Explainable AI (XAI)

Presented by: Yue Ma
yue.ma.anna@rutgers.edu
July 8, 2020

Content

01 XAI Definition

02 The Purpose of XAI

03 Methodologies

04 Implementation

05 Benefits of XAI

01

What is Explainable AI ?

What is XAI? (Definition)

- It is a **post-hoc analysis** (with details & reasons) to clearly explain a black box model and make it easily understood.
- Its meaning depends on who the target **stakeholders** are.

Developers / AI Experts

Ensure/Improve product efficiency, research, new functionalities,…

Examiners / Regulators

Certify model compliance with the legislation in force, audits,…

Business Managers/ CEO

Assess regulatory compliance, Understand corporate AI applications,…

Other Researchers

Trust the model itself, gain scientific knowledge,…

Affected Users

Understand their situation, verify fair decisions,…

What is XAI? (Purposes)

Informativeness

Confidence

Transferability

Fairness

Accessibility

Trustworthiness,
Causality,
Interactivity,
Privacy Awareness

What is XAI? (Purposes)

Informativeness

Problem being solved ≠ Problem being faced

Hence, XAI models should give information about the problem being faced.

→ Extracting information on the inner relations of a model

What is XAI? (Purposes)

Transferability

Models are always bounded by constraints

XAI helps explaining the boundaries that affect a model, which increase understanding & implementation.

→ Understand the inner relations to reuse models

What is XAI? (Purposes)

Models seem incomprehensible to non-expert users

XAI helps end users to get involved in improving & developing certain models.

→ Ease the burden to deal with perplexing algorithm

Accessibility

What is XAI? (Purposes)

Models are expected to be reliable

XAI provides different methods to maintain robustness & stability for specific models.

→ Assure trustworthy interpretations produced by models

Confidence

What is XAI? (Purposes)

Models & algorithm develop rapidly over the years

XAI helps to identify bias in data & give clear visualizations of the relations that affect model results

→ Guarantee fair and ethical analysis in areas such as science, finance and social studies

Fairness

What is XAI? (Purposes)

- Explainable ≠ Trustworthy, and trustworthy models might not be explained. Not easy to quantify.

- Correlation ≠ Causality. XAI can give intuition of possible causal relations or test causality between variables.

- When end users are important, interactivity ensures their ability to modify & get across the models.

- Users data may be deployed against them, so revealing inner relations of complex models helps avoiding a privacy breach.

Trustworthiness,
Causality,
Interactivity,
Privacy Awareness

02

Why is XAI important ?

Why is XAI important?

JP Morgan Case of Loss:
In 2017, the bank was charged $55 million for their mortgage discrimination against the African-American & Hispanic borrowers from 2006 to 2009. The higher interest rate was calculated from *”a model all banks had in 2006 to 2009”*, said the bank’s spokeswoman, Elizabath Seymour.

This affected 53,000 minority borrowers who, on average, paid $1,126 more than white borrowers for a $236,800 loan during this period. Similarly, Wells Fargo paid $175 millions for this issue in 2012.

LAWYER'S
REFERENCE GUIDES
GDPR
The General Data Protection Regulation
DALLAS VICKY - dallas@9amlabs.com

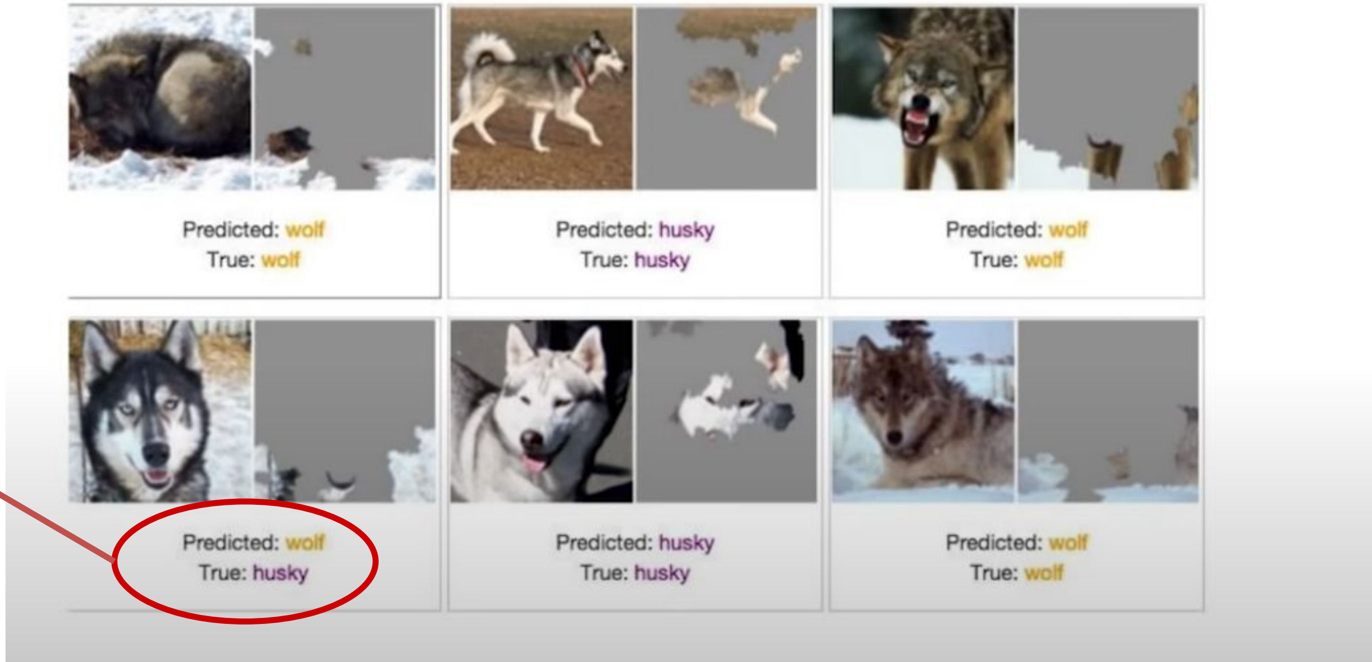
Regulations:

- **Equal Credit Opportunity act in U.S. (1974)**
- **GDPR in EU (2018)**

Why is XAI important?

Classify Husky & Wolf:

Only 1 mistake

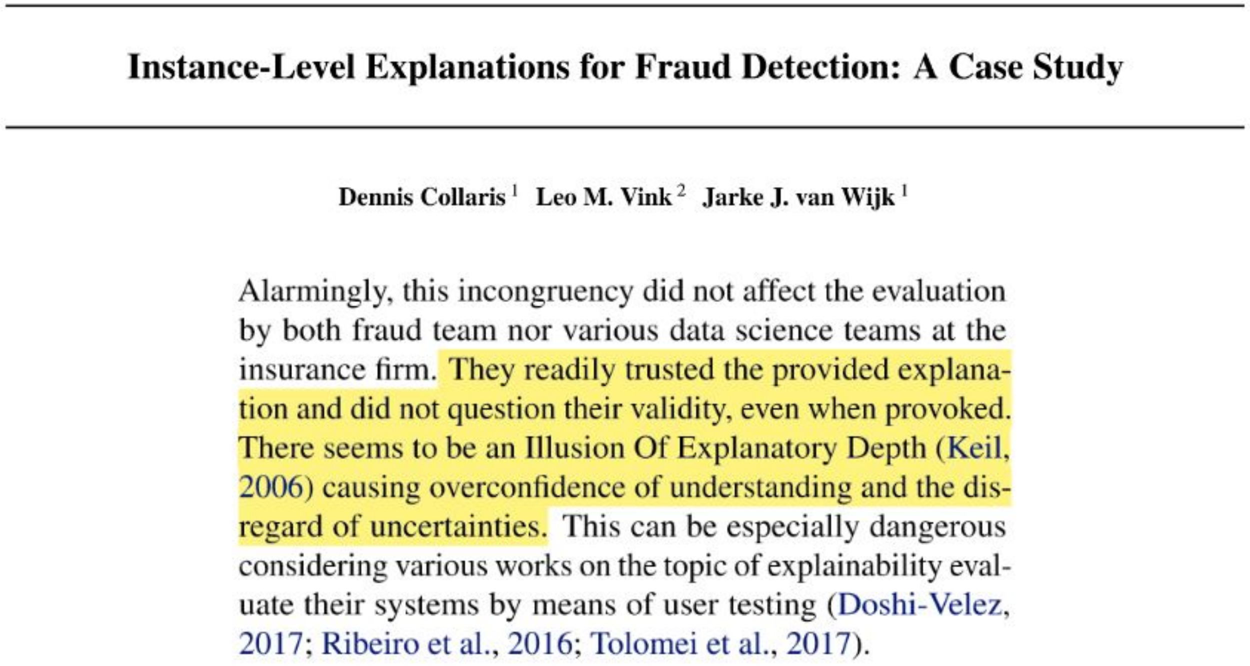


Question?

Can we trust a model & its prediction just because it looks good or accurate?

Why a model or its evaluation could go wrong ?

- **Data Leakage**: signals leak into the training (& validation) data
  - which could poten
  - hard to detect, *i.e.*
- **Data Shift**: training dat x and output y
  - explanation helps t
- **Mismatch** between the d, *i.e. user engagement/retention v.*
- **Not comparing** relative
- **Illusion of explanatory**



Instance-Level Explanations for Fraud Detection: A Case Study

Dennis Collaris[1] Leo M. Vink[2] Jarke J. van Wijk[1]

Alarmingly, this incongruency did not affect the evaluation by both fraud team nor various data science teams at the insurance firm. They readily trusted the provided explanation and did not question their validity, even when provoked. There seems to be an Illusion Of Explanatory Depth (Keil, 2006) causing overconfidence of understanding and the disregard of uncertainties. This can be especially dangerous considering various works on the topic of explainability evaluate their systems by means of user testing (Doshi-Velez, 2017; Ribeiro et al., 2016; Tolomei et al., 2017).

Bottom Line We should be able to **Trust & Understand the insights of** the model when making decisions based on a prediction.

Why is XAI important?

Prevent potential mistakes

Compliance with laws & regulations

Place trust in models

03

Why are the methodologies?

What needs to be done to implement XAI?

1. Classify transparent systems & opaque systems

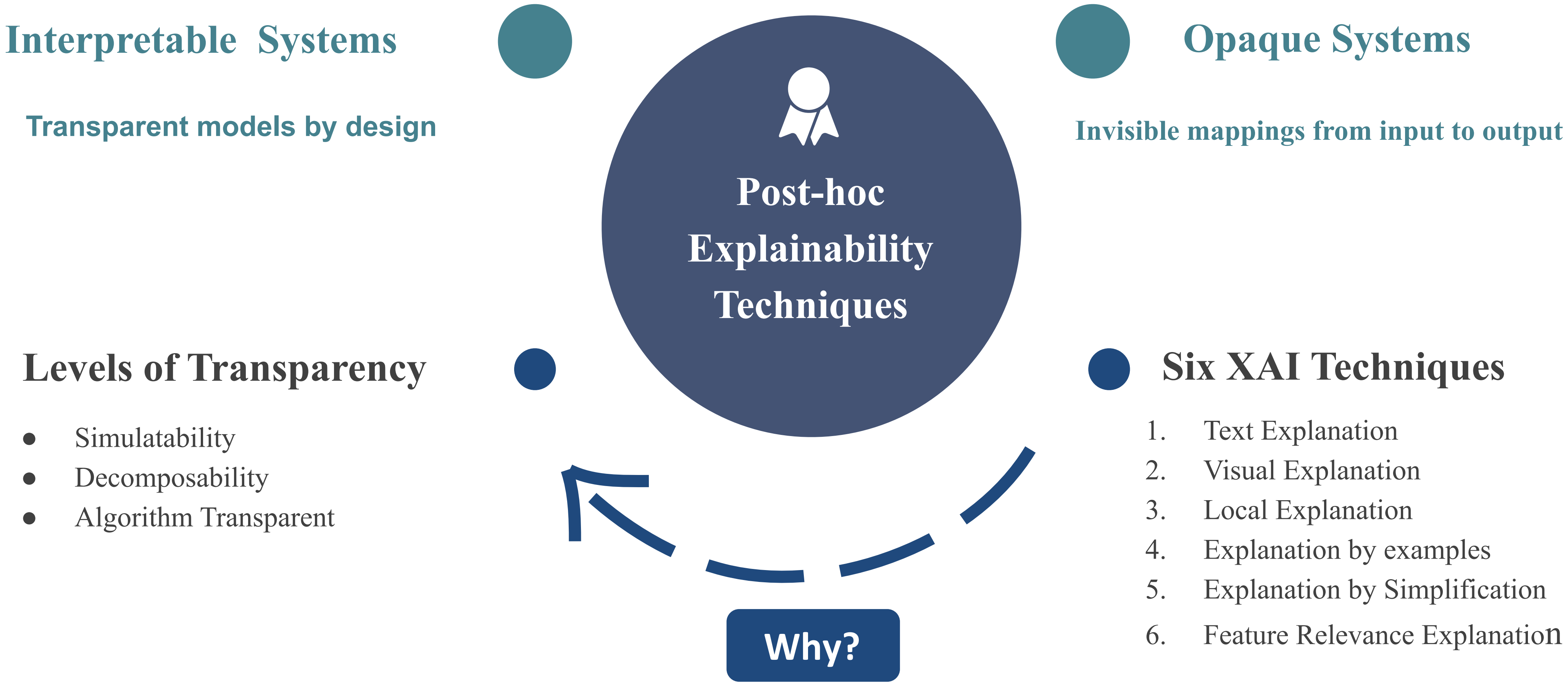
2. Identify levels of transparency

3. Explain why transparent models still need to be explained

4. Understand the common post-hoc explainability techniques

5. Understand the use of model-agnostic v.s. model-specific techniques

What needs to be done to implement XAI?



Interpretable Systems

Transparent models by design

Opaque Systems

Invisible mappings from input to output

Post-hoc Explainability Techniques

Levels of Transparency

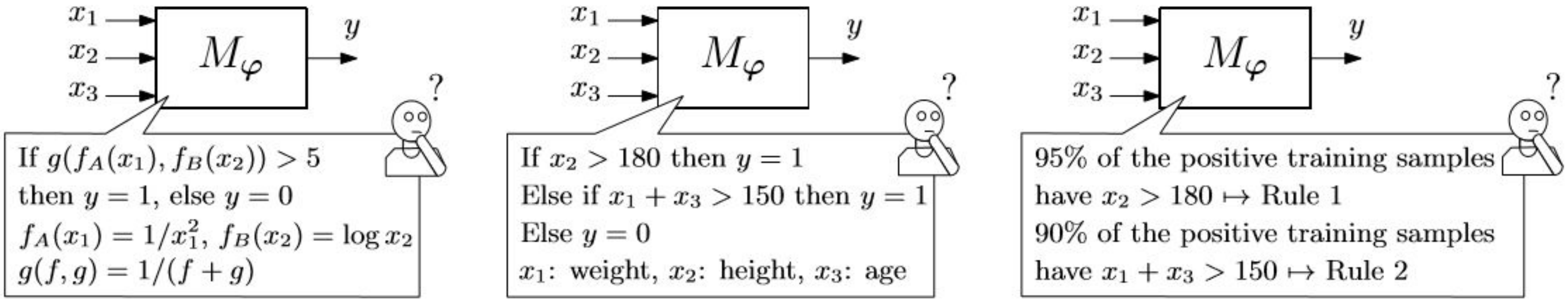
- Simulatability
- Decomposability
- Algorithm Transparent

Six XAI Techniques

1. Text Explanation
2. Visual Explanation
3. Local Explanation
4. Explanation by examples
5. Explanation by Simplification
6. Feature Relevance Explanation

Why?

Levels of Transparency



If $g(f_A(x_1), f_B(x_2)) > 5$
then $y = 1$, else $y = 0$
$f_A(x_1) = 1/x_1^2$, $f_B(x_2) = \log x_2$
$g(f, g) = 1/(f + g)$

(a)

If $x_2 > 180$ then $y = 1$
Else if $x_1 + x_3 > 150$ then $y = 1$
Else $y = 0$
x_1: weight, x_2: height, x_3: age

(b)

95% of the positive training samples
have $x_2 > 180 \mapsto$ Rule 1
90% of the positive training samples
have $x_1 + x_3 > 150 \mapsto$ Rule 2

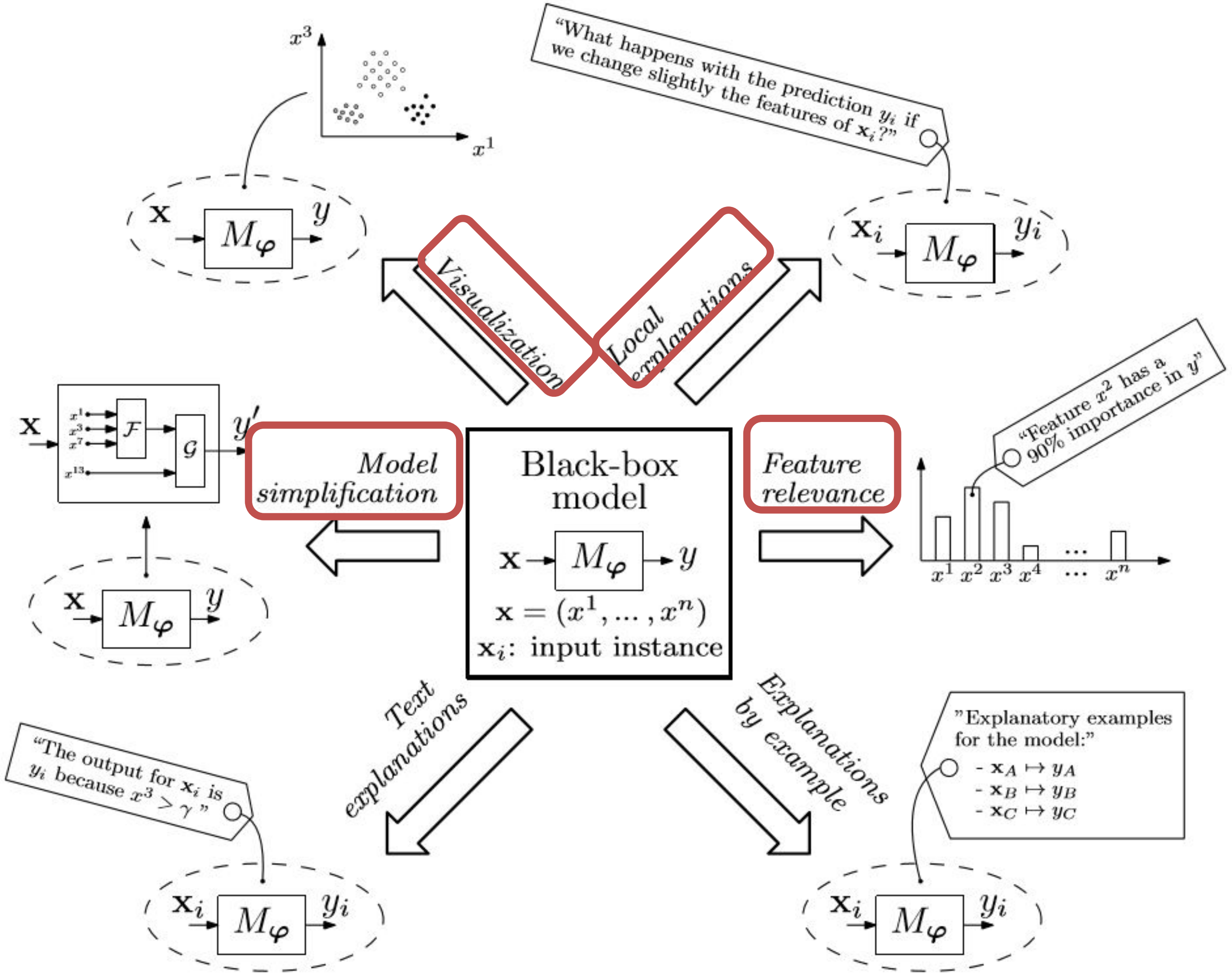
(c)

Simulatability **Decomposability** **Algorithm Transparency**

Why?

- ➢ Transparent properties lost in some circumstances
- ➢ Improve interpretability of the model

Post-hoc Explainability Techniques


Black-box model
$\mathbf{x} \rightarrow M_\varphi \rightarrow y$
$\mathbf{x} = (x^1, \ldots, x^n)$
$\mathbf{x}_i$: input instance
Visualization
Local explanations
Model simplification
Feature relevance
Text explanations
Explanations by example
x^3
x^1
$\mathbf{x}$
M_φ
y
"What happens with the prediction y_i if we change slightly the features of $\mathbf{x}_i$?"
$\mathbf{x}_i$
y_i
x^1 x^3 x^7 x^{13}
$\mathcal{F}$
$\mathcal{G}$
y'
"Feature x^2 has a 90% importance in y"
x^1 x^2 x^3 x^4 … x^n
"The output for $\mathbf{x}_i$ is y_i because $x^3 > \gamma$"
"Explanatory examples for the model:"
- $\mathbf{x}_A \mapsto y_A$
- $\mathbf{x}_B \mapsto y_B$
- $\mathbf{x}_C \mapsto y_C$

□ **Model-Agnostic Techniques**

Model-Agnostic Techniques

Explanation by simplification (most popular)

- Including local explanation
- Based on rule extraction techniques (methods below)
    - **LIME**: explain any classifier by linear approximation of specific observations
    - **G-REX** (Genetic Rule Extraction)
        - a GP (Genetic Programming) framework
        - offers different optimization methods on a variety of predefined *classification and regression* models
        - *plug-in feature:* allows new nodes & fitness functions to extend new models
        - "impact on advertising" problem: results show *high accuracy & comprehensibility*
        - *shortcomings:* computationally expensive; unverified on large data set; look for short rules, thus not representing all complexity of ANN

Model-Agnostic Techniques

Explanation by simplification (most popular)

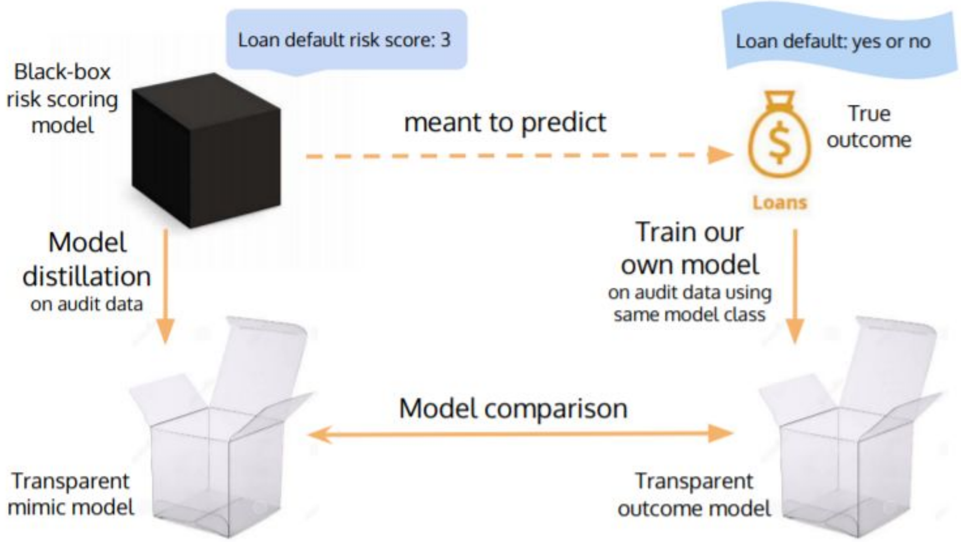
- **Learn rules** in CNF (Conjunctive Normal Form) or DNF (Disjunctive Normal Form)
  - aiming for trade-off between classification accuracy & rule simplicity
  - by learning two-level Boolean rules from dataset with
    - lower level: conjunctions of binary feature build clauses
    - upper level: disjunction of clauses from the prediction
  - objective function for optimization: a weighted combination of (a) classification errors (in Hamming distance) between the current rule and the closest rule that correctly classifies a sample and (b) sparsity.

- **Approximate** a transparent model to a complex one
  - The issues of the complex model should be reflected in the approximation as well
  - performs well on interpreting random forest & neural nets; performance measured by relative accuracy on a test set

Model-Agnostic Techniques

Explanation by simplification (most popular)

- **Distill & audit** black-box models, including
  - black-box models (“teachers”) train transparent models (“students”) to mimic risk scores of the “teachers”
  - a method for model distillation & comparison to audit black-box risk scoring models
  - an statistical test to check if the auditing data is missing key features it was trained with

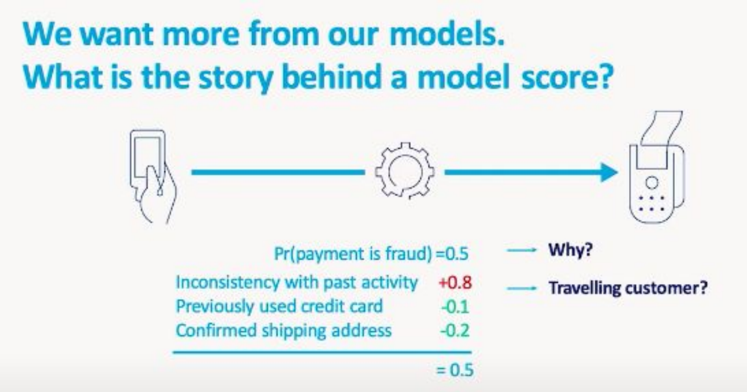

Loan default risk score: 3
Loan default: yes or no
Black-box risk scoring model
meant to predict
True outcome
Loans
Model distillation on audit data
Train our own model on audit data using same model class
Model comparison
Transparent mimic model
Transparent outcome model

source: https://arxiv.org/pdf/1710.06169.pdf

Model-Agnostic Techniques

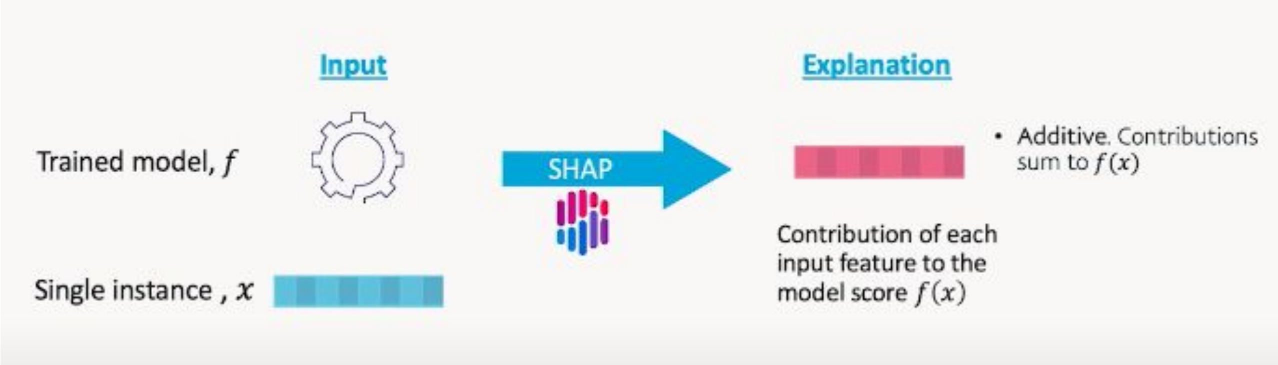
Feature Relevance Explanation (actively developing)

- ➢ Measure and rank the importance, influence or relevance each feature has in predicted outputs
- ➢ Some methods
  - ○ **SHAP** : "locally additive feature importance", game theory



We want more from our models.
What is the story behind a model score?

Pr(payment is fraud) =0.5		→ Why?
Inconsistency with past activity	+0.8	→ Travelling customer?
Previously used credit card	-0.1	
Confirmed shipping address	-0.2	
	= 0.5	

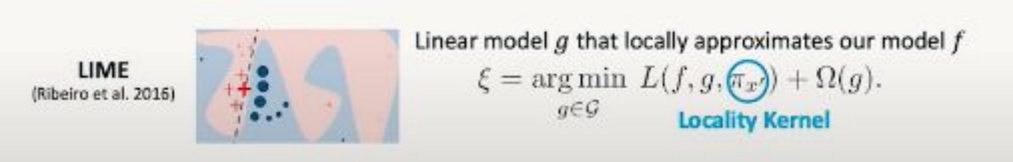


- Additive. Contributions sum to $f(x)$

Contribution of each input feature to the model score $f(x)$

Connection between LIME & SHAP ?

A single unique solution is proven to the class of additive feature attribution;
Thus, the SHAP value is the o**nly possible solution** of the LIME kernel.



Linear model g that locally approximates our model f

$$\xi = \underset{g \in G}{\arg\min}\ L(f, g, \pi_x) + \Omega(g).$$

Source:https://www.youtube.com/watch?v=0yXtdkIL3Xk

Model-Agnostic Techniques

Feature Relevance Explanation (actively developing)

- **Conditional Game Theory & local gradients**
  - Ex. The prediction of Naive Bayes (assuming conditional independence) can be transformed into contribution of individual feature values
  - Local probability gradients can test how changes in each feature contributes to the changes of output

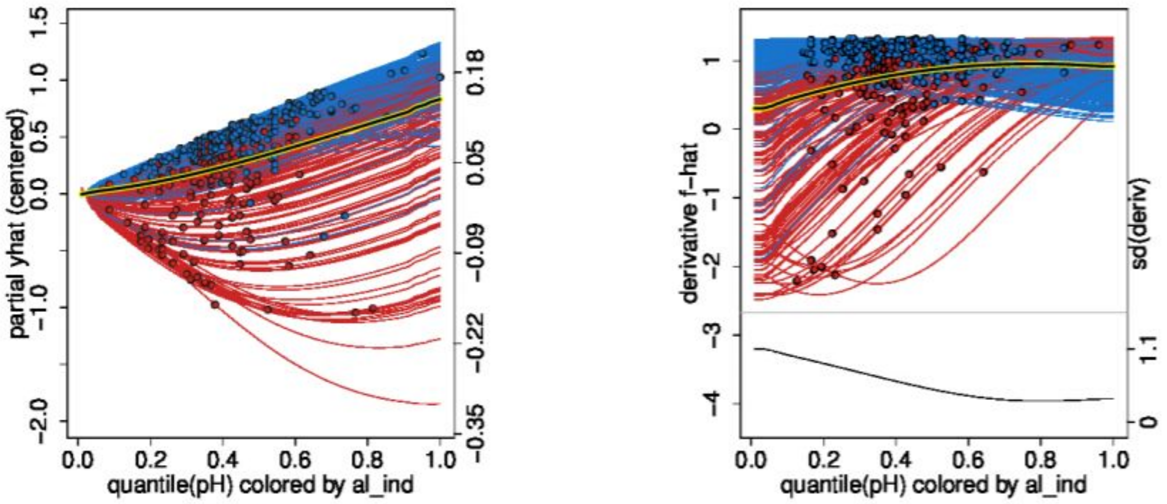
- **Group features** to analyze the relations & dependencies found in the model; to get insights of data
  - *a method to correlate inputs* while measuring influence
- A contrary method: build a **Global SA** (Sensitivity Analysis) with existing SA to extend applicability of methods
- Real-time image saliency method: applied to differentiable image classifiers
- Automatic STRucture IDentification method (ASTRID)
  - finds the largest features subset

    i.e. accuracy of the subset-trained classifier ≈ accuracy of original feature-trained classifier
- Many more methods!! Recent & vibrant (date from 2017 & 2018)

Model-Agnostic Techniques

Visual Explanation (less common)

- visualization approach based on Sensitive Analysis (SA)
  - Global SA & 3 novel SA methods: data-based SA, Monte-Carlo SA, cluster-based SA
  - a novel input importance measure (Average Absolute Deviation)



(a) c-ICE for NN

(b) d-ICE for NN

source: https://arxiv.org/pdf/1309.6392.pdf

- ICE (Individual Conditional Expectation) plots to visualize any supervised learning model

04

How to implement XAI ?

Implementation for Transparent Systems

Logistic/Linear Regression

Rule-based Learning

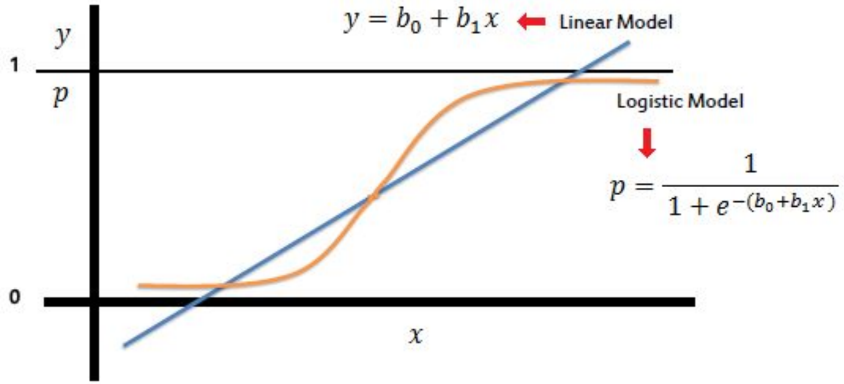
Decision Tree

General Additive Models

KNN

Bayesian Models

Linear/Logistic Regression



source: https://saedsayad.com/logistic_regression.htm

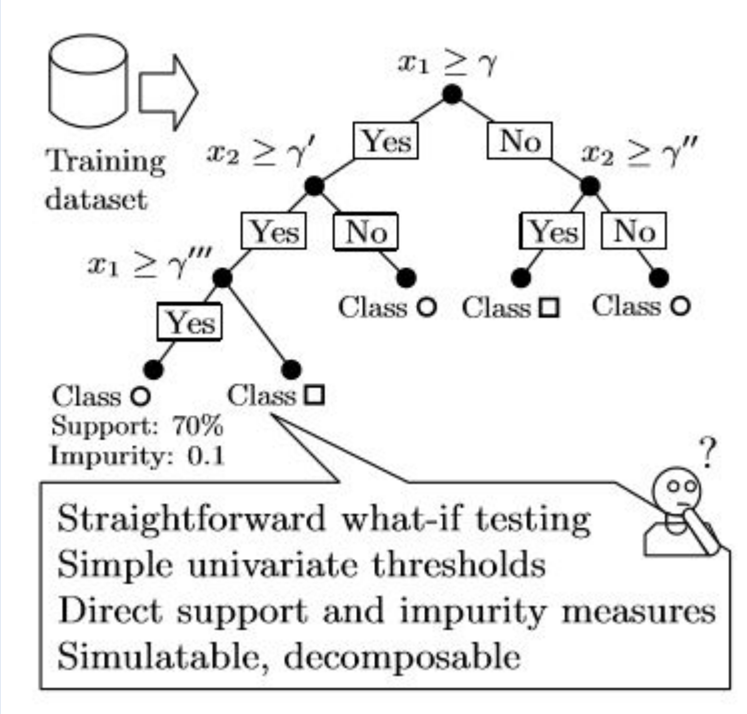
- Logistic (binary dependent variable) vs Linear (continuous dependent variable)
- Social science users: gain enhanced understanding & more accurate estimate with post-hoc explainability techniques
  - eg. when expressing results, users estimate outcomes correctly in 46% of the cases with natural frequencies vs 10% of the cases with probabilities
- Using techniques for:
  - overall model evaluation
  - statistical test for individual predictors
  - good-of-fit statistics
  - validation of predicted probabilities

Remember: Audiences matter!

- Highly engineered features harm decomposability.
- Large size model harms simulatability.

Decision Tree

- Trees help making decisions to support regression & classification
- **Losing transparent properties**:
  - Increasing size beyond human understanding: *simulatability lost*
  - Further increasing size & using complex feature relations: *decomposability lost*
- Tree ensembles overcome the poor generalization properties when a balance between predictive performance is important, *but lose all transparent properties.*



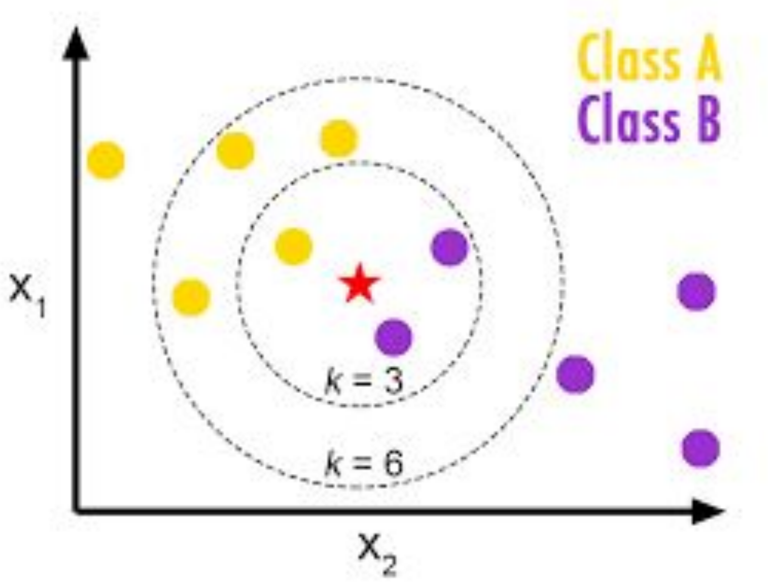
Straightforward what-if testing
Simple univariate thresholds
Direct support and impurity measures
Simulatable, decomposable

KNN

- ➢ KNN predicts the class by voting the classes of its K nearest neighbours

- ➢ *Similar to human decision making*: KNN predictions rely on specific problems being tackled (i.e. calculating distance & similarity between examples) ↔ Human decision rely on past similar cases

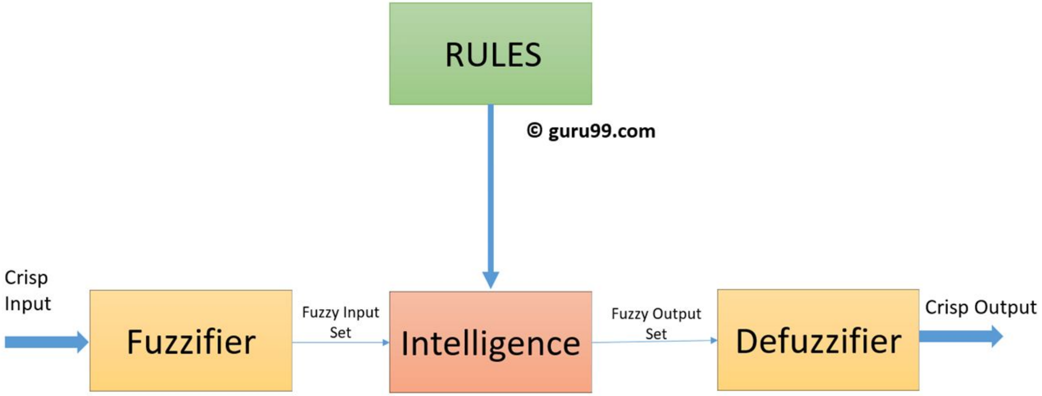
- ➢ **Losing transparent properties**:
  - ○ High K number harms full simulation of the model performed by human user: *simulatability lost*
  - ○ Complex features or distance functions: *decomposability lost*



source: https://medium.com/@equipintelligence/k-nearest-neighbor-classifier-knn-machine-learning-algorithms-ed62feb86582

Rule-based Learning

- Simple rules (eg. if-then) or complex combinations of simple rules
- **Losing transparent properties (trade-off)**:
  - More rules give greater coverage, but harm interpretability.
  - Greater rule specificity base on higher number of antecedents or consequences, thus harming interpretability.
  - Models become *algorithmically transparent only*
- **A potential solution for this**: transform from classical rules to *fuzzy rules* relax the constraints of rule size, because
  - fuzzy rules operate on linguistic terms, thus enhancing model understandability
  - fuzzy rules perform better in uncertainty



source: https://www.guru99.com/what-is-fuzzy-logic.html

General Additive Models (GAM)

- A linear model that *aggregates* some unknown *smooth functions* defined for independent variables to predict the dependent variable
- An *interpretable* model, since users can verify the importance of each variable
- Usage in risk assessment. *Visualization* methods improve interpretation.

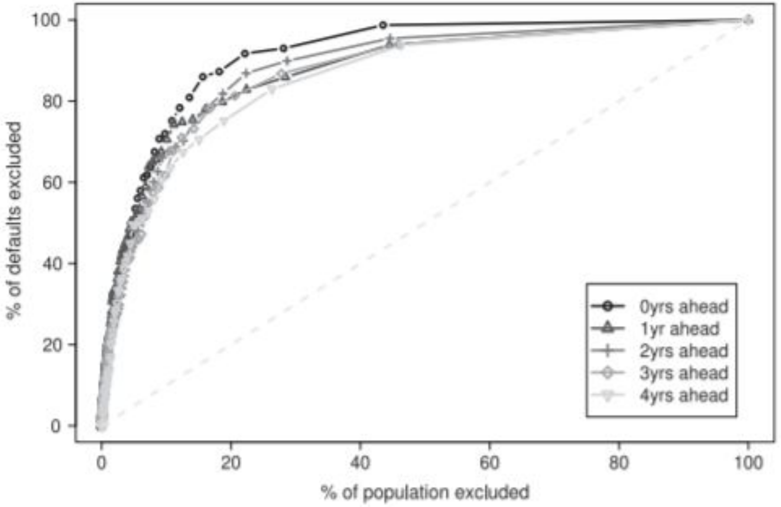


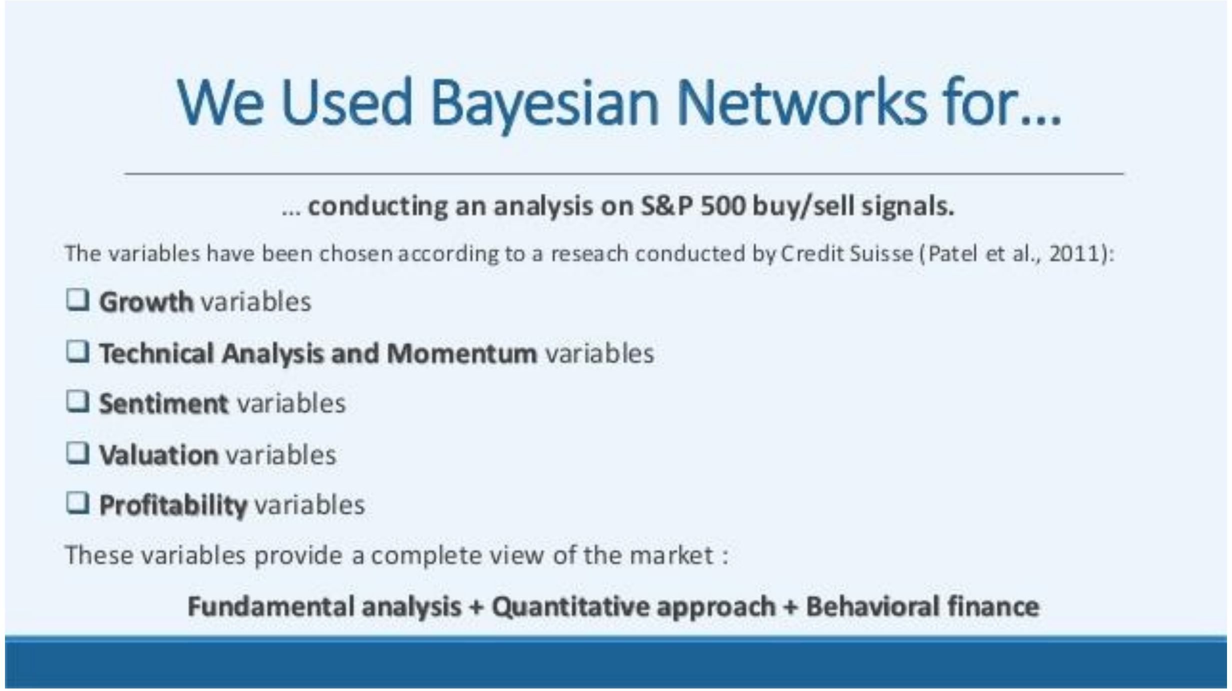
Figure 3. Predictive power depreciation 0–4 years into the future for a GAM model. 1996 data, one-year default horizon.

A default horizon model implying performance depreciation

source: D. Berg, Bankruptcy prediction by generalized additive models, Applied Stochastic Models in Business and Industry 23 (2) (2007) 129–143.

Bayesian Models

- ➢ a *probabilistic* model with links representing *conditional dependencies* between two variables
- ➢ Examples



We Used Bayesian Networks for...

... conducting an analysis on S&P 500 buy/sell signals.

The variables have been chosen according to a reseach conducted by Credit Suisse (Patel et al., 2011):

- ❑ **Growth** variables
- ❑ **Technical Analysis and Momentum** variables
- ❑ **Sentiment** variables
- ❑ **Valuation** variables
- ❑ **Profitability** variables

These variables provide a complete view of the market :

Fundamental analysis + Quantitative approach + Behavioral finance

source: https://www.slideshare.net/AlessandroGreppi3/financial-markets-signal-detection-with-bayesian-networks-phd-dreamt-workshop-17th-march-2016

- ➢ **Losing transparent properties**:
  - ○ when the model is overly complex or has cumbersome variables

Implementation for Opaque Systems

Tree Ensembles

RNN

Support Vector Machine

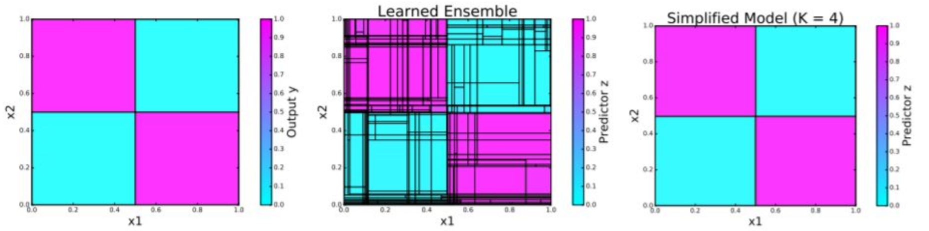
CNN

Multi-layer NN

<u>Alternative Taxonomy</u>

Tree Ensembles

- Cure overfitting by combining different trees to obtain an aggregated prediction/regression
- **But could be too complex to interpret, so**:
  - use *explanation by simplification:*
    - train a simple model from random sample of data labeled by the ensemble model
    - create a simplified Tree Ensemble Learner(STEL)
    - use two models, a simple one for interpretation & a complex one for prediction



(a) Original Data (b) Learned Tree Ensemble (c) Simplified Model

Figure 1. The original data (a) is learned by ATM with 744 regions (b). The complicated ensemble (b) is approximated by four regions using the proposed method (c). Each rectangle shows each input region specified by the model.

source:
https://arxiv.org/pdf/1606.05390.pdf

  - use f*eature relevance:*
    - measure MDA (Mean Decrease Accuracy) or MIE (Mean Increase Error) of the forest when a specific variable is randomly permuted in the out-of-bag sample
    - use a framework to convert the class of an example by descriptively disentangled the variable importance

Tree Ensembles

- ➢ Ensemble strategies:
  - ○ bagging ensembles
  - ○ scarce activity by boosting & stacking classifiers
    - ■ Stacking With Auxiliary Features (SWAF) by harnessing & integrating explanations in stacking ensembles to improve their generalization
  - ○ DeepSHAP: stacking ensembles & multiple classifier systems in addition to Deep Learning models

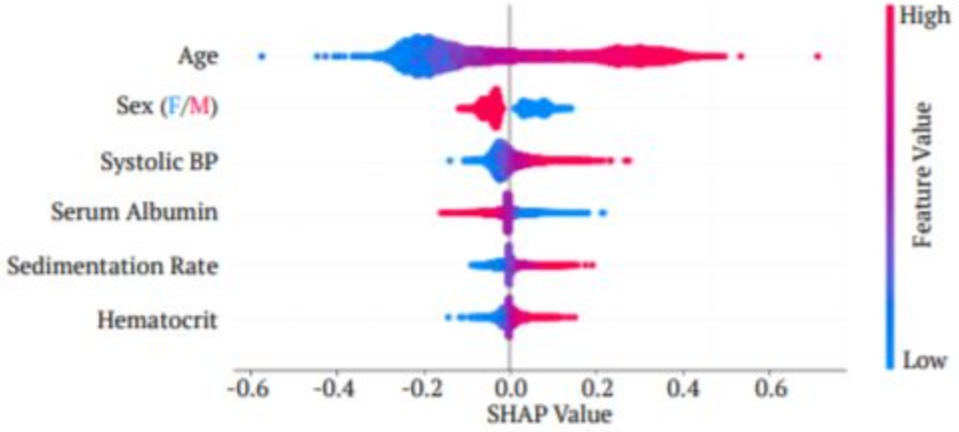


Figure 3: *Summary plot of DeepSHAP attribution values.* Each point is the local feature attribution value, colored by feature value. For brevity, we only show the top 6 features.

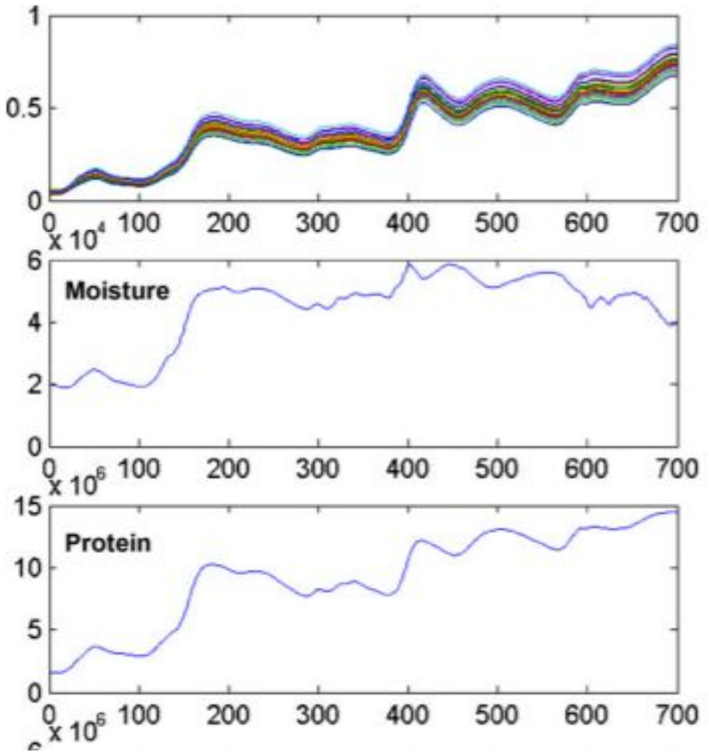
source: https://arxiv.org/pdf/1911.11888.pdf

Support Vector Machine (SVM)

- ➢ SVM, widely used, has excellent prediction & generalization abilities.
  Post-hoc explanability techniques:
  - ○ use *explanation by simplification (4 classes of simplification):*

    1. Build rules based only from support vectors of a trained model
       - ■ use a modified sequential covering algorithm to directly extract rules from a trained SVM
       - ■ generate fuzzy rules, allowing for linguistic understanding
    2. Propose the addition of the SVM's hyperplane
    3. Add the actual training data as a component for rule-building
    4. Use a growing SVC to interpret linear rules of SVM decisions

  - ○ use *visualization:*
    - ■ visualize which and how input variable actually related to the output data
    - ■ combine output of SVM with heatmaps
    - ■ Many studies only account for weight vectors, but **margin** is important as well (can use a statistic to explicitly account for SVM margin)



source: B. ¨Ust¨un, W. Melssen, L. Buydens, Visualisation and interpretation of support vector regression models, Analytica Chimica Acta 595 (1-2) (2007) 299–309.

Support Vector Machine (SVM)

- use *local explanation:* a newer approach
- use *explanation by examples*

Note: Bayesian system can be adopted as a post-hoc technique to explain SVM decisions

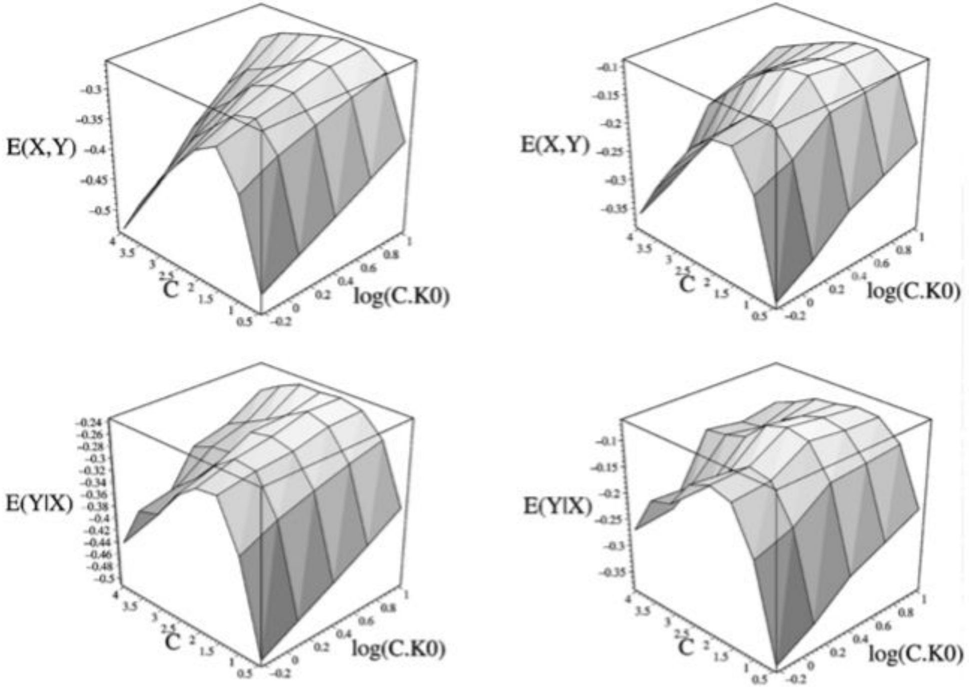


Figure 6. The evidence for an SVM with an RBF kernel trained on the data set of figure 4, as a function of the noise parameter C and the kernel amplitude K_0; the lengthscale $l = 0.05$ was kept fixed. The rows and columns show the same quantities as in figure 5. Along the K_0 axis, $\log_{10} CK_0$ is shown rather than just $\log_{10} K_0$, because a constant value of CK_0 corresponds to the same conventional SVM (maximum a posteriori) solution, independently of C. Note that the same is not true of the evidence.

source: P. Sollich, Bayesian methods for support vector machines: Evidence and predictive class probabilities, Machine learning 46 (1-3) (2002) 21–52.

Multi-layer NN

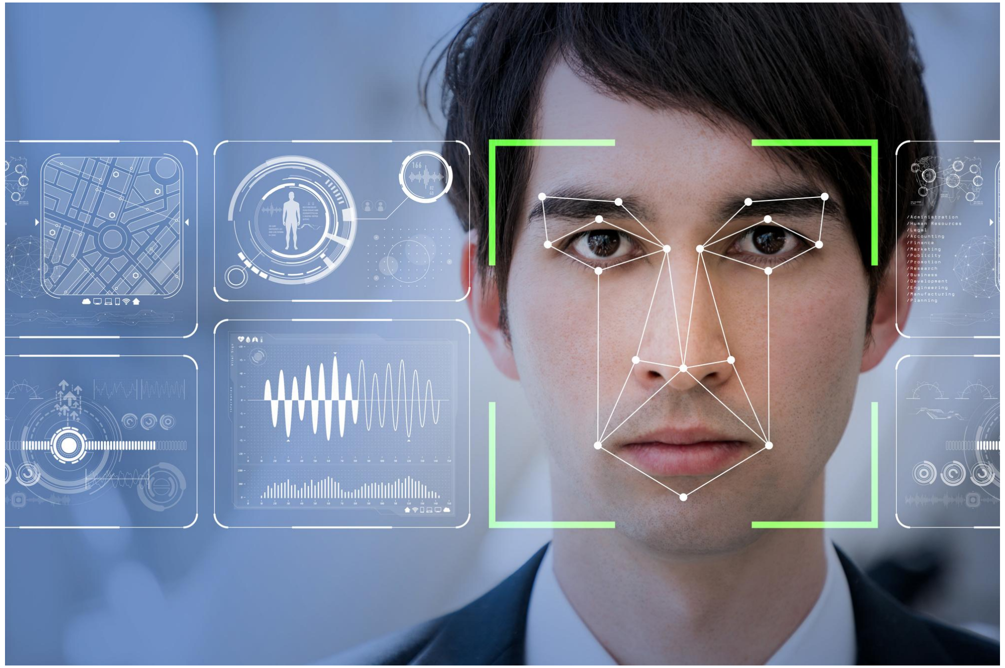
- **Post-hoc explainability techniques**:
  - use *explanation by simplification:*
    - DeepRed Algorithm: adding more decision trees & rules
    - Interpretable Mimic Learning: extracting models by gradient boosting trees
  - use *text explanation, local explanation & visualization*
  - use *feature relevance:*
    - Deep Taylor Decomposition: consider each neuron as an object to be decomposed & expanded, then aggregate & back-propagate these decomposition
    - DeepLIFT: compute importance scores in a multi-layer NN
    - Note: some axioms of feature relevance techniques have been violated in practices by most methods.

Take Home Message

- Theoretical soundness are important when applying post-hoc explanability techniques.
- New explanation methods that are theoretically correct can be explored.

CNN

Facial recognition techniques in the banking & financial industry not only provide great convenient & secure services for customers, but also allow banks to track transactions & fraud efficiently, thus reducing multiple risks.



CHASE

HSBC

USAA®

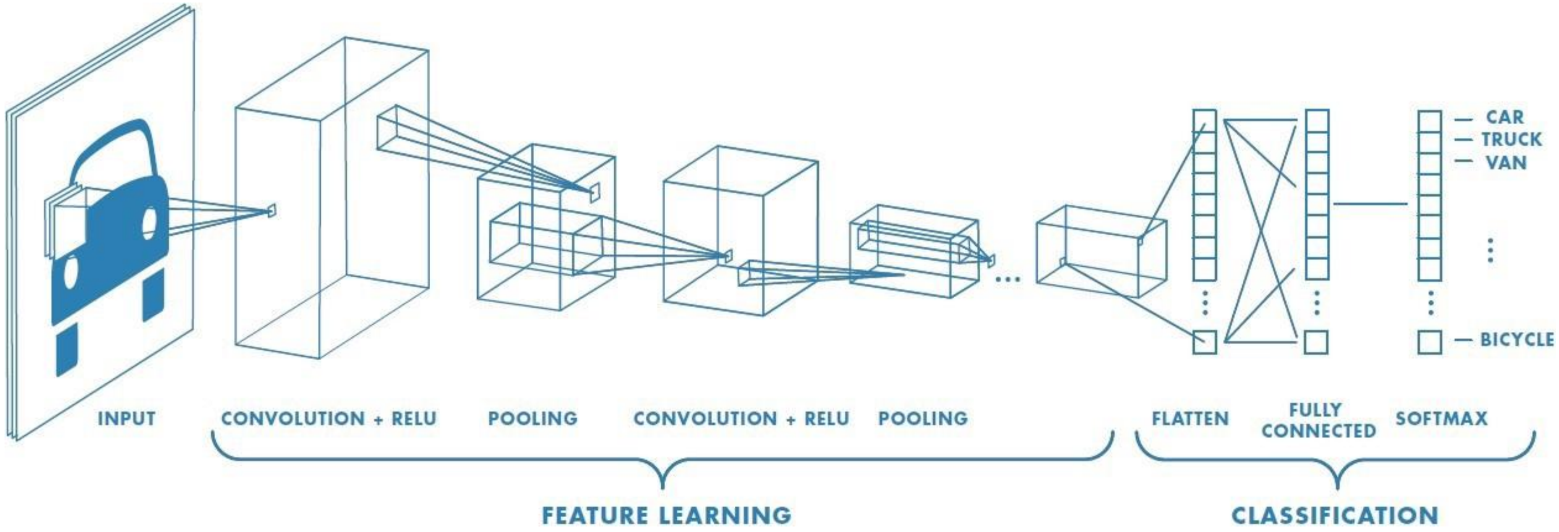
Alipay™

G Pay

CNN

- A sequence of convolutional layers that automatically learn higher level features

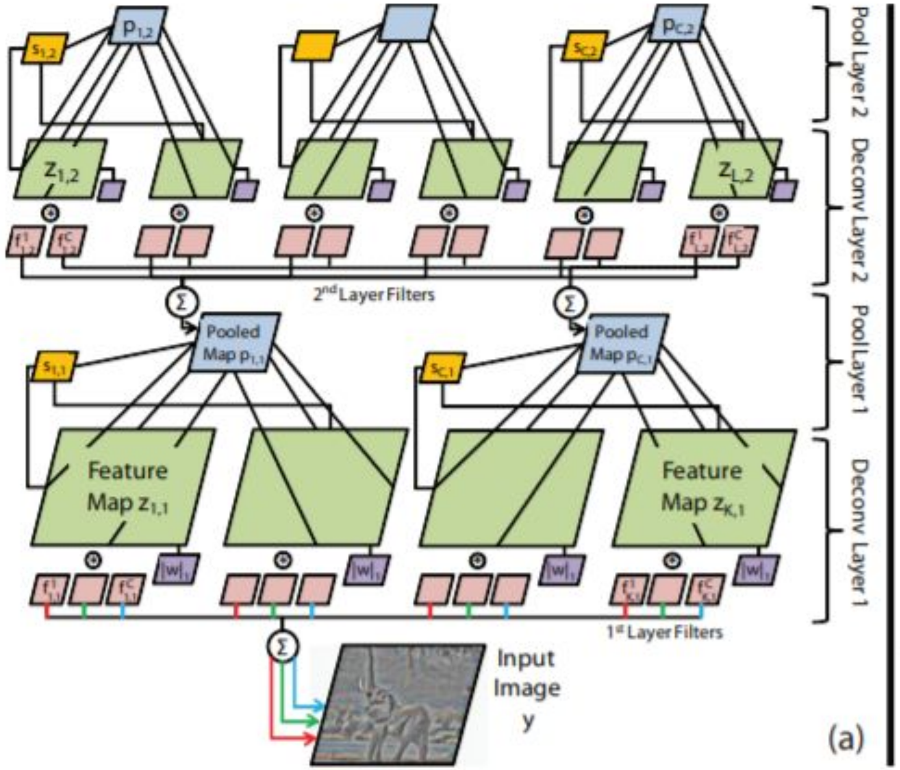
source: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53



- 2 categories to understand how CNN works:
  - understand the decision process by how & which input maps to the output
  - interpret how the intermediate layers see the external world in general

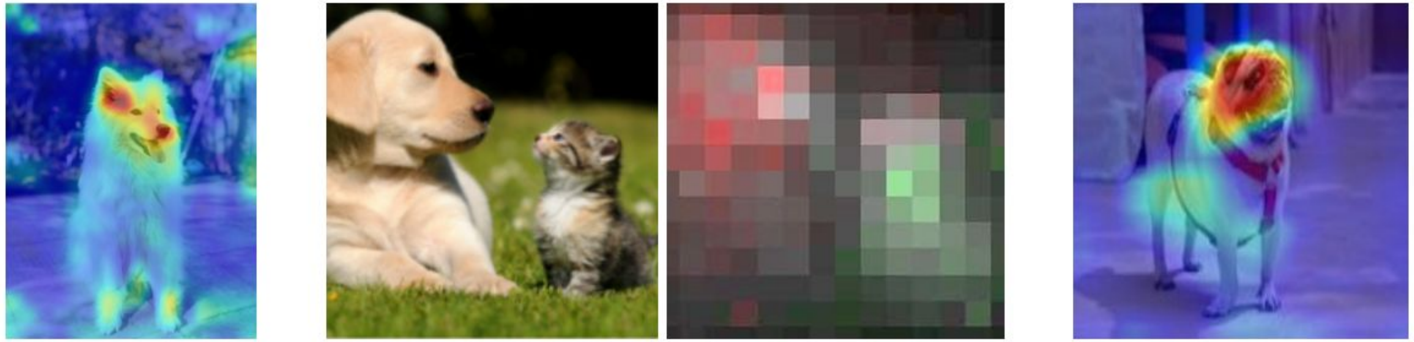
CNN

- ➢ Detailed methods:
  - ○ each layer outputs some feature maps when an input image runs feed-forward through a CNN



CNN

- ➢ Detailed methods:
  - ○ feed a feature map from a selected layer & reconstruct the maximum activations
  - ○ simplifying both the CNN architecture & the visualization method
  - ○ Layer-wise Relevance Propagation (LRP) technique
  - ○ Grad-CAM (Gradient-weighted Class Activation Mapping)



(a) Heatmap [168] (b) Attribution [293] (c) Grad-CAM [292]

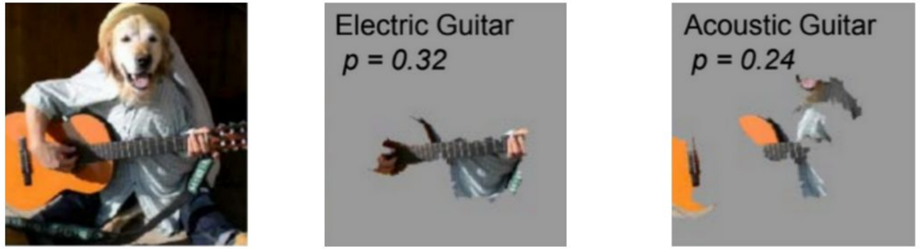
Figure 7: Examples of rendering for different XAI visualization techniques on images.

Take Home Message 1

- ● Visualization mixed with feature relevance methods are the most popular explainability approach for CNN.

CNN

- Detailed methods:
  - A much simpler approach in LIME framework
    - divide input image into interpretable components & create a set of perturbed instances
    - run each perturbed instance through the model & get a probability
    - locally weighted; present the superpixels with the highest positive weights as an explanation



(a) Original image (b) Explaining *electric guitar* (c) Explaining *acoustic guitar*

Figure 9: Examples of explanation when using LIME on images [71].

Take Home Message 2

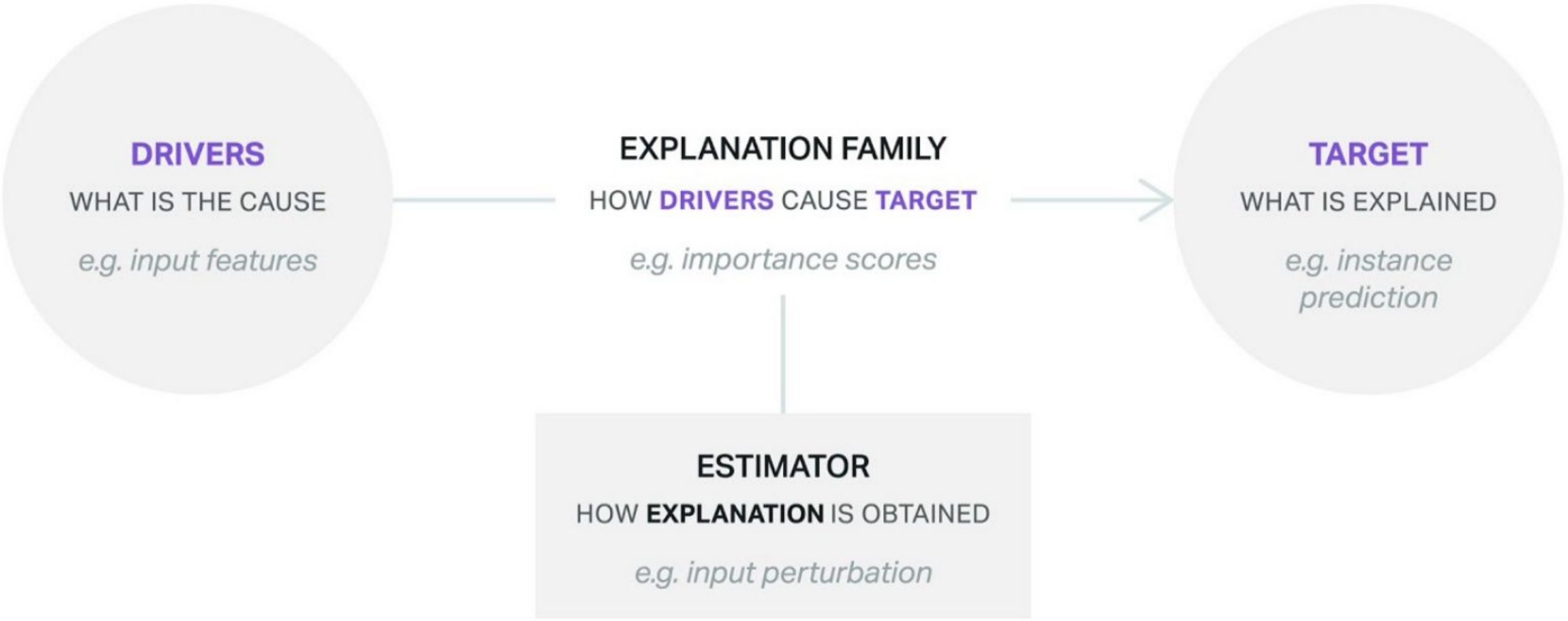
- CNN models are the fundamental of computer vision tasks, such as facial recognition & image classification.

RNN

- ➢ Used extensively for predictive problems defined over sequential data
  - ○ NLP (eg. chatbot, social media analysis)
  - ○ Time series analysis

- ➢ Explaining RNN by
  - ○ understanding what a RNN model has learned (via feature relevance)
    - ■ LRP(Layer-wise Relevance Propagation): proposed a specific propagation rule working with multiplicative connections like LSTM
    - ■ interpretable features learned by fitting Gradient Boosting Trees to the trained LSTM network
  - ○ modifying RNN architecture to provide insights of decision (via local explanation)
    - ■ RETAIN (REverse Time AttentIoN) model: detect influential pattern
    - ■ SISTA (Sequential Iterative Soft- Thresholding Algorithm): models a sequence of correlated observations with a sequence of sparse latent vectors

Alternative Taxonomy



DRIVERS

WHAT IS THE CAUSE

e.g. input features

EXPLANATION FAMILY

HOW **DRIVERS** CAUSE **TARGET**

e.g. importance scores

TARGET

WHAT IS EXPLAINED

e.g. instance prediction

ESTIMATOR

HOW **EXPLANATION** IS OBTAINED

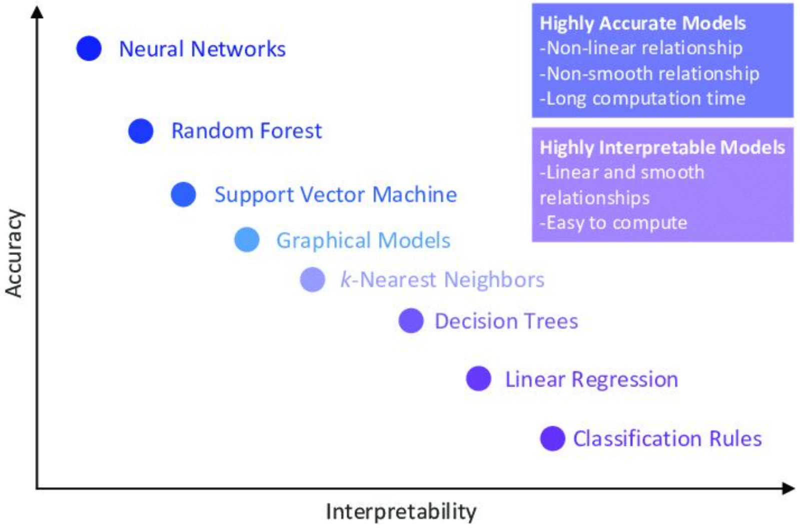
e.g. input perturbation

05

What are the benefits ?

Challenges

Trade-off between interpretability & performance



Is it possible to be both accurate & comprehensive?

XAI & Adversarial Machine Learning

Data model functionality stealing is possible

Ethical Issues

Fairness & Discrimination

User Privacy & Protection

Model Accountability

https://github.com/firmai/ml-fairness-framework

Reference

[1] https://www.youtube.com/watch?v=rPSiEDYcXr4&t=562s
[2] https://www.youtube.com/watch?v=I0yrJz8uc5Q&t=1s
[3] https://www.youtube.com/watch?v=n_mwYWfI_sI
[4] https://www.youtube.com/watch?v=vz_fkVkoGFM
[5] https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f
[6] https://www.sciencedirect.com/science/article/pii/S1566253519308103?casa_token=xAXa2ungOZwAAAAA:jjVbf1Dfq2wj8m-BrowifL4xzvBbeDEYfPeD4gBUxLTqSGuPPNgH6GBDg-zxKwKRpbsOIs2Z
[7] https://arxiv.org/pdf/1806.07129.pdf
[8] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.4124&rep=rep1&type=pdf
[9] https://www.researchgate.net/profile/Ulf_Johansson5/publication/228766101_Accuracy_vs_comprehensibility_in_data_mining_models/links/0deec52ff78cf32bc7000000/Accuracy-vs-comprehensibility-in-data-mining-models.pdf
[10] https://arxiv.org/pdf/1602.04938.pdf
[11] https://www.times-standard.com/2017/01/18/jpmorgan-settles-mortgage-discrimination-lawsuit/
[12] https://www.mcall.com/business/mc-fast-facts-jp-morgan-chase-mortgage-settlement-20170118-story.htm
[13] https://arxiv.org/pdf/1706.09773.pdfl
[14] https://arxiv.org/pdf/1606.05798.pdf
[15] https://arxiv.org/pdf/1710.06169.pdf
[16] https://www.youtube.com/watch?v=0yXtdkIL3Xk
[17] http://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf

https://github.com/firmai/ml-fairness-framework
https://github.com/firmai/ml-fairness-framework/blob/master/assets/Screen%20Shot%202020-04-01%20at%2010.33.01%20AM.png

THANKS

Any Questions?