Explainable Machine Learning

Introduction

Shim Jaewoong

jaewoong@seoultech.ac.kr

Course Introduction

Course Information

- 설명가능한 기계학습
  - '설명가능 기계학습' 관련 핵심적인 개념들을 다루며, 관련 방법론들이 현실에서 어떻게 응용되는지를 알아본다.
  - 학생들은 직접 논문을 읽고 발표하는 시간을 가지며, 그 내용에 대해 **논의**한다.
  - Research proposal을 수행한다.

- Prerequisites
  - Probability, statistics, linear algebra
  - Python programming
  - **Machine learning / Deep learning**

Course Information

- General Information
  - Class time : 월 7~9 (15:00PM ~ 18:00PM)
  - Location : 프론티어관 501호
  - Language : Korean

- Instructor
  - 심재웅, 다산관 208호, jaewoong@seoultech.ac.kr
  - Office hour : by appointment via e-mail

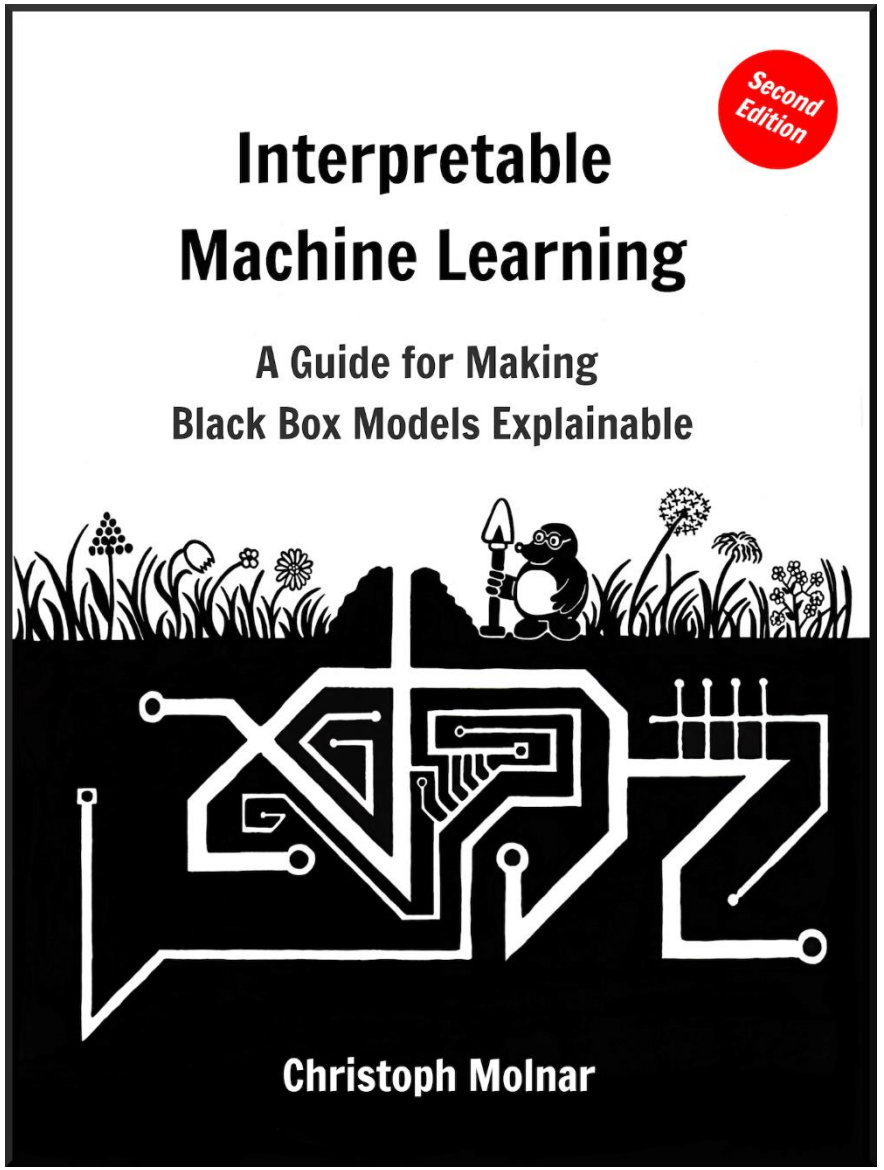
Course Schedule (tentative)

- 강의 + 논문
  - **강의**
    - Lecture Slide / online book
  - **논문 발표 & discussion**
    - 논문 List 제공 예정

** 프로그래밍 실습 X*

Course Information

- Main text book
  - **Interpretable Machine Learning (2nd edition), Christoph Molnar.**
    - Link: https://christophm.github.io/interpretable-ml-book/
  - Other papers …



Course Schedule (tentative)

주차	topics	비고
1	introduction	
2	Interpretability. (machine learning review)	
3	interpretable model. (machine learning review)	
4	Global model-agnostic	**Quiz**
5	local model-agnostic (LIME, SHAP, …)	
6	local model-agnostic (LIME, SHAP, …)	
7	Neural Network Interpretation	
8	Neural Network Interpretation	paper 선정
9	Additional topic, Q&A	
10	**Midterm exam**	
11	paper presentation & discussion	
12	paper presentation & discussion	팀구성
13	paper presentation & discussion	
14	paper presentation & discussion	
15	**팀별 Research proposal 발표**	

Grading

Attendance	Midterm exam	Review quiz	Research proposal	Paper discussion
10%	40%	10%	20%	20%

- Late submissions will NOT be accepted.
- Exams are closed-book and closed-note.
- Final grades will be assigned based on the overall class performance.

Paper Discussion

- 지정된 논문 리스트 중 선정하여 발표 (개인별)
- 개인당 25분 (15분 발표, 10분 질의 응답)
- 8주차에 논문 할당 예정
- E-class로 발표자료 제출

- *발표자가 아닌 모든 학생들도 논문을 읽고 참석해야 하며, 각자 리뷰를 작성하여 제출 (양식 제공 예정), 발표 후 질문과 답변을 통해 감점/가점*

Research Proposal (Team)

- 구성 : 최대 3인 1조
  - E-class를 통해 조 편성 예정

- 주제
  - '설명가능 인공지능'과 관련된 본인의 연구 주제 제안
  - 연구 배경, 연구 필요성과 목적, 관련 연구, 연구 방법론, 평가 방법, 기대 효과 등

- 평가
  - 주제의 참신성, 차별성, 연구 방법의 구체성, 내용 전달력 등

Academic Integrity

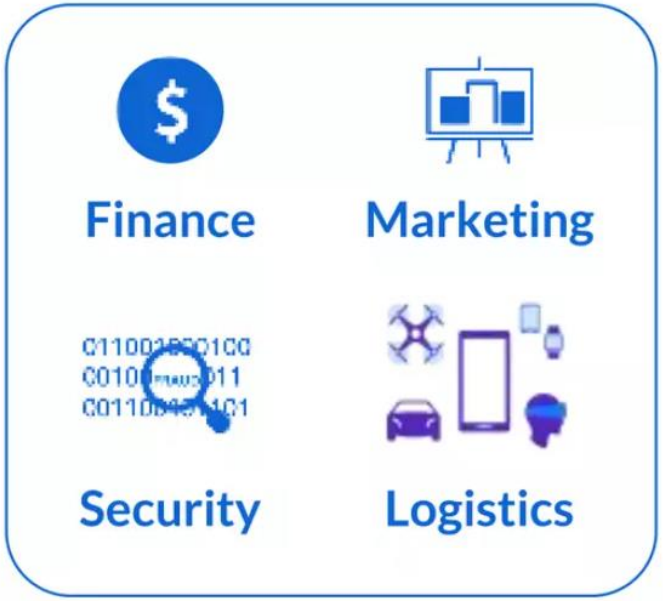
- Students are responsible for maintaining high standards of academic integrity in all of their class activities.
- Cheating or plagiarism in any form will not be tolerated.
- Any violation of academic integrity is a serious offense and is therefore subject to an appropriate sanction or penalty.

Intro

Need for Explainable ML

Current AI Systems

- Machine Learning centric today.
- ML Models are **opaque, non-intuitive** and **difficult to understand**

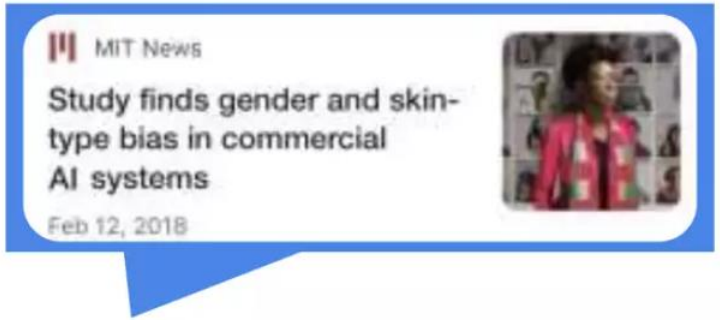


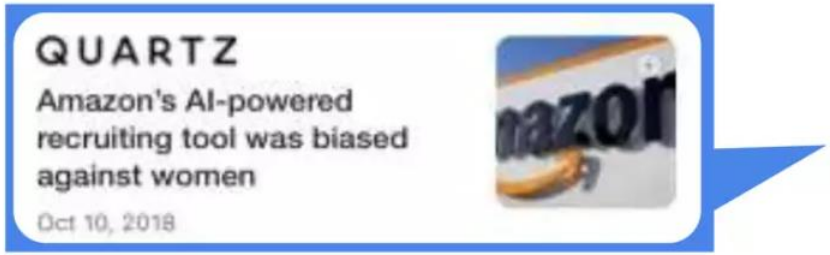
User

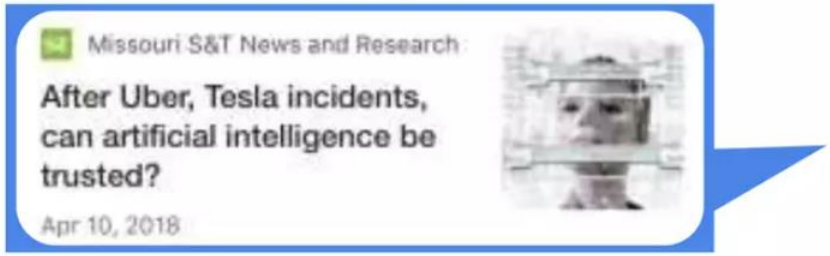
- Why did you do that?
- Why not something else?
- When do you succeed or fail?
- How do I correct an error?
- When do I trust you?

Explainable AI and ML *is essential for future customers to understand, trust, and effectively manage the emerging generation of AI applications*

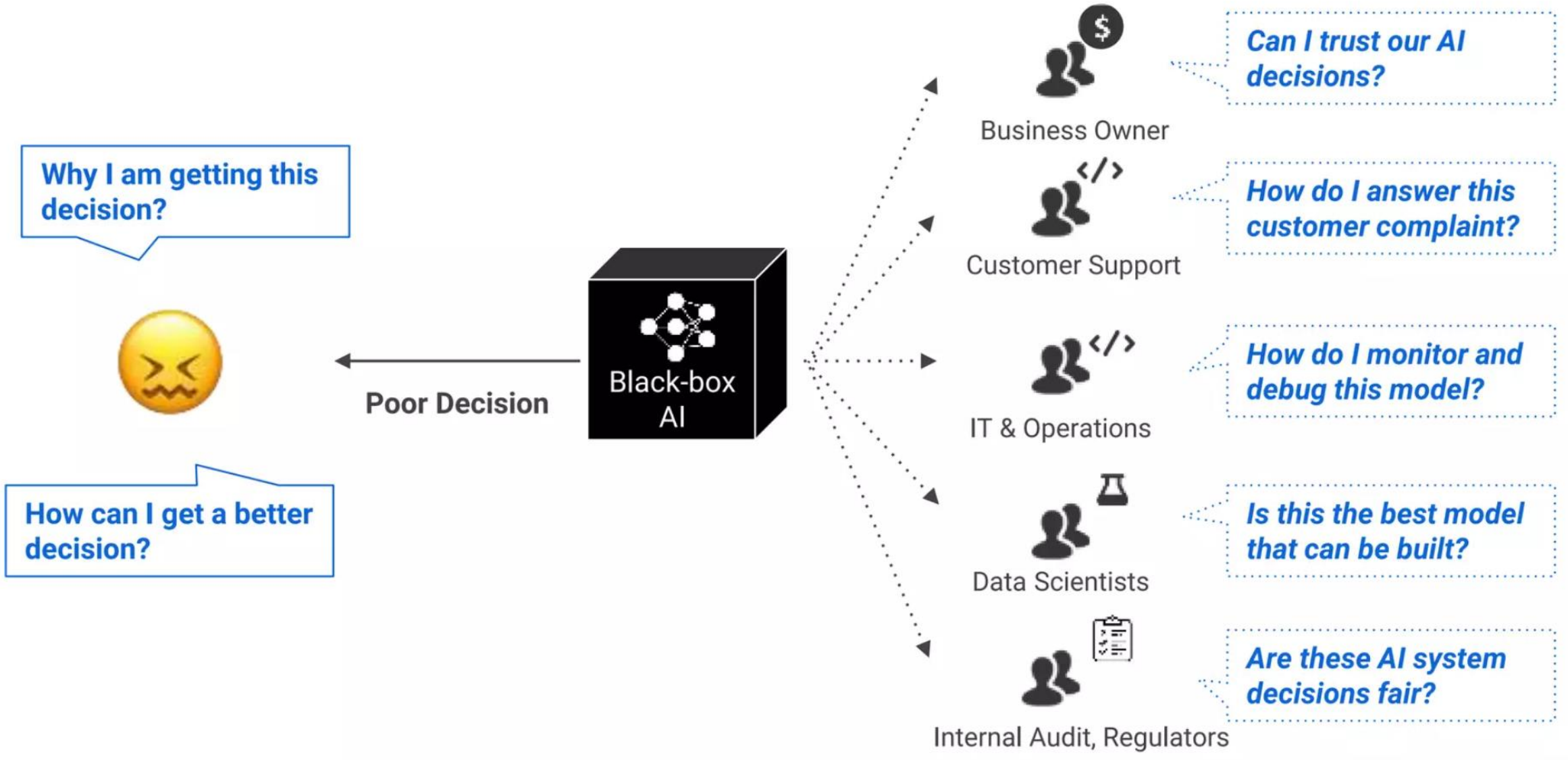
Black-box AI creates business risk for Industry





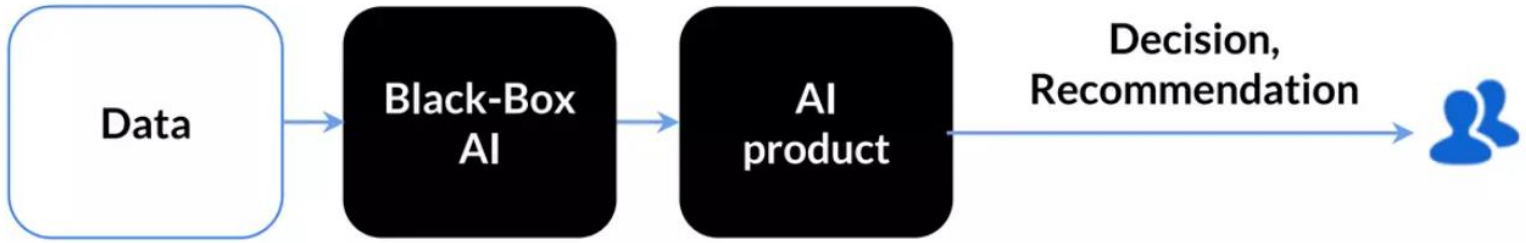


Black-box AI creates confusion and doubt



Explainable AI ?

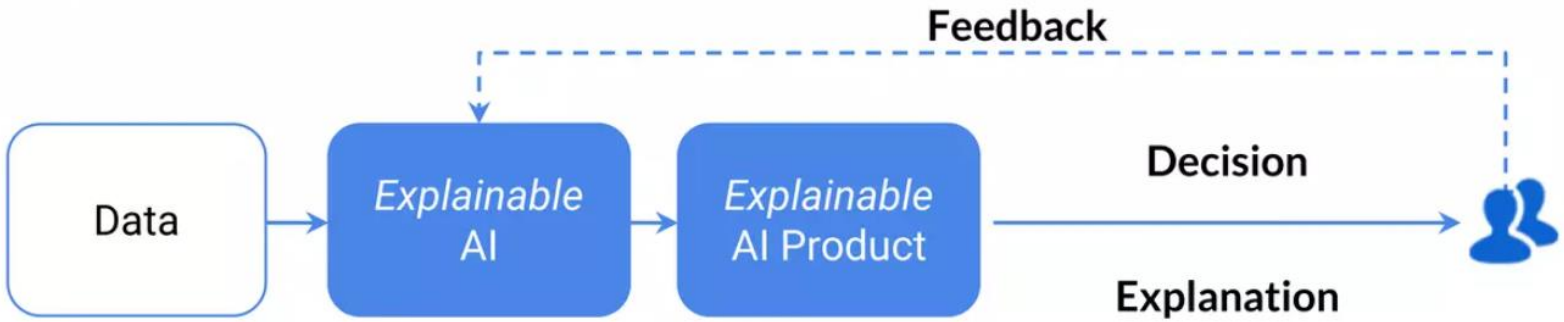
Black Box AI



Confusion with Today's AI Black Box

- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

Explainable AI



Clear & Transparent Predictions

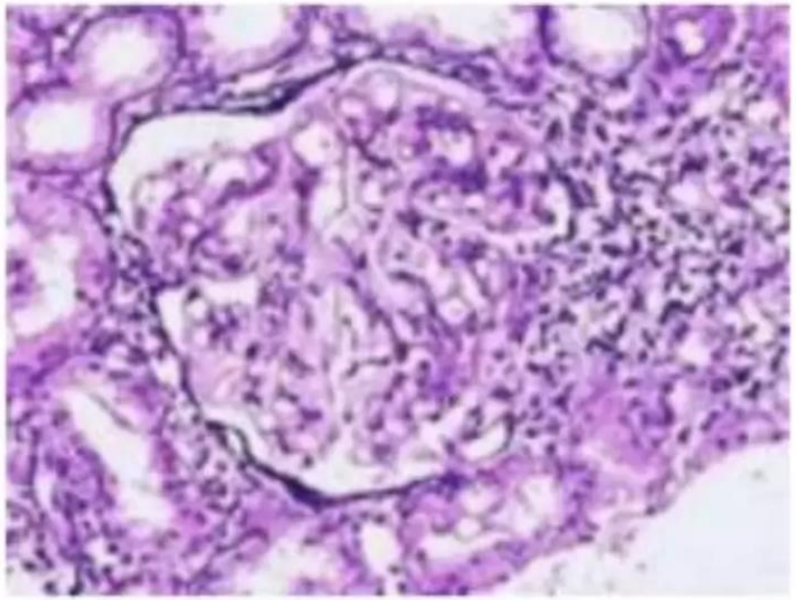
- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Why Explainability: Verify the ML model / system

Wrong decisions can be costly and dangerous

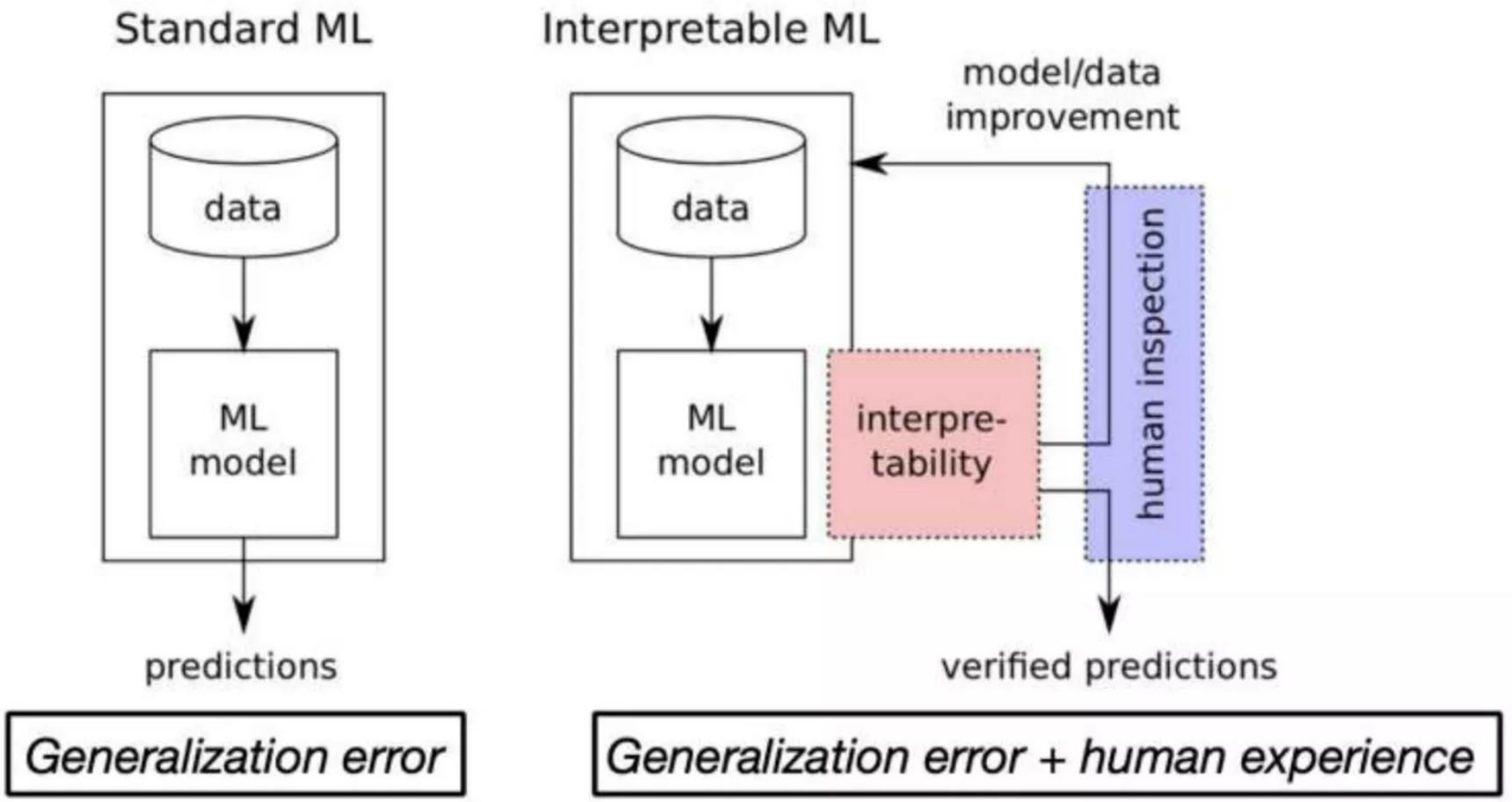
"Autonomous car crashes, because it wrongly recognizes …"

"AI medical diagnosis system misclassifies patient's disease …"



Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Improve ML model

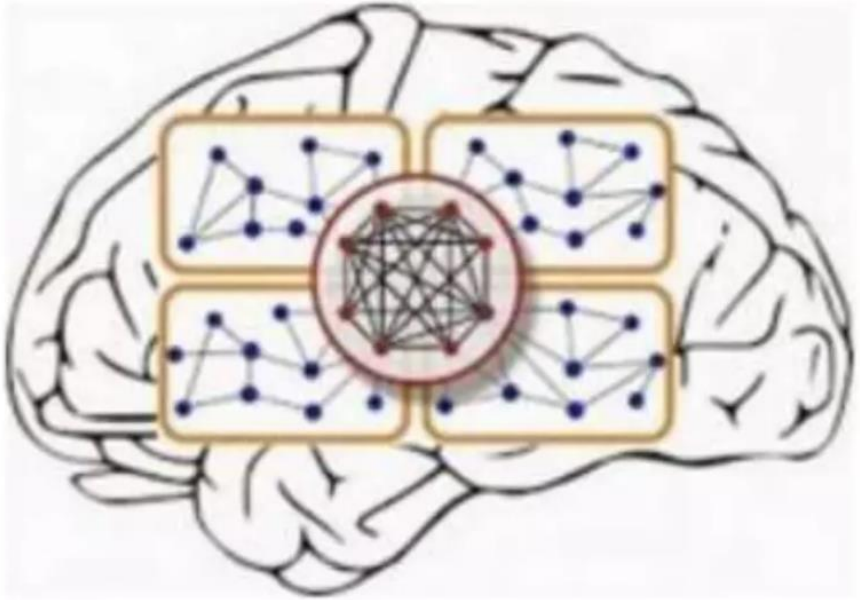


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Learn new insights

"It's not a human move. I've never seen a human play this move." (Fan Hui)

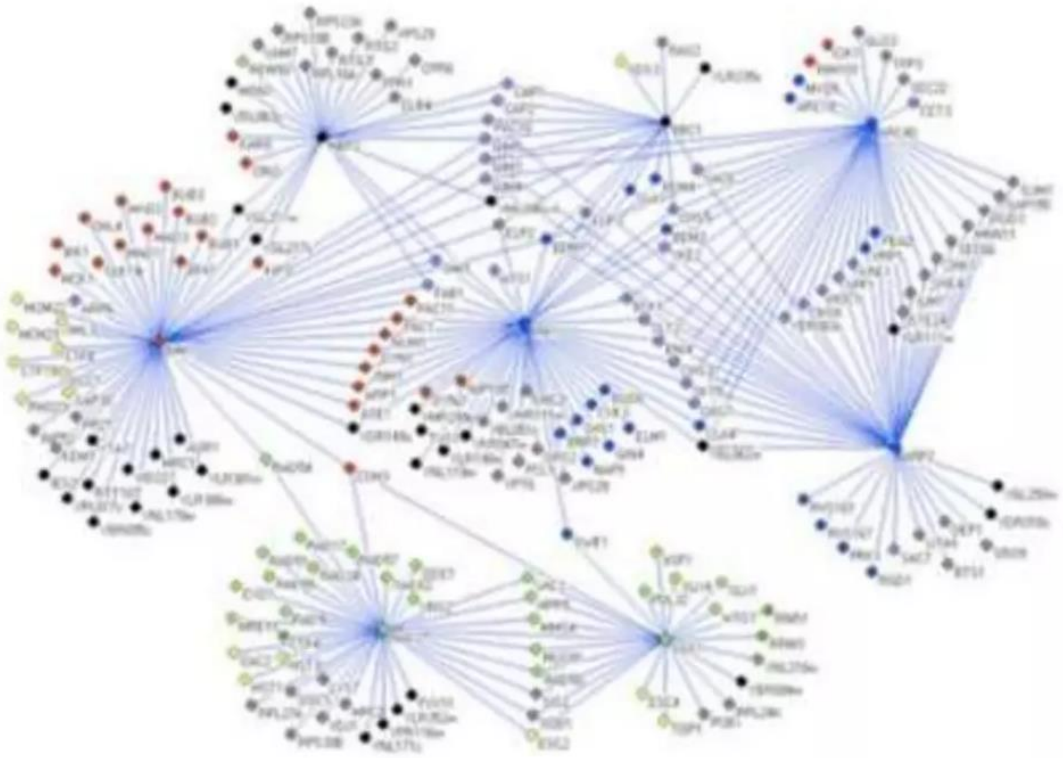
Old promise:
"Learn about the human brain."

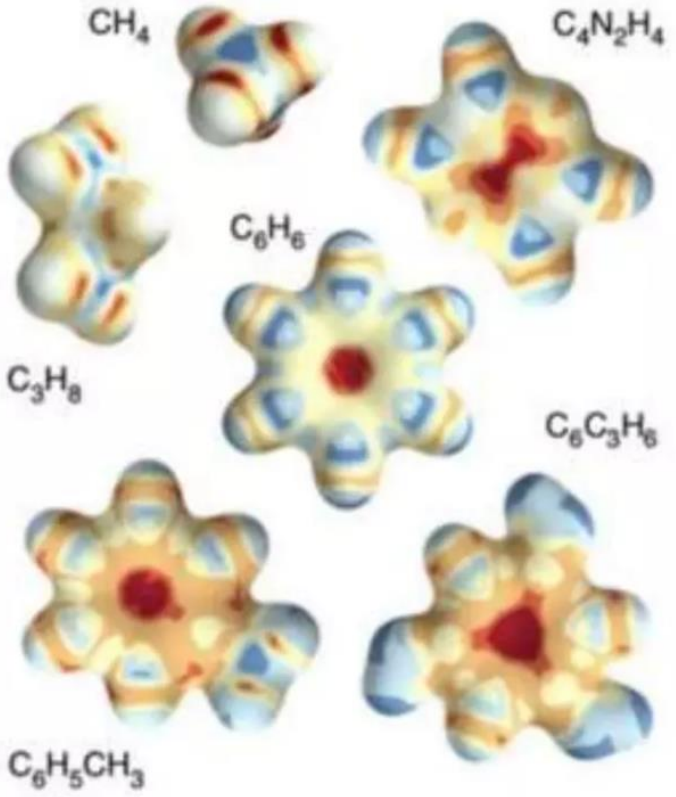


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Learn insights in the Sciences

Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites …)





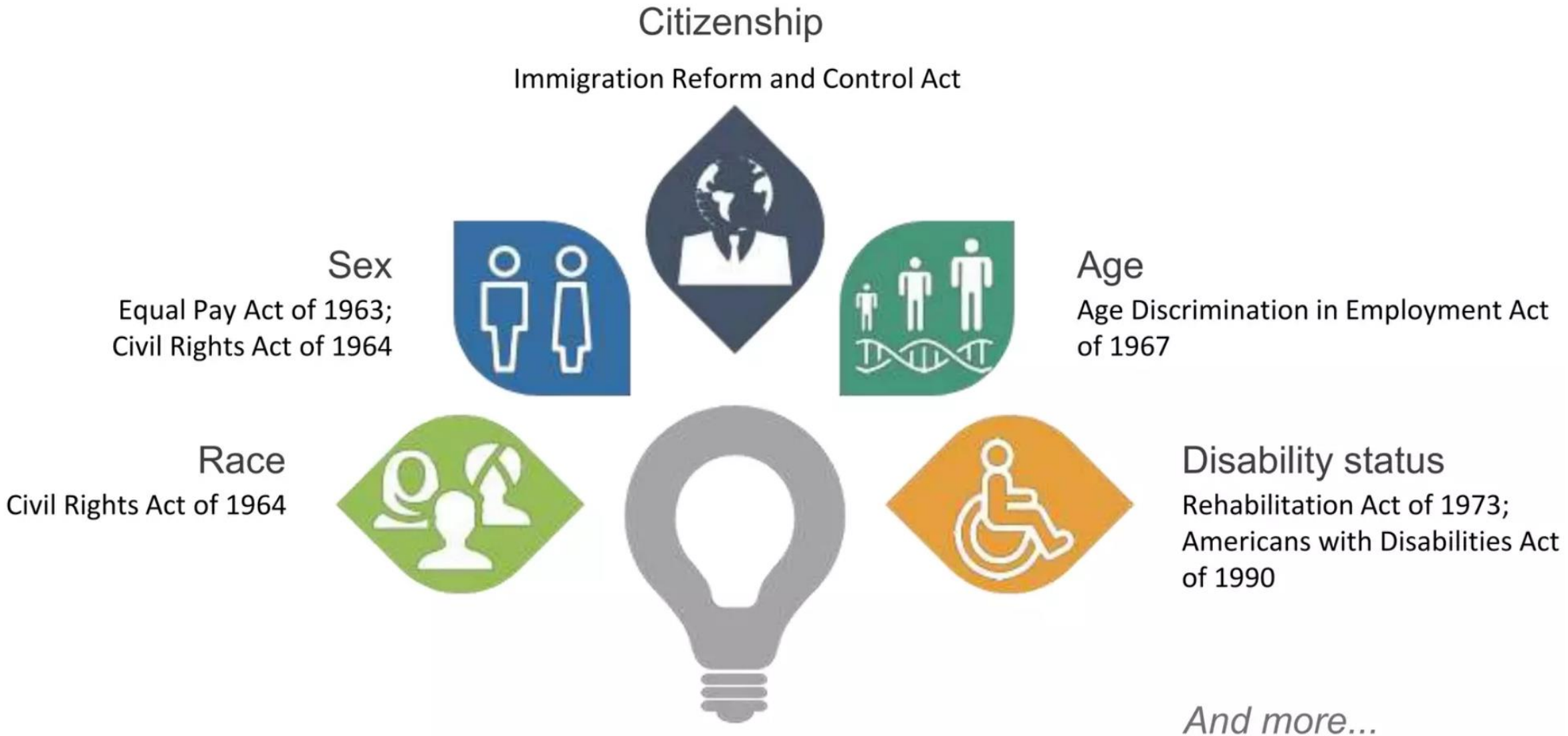
Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Debug Mis-predictions

Top label: **“clog”**

Why did the network label this image as **“clog”**?

Why Explainability: Laws against Discrimination



Growing Global AI Regulation

- **GDPR**: Article 22 empowers individuals with the **right to demand an explanation of how an automated system made a decision** that affects them.
- **Algorithmic Accountability Act 2019**: Requires companies to **provide an assessment of the risks** posed by the automated decision system to the **privacy** or **security** and the risks that contribute to **inaccurate, unfair, biased, or discriminatory decisions** impacting consumers
- **California Consumer Privacy Act:** Requires companies to **rethink their approach to capturing, storing, and sharing personal data** to align with the new requirements by January 1, 2020.
- **Washington Bill 1655**: Establishes guidelines for the use of automated decision systems to protect consumers, improve transparency, and create more market predictability.
- **Massachusetts Bill H.2701**: Establishes a commission on **automated decision-making, transparency, fairness, and individual rights**.
- **Illinois House Bill 3415:** States predictive data analytics determining creditworthiness or hiring decisions **may not include information that correlates** with the applicant race or zip code.