

# **eXplainable Method For High Complexity Model**

Data Scientist  
안건이

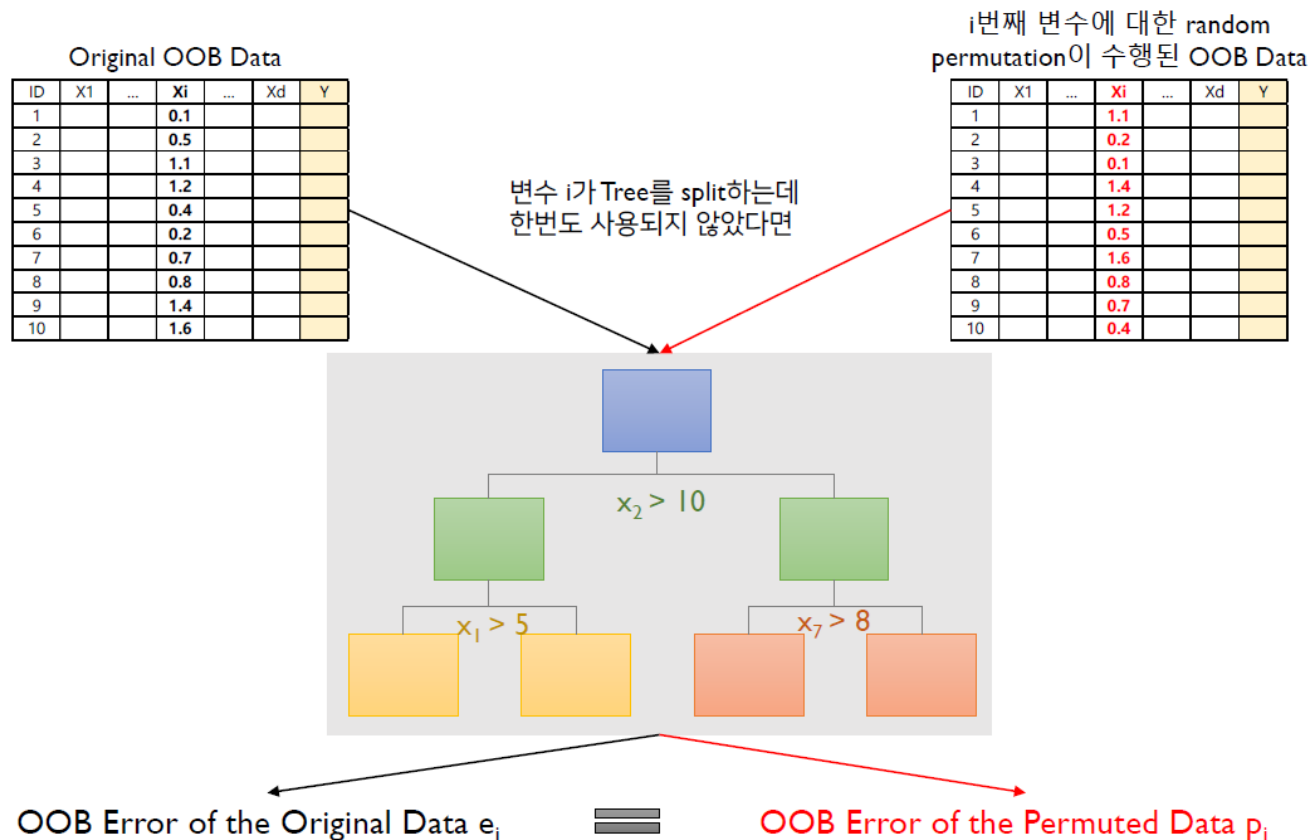
# 목차

---

- Global vs Local Importance
- LIME
- SHAP
- Code 실습

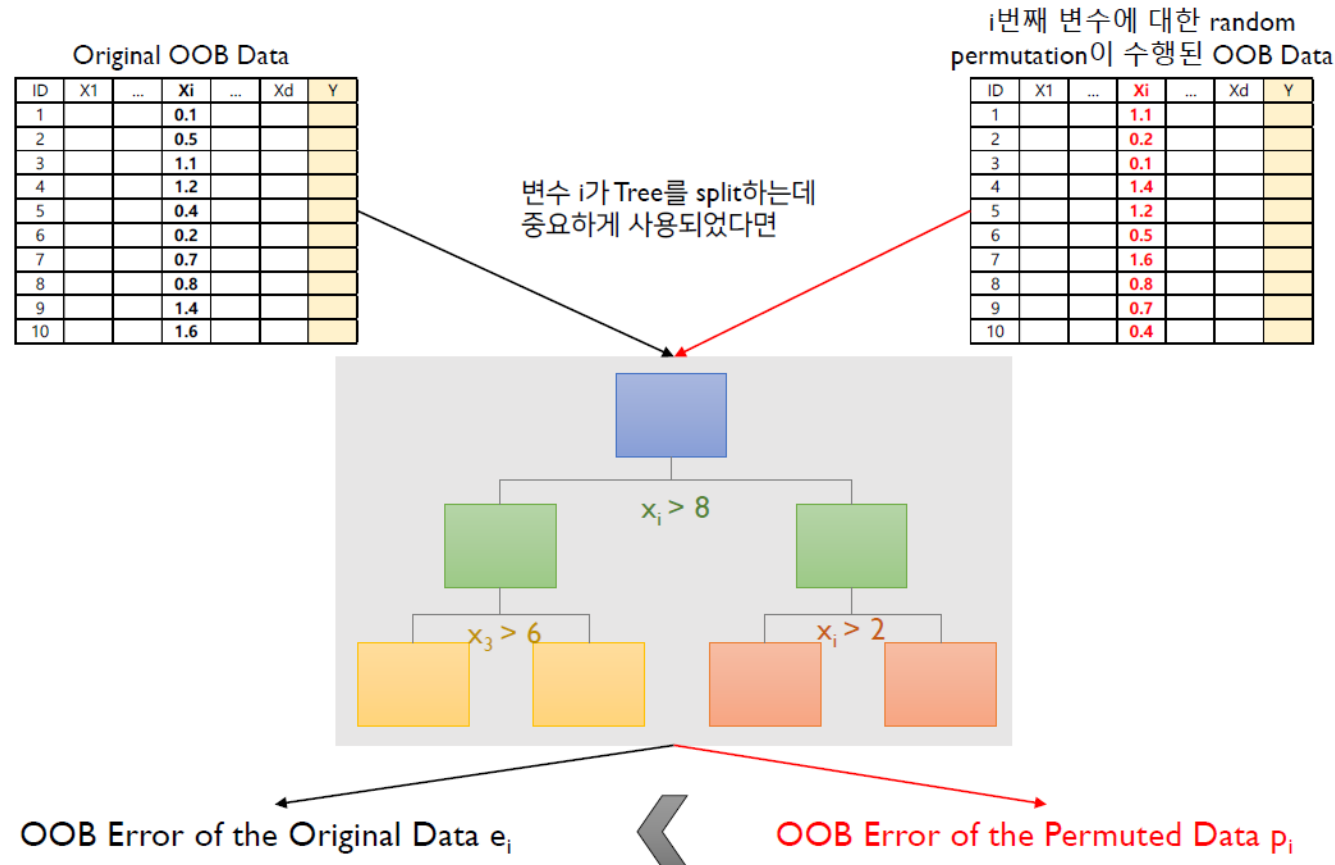
# Importance Scores

- Y에 얼마만큼 영향을 미치는지에 대한 중요도
  - $Y ? \rightarrow 10\% \text{ or } 50\% \text{ or } 100\%$
  - 여태 배웠던 기법은 모두 Y 전체에 영향을 미치는 중요도를 배움
- Random Forest
  - Variable Importance
  - Step 1 : Compute the OOB(Out of bag) error for the original dataset
  - Step 2 : Compute the OOB error for the dataset in which the variable  $x_i$  is permuted  $p_i$
  - Step 3 : Compute the variable importance based on the mean and standard deviation of over all trees in the population



# Importance Scores

- Y에 얼마만큼 영향을 미치는지에 대한 중요도
  - $Y ? \rightarrow 10\% \text{ or } 50\% \text{ or } 100\%$
  - 여태 배웠던 기법은 모두 Y 전체에 영향을 미치는 중요도를 배움
- Random Forest
  - Variable Importance
  - Step 1 : Compute the OOB(Out of bag) error for the original dataset
  - Step 2 : Compute the OOB error for the dataset in which the variable  $x_i$  is permuted  $p_i$
  - Step 3 : Compute the variable importance based on the mean and standard deviation of over all trees in the population



# Importance Scores

- Y에 얼마만큼 영향을 미치는지에 대한 중요도
  - Y ?  $\rightarrow$  10% or 50% or 100%
  - 여태 배웠던 기법은 모두 Y 전체에 영향을 미치는 중요도를 배움
- Random Forest
  - Variable Importance
  - Step 1 : Compute the OOB(Out of bag) error for the original dataset
  - Step 2 : Compute the OOB error for the dataset in which the variable  $x_i$  is permuted  $p_i$
  - Step 3 : Compute the variable importance based on the mean and standard deviation of over all trees in the population

✓ 랜덤 포레스트에서 변수의 중요도가 높다면

- 1) Random permutation 전-후의 OOB Error 차이가 크게 나타나야 하며,
  - 2) 그 차이의 편차가 적어야 함
- 
- m번째 tree에서 변수 i에 대한 Random permutation 전후 OOB error의 차이

$$d_i^m = p_i^m - e_i^m$$

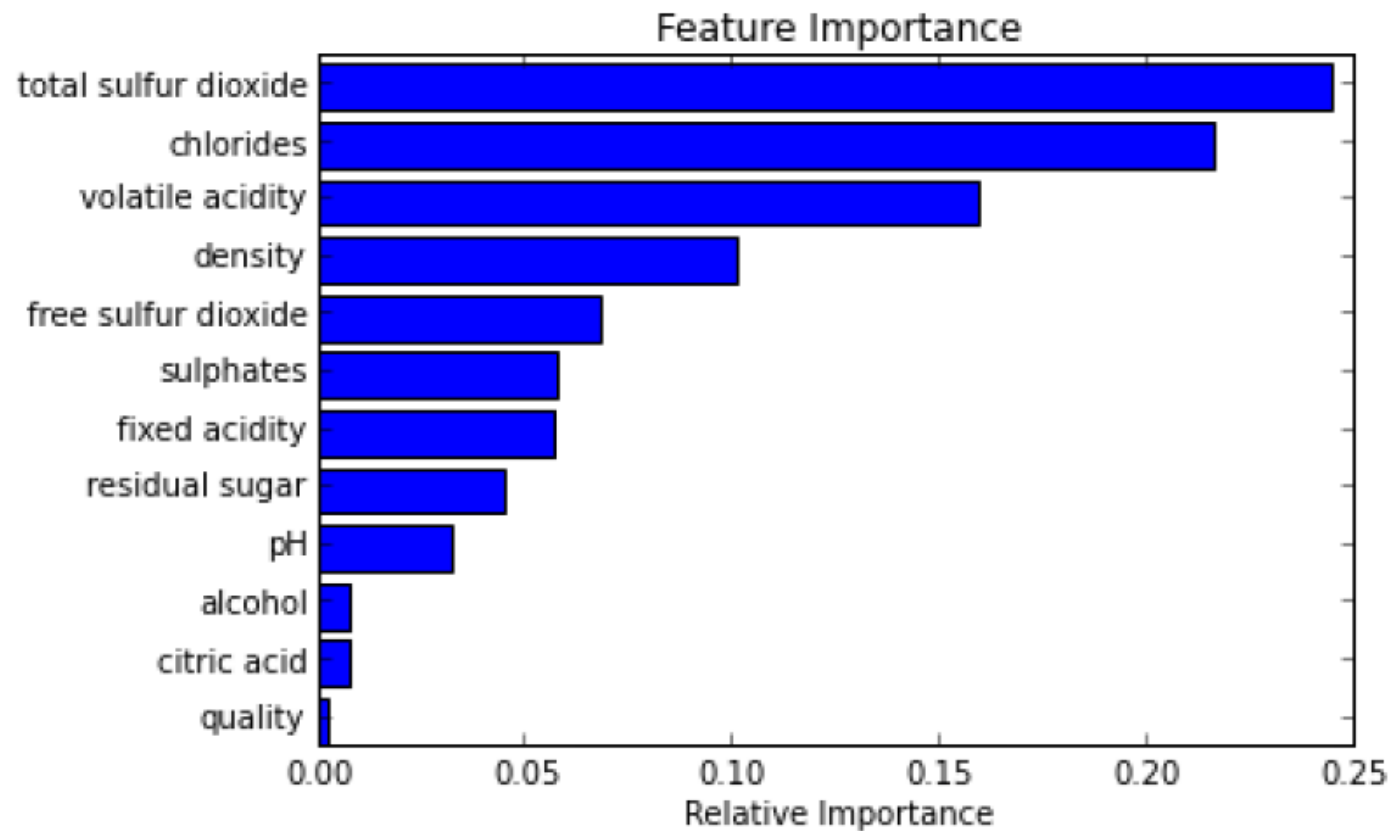
- 전체 Tree들에 대한 OOB error 차이의 평균 및 분산

$$\bar{d}_i = \frac{1}{m} \sum_{i=1}^m d_i^m, \quad s_i^2 = \frac{1}{m-1} \sum_{i=1}^m (d_i^m - \bar{d}_i)^2$$

- i번째 변수의 중요도:  $v_i = \frac{\bar{d}_i}{s_i}$

# Importance Scores

- Y에 얼마만큼 영향을 미치는지에 대한 중요도
  - Y ? → 10% or 50% or 100%
  - 여태 배웠던 기법은 모두 Y 전체에 영향을 미치는 중요도를 배움
- Random Forest
  - Variable Importance
  - Step 1 : Compute the OOB(Out of bag) error for the original dataset
  - Step 2 : Compute the OOB error for the dataset in which the variable  $x_i$  is permuted  $p_i$
  - Step 3 : Compute the variable importance based on the mean and standard deviation of over all trees in the population



**Global**

**vs**

**Local**

## Global

vs

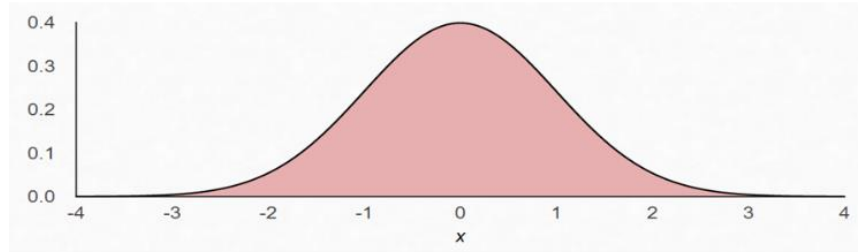
## Local

- Y 전체(Global)
- Y에 영향을 미치는 중요도
  - 효율에 영향을 미치는 중요도
  - 공부에 영향을 미치는 중요도

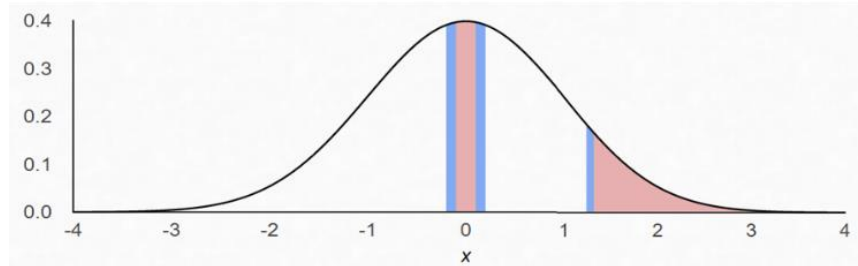
- 특정 Y (Local)
- 특정 Y에 영향을 미치는 중요도
  - 초고효율에 영향을 미치는 중요도
  - 고득점에 영향을 미치는 중요도



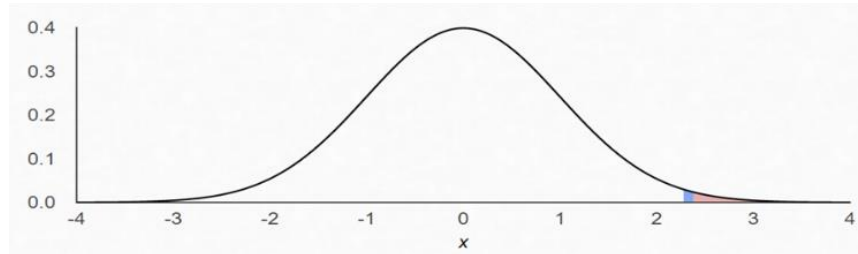
## Global vs Local



Case 1



Case 2



Case 3

## Global vs Local

- 기존 분석의 중요인자 추출 기법은 전체 데이터를 대변하는 중요 인자를 추출함
- 우리의 목적은 고효율 군을 결정 짓는 중요 인자를 추출하는 것
  - 더 나아가 상위 0.1% 초고효율군을 결정 짓는 중요 인자를 추출
- 전체 Y를 대변하는 중요 인자(Global Interpretability)와 특정 데이터에 대한 중요 인자(Local Interpretability)는 다른 결과를 불러올 수 있음
- 특정 Y(내가 원하는 데이터)에 대한 해석력을 얻기 위해서는 Black Box Model을 열어 봐야함
- Interpretable Machine Learning (IML)을 통해 Black Box Model을 열어 봄

Global

vs

Local

# HOW ????

- 기존 분석의 중요인자 추출 방법은 전체 데이터를 대변하는 중요인자를 추출함
- 우리의 목적은 고효율 군을 구성하는 중요인자 추출하는 것
  - 더 나아가 상위 0.1% 초고효율군을 결정 짓는 중요인자를 추출
- 전체 Y를 대변하는 중요인자(Global Interpretability)와 특정 데이터에 대한 중요인자(Local Interpretability)는 다른 결과를 불러올 수 있음
- 특정 Y(내가 원하는 데이터)에 대한 해석력을 얻기 위해서는 Black Box Model을 열어 봐야함
- Interpretable Machine Learning (IML)을 통해 Black Box Model을 열어 봄

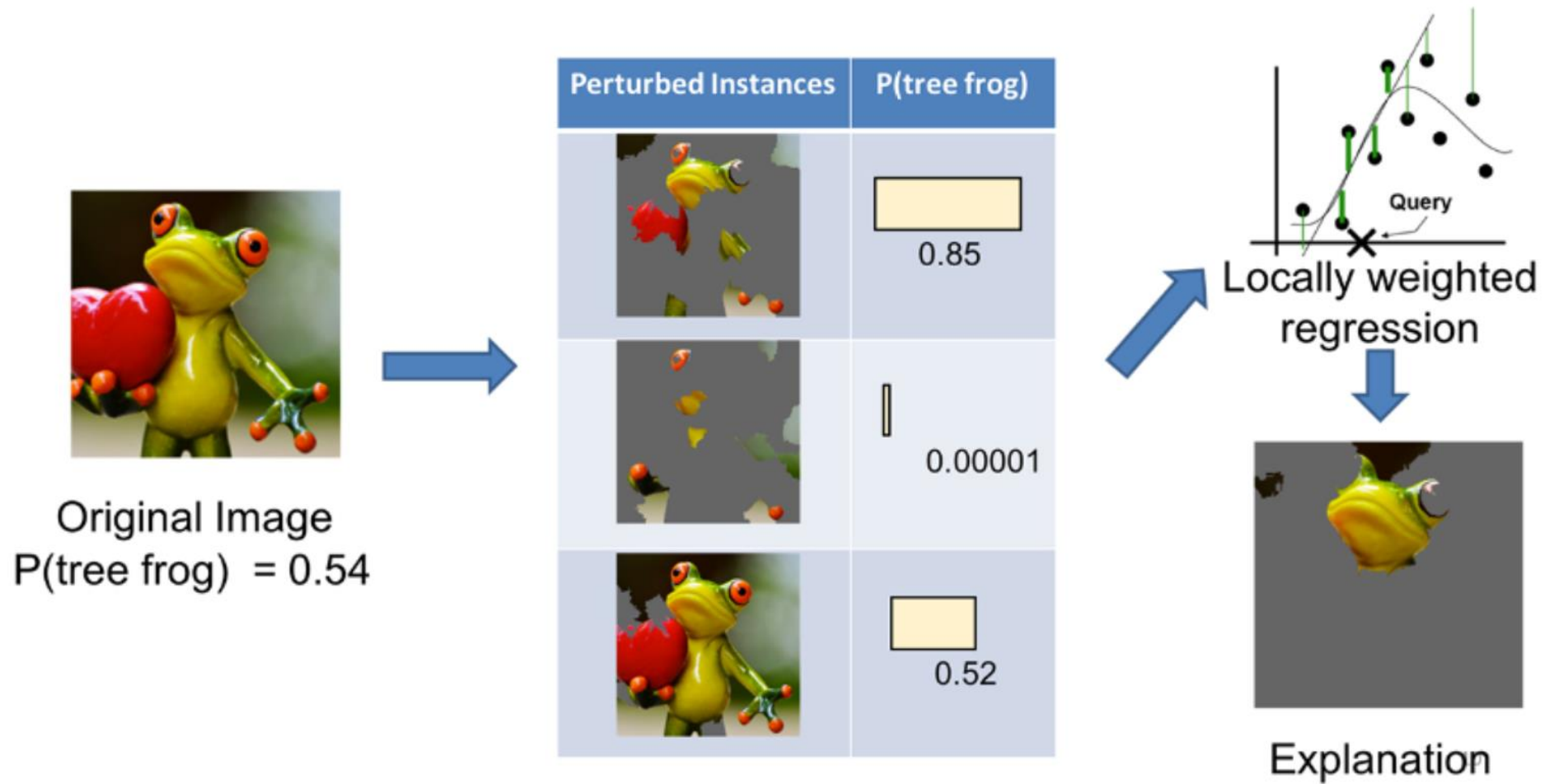
- LIME

## Local Interpretable Model-agnostic Explanation

- Local : 특정 Y
- Interpretable : 해석 가능한
- Model-agnostic : Model Free (어떠한 Model 이라도 적용 가능)
- Explanation : 설명할 수 있음
- etc : 숫자, 텍스트, 이미지 등 모든 데이터에 적용 가능

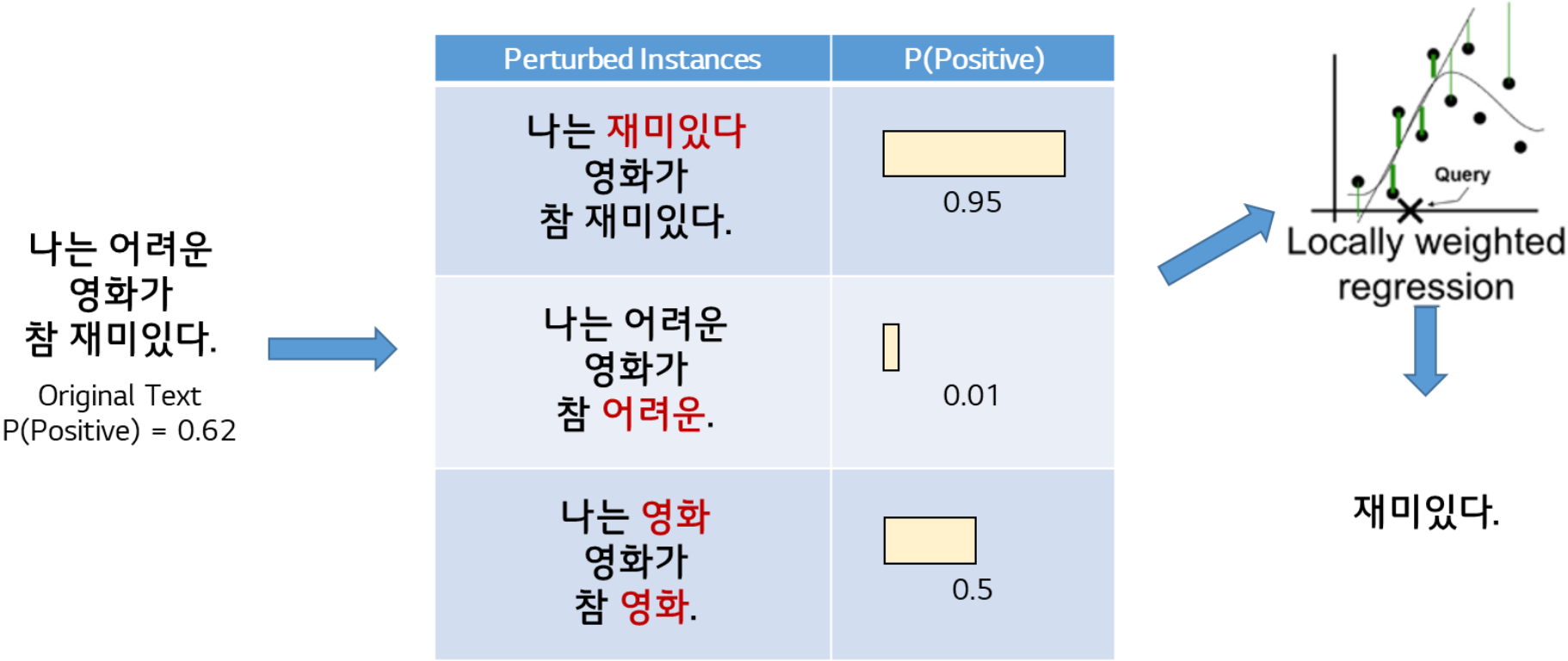
# Interpretable Machine Learning - LIME

- LIME
  - IMAGE



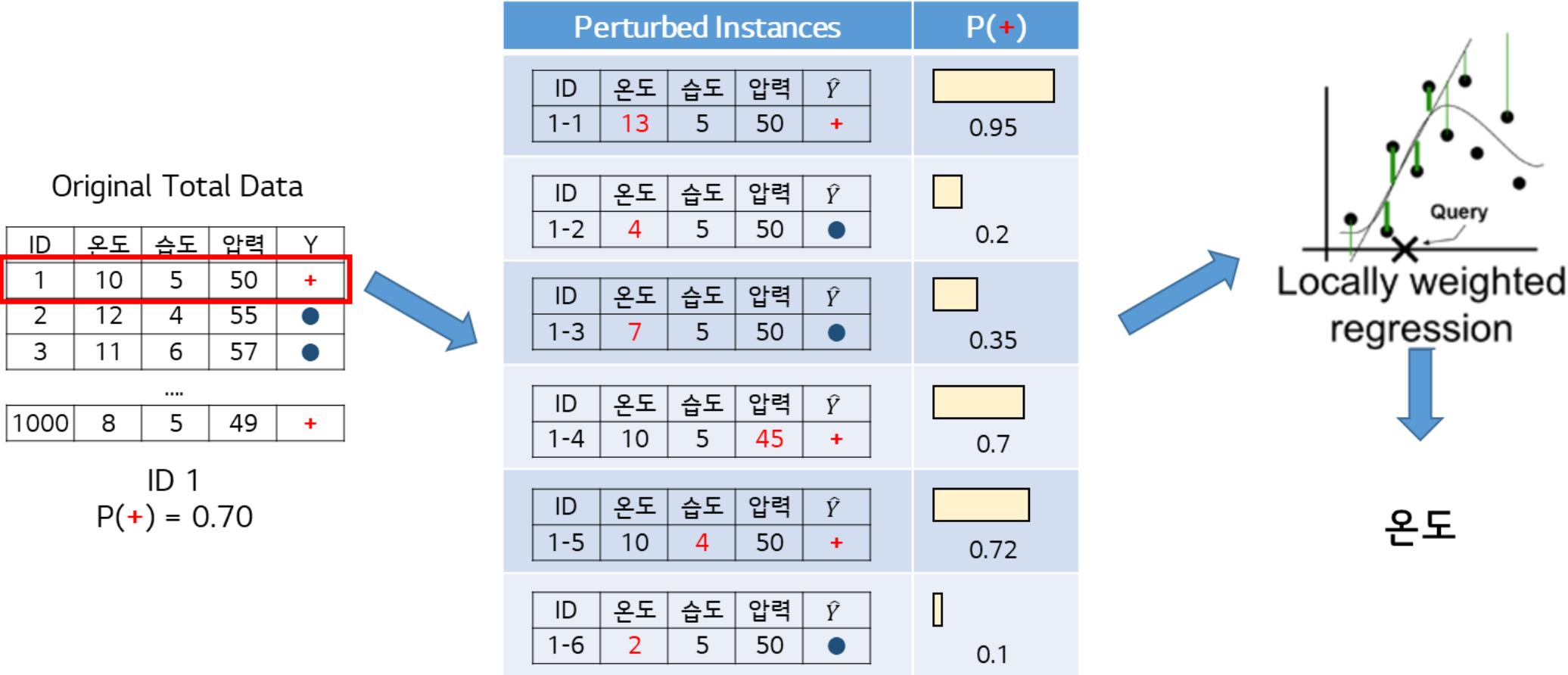
# Interpretable Machine Learning - LIME

- LIME
  - TEXT



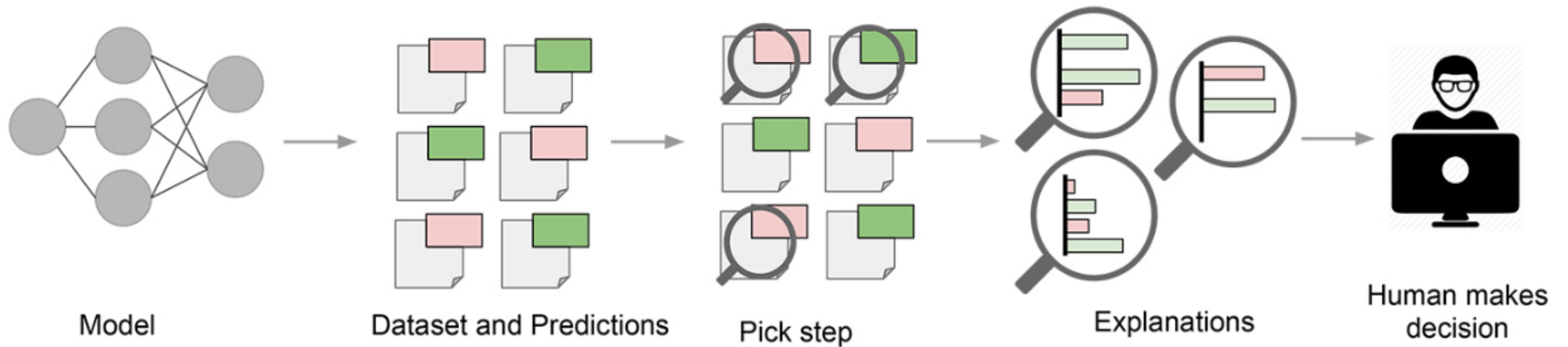
# Interpretable Machine Learning - LIME

- LIME
  - Numeric



# Interpretable Machine Learning - LIME

- LIME
  - Process



*Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.*



# Interpretable Machine Learning - LIME

- LIME
  - Step 1. Modeling
    - 평상시 우리가 돌리는 모델을 사용하여 학습을 진행함
    - Weight based Model (Regression, Logistic Regression ...), Tree based Model (Random Forest, Gradient Boosting Tree, Xgboost ...), Deep Learning (CNNs, RNNs ...) 모든 모델을 사용해도 됨
    - LIME은 모델에 대한 Scalability를 보장함

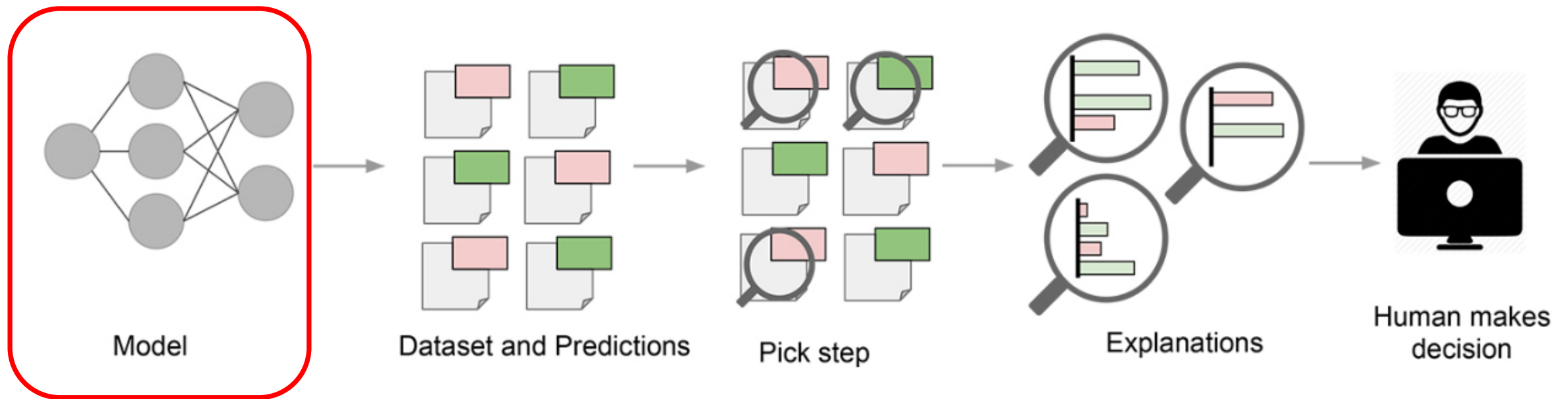


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

# Interpretable Machine Learning - LIME

- LIME
  - Step 2. Dataset and Predictions
    - 학습이 완료된 모델을 사용하여 학습데이터 또는 검증데이터에 대하여 예측을 진행함

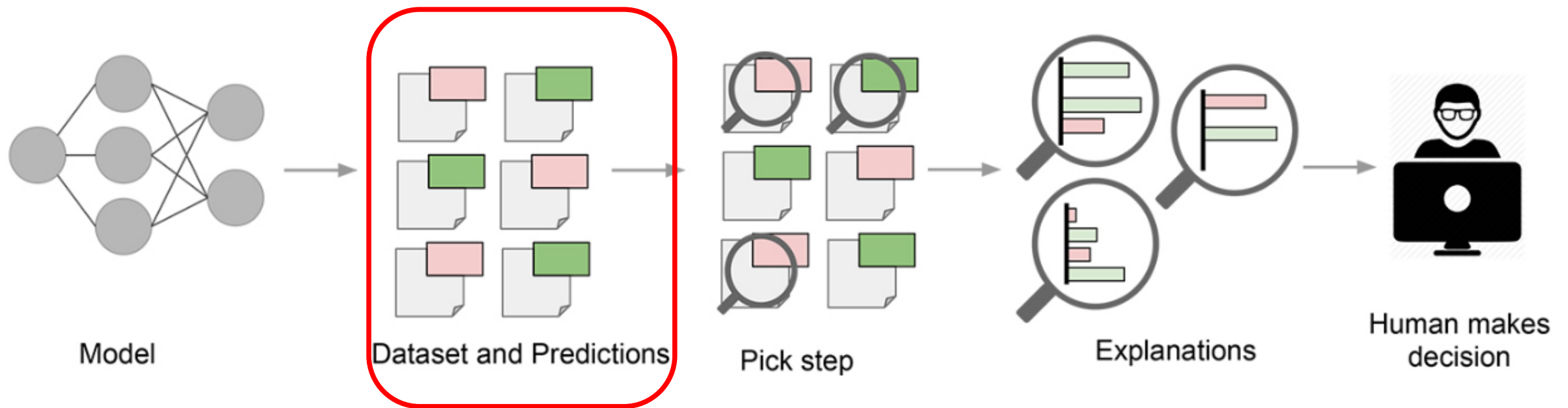


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

# Interpretable Machine Learning - LIME

- LIME
  - Step 3. Picking step
    - 우리가 Targeting한 데이터를 추출함
    - 예를 들어, 초 고효율군을 학습데이터와 검증데이터에서 추출함
    - 더 나아가 확실한 Pattern이 있는 데이터를 추출하기 위하여 (학습이 잘된 데이터) MSE가 낮은 데이터를 추출함

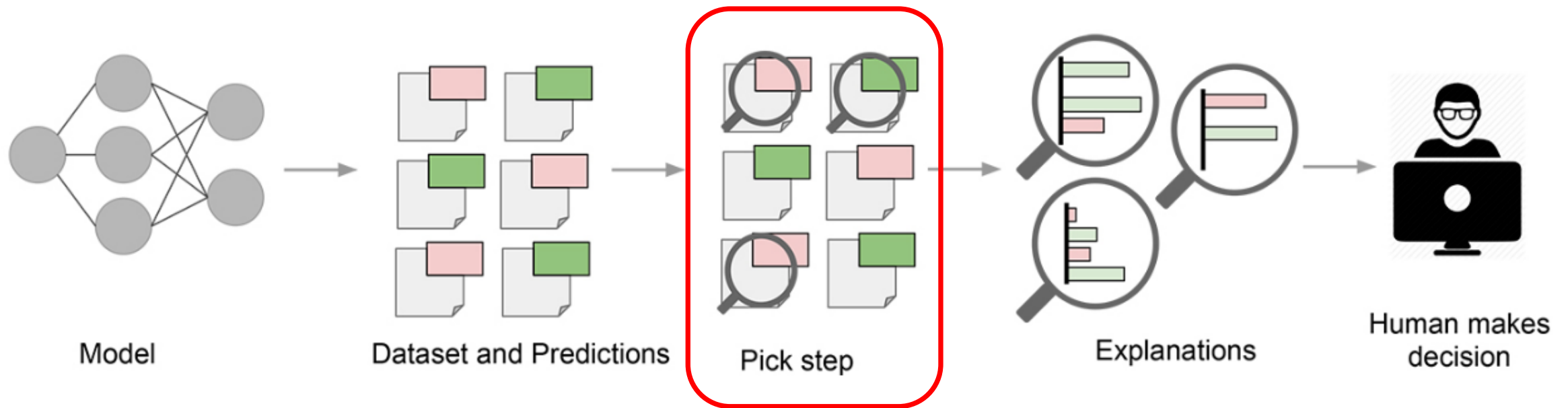


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

# Interpretable Machine Learning - LIME

- LIME
  - Step 4. Explanations
    - Pick step에서 추출한 데이터에 대해 중요인자를 도출함
    - 도출할 때는 LIME 또는 SHAP를 사용하여 도출함

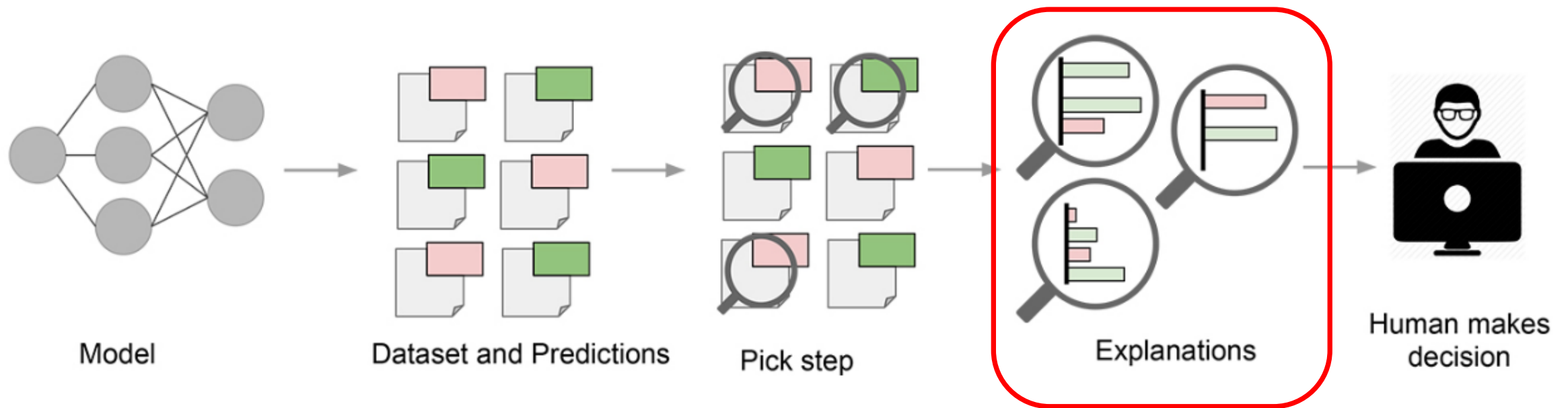


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

# Interpretable Machine Learning - LIME

- LIME
  - Step 5. Human makes decision
    - Explanations step에서 도출한 중요인자를 바탕으로 실제로 중요한지에 대한 여부를 판단함
    - 판단하기 위해서 중요인자와 Y와의 상관관계, plotting을 통하여 Insight를 도출함

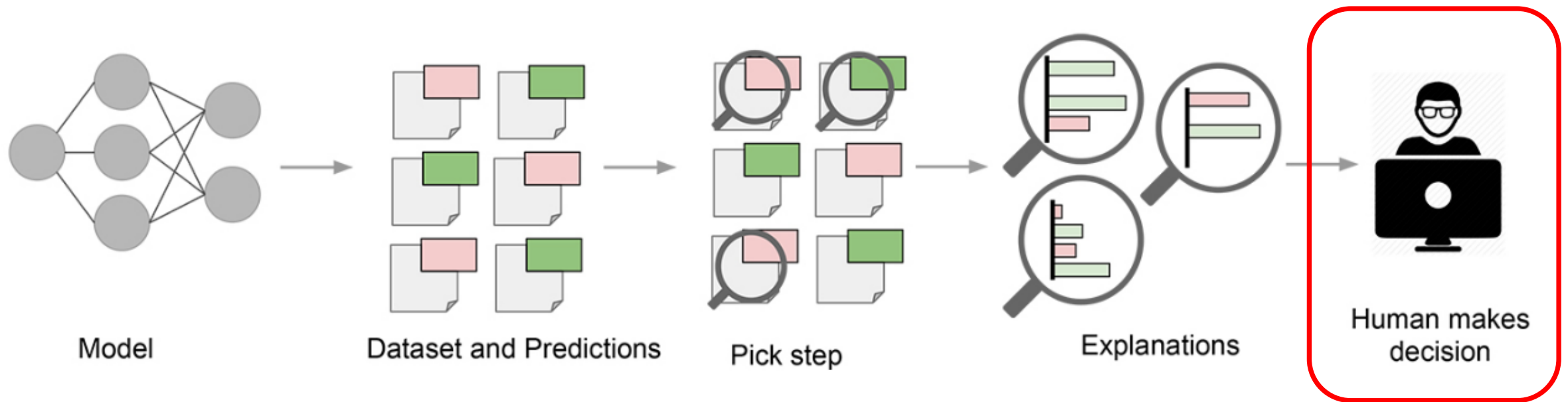


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

# Interpretable Machine Learning - LIME

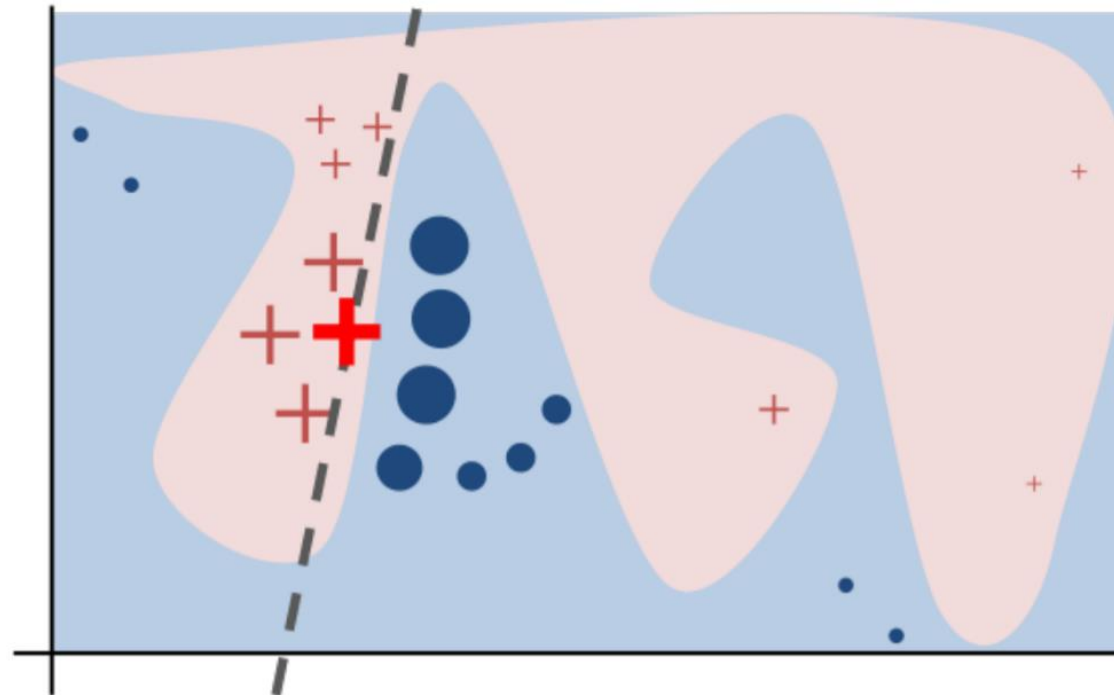
- LIME
  - Step 5. Human makes decision
    - Explanations step에서 도출한 중요인자를 바탕으로 실제로 중요한지에 대한 여부를 판단함
    - 판단하기 위해서 중요인자와 Y와의 상관관계, plotting을 통하여 Insight를 도출함



Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

# Interpretable Machine Learning - LIME

- LIME
  - 가장 진한 빨간색 "+"는 우리가 선택한 데이터
    - 우리는 이 데이터에 대해 중요인자를 알고 싶음
  - 먼저, 가장 진한 빨간색 "+" 데이터에 대하여 변수의 값을 바꿔가며 여러 개의 데이터를 생성함 (Perturbation)
  - 앞서 Image에서는 특정 영역의 값을 지우며 데이터를 변형 시켰고, Text에서는 특정 단어들을 바꿔가며 데이터를 변형시켰음
  - Numeric 데이터에서는 그림과 같이 변수의 값들을 조금씩 바꿔가며 여러 개의 데이터를 생성함
  - 이렇게 생성된 데이터는 그림에서 "+", "o"이며, 우리가 선택한 데이터 주변에 분포하게 됨
  - 교란된 데이터는 미리 학습하여 만든 모형으로 교란된 데이터의 Y를 결정해주고(X는 선택한 데이터를 바탕으로 변수의 값을 바꿔가며 생성해 냈음), 우리가 선택한 데이터와 교란된 데이터를 바탕으로 새로운 모델을 생성해 줌(Locally weighted regression). 이 모델은 그림에서 "----"에 해당함



# Interpretable Machine Learning - LIME

- LIME
  - Locally approximation 과정

$$\varepsilon(g) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

- $f$  : 전체 데이터를 사용하여 학습 시켜 놓은 모델 (Global Model, [그림 4]에서 파란색과 분홍색 영역을 나누는 Boundary)
- $g$  : 선택한 데이터와 교란된 데이터를 바탕으로 Locally approximation할 새로운 Simple 모형 (논문에서는 Lasso만 사용하여 K개의 변수를 사용함, Simple 모형이고 데이터 수가 적기 때문에 복잡한 모델을 사용하지 않아도 됨, [그림 4]에서  $g$ 는 "-----"임)
- $G$  :  $g$ 에 사용할 수 있는 모델들의 집합(Lasso, Decision Tree, Random Forest 등등, 변수 중요도를 구할 수 있는 Simple 모델)
- $\pi_x$  : 내가 선택한 데이터와 교란된 데이터와의 거리를 가중치로 환산한 것(거리가 멀면 가중치가 낮고, 거리가 가까우면 가중치가 크다)
- $\Omega(g)$  : 새로 생성하는 Simple 모형의 복잡도 (Lasso면 Penalty term  $\lambda$ , Tree 계열이면 Depth)

$$L(f, g, \pi_x) = \sum_{z, z'} \pi_x(z) \{f(z) - g(z')\}^2$$

$$\text{Where } \pi_x(z) = \exp\left(\frac{-D(x, z)^2}{\sigma^2}\right)$$



# Interpretable Machine Learning – SHAP





- SHAP
  - SHapley Additive exPlanations
  - LIME 개념 + 경제학 개념
    - 노벨 경제학상을 받은 Shapley Values(게임이론)를 접목시킴
- Core는 효율성이라는 바람직한 특성을 가지고 있지만, 바람직하지 못한 특성도 보유함
- 즉, Core가 존재하지 않을 수도 있고, 아주 커다란 core를 가지고 있을 수도 있음
- 따라서 “유일한” 해를 제공하지 못하는 단점이 있음
- 이러한 단점이 없는 다른 해들이 많이 개발 되었으며, 그 중 가장 잘 알려진 것이 SV(Shapley Value)임
- Shapley Value는 UCLA의 Lloyd Shapley의 이름을 딴 이름임



Lloyd Shapley, 1980  
(1923~, 87세)

# Interpretable Machine Learning – SHAP

- SHAP
  - Shapley Value에 대해 알기 위해서는 먼저 게임이론에 대해 이해해야 함
  - 게임이론이란 우리가 아는 게임을 말하는 것이 아닌 여러 주제가 서로 영향을 미치는 상황에서 서로 어떤 의사결정이나 행동을 하는지에 대해 이론화한 것을 말함

	A DEFECT	A COOPERATE
B DEFECT		
B COOP- ERATE		

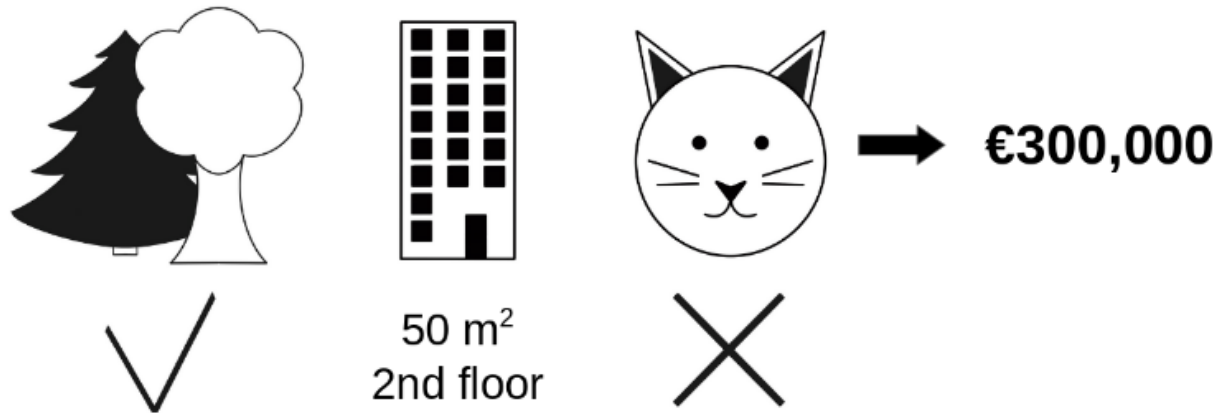
상호 작용

- SHAP
  - Shapley Value에 대해 알기 위해서는 먼저 게임이론에 대해 이해해야 함
  - 게임이론이란 우리가 아는 게임을 말하는 것이 아닌 여러 주제가 서로 영향을 미치는 상황에서 서로 어떤 의사결정이나 행동을 하는지에 대해 이론화한 것을 말함

Y에 영향을 미친 X's을  
Shapley Value를 활용하여  
중요도 계산을 해보자.

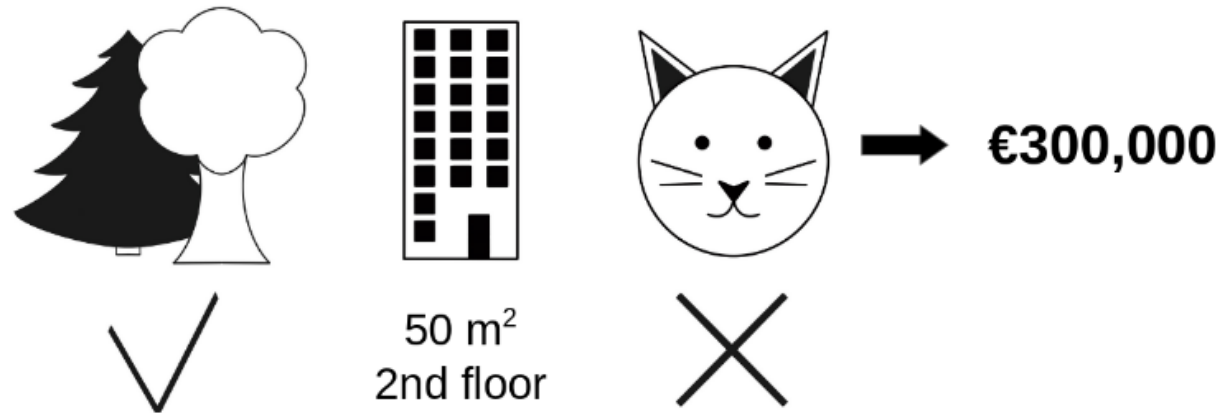
# Interpretable Machine Learning – SHAP

- 예시
  - 집 값: 300,000 유로
    - 공원(O)
    - 50평
    - 2층
    - 고양이 출입금지
  - 주변 평균 아파트 시세: 310,000 유로
- 과연 어떤 인자가 주변 시세보다 10,000 유로를 낮게 했을까?



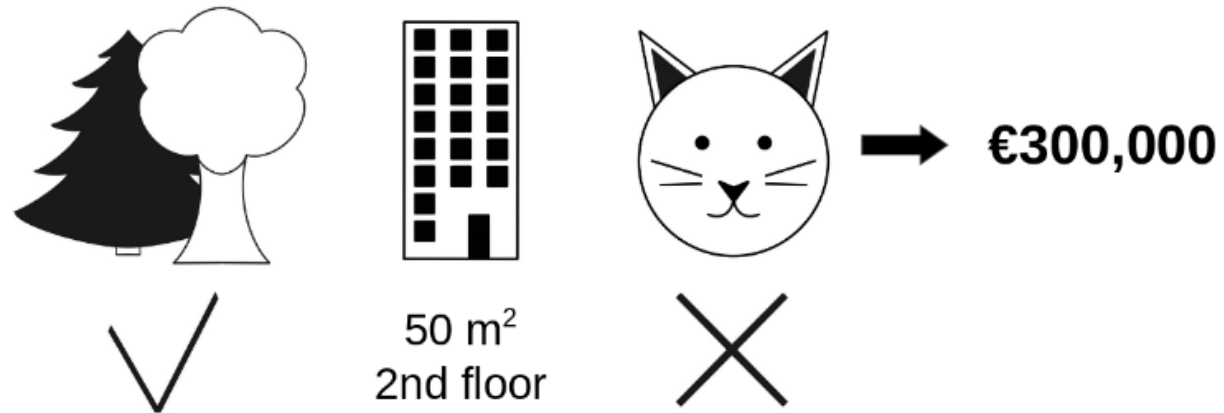
# Interpretable Machine Learning – SHAP

- 예시
  - 집 값: 300,000 유로
    - 공원(O)
    - 50평
    - 2층
    - 고양이 출입금지
  - 주변 평균 아파트 시세: 310,000 유로
- LIME의 경우 가중치(선형회귀의 경우 Coefficient를 말함)을 곱한 값
- 선형모델은 인과성을 가지고 있기 때문에 이렇게 단순히 가중치를 통해서 영향력을 바로 확일 할 수 있음



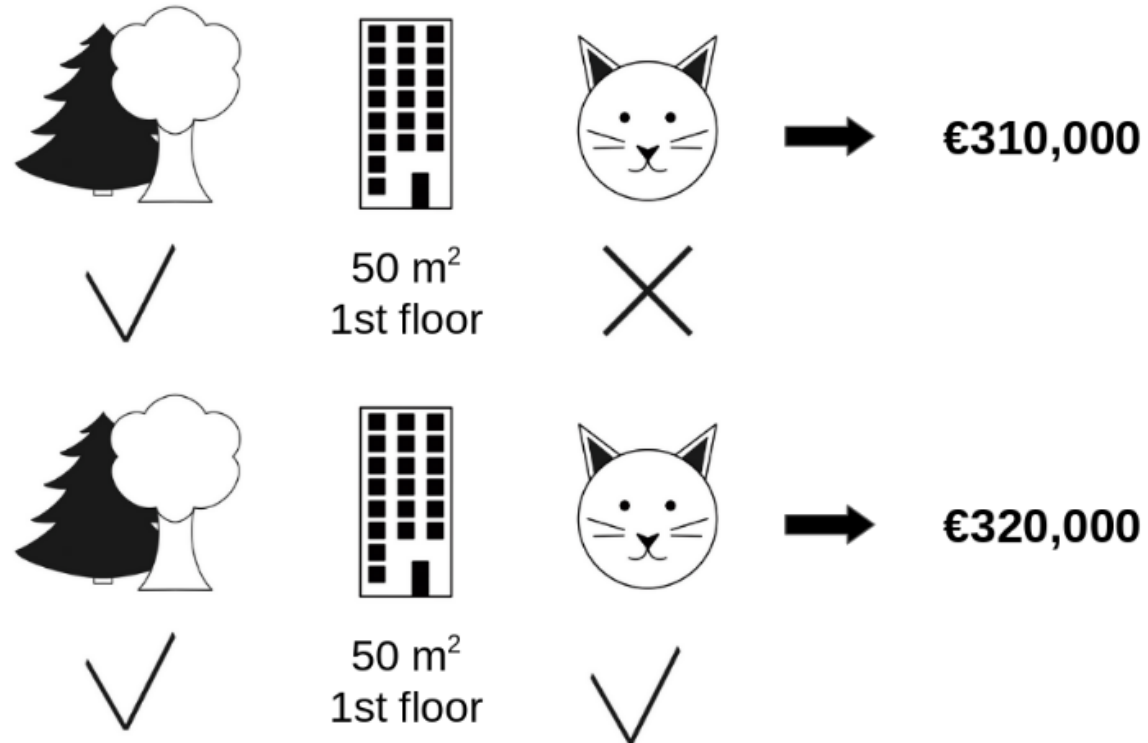
# Interpretable Machine Learning – SHAP

- 예시
  - 집 값: 300,000 유로
    - 공원(O)
    - 50평
    - 2층
    - 고양이 출입금지
  - 주변 평균 아파트 시세: 310,000 유로
- Shapley Value는 모든 가능한 조합에 대해서 하나의 특성의 기여도를 종합적으로 합한 값



# Interpretable Machine Learning – SHAP

- 예시
  - 1. 공원(0)과 50평을 추가 했을 때 고양이 출입금지의 기여도를 평가한다고 가정
    - 공원(0)과 50평 그리고 고양이 출입금지를 사용하여 무작위로 여러 아파트들에 대한 예측을 해보고 총 수 특성에 대한 가치를 확인함
  - 2. 2층에서 1층으로 변경 되었음에도 가격은 변하지 않았음
  - 3. 고양이 출입허용으로 변경 하였을 때 가격이 10,000 유로 상승함
  - 따라서, 고양이 출입금지의 기여도는 -10,000 유로 라는 것을 확인 할 수 있음
- 위와 같은 계산 과정을 모든 가능한 연합에 대해 반복함. SV는 모든 가능한 연합에 대한 모든 한계 기여도의 평균임



# Interpretable Machine Learning – SHAP

- 예시
  - 고양이 출입 금지의 SV를 계산하기 위한 모든 특성들의 조합은 아래와 같음
  - 각 연합에 대해서 고양이 출입 금지가 포함된 연합과 포함되지 않은 연합의 아파트 예측 가격을 계산하고 한계 기여도를 계산하여 차이를 구함
  - SV는 한계 기여도의 평균임
  - 기계학습에서 예측치를 구하기 위해 연합에 포함되지 않은 특성 값은 아파트 데이터 셋에서 무작위로 추출해서 대체함

- No Feature Values

- 공원(0)

- 50평

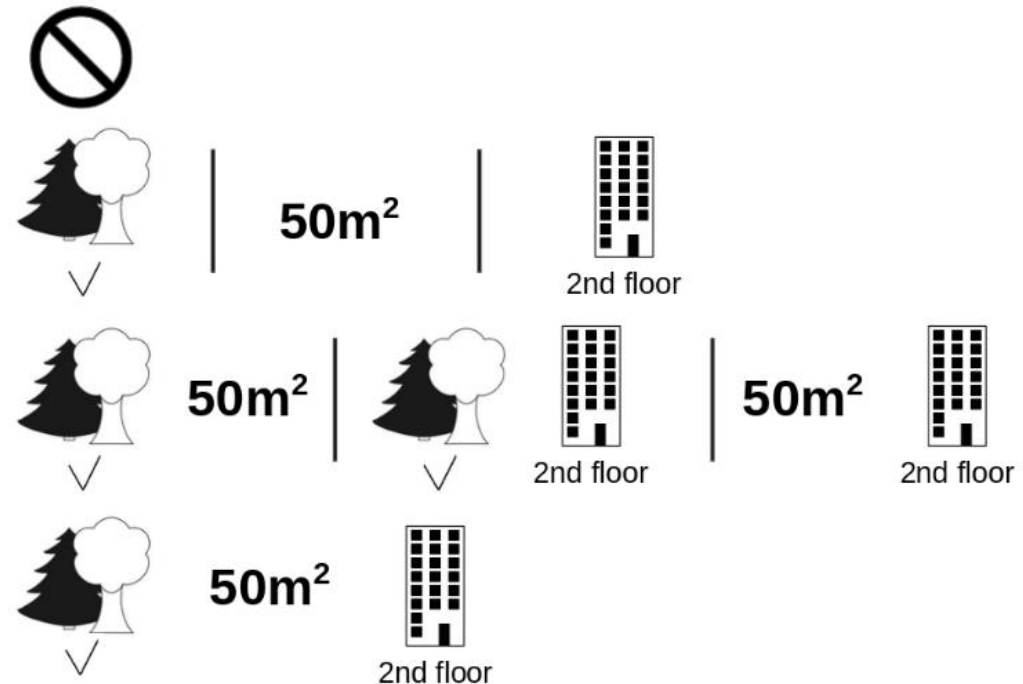
- 2층

- 공원(0) + 50평

- 공원(0) + 2층

- 50평 + 2층

- 공원(0) + 50평 + 2층





Q & A