COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Interpretable Machine Learning: Concepts and Techniques

Xia "Ben" Hu

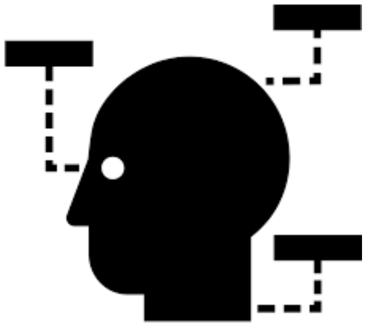
Assistant Professor and Lynn '84 and Bill Crane '83 Faculty Fellow

Department of Computer Science and Engineering

Texas A&M University

Email: hu@cse.tamu.edu

Human-Centric Machine Learning



How to enable ***interpretable*** and ***Interactive*** machine learning?

Interpretable Machine Learning
(IML)

Provide explanations for human to ***easily understand*** the system

How to enable ***automated*** knowledge discovery and learning?

Automated Machine Learning
(AutoML)

Provide convenience for human to ***easily build*** the system

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

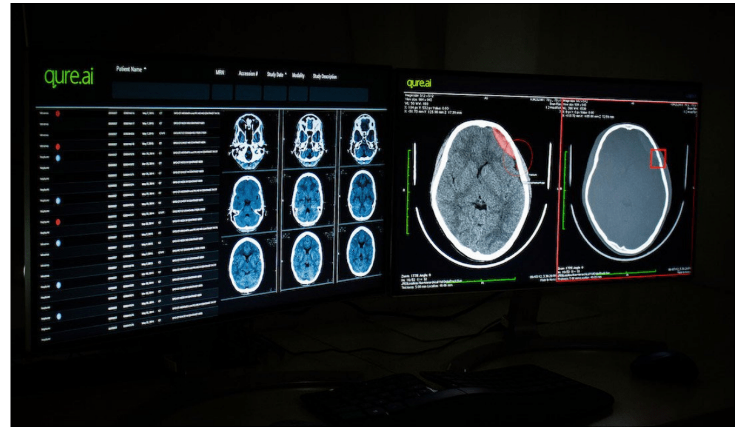
OUTLINE

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

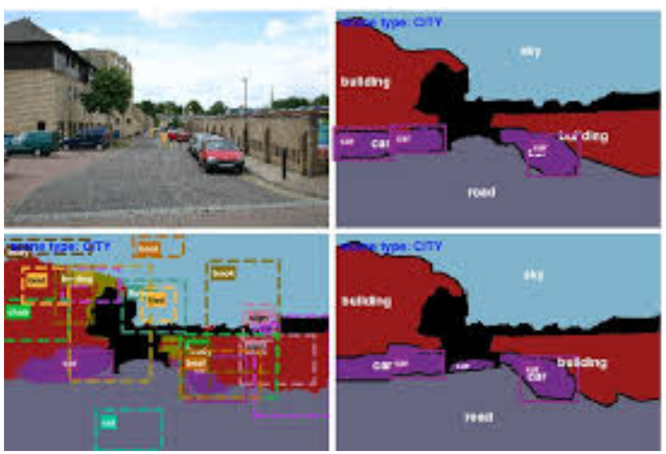
Machine Learning is Everywhere

Playing Go

Medical Diagnosis



Scene Understanding



Voice Recognition

Machine Learning is Everywhere

Playing Go

Medical Diagnosis

What have been learned inside the models?

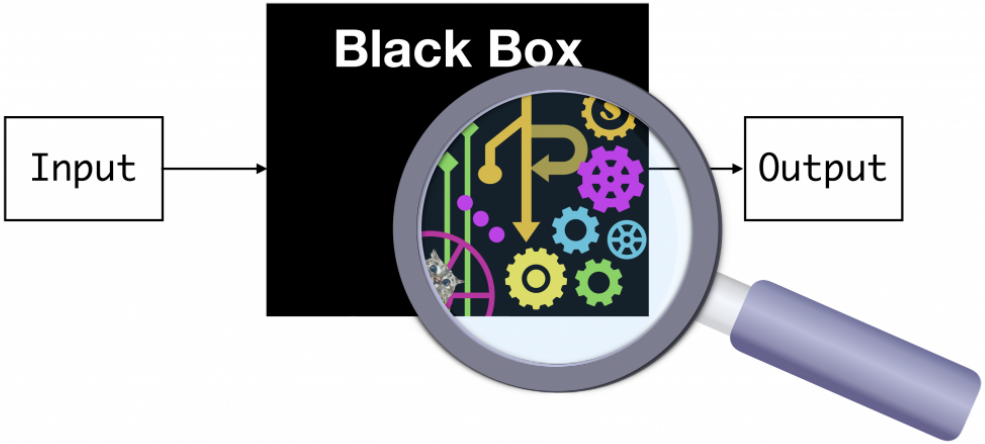
Why Interpretable Machine Learning

Safety of AI Models

Trust of AI Decision

Policy and Regulation

What is Interpretable Machine Learning

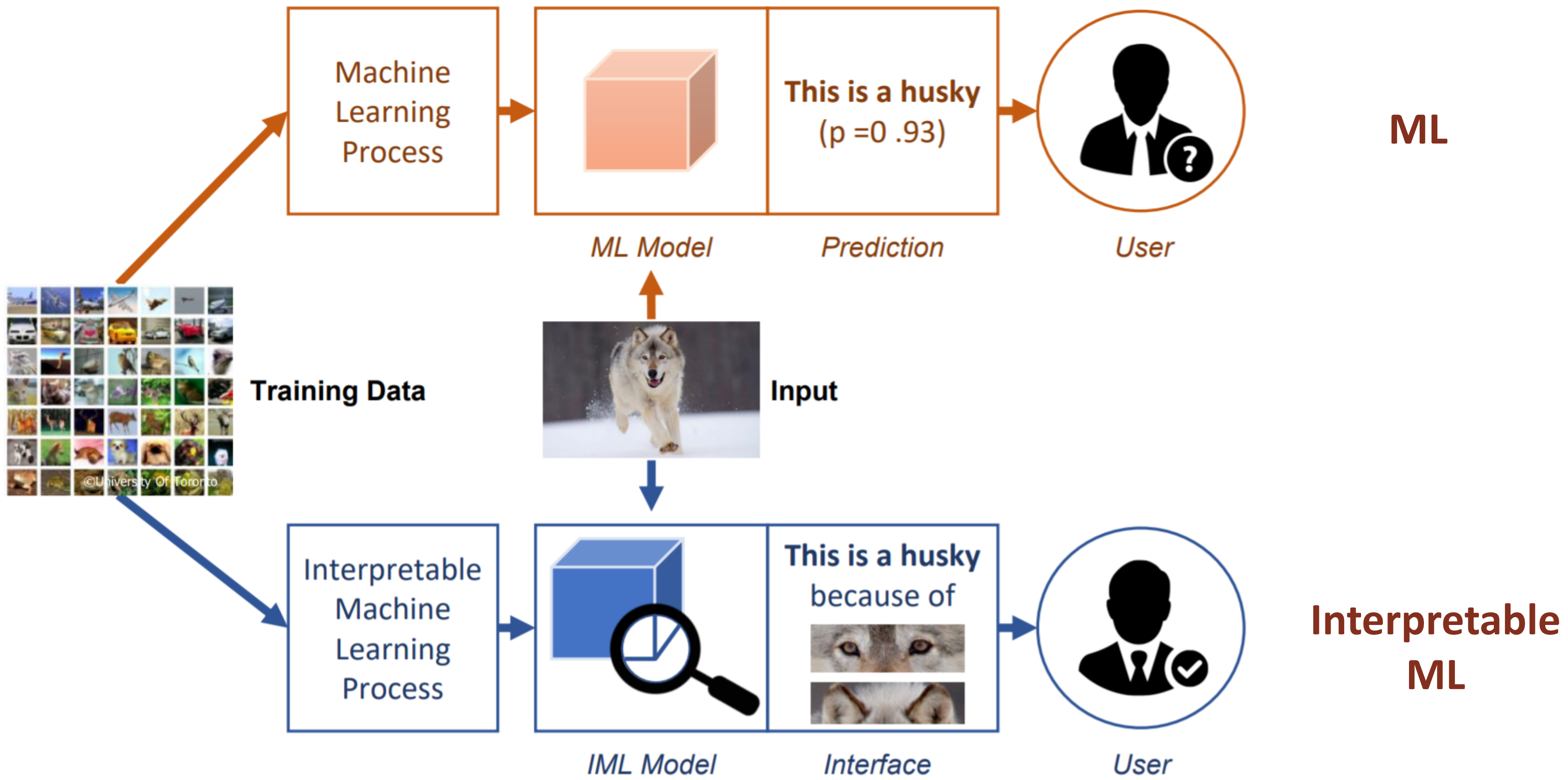


"Interpretable Machine Learning is the ability to explain or to present the behavior of a black-box ML model in understandable terms to a human" [1]

"We define interpretable machine learning as the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model. Here, we view knowledge as being relevant if it provides insight for a particular **audience** into a **chosen problem**. These insights are often used to guide communication, actions, and discovery." [2]

[1] Bang, Seojin, et al. "Explaining a black-box using deep variational information bottleneck approach." arXiv preprint arXiv:1902.06918 (2019).
[2] Murdock et al. "Interpretable machine learning: definitions, methods, and applications", PNAS 2019.

What is Interpretable Machine Learning



Interpretable Machine Learning

- Model-agnostic explanation
  - Broadly applicable to various machine learning models
  - Treating a model as a black-box
  - Does not inspect internal model parameters

- Model-specific explanation
  - Specifically designed for each model
  - Usually require examining internal structures and parameters

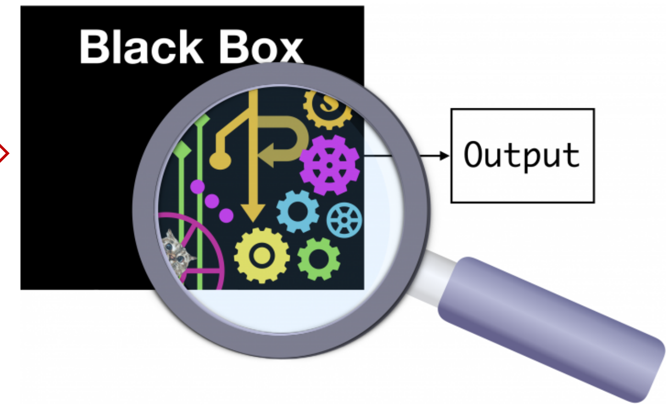
Model-agnostic explanation (permutation-based)

0	0	0	1	2
1	0	0	2	1
2	0	1	0	2
3	0	1	2	0
4	0	2	0	1
5	0	2	1	0
6	1	0	0	2
7	1	0	2	0
8	1	2	0	0
9	2	0	0	1
10	2	0	1	0
11	2	1	0	0

Input *X*

0	0	0	1	2
1	0	0	2	1
2	0	1	0	2
3	0	1	2	0
4	0	2	0	1
5	0	2	1	0
6	1	0	0	2
7	2	0	1	0
8	1	2	0	0
9	2	0	0	1
10	2	0	1	0
11	2	1	0	0

Permutated input X_p



Permutation feature importance

- For each feature, do permutation, and then retrain the model
- Repeating this for n times for each feature, and compare the model accuracy
- Rank accuracy

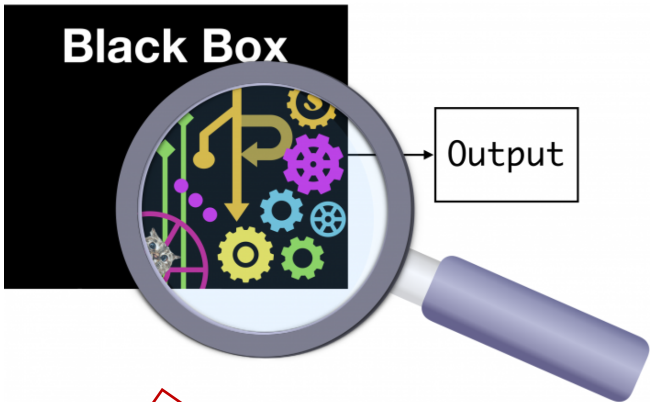
Model-agnostic explanation (permutation-based)

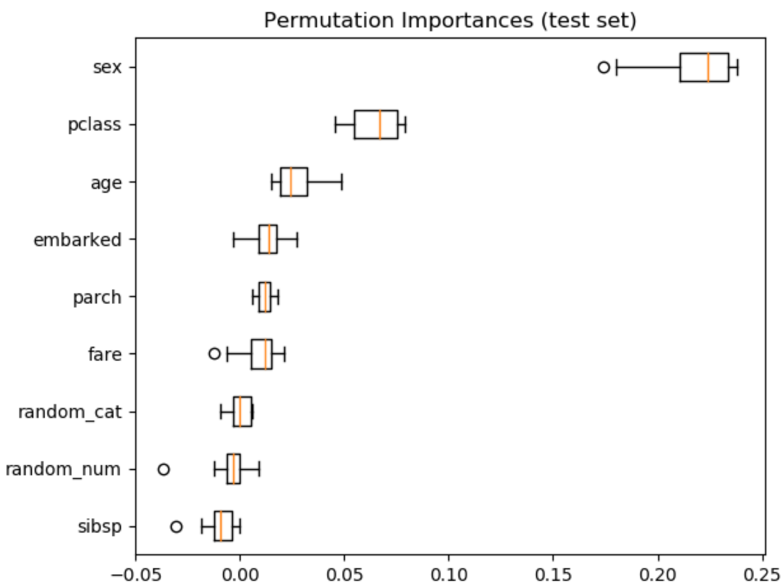
0	0	0	1	2
1	0	0	2	1
2	0	1	0	2
3	0	1	2	0
4	0	2	0	1
5	0	2	1	0
6	1	0	0	2
7	1	0	2	0
8	1	2	0	0
9	2	0	0	1
10	2	0	1	0
11	2	1	0	0

Input X

0	0	0	1	2
1	0	0	2	1
2	0	1	0	2
3	0	1	2	0
4	0	2	0	1
5	0	2	1	0
6	1	0	0	2
7	2	0	1	0
8	1	2	0	0
9	2	0	0	1
10	2	0	1	0
11	2	1	0	0

Permutated input X_p





Feature importance vector

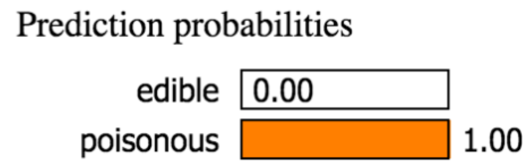
Model-specific explanation

Feature	Value			
odor=foul	1	0	1	0
gill-size=broad	0	1	1	0
stalk-surface-above-ring=silky	0	0	0	1
spore-print-color=chocolate	1	1	1	1
stalk-surface-below-ring=silky	0	1	0	1

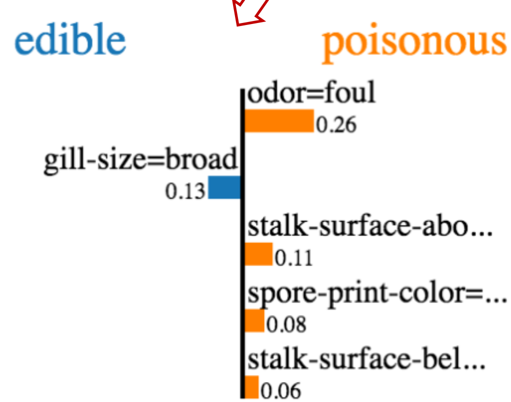
Input *X*

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

LASSO Model



Output *y*

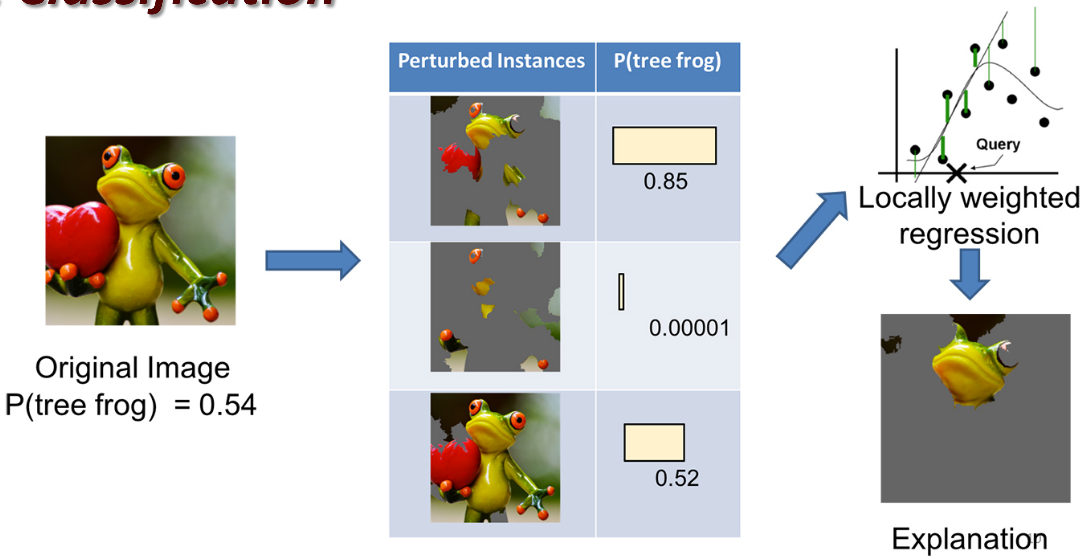


Interpretation to LASSO model: Feature importance vector of The linear weight β

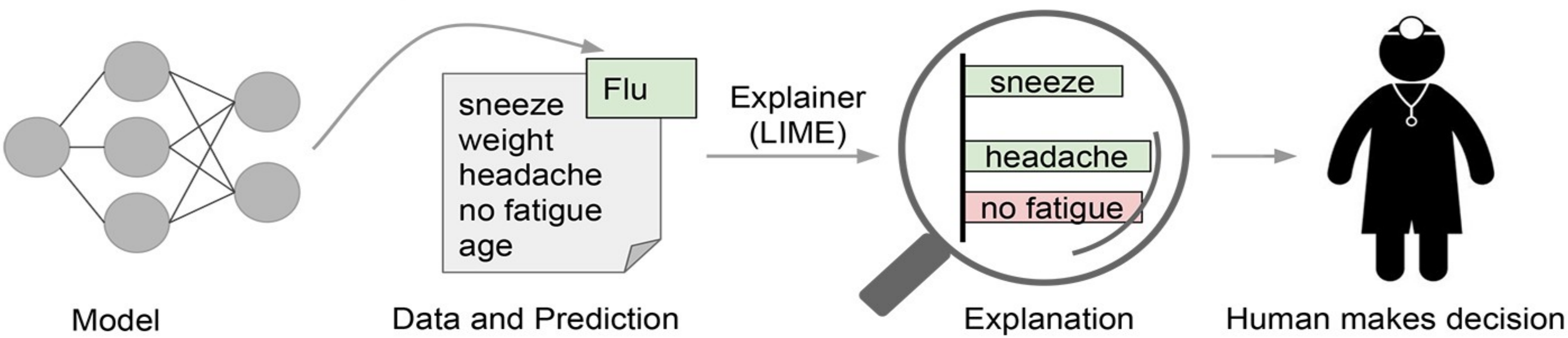
https://github.com/marcotcr/lime

More Examples of IML

❶ *Image Classification*



❷ *Medical Diagnosis*



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." KDD. 2016.

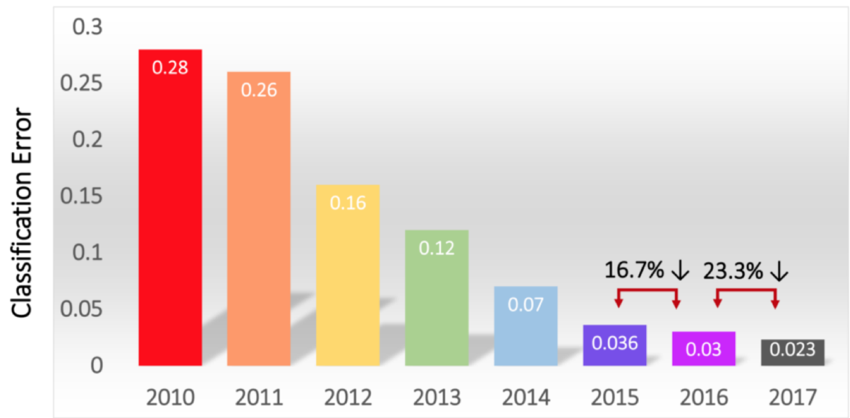
OUTLINE

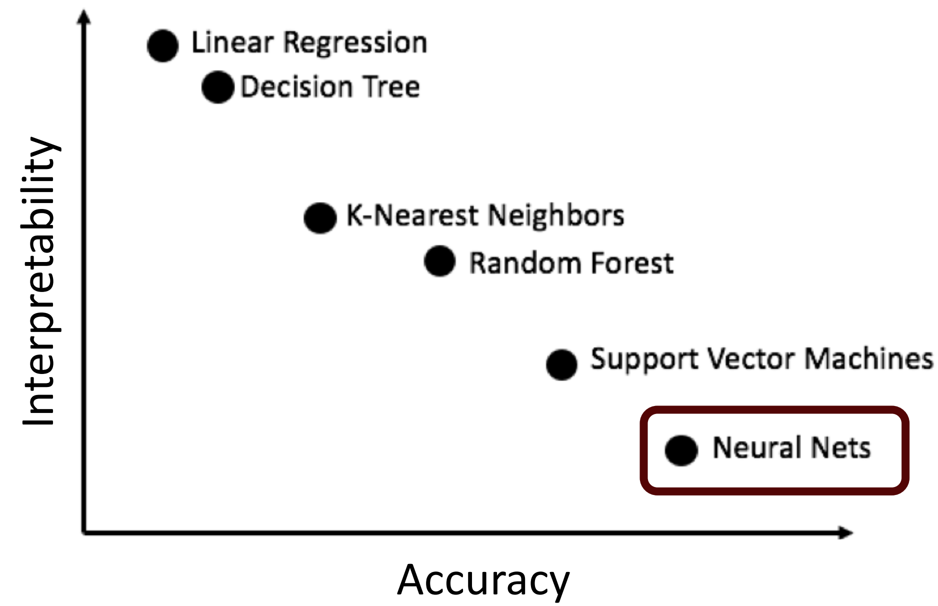
1. Introduction to Interpretable Machine Learning
2. **Interpretable Deep Learning**
3. Evaluation of Interpretation
4. Applications To Four Domains
    - Explaining CNN for Image Classification
    - Explaining Recommender System
    - Explaining Outlier Detection System
    - Demo for Interpretable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

Interpretable Deep Learning

IMAGENET

Classification Results (CLS)

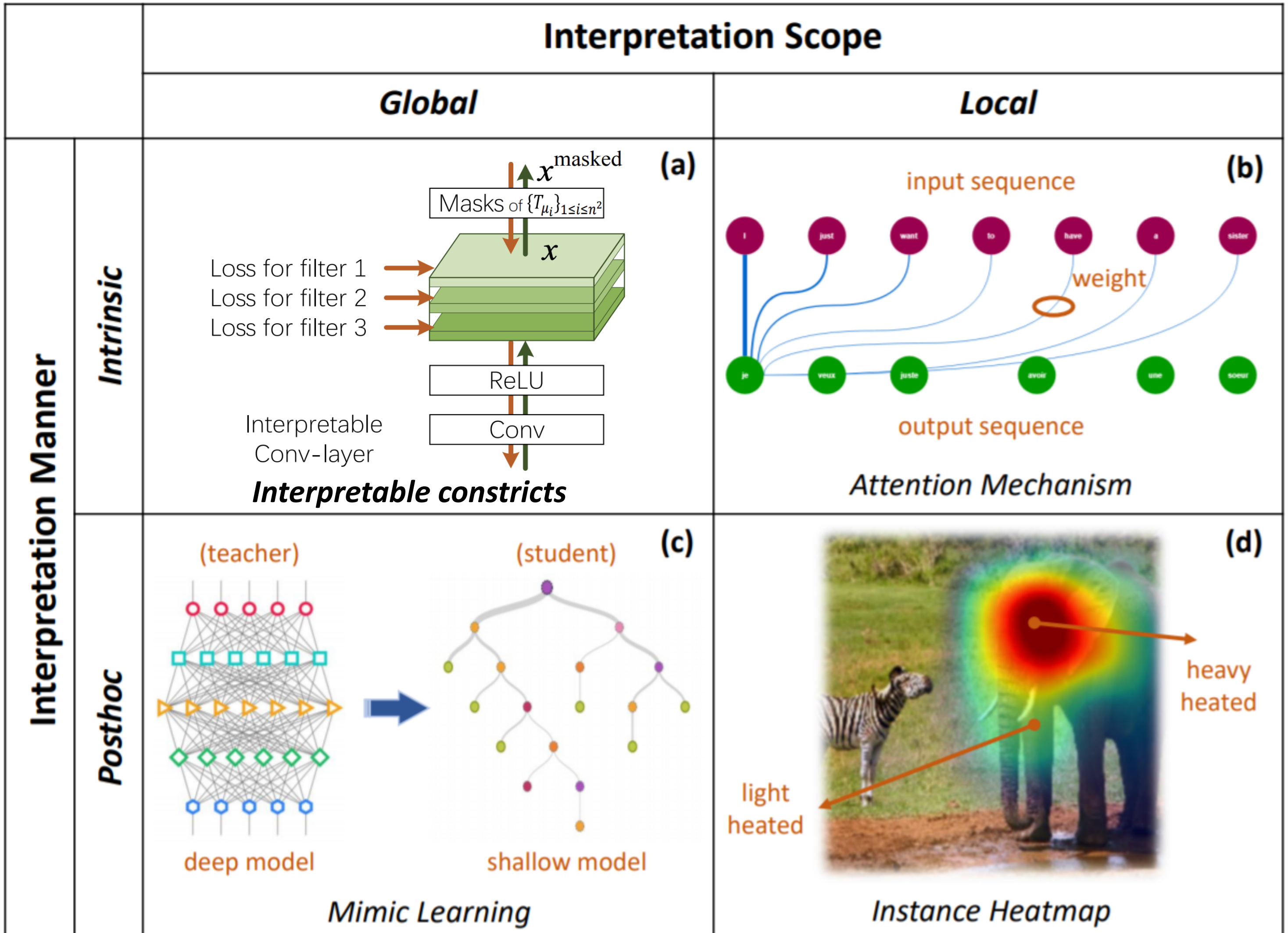




DNNs make lots of *progresses*

DNNs are regarded as *black boxes*

Interpretable Deep Learning Categorization

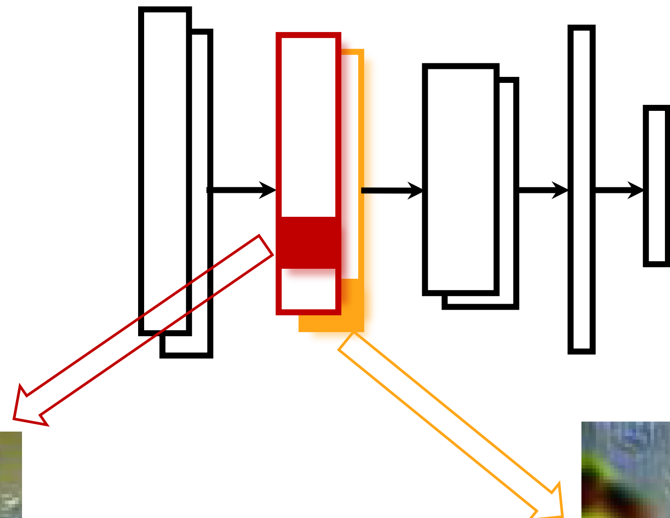


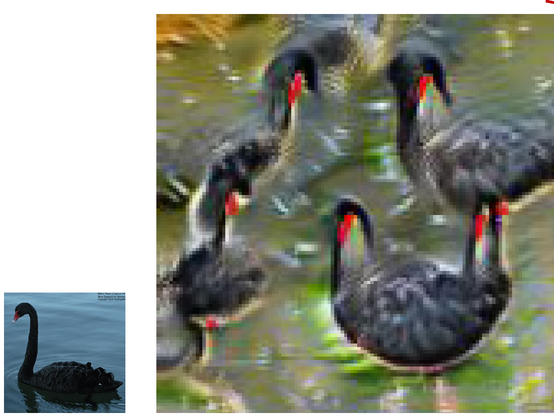
Post-hoc Explanation (Global)

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

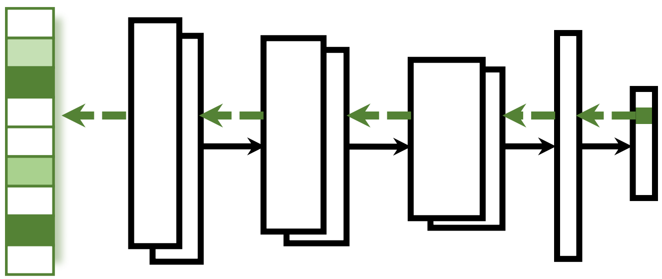
Activation maximation

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





Post-hoc Explanation (Local)

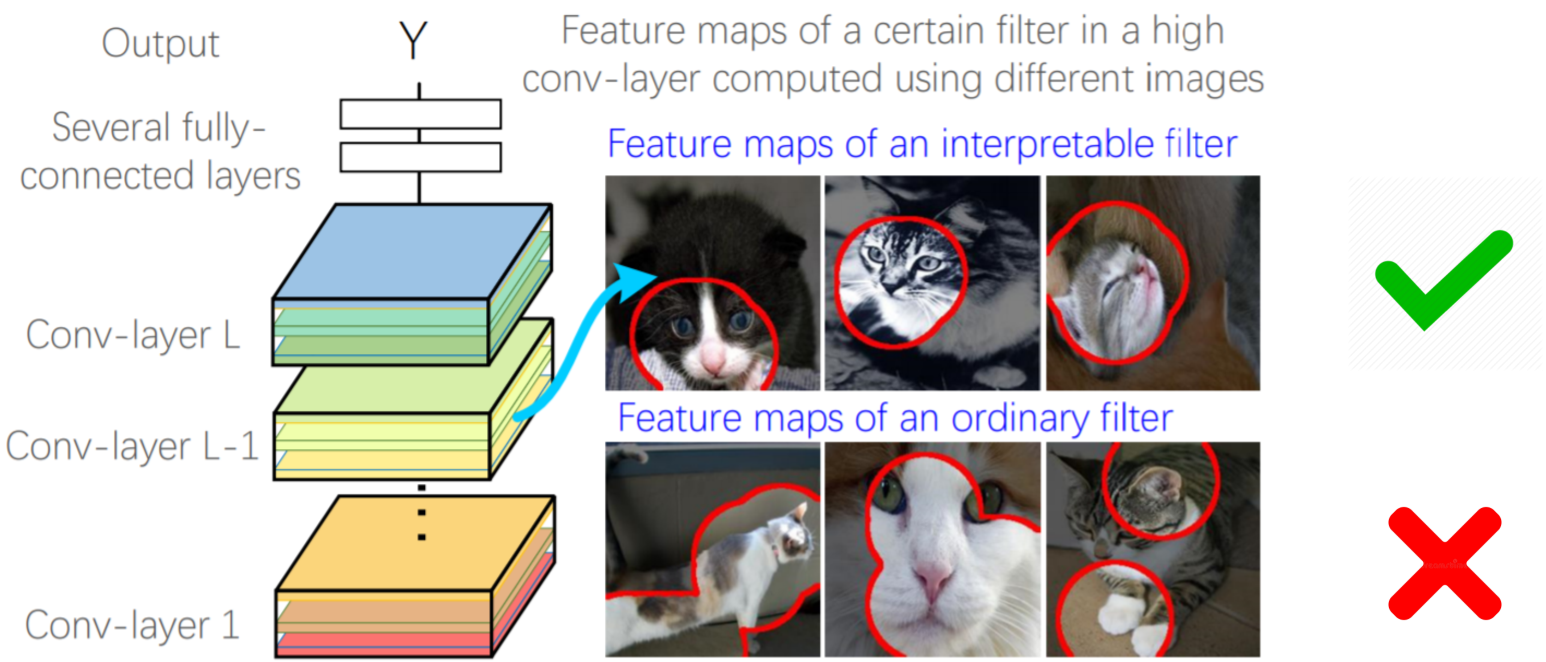


Post-hoc Local Interpretation

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

Intrinsic Interpretable Model (Global)

Globally interpretable models that offer a certain extent of transparency about what is going on inside a model.

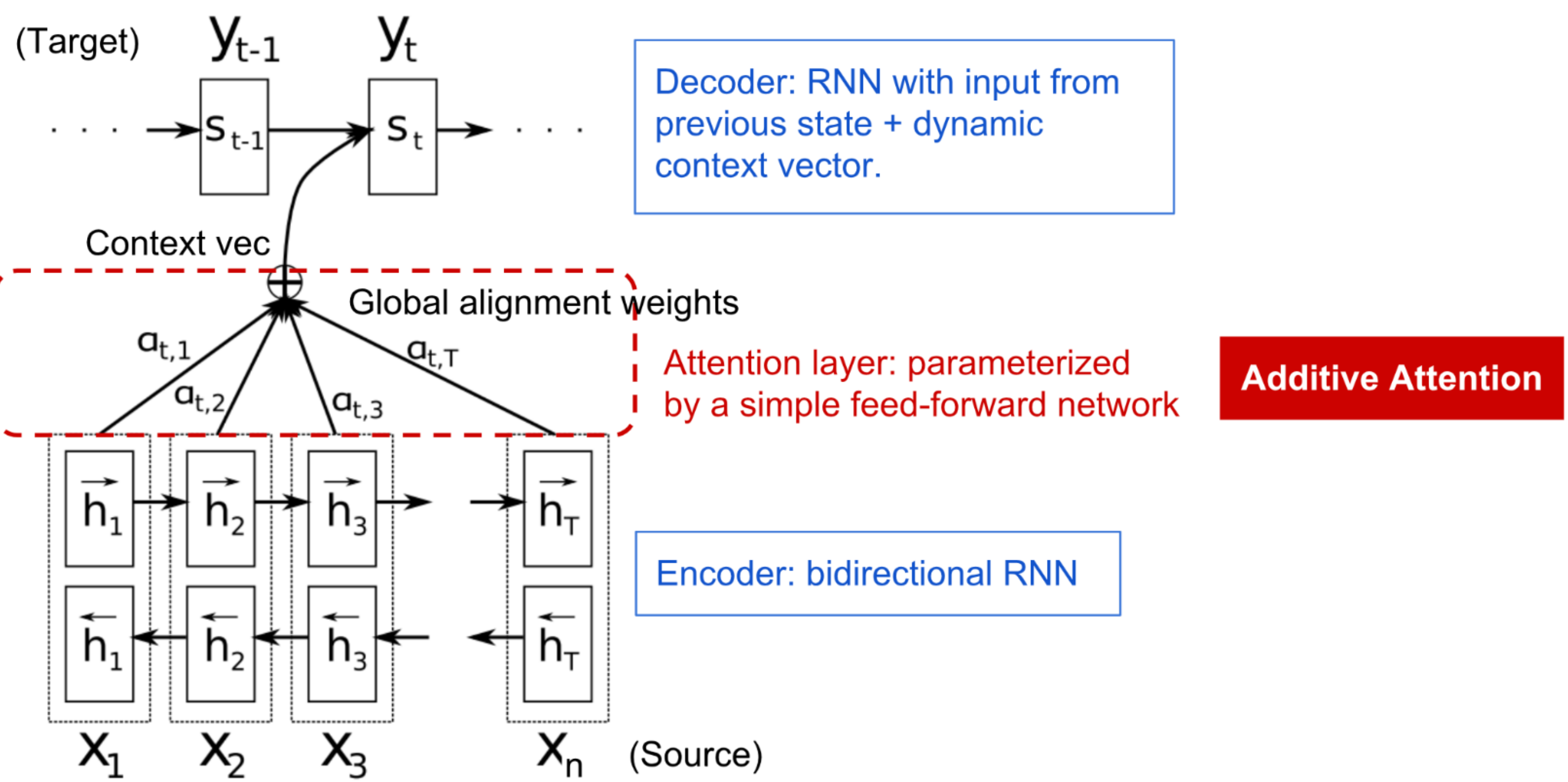


In interpretable CNN, each filer in high-layers represents a specific object part.

Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu. "Interpretable convolutional neural networks." CVPR. 2018.

Intrinsic Interpretable Model (Local)

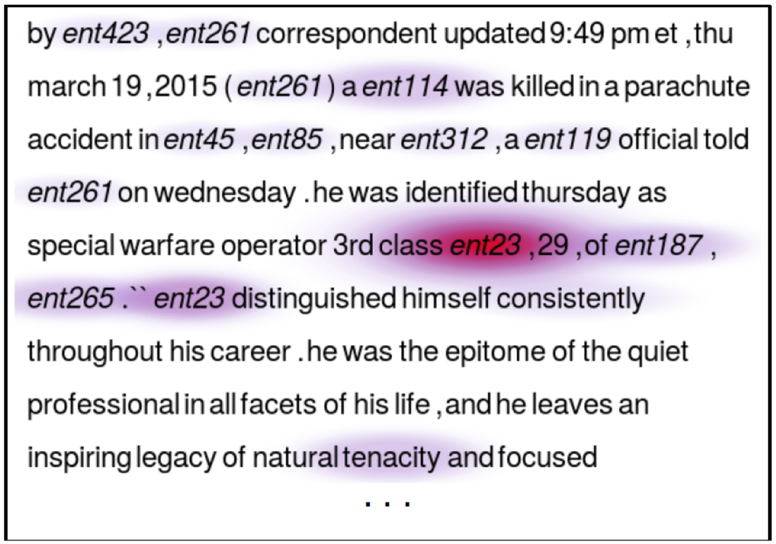
Designing more justified model architectures that could explain why a specific decision is made



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

Intrinsic Interpretable Model (Local)

Designing more justified model architectures that could explain why a specific decision is made



Interpretation heatmap

Interpretation Visualization

--- Contribution score for each feature in input

--- Deeper color in the heatmap means higher contribution

Hermann, Karl Moritz, et al. "Teaching machines to read and comprehend." Advances in neural information processing systems. 2015.

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 for Interpretable Machine Learning

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

Fidelity

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

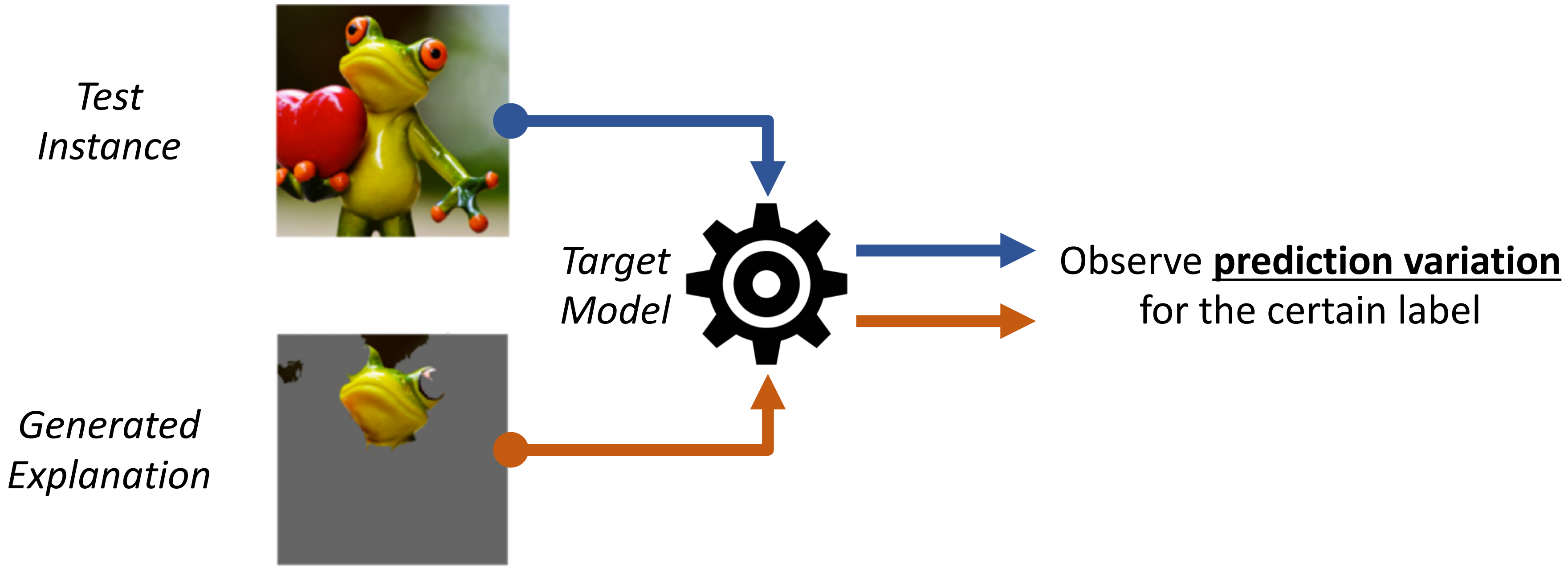
Are the generated explanations ***friendly*** to the human users?

Persuasibility

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

Philosophy of Fidelity Evaluation

Ablation Analysis

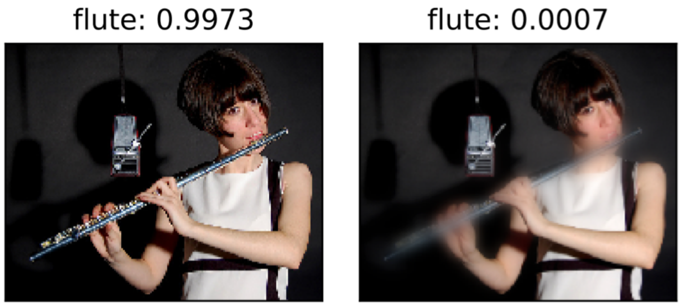


If the generated explanation is **faithful** to the target model, the **prediction variation** should be **small**.

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

Positive (99.74%)

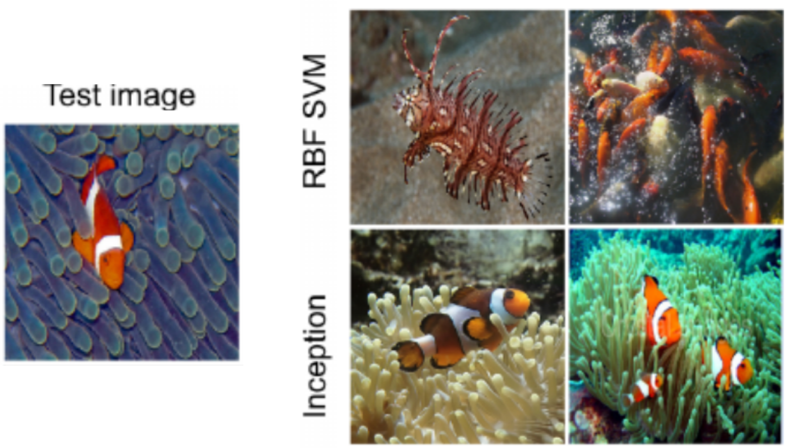
Occasionally melodramatic, it 's also extremely effective.

Negative (99.00%)

Occasionally melodramatic, it 's also terribly effective.

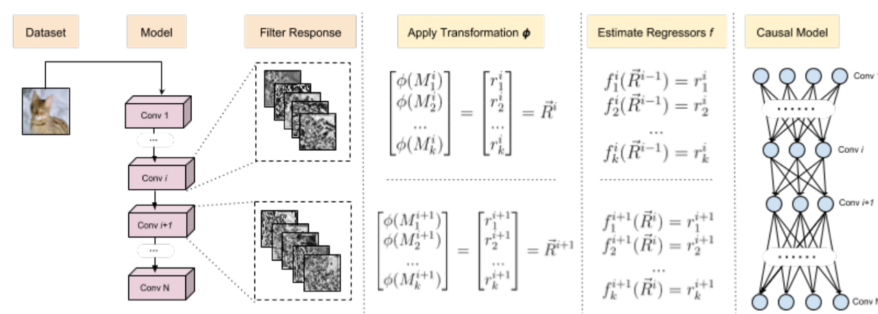
Du, Mengnan, et al. “On attribution of recurrent neural network predictions via additive decomposition.” The WebConf, 2019.

Training Data



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

Model Component

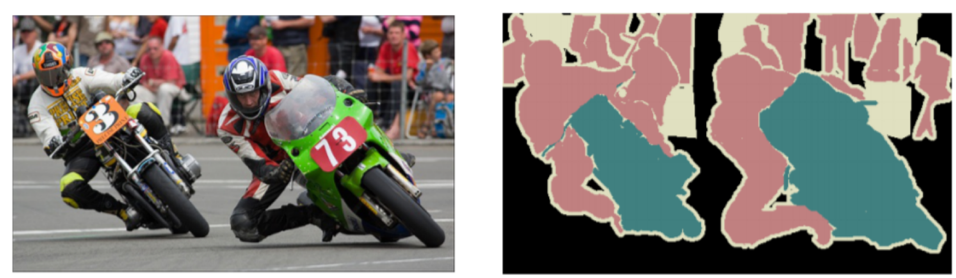


Narendra, Tanmayee, et al. "Explaining deep learning models using causal inference." arXiv, 2018.

Persuasibility Evaluation with Image Bounding

Evaluation with Bounding Box

Evaluation with Semantic Segmentation



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

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

Persuasibility Evaluation 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*.

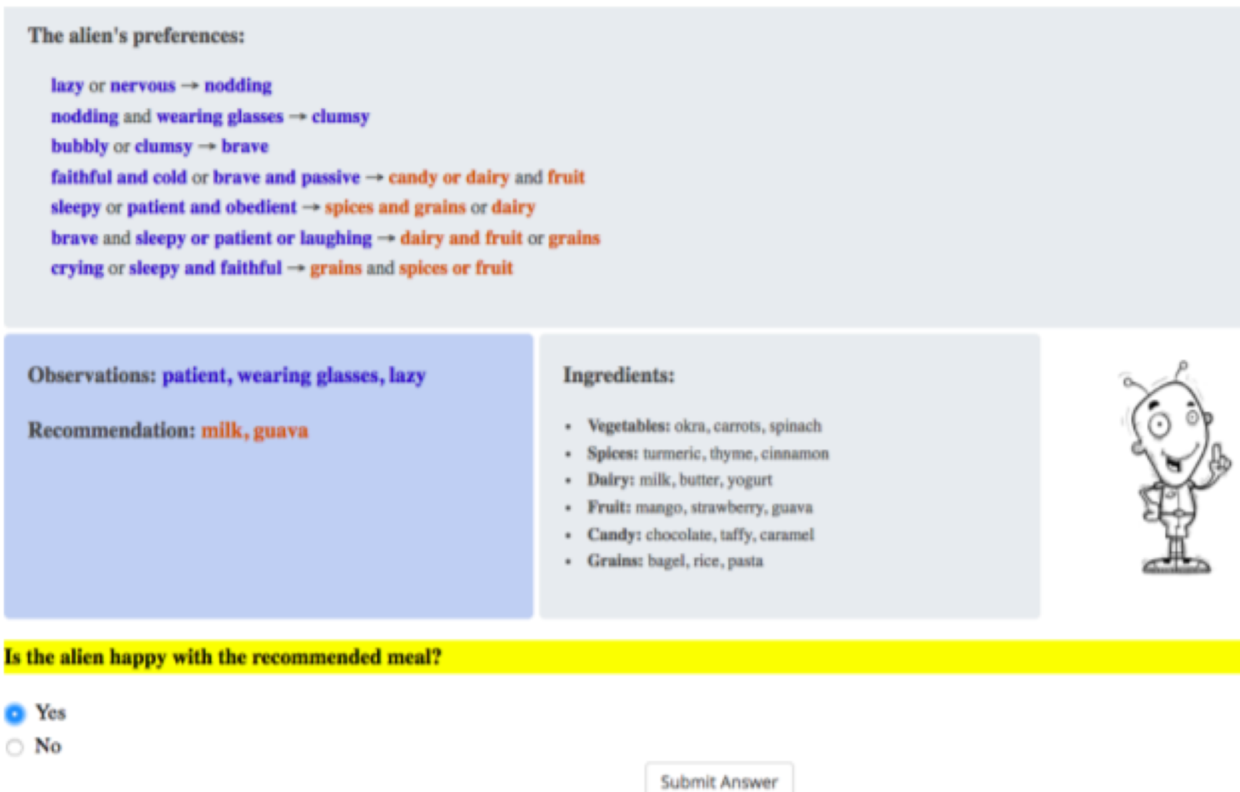
Task: beer appearance **Label**: positive

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.

Du, Mengnan, et al. "Learning credible deep neural networks with rationale regularization." ICDM, 2019.

Persuasibility Evaluation 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

Interpretable Machine Learning

Task

Classification / Ranking / Outliers

“ Why a specific instance has been classified into a certain category? ”

“ Which part of data are most responsible for a specific prediction ”

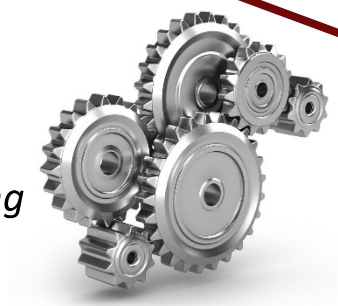
Developer / Expert / End-User

Stake Holder

Data

Image / Text / Tabular

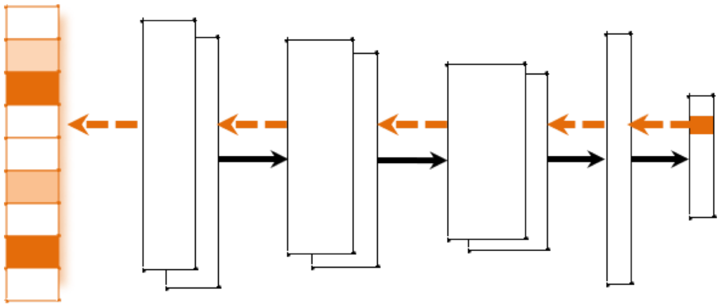
“ What are the model looking for when making the predictions? ”



Model

CNN / CF / OD

Post-Hoc CNN Interpretation



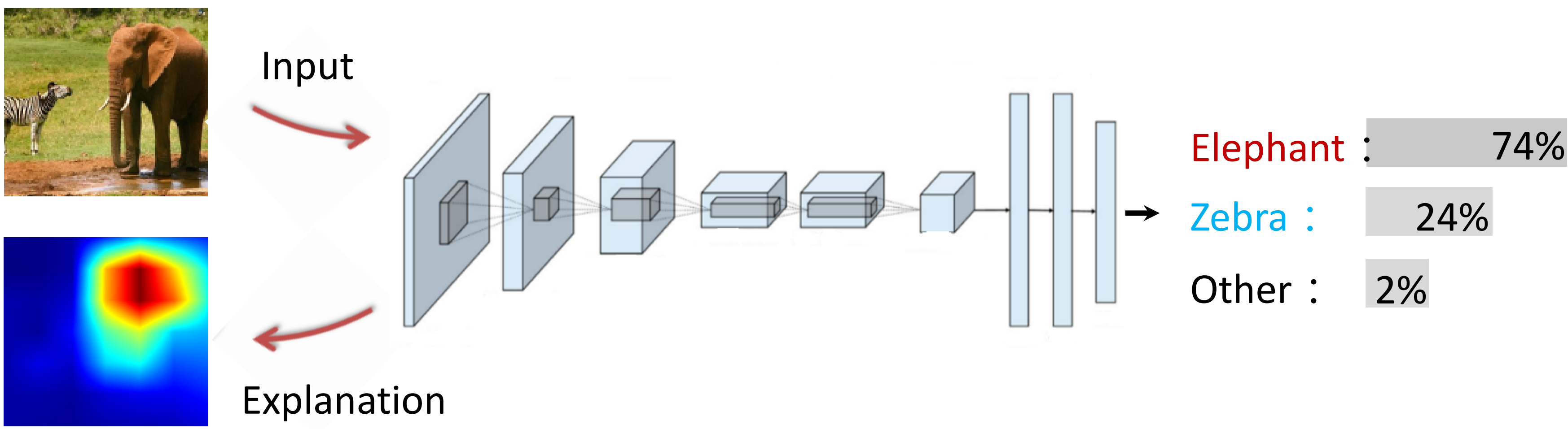
Key factors

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

--- The prediction of DNN

Post-hoc Interpretation

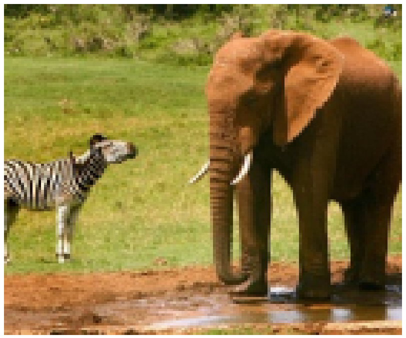
--- Contribution score for each feature in input

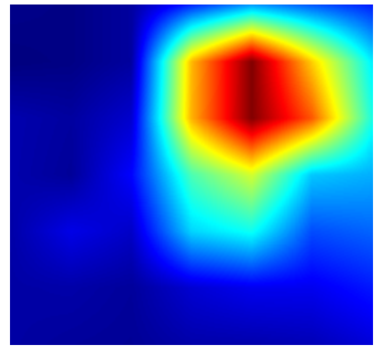


*Interpretation is to find **evidence** for prediction "elephant"*

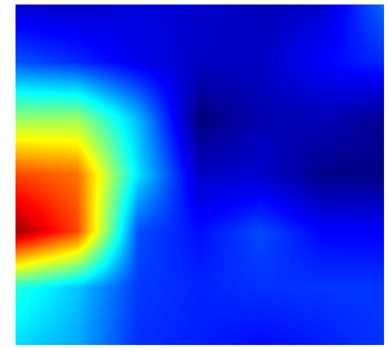
Motivation: Using ***deep representations*** in intermediate layers to derive interpretations

Challenges





Elephant



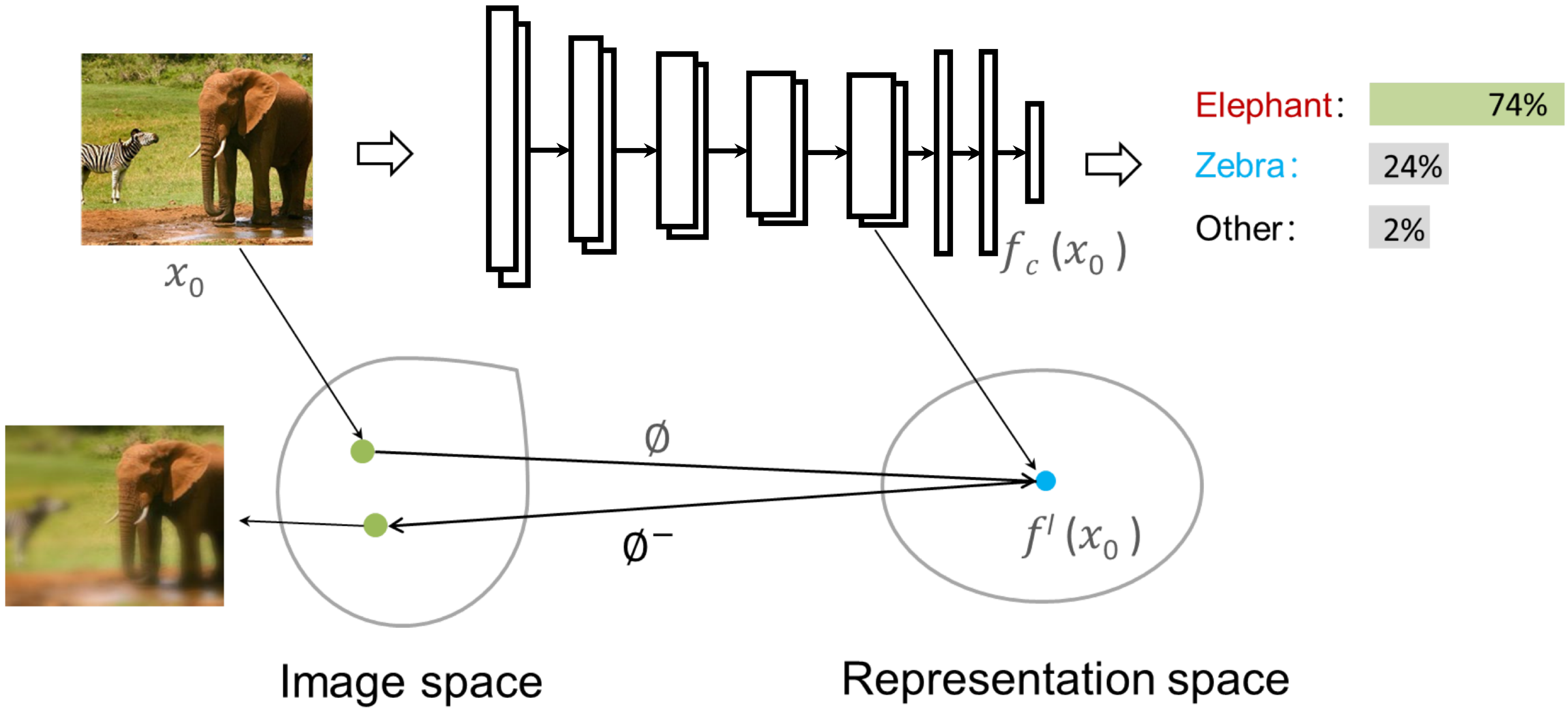
Zebra

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

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

Representation Inversion

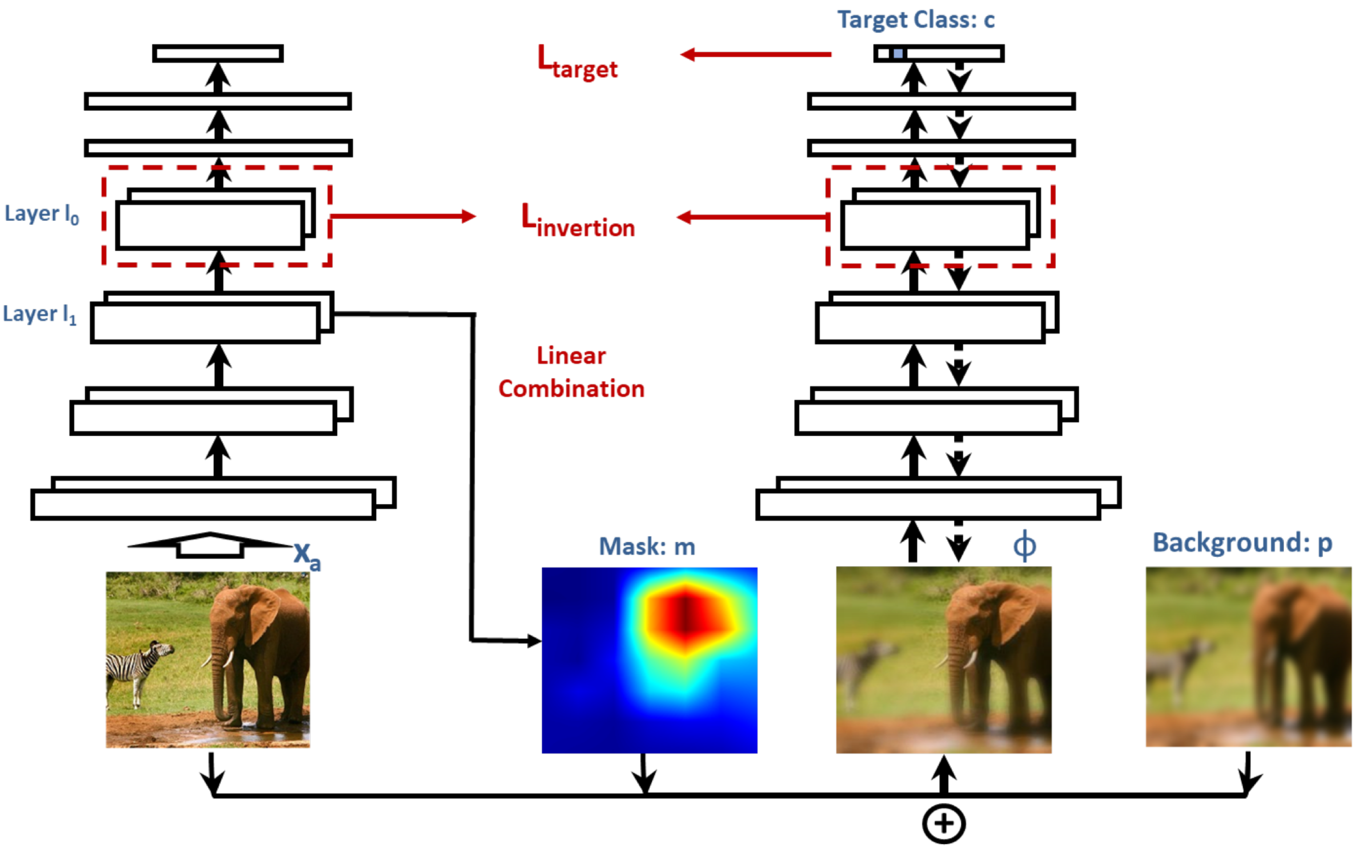
Sub-network ∅ maps **input $\boldsymbol{x_0}$** to a **representation $\boldsymbol{f^l(x_0)}$**



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

Aravindh Mahendran and etc, “Understanding deep image representations by inverting them”. CVPR, 2015.

The Proposed Model



- ***Guided feature inversion*** to preserve the object location in a mask
- ***Model target neuron in output layer*** to get class-discriminative interpretation
- ***Regularization by inner layers*** to further reduce artifacts

Guided Feature Inversion

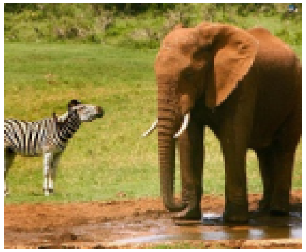
Representation of the **inverted input**

Representation of the **original input**

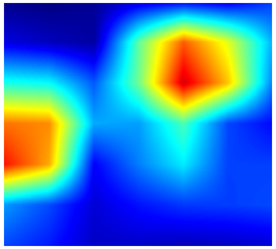
$$L_{\text{inversion}}(\mathbf{x}_a, \mathbf{m}) = \|\mathbf{f}^{l_0}(\boxed{\Phi(\mathbf{x}_a, \mathbf{m})}) - \mathbf{f}^{l_0}(\mathbf{x}_a)\|^2 + \alpha \cdot \frac{1}{d} \sum_{i=1}^{d} \mathbf{m}_i$$

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

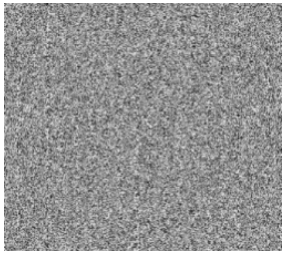
Inverted input



Original input

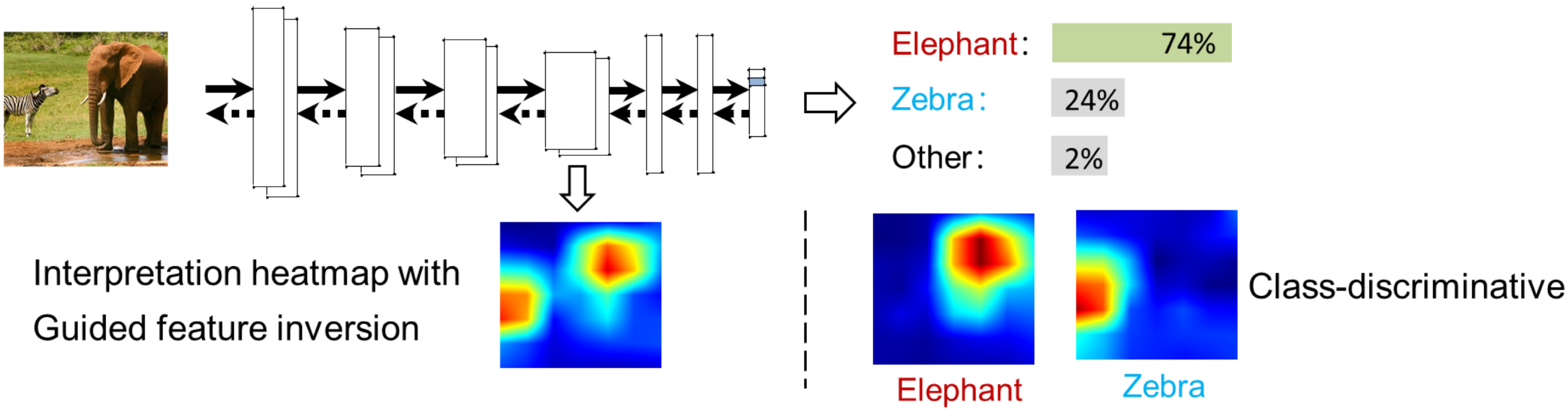


Weight matrix



Noise

Class-Discriminative



Output of the **target neuron**

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

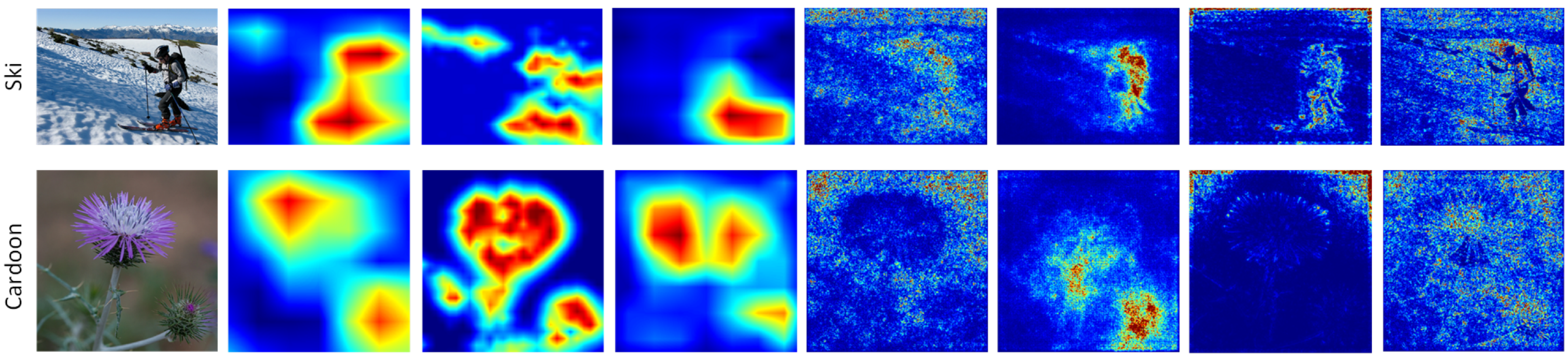
highlight

suppress

Accurate Interpretation (1/3)

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

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

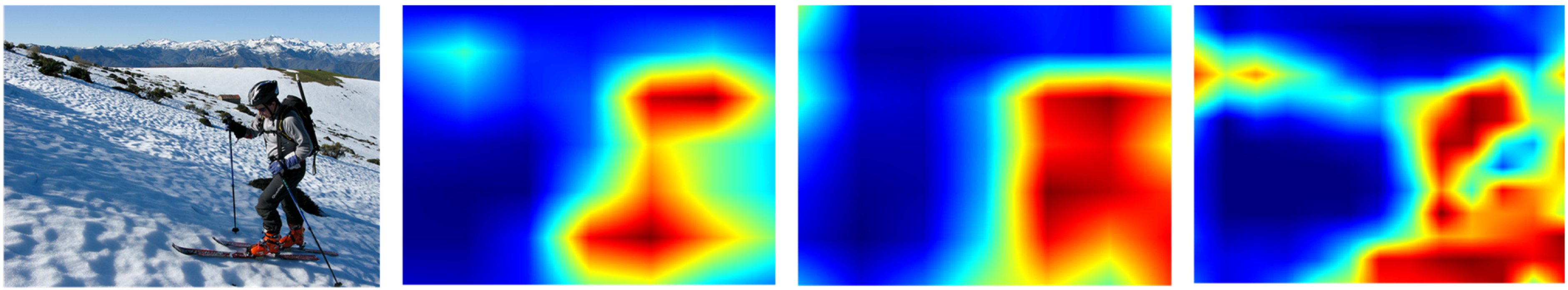


(a) Input (b) Ours (c) Mask (d) Grad-CAM (e) Grad (f) SmoothGrad (g) GuidedBP (h) Integrated

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

Accurate Interpretation (2/3)

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



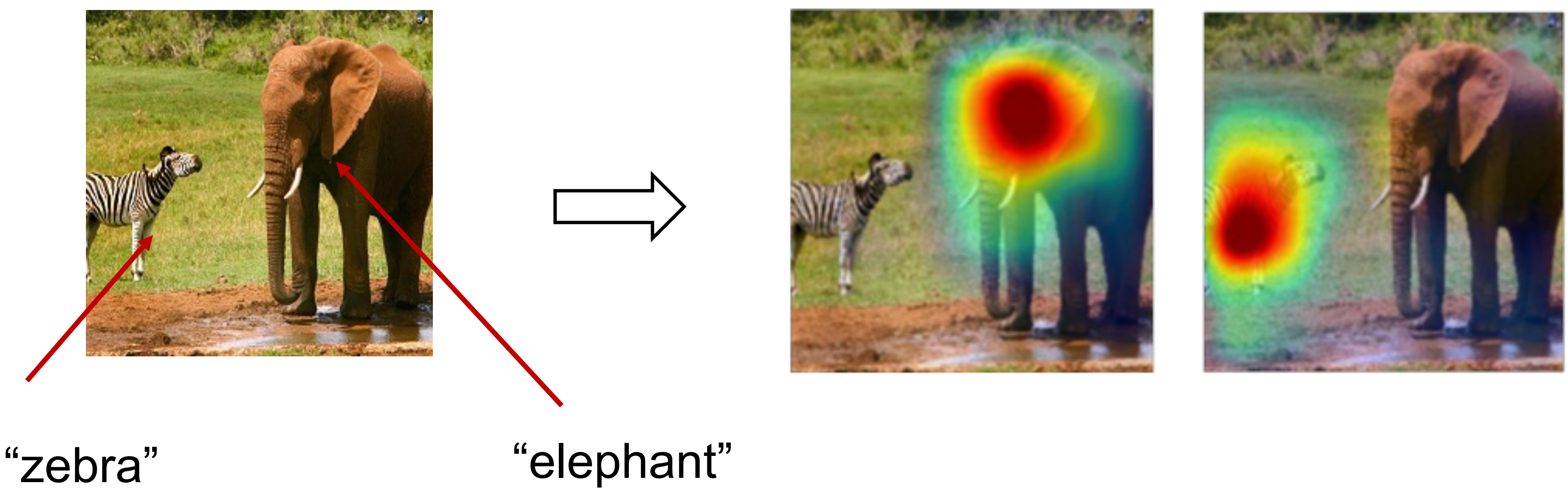
Input VGG-19 ResNet-18 AlexNet

Interpretations help ***capture the pros and cons***

of different network architectures.

Accurate Interpretation (3/3)

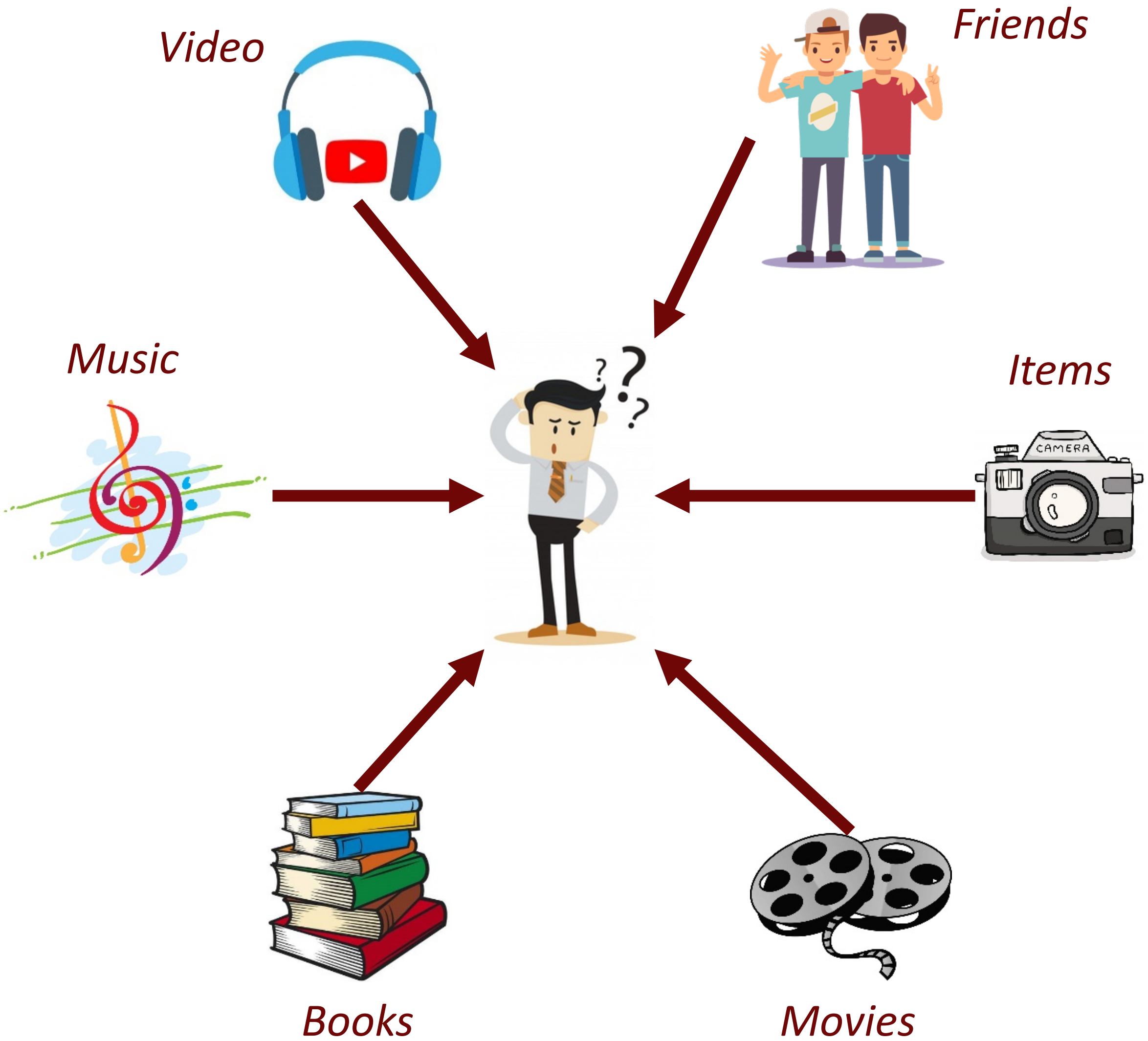
Visualization for input with multiple foreground objects



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*

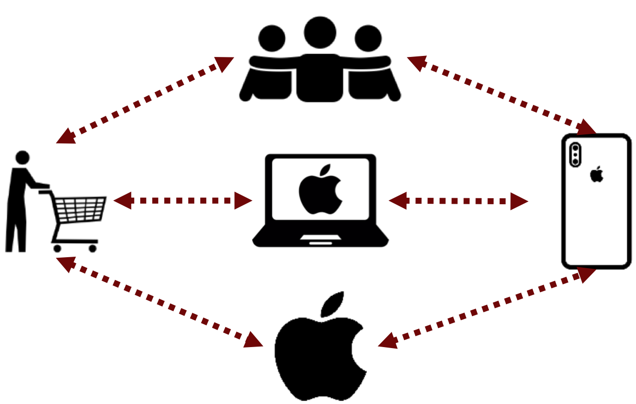
Challenges

1. The latent factors of users and items learned by recommender systems are simply the **uninterpreted vectors** to humans.

$$\begin{pmatrix}9.1\\1.2\\6.3\\\dots\\2.5\end{pmatrix}\quad\begin{pmatrix}1.1\\8.7\\4.9\\\dots\\4.5\end{pmatrix}\quad\begin{pmatrix}7.6\\7.5\\1.3\\\dots\\5.5\end{pmatrix}\longrightarrow$$

What does it mean?

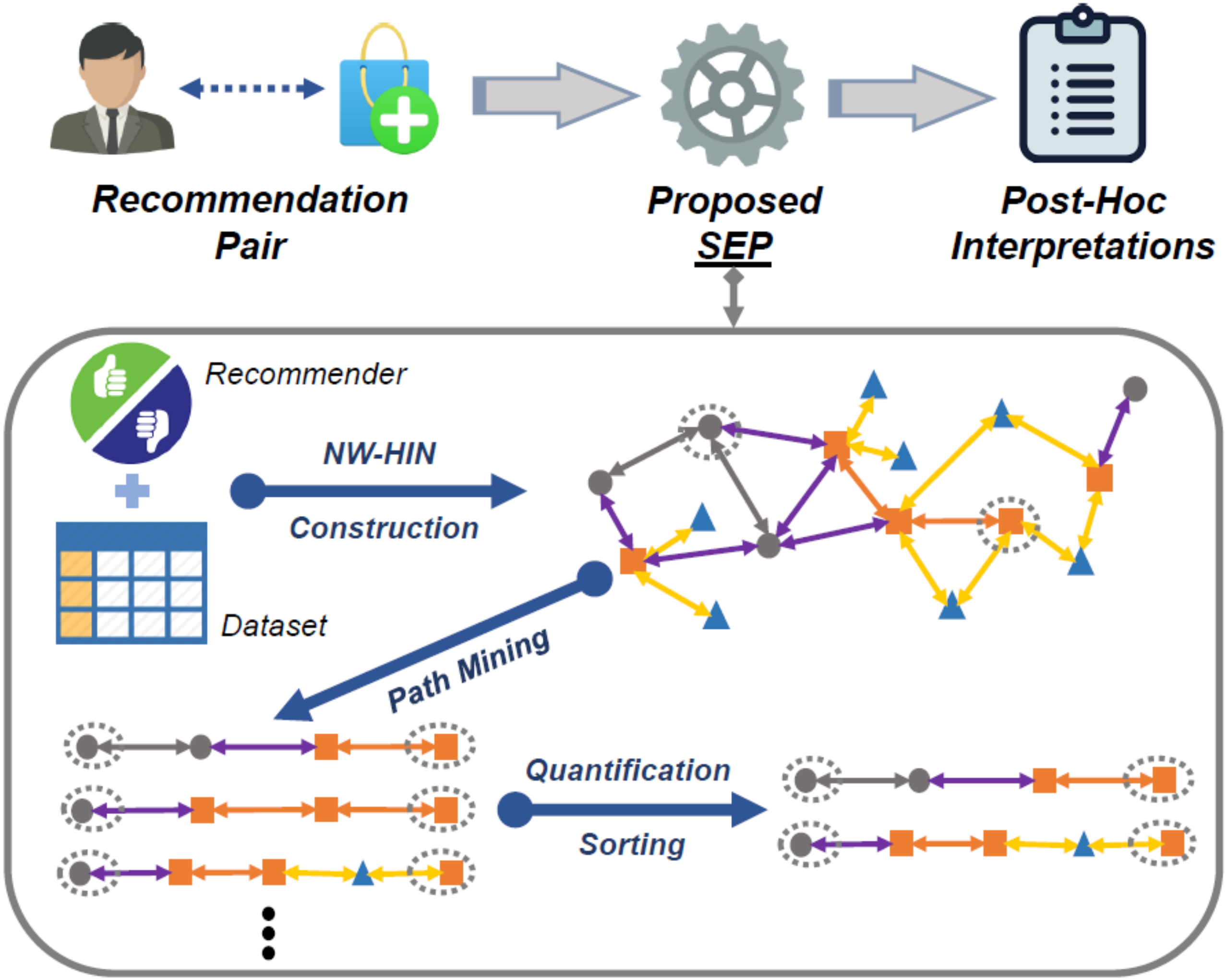
2. The possible interpretations for each recommendation can be rather diversified, and **appropriate selections** would be difficult.



→ ***Which interpretation would be influential?***

Proposed Framework

SEP → Sorted Explanation Path



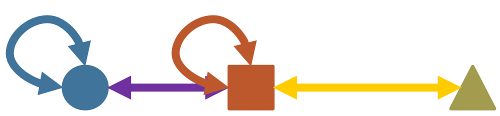
HIN Components

Our Constructed HIN Structure ---

Node Type: User Node   Item Node   Aspect Node

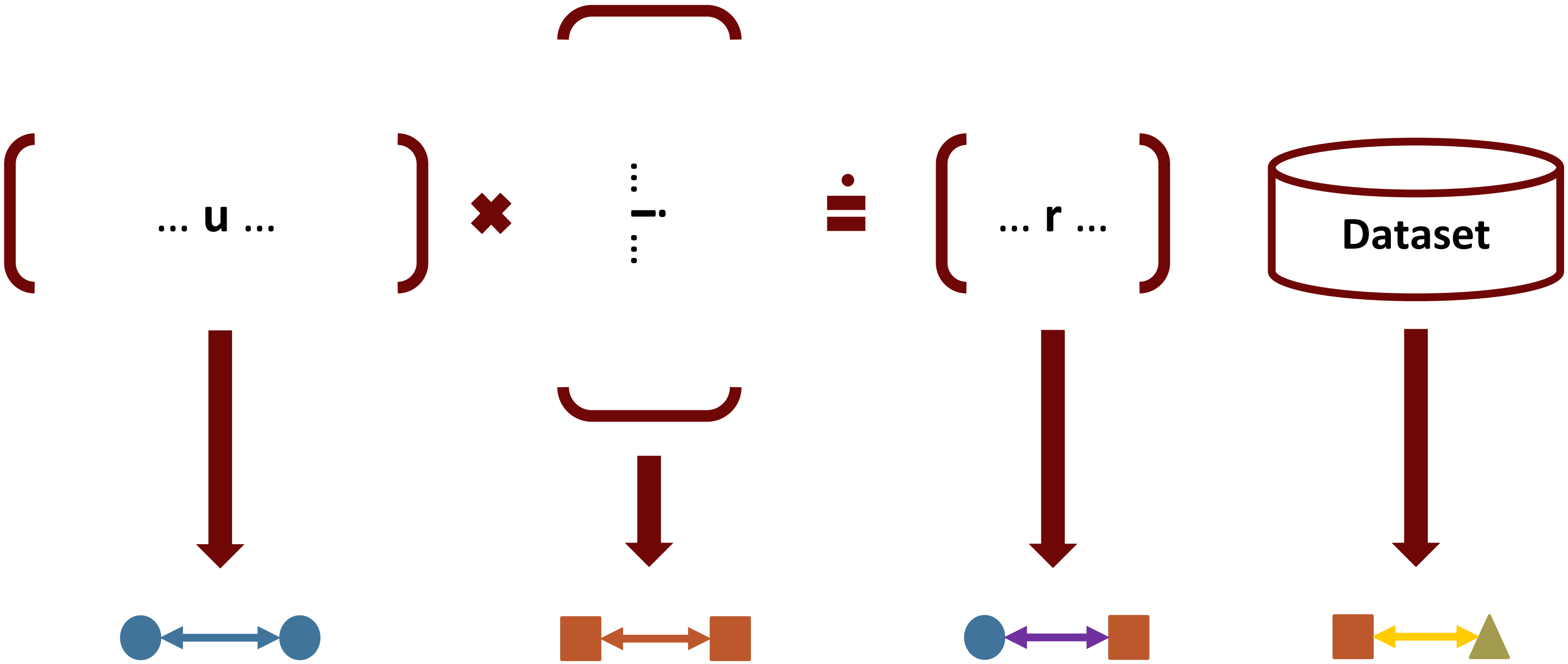
Link Type: User-User   Item-Item   User-Item   Item-Aspect

Network Schema:



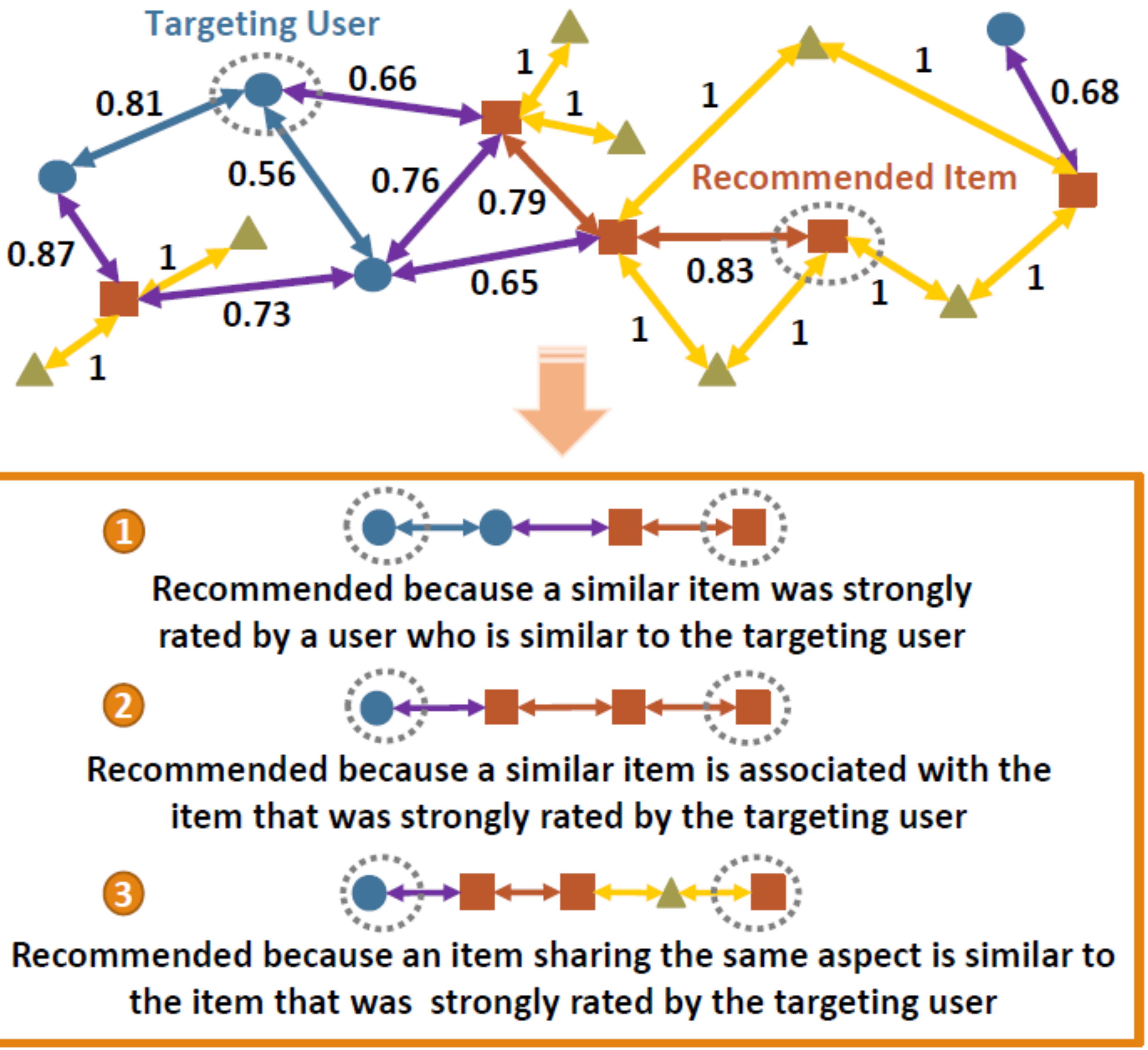
HIN Construction

Latent-Factor Recommender System ---



User-User Similarity	*Item-Item Similarity*	*User-Item Strength*	*Item-Aspect Relevance*
Pearson Coefficients	Pearson Coefficients	Unified Score	Binary Indicator

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***

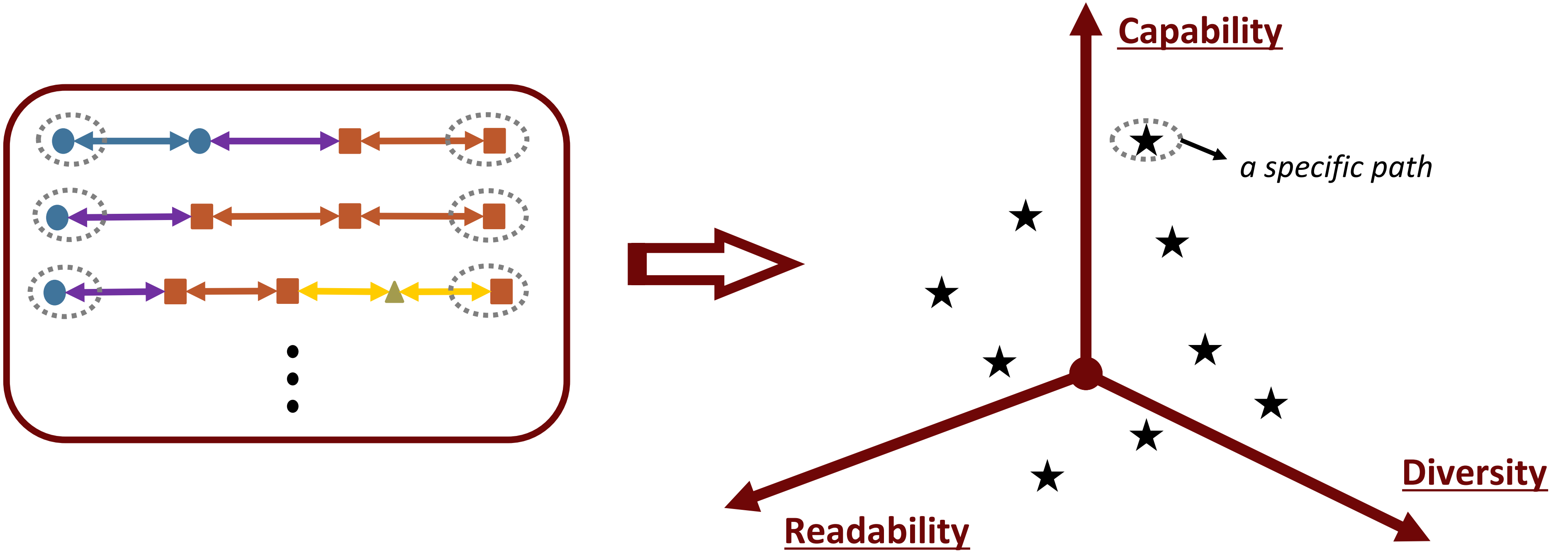
Path Quantification

For each explanation path k, we have →

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

Candidate Path Set

Candidate Ranking Space



Experimental Designs

We use the model built by Non-negative Matrix Factorization (NMF) as the targeting recommender systems

1. ***Applicability***

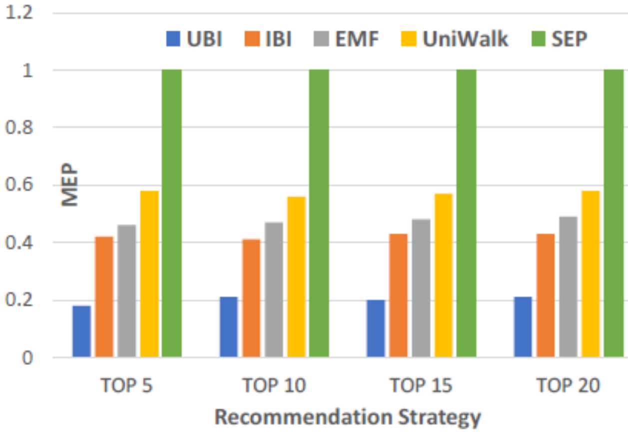
Mean Explainability Precision (MEP) & Mean Explainability Recall (MER)

$$\mathrm{MEP} = \sum_{u \in \mathcal{U}} \frac{|\mathcal{I}_u^{ir}|}{|\mathcal{I}_u^{r}|} \Big/ |\mathcal{U}| \,, \quad \mathrm{MER} = \sum_{u \in \mathcal{U}} \frac{|\mathcal{I}_u^{ir}|}{|\mathcal{I}_u^{i}|} \Big/ |\mathcal{U}|$$

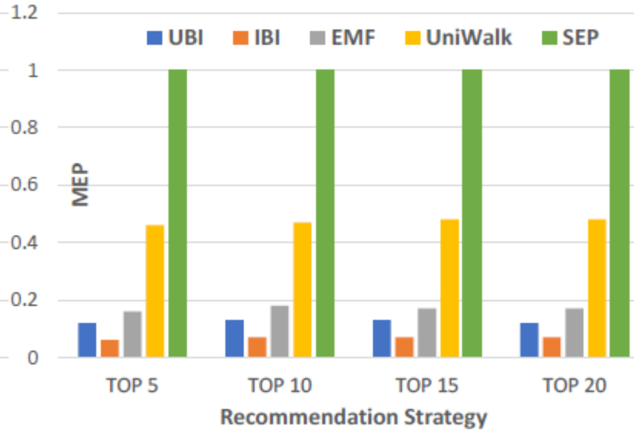
2. ***Effectiveness***

We knock out from training data the objects that appear in the interpretation results, and then retrain the whole system with the modified training data.

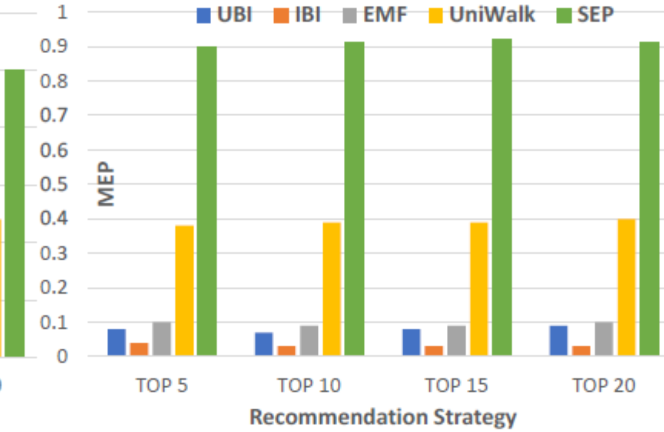
Applicability



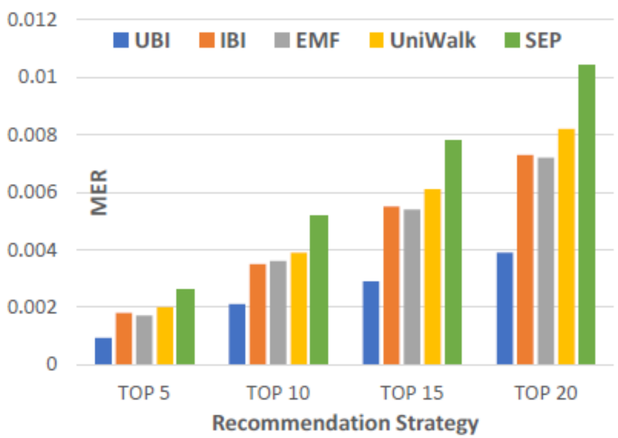
(a) MEP performance in ML-100K.



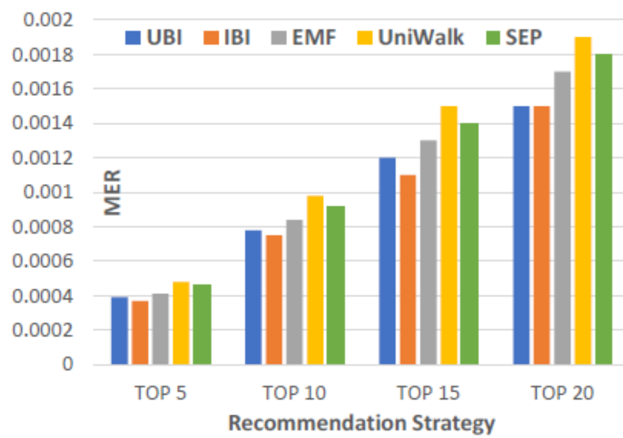
(b) MEP performance in ML-1M.



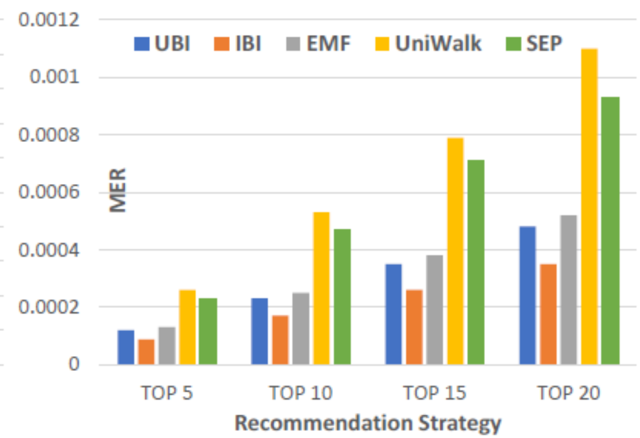
(c) MEP performance in GB-10K.



(d) MER performance in ML-100K.



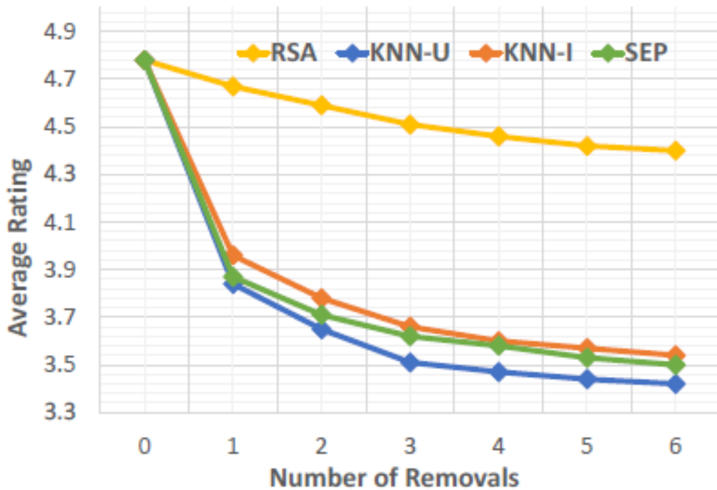
(e) MER performance in ML-1M.



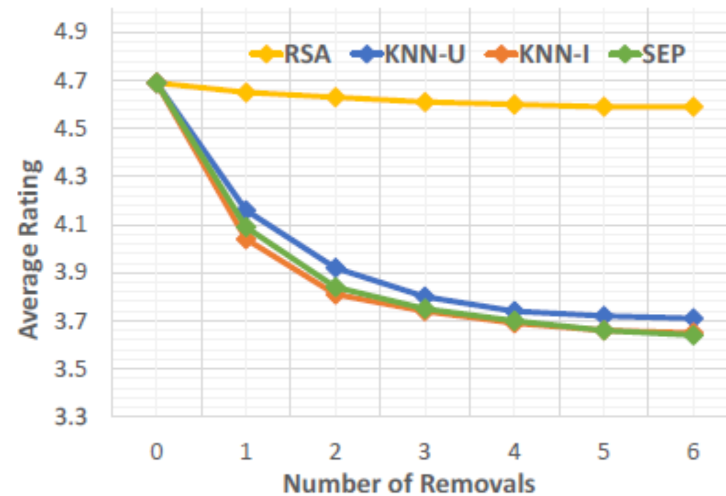
(f) MER performance in GB-10K.

The proposed SEP method is **superior in MEP performance**, and somewhat **competitive in MER performance**

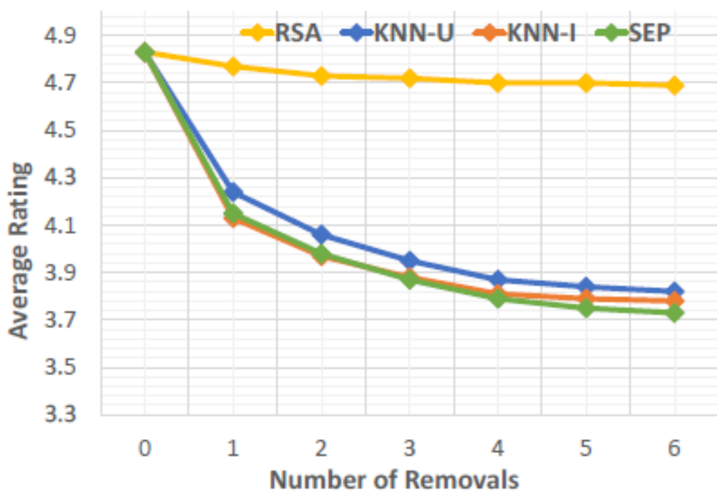
Effectiveness



(a) Average ratings in ML-100K.



(b) Average ratings in ML-1M.



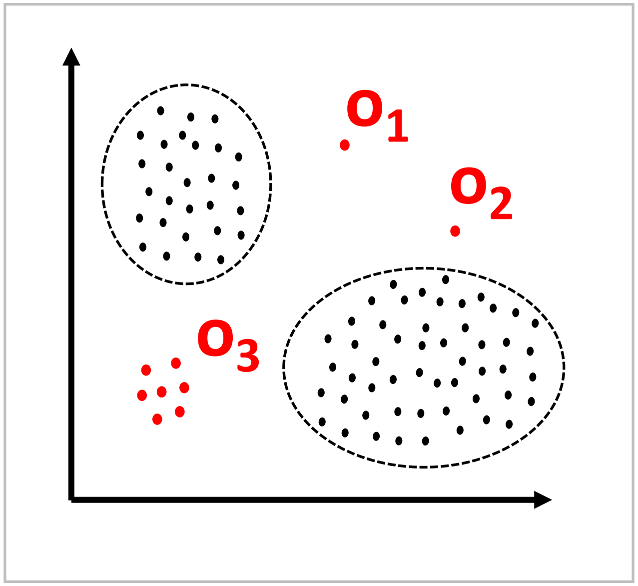
(c) Average ratings in GB-10K.

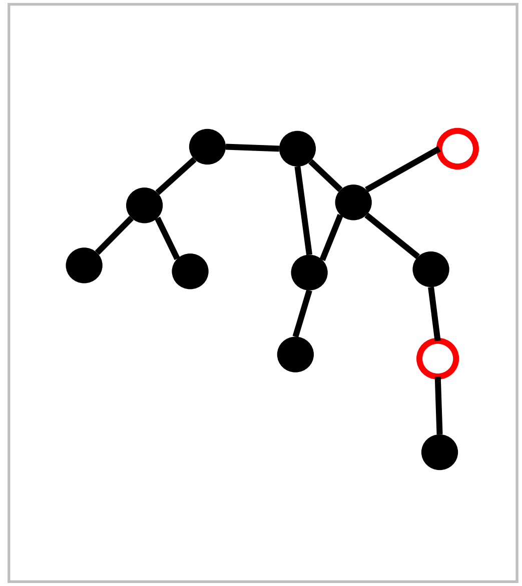
The interpretations generated from SEP method are **influential to the targeting recommender system**, which indicates the effectiveness of the proposed method.

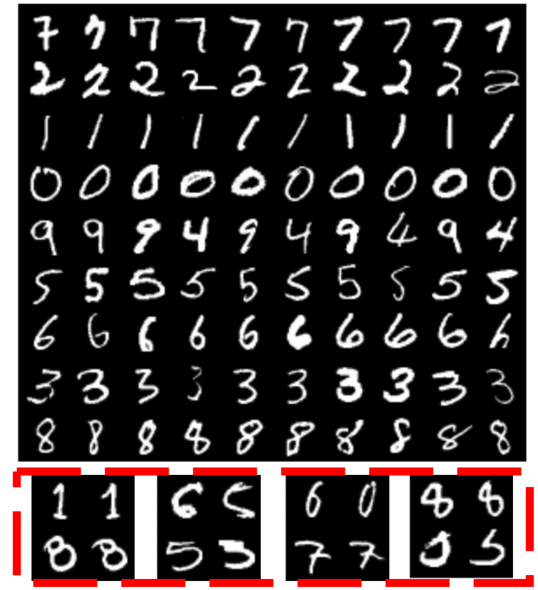
OUTLINE

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?



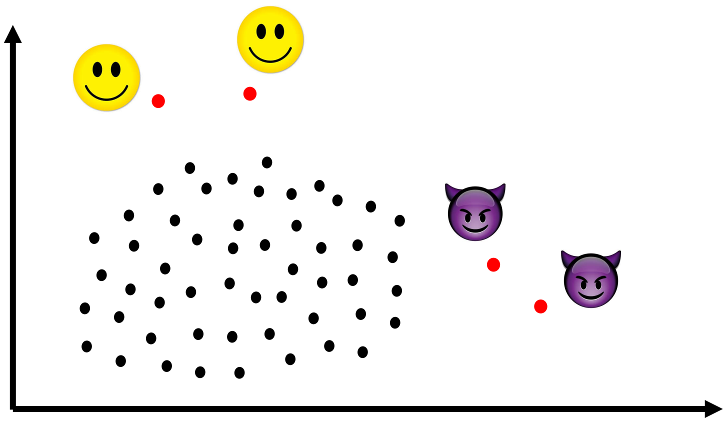




The noteworthy objects with patterns or behaviors that significantly deviate from the chosen background (or context)

Why is Interpretation Needed?

- Hard to tell whether the detected outliers are relevant to the application scenario
- Existing metrics such as ROC AUC and nDCG are unstable or limited in measuring the performance
- Outlier Detection for UnitedHealthcare



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

Given a dataset $\mathbf{X} = \{\mathbf{x}_\mathrm{n}\}$ 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;

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

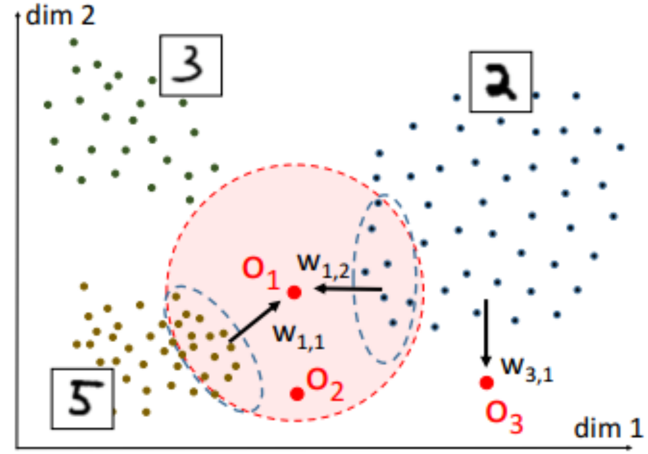
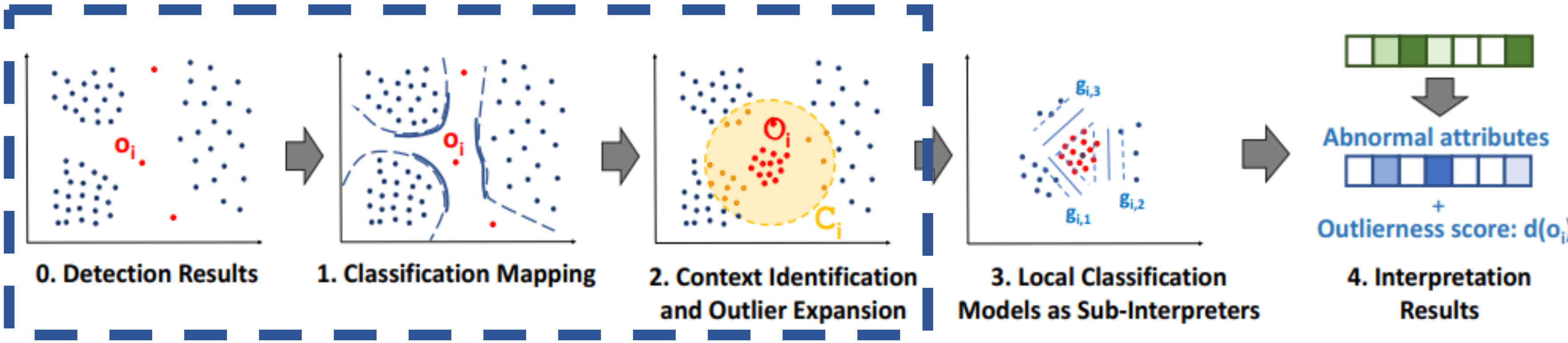


Figure 1: An toy example explaining why context clustering is needed

Proposed Framework



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

$$\min_{f} \mathcal{L}(h, f; \mathcal{O}, \mathcal{X} - \mathcal{O}) \Rightarrow \min_{f} \sum_{i} \mathcal{L}(h, f; \mathbf{o}_i, \mathcal{C}_i)$$

Decomposition

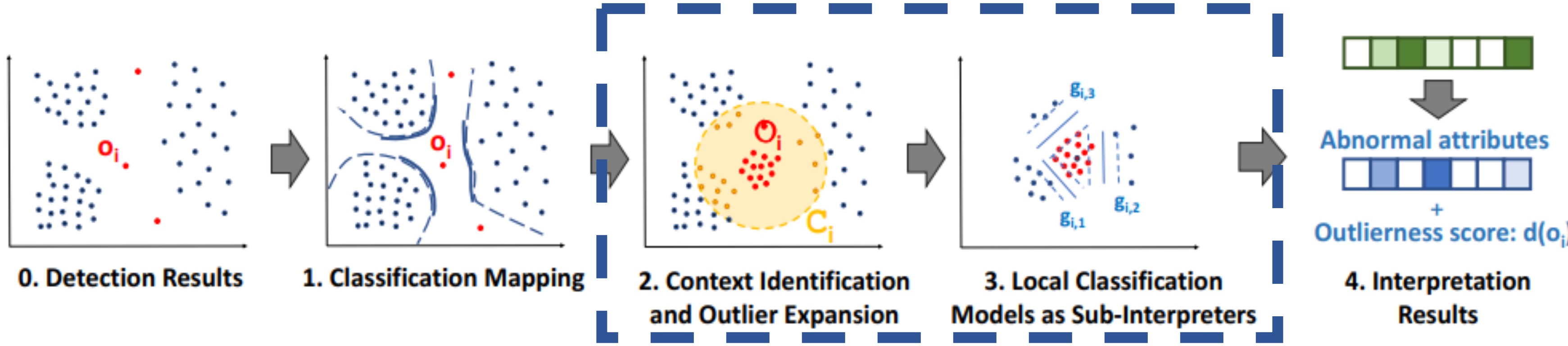
$$\Rightarrow \sum_{i} \min_{g_i} \mathcal{L}(h, g_i; \mathbf{o}_i, \mathcal{C}_i)$$

g_i: Local boundary

$$\Rightarrow \sum_{i} \min_{g_i} \mathcal{L}(h, g_i; \mathcal{O}_i, \mathcal{C}_i)$$

Synthetic sampling

Proposed Framework



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

$$
\begin{aligned}
\boxed{P^{err}(\mathcal{O}_i, \mathcal{C}_i)} &= P(\mathcal{O}_i)\int_{\mathcal{C}_i} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x} + P(\mathcal{C}_i)\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_i)d\mathbf{x} \\
&\approx \Big(\sum_{l\in[1,L]} P(\mathcal{O}_i)\int_{\mathcal{C}_{i,l}} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x}\Big) + \Big(\sum_{l\in[1,L]} P(\mathcal{C}_{i,l})\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_{i,l})d\mathbf{x}\Big) \\
&= \sum_{l\in[1,L]} \Big(P(\mathcal{O}_i)\int_{\mathcal{C}_{i,l}} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x} + P(\mathcal{C}_{i,l})\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_{i,l})d\mathbf{x}\Big) \\
&\approx \sum_{l\in[1,L]} \boxed{P^{err}(\mathcal{O}_{i,l}, \mathcal{C}_{i,l}).}
\end{aligned}
$$

Cluster-wise decomposition

$g_{i,l}$

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

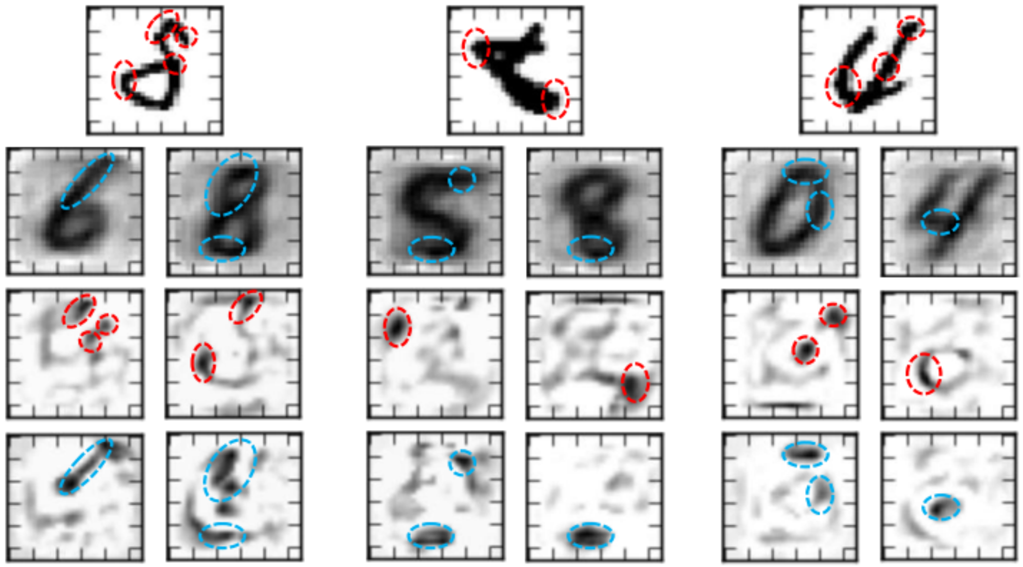
A Case Study

Query Outliers

Context Clusters
(Two for each query)

Positive Outlying
Regions

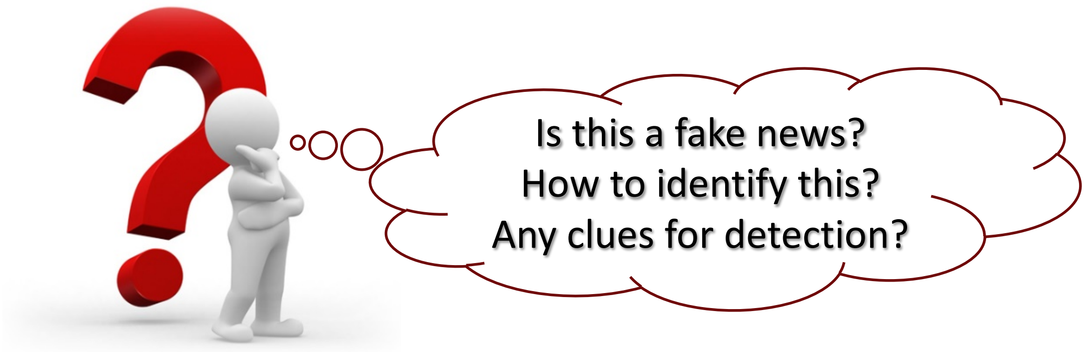
Negative Outlying
Regions



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



Challenges:

Beyond Text Classifications ---

❖ More challenging given heterogeneous types of information

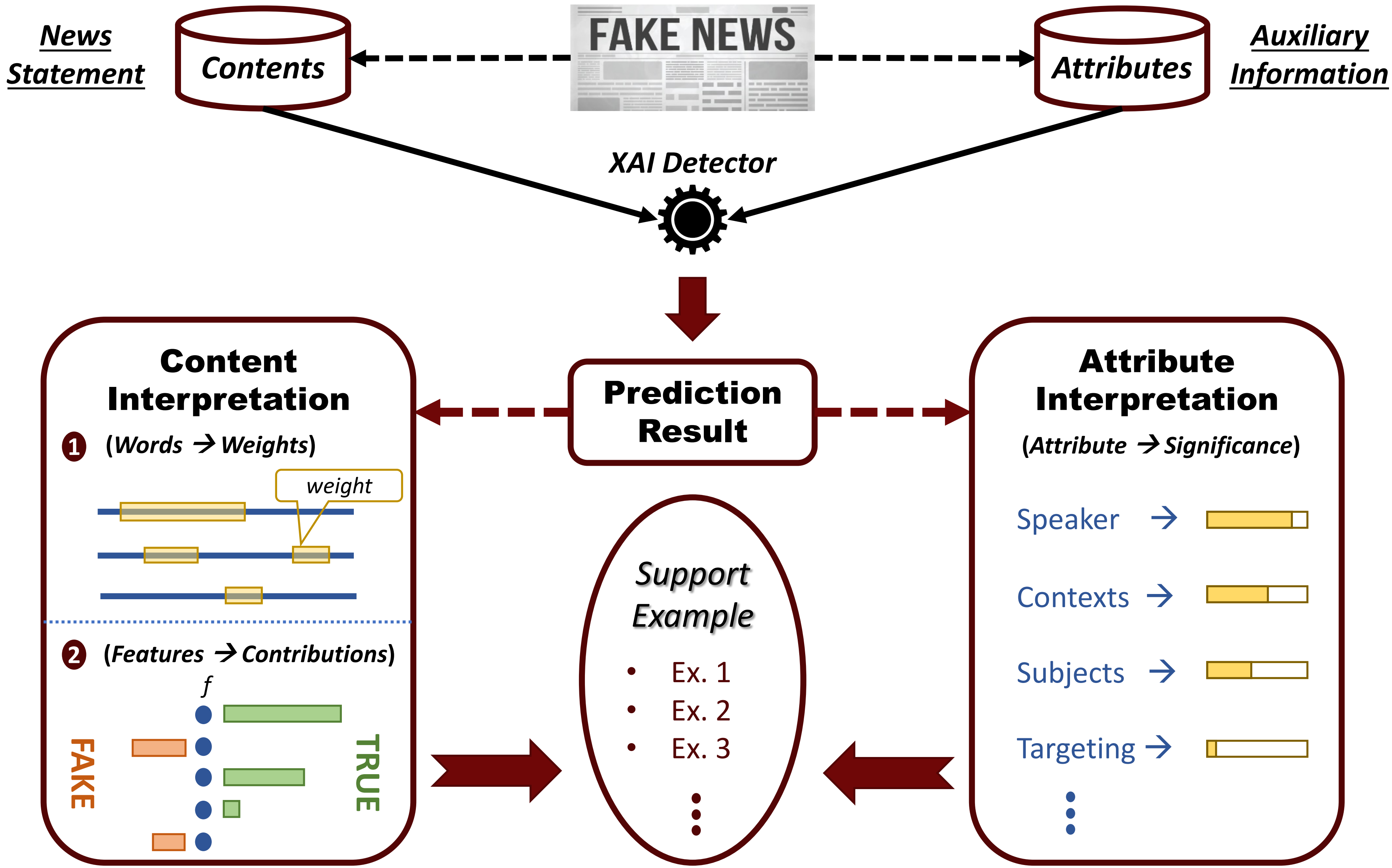
Hard to Achieve Effective Interpretations ---

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

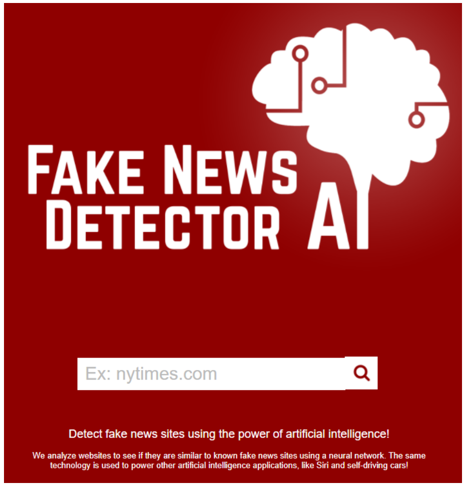
Beyond the News Itself ---

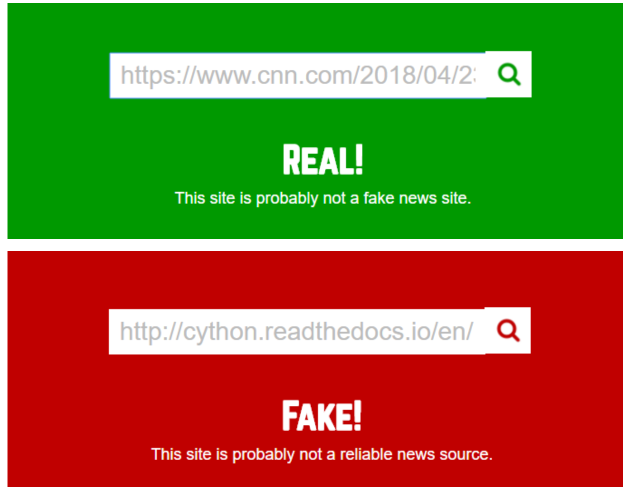
❖ Further supports are needed to convince people about the interpretations

Interpretable Fake News Detection



Interpretable Fake News Detection





Current AI Model on *Fake News Detection*

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

Prediction Labels Only

Labels

TRUE

FAKE

+

Interpretations

Key Components

Word Attributions

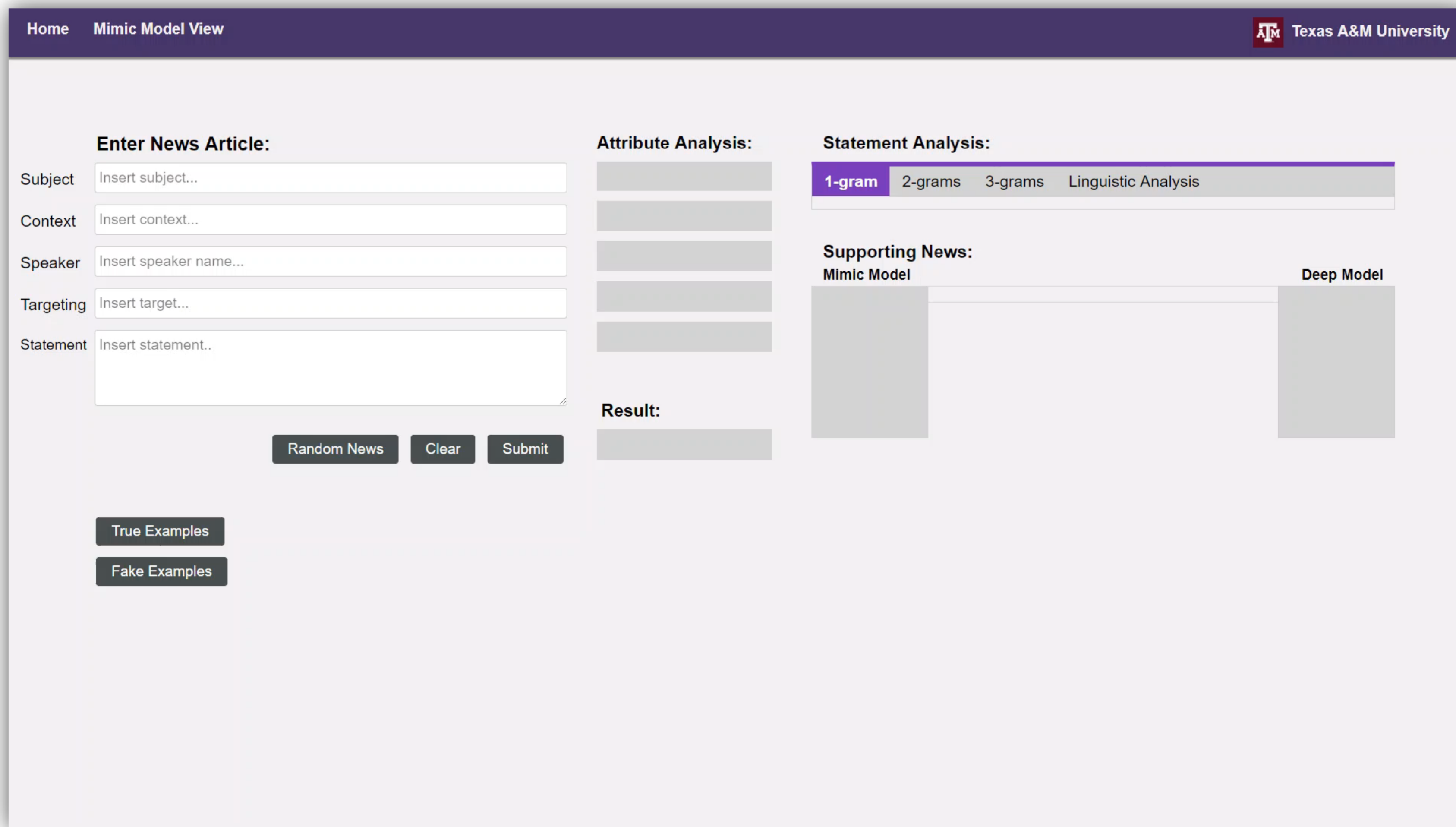
Linguistic Features

Proposed XAI Model on *Fake News Detection*

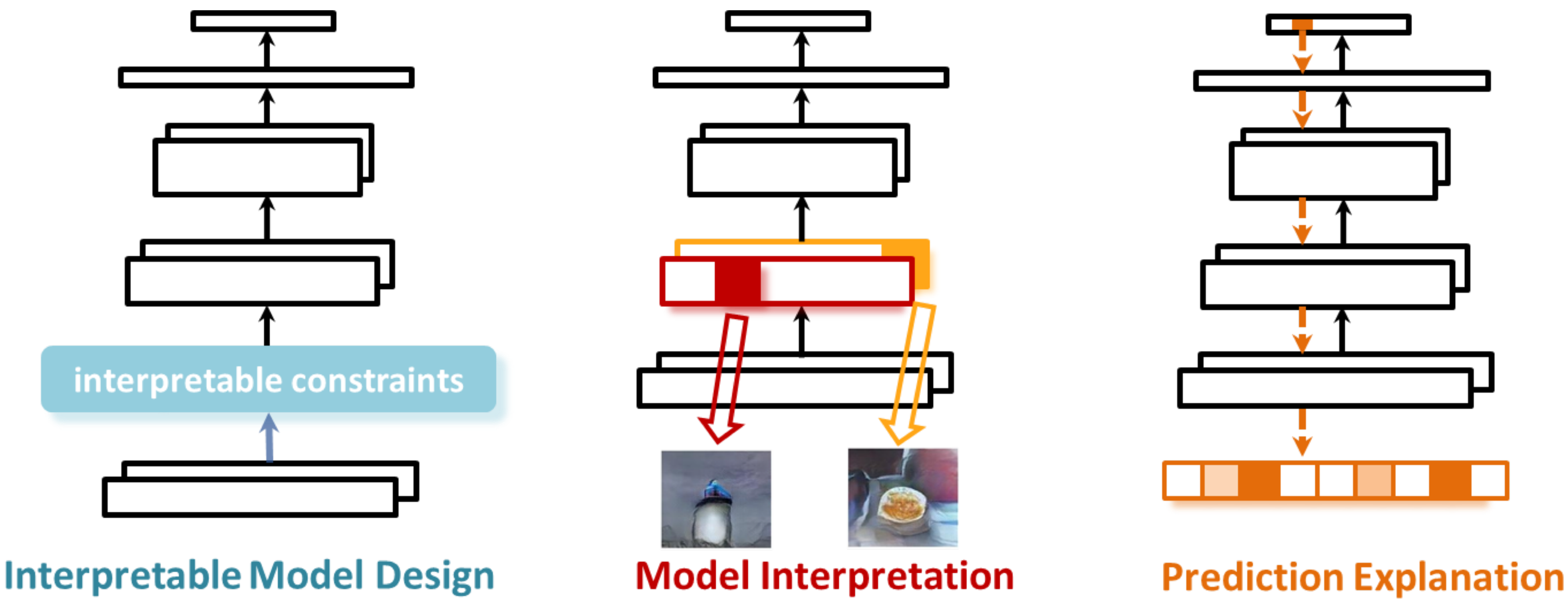
Prediction Labels
+
Interpretations
+
Meta-Interpretations

Demo for Interpretable Fake News Detection

Available at: http://csedatasrv.cs.tamu.edu:3001



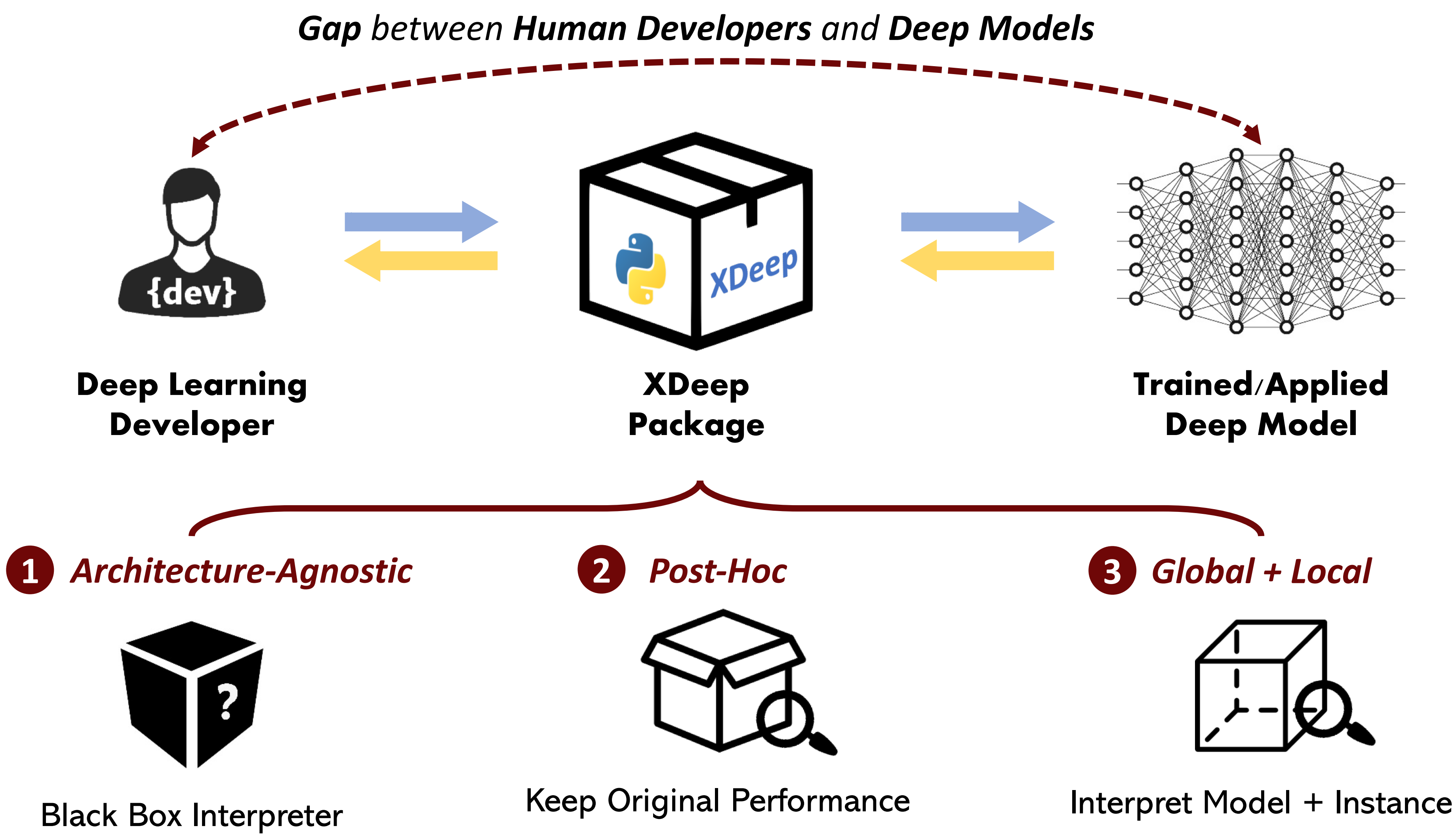
A Recent Survey



Mengnan Du, Ninghao Liu and Xia Hu. Techniques for Interpretable Machine Learning, CACM, 2020.

XDeep

--- A Python Package for Interpretable Deep Learning



A Long Way to Go

Tweet

Geoffrey Hinton
@geoffreyhinton

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

14:37 · 2/20/20 · Twitter Web App

204 Retweets **791** Likes

Yann LeCun
February 5 at 06:35 ·

A good example is how a wing causes lift. The computational fluid dynamics model, based on Navier-Stokes equations, works just fine. But there is no completely-accurate intuitive "explanation" of why airplanes fly.
Is it because of Bernoulli principle?
Because a wing deflects the air downwards?
Because the air above the wing want to keep going straight but by doing so creates a low-pressure region above the wing that forces the flow downwards sucks the wing upwards?
All of the above, but none of the above by itself.

Now, if there ever was a life-critical physical phenomenon, it is lift production by an airliner wing.
But we don't actually have a "causal" explanation for it, though we do have an accurate mathematical model and decades of experimental evidence.

You know what other life-critical phenomena we don't have good causal explanations for?
The mechanism of action of many drugs (if not most of them).
An example? How does lithium treat bipolar disorder?
We do have considerable empirical evidence provided by extensive clinical studies.

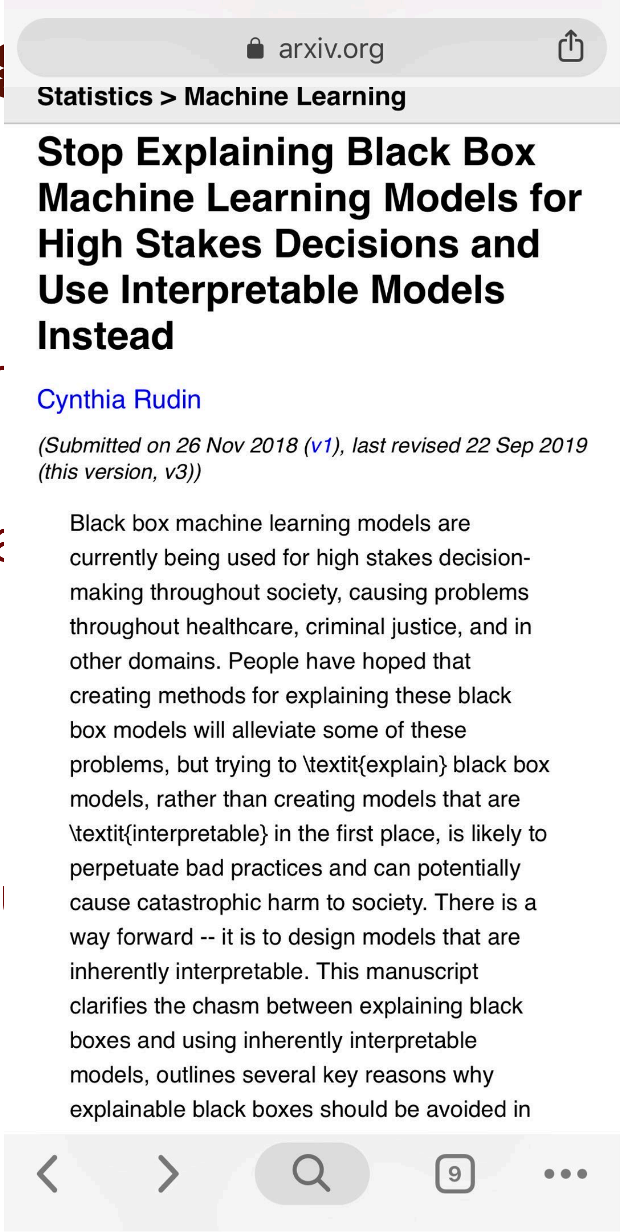
This is not to say that causality is not an important area of research for AI.
It is.

But sometimes, requiring explanability is counterproductive.

Write a comment...

Interpretable Machine Learnin

- Model-agnostic explanation
  - Broadly applicable to various machir
  - Treating a model as a black-box
  - Does not inspect internal model par
- Model-specific explanation
  - Specifically designed for each model
  - Usually require examining internal str


arxiv.org

Statistics > Machine Learning

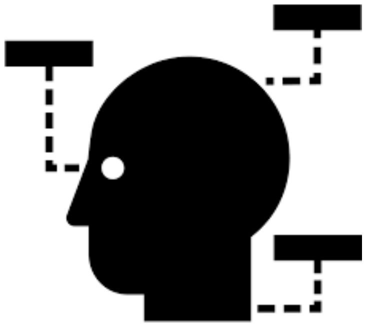
Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Cynthia Rudin

(Submitted on 26 Nov 2018 (v1), last revised 22 Sep 2019 (this version, v3))

Black box machine learning models are currently being used for high stakes decision-making throughout society, causing problems throughout healthcare, criminal justice, and in other domains. People have hoped that creating methods for explaining these black box models will alleviate some of these problems, but trying to \textit{explain} black box models, rather than creating models that are \textit{interpretable} in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society. There is a way forward -- it is to design models that are inherently interpretable. This manuscript clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in

Human-Centric Machine Learning



How to enable ***interpretable*** and ***Interactive*** machine learning?

Interpretable Machine Learning
(IML)

Provide explanations for human to ***easily understand*** the system

How to enable ***automated*** knowledge discovery and learning?

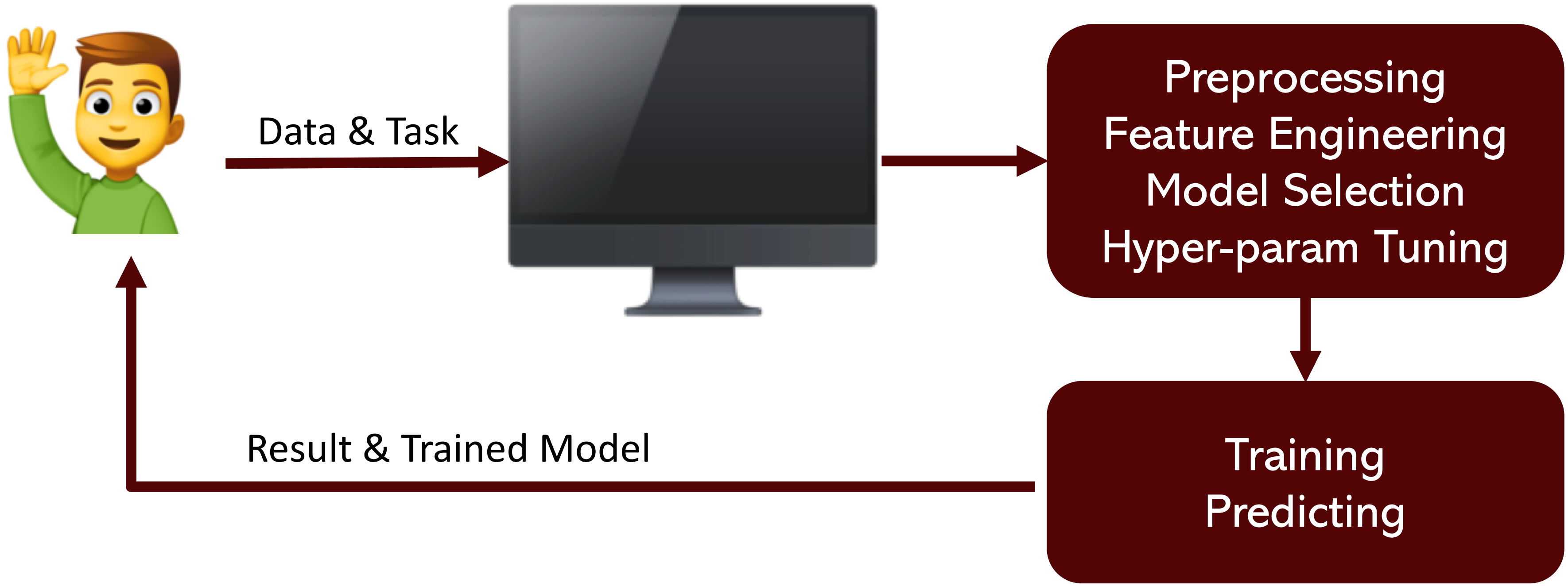
Automated Machine Learning
(AutoML)

Provide convenience for human to ***easily build*** the system

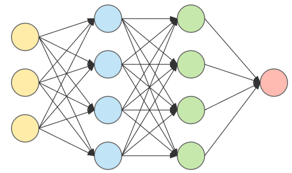
What is AutoML

Make machine learning an ***accessible tool*** ---

- to domain experts and data scientists
- by automating the ***end-to-end*** *process from data to the result*.

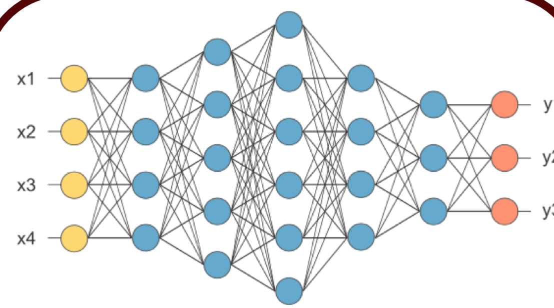


Automated Deep Learning



Learning rate: 0.1
Batch size: 128
Epoch: 20

Predicted Labels



Learning rate: 0.025
Batch size: 64
Epoch: 200

Predicted Labels

Given a dataset, find the ***best neural architecture and hyper-params***

and ***produce the prediction results***.

- We developed an AutoML System named ***Auto-Keras***;
- It provides ***easy-to-use solutions*** to deep-learning tasks;

Single GPU

Open-Source

```
import autokeras as ak

clf = ak.ImageClassifier()
clf.fit(x_train, y_train)
results = clf.predict(x_test)
```

Concise Interface

- Visit www.autokeras.com for more information.

AutoKeras: an Open-Source AutoML System

Machine Learning Platform Ecosystem

TensorFlow

Keras

KerasTuner

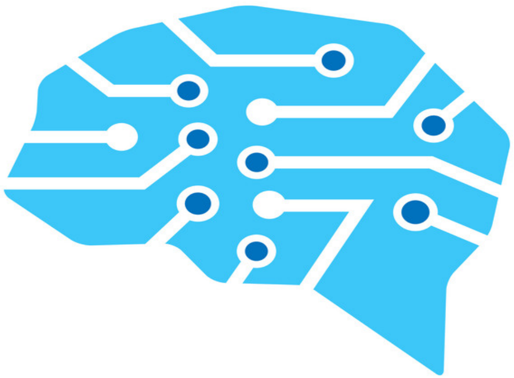
Auto-Keras

A Spectrum of Platform APIs

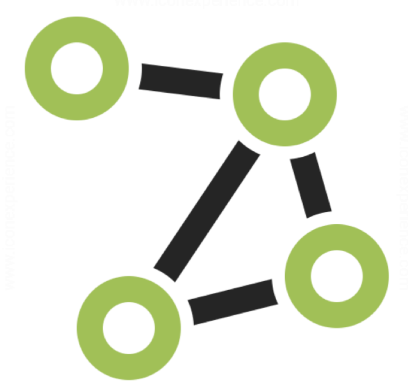
Configurable ⟷ **Simple**

Data Analytics at Texas A&M (DATA) Lab

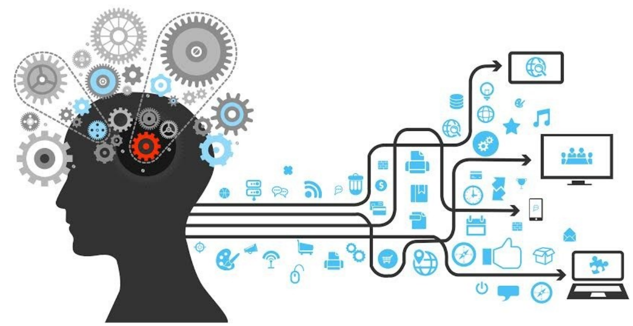
Interpretable Machine Learning



Automated Machine Learning



Network Analytics



Data Mining for Social Good

Acknowledgements

- DATA Lab and Collaborators

Data Analytics at Texas A&M (DATA Lab)

- Funding Agencies

  --- Defense Advanced Research Projects Agency (DAPRA)

  --- National Science Foundation (NSF)

  --- Industrial Sponsors (Adobe, Alibaba, Apple, JP Morgan, etc.)

- Everyone attending the talk!