Explainable Artificial Intelligence

Demystifying the Hype

About Me

Dipanjan Sarkar

Data Science Lead, Author, Google Developer Expert - ML

APPLIED MATERIALS®

Experts
Machine Learning

Springboard

{Propulsion}

MUST

Necessity for Explainable AI

What do you see?

How about now?

What would you say now?

The Problem with Machine Learning

Types of Bias in your Models & Data

1. **Sample Bias**

   Selective sampling of data e.g keeping majority samples of well-off people in a loan-lending model

2. **Racial Bias**

   Intentionally or often unintentionally training models on skewed data of certain demographics e.g. facial recognition on caucasian men/women

3. **Association Bias**

   The classic data skewness leading to models giving preference to a specific culture or gender e.g. AI in recruitment

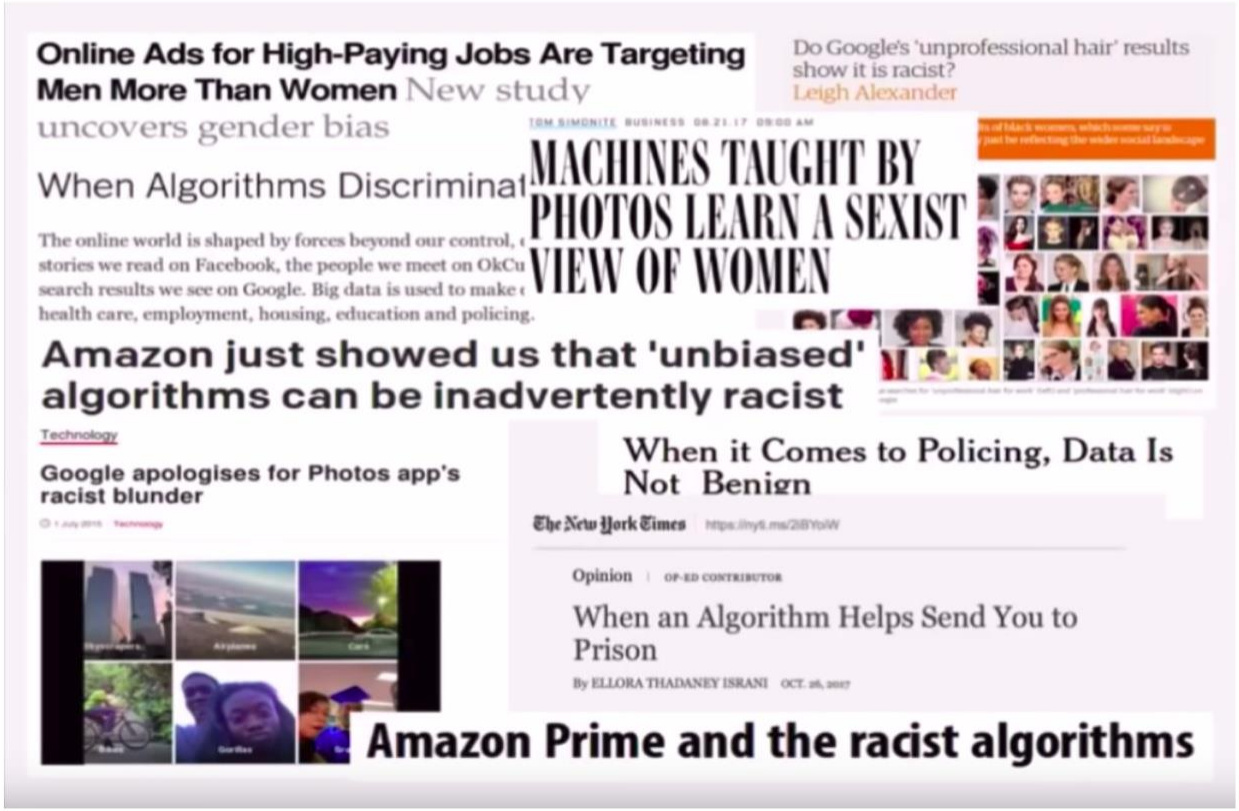
4. **Measurement Bias**

   Data collected from a local sample or device and stark differences in real-world data e.g cancer detection model in a hospital

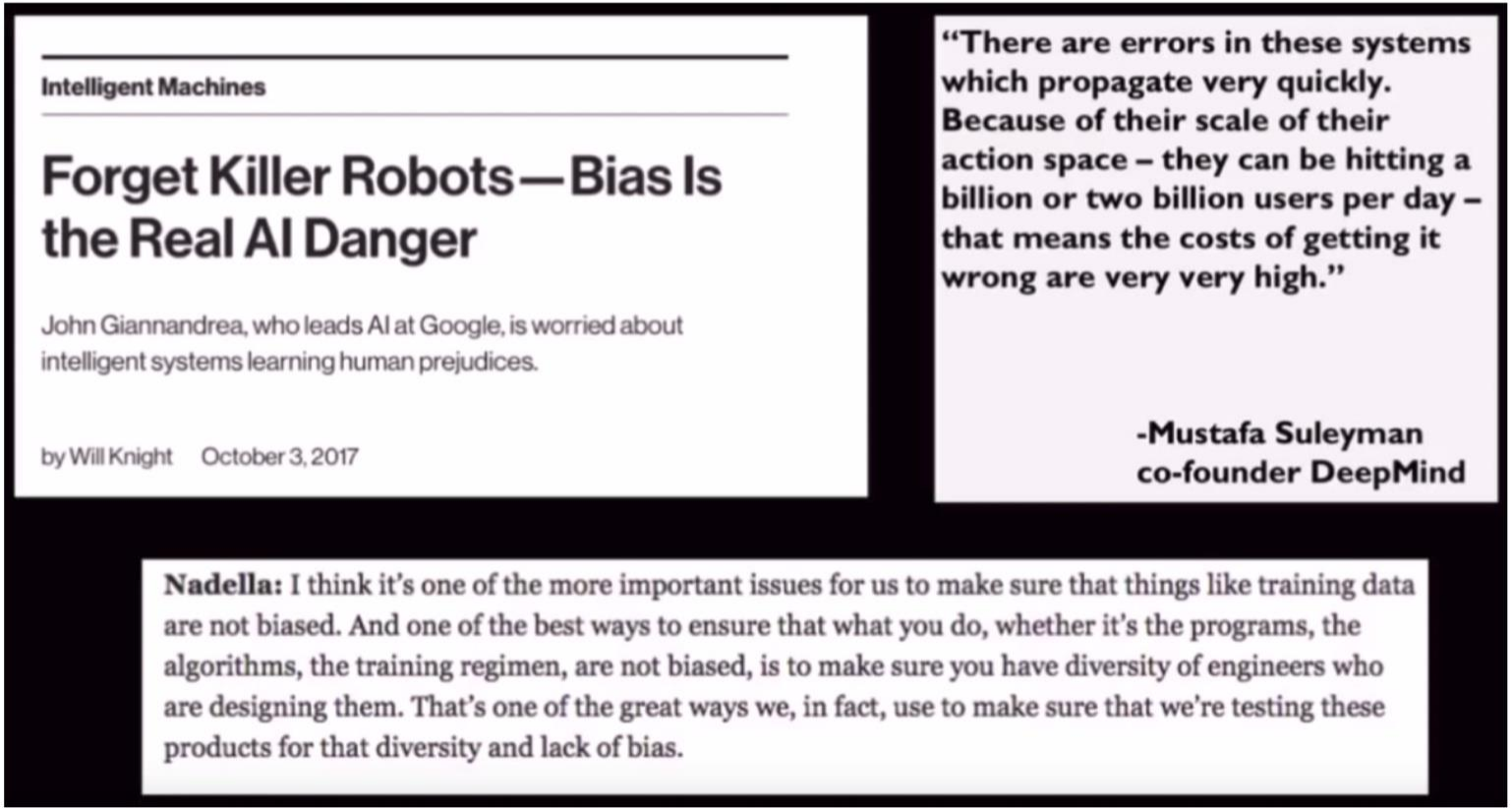
5. **Confirmation Bias**

   Happens when conscious or unconscious subjective thoughts control data annotation, sampling and modeling

The Trouble with Bias in ML Systems



The Trouble with Bias in ML Systems



Intelligent Machines

Forget Killer Robots—Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.

by Will Knight October 3, 2017

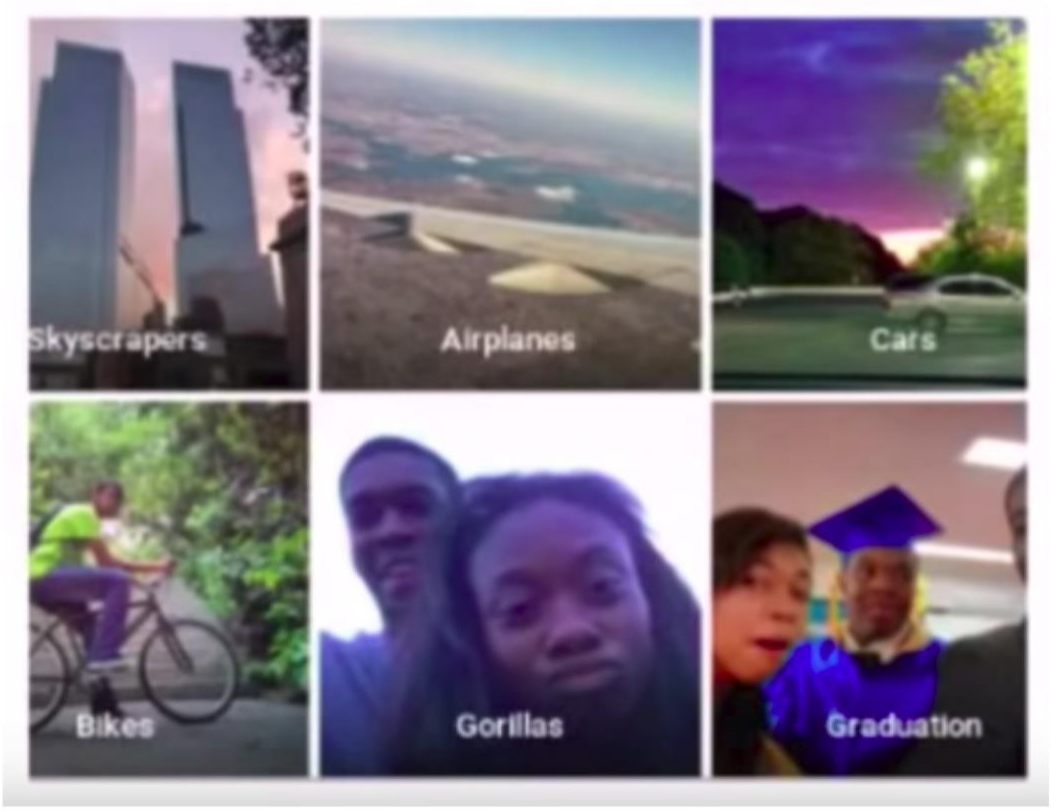
"There are errors in these systems which propagate very quickly. Because of their scale of their action space – they can be hitting a billion or two billion users per day – that means the costs of getting it wrong are very very high."

-Mustafa Suleyman
co-founder DeepMind

Nadella: I think it's one of the more important issues for us to make sure that things like training data are not biased. And one of the best ways to ensure that what you do, whether it's the programs, the algorithms, the training regimen, are not biased, is to make sure you have diversity of engineers who are designing them. That's one of the great ways we, in fact, use to make sure that we're testing these products for that diversity and lack of bias.

Sample Bias Implications

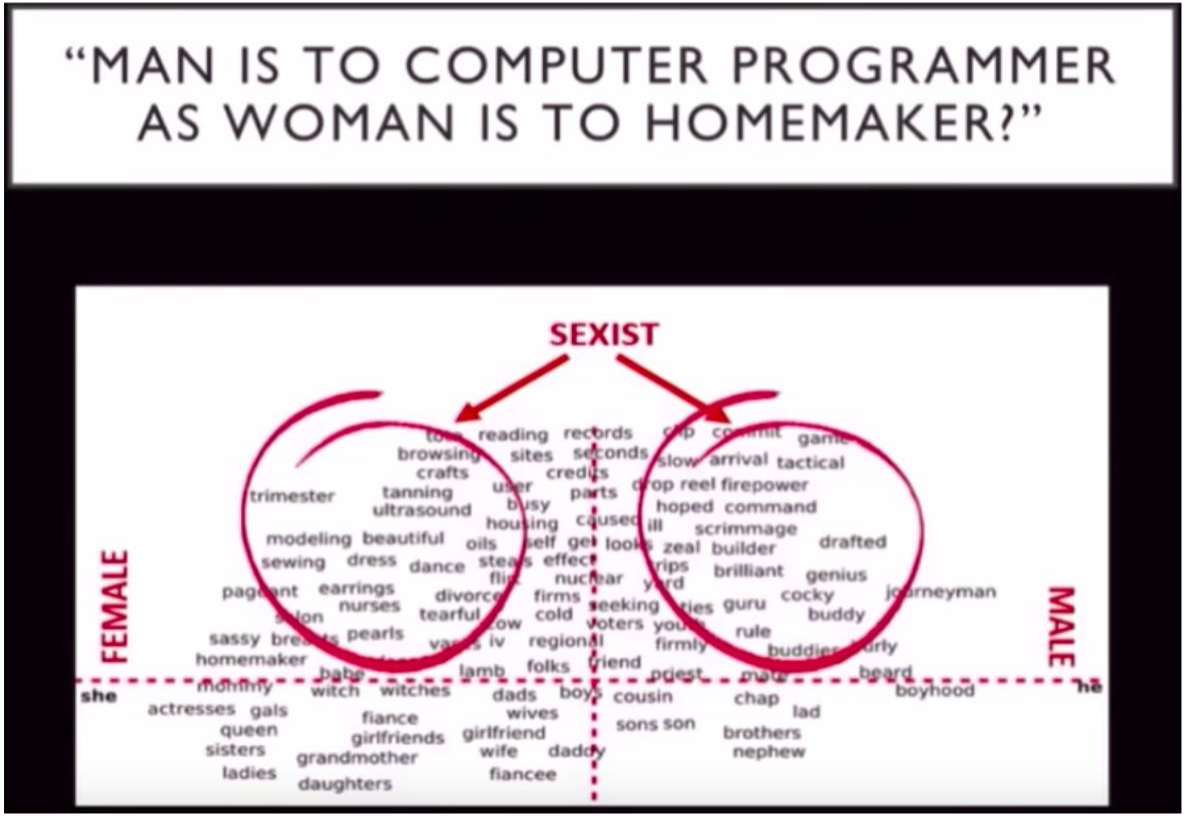
Racial Bias Implications



Sample, Association, Racial Bias Implications

Sample Bias Implications

Association Bias Implications



Association Bias Implications

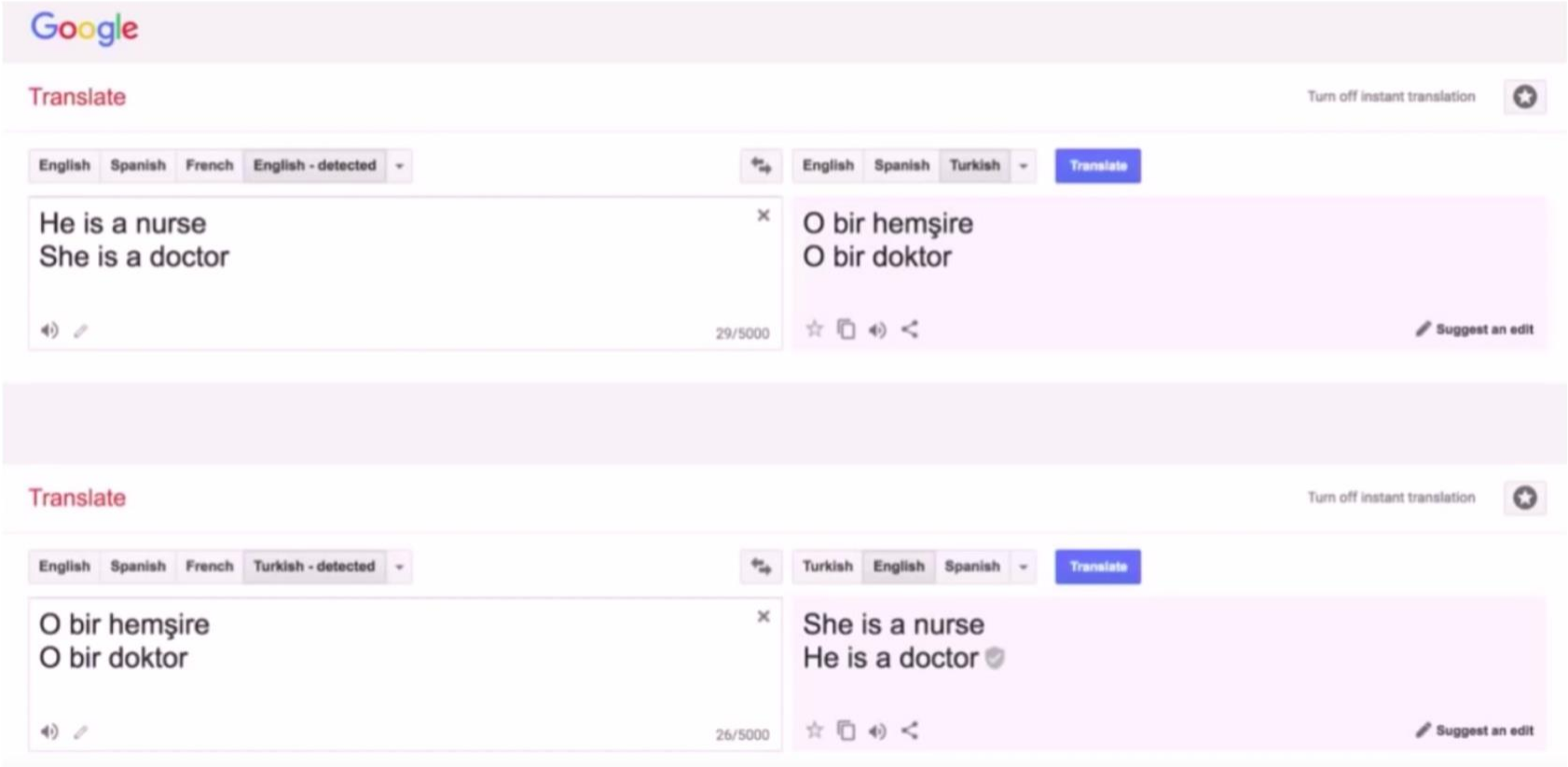
```
In [7]: model.most_similar(positive=['computer_programmer', 'woman'], negative=['man'])

Out[7]: [('homemaker', 0.5627118945121765),
         ('housewife', 0.5105047225952148),
         ('graphic_designer', 0.505180299282074),
         ('schoolteacher', 0.49794942140579224),
```

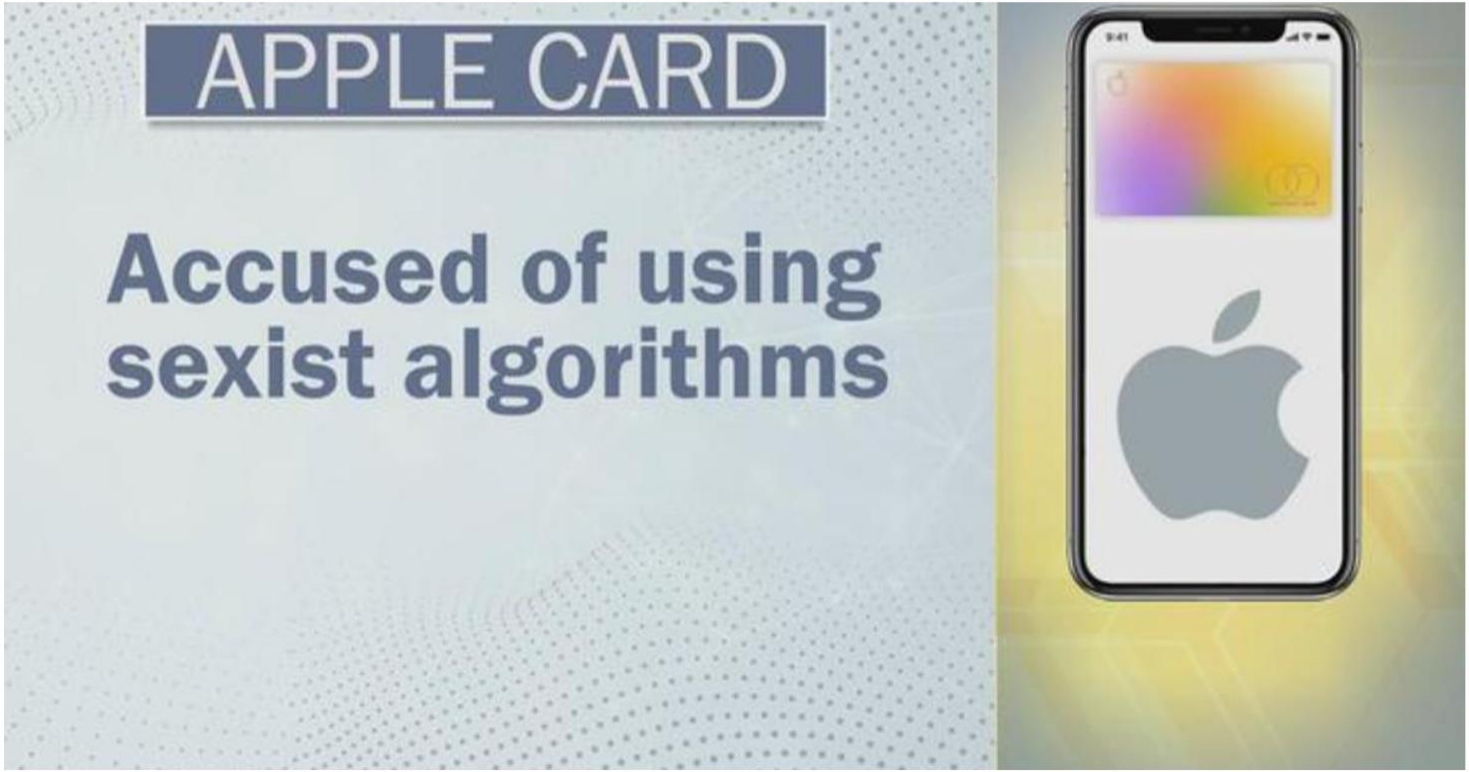
```
In [10]: model.most_similar(positive=['mexicans'], topn=30)

Out[10]: [('hispanics', 0.7345616817474365),
          ('latinos', 0.6618988513946533),
          ('ILLEGALS', 0.6574230194091797),
          ('LEGAL_immigrants', 0.6541558504104614),
          ('mexican', 0.6493428945541382),
          ('thats_ok', 0.6343405246734619),
          ('americans', 0.6324713230133057),
          ('illegals', 0.6298996210098267),
          ('ILLEGAL_aliens', 0.6289116144180298),
```

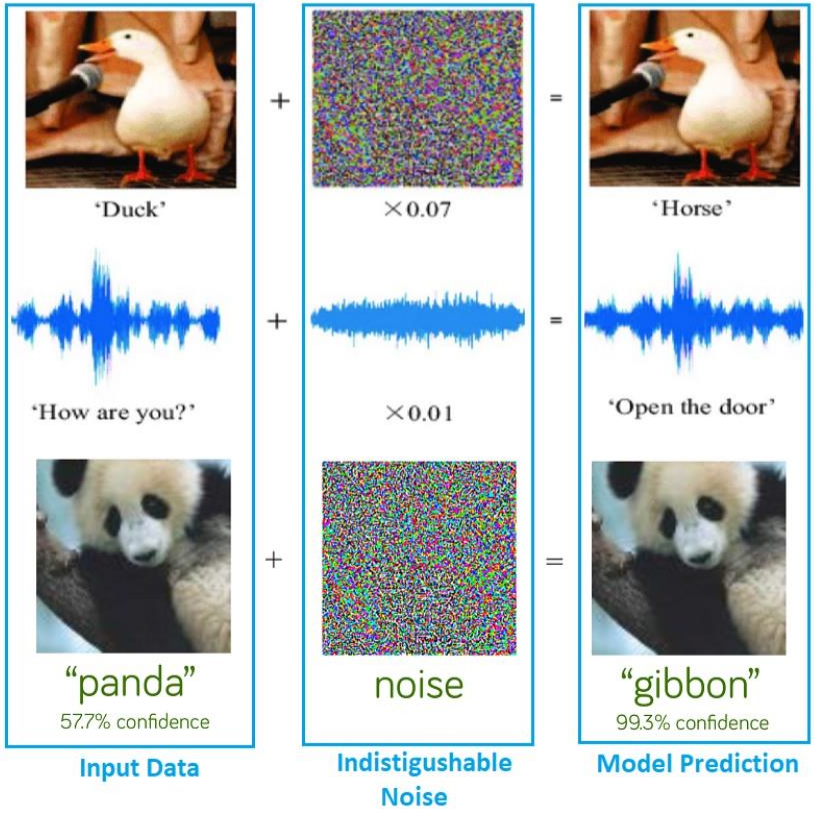
Association Bias Implications



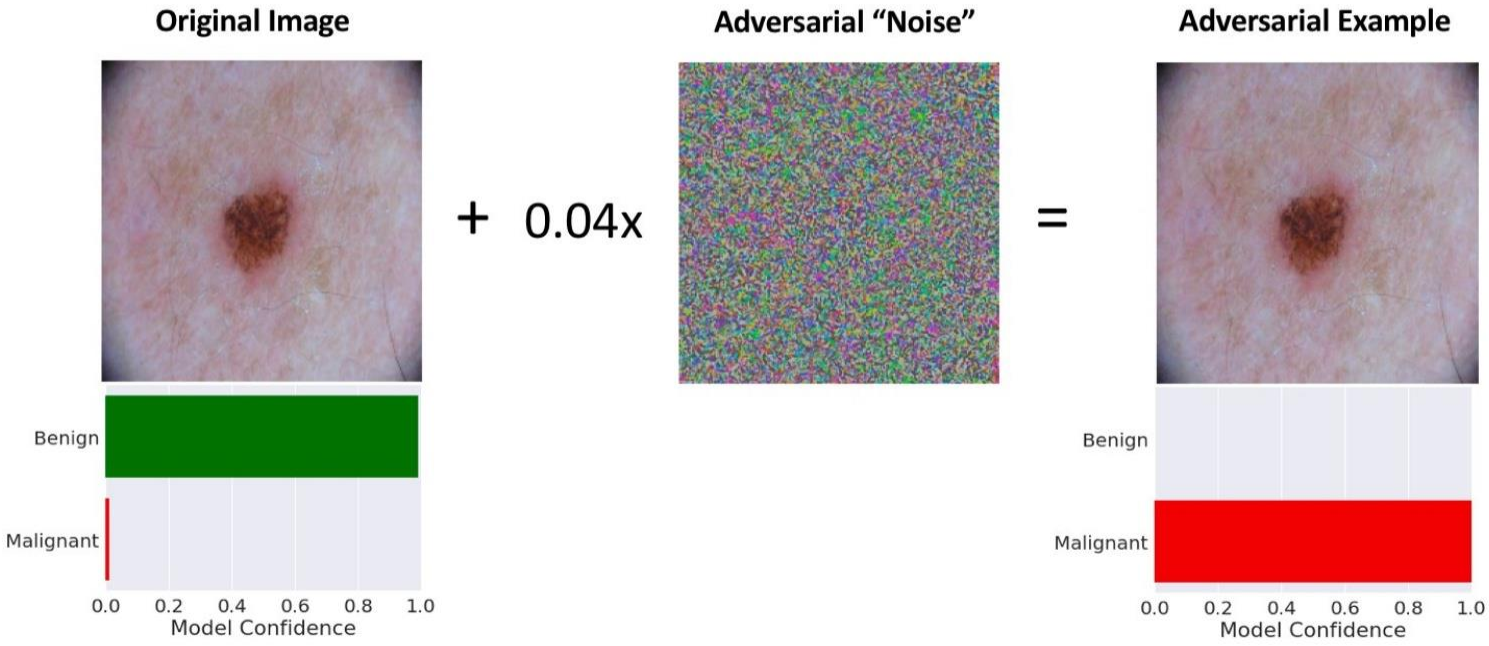
Association Bias Implications



Degrading ML Models with Adversarial Attacks



Degrading ML Models with Adversarial Attacks




Responsible AI Systems

Why Responsible AI?

1. **Potential Disruptive Impact**

   AI has the capability to automate a lot of processes and disrupt existing industries rapidly

2. **Potential Bias in Decision Making**

   Bias in data leads to bias in machine learning models and in turn biased model decisions

3. **Lack of Control over Model Decisions**

   Relying completely on black-box ML model decisions can lead to inherent problems without proper human-in-the-loop processes

4. **Lack of Privacy & Security**

   Lack of following proper guidelines to protect and anonymize sensitive user data and ML models could lead to disasters if they were breached

Responsible AI Principles

Fairness

AI systems should be fair and inclusive to everyone

Explainability

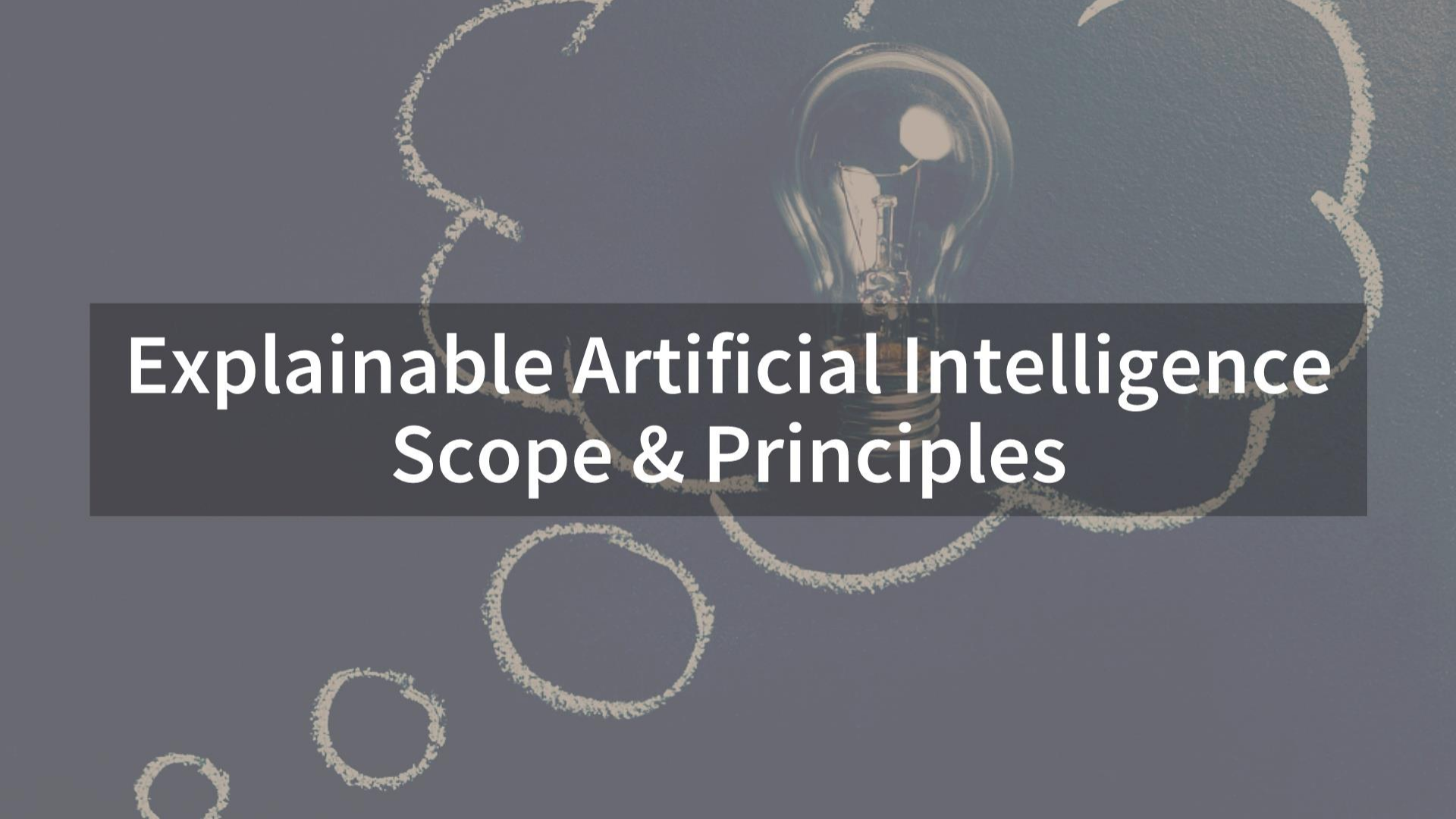
AI systems should be transparent, accountable & trustable

Privacy

AI systems should have regulations training on sensitive data

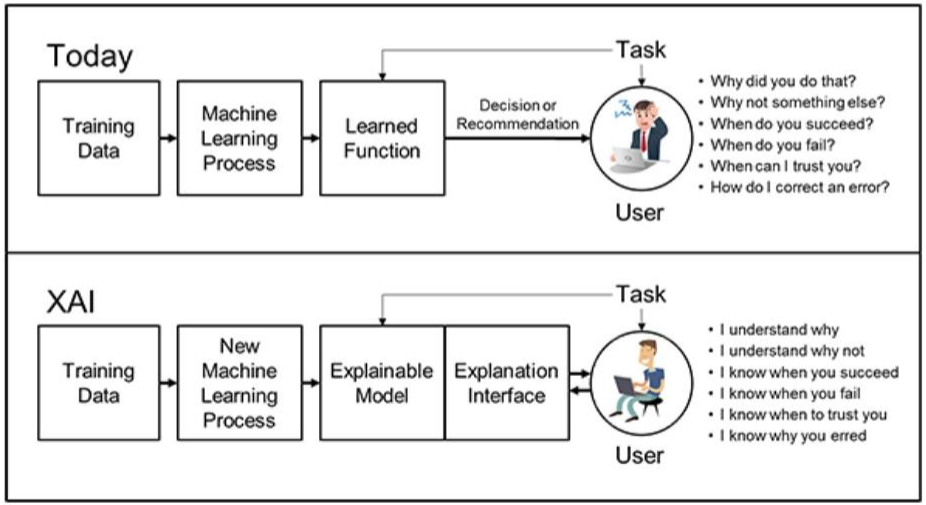
Security

AI systems should be robust to threats and attacks



Explainable Artificial Intelligence
Scope & Principles

Explainable AI (XAI) Principles



- Unbox the opacity of black-box models to make them as explainable (white-box) as possible
- Build more explainable models, while maintaining a high level of performance
- Enable human users to understand and trust model decisions to stay fair and accountable

Defining Explainability or Interpretability in XAI

- **What drives model predictions?**

  We should have the ability to find out which features are driving the overall decision-making policies of the model and key feature distributions. This ensures **fairness** of the model.

- **Why did the model take a certain decision?**

  We should also be able to discover and justify **why** certain key features were responsible in driving specific decisions taken by a model during predictions. This ensures **accountability** and **reliability** of the model.

- **How can we trust model predictions?**

  We should be able to evaluate and validate any data point and how a model takes decisions on it. This should be reproducible and demonstrable. This ensures **transparency** of the model.

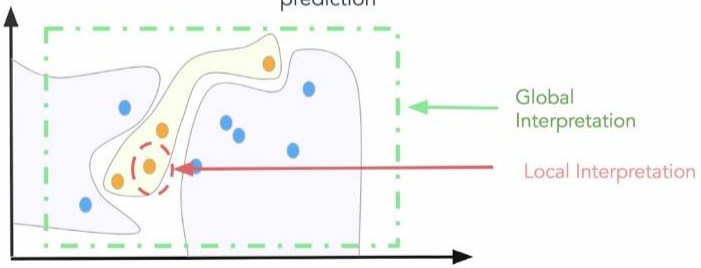
Scope of Model Interpretation

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction



- **Global Interpretations**

  This is all about trying to understand **"How does the model make predictions?"** and **"How to comprehend and interpret the whole model at once?"**

- **Local Interpretations**

  This is all about trying to understand **"Why did the model make specific decisions for a specific instance or data point?"**

Traditional Techniques for Model Interpretation

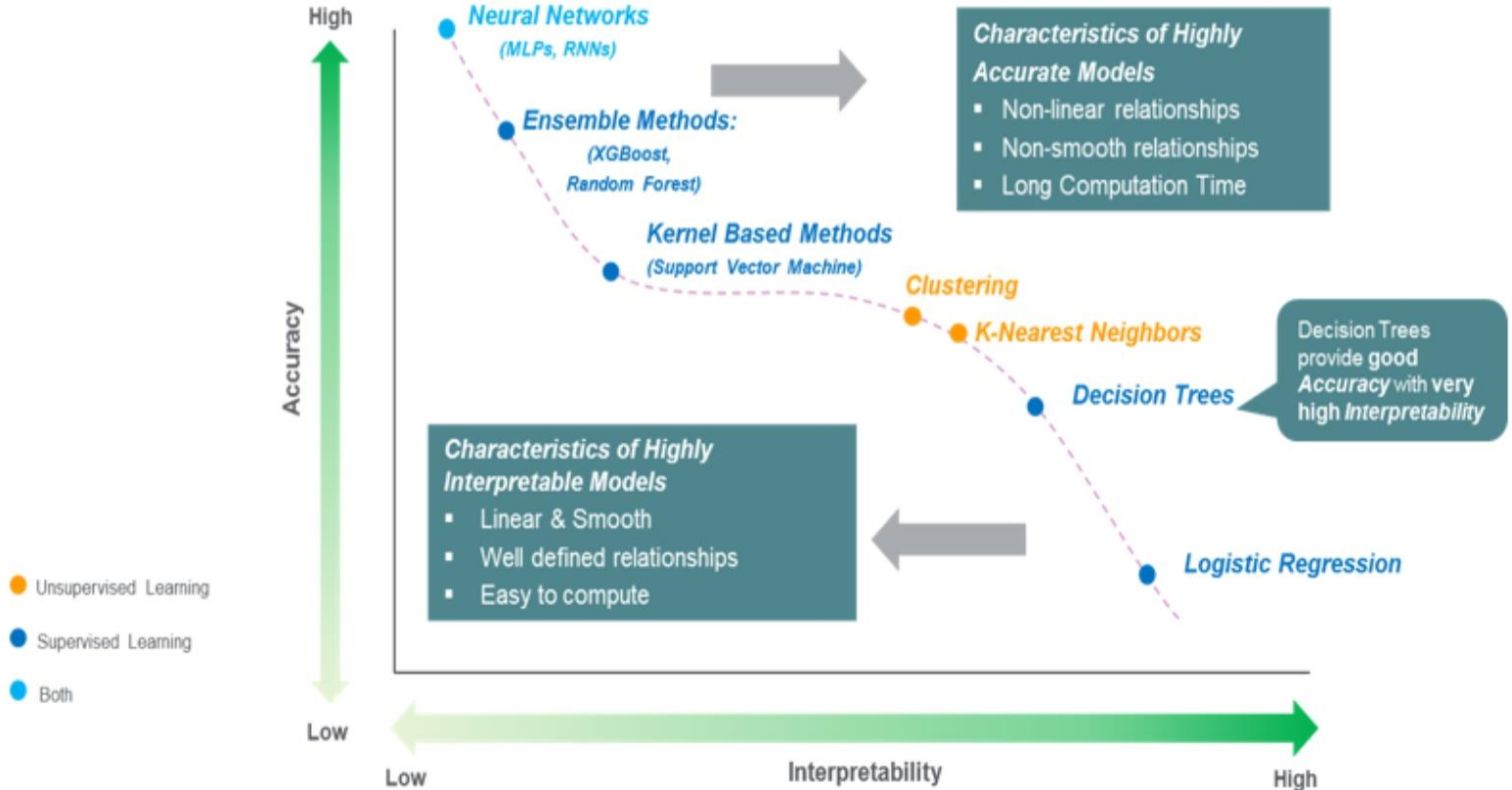
- **Exploratory Analysis and Visualization**

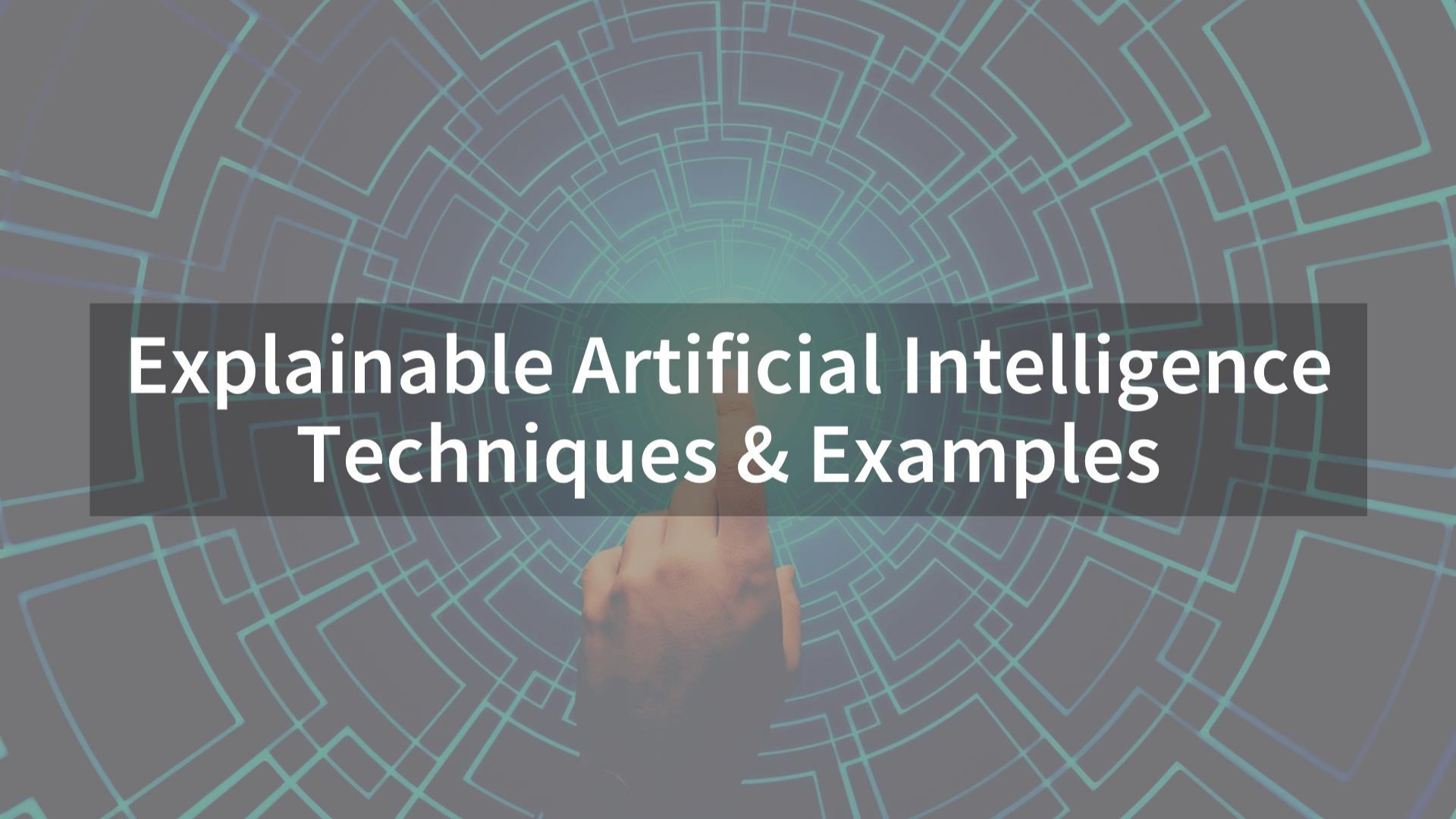
  These include techniques like clustering, dimensionality reduction, lift & gain analysis

- **Interpretable Models**

  These include linear models like linear or logistic regression and non-linear models like decision trees

The Accuracy vs. Interpretability Trade-off




Explainable Artificial Intelligence
Techniques & Examples

Techniques for Interpreting ML Models - Structured Data

1. **Using Interpretable Models**

   Use models which are interpretable like linear models or decision trees or RuleFit models

2. **Model Feature Importances**

   Feature importance is generic term for the degree to which a predictive model relies on a particular feature. This can be model specific or model agnostic

3. **Partial Dependence Plots**

   Partial Dependence describes the average marginal impact of a feature on model prediction, holding other features in the model constant and perturbing the feature value

4. **Individual Conditional Expectation Plots**

   An ICE plot visualizes the dependence of the prediction on a feature for each instance separately, resulting in one line per instance, compared to one line overall in partial dependence plots

5. **Global Surrogate Models**

   Model-agnostic surrogate models approximate the predictions of the underlying model as closely as possible while being interpretable by fitting interpretable models on the source model predictions

6. **Local Interpretable Model-agnostic Explanations (LIME)**

   LIME focuses on fitting local surrogate models to explain how single prediction decisions were made.

7. **Shapley Values and SHapley Additive exPlanations (SHAP)**

   SHAP values try to explain the output of a model (function) as a sum of the effects of each feature being introduced into a conditional expectation (avg over different orderings)

Model Agnostic XAI Principles

- **LIME and SHAP are examples of model agnostic, additive feature attribution techniques**
- **Additive feature attribution principles:**
  - Can decompose a model prediction such that it is a linear sum of individual contributions of features

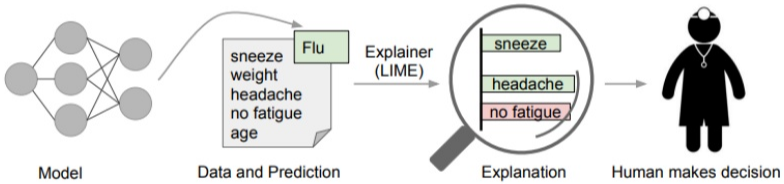
Model Agnostic XAI Principles

- **Additive decomposition should work for non-linear complex models too to be truly model agnostic**
- **If our model's objective function *f* is non-linear then global decomposition is impossible**
  - Focus is on local approximation of ***f*** around a sample of interest to be explained ***i*** using a linear function - locally faithful model
  - Local interpretation focuses on interpreting or explaining a specific sample of interest **- *i***

Model Agnostic XAI Principles

- If our original feature space *x* is too complex, a feature transformation can be optionally done before interpretations
- Focus then lies on approximating *f* around a sample *i* with a linear approximation model *g* in a simpler, transformed feature space

LIME Methodology



Loss function

Regularizer

$$\xi = \underset{g \in \mathcal{G}}{\arg\min}\ L(f, g, \pi_{x'}) + \Omega(g)$$

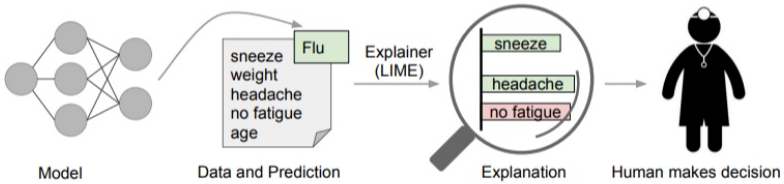
Local kernel

LIME Objective

- LIME focuses on training local surrogate models to explain individual predictions
- LIME tests what happens to the predictions when you give variations of your data into the machine learning model
- LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model
- LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest
- The learned model should be a good approximation of the machine learning model predictions locally

https://christophm.github.io/interpretable-ml-book/lime.html#lime-for-tabular-data

LIME Methodology



Loss function

Regularizer

$$\xi = \arg\min_{g \in \mathcal{G}} L(f, g, \pi_{x'}) + \Omega(g)$$

Local kernel

LIME Objective

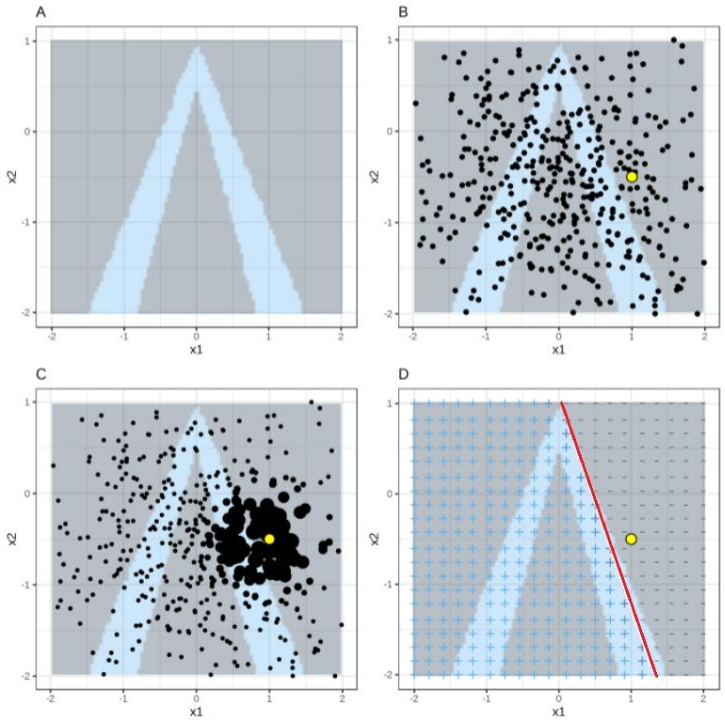
- The explanation model for instance *x* is the model *g* (e.g. linear regression model) that minimizes loss *L* (e.g. mean squared error)
- The loss *L* measures how close the explanation is to the prediction of the original model *f* (e.g. a gradient boosting model)
- Model complexity Ω(g) is based on the number of features
- The proximity measure *πx* defines how large the neighborhood around instance *x* do we consider for the explanation

https://christophm.github.io/interpretable-ml-book/lime.html#lime-for-tabular-data

LIME Workflow

1. Train black box (usually complex non-linear) model on data
2. Choose your instance of interest for which you want to explain the prediction from the black box model
3. Perturb your dataset to generate data points around the instance of interest
4. Get black box model predictions for these points
5. Weigh the samples by their proximity to the instance of interest e.g using an exponential smoothing kernel
6. Fit a weighted, interpretable linear approximation model on this dataset with variations
7. Explain predictions by interpreting the local model

LIME on Tabular Data - Example



- **A:** Black-box model (random forest) decision boundary for binary classification
- **B:** Generate perturbed samples around point of interest (yellow point)
- **C:** Use kernel function to weight samples based on proximity to point of interest
- **D:** Use a local linear approximation model to generate local decision boundary and interpretations

https://christophm.github.io/interpretable-ml-book/lime.html#lime-for-tabular-data

LIME on Text Data - Example

Spam Classification Dataset

	CONTENT	CLASS
267	PSY is a good guy	0
173	For Christmas Song visit my channel! ;)	1

Sample perturbations around point of interest

	For	Christmas	Song	visit	my	channel!	;)	prob	weight
2	1	0	1	1	0	0	1	0.17	0.57
3	0	1	1	1	1	0	1	0.17	0.71
4	1	0	0	1	1	1	1	0.99	0.71
5	1	0	1	1	1	1	1	0.99	0.86
6	0	1	1	1	0	0	1	0.17	0.57

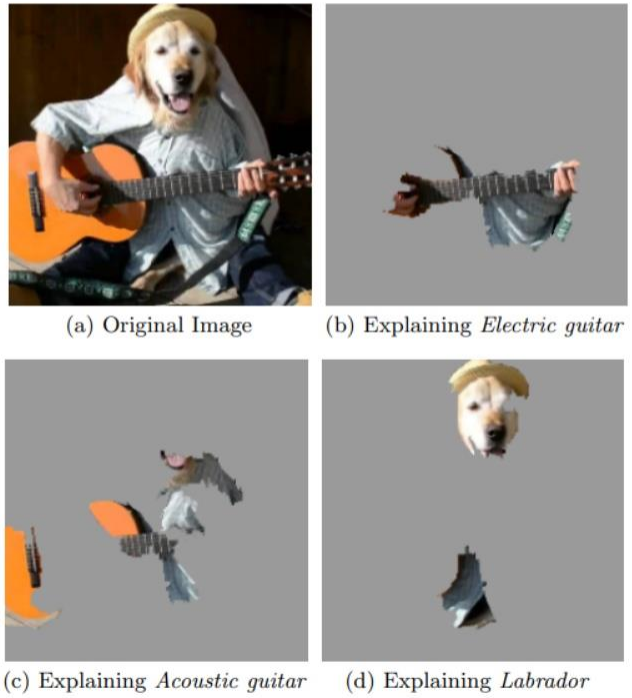
Estimate feature importances with local linear approximation model

case	label_prob	feature	feature_weight
1	0.1701170	is	0.000000
1	0.1701170	good	0.000000
1	0.1701170	a	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	Christmas	0.000000

https://christophm.github.io/interpretable-ml-book/lime.html

- **Select data point of interest**
- **Generate perturbed samples around point of interest**
  - Each row is a variation
- **Generate black box model predictions and weigh samples**
  - **"prob"** column shows the predicted probability of spam for each of the sentence variations
  - **"weight"** column shows the proximity of the variation to the original sentence, calculated as 1 minus the proportion of words that were removed
- **Use a local linear approximation model to generate local decision boundary and interpretations**
  - The word **"channel"** indicates a high probability of spam

LIME on Image Data - Example



(a) Original Image (b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

- We don't perturb individual pixels, since many more than one pixel contribute to one class
- Randomly changing individual pixels would probably not change the predictions by much
- Variations of images are created by segmenting the image into "superpixels" and turning them off or on
  - Superpixels are interconnected pixels with similar colors and can be turned off by replacing each pixel with a user-defined color such as gray
  - The user can also specify a probability for turning off a superpixel in each permutation

"Why Should I Trust You?": Explaining the Predictions of Any Classifier, S. Singh, et al.
https://christophm.github.io/interpretable-ml-book/lime.html

Shapley Values Methodology

Lloyd Shapley

Nobel Prize in 2012

- **The Shapley value, coined by Shapley, is a method for assigning payouts to players depending on their contribution towards the total payout**
- **Players cooperate in a coalition and obtain a certain gain from that cooperation**
  - The **'game'** is the prediction task for a single instance of the dataset
  - The **'gain'** is the actual prediction for this instance minus the average prediction of all instances
  - The **'players'** are the feature values of the instance, which collaborate to receive the gain
- **The Shapley value of a feature is the average marginal contribution of a feature value over all possible coalitions**
  - Coalitions are basically combinations of features which are used to estimate the shapley value of a specific feature

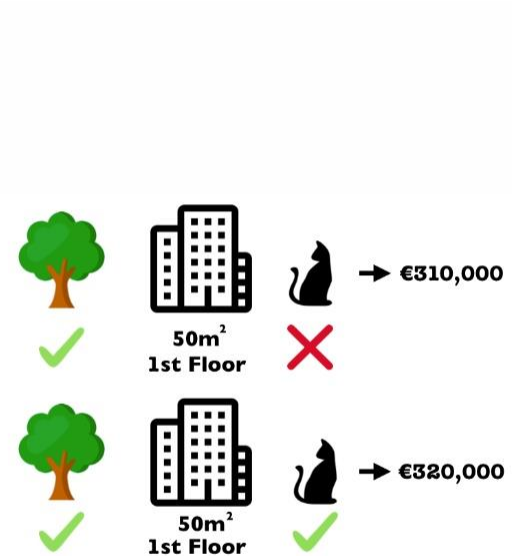
https://christophm.github.io/interpretable-ml-book/shapley.html
https://www.youtube.com/watch?v=-taOhqkiuIo

Shapley Values Example

- **The predicted price for a 50 m2, 2nd floor apartment with a nearby park and cat ban is €300,000**
  - Goal is to explain how each of these feature values contributed to the prediction
- **How much has each feature value contributed to the prediction compared to the average prediction**
  - The average prediction for all apartments is €310,000
  - Feature values park-nearby, cat-banned, area-50 and floor-2nd worked together to achieve the prediction of €300,000
  - Goal is to explain the difference between the actual prediction (€300,000) and the average prediction (€310,000): a difference of -€10,000
- **One possible answer: The park-nearby contributed €30,000; size-50 contributed €10,000; floor-2nd contributed €0; cat-banned contributed -€50,000**
  - The contributions add up to -€10,000, the final prediction minus the average predicted apartment price

https://christophm.github.io/interpretable-ml-book/shapley.html
https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739

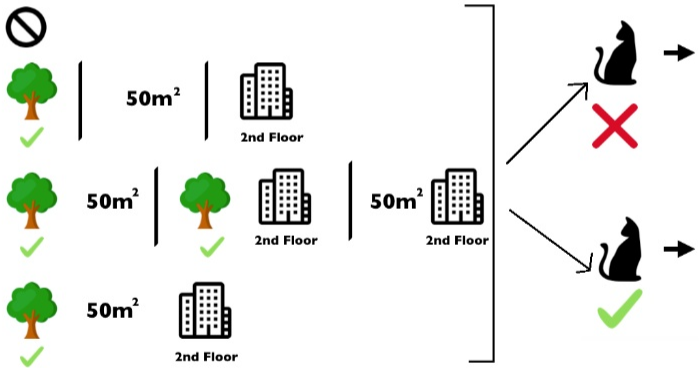
Shapley Values Example



- **Shapley value is the average marginal contribution of a feature value across all possible coalitions**
- **We evaluate the contribution of the cat-banned feature value when it is added to a coalition of park-nearby and size-50**
  - We simulate that only park-nearby, cat-banned and size-50 are in a coalition
  - The value floor-2nd was replaced by the randomly drawn floor-1st from the dataset
  - We predict the price of the apartment with this combination (€310,000)
  - We remove cat-banned from the coalition by replacing it with cat-allowed (random sample)
  - We predict the apartment price for the coalition of park-nearby and size-50 (€320,000)
  - The contribution of cat-banned was €310,000 - €320,000 = -€10.000
  - This estimate depends on the values of the randomly drawn apartment that served as a “donor” for the cat and floor feature values
  - We will get better estimates if we repeat this sampling step and average the contributions

https://christophm.github.io/interpretable-ml-book/shapley.html
https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739

Shapley Values Example

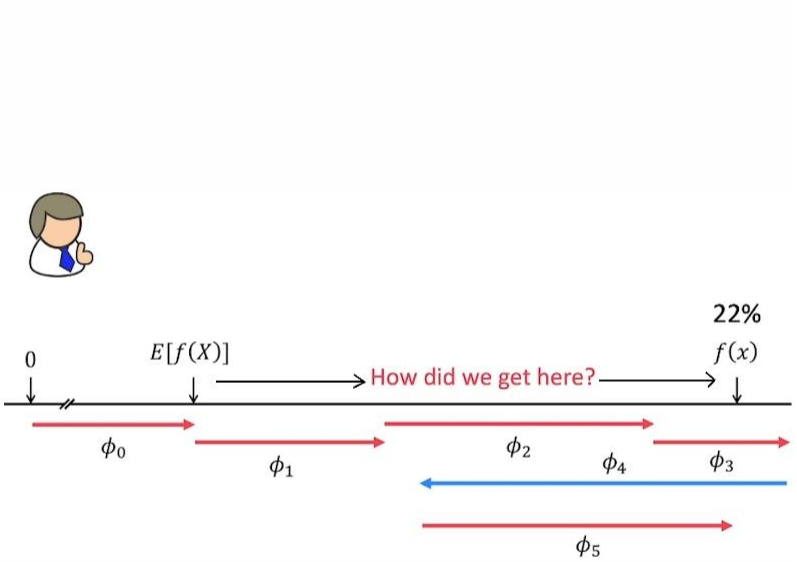


- The Shapley value is the average of all the marginal contributions to all possible coalitions
  - The computation time increases exponentially with the number of features
- Figure shows all coalitions of feature values that are needed to determine the Shapley value for cat-banned
  - No feature values
  - park-nearby
  - size-50
  - floor-2nd
  - park-nearby+size-50
  - park-nearby+floor-2nd
  - size-50+floor-2nd
  - park-nearby+size-50+floor-2nd
- For each of these coalitions we compute the predicted apartment price with and without the feature value cat-banned and take the difference to get the marginal contribution
- The Shapley value is the (weighted) average of marginal contributions

https://christophm.github.io/interpretable-ml-book/shapley.html

https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739

SHAP Methodology

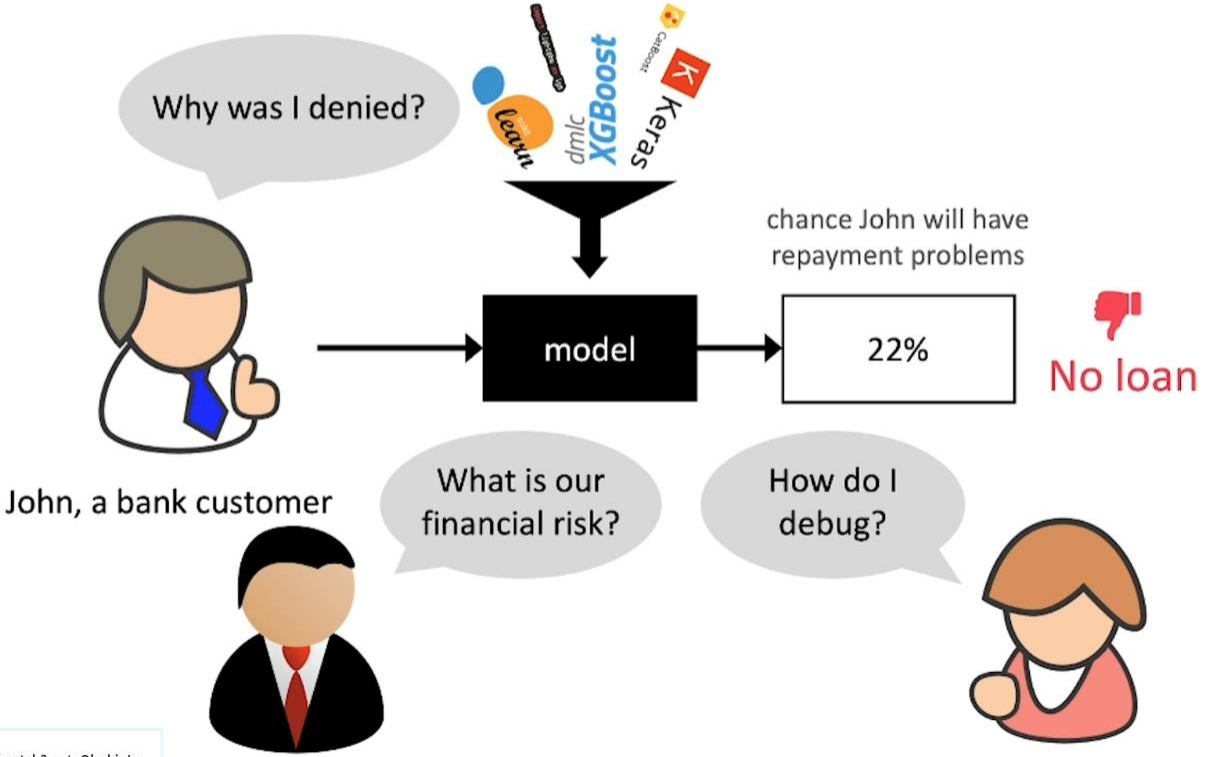


- SHAP (SHapley Additive exPlanations) is an enhancement on the Shapley values
- SHAP assigns each feature an importance value for a particular prediction. Its novel components include:
  - Identification of a new class of additive feature importance measures
  - Results showing there is a unique solution in this class with a set of desirable properties
- SHAP values try to explain the output of a model (function) as a sum of the effects of each feature being introduced into a conditional expectation
- Importantly, for non-linear functions the order in which features are introduced matters
- The SHAP values result from averaging over all possible orderings

https://christophm.github.io/interpretable-ml-book/shap.html

https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739

SHAP Example



Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

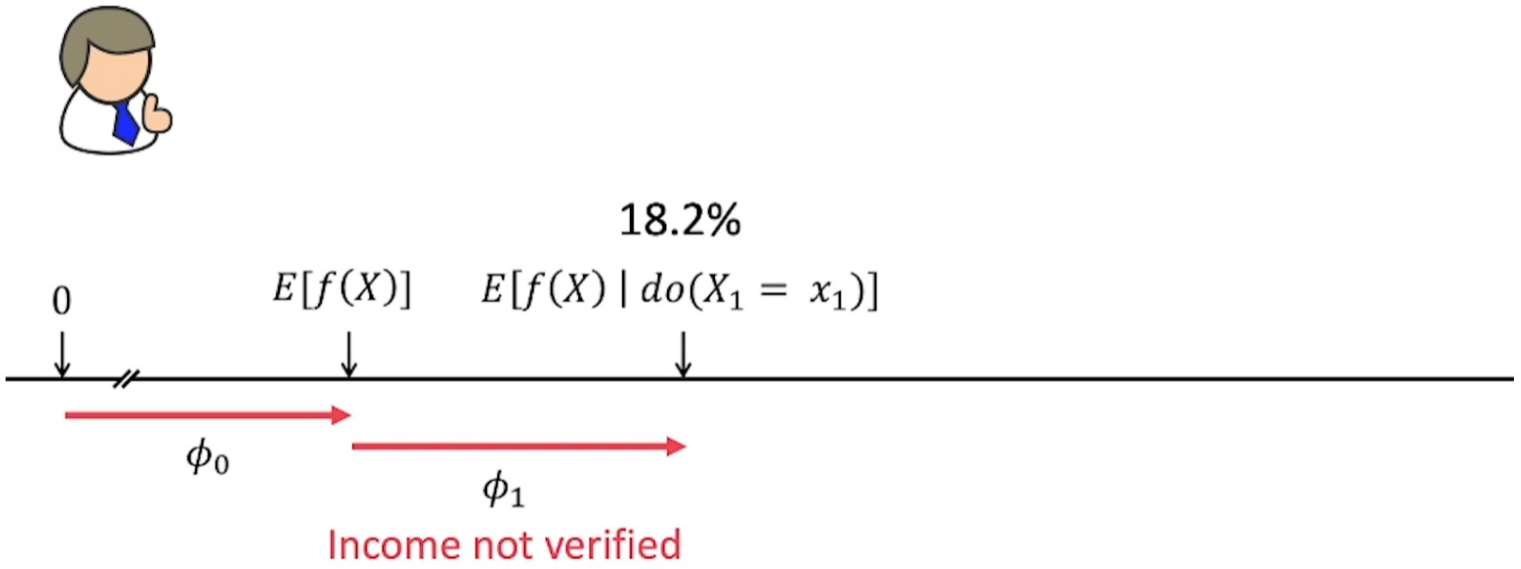
SHAP Example

Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

SHAP Example

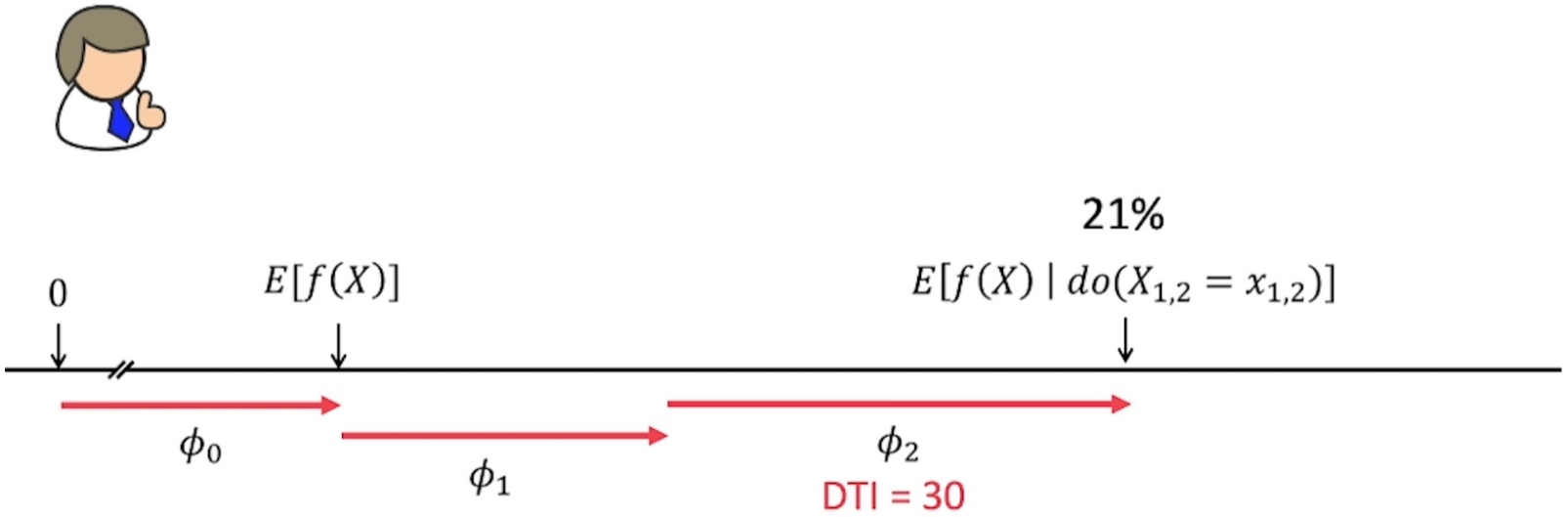
Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

SHAP Example



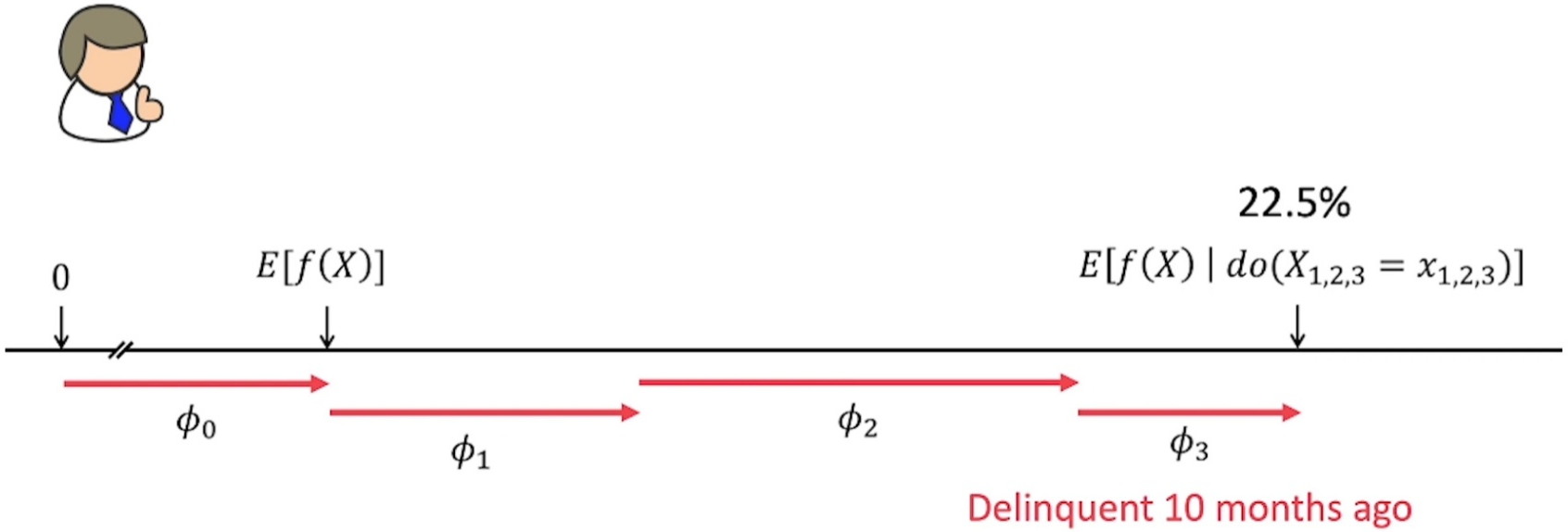
Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

SHAP Example



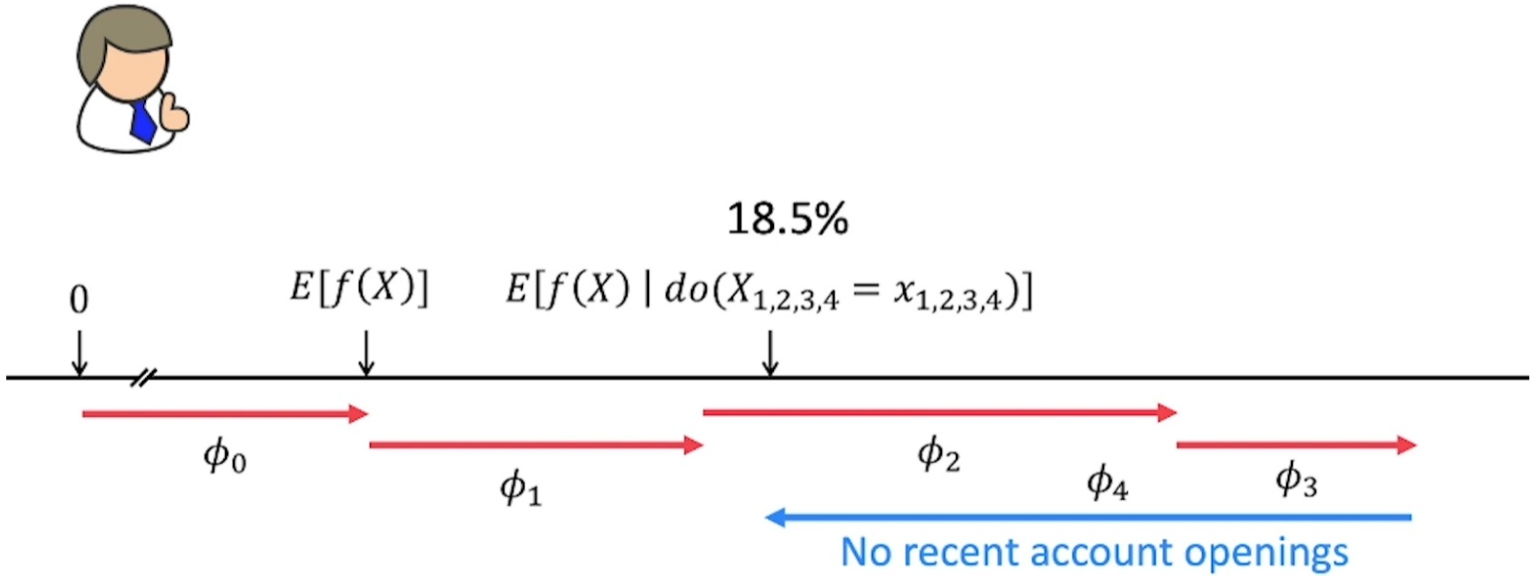
Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

SHAP Example



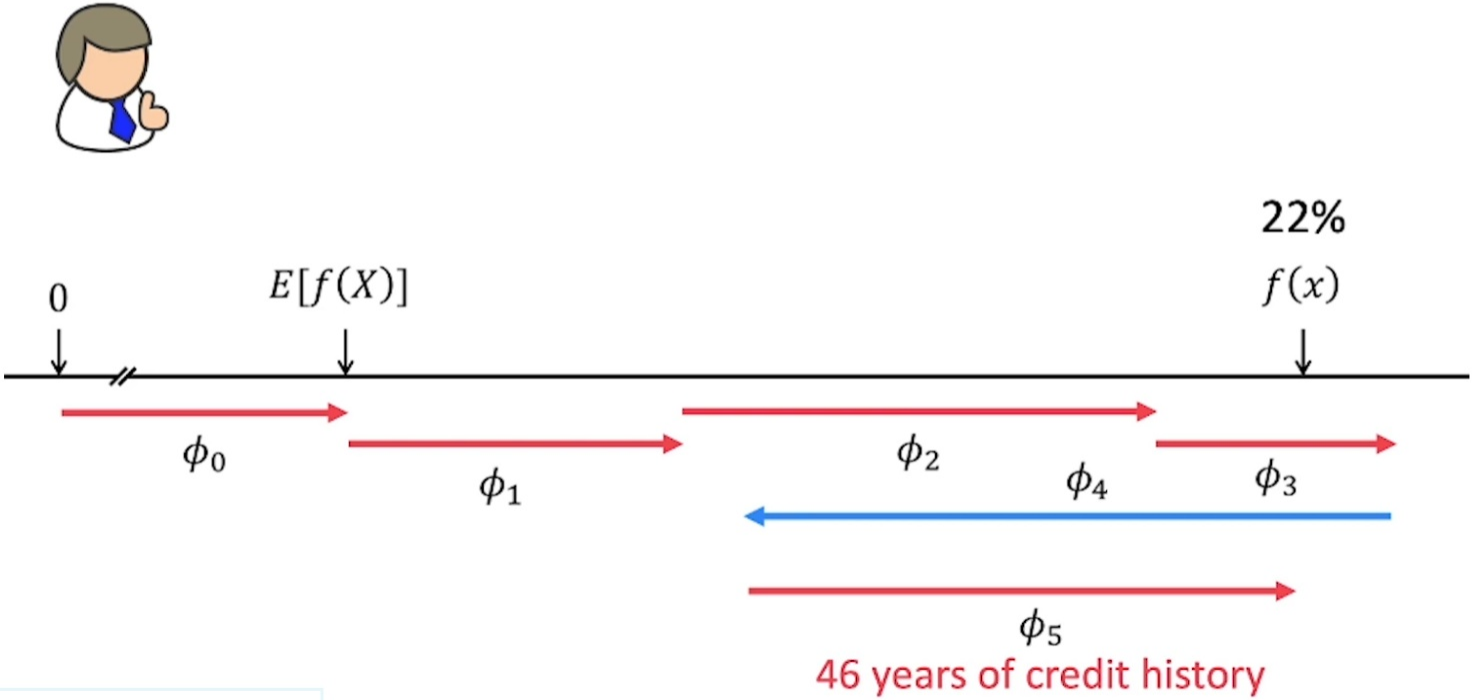
Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

SHAP Example



Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

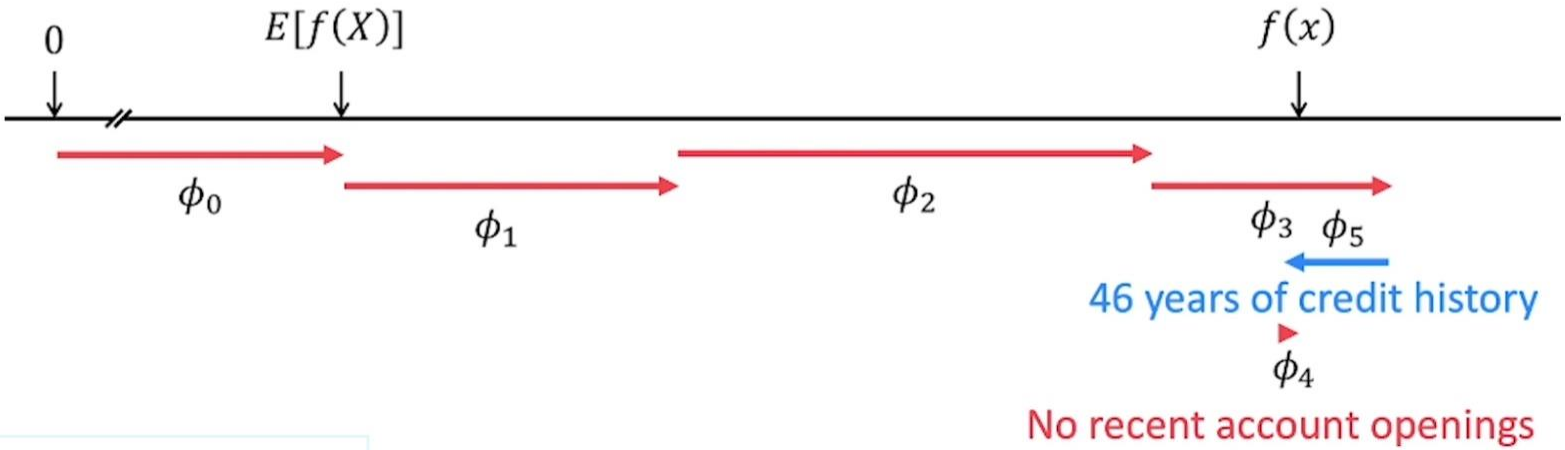
SHAP Example



Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

SHAP Example

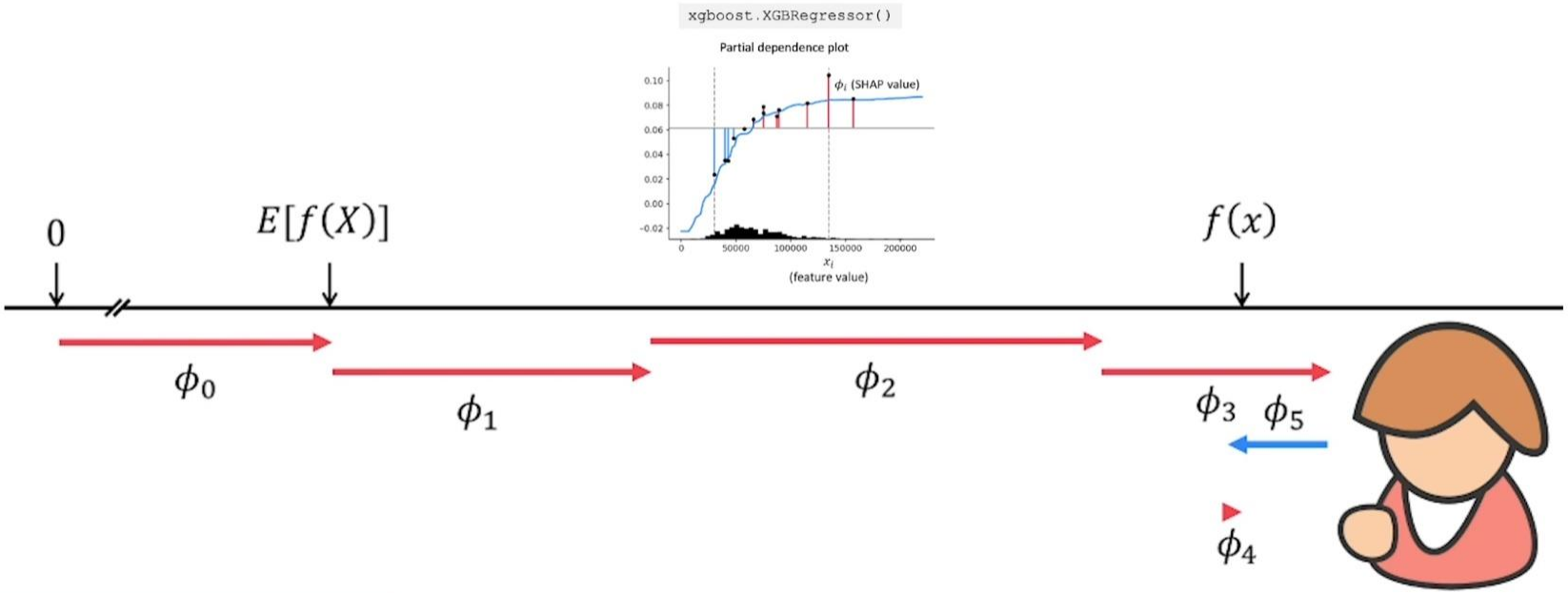
The order matters!



Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

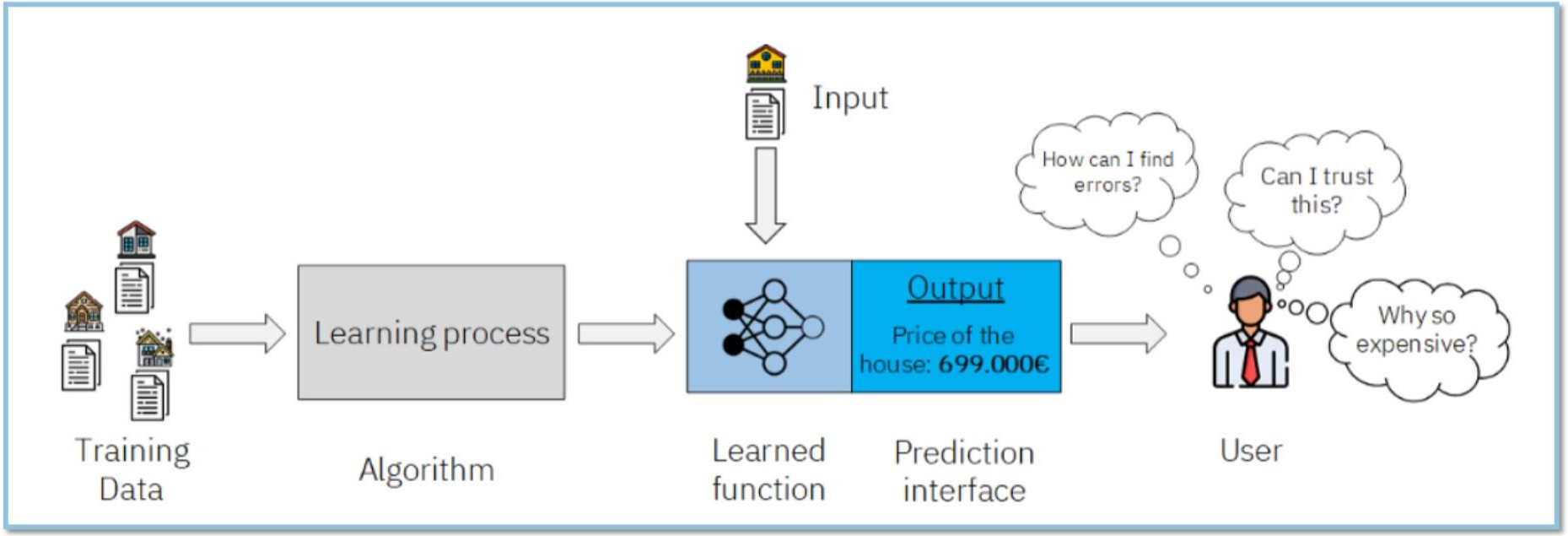
SHAP Example

Shapley values result from **averaging over all N! possible orderings.**

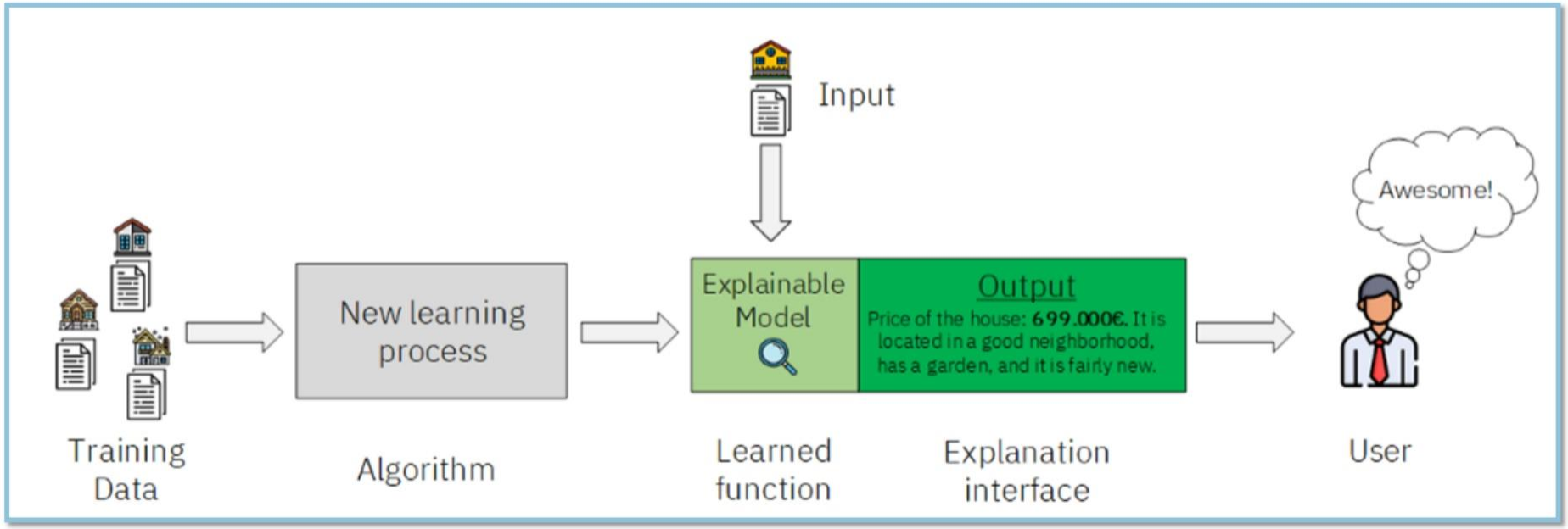


Science behind InterpretML: https://www.youtube.com/watch?v=-taOhqkiuIo

House Price Prediction - Standard ML Workflow



House Price Prediction - XAI Workflow



House Price Prediction - Dataset Details

Boston house prices dataset

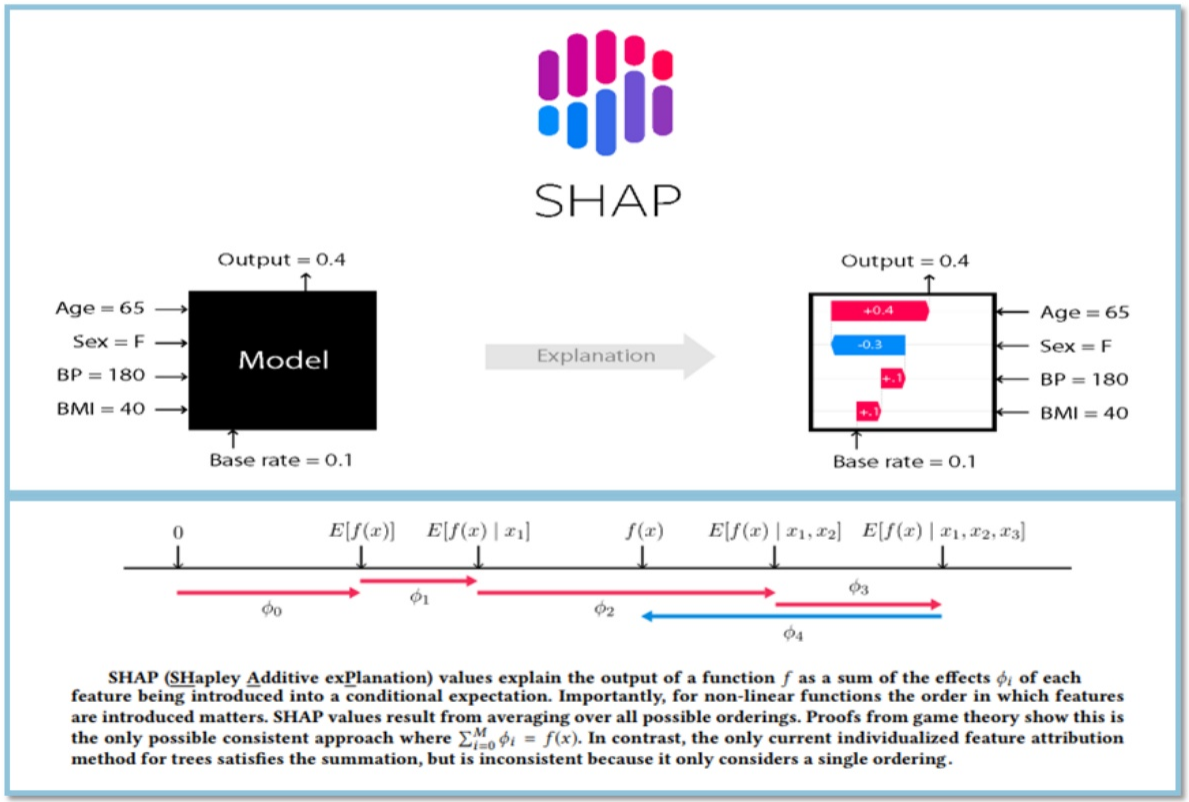
Data Set Characteristics:

Number of Instances:	506
Number of Attributes:	13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.
Attribute Information (in order):	• CRIM per capita crime rate by town • ZN proportion of residential land zoned for lots over 25,000 sq.ft. • INDUS proportion of non-retail business acres per town • CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) • NOX nitric oxides concentration (parts per 10 million) • RM average number of rooms per dwelling • AGE proportion of owner-occupied units built prior to 1940 • DIS weighted distances to five Boston employment centres • RAD index of accessibility to radial highways • TAX full-value property-tax rate per $10,000 • PTRATIO pupil-teacher ratio by town • B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town • LSTAT % lower status of the population • MEDV Median value of owner-occupied homes in $1000's
Missing Attribute Values:	None
Creator:	Harrison, D. and Rubinfeld, D.L.

Sample Data:

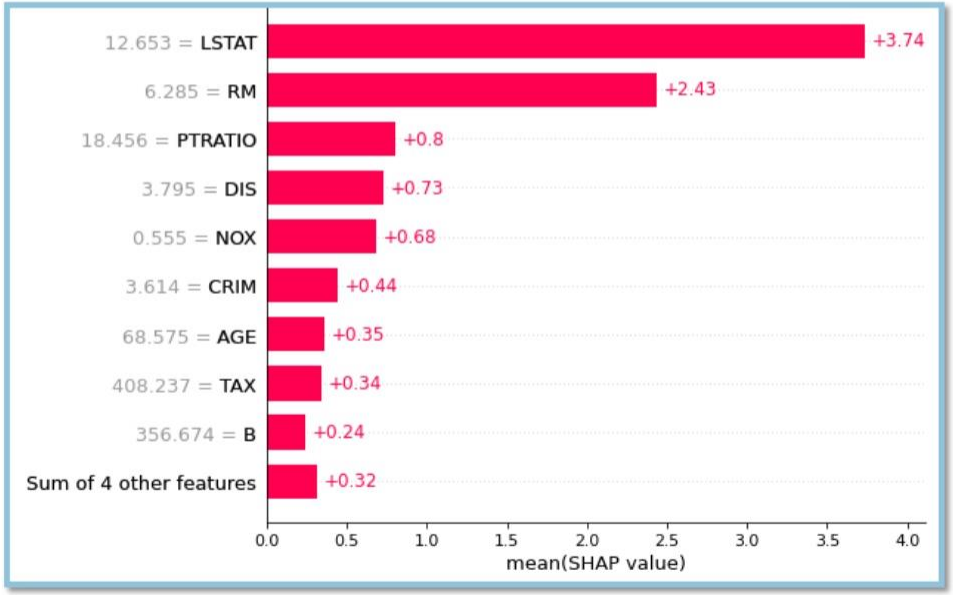
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2
5	0.02985	0.0	2.18	0.0	0.458	6.430	58.7	6.0622	3.0	222.0	18.7	394.12	5.21	28.7
6	0.08829	12.5	7.87	0.0	0.524	6.012	66.6	5.5605	5.0	311.0	15.2	395.60	12.43	22.9
7	0.14455	12.5	7.87	0.0	0.524	6.172	96.1	5.9505	5.0	311.0	15.2	396.90	19.15	27.1
8	0.21124	12.5	7.87	0.0	0.524	5.631	100.0	6.0821	5.0	311.0	15.2	386.63	29.93	16.5
9	0.17004	12.5	7.87	0.0	0.524	6.004	85.9	6.5921	5.0	311.0	15.2	386.71	17.10	18.9

House Price Prediction – XAI Interface with SHAP



SHAP (SHapley Additive exPlanation) values explain the output of a function f as a sum of the effects ϕ_i of each feature being introduced into a conditional expectation. Importantly, for non-linear functions the order in which features are introduced matters. SHAP values result from averaging over all possible orderings. Proofs from game theory show this is the only possible consistent approach where $\sum_{i=0}^{M} \phi_i = f(x)$. In contrast, the only current individualized feature attribution method for trees satisfies the summation, but is inconsistent because it only considers a single ordering.

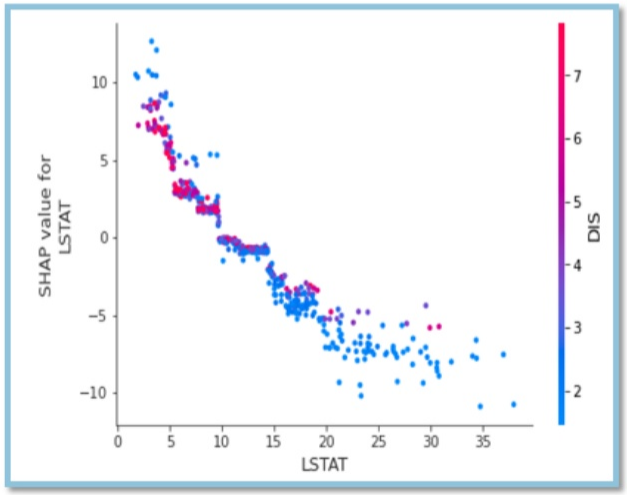
House Price Prediction – Global Interpretation



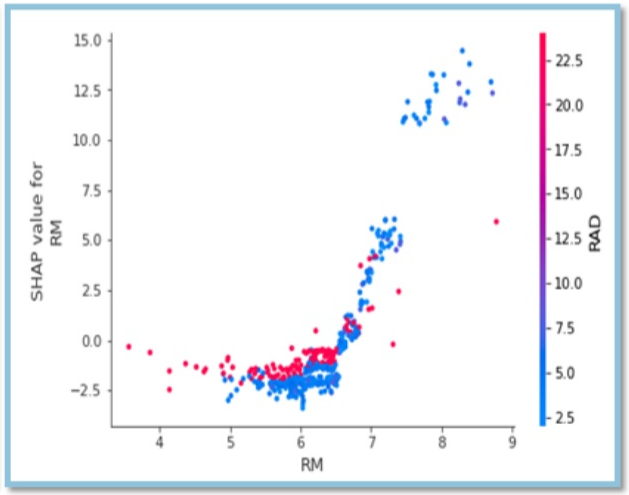
Top features driving house prices up globally

- LSTAT: % Lower Status of Population
- RM: Average Rooms per Dwelling
- PTRATIO: Pupil-Teacher Ratio by Town

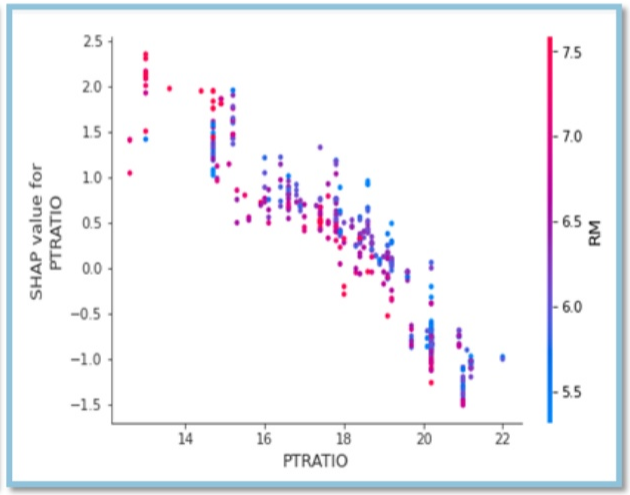
House Price Prediction – Dependence Plots



High LSTAT drives Prices Down



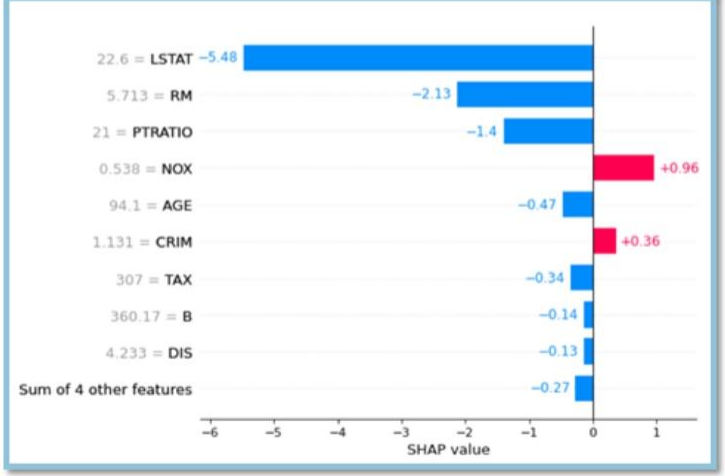
High RM drives Prices Up



High PTRATIO drives Prices Down

House Price Prediction – Local Interpretation

```
Actual House Price ($1000s: 12.7
Predicted House Price ($1000s): 13.480962
Key Features Affecting Prediction:
```



Local Important Features

Features Driving Prediction Prices

XAI Interpretation

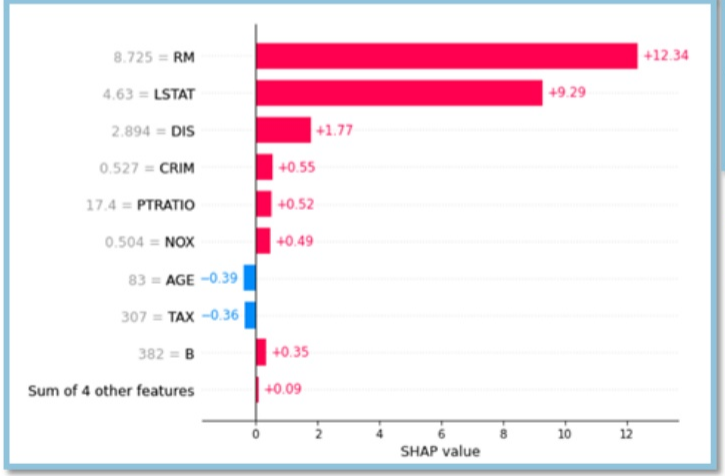
House Price is 13481$
because
LSTAT is 22.6,
RM is 5.71 and
PTRATIO is 21

XAI Human Explanation

House Price is 13481$
because of
a higher % in lower status population (22.6),
a lower number of rooms on avg. (5.71)
and a high pupil-teacher ratio in schools nearby (21)

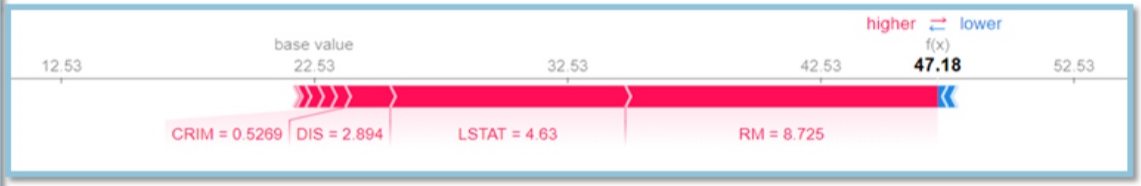
House Price Prediction – Local Interpretation

```
Actual House Price ($1000s: 50.0
Predicted House Price ($1000s): 47.177902
Key Features Affecting Prediction:
```



Local Important Features

Features Driving Prediction Prices



XAI Interpretation

House Price is 47178$
because
RM is 8.725,
LSTAT is 4.63 and
DIS is 2.894

XAI Human Explanation

House Price is 47178$
because of
a higher number of rooms on avg. (8.725)
a lower % in lower status population (4.63)
and a shorter distance to the major employment centres in Boston (2.894)

Census Income Prediction - Dataset Details

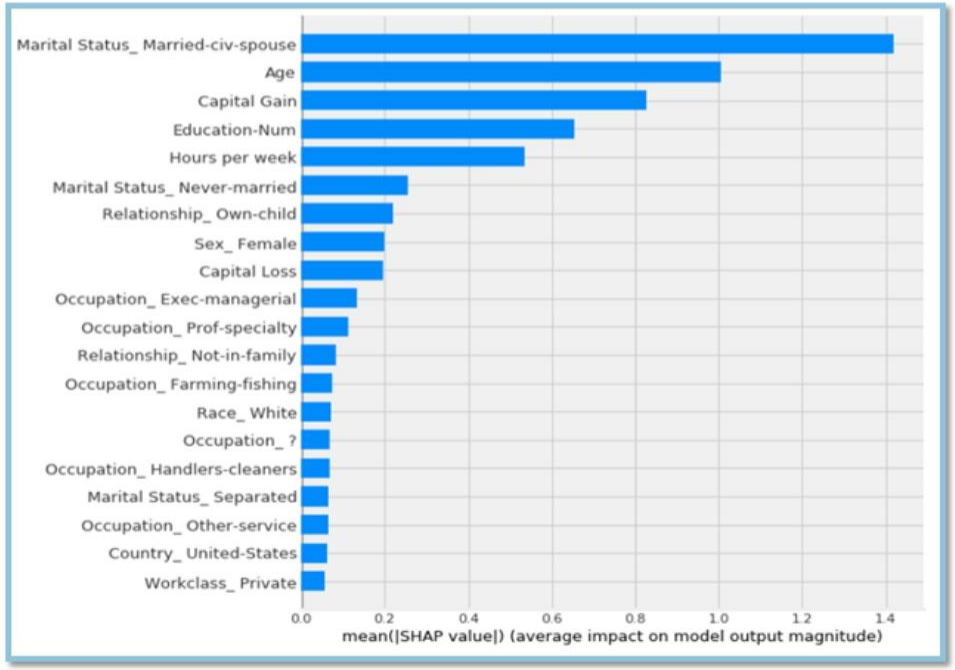
Census Income Dataset:

Attribute Name	Type	Description
Age	Continuous	Represents age of the person
Workclass	Categorical	Represents the nature of working class\category (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
Education-Num	Categorical	Numeric representation of educational qualification. Ranges from 1-16. (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool)
Marital Status	Categorical	Represents the marital status of the person (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)
Occupation	Categorical	Represents the type of profession\job of the person (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
Relationship	Categorical	Represents the relationship status of the person (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
Race	Categorical	Represents the race of the person (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
Sex	Categorical	Represents the gender of the person (Female, Male)
Capital Gain	Continuous	The total capital gain for the person
Capital Loss	Continuous	The total capital loss for the person
Hours per week	Continuous	Total hours spent working per week
Country	Categorical	The country where the person is residing
Income Label (labels)	Categorical (class label)	The class label column is the one we want to predict (False: Income <= $50K & True: Income > $50K)

Sample Data:

	Age	Workclass	Education-Num	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country
0	39.0	State-gov	13.0	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0	40.0	United-States
1	50.0	Self-emp-not-inc	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	13.0	United-States
2	38.0	Private	9.0	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0	40.0	United-States
3	53.0	Private	7.0	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0.0	40.0	United-States
4	28.0	Private	13.0	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0.0	40.0	Cuba

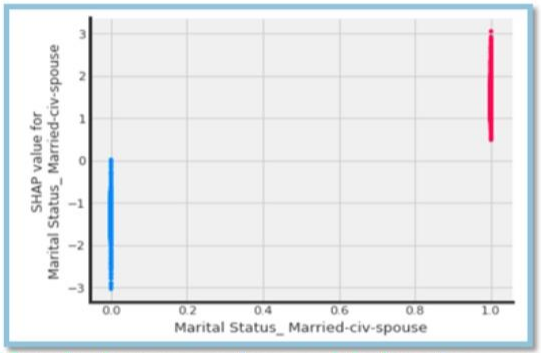
Census Income Prediction – Global Interpretation



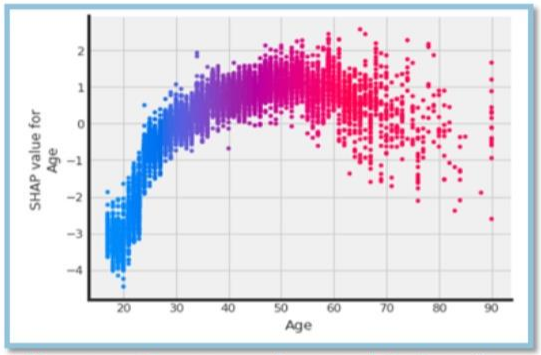
Top features driving model to predict that people will earn > $50K are:

- Marital Status: Marital status of the person is Married-civ-spouse
- Age: The age of the person
- Capital Gain: Total Capital Gain of the person
- Education-Num: The education level of the person (higher is better)

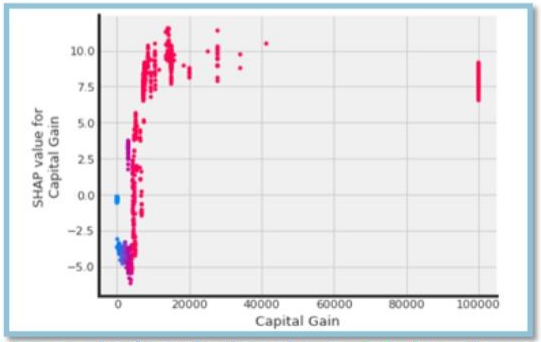
Census Income Prediction – Dependence Plots



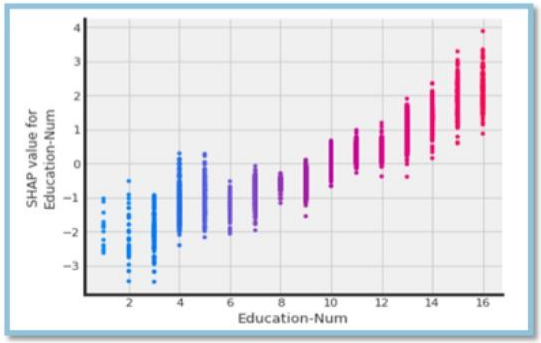
Married person has a higher income



Middle-aged person has a higher income



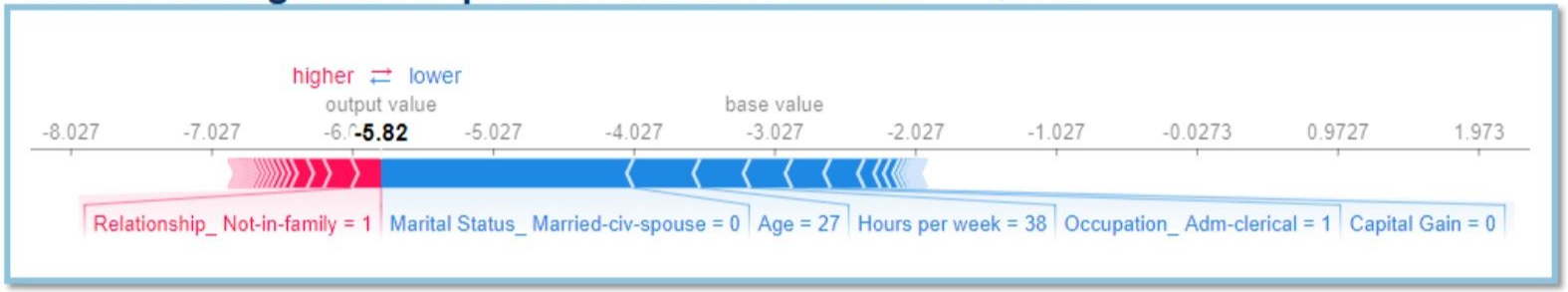
High capital gain leads to higher income



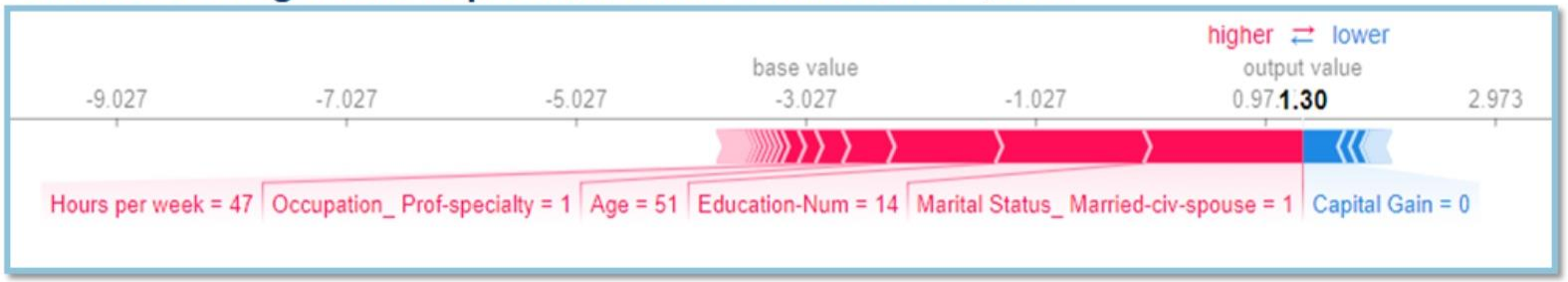
Higher education leads to higher income

Census Income Prediction – Local Interpretation

Features driving Model to predict Person A's income <= $50K



Features driving Model to predict Person B's income > $50K



Techniques for Interpreting DL Models - Image Data

1. ## Visualizing Activation Layers

   Visualize how a given input comes out of specific activation layers. Explores which feature maps are getting activated in the model

2. ## Occlusion Sensitivity

   Visualize how parts of the image affects neural network's confidence by occluding \ hiding parts of the image iteratively

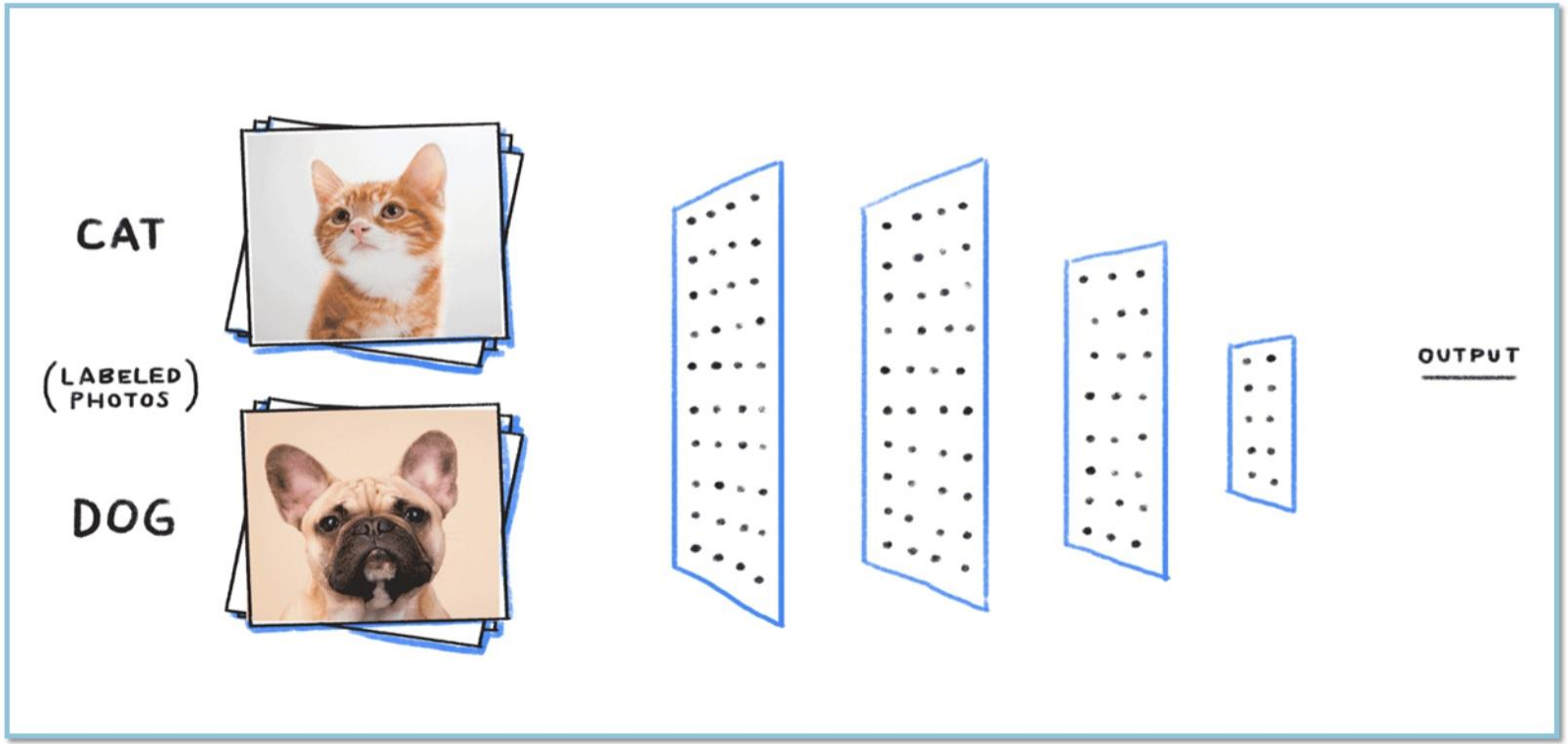
3. ## SHAP Gradient Explainer

   Combines ideas from activation gradients and SHAP values into a single expected value equation

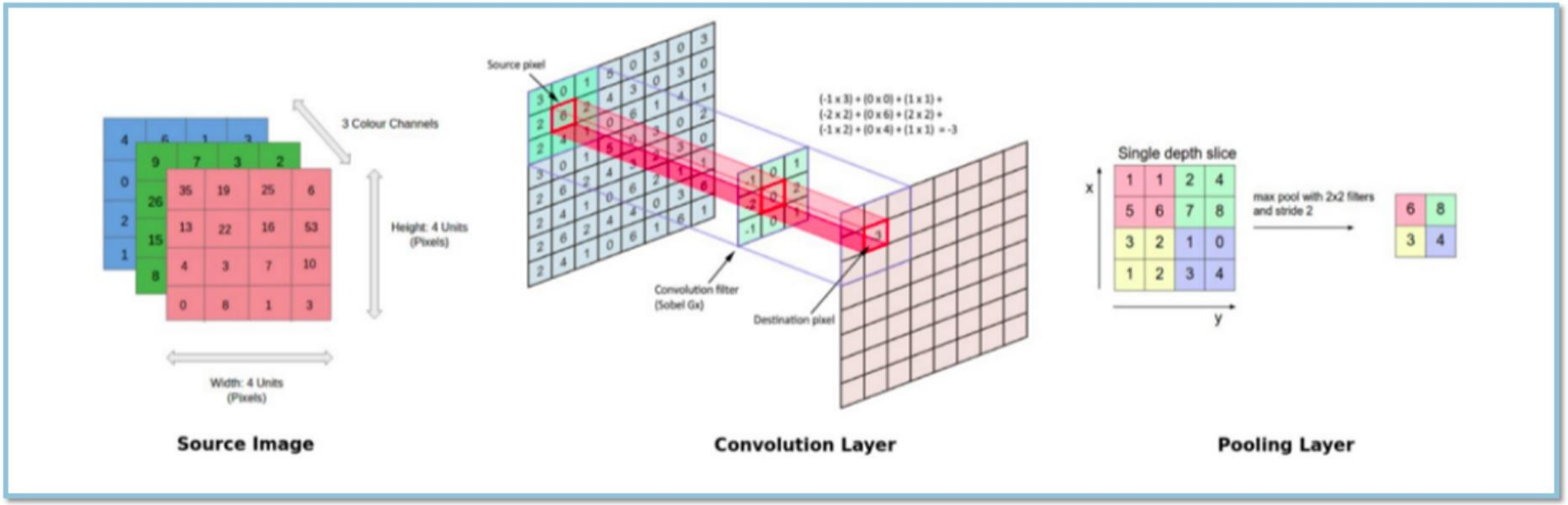
4. ## Grad-CAM

   Visualize how parts of the image affects neural network's output by looking into the gradients backpropagated to the class activation maps

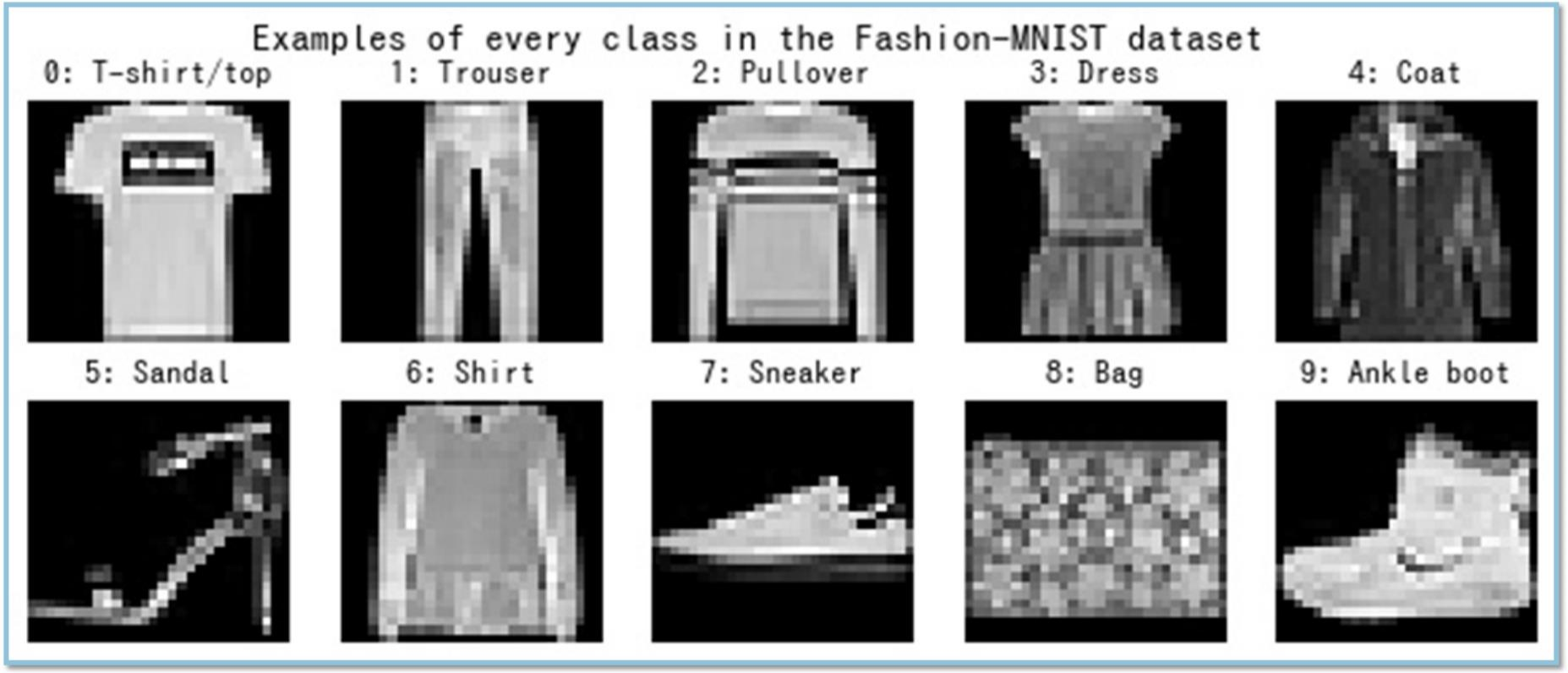
Convolutional Neural Networks



Convolutional Neural Networks



Fashion Apparel Prediction – Dataset Examples



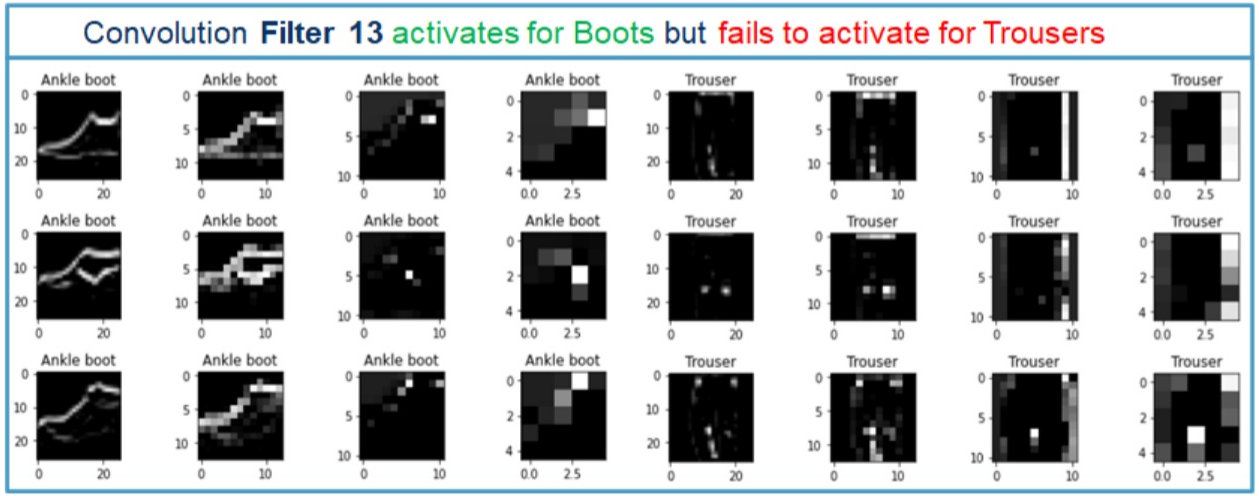
Fashion Apparel – Visualizing Activation Layers

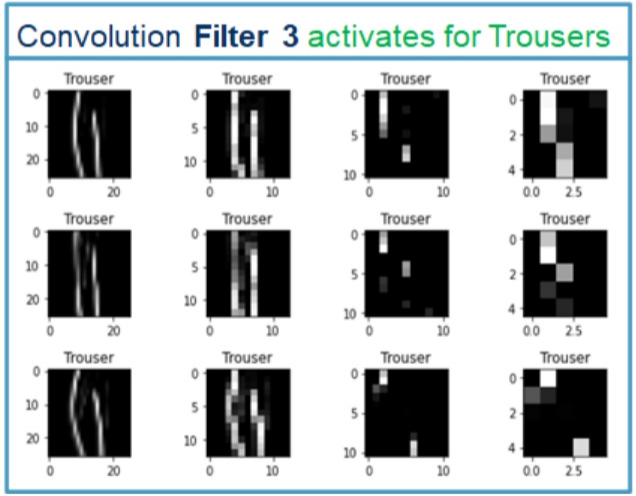
Input Image

Output From Conv2D

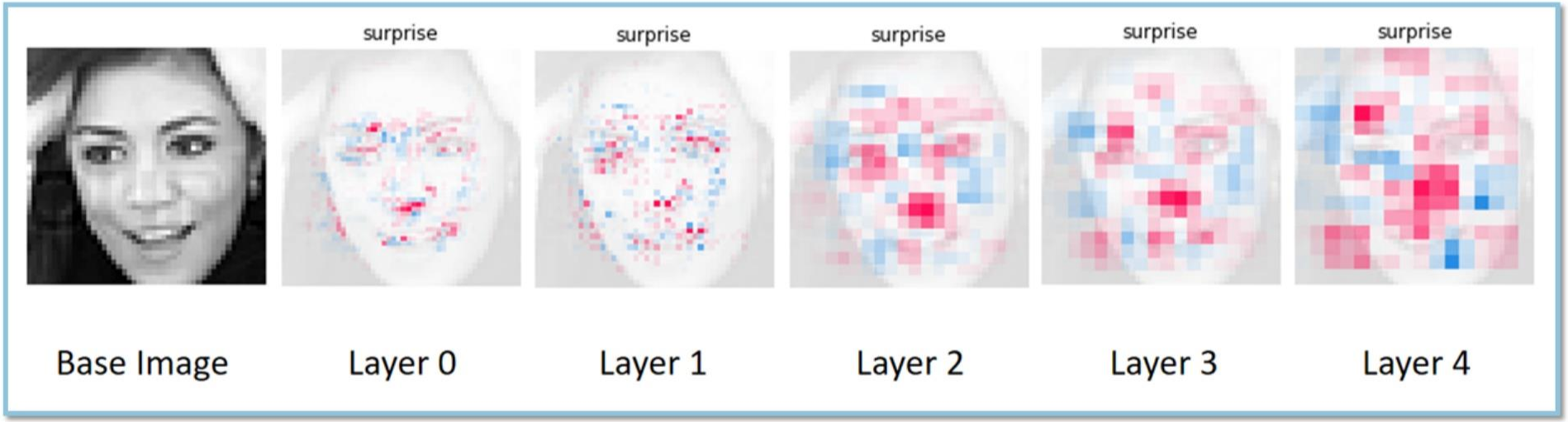
(Feature Maps after ReLU Processing)

Fashion Apparel – Visualize CNN Activations



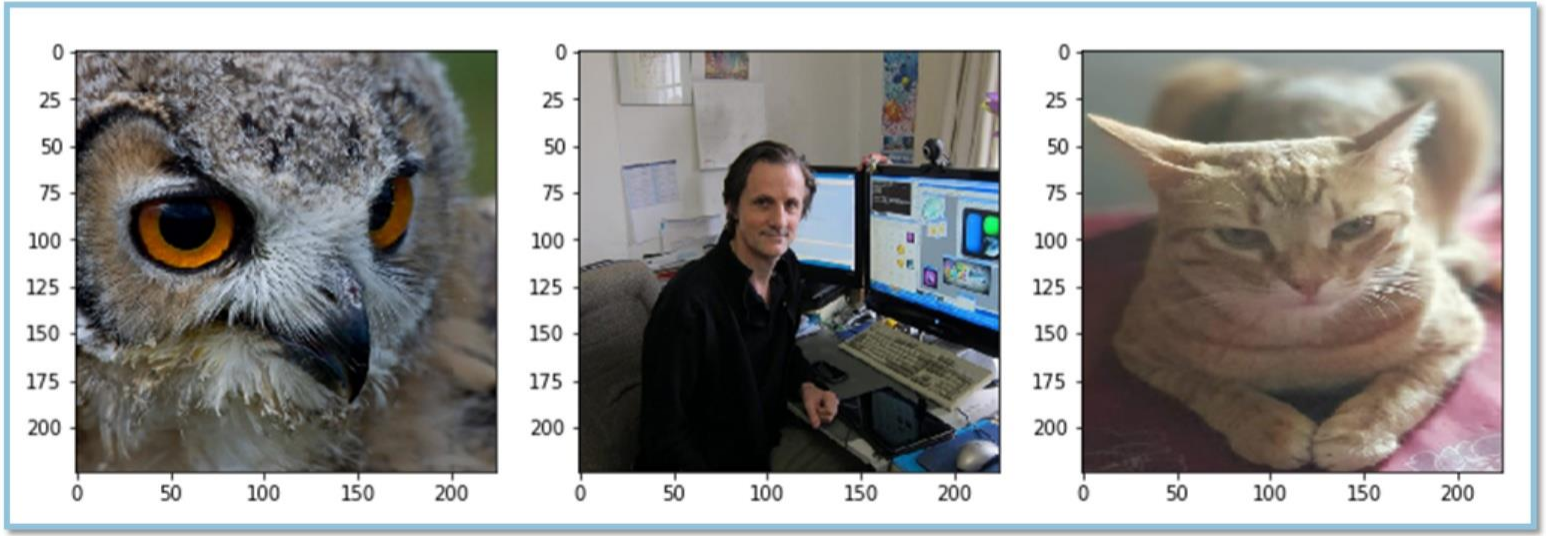


SHAP Gradient Explainer

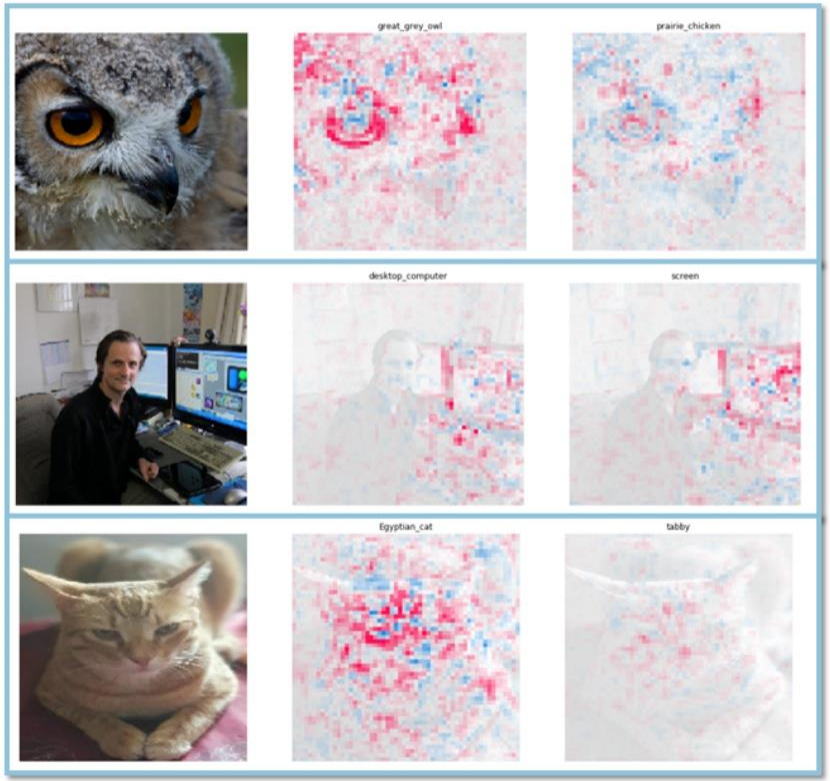


- Red pixels represent positive SHAP values that increase the probability of the predicted class
- Blue pixels represent negative SHAP values the reduce the probability of the predicted class

SHAP Gradient Explainer – Sample Input Data



SHAP Gradient Explainer – Prediction Explanations

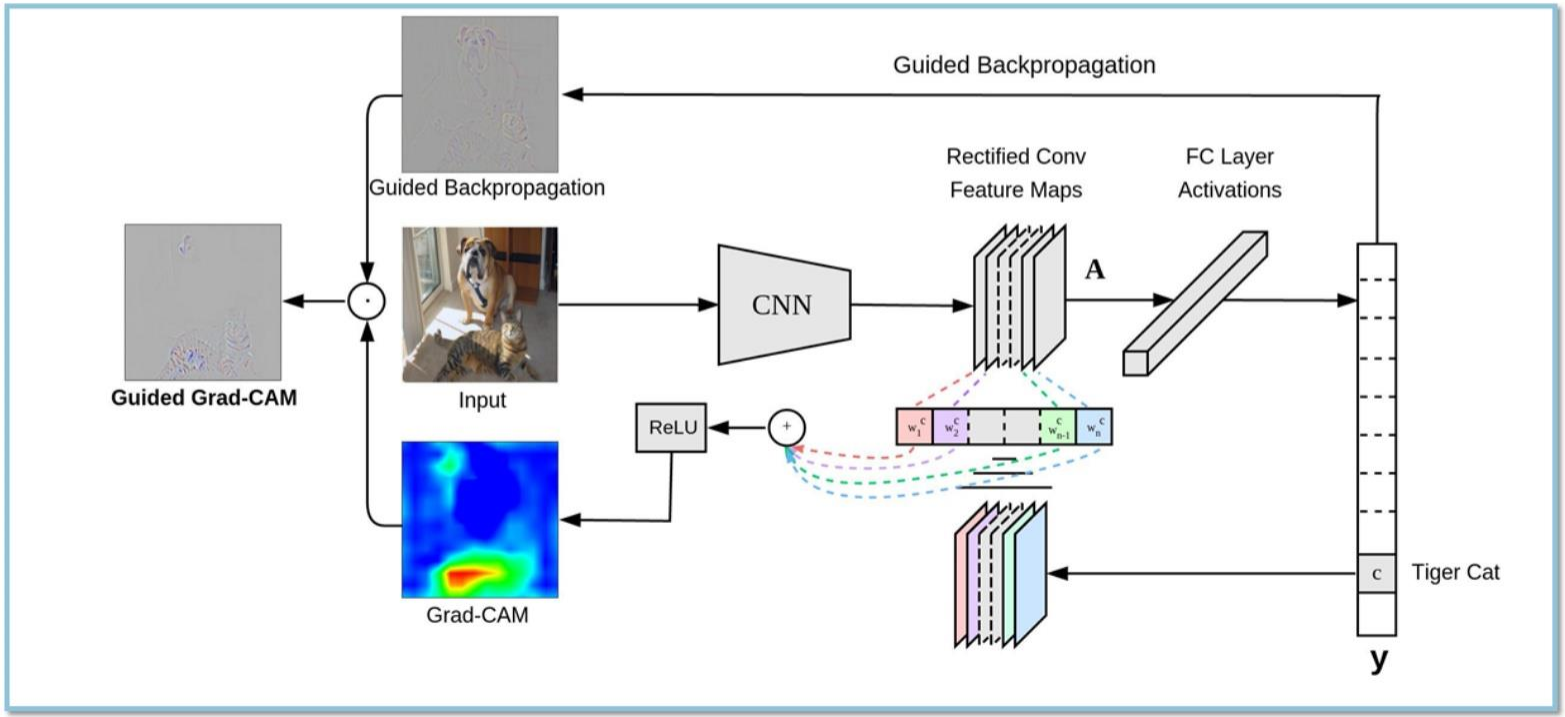


- 1st most probable prediction: Great grey owl which focuses on facial structure and texture of the image
- 2nd most probable prediction: Prairie chicken which focuses more on the eyes of the image

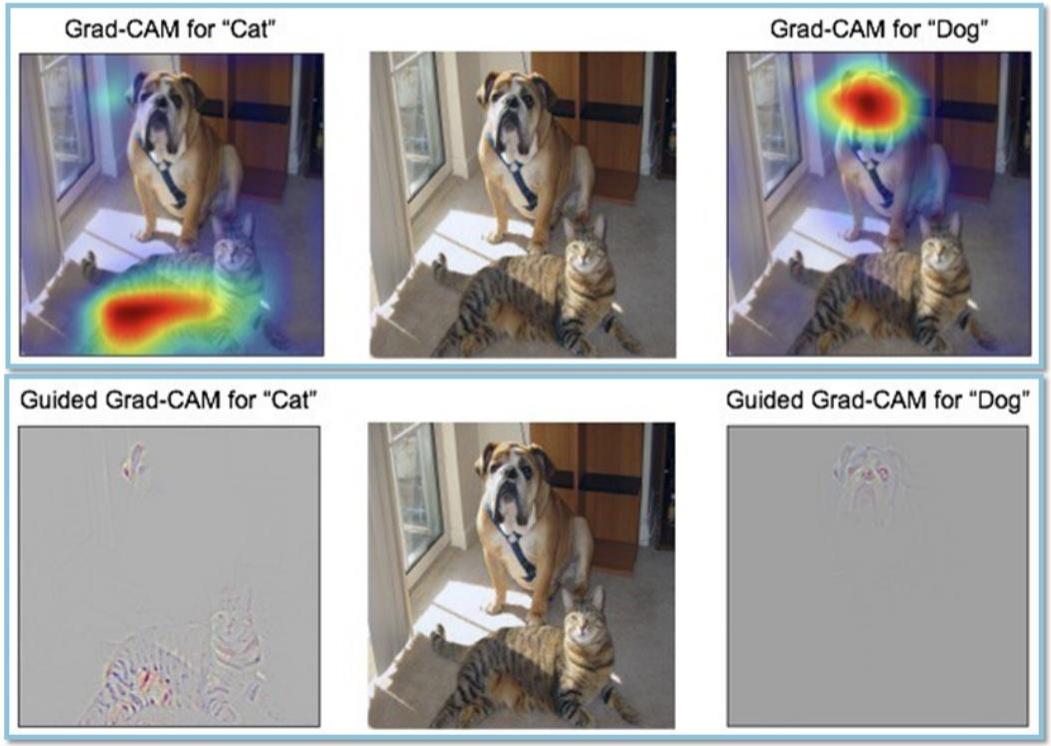
- 1st most probable prediction: Desktop computer which focuses on screen, keyboard, mouse of the image
- 2nd most probable prediction: Screen which focuses more on the screen area of the image

- 1st most probable prediction: Egyptian cat which focuses on overall structure and texture of the image
- 2nd most probable prediction: Tabby cat which focuses more on the facial texture \ pattern of the image

Grad-CAM – CNN Prediction Explanations



Grad-CAM – CNN Prediction Explanations



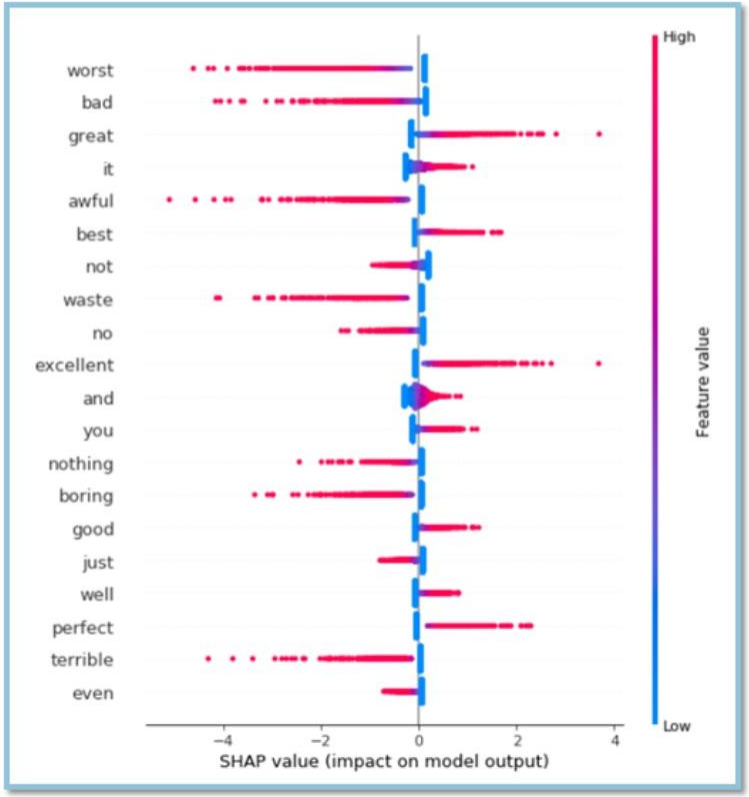
Techniques for Interpreting ML Models - Text Data

1. LIME
2. SHAP

Predicting Movie Review Sentiment – Sample Dataset

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

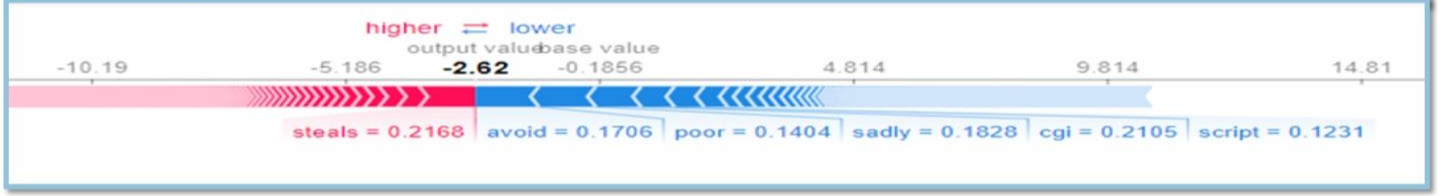
SHAP Explainer – Global Interpretations for Sentiment



SHAP Explainer – Local Interpretations for Sentiment

Features driving model to predict Negative sentiment

Review: mr perlman gives a standout performance as usual sadly he has to struggle with an underwritten script and some nonsensical set pieces larsen is in die hard mode complete with singlet and bulging muscles i am sure he could do better but seems satisfied to grimace and snarl through his part the lovely erika is very decorative even though fully clothed and shows some signs of getting acting at last sfx are mainly poor cgi and steals from other movies the shootouts are pitiful worthy of the a teamnot even worth seeing for perlman avoid
Actual: Negative
Predicted: Negative



Features driving model to predict Positive sentiment

Review: this i think is one of the best pictures ever made it is so pure and beautiful it really touched me i am glad david lynch proved that a film does not necessarily need sfx a twisting complicated plot or flashy images way to go dave i would like to see cronenberg do that
Actual: Positive
Predicted: Positive


Hands-on Tutorials
LIME & SHAP

Final Words

- Models are not biased by themselves, the roots lie in bias in data and humans
- Focus on building robust and explainable models for critical systems
- Explainability is not necessary for all ML systems
- Constraint Optimization and Adversarial Learning are key to robust ML systems
- Responsible AI principles are critical to ensure progress and adaptation of AI

Stay in Touch!

LinkedIn
linkedin.com/in/dipanzan

GitHub
github.com/dipanjanS

Blog
blog.djsarkar.ai

Medium
djsarkar.medium.com