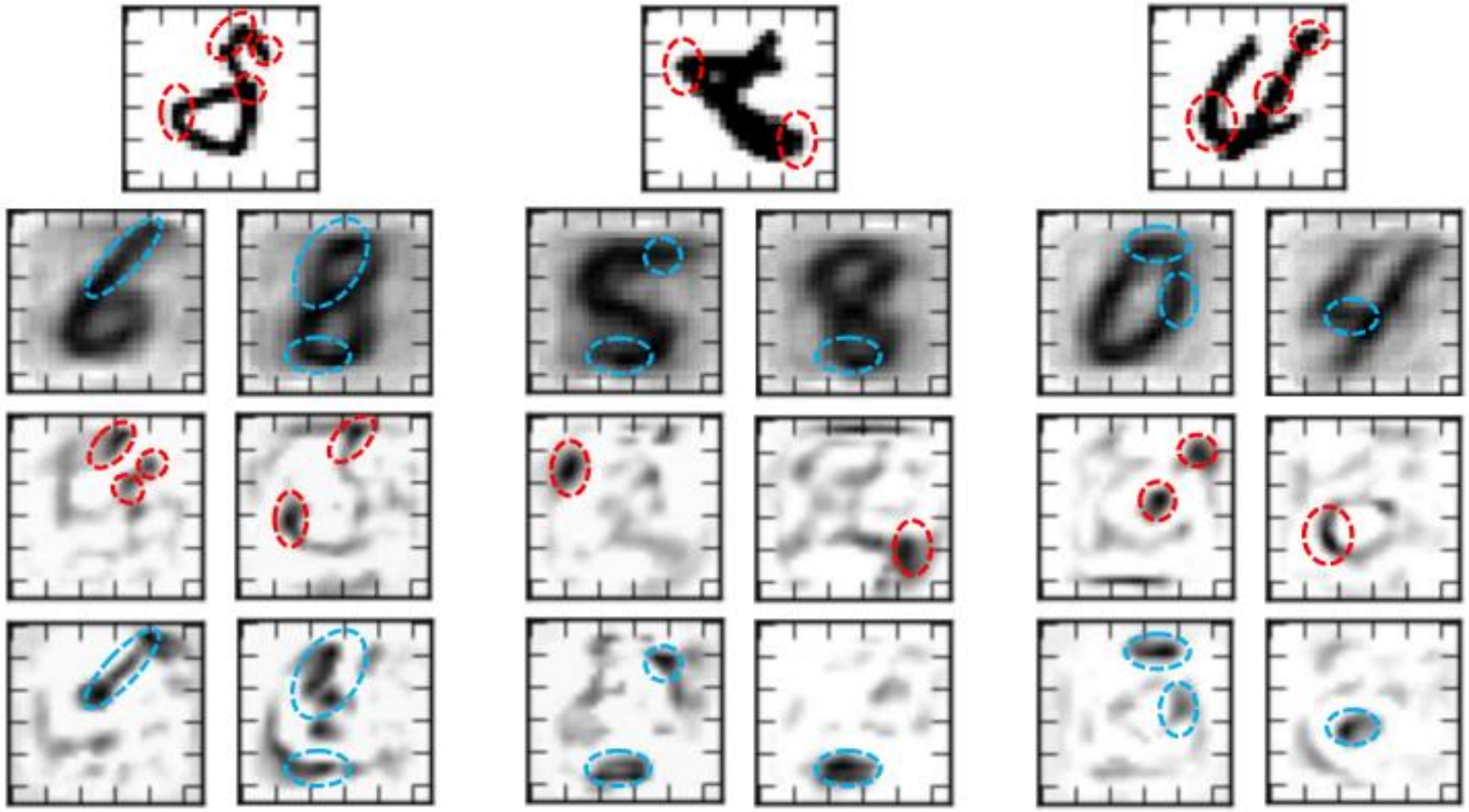
Experiments - Case Study

Query Outliers

Context Clusters
(Two for each query)

Positive Outlying
Regions

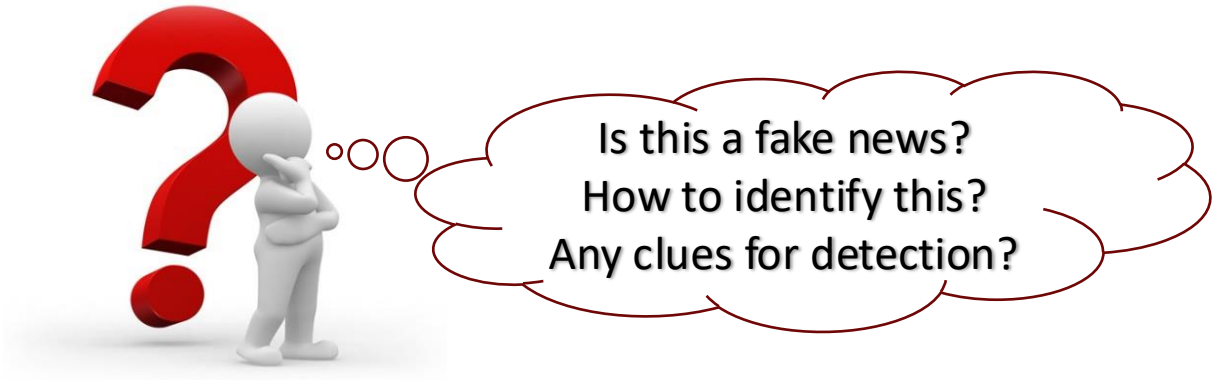
Negative
Outlying Regions



Outline

1. Introduction to Interpretable Machine Learning
2. Interpretable Deep Learning
3. Evaluation of Interpretation
4. Applications To Four Domains
   - *Explaining CNN for Image Classification*
   - *Explaining Recommender System*
   - *Explaining Outlier Detection System*
   - *Demo for Interpretable Fake News Detection*

Interpretable Fake News Detection



Beyond Text Classifications ---

❖ More challenging given heterogeneous types of information

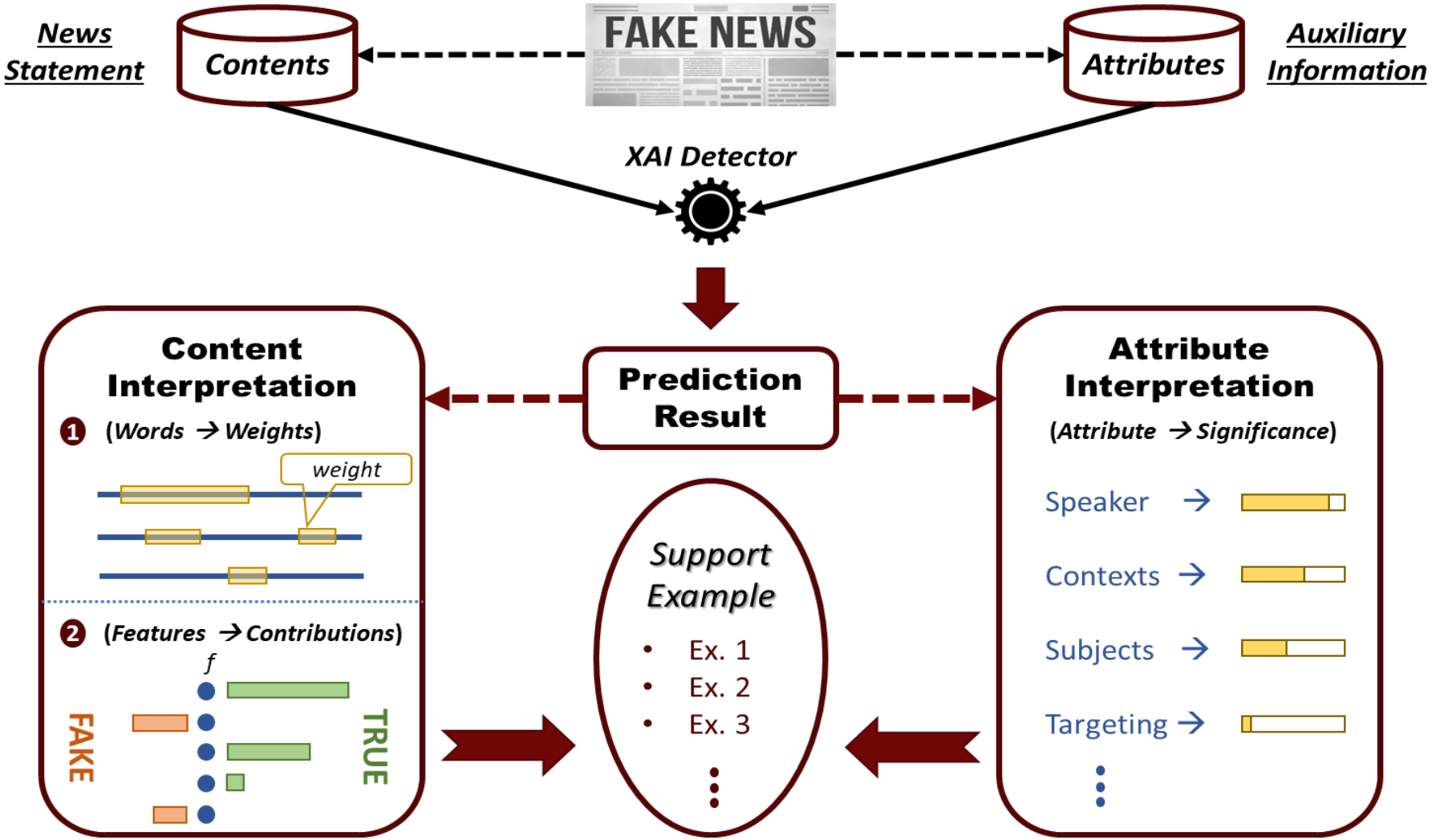
Hard to Achieve Effective Interpretations ---

❖ Various aspects including the person, the statement or the other contexts

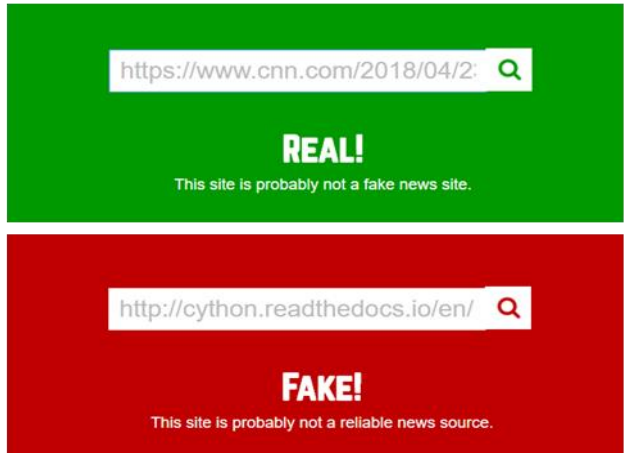
Beyond the News Itself ---

❖ Further supports are needed to convince people about the interpretations

Our Built System



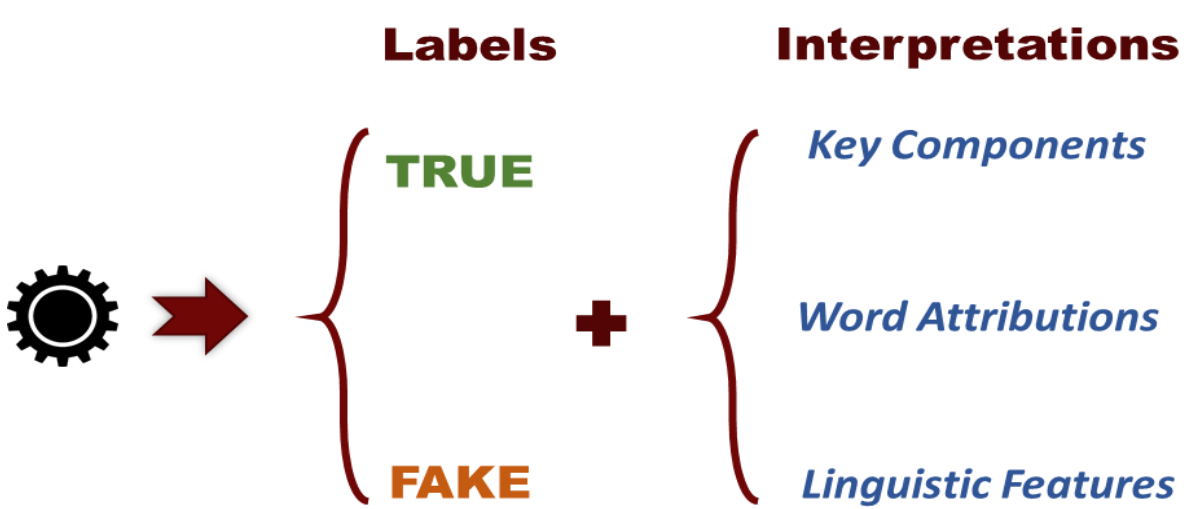
Interpretable System Outputs



Current AI Model
on *Fake News Detection*

(*http://www.fakenewsai.com/*)

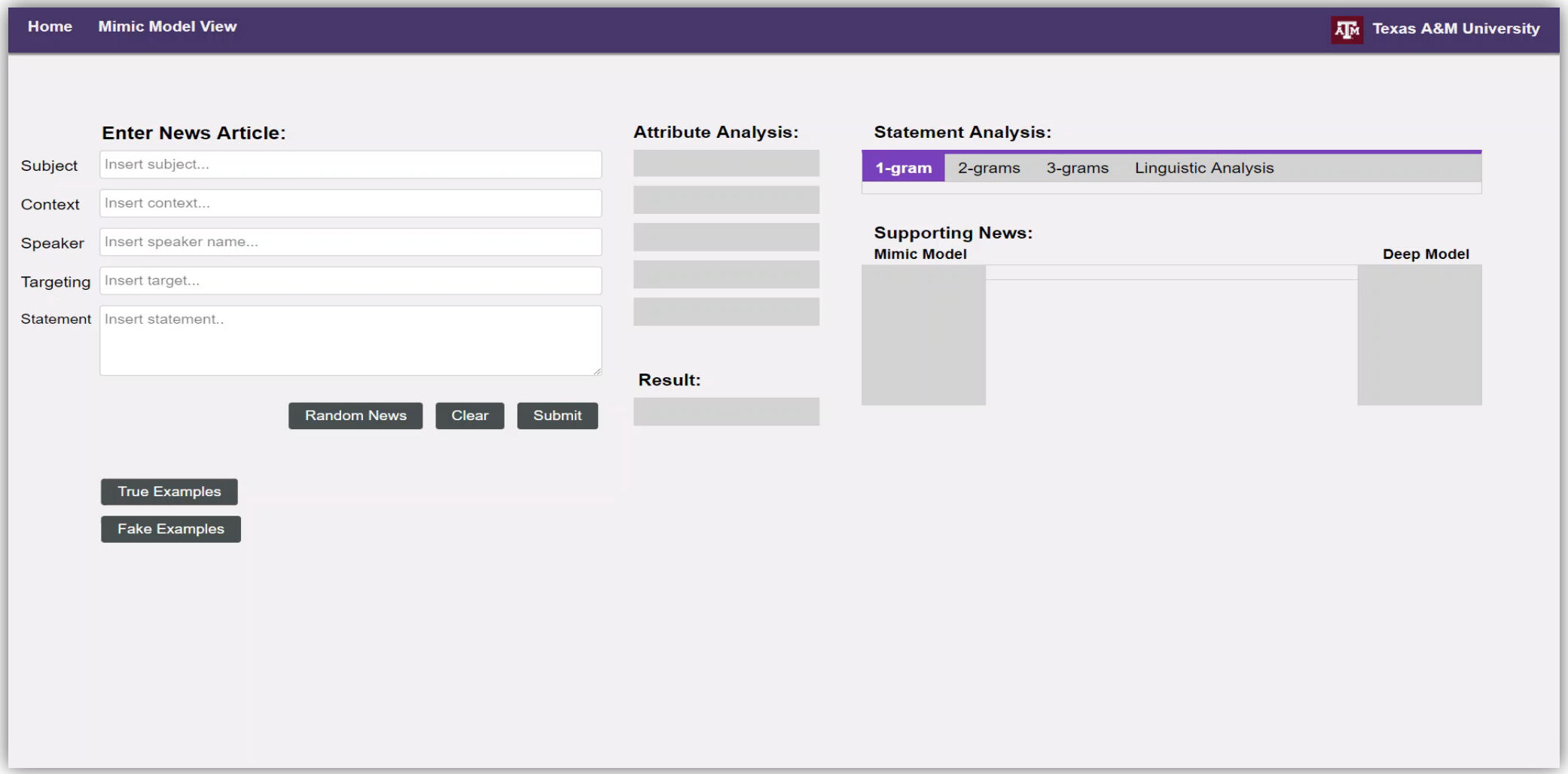
Prediction Labels Only



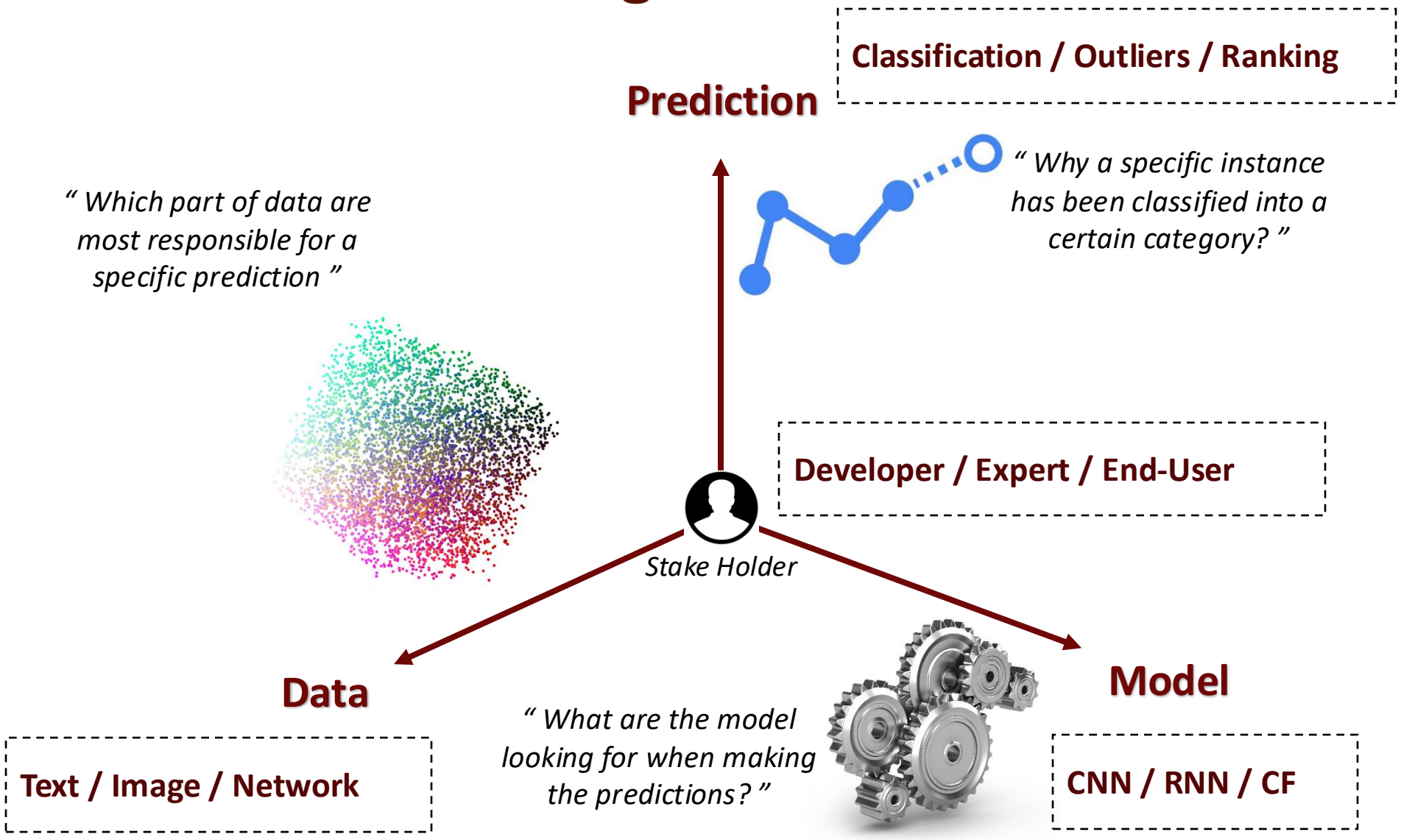
Proposed XAI Model
on *Fake News Detection*

Prediction Labels
+
Interpretations
+
Meta-Interpretations

Demo Video


Home
Mimic Model View
Texas A&M University
Enter News Article:
Subject
Insert subject...
Context
Insert context...
Speaker
Insert speaker name...
Targeting
Insert target...
Statement
Insert statement..
Random News
Clear
Submit
True Examples
Fake Examples
Attribute Analysis:
Result:
Statement Analysis:
1-gram
2-grams
3-grams
Linguistic Analysis
Supporting News:
Mimic Model
Deep Model

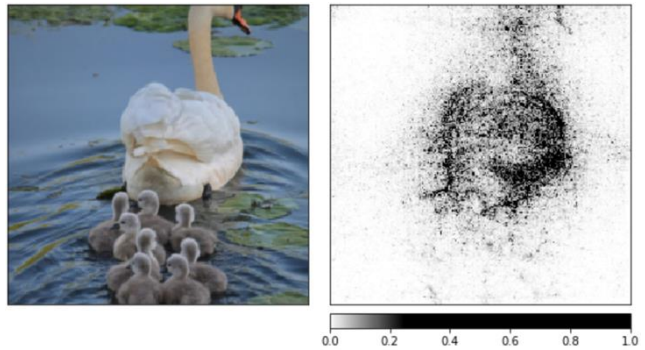
Interpretable Machine Learning



Captum Open-source Package for XAI

- Captum is an open source, extensible library for model interpretability built on PyTorch
- For example:

```
>>> net = ImageClassifier()
>>> ig = IntegratedGradients(net)
>>> input = torch.randn(2, 3, 32, 32, requires_grad=True)
>>> # Computes integrated gradients for class 3.
>>> attribution = ig.attribute(input, target=3)
```



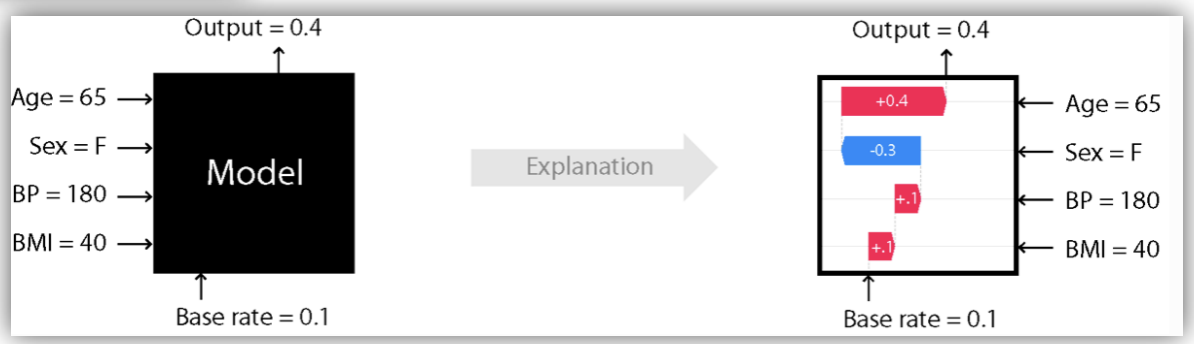

```
>>> net = SimpleClassifier()

>>> # Generating random input with size 1 x 4 x 4
>>> input = torch.randn(1, 4, 4)

>>> # Defining KernelShap interpreter
>>> ks = KernelShap(net)
>>> # Computes attribution, with each of the 4 x 4 = 16
>>> # features as a separate interpretable feature
>>> attr = ks.attribute(input, target=1, n_samples=200)
```

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.96)	pos	1.29	it was a fantastic performance ! #pad
pos	pos (0.87)	pos	1.56	best film ever #pad #pad #pad #pad
pos	pos (0.92)	pos	1.14	such a great show ! #pad #pad
neg	neg (0.29)	pos	-1.11	it was a horrible movie #pad #pad
neg	neg (0.22)	pos	-1.03	i 've never watched something as bad
neg	neg (0.07)	pos	-0.84	that is a terrible movie . #pad

SHAP Open-source Package for XAI

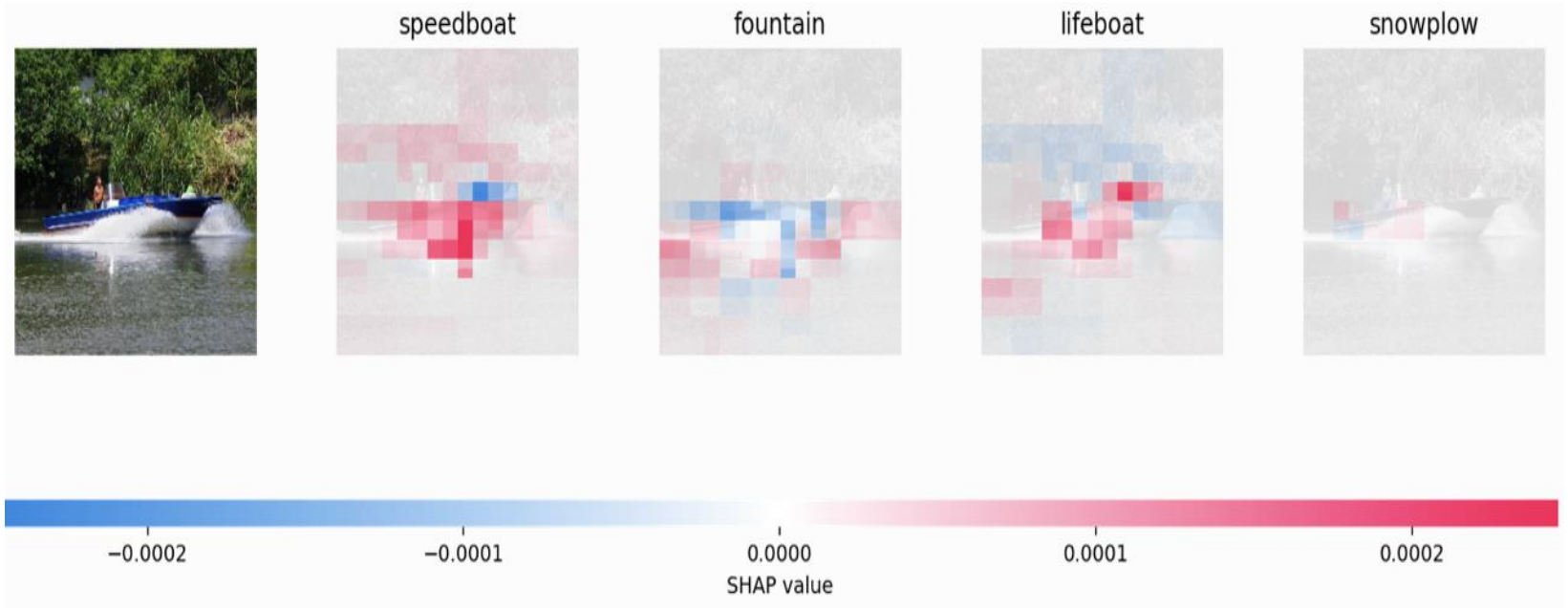
- SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model.
- For example:

```
def f(x):
    tmp = x.copy()
    preprocess_input(tmp)
    return model(tmp)

# define a masker that is used to mask out partitions of the input image.
masker = shap.maskers.Image("inpaint_telea", X[0].shape)

# create an explainer with model and image masker
explainer = shap.Explainer(f, masker, output_names=class_names)

# here we explain two images using 500 evaluations of the underlying model to estimate the
shap_values = explainer(X[1:3], max_evals=100, batch_size=50, outputs=shap.Explanation.arg
```

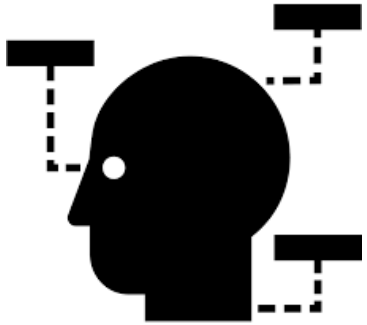


Open-source Package for XAI

- SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model.
- For example:



Human-Centric Machine Learning Algorithms and Systems



How to facilitate ***trustworthy AI?***

Interpretable ML
Fairness in ML

Enable transparency for human to *understand and trust* ML systems

How to build ***automated and efficient*** ML systems***?***

AutoML
Efficient Foundation Models

Democratize ML systems to be easily used by all

Acknowledgements

- DATA Lab Members and Collaborators
- Funding Agencies
    - --- *Federal Agencies (DARPA, NSF, NIH and DoT)*
    - --- *Industrial Sponsors (Apple, Google, JP Morgan, Meta, Samsung etc.)*
- Everyone attending the talk!