

12: Interpretable and Explainable AI - Part 2

Juho Kim & Jean Young Song

Human-AI Interaction KAIST Fall 2020 | kixlab.org/courses/human-ai

Administrative Notes


- **Project Pitch Feedback Meetings**
 - During class time on 10/15, 15 mins per team
 - Schedule announced on website/Campuswire
 - Bring your initial ideas for short intro & discussion.

- Assignment #2 will be announced on **10/15 (due 11/5)**.

Tentative Project Ideas

- tentative (recommendation system, intelligent agent,,?)
- Machine learning bias interactive visualization.
- User-modulation of AI behavior via visual attention interface
- AI tutor that help children answer academic questions (just assist, not giving the solution)
- Window's lockscreen suggestion application based on user's facial expression
- AI that recommends University students' Majors
- explainability, interpretability (of AI decision like AI judge or AI employee arrangement)
- "AI-assisted for KAIST admission(freshmen). Because for now, we have to directly call or e-mail the staff. We are thinking of collecting previous years' official data. Our track is gonna be chatbot, and our intention is to help applicants to get info. We are open to any suggestions from you...!!!"
- Tip of the tongue helper using dictionary API and NLP

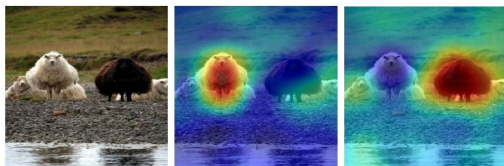
Previously on CS492F...

 **Geoffrey Hinton**
@geoffreyhinton

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

5:37 AM · Feb 21, 2020 · Twitter Web App

1.1K Retweets 613 Quote Tweets 5.2K Likes

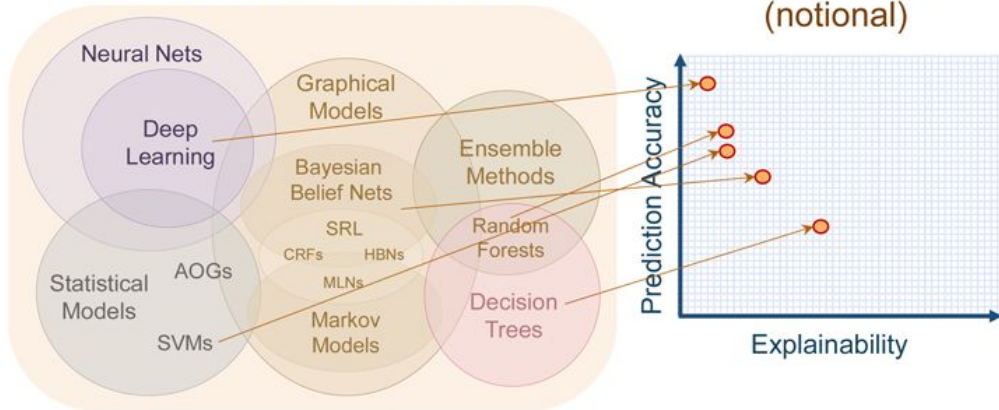


(a) Sheep - 26%, Cow - 17% (b) Importance map of 'sheep' (c) Importance map of 'cow'



(d) Bird - 100%, Person - 39% (e) Importance map of 'bird' (f) Importance map of 'person'

Learning Techniques (today)



Explainability (notional)

Why these ads? ×

Ads from Amazon.com were shown to you based on:

- Your current search terms

Ads from Hotel Restaurant Supply were shown to you based on:

- Your current search terms

Ads from KaTom Restaurant Supply were shown to you based on:

- Your current search terms

Ads from WebstaurantStore.com were shown to you based on:

- Your current search terms

[LEARN MORE](#) [ADS SETTINGS](#)

Today's Learning Objectives

After today's class, you should be able to...

- Understand the multi-dimensional nature of interpretability.
- Practice thinking about what degree and kind of interpretability would be appropriate for a given context.
- Consider interpretability in all phases of design, beyond the AI model.
- Apply knowledge about interpretability in creating a model card.

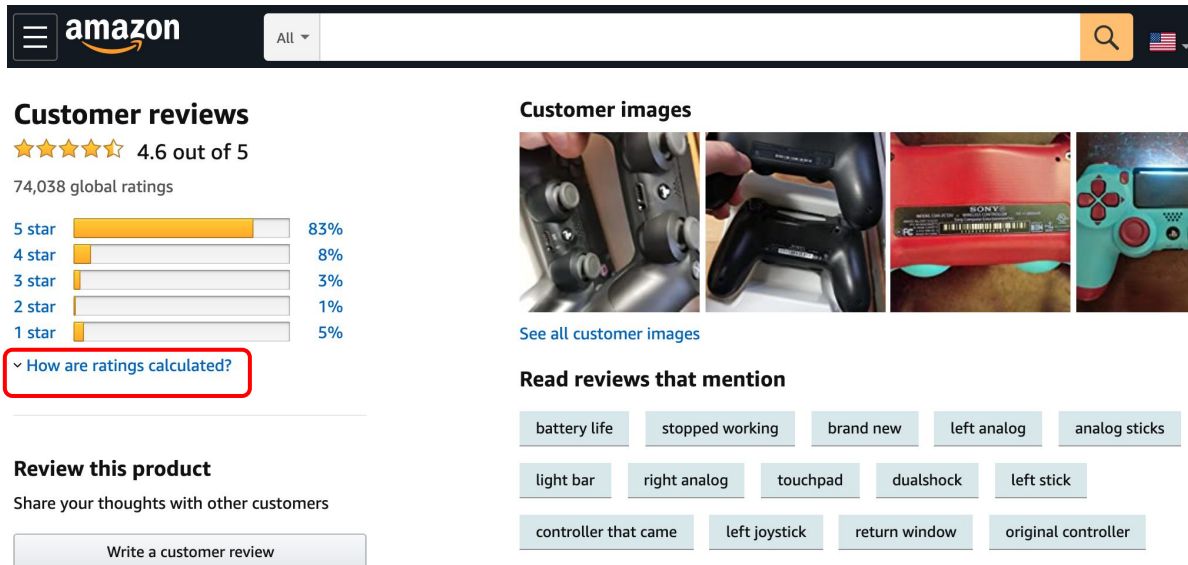
Reflection on the last in-class activity

Instruction



- Choose **two examples** from the given three examples of explainable AI, and do the following:
 - See if the explanation is enough or not for you.
 - If not enough, **explain why it is not enough and what information is missing.**
 - Try to **provide a more satisfying explanation** and **explain why it is more satisfying than the original one.**
 - **Review & Improve:** Go to **team {your team number+1}'s slide** and read **one of** the other team's explanations.
 - Discuss if the new explanation is good or bad. If good, comment why it is good. If bad, **try to improve the other team's explanation.**

Example 1

- Go to Amazon.com, and click any product.
- Go to Customer reviews.
- Read “How are ratings calculated?”



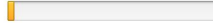
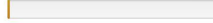
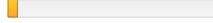


The screenshot shows the Amazon product page for a DualShock 4 controller. The top navigation bar includes the Amazon logo, a search bar, and a location selector for the United States. The main content area is divided into two columns. The left column features the 'Customer reviews' section, which displays a 4.6 out of 5 star rating based on 74,038 global ratings. A bar chart shows the distribution of ratings: 5 stars (83%), 4 stars (8%), 3 stars (3%), 2 stars (1%), and 1 star (5%). A red-bordered button labeled 'How are ratings calculated?' is highlighted. Below the reviews is a 'Review this product' section with a 'Write a customer review' button. The right column features the 'Customer images' section, which shows four images of the controller from different angles. Below the images is a link to 'See all customer images'. The 'Read reviews that mention' section lists various keywords such as 'battery life', 'stopped working', 'brand new', 'left analog', 'analog sticks', 'light bar', 'right analog', 'touchpad', 'dualshock', 'left stick', 'controller that came', 'left joystick', 'return window', and 'original controller'.

amazon All ▾  

Customer reviews

★★★★☆ 4.6 out of 5
74,038 global ratings

5 star		83%
4 star		8%
3 star		3%
2 star		1%
1 star		5%


[How are ratings calculated?](#)

Review this product

Share your thoughts with other customers

[Write a customer review](#)

Customer images



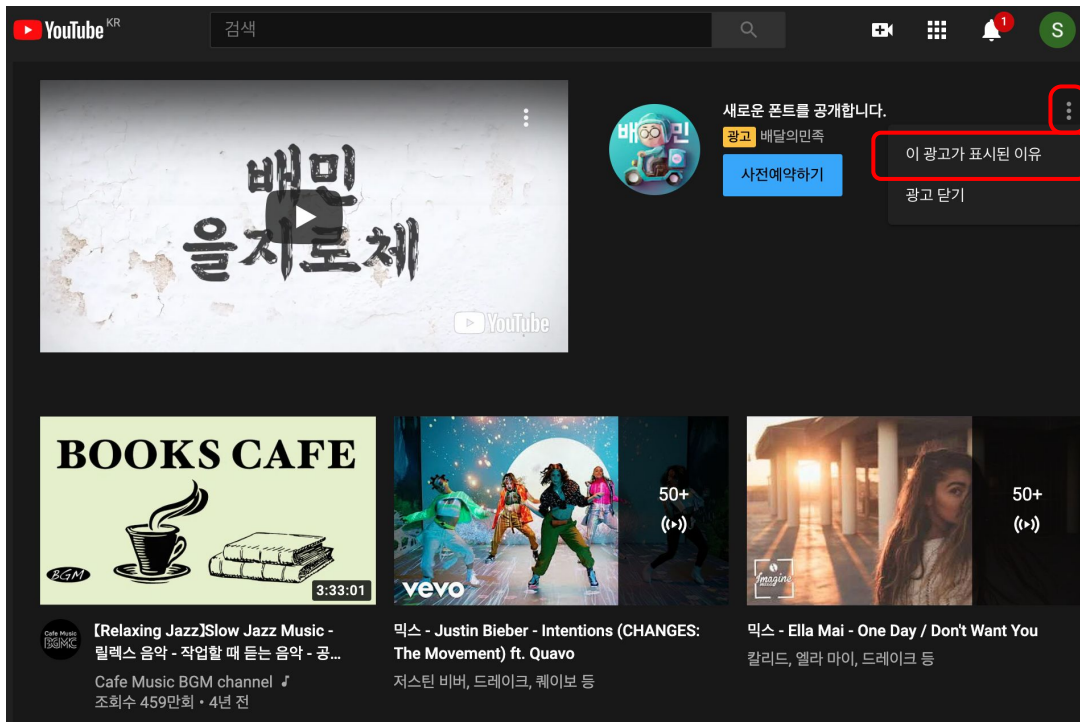
[See all customer images](#)

Read reviews that mention

battery life stopped working brand new left analog analog sticks
light bar right analog touchpad dualshock left stick
controller that came left joystick return window original controller

Example 2

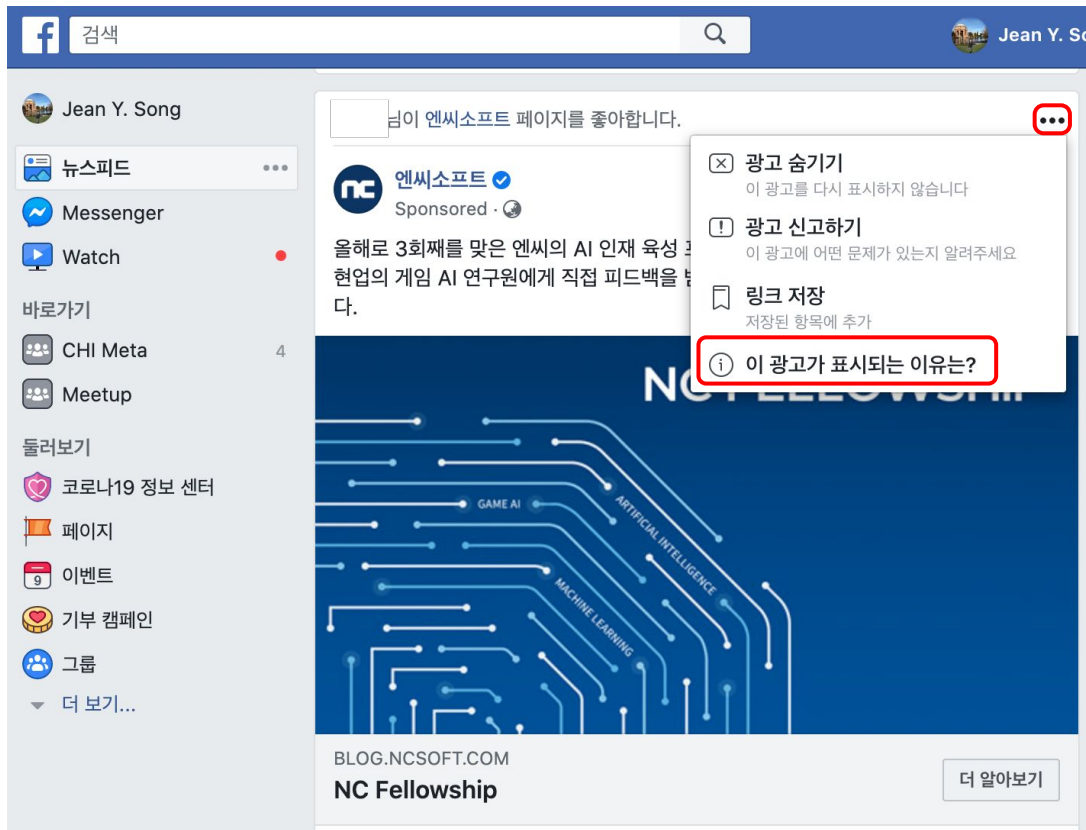
- Go to YouTube.
- If an Ad is shown, click the three dot button at the right and click “Why this ad?”
- Read why this ad is shown.



The screenshot shows the YouTube interface in Korean. The main video player displays a Korean advertisement with the text "배민을지로체" (Bemart Jeonju style) and a play button icon. To the right of the video, a menu is open, showing options like "새로운 폰트를 공개합니다." (We've released a new font.), "광고 배달의민족" (Ad Delivery of Bemart), "사전예약하기" (Pre-order), and "이 광고가 표시된 이유" (Why this ad is shown), which is highlighted with a red box. Below the video player, there are three recommended video thumbnails: "BOOKS CAFE" (Relaxing Jazz), "믹스 - Justin Bieber - Intentions (CHANGES: The Movement) ft. Quavo", and "믹스 - Ella Mai - One Day / Don't Want You".

Example 3

- Go to Facebook.
- If an Ad is shown, click the three dot button at the right and click “Why this ad?”
- Read why this ad is shown.



The screenshot shows a Facebook interface. At the top, there is a search bar with the text '검색' and a profile picture of 'Jean Y. Song'. Below the search bar, the profile name 'Jean Y. Song' is visible. On the left side, there is a navigation menu with options like '뉴스피드', 'Messenger', 'Watch', '바로가기', 'CHI Meta', 'Meetup', '둘러보기', '코로나19 정보 센터', '페이지', '이벤트', '기부 캠페인', '그룹', and '더 보기...'. The main content area shows a post from 'NC Fellowship' (verified account) with the text '올해로 3회째를 맞은 엔씨의 AI 인재 육성 프로그램. 현업의 게임 AI 연구원에게 직접 피드백을 받는다.' and a blue background with circuit-like patterns and the text 'GAME AI', 'ARTIFICIAL INTELLIGENCE', 'MACHINE LEARNING'. Below the post, there is a link 'BLOG.NCSOFT.COM' and the text 'NC Fellowship'. A '더 알아보기' button is in the bottom right. A red circle highlights a three-dot menu button in the top right corner of the post. A white menu is open over the post, containing options: '광고 숨기기' (Hide ad), '광고 신고하기' (Report ad), '링크 저장' (Save link), and '이 광고가 표시되는 이유는?' (Why is this ad shown?). The '이 광고가 표시되는 이유는?' option is highlighted with a red rectangle.

Result Highlights

- Amazon: “no” x 8
- YouTube: “no” x 7
- **Facebook: “no” x 2 + “yes”/“okay” x 3**
- Common observations
 - Simple listing of factors isn’t informative enough. (Still better than not listing any.)
 - Be specific and avoid unclear terminology such as trustworthy, recent, “your activity”.
 - Too much personal information is used to make a decision. (But being transparent about is better?)

Multiple Faces of Interpretability

“Right to Explanation”

- “A right to be given an explanation for an output of the algorithm” [Wikipedia]
- Credit score, Criminal justice, ...
- EU’s GDPR (enacted 2016, taking effect 2018)
 - *“In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, **to obtain an explanation of the decision reached after such assessment and to challenge the decision.**”*

Scenario: Age-Guessing AI

- Here's an AI that predicts how old you are based on your photo. It tells you you're... 53.
- What kind of information / explanation would you like to demand?

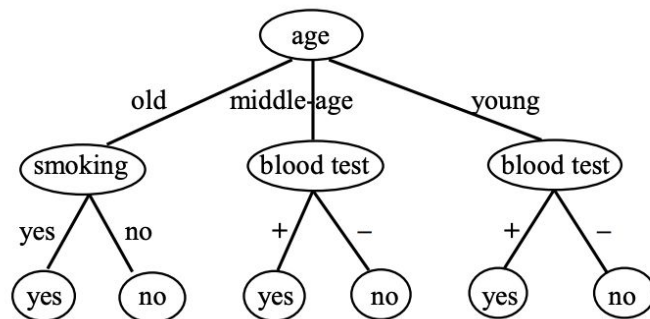
Scenario: Age-Guessing AI

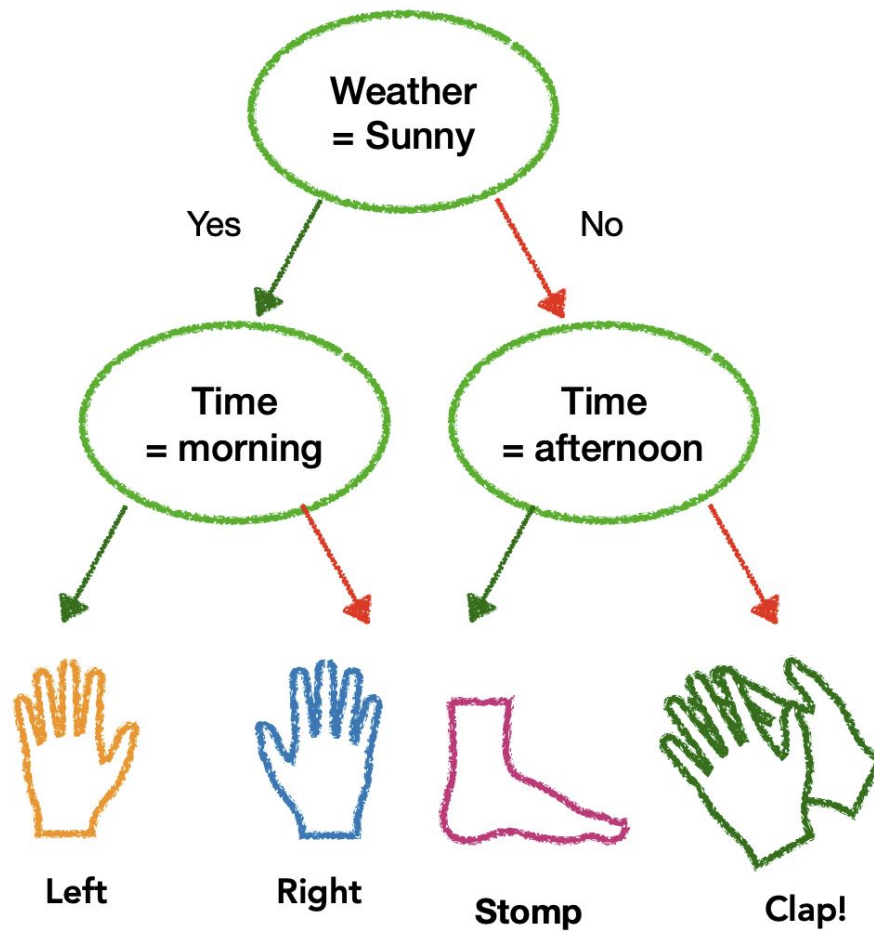
- Here's an AI that predicts how old you are based on your photo. It tells you you're... 53.
- What questions will it be able to answer?
- What questions will it not be able to answer?
- What considerations should be made w.r.t. interpretability if this AI was used by insurance companies? Hospitals? Job placement decisions?

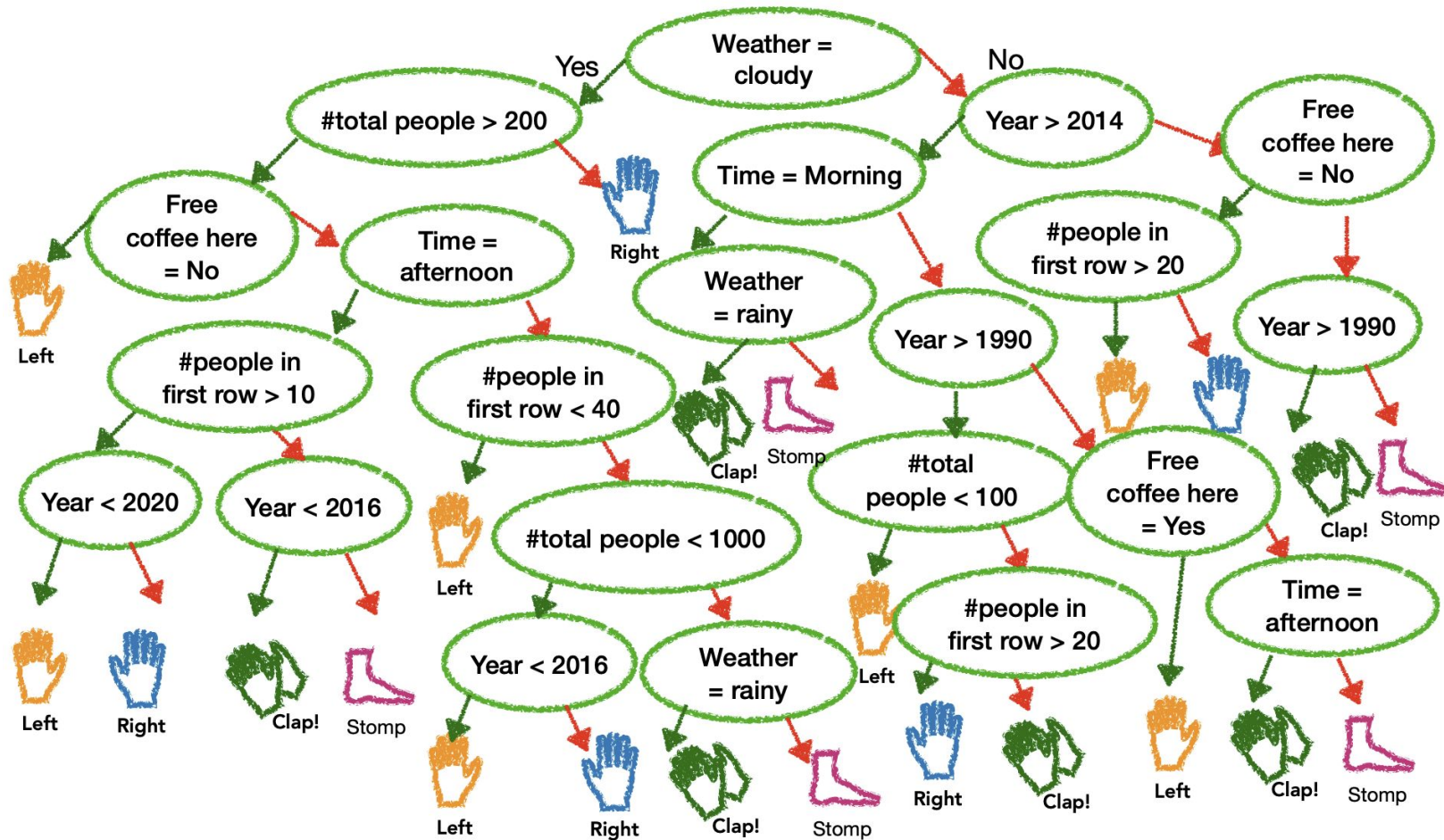
Scenario: Interpretability of Decision Trees

They are known to be interpretable. Really?

- Observations/Features as branches
- Class labels as leaves
- Graphical structure: easy to follow a path that leads to a decision
- (Relatively) nodes closer to the root indicate higher importance.
- Small no. of attributes represented



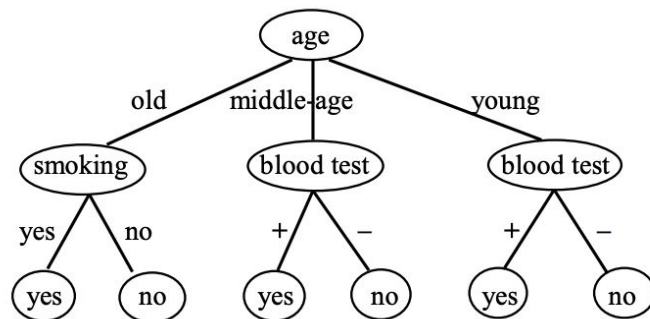




Scenario: Interpretability of Decision Trees

They are known to be interpretable. Really?

- Yes and No.
- What aspects of decision trees make it difficult to interpret?
 - Depth-relevance relationship doesn't always hold true.
 - Duplicate / irrelevant attributes may exist due to their algorithmic structure.



**Simpler models don't necessarily
guarantee interpretability.**

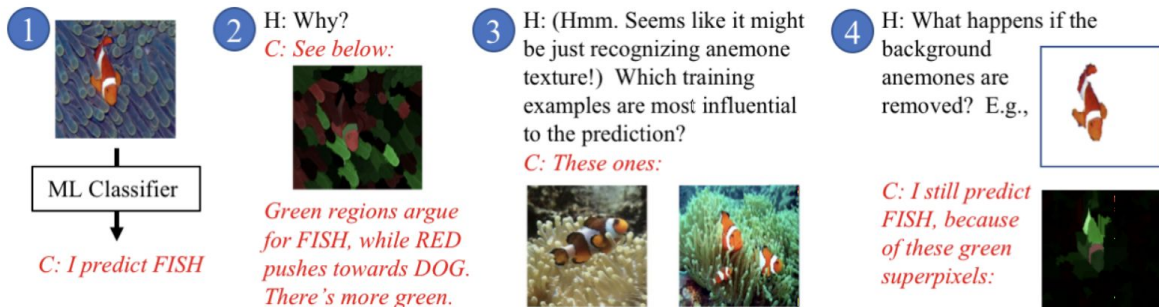
How to help humans interpret better is key.

Discussion: Is a single explanation enough?

- Can you think of cases where a single version of explanation is not enough?
 - Different user needs. What user dimensions should we consider to make a user-centered explanation?
 - Multiple stakeholders. How should multiple versions of explanation be shown to users?

Discussion: Is interactive explanation useful?

- Can you think of cases where a user would want to interactively request / explore explanations?
 - Follow-up questions, further investigation
 - Additional rationale, testing additional input
 - Provide feedback



Discussion: Is an explanation always desired?

- Can you think of cases where explanation is not necessary at all or even harmful?
 - No consequences (or domain too complex anyway).
 - Well-known problems.
 - Personal/Proprietary information.
 - Prevent gaming.

Designing for Interpretability

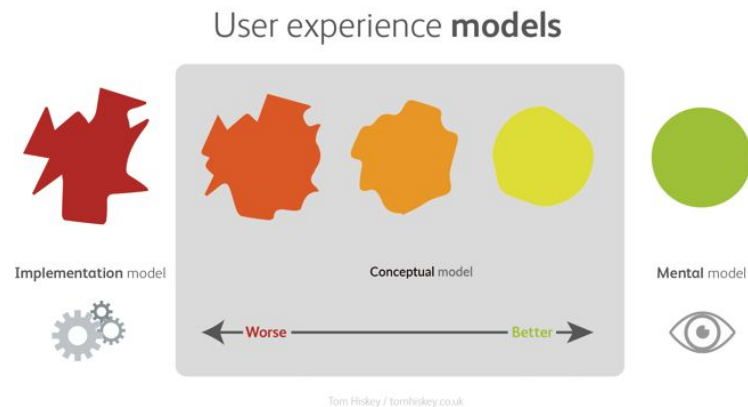
Interpretability from a design perspective

- A way to bridge the implementation model and the mental model
- Model being able to explain

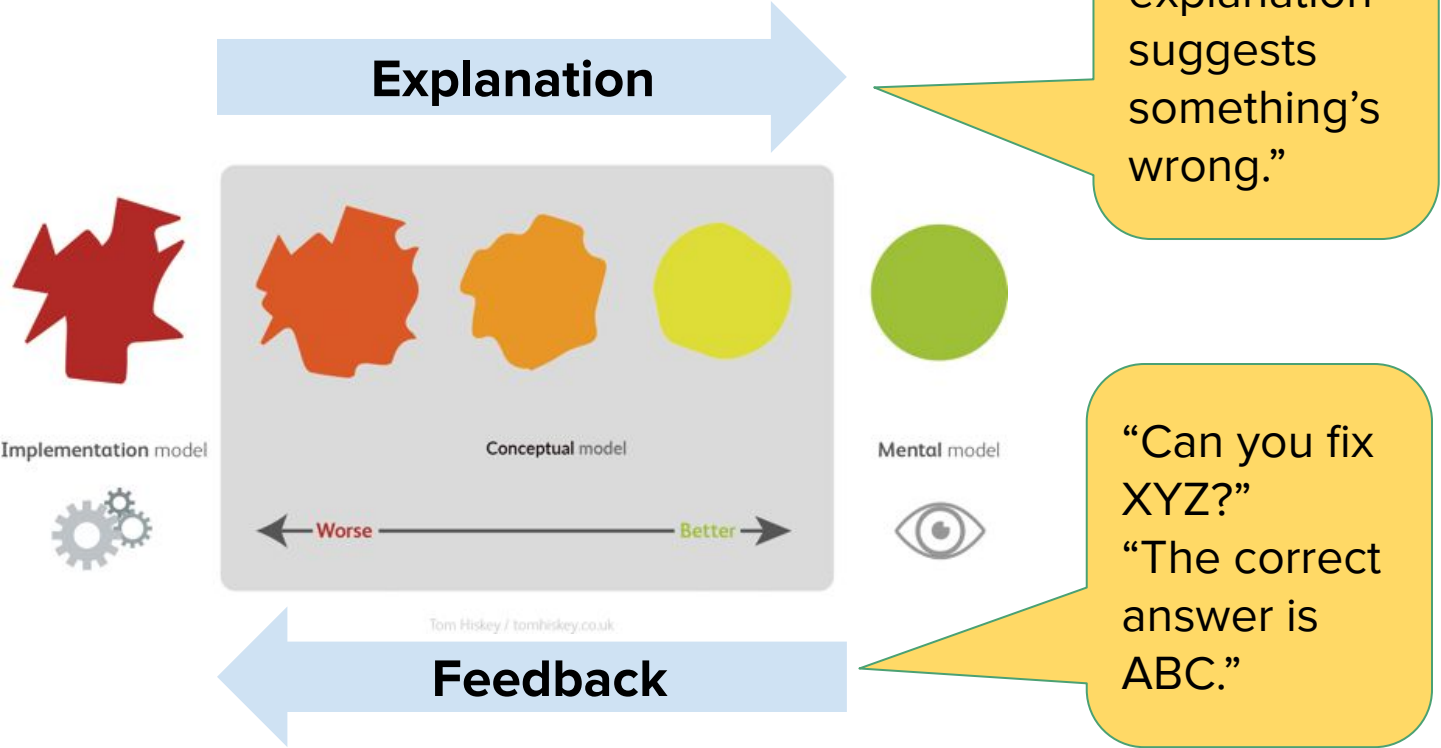
VS

User being able to understand

- These are two different kinds of Interpretability (and probably why AI and HCI communities don't collaborate well).



Explanation and Feedback



Design Considerations for Explanation & Feedback

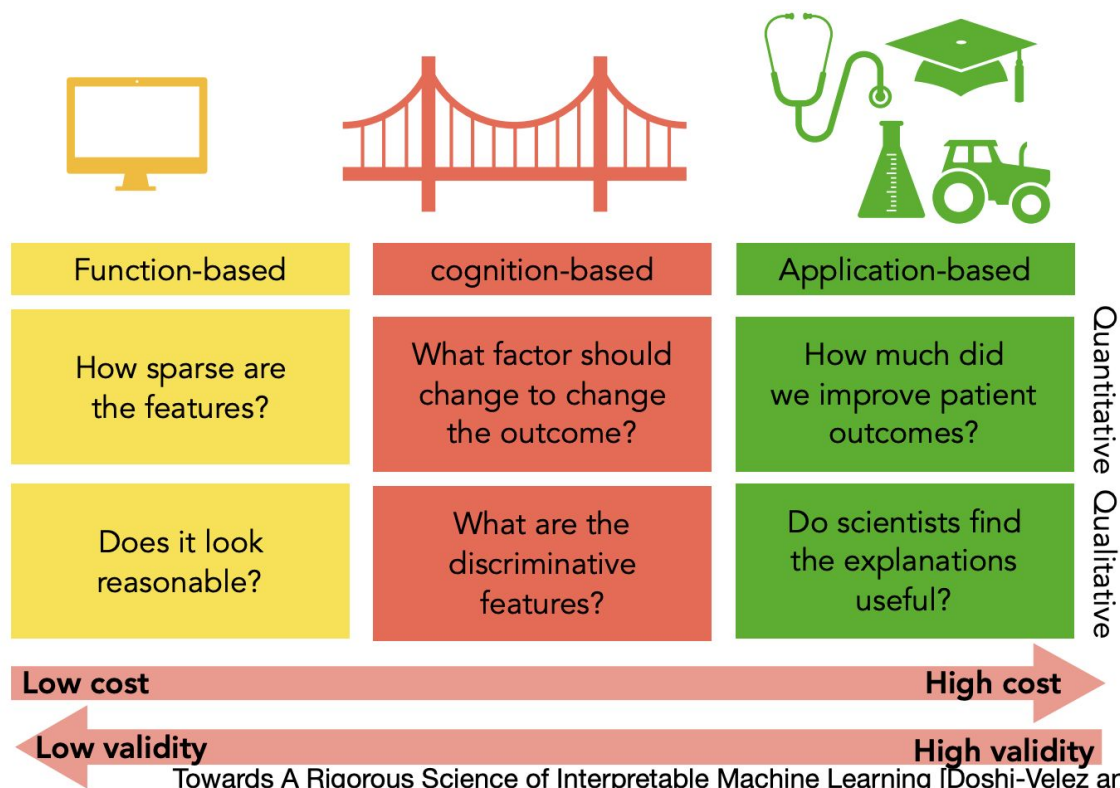
	Explanation (AI to user)	Feedback (user to AI)
UI	How is an explanation presented?	How can the user provide feedback?
Model	Can the model generate an explanation?	How does the model incorporate user feedback?
Data	What does the data look like?	How does data collection, processing, filtering... react to user feedback?

Can Explanation and Feedback Complement?

- Binary text classification, simple explanation
- Low-quality (~75% accuracy) vs High-quality (~90%) models
- Explanations (with/without) X Feedback (none, instance-, feature-level)
- Main findings
 - Users wanted the opportunity to provide feedback.
 - Low-quality model: feedback reduced frustration and increased trust & acceptance, but explanations had the opposite effect.
 - When users provided detailed feedback, they expected more improvement.

**Interpretability isn't just about a model.
Interpretability requires careful
considerations in all stages of AI design.**

Evaluating Interpretability



Towards A Rigorous Science of Interpretable Machine Learning [Doshi-Velez and K. 18]

Model Cards

- Short document about a trained ML model, with its intended use and performance characteristics.
 - Goal: To help users decide whether and how to apply the model to their context.
- A structured communication medium (like a spec sheet for hardware devices and electrical components) to be shared across different stakeholders.

ACTIVITY: Let's crowdsource model card generation.

- Let's fill in the missing information to complete a model card for an image cropping AI.
- Groups of 3-4, 20 mins

yellkey.com/pattern

Resources

- [Google's Responsible AI Practices](#)
- [CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision](#)
- [Google's Model Cards Documentation](#)
- Doshi-Velez, Finale, and Been Kim. ["Towards a rigorous science of interpretable machine learning."](#) arXiv preprint arXiv:1702.08608 (2017).