Scrutinizing the Weakness and Strength of AI System

Wenbo Guo[1], Xinyu Xing[1], Jimmy Su[1]

1. JD Security Research Center

京东安全

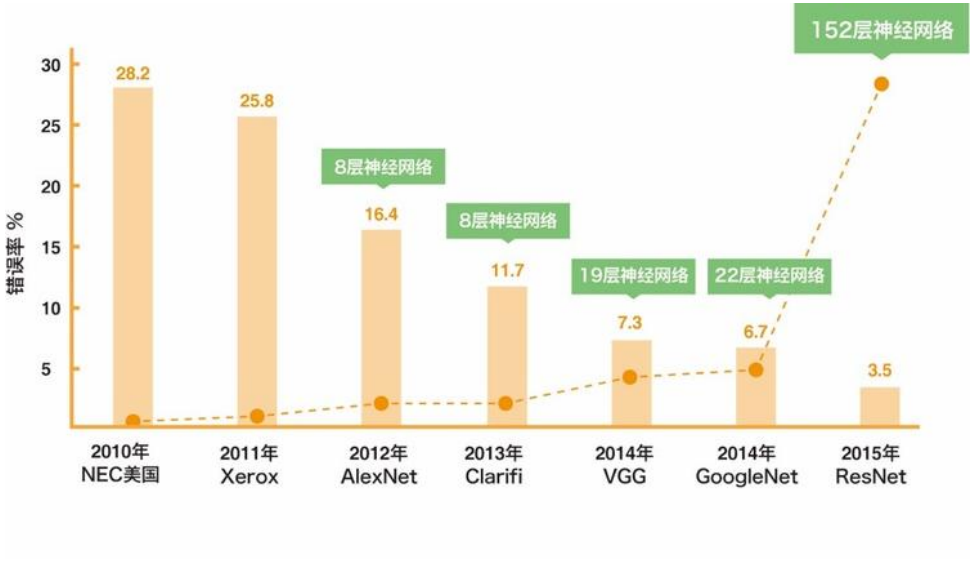
京东安全

Roadmap

- Background.
- What is model explanation.
- Existing explanation techniques.
- Proposed explanation techniques.
- Evaluation results.
- Summary.

AI System – Deep learning

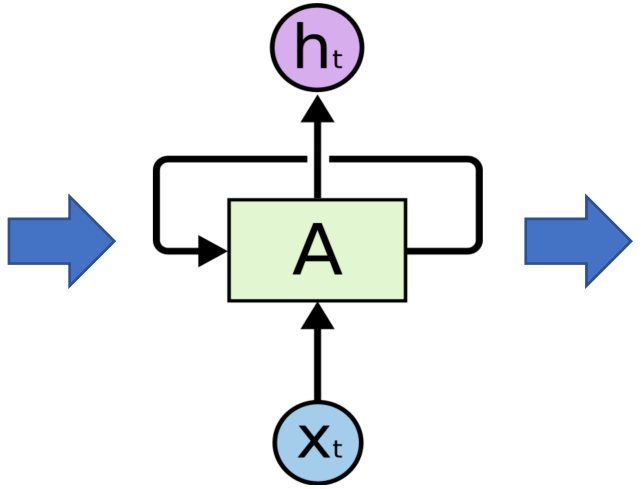
- Deep learning – High performance machine learning models.
  - Computer vision (CNN wins the ILSVRC contests).
  - Natural language processing (Seq-seq model for machine translation).
  - Alpha Go (Deep reinforcement learning).



Deep learning in security application

- Current applications.
  - Binary Analysis (USENIX 15, USENIX 17, CCS 17).
  - Malware Classification (KDD 17).
  - Network Intrusion Detection (WINCOM 16).

```
00000000004005a1 <mul_inv>:
  4005a1: push   %rbp
  4005a2: mov    %rsp,%rbp
  4005a5: mov    %edi,-0x24(%rbp)
  4005a8: mov    %esi,-0x28(%rbp)
  4005ab: mov    -0x28(%rbp),%eax
...
  400615: jns    40061d <mul_inv+0x7c>
  400617: mov    -0xc(%rbp),%eax
  40061a: add    %eax,-0x8(%rbp)
  40061d: mov    -0x8(%rbp),%eax
  400620: pop    %rbp
  400621: retq
```

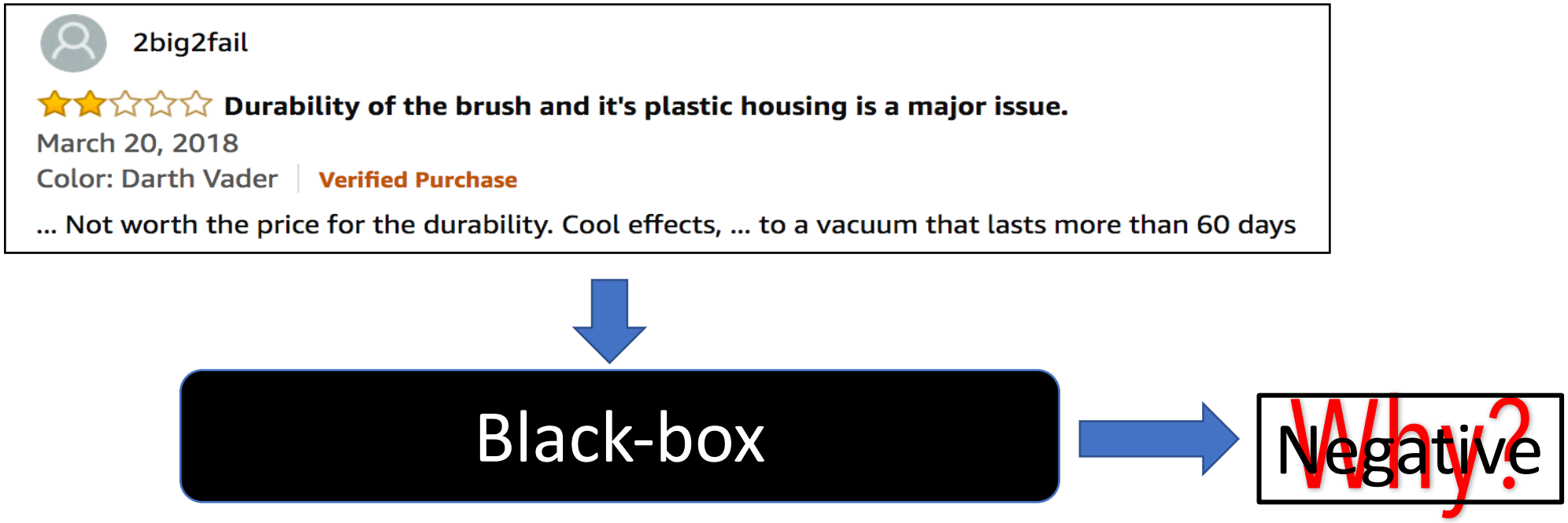


Function Start

```
00000000004005a1 <mul_inv>:
  4005a1: push   %rbp
  4005a2: mov    %rsp,%rbp
  4005a5: mov    %edi,-0x24(%rbp)
  4005a8: mov    %esi,-0x28(%rbp)
  4005ab: mov    -0x28(%rbp),%eax
...
  400615: jns    40061d <mul_inv+0x7c>
  400617: mov    -0xc(%rbp),%eax
  40061a: add    %eax,-0x8(%rbp)
  40061d: mov    -0x8(%rbp),%eax
  400620: pop    %rbp
  400621: retq
```

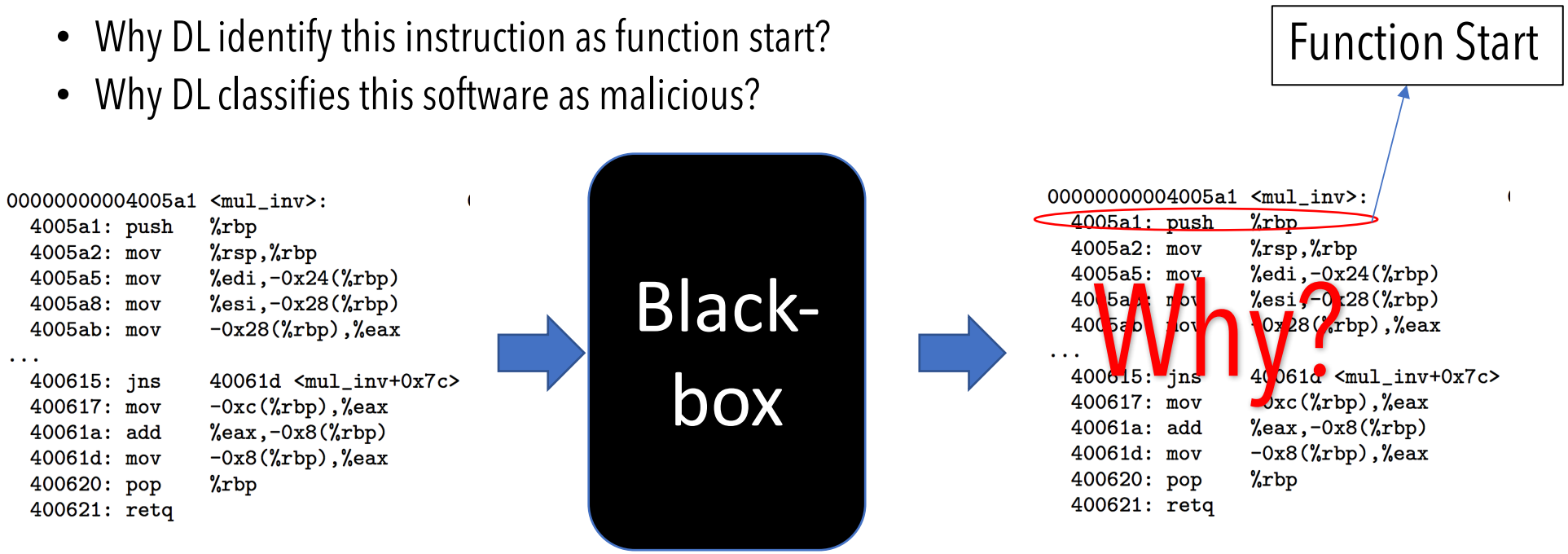
Why not Deep Learning

- Lack of transparency of deep neural network.
  - Contains hundreds of thousands of neurons.
  - High classification accuracy but low interpretability.



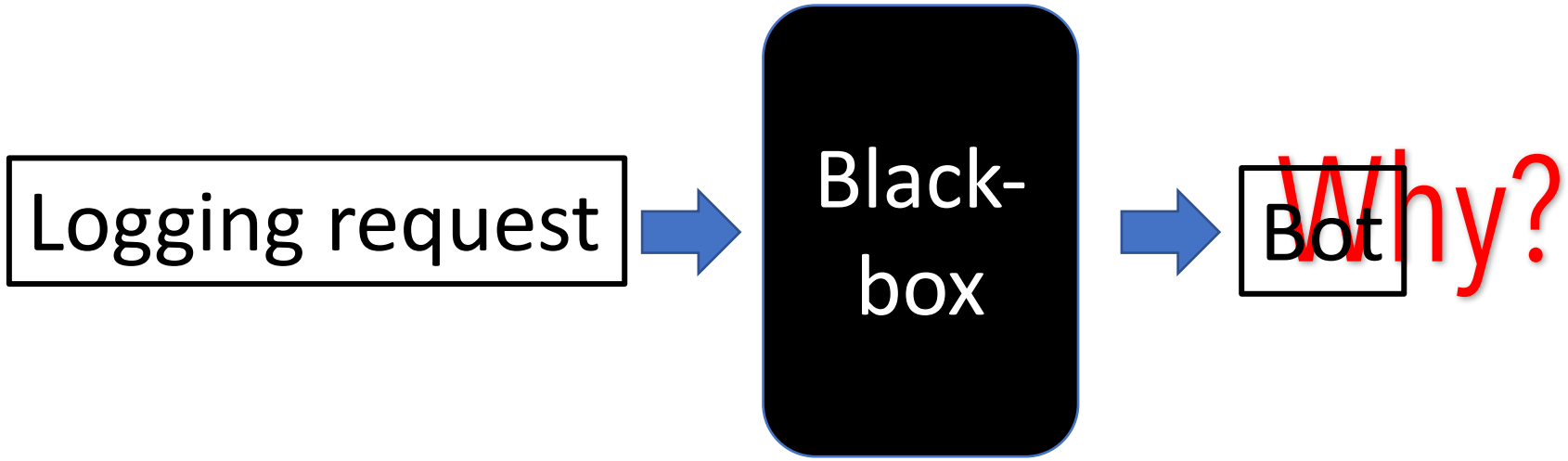
Why not Deep Learning

- Lack of transparency of deep neural network.
  - User cannot understand the behavior of the model.
    - Why DL identify this instruction as function start?
    - Why DL classifies this software as malicious?

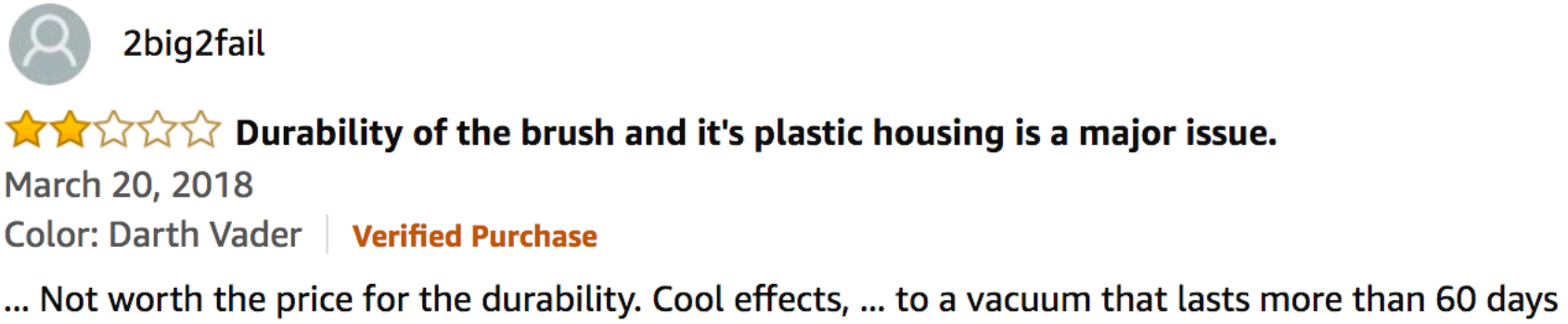


Why not Deep Learning

- Lack of transparency of deep neural network.
  - User cannot build trust of the model.
  - E.g., AI bot detection system.
    - Cannot reject a logging request without reason.



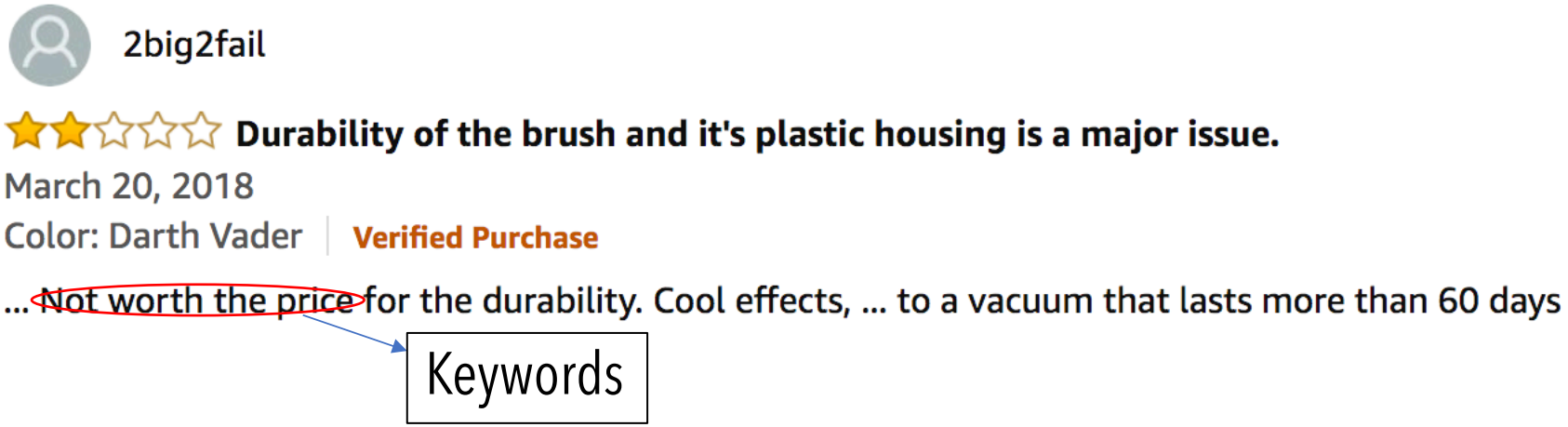
Open up the black-box



- Human: this is a negative comment, because I saw the words: "**not worth the price**".
- DL: this is a negative comment, **but no reason**.
- Explaining DL model: reasoning the decision making process of a DL model.

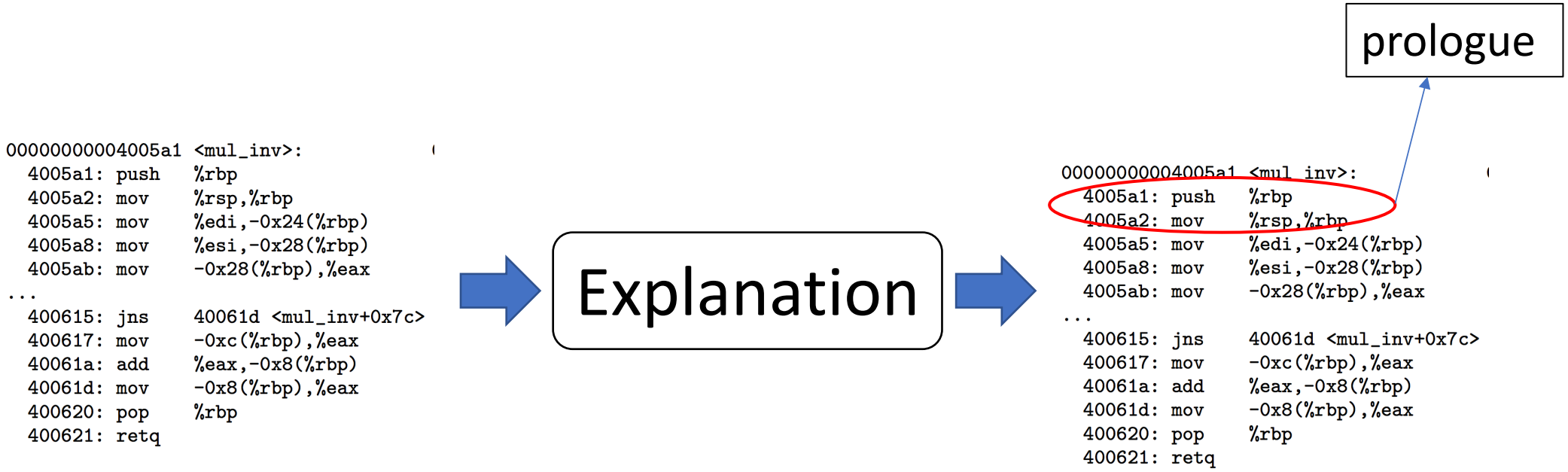
Interpreting Deep Learning models

- What is an explanation?
  - Given a testing sample, identifying a set of important features make key contributions to classification results.
    - Image recognition: a group of important pixels.
    - Sentiment analysis: Key words.



Interpreting Deep Learning models

- For security applications.
  - Which parts of the program make DL identify this instruction as a functions start?



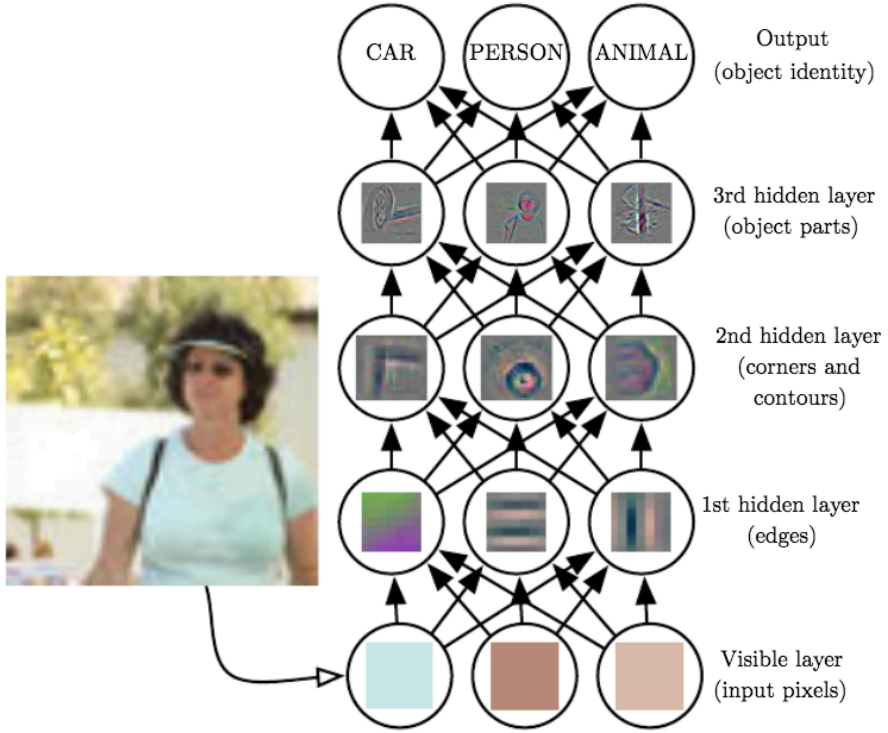
Interpreting Deep Learning models

- Challenges.
  - Complex network architectures: contains hundreds of thousands of neurons.
  - Varied network structures: so many variances of the basic network architectures.

- Existing techniques.
  - White-box explanation.
  - Black-box explanation.

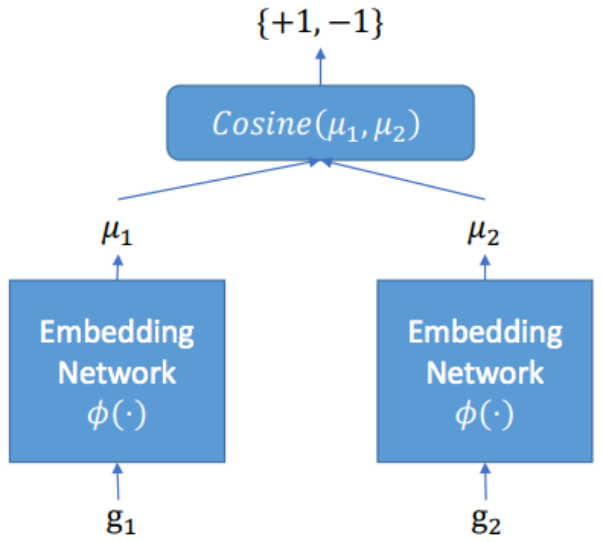
White-box Explanation

- High level idea.
  - Dissect the neural networks and find out how information propagate in the networks.



Why not white-box in security?

- Current white-box techniques are mainly designed for basic network architecture.
  - More and more complex architectures has been adopted in computer security.
    - Applying a hybrid network to detect the vulnerability of binary codes (CCS 17).
- Plenty of variances of basic network architectures.
  - Recurrent network: Simply RNN, GRU, LSTM.

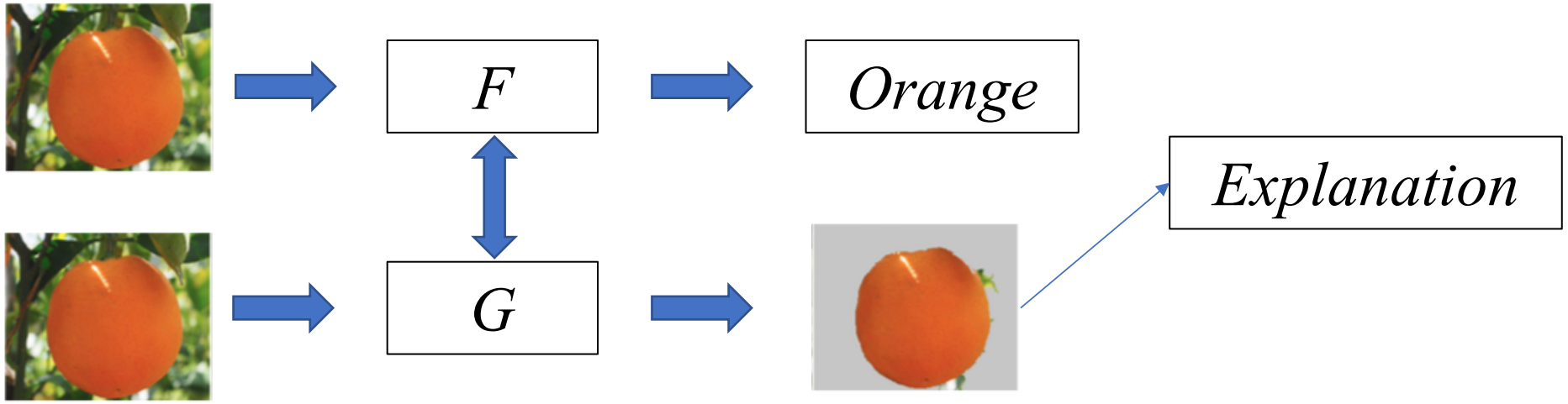


Why not white-box in security?

- The network structures are not available.
  - VirtusTotal
  - Dyninst
- The hidden layer representations cannot be understand.
  - Different from images, the hidden representations of binary code can not be interpreted.

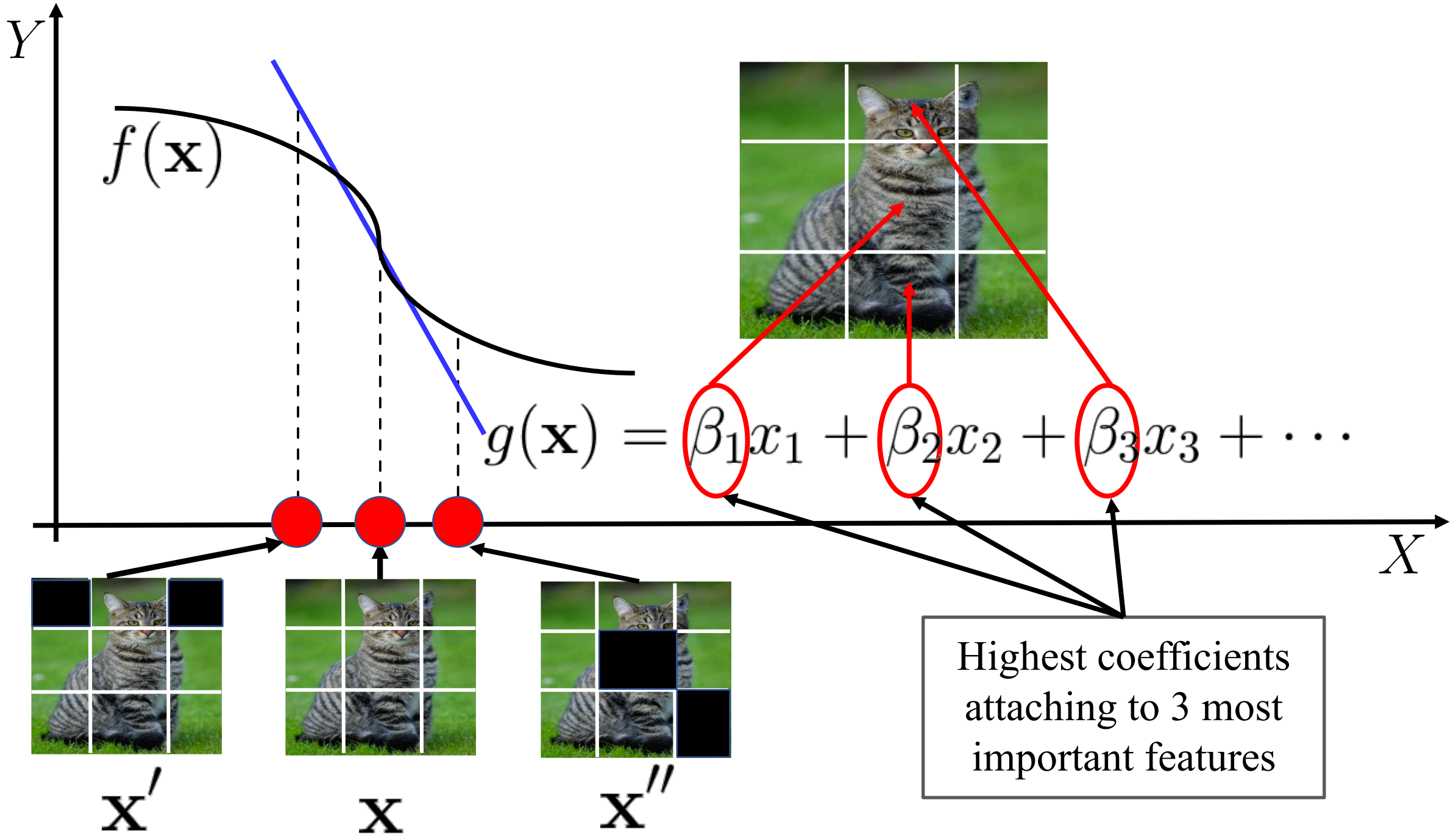
Black-box Explanation

- Treat deep learning methods as a function *F*.
- Approximate *F* with simple interpretable models *G*.
- Draw important features from *G* as the explanation for *F*.



Black-box Explanation

- How to generate an explanation.
  - Approximate the deep learning model with Linear regression: LIME (KDD 16), SHAP (NIPS 17).
  - Inspect the regression coefficients.
  - Pinpoint the features that corresponding to the highest coefficients.

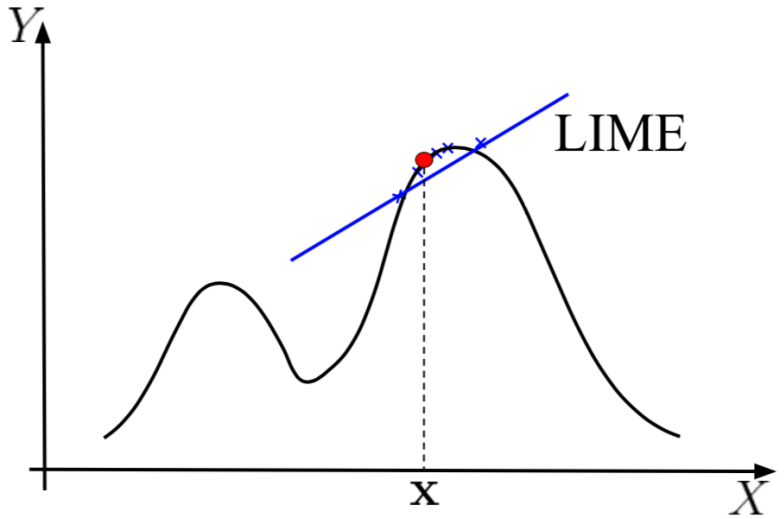

Y
f(x)
g(x) = β1x1 + β2x2 + β3x3 + ···
X
x′
x
x″
Highest coefficients attaching to 3 most important features

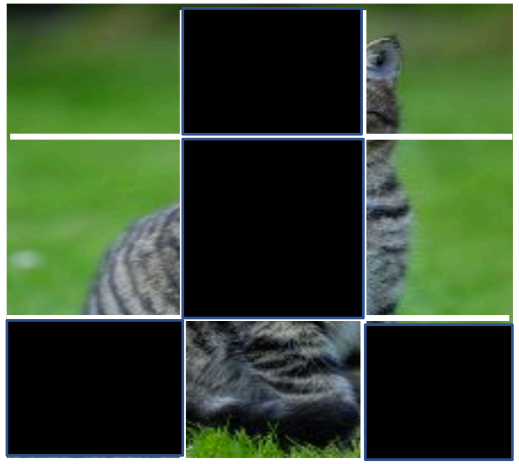
Challenges.

- How to generate a **precise** explanation.
  - For security applications, precise explanation is important.
    - Pinpoint the wrong features will lead to serious problems.
    - E.g., defense against AI black market system.
  - Precise approximation results in correct explanation.

Challenges.

- How to obtain a nearly perfect approximation.
  - Deep Learning model is highly non-linear.
  - Simple **linear approximation** is not a good choice (e.g., LIME).





Challenges.

- How to select the most important features.
  - For high dimensional data, simply ranking the regression coefficient is not enough.
  - Filter out the unimportant features while fitting the approximation model.

Our technique: high level idea.

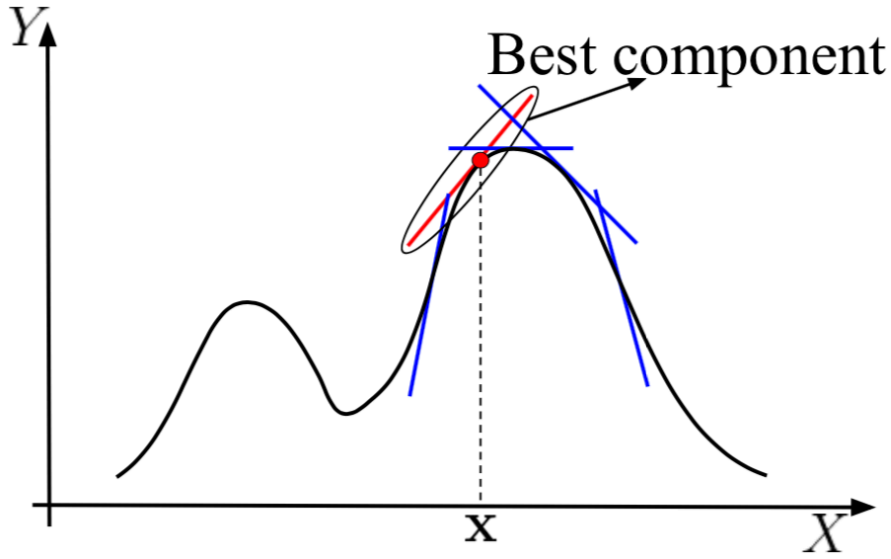
- Dirichlet process mixture regression model with multiple elastic nets.
  - **Precise approximation**.
    - mixture regression model: approximate arbitrary decision boundary.
    - elastic net: enable mixture model to deal with high dimensional and highly correlated data.
  - **Correct features**.
    - elastic net: Only select the most important features.
  - Multiple elastic net.
    - enable mixture model to accommodate different types of data correlation.

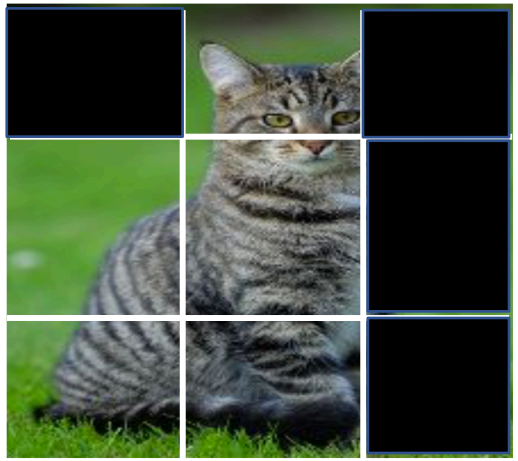
Our technique: how to derive an explanation.

- Given a well trained DL model and a set of data samples.
- Use the data samples and the corresponding model outputs to fit a approximation model *G*.
- Leverage MCMC to inference the parameters in the approximation model *G*.

Our technique: how to derive an explanation.

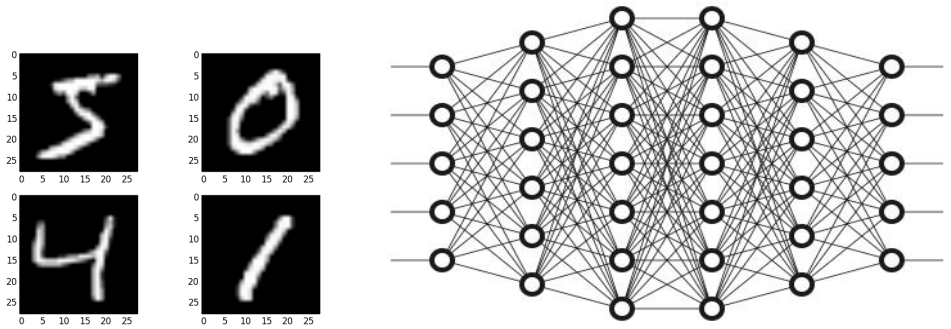
- Find the mixture component that the data sample lies in.
- Collect the top important features according to the regression coefficients of that component.



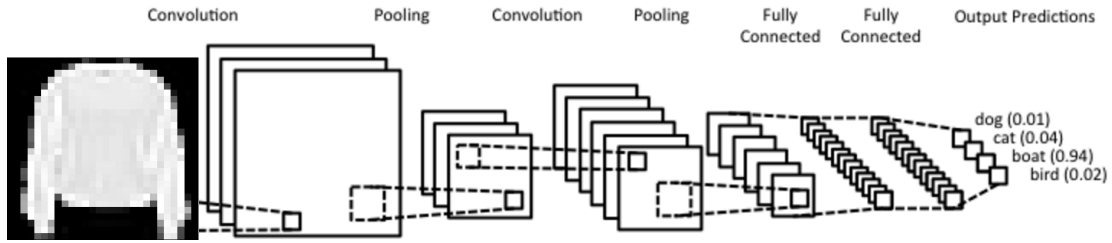


Evaluation: Image recognition

- MNIST: hand written digitals (Multilayer perceptron).
- Fashion-MNIST: Fashion products (Convolutional neural networks).



MLP on MNIST



CNN on Fashion-MNIST

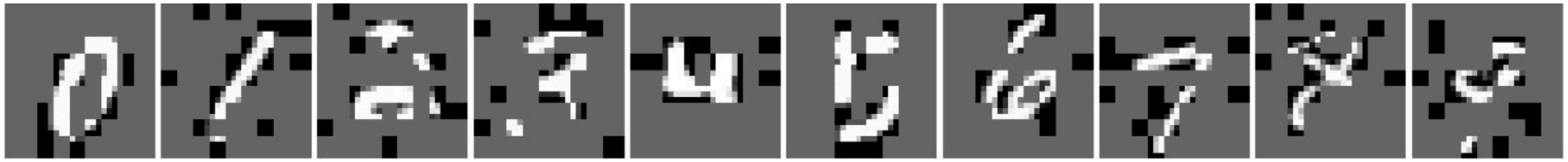
Explanation results: MNIST

Original images:

Our technique:

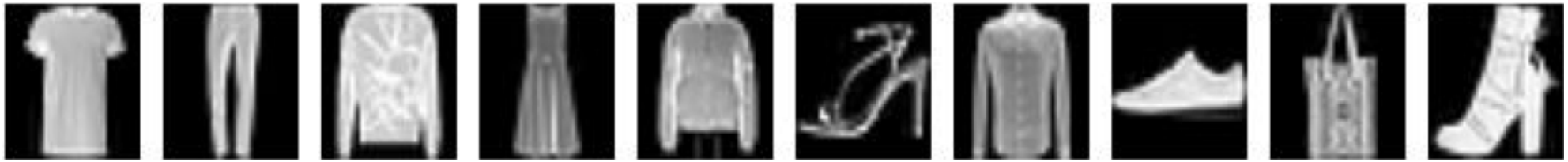
LIME (KDD 16):

SHAP (NIPS 17):

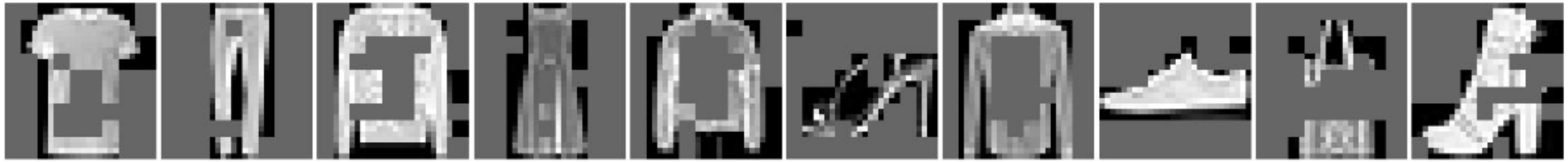


Explanation results: Fashion-MNIST

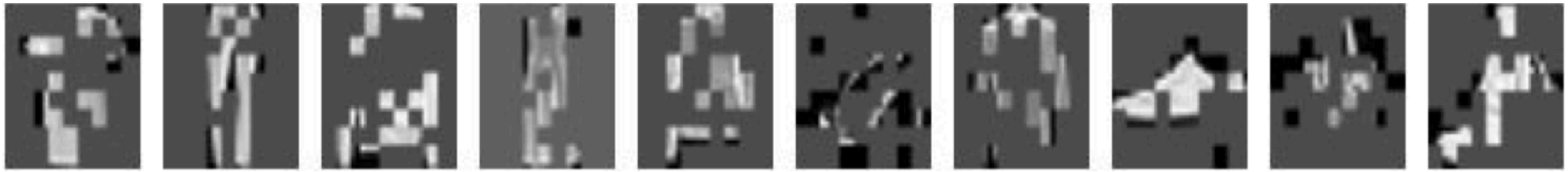
Original images:



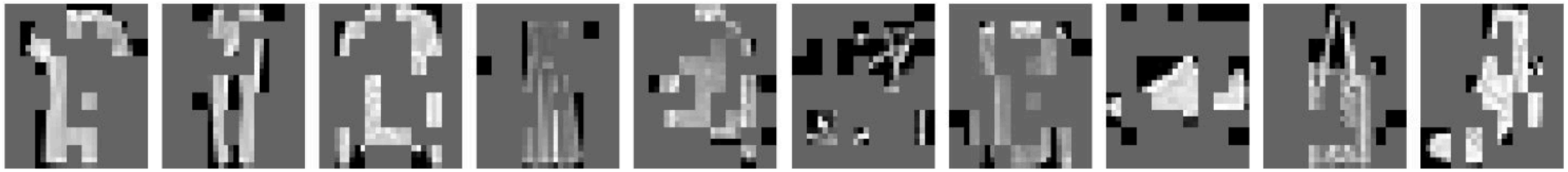
Our technique:



LIME (KDD 16):

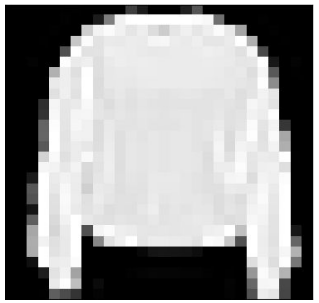


SHAP (NIPS 17):

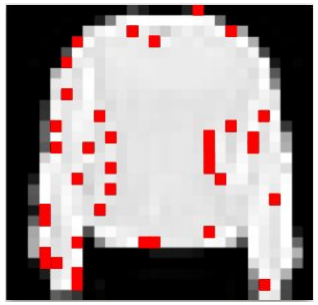


Scrutinize model weakness – adversarial samples

- Carry the right semantic but do not contain the features identified by our approach.
- How to generate testing samples:
  - Nullify the top important features identified by our approach among positive samples.
  - If we select the correct features: DL model will misclassify the generated samples.



Original image

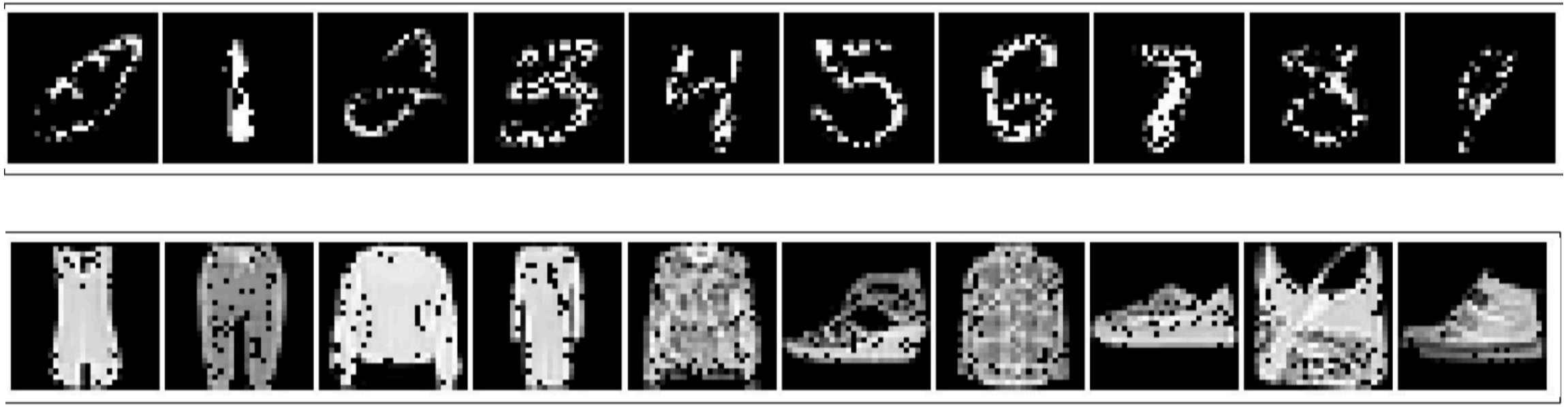


Important features

Adversarial samples

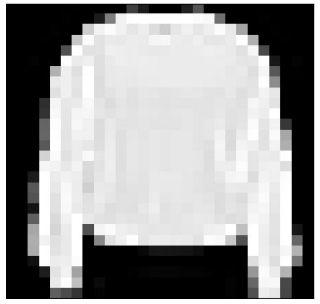
Scrutinize model weakness – pathological samples

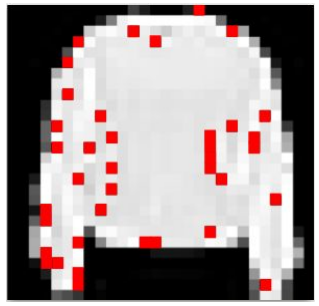
- Models classify these samples into wrong classes.

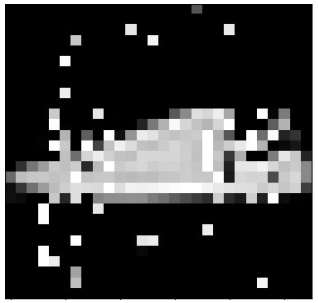


Scrutinize model weakness – pathological samples

- Contain the important features, but are unclassifiable to human.
- How to generate testing samples:
  - Replace the feature values of the negative sample with those of one positive sample.
  - If we select the correct features: DL model will classify the generated samples as positive.


Original image


Important features

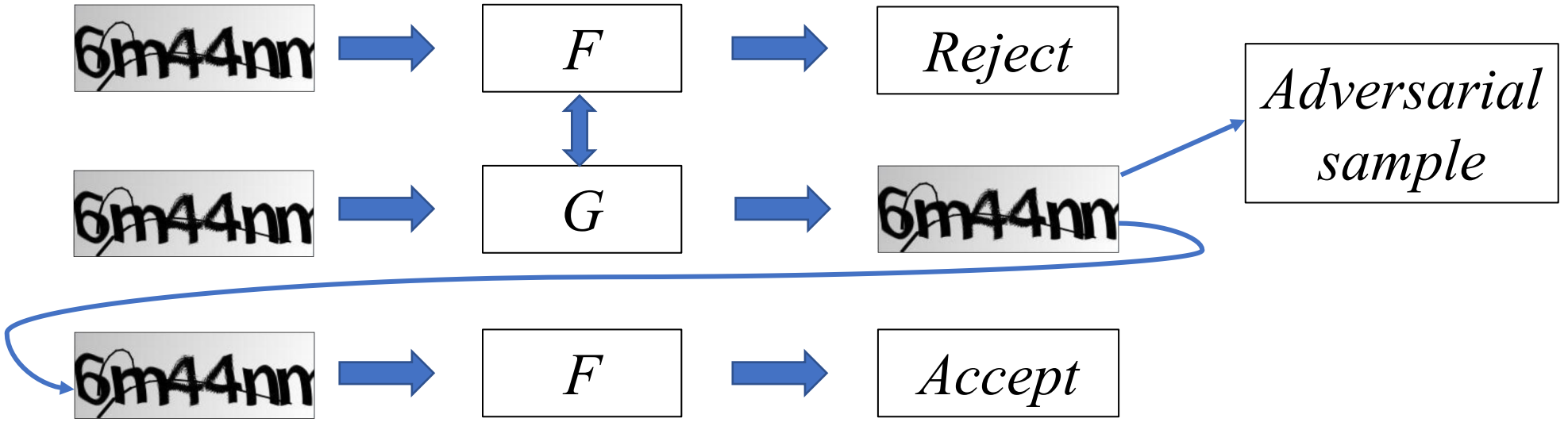

Pathological sample

Scrutinize model weakness – adversarial samples

- Models classify these samples into wrong classes.

Potential application: black market

- Black market system uses AI for identity verification (Image verification).
- Our technique identifies the important pixels in images without model architectures.
- Generate adversarial samples to bypass the black market system.



Summary

- Interpreting Deep Learning model.
  - Provide reasons for model decision.
  - Help user understand model behavior and build trust of the model.
- High **Precision** Black-box explanation technique.
  - Establish explanations to individual model outputs.
  - Scrutinize the model weakness and explore blind spots.
  - Potential to be applied to security applications (e.g., black market).

Thank you very much!

京东安全