

# Why Should I Trust You?

## LIME

(**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations)

박이삭

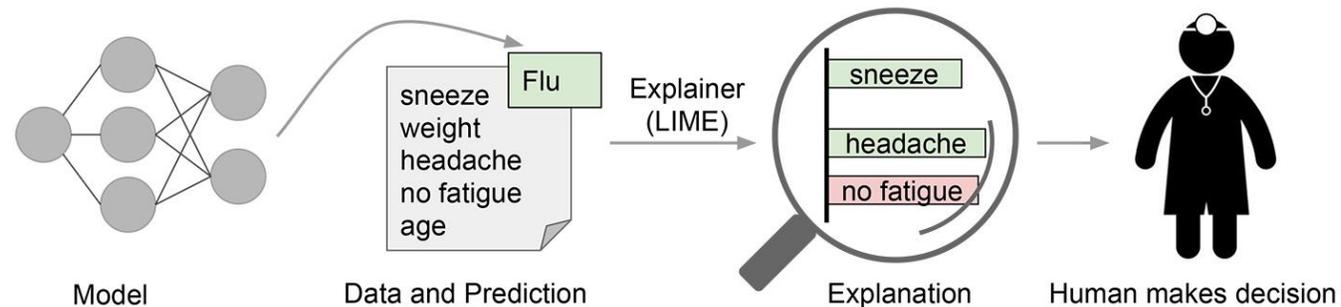
2018.10.29

1. Introduction
2. Local Interpretable Model-Agnostic Explanations
3. 코드
4. Sub-modular Pick (SP LIME)
5. 실험

- <https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime> : 설명 글
- <https://dreamgonfly.github.io/2017/11/05/LIME.html> 설명글2
- <https://github.com/marcotcr/lime> : 깃허브
- <https://arxiv.org/pdf/1602.04938.pdf> :논문 주소
- <https://www.youtube.com/watch?v=CY3t11vuuOM> 동영상

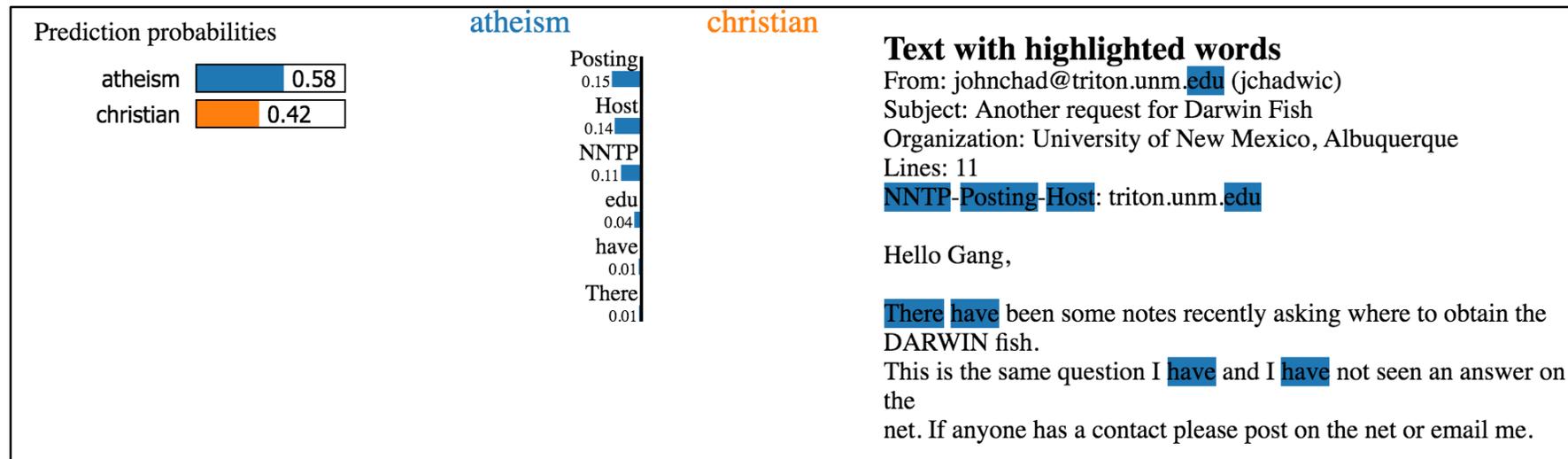
## 1. Introduction

- 머신 러닝 모델에 대해서 예측의 이유를 설명하는 것은 어렵다.
- 모델이 복잡해질수록 예측의 정확도는 올라가지만, 결과의 해석은 어려워 짐. (블랙박스라고 불리는 이유)
- 하지만 모델이 '왜' 그렇게 작동하는지 아는 것은 중요!!
- 의사가 "인공 지능이 이렇게 하래"라며 환자를 수술하지는 않겠죠
- 예측을 하는 것만으로 끝나지 않고, '**Explainer**'가 있어, 모델에서 가장 중요했던 변수를 강조해주면 우리는 모델을 신뢰하거나 또는 신뢰하지 않기로 결정하기가 좀 더 쉬울 것.



## 1. Introduction - 텍스트 예시

- 20 newsgroups 데이터셋 - '기독교', '무신론' 분류 문제
- 서로 많은 단어들을 공유하기 때문에 분류하기 어려움
- 하지만, 랜덤 포레스트 사용 시, 92.4%의 높은 정확도를 보임.
- 이때, LIME의 설명을 들여다보면 모델을 다시 보게 될 것.



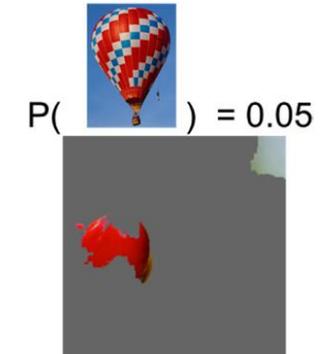
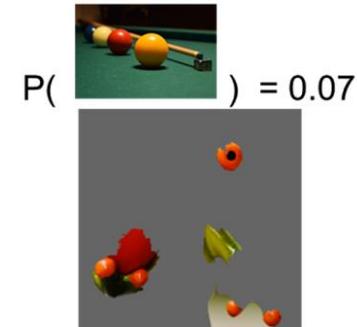
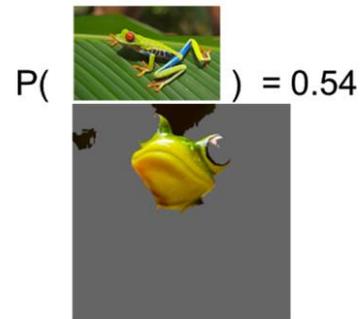
'Posting'이 기독교와 무신론자를 가르는 중요한 단어일까?

## 1. Introduction - 이미지 예시

- 구글의 Inception V3
- 주어진 이미지가 개구리일 확률이 가장 높다고 예측. 다음으로 당구대, 풍선을 좀 더 낮은 확률로 예측

- **Why?**

- 주어진 이미지를 나눠서 보면 눈과 손이 당구공을
- 개구리 손에 있는 하트는 풍선을 닮음



LIME을 통해 왜 모델이 이런 예측을 했는지 이해할 수 있다.

# CAN YOU BUILD YOUR TRUST BASED ON ACCURACY?

Only 1  
mistake!

 Predicted: <b>wolf</b> True: <b>wolf</b>	 Predicted: <b>husky</b> True: <b>husky</b>	 Predicted: <b>wolf</b> True: <b>wolf</b>
 Predicted: <b>wolf</b> True: <b>husky</b>	 Predicted: <b>husky</b> True: <b>husky</b>	 Predicted: <b>wolf</b> True: <b>wolf</b>

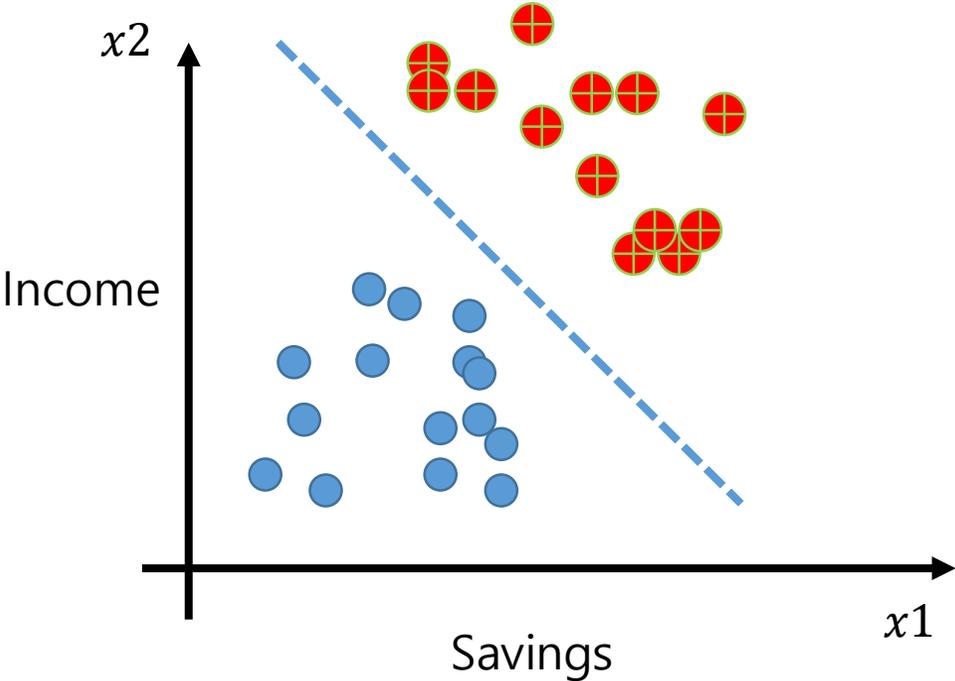


(a) Husky classified as wolf



(b) Explanation

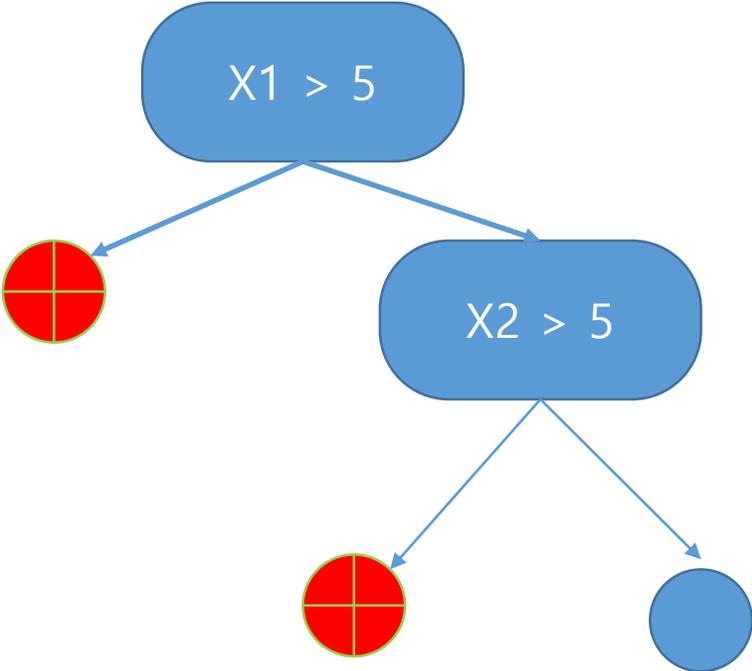
## 2. Local Interpretable Model-Agnostic Explanations



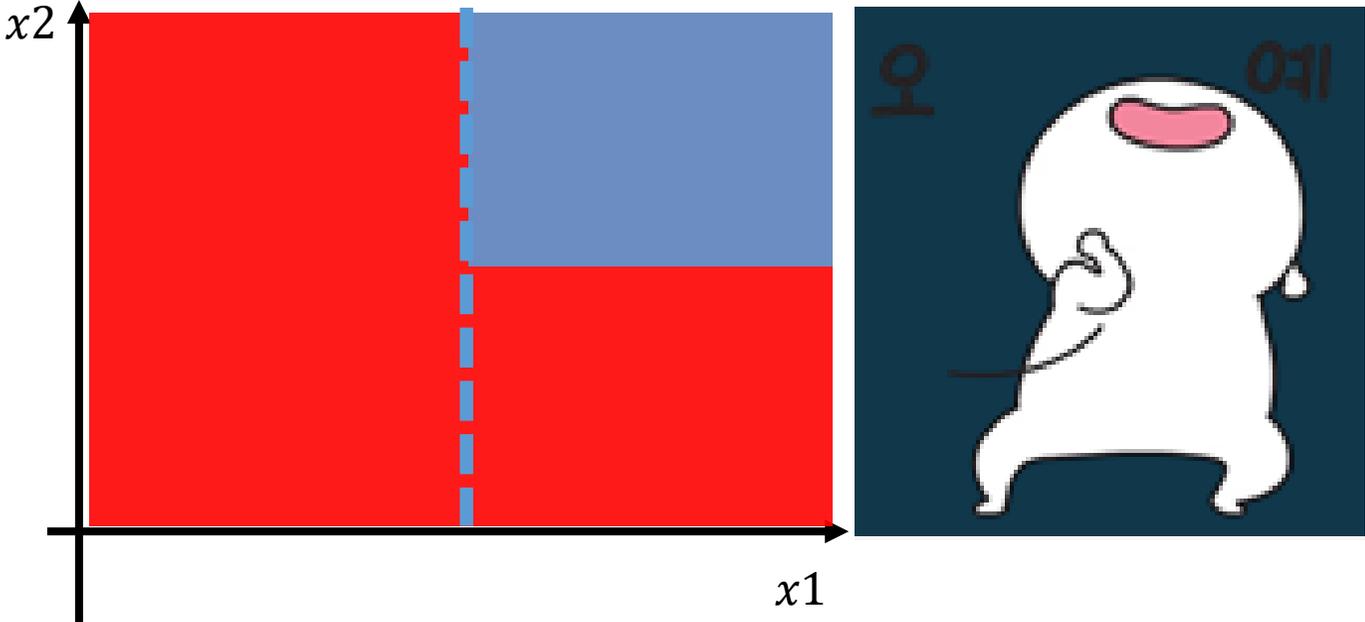
회귀 직선은 간단히 해석할 수 있다.

*If  $10x_1 + x_2 - 15 > 0$  then* ●  
*else* ●+

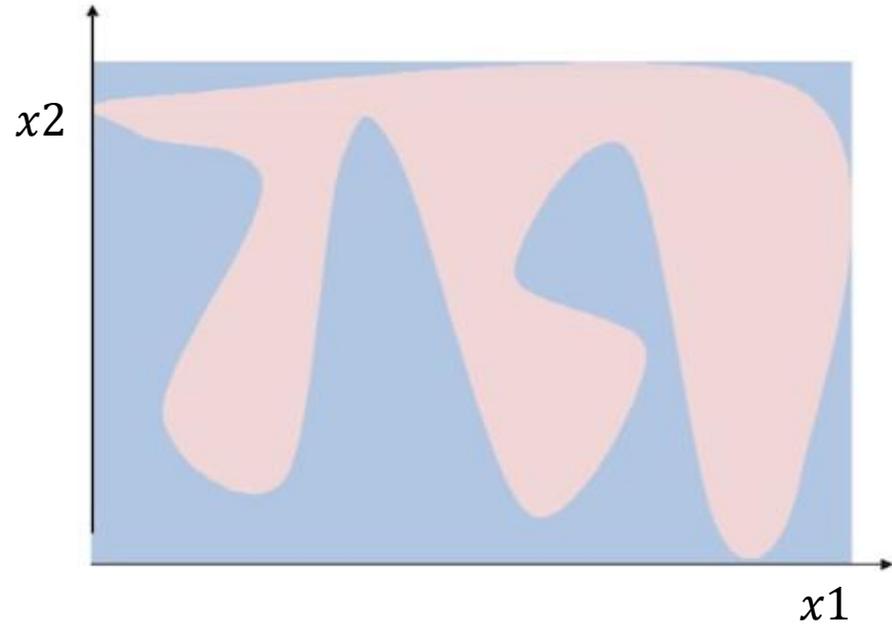
2. Local Interpretable Model-Agnostic Explanations



DT정도야~ 간단히 해석할 수 있다.

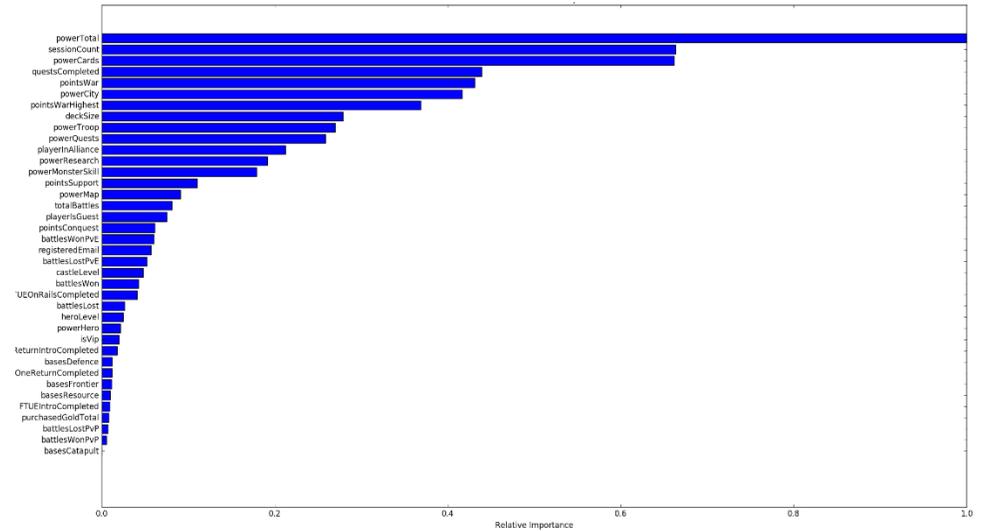
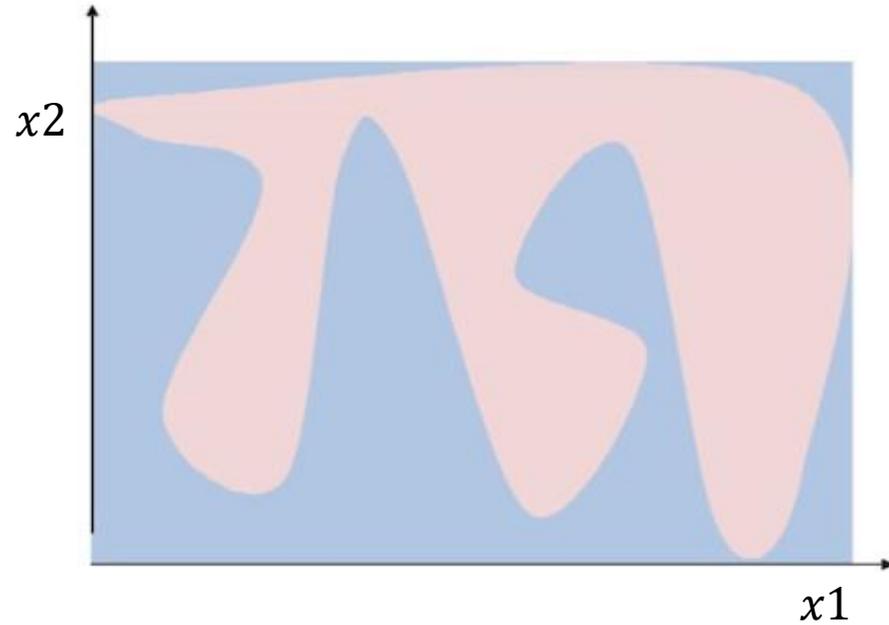


## 2. Local Interpretable Model-Agnostic Explanations



만일, 이렇게 복잡하고 해석하기 힘든 모델이라면?

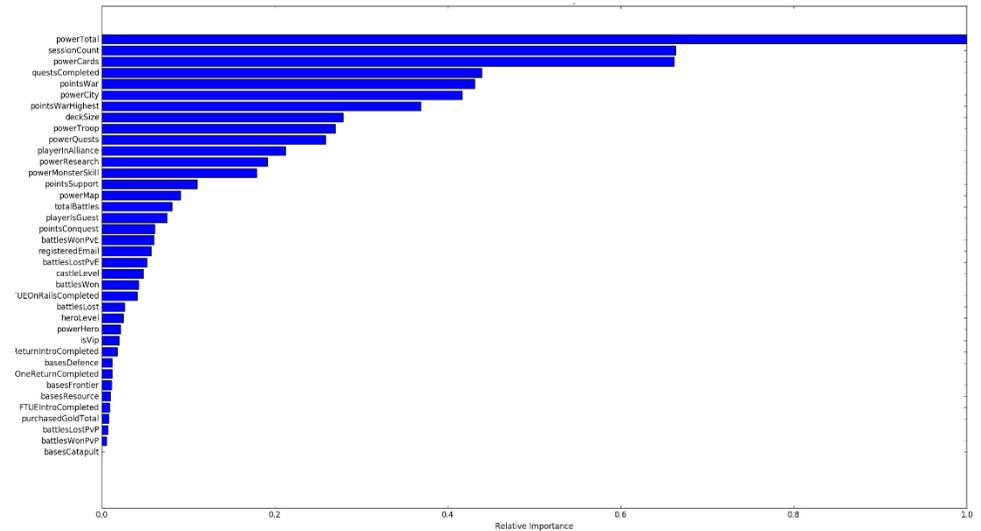
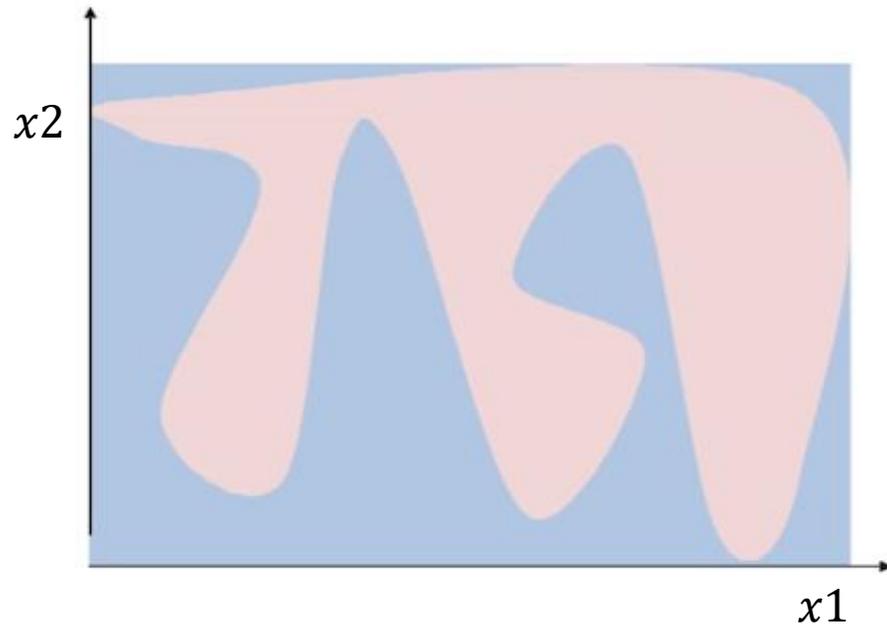
## 2. Local Interpretable Model-Agnostic Explanations



만일, 이렇게 복잡하고 해석하기 힘든 모델이라면?

- Feature importance만 그리지 말고
- **Local** 부분을 먼저 살펴 봅시다. ← Key idea

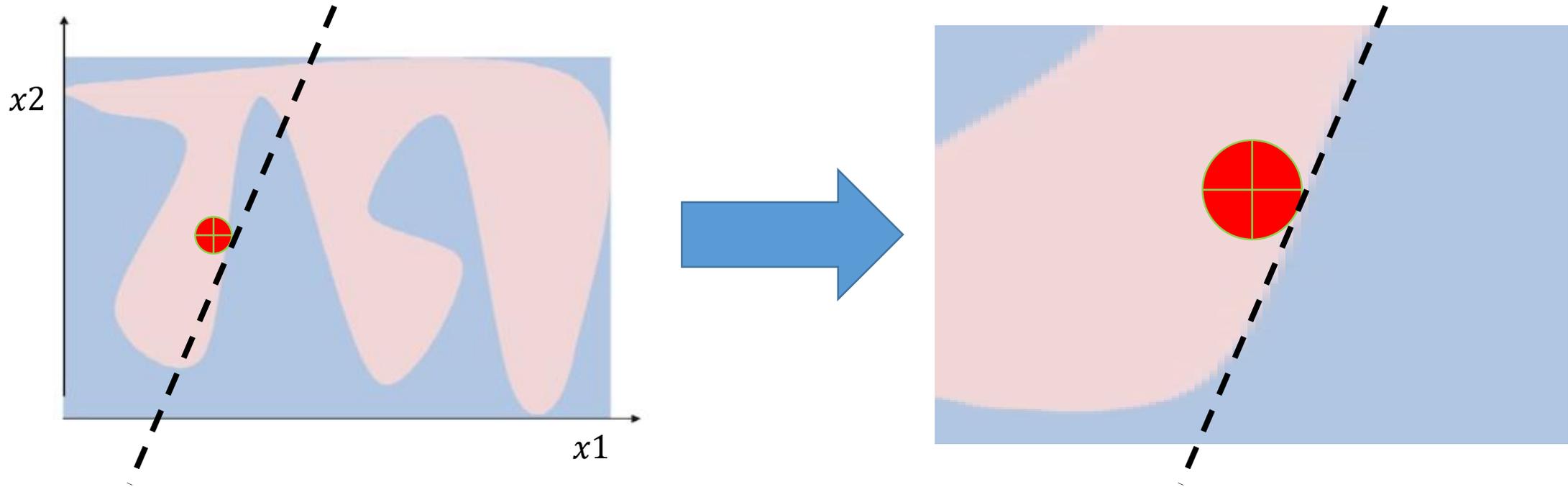
## 2. Local Interpretable Model-Agnostic Explanations



만일, 이렇게 복잡하고 해석하기 힘든 모델이라면?

- Feature importance만 그리지 말고
- **Local** 부분을 먼저 살펴 봅시다. ← Key idea

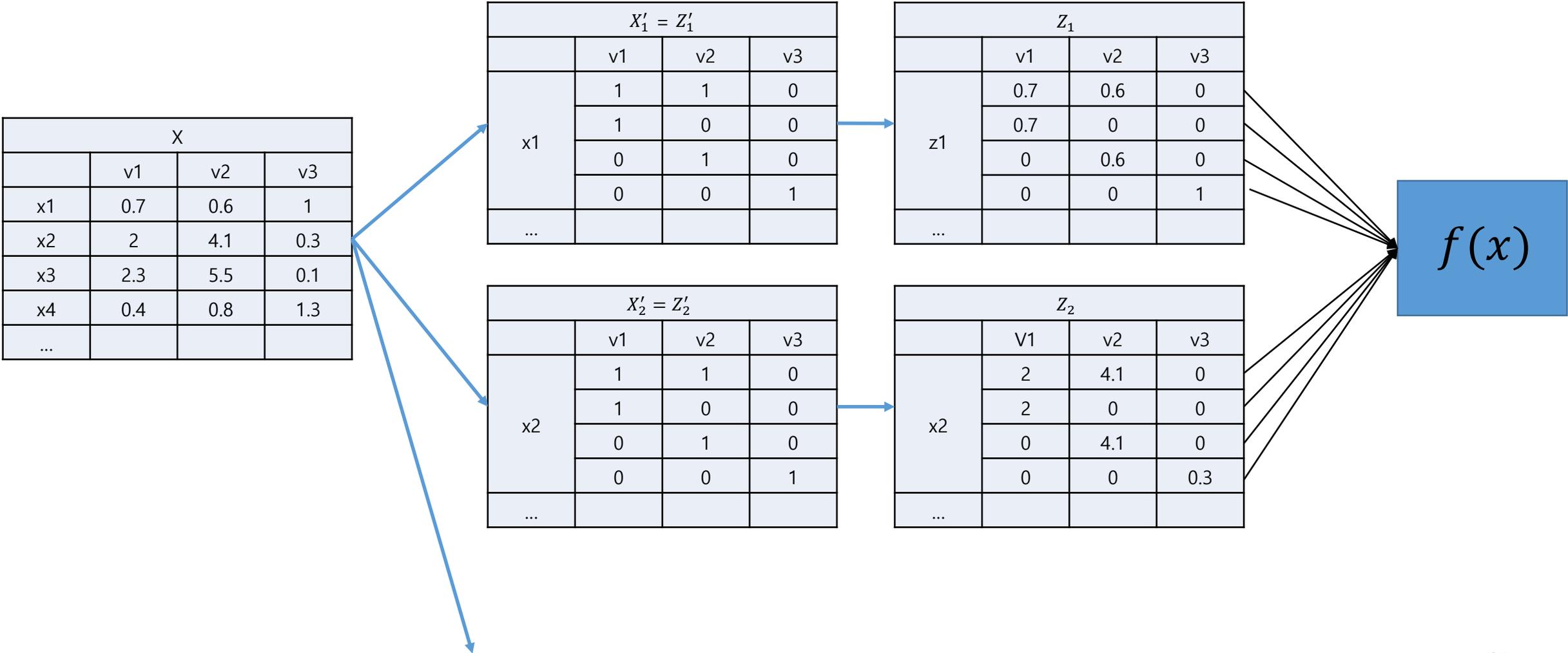
## 2. Local Interpretable Model-Agnostic Explanations



- 모델 전체로 봤을 때, 데이터에 대한 직선(해석가능한 모델 =  $g$ )은 매우 안 맞지만,
- Local에서는 매우 잘 맞을 것 (가정)
- 직선을 만들기 위해선, 주변 데이터들이 필요.

## 2. Local Interpretable Model-Agnostic Explanations

<주변 데이터 만들기 예시>



## 2. Local Interpretable Model-Agnostic Explanations

<주변 데이터 만들기 예시>

X			
	v1	v2	v3
x1	0.7	0.6	1
x2	2	4.1	0.3
x3	2.3	5.5	0.1
x4	0.4	0.8	1.3
...			

$X'_1=Z'_1$			
	v1	v2	v3
x1	1	1	0
	1	0	0
	0	1	0
	0	0	1
...			

$X'_2=Z'_2$			
	v1	v2	v3
x2	1	1	0
	1	0	0
	0	1	0
	0	0	1
...			

1. 사람이 해석 할 수 있는 interpretable representation vector를 만든다.

- Text의 경우  $x'$ 은 단어의 absent or presence 가 된다.
- 이미지의 경우 Super-pixel로 정의



Original Image



Interpretable Components

2. X에 맞게 recover 한 뒤, 모델에 넣어 label을 만든다.
3. sample Z는 X와의 거리에 따라 가중치를 매김  $\pi_x(z)$
4. 해석 가능한 모델  $g \in G$ 를 이용해  $g(z')$ 을 만든다  
(논문은 Lasso만 사용해 K개의 변수를 이용)
5. 따라서, Loss는 다음과 같이 정의된다.

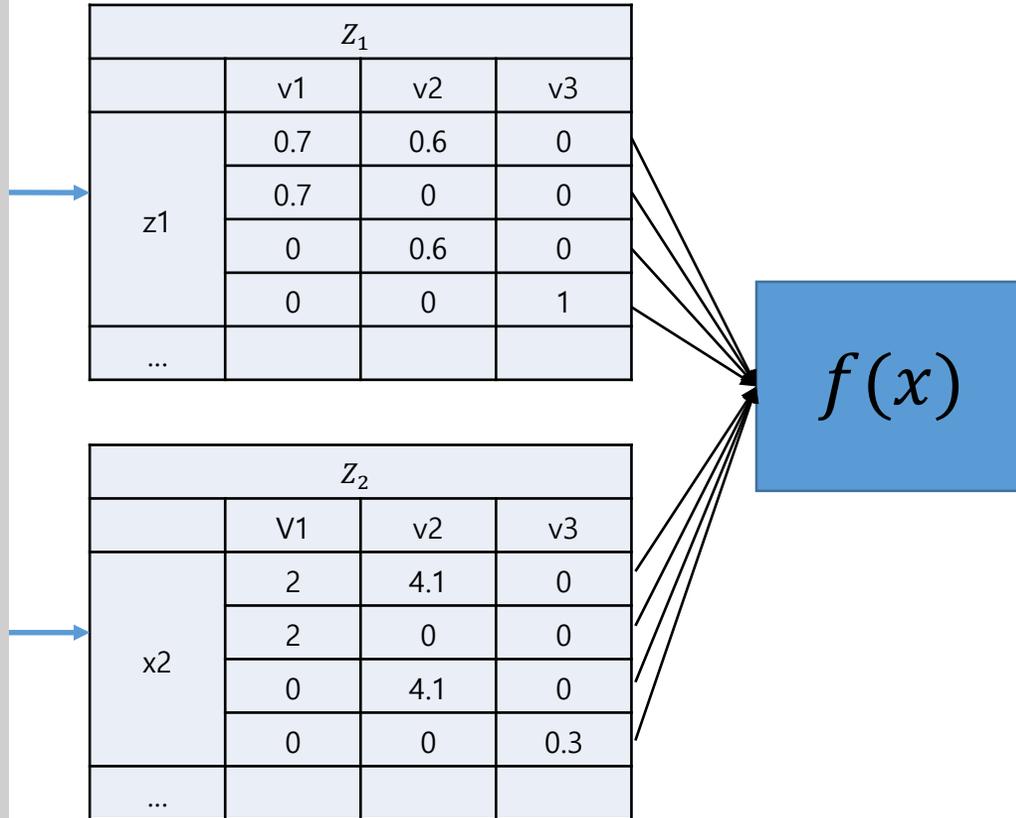
$$L(f, g, \pi_x) = \sum_{z, z'} \pi_x(z) \{f(z) - g(z')\}^2$$

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

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

모델의 설명은  $L$ 과  $\Omega$ 를 최소로 하는 모델  $g$ 를 찾는 것이다.

( $\Omega$ 는 Lasso의 경우 0이 아닌 weight의 수 Tree의 경우 depth에 해당)



## 2. Local Interpretable Model-Agnostic Explanations

### 장점

1. Model-agnostic(모델에 관계없다) 하다
2. 해석이 용이하다.
  - 1) 직선을 찾았을 때, (g)
  - 2) 회귀계수를 사용할 수 있음(영향력 파악)

### 단점

1. Sample Z가 많아질 수록 속도가 느려짐
2. 매우 비선형 모델이라면 g가 잘 작동할지 의문  
(iris 데이터에서 실제로 발생)

### 3. 코드

---

## 4. Submodular Pick for Explaining Models – SP-LIME

- 단일 예측의 설명은 사용자에게 모델에 대한 이해를 제공하지만, 모델 전체에 대한 신뢰도를 평가하기에는 충분하지 않다.
- 우리는 개별 사례를 설명함으로써 그 모델에 대한 전체적인 이해를 제안한다. ← model agnostic 하게  
→ 데이터 셋이 크다면 많은 수의 설명을 검토할 수 없다.
- 따라서  $B$ 개의 instance만을 살펴본다. (최소  $B$ 개는 모델을 이해할 때 꼭 봐야 함) ← *Pick step*
- $B$ 개를 고를 때는
  - 1) 모델을 설명할 수 있는 대표성이 있는 sample을 골라야 함
  - 2)  $B$ 개는 다양한 특성을 지닌 sample을 포함 하도록 뽑는다
- SP-LIME에 의해 뽑힌 서브셋의 Explanation을 보고 모델의 전반적인 작동 방식을 신뢰할 것인지 평가

## 4. Submodular Pick for Explaining Models – SP-LIME

<iris 예시>

	exp number	petal length (cm)	petal width (cm)	sepal length (cm)	sepal width (cm)
setosa	0	-0.082228	-0.236923	-0.032294	0.012060
setosa	1	-0.090088	-0.260716	-0.025831	0.012220
setosa	2	-0.061403	-0.203016	-0.032648	0.008165
setosa	3	-0.115855	-0.326017	-0.043795	0.021527
setosa	4	-0.108132	-0.319248	-0.043905	0.014563
versicolor	0	-0.180320	0.074353	0.047606	-0.000603
versicolor	1	-0.130953	0.108325	0.049563	-0.012690
versicolor	2	-0.224952	0.035047	0.027544	-0.009744
versicolor	3	-0.009624	0.215826	0.085704	-0.021219
versicolor	4	-0.032447	0.207419	0.082044	-0.020580
virginica	0	0.262549	0.162570	-0.015312	-0.011457
virginica	1	0.221041	0.152391	-0.023731	0.000470
virginica	2	0.286355	0.167969	0.005104	0.001579
virginica	3	0.125479	0.110191	-0.041909	-0.000308
virginica	4	0.140578	0.111829	-0.038139	0.006017

	exp number	petal length (cm)	petal width (cm)	sepal length (cm)	sepal width (cm)
setosa	0	-0.082228	-0.236923	-0.032294	0.01206
setosa	1	-0.090088	-0.260716	-0.025831	0.01222
setosa	2	-0.061403	-0.203016	-0.032648	0.008165
setosa	3	-0.115855	-0.326017	-0.043795	0.021527
setosa	4	-0.108132	-0.319248	-0.043905	0.014563
versicolor	0	-0.18032	0.074353	0.047606	-0.000603
versicolor	1	-0.130953	0.108325	0.049563	-0.01269
versicolor	2	-0.224952	0.035047	0.027544	-0.009744
versicolor	3	-0.009624	0.215826	0.085704	-0.021219
versicolor	4	-0.032447	0.207419	0.082044	-0.02058
virginica	0	0.262549	0.16257	-0.015312	-0.011457
virginica	1	0.221041	0.152391	-0.023731	0.00047
virginica	2	0.286355	0.167969	0.005104	0.001579
virginica	3	0.125479	0.110191	-0.041909	-0.000308
virginica	4	0.140578	0.111829	-0.038139	0.006017

## 4. Submodular Pick for Explaining Models – SP-LIME

SP-LIME의 접근법을 수식으로 설명하기 위해 몇가지 정의를 짚어본다.

- $W : n \times d'$  **Explanation Matrix.**
  - 각 인스턴스의 interpretable component의 local importance
  - 예를 들어 linear model을 explanation으로 사용한 경우  $W_{ij} = |w_{g_{ij}}|$
- $I_j : W$ 의  $j$ 번째 컬럼의 **global importance**
  - 많은 인스턴스를 설명할수록 global importance가 높다고 함
  - linear model의 예에서는  $I_j = \sqrt{\sum_{i=1}^n w_{ij}}$ 로 정의 가능
- **coverage**  $c(V, W, I)$ 
  - $W, I$ 가 주어졌을때 set  $V$ 에 속한 인스턴스에 한번이라도 나타난 feature의 total importance

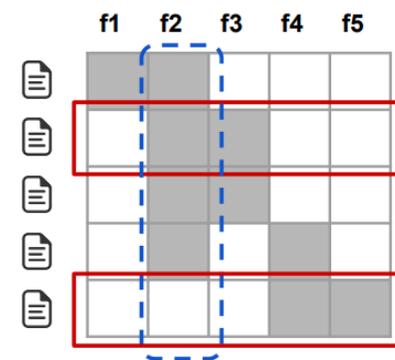


Figure 5: Toy example  $W$ . Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

### 다음 3가지를 실험

- (1) Are the explanations faithful to the model,  
→ 모델을 충실히 설명했는지.
- (2) Can the explanations aid users in ascertaining trust in predictions,  
→ 설명이 예측에 대한 신뢰를 확인하는 데 도움이 될 수 있는가?
- (3) Are the explanations useful for evaluating the model as a whole  
→ 설명은 모델 전체를 평가하는데 유용한지

### 다음 3가지를 실험

- (1) Are the explanations faithful to the model,  
→ 모델을 충실히 설명했는지.
- (2) Can the explanations aid users in ascertaining trust in predictions,  
→ **설명이 예측에 대한 신뢰를 확인하는 데 도움이 될 수 있는가? (이미 확인)**
- (3) Are the explanations useful for evaluating the model as a whole  
→ 설명은 모델 전체를 평가하는데 유용한지

### 다음 3가지를 실험

(1) Are the explanations faithful to the model,

→ **모델을 충실히 설명했는지.**

(2) Can the explanations aid users in ascertaining trust in predictions,

→ 설명이 예측에 대한 신뢰를 확인하는 데 도움이 될 수 있는가?

(3) Are the explanations useful for evaluating the model as a whole

→ 설명은 모델 전체를 평가하는데 유용한지

### HOW?

Books, DVD 대여 데이터를 이용해 감성분석을 각각 실시 (n=2,000)

- Train = 1600, Test = 400
- Classify product review as **Positive** VS **Negative**
- Decision trees (**DT**),
- Logistic regression with L2 regularization (**LR**),
- Nearest neighbors (**NN**),
- Support vector machines with RBF kernel (**SVM**),
- Random forests (with 1000 trees) (**RF**),
- All using bag of words as features with the average word2vec embedding



Use Sklearn -default parameter

## 5. 실험

### HOW?

Books, DVD 대여 데이터를 이용해 감성분석을 실시 (n=2,000)

#### ✓ 평가방식

- ✓ 400개의 테스트 셋에서 중요한 단어는 이미 파악하고 있음
- ✓ 각 모델 별 중요하다고 생각되는 10개의 단어를 추출

- ✓ Average Recall 값을 계산  $\left(\frac{\text{Predict True}}{\text{실제 True}}\right)$

**Random** : 변수(단어)를 랜덤으로 선정

**Greedy** : remove features that contribute the most to the predicted class until the prediction changes

**Parzen**: ... 커널 밀도 추정방식? 아 돈 노..

( 관련자료 → [링크](#) )

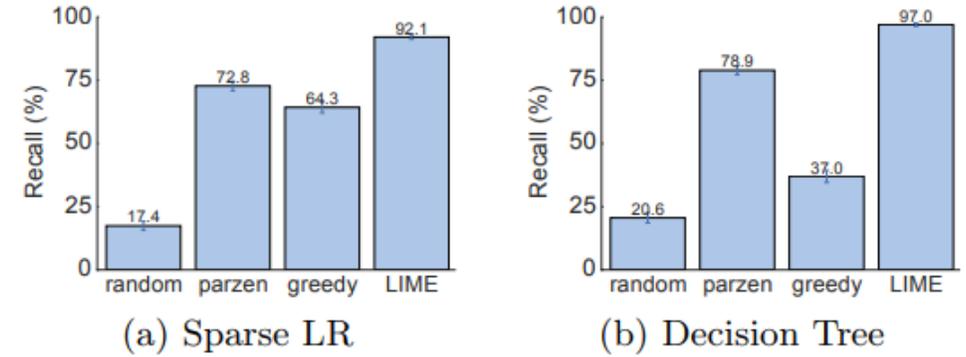


Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

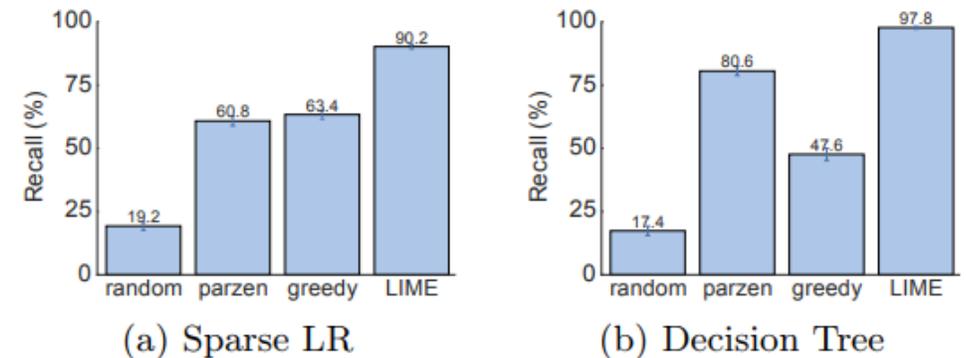


Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

### 다음 3가지를 실험

- (1) Are the explanations faithful to the model,  
→ 모델을 충실히 설명했는지.
- (2) Can the explanations aid users in ascertaining trust in predictions,  
→ 설명이 예측에 대한 신뢰를 확인하는 데 도움이 될 수 있는가?
- (3) Are the explanations useful for evaluating the model as a whole  
→ **설명**은 **모델 전체를 평가하는데 유용한지**

## 5. 실험

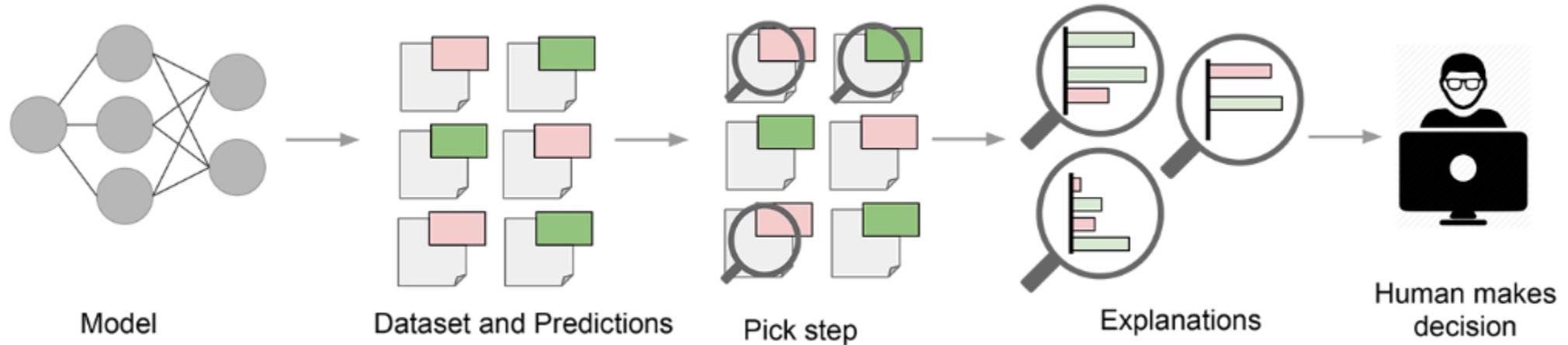
### HOW?

- Text 중, 신뢰할 수 없는 단어(변수)를 25%를 지정
- 4가지 방법별로 중요한 단어가 이들을 포함했는지 확인
- F1 score로 측정

**Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.**

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	<b>96.6</b>	<b>94.5</b>	<b>96.2</b>	<b>96.7</b>	<b>96.6</b>	<b>91.8</b>	<b>96.1</b>	<b>95.6</b>

## 5. Summary



**LIME:** 모델의 개별 예측값을 설명하기 위한 알고리즘  
복잡한 모델을 해석이 가능한 심플한 모델으로 locally approximation을 수행하여 설명

**SP-LIME:** 모델 자체의 신뢰 문제를 풀기위해 대표적인 인스턴스를 선택. 중복되는 정보들을 담고있는 인스턴스들은 제외하고 중요 정보를 담고 있는 소수의 인스턴스를 추려내는 과정

## 6. SHAP



Q.

A,B,C 세 사람은 택시를 타고 집으로 가려고 하는데..

택시 하나를 이용해 집을 가려고 한다.

각자 내는 요금은 위 그림과 같고 택시 하나를 이용하면, 42원 밖에 들지 않는다면

서로 얼마만큼의 이득을 분배해야 될까?

# 6. SHAP



## Contributions

요금	인원
6	{A}
12	{B}
42	{C}
12	{A,B}
42	{A,C}
42	{B,C}
42	{A,B,C}

## Shapley Value Calculation

경우의 수 (순열)	A	B	C
(A,B,C)	6	6	30
(A,C,B)	6	0	36
(B,A,C)	0	12	30
(B,C,A)	0	12	30
(C,A,B)	0	0	42
(C,B,A)	0	0	42
$\pi$	2	5	35

모든 조합의 평균

- SHAP는 Y를 두고 변수들간 기여도를 측정하는 방식.
- LIME은 X의 국소적인 부분을 자세히 보는 방식

변수의 중요도를 측정하는 방식이 다르기 때문에 항상 SHAP > LIME인 상황은 아니다.

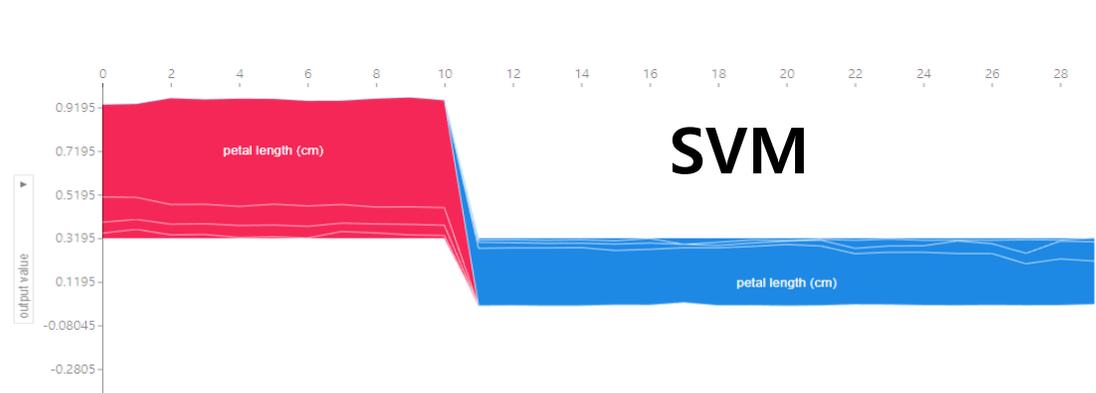
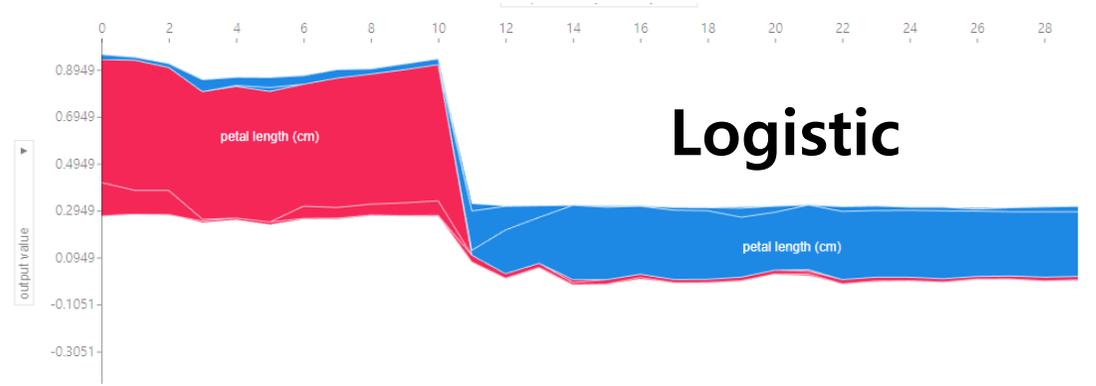
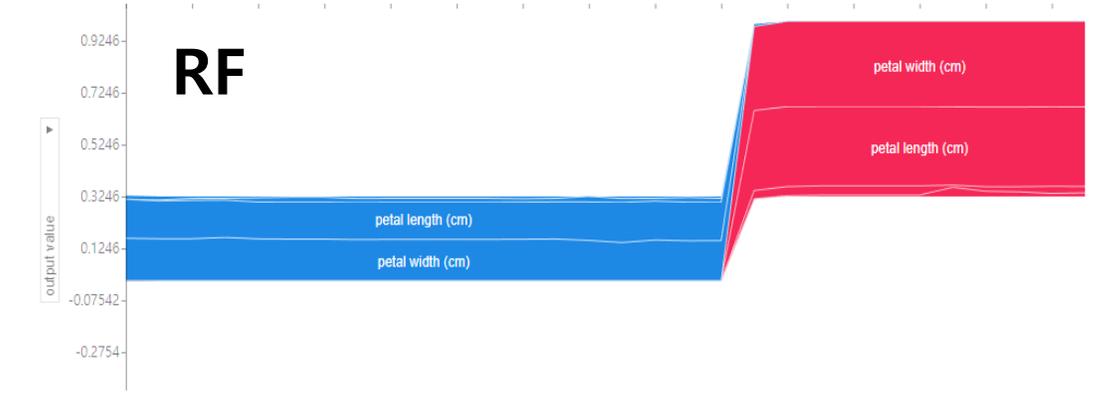
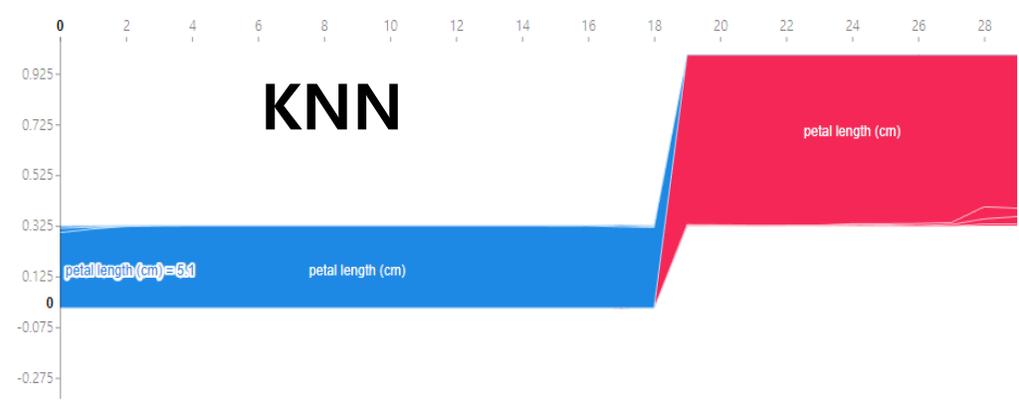
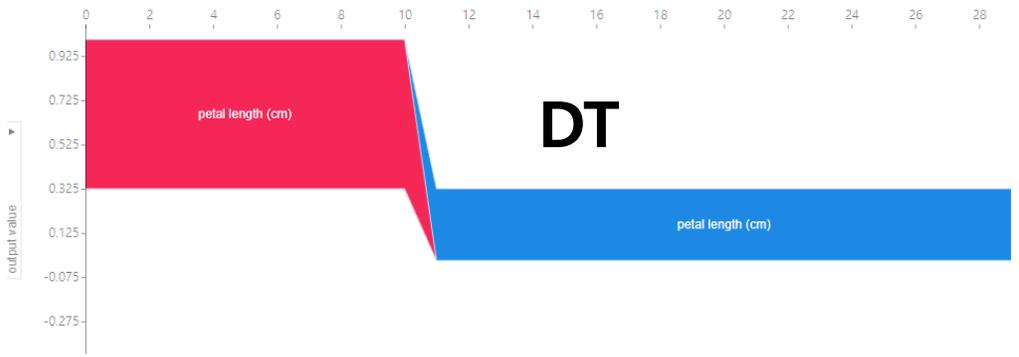
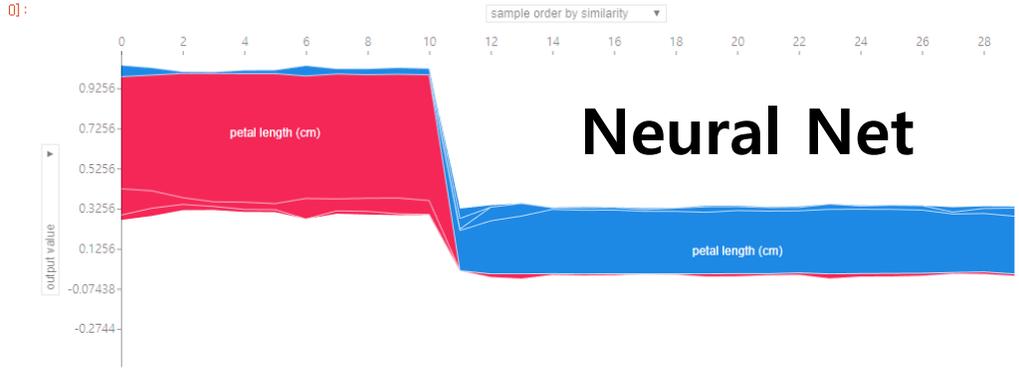
### LIME의 장점:

1) .

### SHAP의 장점:

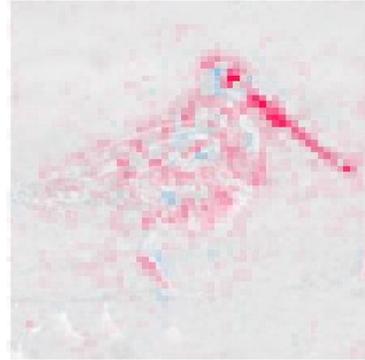
- 1) 교호작용 측정 가능
- 2) 만약  $V_1$  변수가 중요했다면, 얼마 이상이 되어 중요한지 측정 가능.
- 3) 해석이 비교적 쉽다 (개인적)
- 4) Model-Agnostic이지만, 모델별 비교 가능(다음 페이지)

0]:





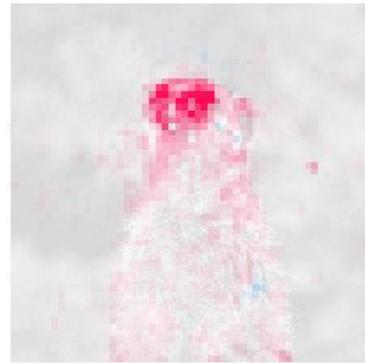
dowitcher



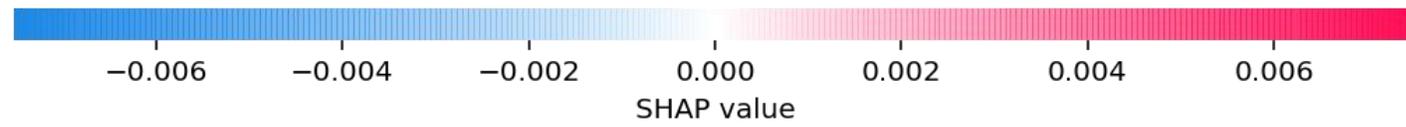
red-backed\_sandpiper



meerkat



mongoose



$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

요금	인원
6	{A}
12	{B}
42	{C}
12	{A,B}
42	{A,C}
42	{B,C}
42	{A,B,C}

$$1) S = \{B\} \quad \frac{1 \cdot (3-1-1)!}{3!} (v(A,B) - v(B)) = \frac{1}{6} (12 - 12)$$

$$2) S = \{C\} \quad \frac{1 \cdot (1)}{3!} (v(A,C) - v(C)) = \frac{1}{6} (42 - 42)$$

$$3) S = \{B,C\} \quad \frac{2! \cdot 1}{3!} (v(A,B,C) - v(B,C)) = \frac{2}{6} (42 - 42)$$

$$4) S = \emptyset \quad \frac{1 \cdot 2!}{3!} (v(A) - 0) = \frac{2}{6} (6 - 0)$$

$$\phi_A = 2$$