



Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI



이상용 / 2020-02-14



Computational Data Science LAB



CONTENTS

1. Introduction
 2. Explainability: What, Why, What For and How?
- 
- 

01 | Introduction

- ✓ 최근 몇 년 동안 Artificial Intelligence (AI)는 복잡한 문제에서도 전대미문의 성능을 보이고 있기 때문에 주목할 만한 움직임을 보이고 있음
- ✓ 좋은 성능을 보이는 AI는 인간의 'Decision'을 대신하기 위해서 여러 도메인에서 AI의 활용범위가 넓어지고 있음
- ✓ 하지만, 특정 분야에서는 직접 해석할 수 없고, 신뢰할 수 없는 기법을 사용하는 것을 주저함
- ✓ 따라서, AI의 결과를 해석하고 설명가능한 모델을 만들기 위해서 eXplainable AI (XAI)를 활발히 연구 중

02 | Explainability: What, Why, What For and How?

Terminology Clarification

- Interpretability
 - ✓ 주어진 모델이 의미를 제공하거나 설명하는 수동적인 특성
(모델이 어떠한 이유로 결과를 도출한 지 몰라도 결과에 대한 해석 가능)
- Explainability
 - ✓ 모델의 내부 기능을 명확히 하거나 모델에 의해 취해진 모든 조치나 절차를 나타내는 것으로 모델의 능동적
(Black-box에서 white-box로의 메커니즘 이해 → 실제로 무슨 일이 일어나는지 설명할 수 있는 것)

02 | Explainability: What, Why, What For and How?

What? & Why?

- What?

- ✓ XAI는 모델이 어떻게 기능하는지 이해하기 쉽도록 상세한 설명 또는 이유를 만들어내는 것

- Why?

- ✓ 첫 번째, research community와 business sectors의 gap차이
(프로세스의 디지털 전환에서 뒤쳐져 온 분야)
- ✓ 두 번째, 과학계와 사회에서는 performance에 의해서만 관심이 가져지는 것이 아님

02 | Explainability: What, Why, What For and How?

What for?

- What for?

XAI Goal	Main target audience (Fig. 2)	References
Trustworthiness	Domain experts, users of the model affected by decisions	[5, 10, 24, 32, 33, 34, 35, 36, 37]
Causality	Domain experts, managers and executive board members, regulatory entities/agencies	[35, 38, 39, 40, 41, 42, 43]
Transferability	Domain experts, data scientists	[5, 44, 21, 26, 45, 30, 32, 37, 38, 39, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85]
Informativeness	All	[5, 44, 21, 25, 26, 45, 30, 32, 34, 35, 37, 38, 41, 46, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 63, 64, 65, 66, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 86, 87, 88, 89, 59, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154]
Confidence	Domain experts, developers, managers, regulatory entities/agencies	[5, 45, 35, 46, 48, 54, 61, 72, 88, 89, 96, 108, 117, 119, 155]
Fairness	Users affected by model decisions, regulatory entities/agencies	[5, 24, 45, 35, 47, 99, 100, 101, 120, 121, 128, 156, 157, 158]
Accessibility	Product owners, managers, users affected by model decisions	[21, 26, 30, 32, 37, 50, 53, 55, 62, 67, 68, 69, 70, 71, 74, 75, 76, 86, 93, 94, 103, 105, 107, 108, 111, 112, 113, 114, 115, 124, 129]
Interactivity	Domain experts, users affected by model decisions	[37, 50, 59, 65, 67, 74, 86, 124]
Privacy awareness	Users affected by model decisions, regulatory entities/agencies	[89]

02 | Explainability: What, Why, What For and How?

- What for?
 - ✓ Trustworthiness: 어떤 문제에 직면했을 때 의도한대로 작동되는 것
 - ✓ Causality: 데이터 변수 간의 인과관계를 찾는 것
 - ✓ Transferability: ‘모델’ 내에서 일어나는 inner relations의 이해만으로 다른 문제에서 사용가능한 것
 - ✓ Informativeness: ‘모델’의 inner relations을 추출하기 위한 것 (의사결정 지원)
 - ✓ Confidence: 항상 robust하고 stable한 모델을 만들기 위한 것
 - ✓ Fairness: 노출되었던 데이터의 편향을 강조하는 것 (인간의 생명을 수반하는 분야에서 중요)
 - ✓ Accessibility: 사용자의 접근성으로, 이해할 수 없는 알고리즘을 비전문적 사용자가 다룰 때 부담을 줄이는 것
 - ✓ Interactivity: 사용자와 상호작용 할 수 있는 것
 - ✓ Privacy awareness: Privacy를 평가하는 것 (권한이 없는 제3자의 모델이 설명하는 능력은 privacy 침해 가능)

02 | Explainability: What, Why, What For and How? How?

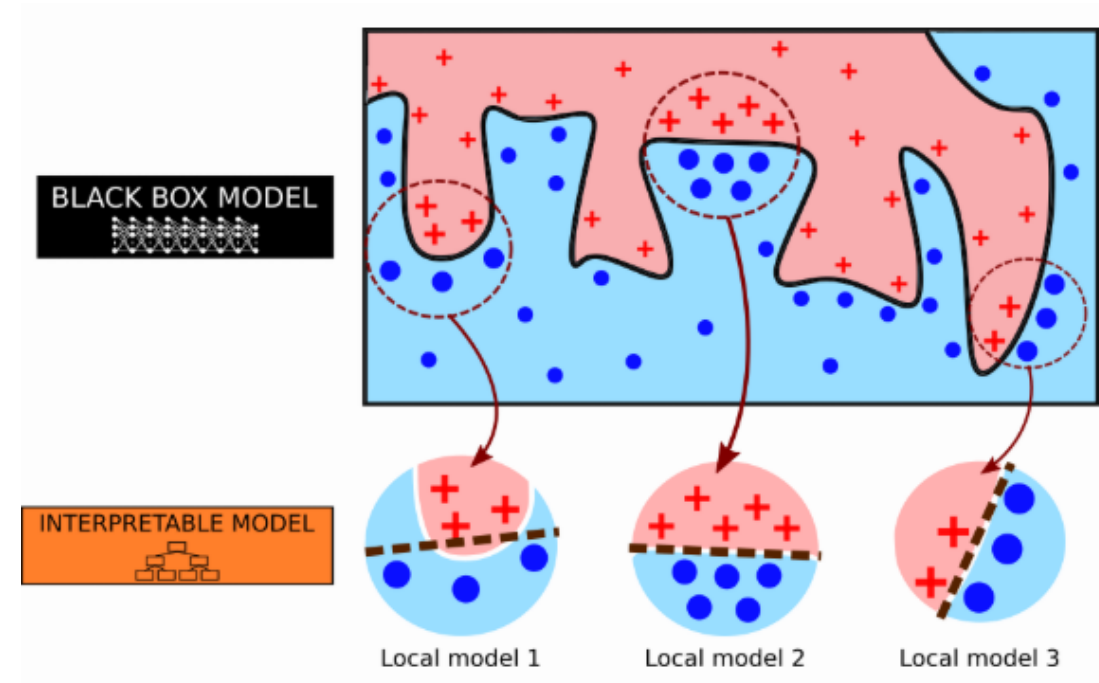
- Intrinsic vs. Post-hoc

- ✓ 모델 자체적으로 해석력을 확보하고 있을 때 'Transparency' 를 가지고 있다고 표현 함
- ✓ 내재적으로 transparency를 가지고 있는 모델: 'Intrinsic'
- ✓ Ex) 의사결정나무

- ✓ 복잡성이 높은 모델은 학습이 끝난 뒤 모델에 대한 해석이 가능: 'Post-hoc'
- ✓ Ex) 신경망

02 | Explainability: What, Why, What For and How? How?

- Model-specific vs. Model-agnostic
 - ✓ 특정한 종류의 모델에 작동: 'Model-specific'
 - ✓ 범용적으로 어떤 모델에도 작동: 'Model-agnostic'
- Local vs. Global
 - ✓ 모델의 예측 결과에 대해 전역적으로 설명: 'Global'
 - ✓ 모델의 예측 결과에 대해 국소적으로 설명: 'Local'

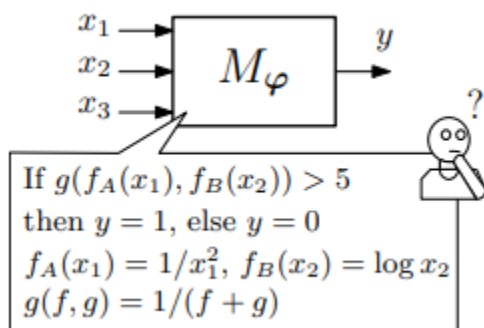


02 | Explainability: What, Why, What For and How?

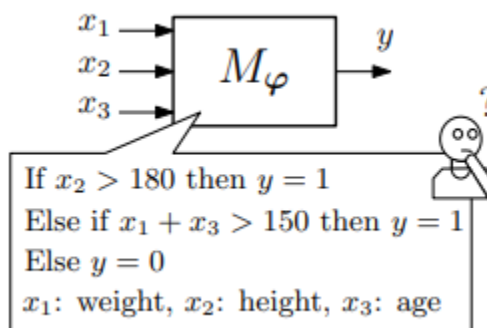
• *Levels of Transparency* in Machine Learning Models

- ✓ Simulatability: 모델의 시뮬레이션 능력
- ✓ Decomposability: 모델의 Input, parameter, calculation을 설명할 수 있는 능력 (모든 input이 쉽게 해석 가능)
- ✓ Algorithmic Transparency: 사용자가 모델의 input data로 부터 output까지의 process를 이해하는 능력

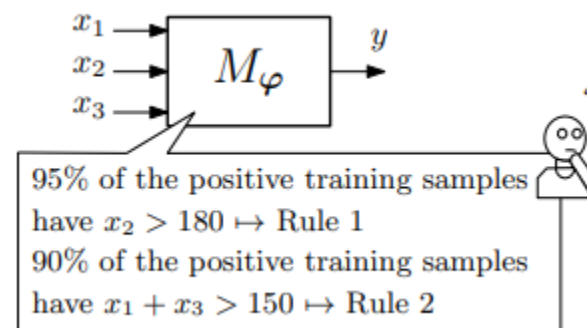
Ex) Linear model의 경우, 모델이 직면한 상황에 대하여 어떻게 작용할 지 알 수 있음



Simulatability



Decomposability



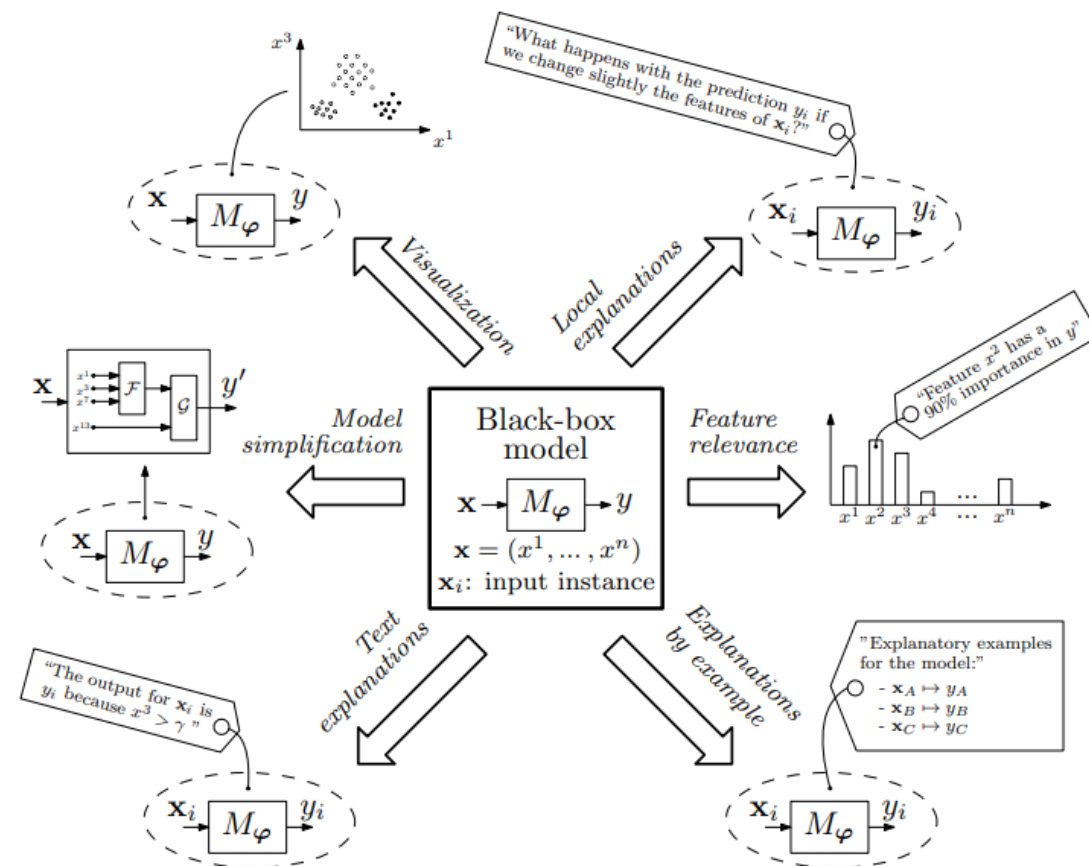
Algorithmic transparency

02 | Explainability: What, Why, What For and How? How?

- *Post-hoc Explainability* Techniques for Machine Learning Models
 - ✓ Text explanations: 모델의 결과를 설명하는데 도움이 되는 텍스트를 생성
 - ✓ Visual explanation: 모델의 동작을 시각화 하는 것 (dimensionality reduction techniques을 자주 사용)
 - ✓ Local explanations: 모델의 예측 결과에 대한 전체 솔루션 공간보다 덜 복잡한 하위 공간의 설명을 제공
 - ✓ Explanations by example: 모델의 동작을 설명하기 위해 데이터 셋에서 특정한 인스턴스를 선택하는 것
 - ✓ Explanations by simplification: 기존의 모델에 기초하여 새로운 시스템을 재구축 하는 것 (단순하고 유사한 성능)
 - ✓ Feature relevance explanation: relevance score(sensitivity)을 계산하여 모델 내부 기능을 명확히 하는 것

02 | Explainability: What, Why, What For and How? How?

- *Post-hoc Explainability* Techniques for Machine Learning Models



02 | Explainability: What, Why, What For and How?

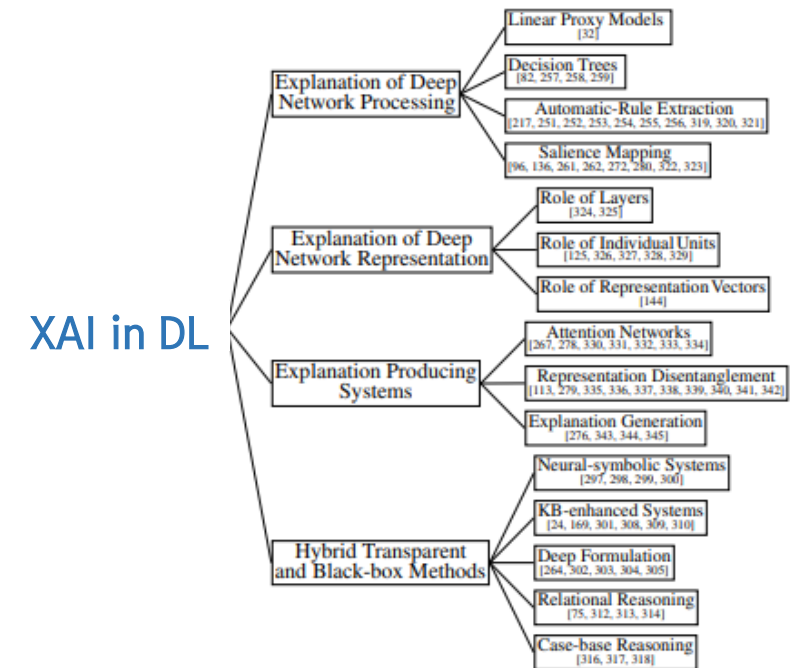
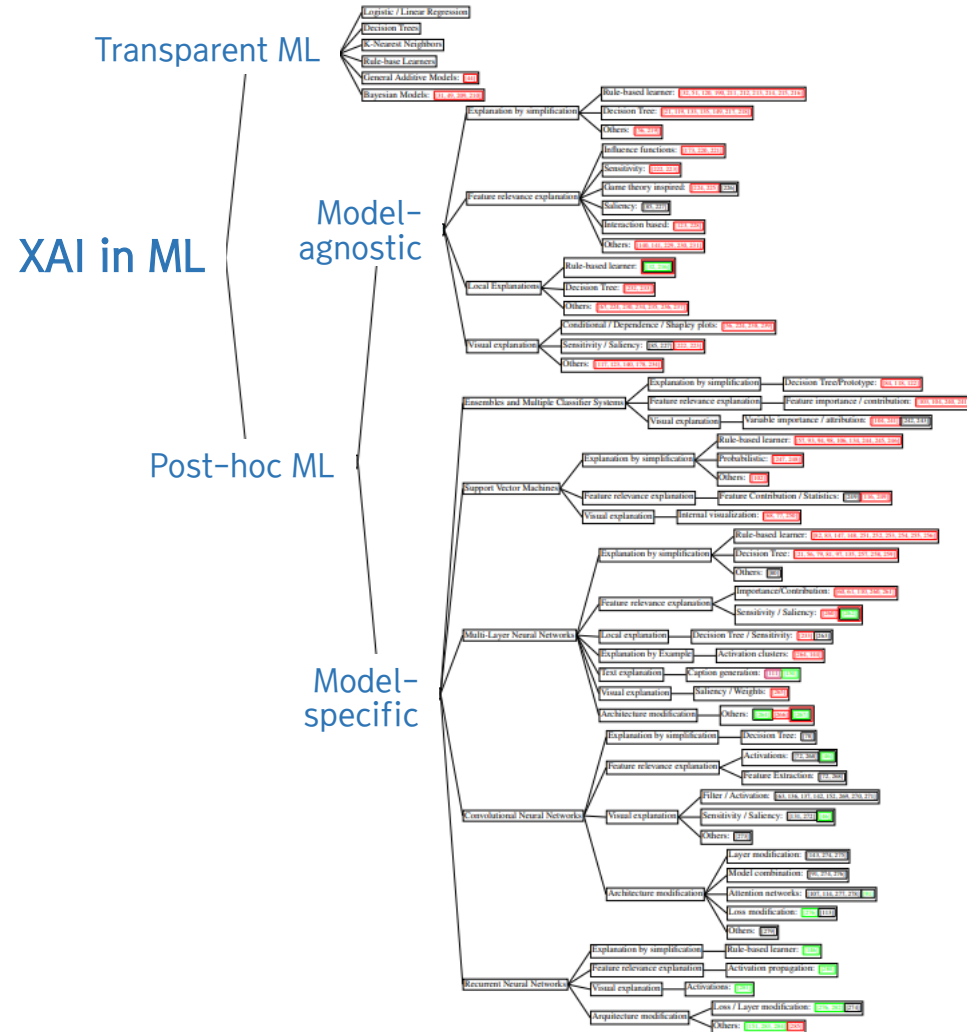
Transparent ML

Post-hoc ML

Model	Transparent ML Models			Post-hoc analysis
	Simulatability	Decomposability	Algorithmic Transparency	
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools	Not needed
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process	Not needed
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model	Not needed
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour	Not needed
General Additive Models	Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding	Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model	Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools	Not needed
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools	Not needed
Tree Ensembles	✗	✗	✗	Needed: Usually <i>Model simplification</i> or <i>Feature relevance</i> techniques
Support Vector Machines	✗	✗	✗	Needed: Usually <i>Model simplification</i> or <i>Local explanations</i> techniques
Multi-layer Neural Network	✗	✗	✗	Needed: Usually <i>Model simplification</i> , <i>Feature relevance</i> or <i>Visualization</i> techniques
Convolutional Neural Network	✗	✗	✗	Needed: Usually <i>Feature relevance</i> or <i>Visualization</i> techniques
Recurrent Neural Network	✗	✗	✗	Needed: Usually <i>Feature relevance</i> techniques



02 | Explainability: What, Why, What For and How?



Q&A

감사합니다.