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

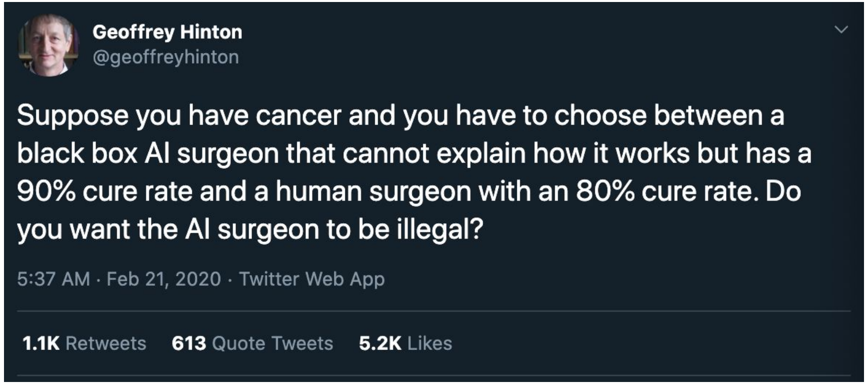
- **Project Pitch Feedback Meetings**
  - During class time on 10/15, 15 mins per team
  - Schedule announced on website/Campuswire
  - Bring your initial ideas for short intro & discussion.

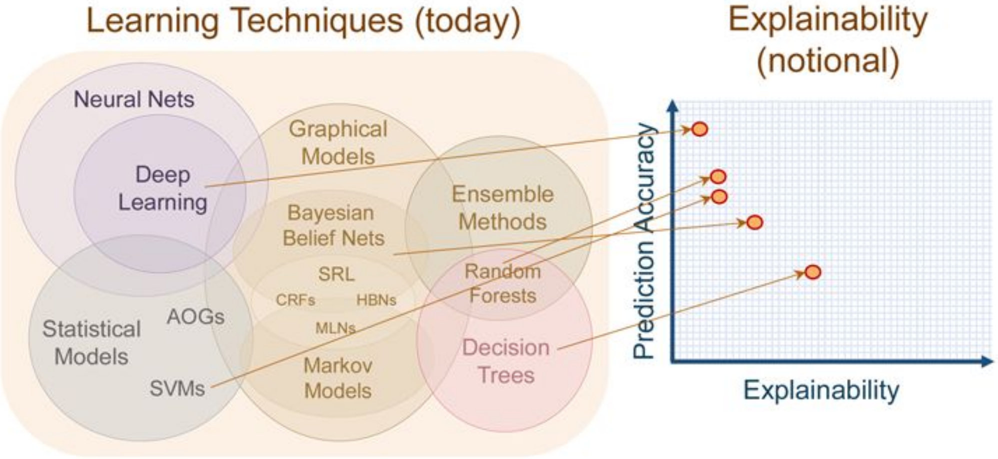
- Assignment #2 will be announced on **10/15 (due 11/5).**

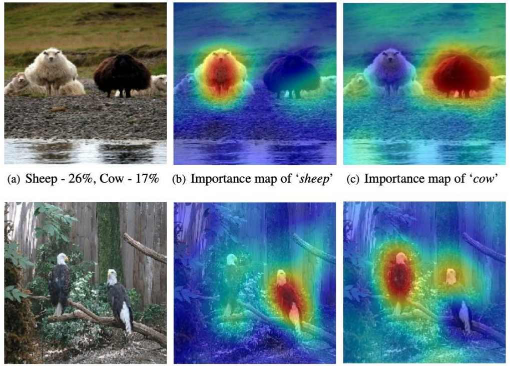
Tentative Project Ideas

- 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...

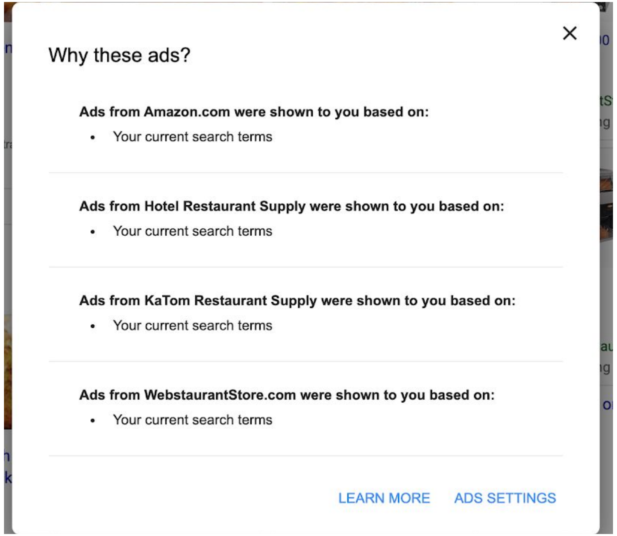






(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

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



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 interprebility in all phases of design, beyond the AI model.
- Apply knowledge about interpretability in creating a model card.

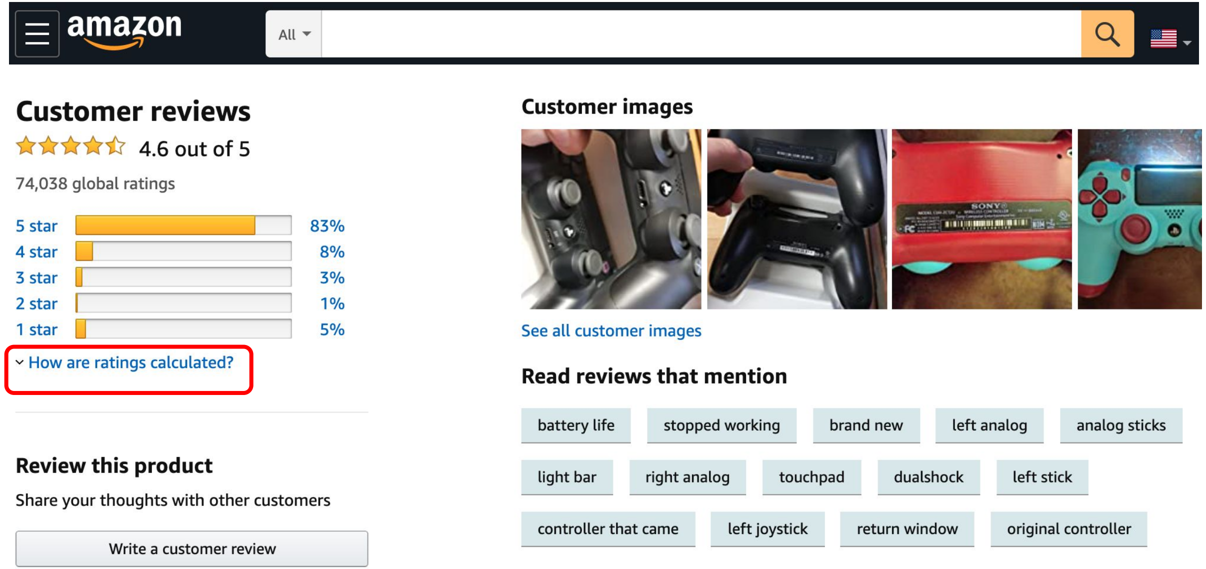
Reflection on the last in-class activity

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**.

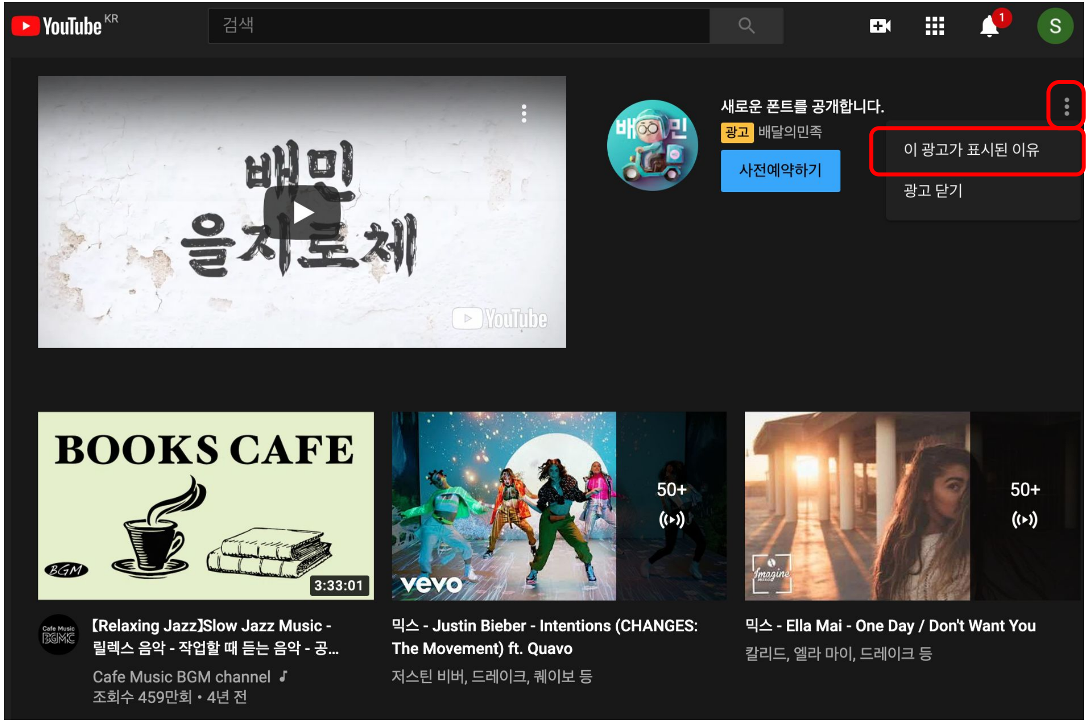
Example 1

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



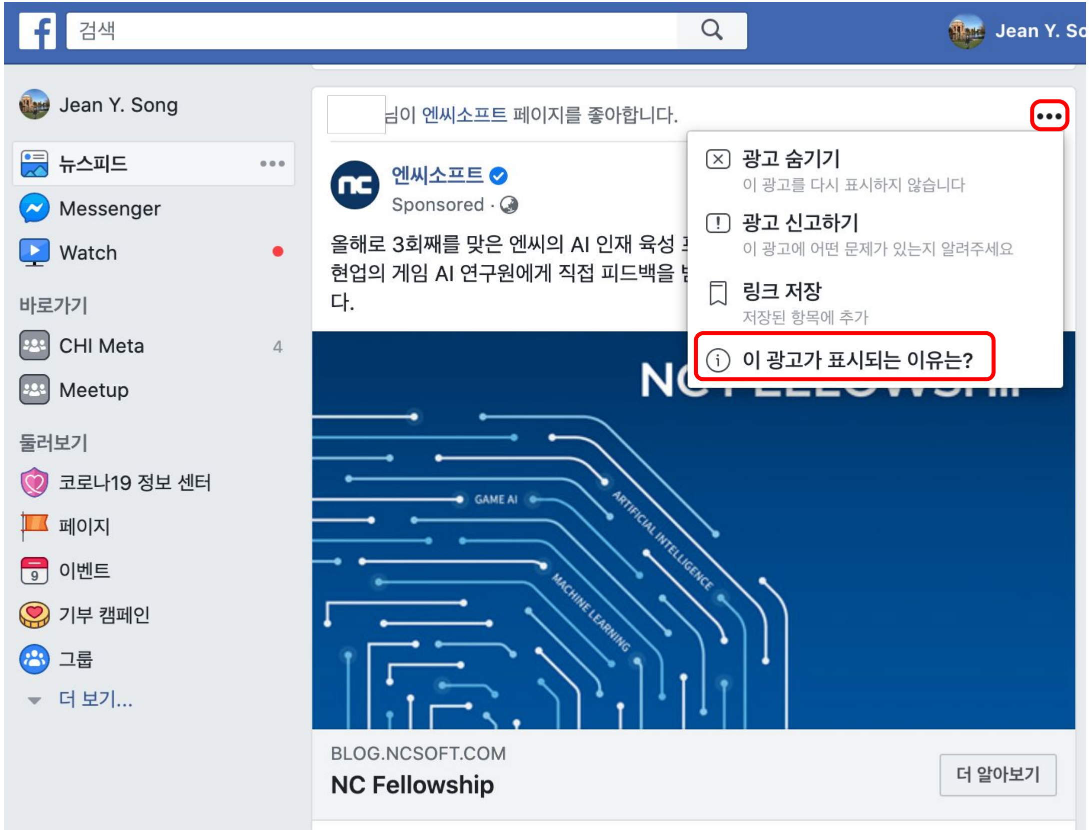
Example 2

- 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.



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.



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,* ***<u>to obtain an explanation of the decision reached after such assessment and to challenge the decision.</u>”***

https://en.wikipedia.org/wiki/Right_to_explanation

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?

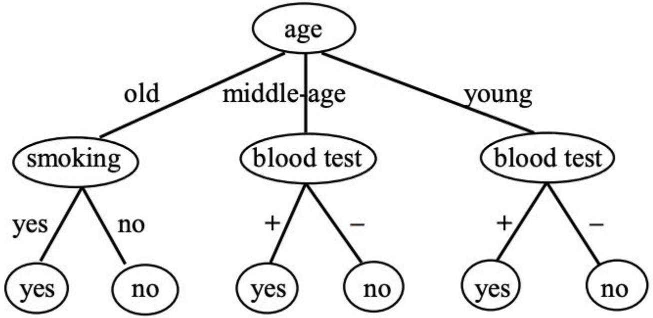
Scenario: Age-Guessing AI

- 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?

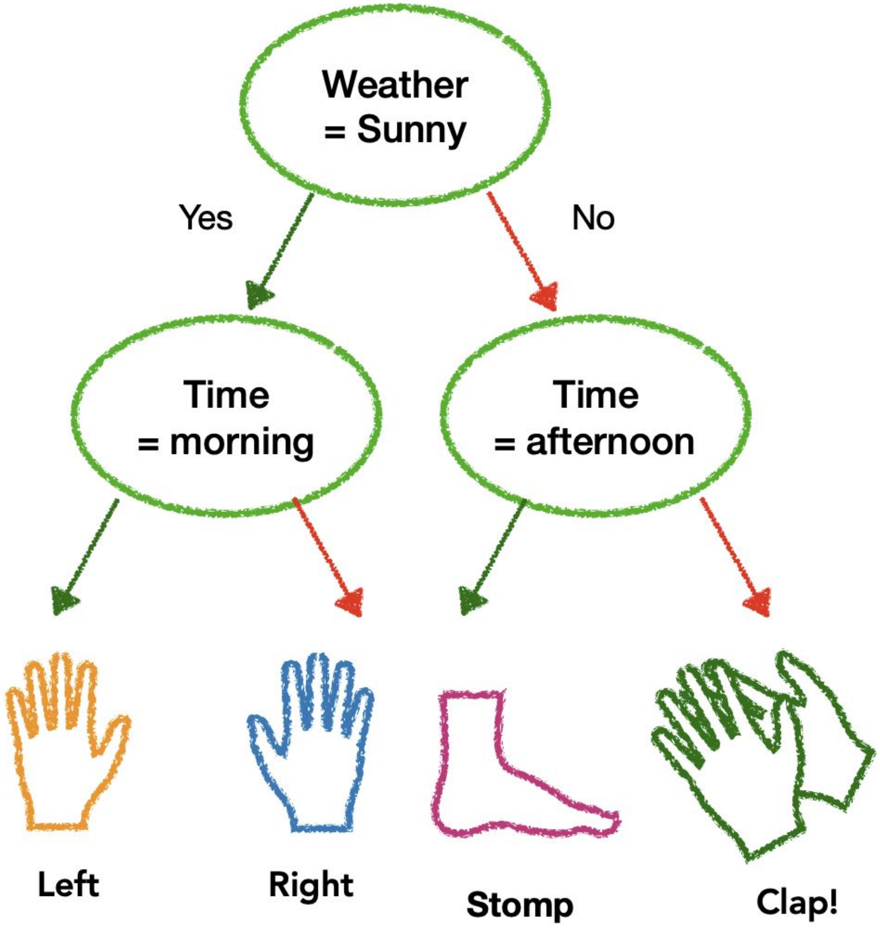
Scenario: Interpretability of Decision Trees

They are known to be interpretable. Really?

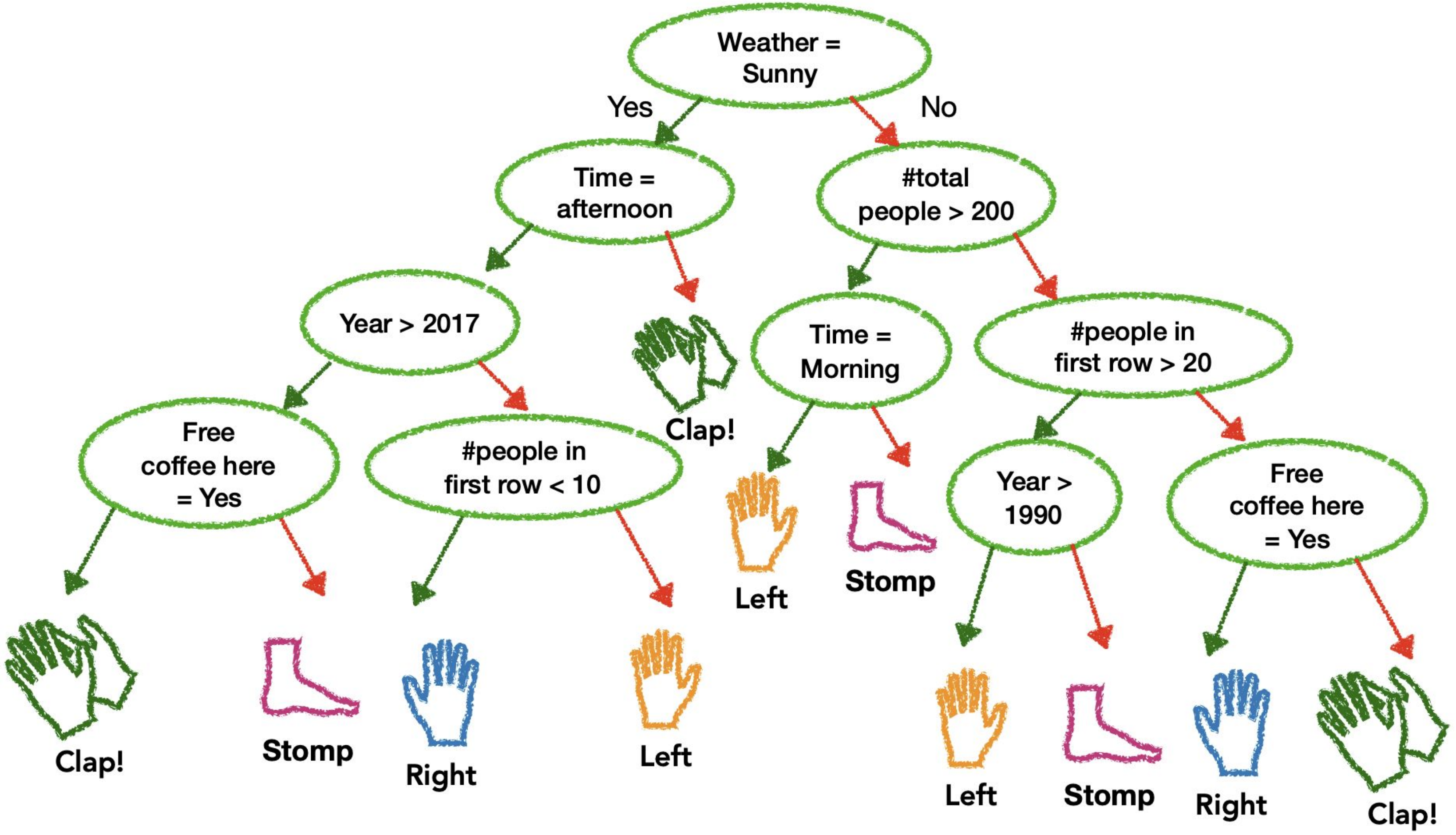
- Observations/Features as branches
- Class labels as leaves
- Graphical structure: easy to follow a path that leads to a decision
- (Relatively) nodes closer to the root Indicate higher importance.
- Small no. of attributes represented



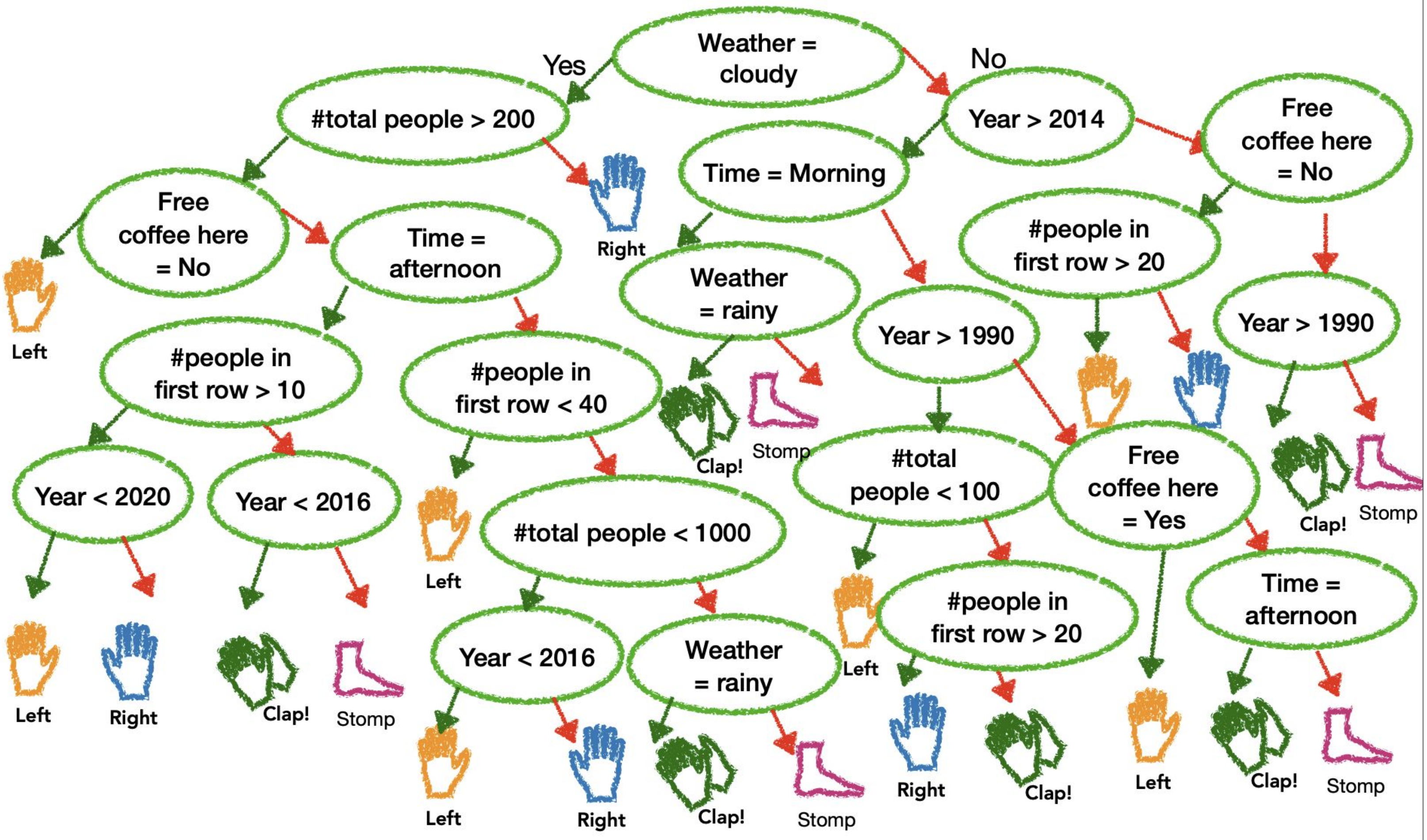
Freitas, Alex A. "Comprehensible classification models: a position paper." ACM SIGKDD explorations newsletter 15.1 (2014): 1-10.



Example from Been Kim's "Introduction to Interpretable Machine Learning" tutorial



Example from Been Kim's "Introduction to Interpretable Machine Learning" tutorial

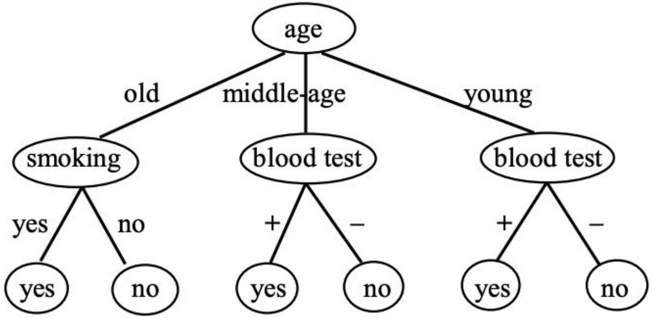


Example from Been Kim's “Introduction to Interpretable Machine Learning” tutorial

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.



Freitas, Alex A. "Comprehensible classification models: a position paper." ACM SIGKDD explorations newsletter 15.1 (2014): 1-10.

Simpler models don’t necessarily guarantee interpretability.

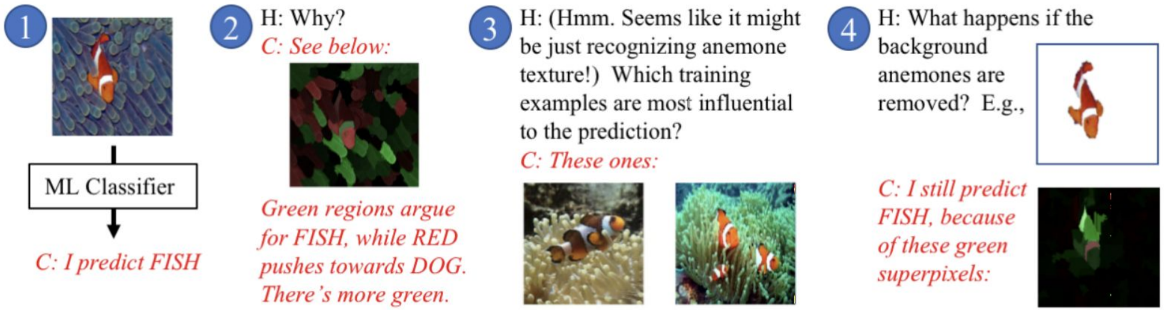
How to help humans interpret better is key.

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



Weld, Daniel S., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of the ACM 62.6 (2019): 70-79.

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.

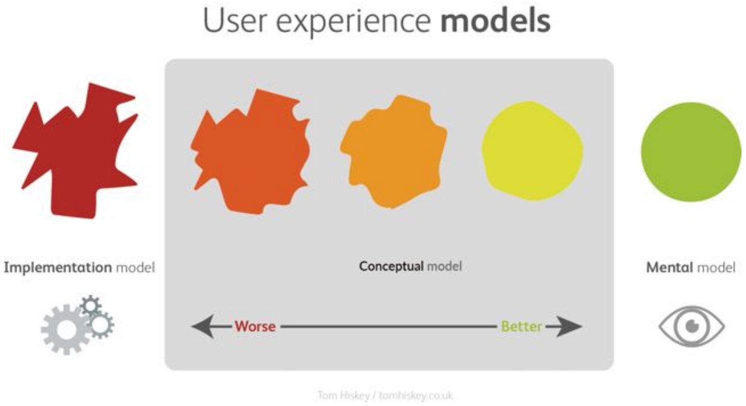
Designing for Interpretability

Interpretability from a design perspective

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

  vs

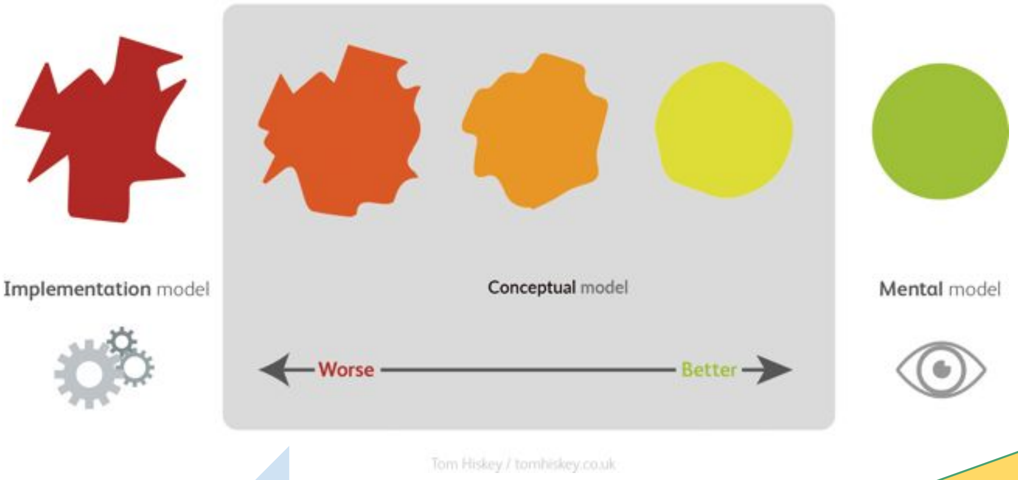
  User being able to understand
- These are two different kinds of Interpretability (and probably why AI and HCI communities don't collaborate well).



Content from https://www.fatml.org/resources/principles-for-accountable-algorithms

Explanation and Feedback

Explanation



Feedback

"Okay, the explanation suggests something's wrong."

"Can you fix XYZ?"
"The correct answer is ABC."

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?

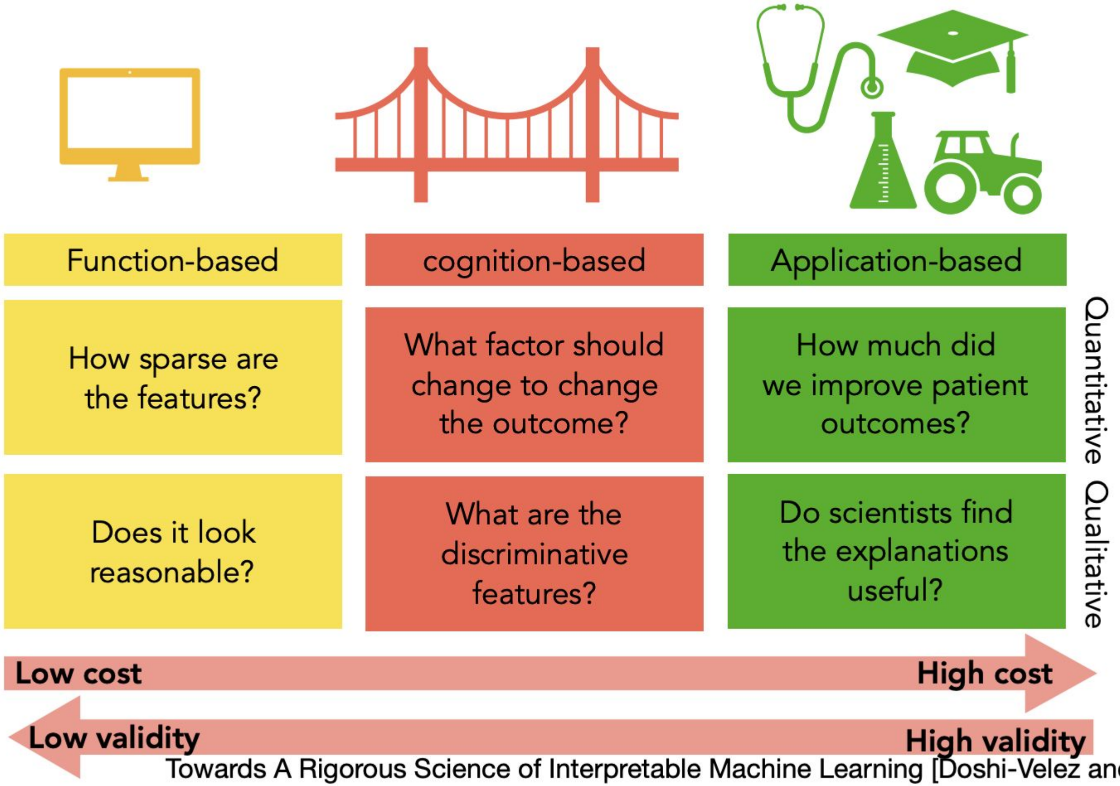
Can Explanation and Feedback Complement?

- 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.

Smith-Renner, Alison, et al. "No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML." Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.

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]

Example from Been Kim's ["Introduction to Interpretable Machine Learning" tutorial]

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.

Mitchell, Margaret, et al. "Model cards for model reporting." Proceedings of the conference on fairness, accountability, and transparency. 2019.
https://modelcards.withgoogle.com/about

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).