



University
of Glasgow
Oilthigh Ghlaschu

Transparency and Explainable AI (XAI)

Prof Simone Stumpf

Simone.Stumpf@glasgow.ac.uk

WORLD
CHANGING
GLASGOW

A WORLD
TOP 100
UNIVERSITY



University
of Glasgow
Oilthigh Ghlaschu

History



Not a new thing...

- Lots of work on explanations back in the 1980s and 1990s
 - neoMYCIN, etc to explain expert systems
- Ben Shneiderman versus Pattie Maes debate in the 1990s
 - Transparency and trust important
- Also work on autonomous agents and decision-making in the early 2000s
- But really ramped up since 2010s with rise in machine learning and black-box models

Mycin: Is Meningitis a hospital acquired illness?

Doctor: why?

Mycin: [2.0] this will aid in determining the organisms that might be causing the infection

It has already established

[2.1] the infection that requires therapy is meningitis

[2.2] organisms were not seen on the stain of this pending csf culture

[2.3] the type of infection is bacterial

[2.4] neurosurgery is not one of the relevant items of patient 100

[2.5] the age of patient 100 is greater than 25

therefore if [2.6] the infection was acquired while the patient was in hospital then there is evidence that the organisms that might be causing the infection are E coli(0.75) staphylococcus-coag-pos(0.3) pseudomonas-aeruginosa(0.3) klebsiella-pneumoniae(0.5)

Doctor: How 2.3

Mycin: The following rules concluded about the likelihood that the type of the infection is bacterial:

[2.7] RULE148 (0.19)

[2.8] RULE500 (0.51)

[2.9] RULE501 (0.95)

[3.0] RULE502 (0.97)

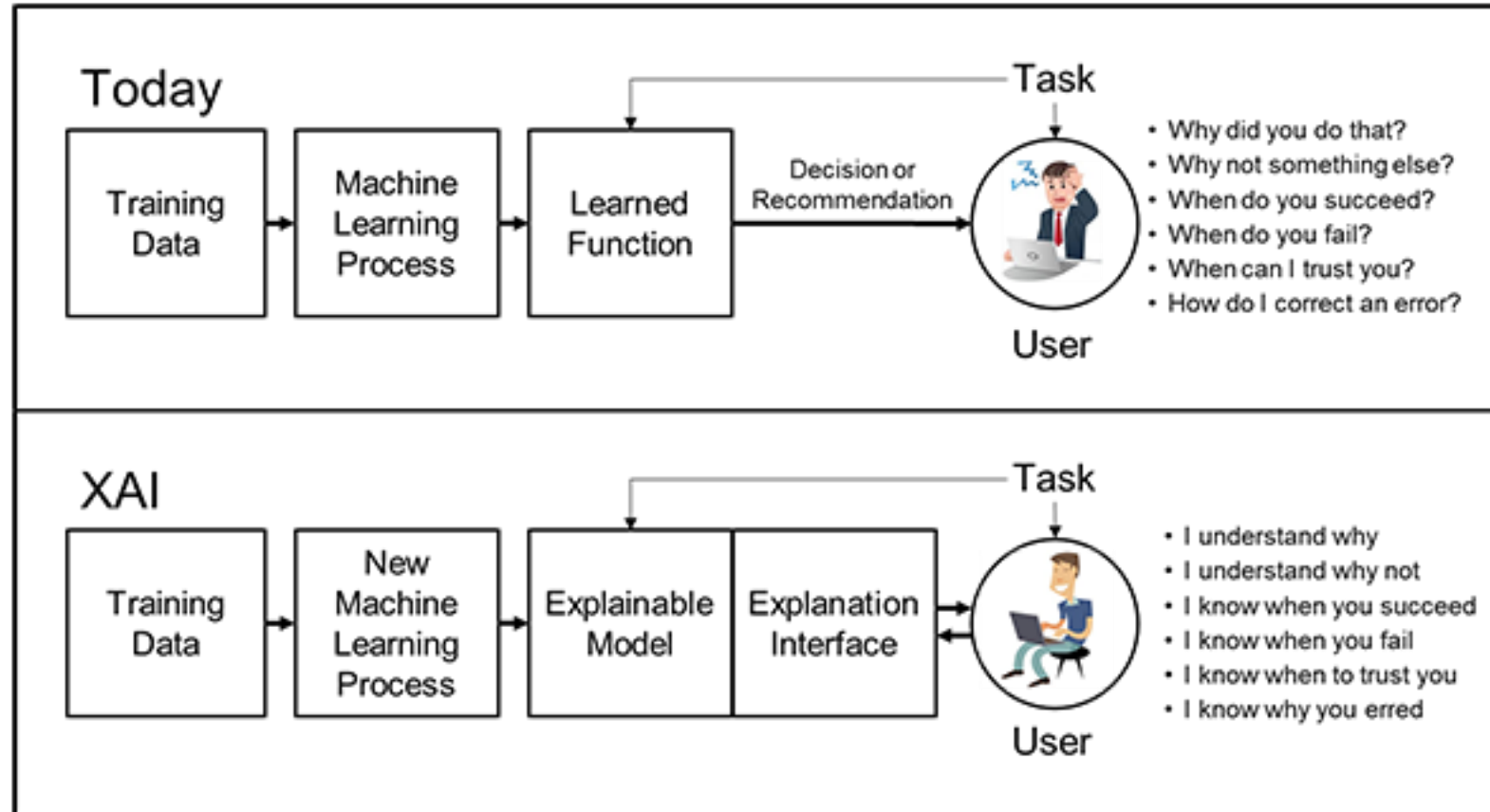
[3.1] RULE526 (0.98)

[3.2] RULE504 (0.97)

[3.3] RULE524 (0.98)



Explainable AI (XAI) vision (2016)



Calibrated /
appropriate
trust



Regulation, standards and guidelines

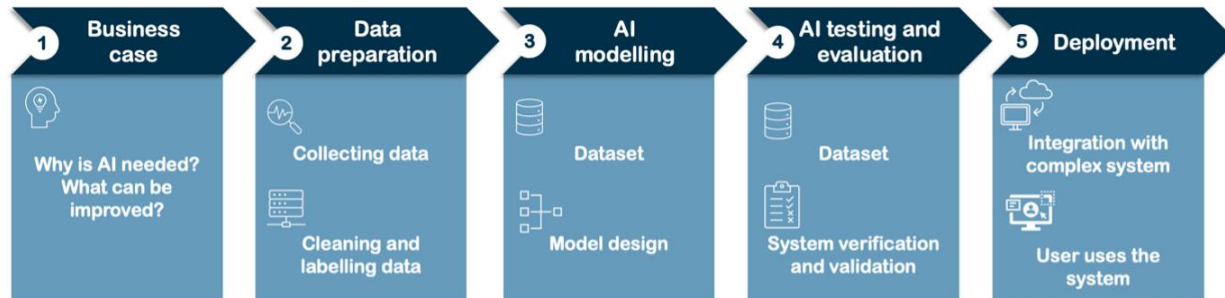
- Industry-led by tech giants
 - Microsoft's Guidelines for Human-AI Interactions
 - Google's Responsible AI practices and People+AI Handbook
 - IBM's Everyday Ethics for Artificial Intelligence
 - Fujitsu's AI Ethics Impact Assessment Practice Guide
- EU Assessment List for Trustworthy Artificial Intelligence (ALTAI)
- EU AI Act (ratified earlier this year)
 - High risk applications need to be assessed and **transparent**
 - Generative AI will have to be **transparent**



So what is AI “transparency”?

- How the AI model works
- Why a specific prediction was made by the AI ...or not

- Currently somewhat overlooked:
 - Why was the model developed in the first place
 - What training data was used to develop the model
 - How was the model evaluated
 - How good is it
 - What biases or blind spots does it have
 - What decisions about the AI were made during its development





University
of Glasgow
Oilthigh Ghlaschu

Explainable AI (XAI)



University
of Glasgow
Oilthigh Ghlaschu

Motivation for XAI

Model understanding is absolutely critical in several domains, particularly those involving *high potential for harm*, to support **debugging**, **bias detection** and **recourse**





University
of Glasgow
Oilthigh Ghlaschu

Lots of work to make AI ‘explainable’

[Molnar 2022]

- Global explanations:
 - Exposing the model
- Local explanations:
 - Exposing (combination of) features that contribute to a decision



University
of Glasgow
Oilthigh Ghlaschu

Local explanations



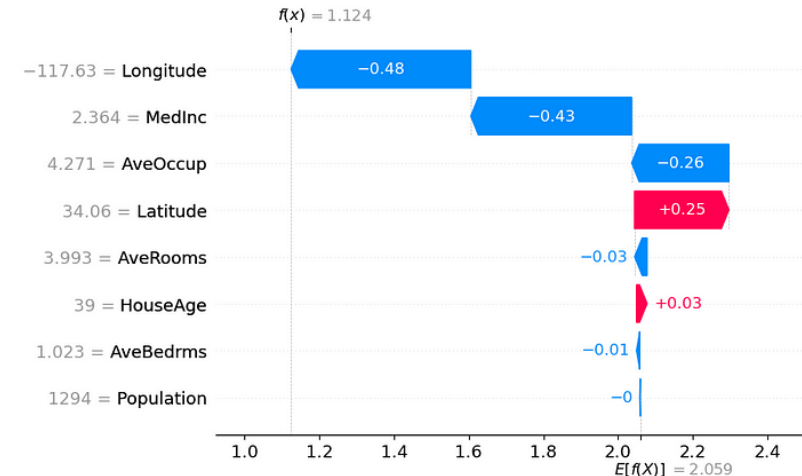
University
of Glasgow
Oilthigh Ghlaschu

LIME: Local Interpretable Model-Agnostic Explanations

- Explains important feature that led to a decision
- Uses a post-hoc explanation on a simplified model
- Another popular method which outputs feature importances: SHAP



[Ribeiro et al. KDD 2016]





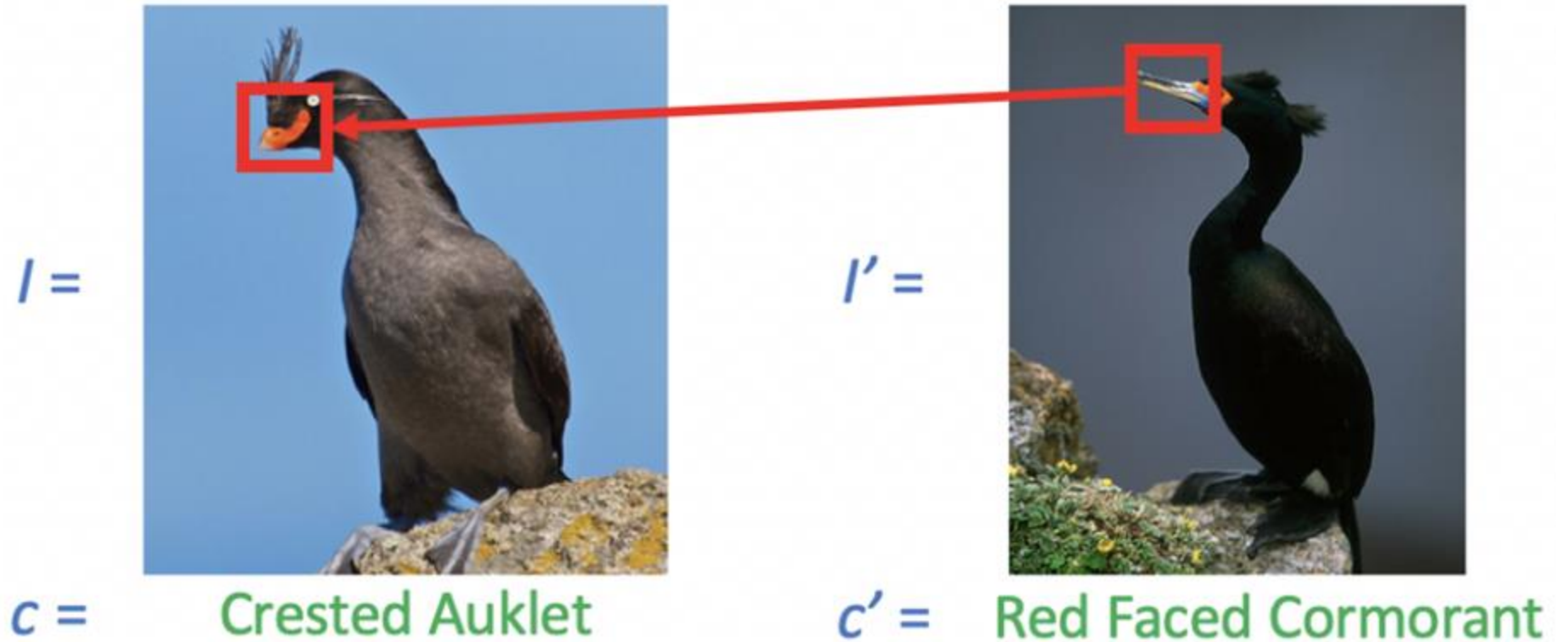
Prototypes/Example

- Use examples (synthetic or natural) to explain individual predictions
 - Identify instances in the training set that are responsible for the prediction of a given test instance
 - Identify examples (synthetic or natural) that strongly activate a function (neuron) of interest



Counterfactual Explanations

What features need to be changed and by how much to flip a model's prediction?

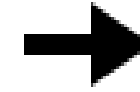
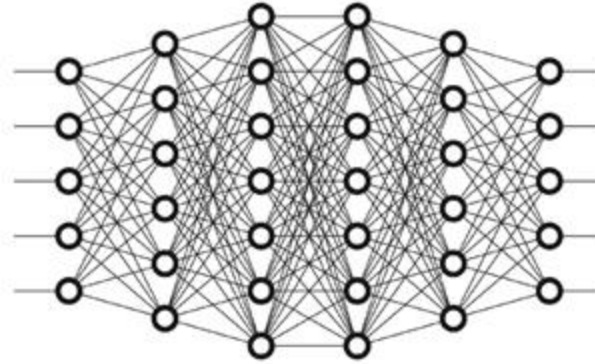




University
of Glasgow
Oilthigh Ghlaschu

Saliency Maps

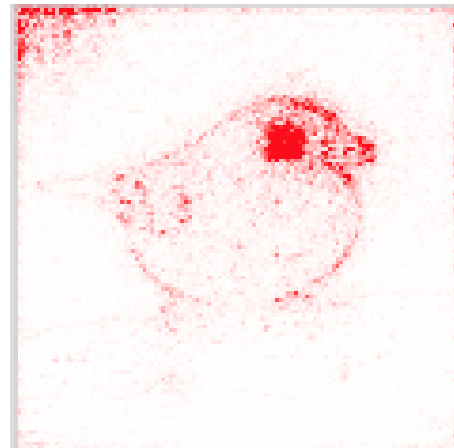
Input



Prediction

Junco Bird

What parts of the input are most relevant for the model's prediction: **'Junco Bird'**?

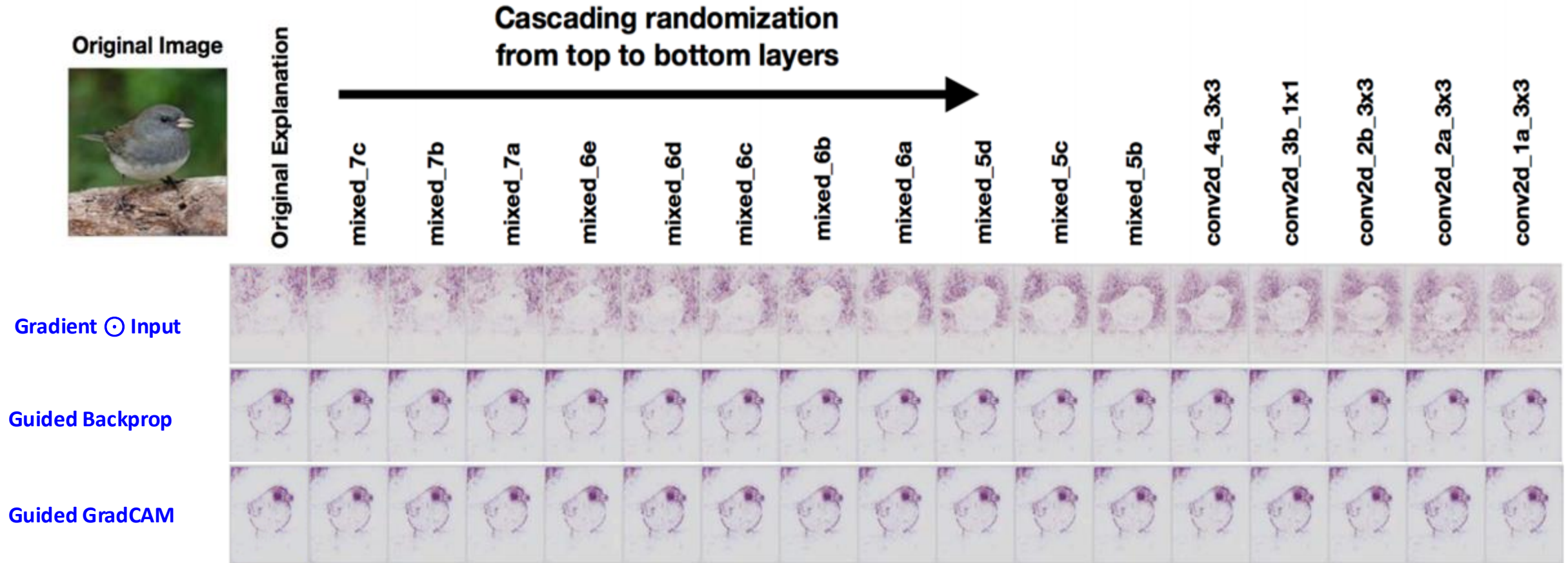


Saliency Map



But beware: “explanation” might be misleading

Model parameter randomization test





University
of Glasgow
Oilthigh Ghlaschu

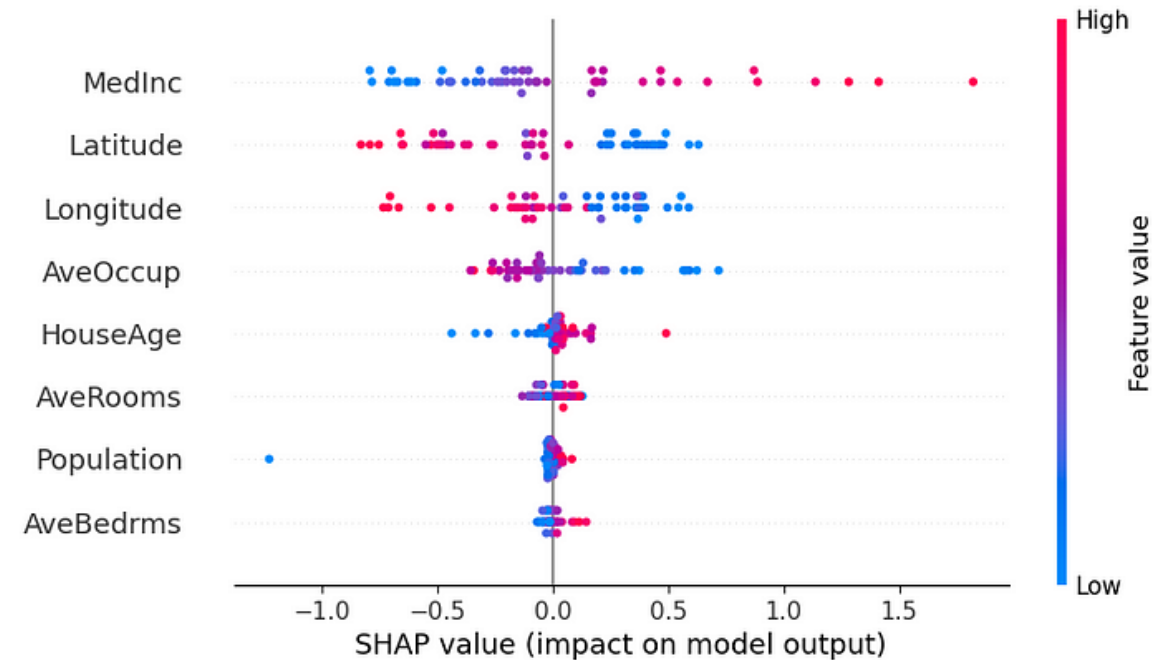
Global explanations



University
of Glasgow
Oilthigh Ghlaschu

SHAP (SHapley Additive exPlanations)

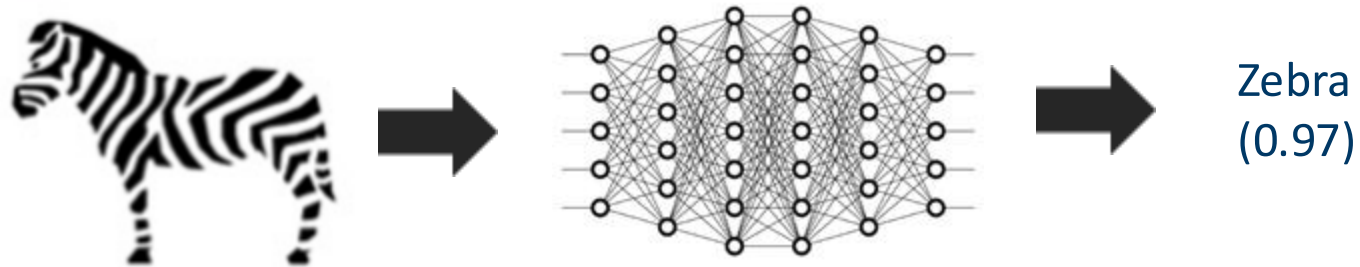
- Explains model overall i.e. the importance of features globally and how feature values contribute to a decision





University
of Glasgow
Oilthigh Ghlaschu

Representation Based Explanations



How important is the notion of “stripes” for this prediction?

[Kim et. al., 2018]



University
of Glasgow
Oilthigh Ghlaschu

Human-Centric Explanations



Explainability versus Interpretability

- Explainability = **system-centric** ability of an AI system to explain itself
 - Measured through fidelity, consistency, etc – technical metrics
- Interpretability = **human-centric** ability of a user to build an appropriate *mental model* that guides interaction with the AI system
 - Understanding of how the system works
 - Being able to use the system successfully
 - Being able to 'trouble-shoot' system and fix 'mistakes'
 - Measured through subjective and objective understanding, satisfaction, acceptability, trust/reliance

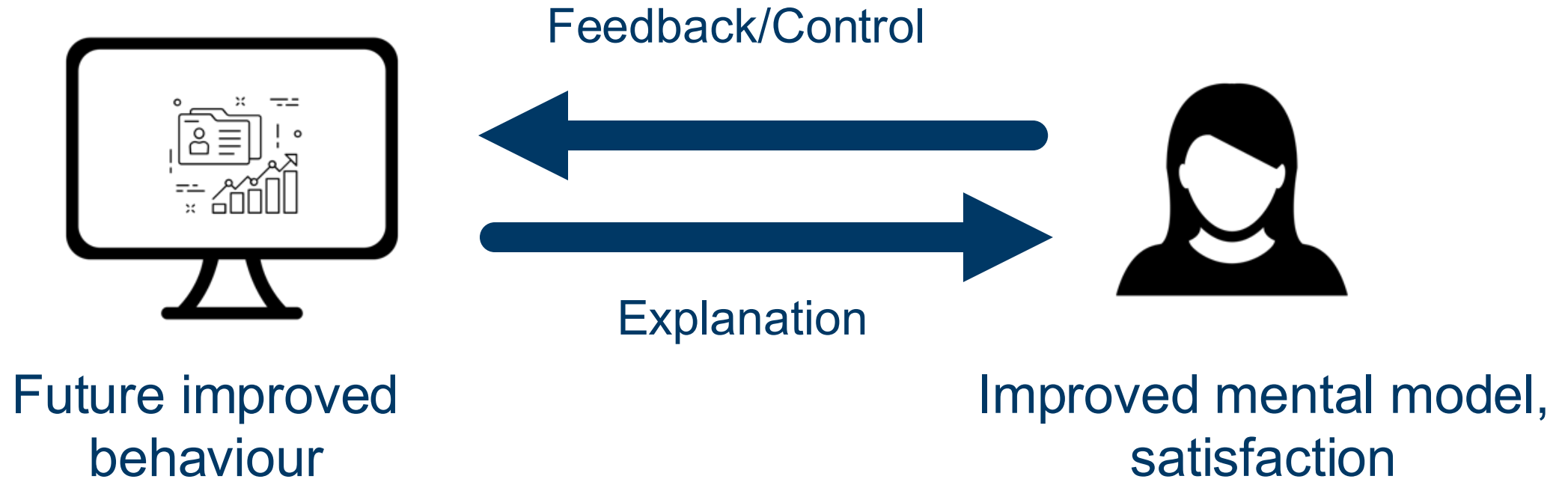
For mental model see:

- Norman 1983
- Johnson-Laird 1983



University
of Glasgow
Oilthigh Ghlaschu

Explanatory debugging for interactive machine learning



See:

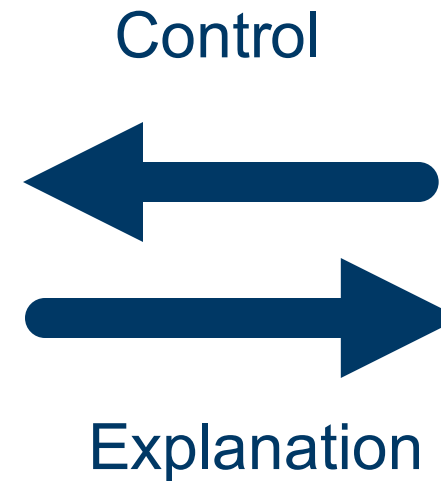
- Stumpf et al. IJHCS 2009
- Kulesza et al. TiiS 2011
- Kulesza et al. CHI 2012
- Das et al. AI 2013
- Kulesza et al. IUI 2015



University
of Glasgow
Oilthigh Ghlaschu

Explanatory debugging principles

- Explanation
 - Iterative
 - Sound
 - Complete
 - Don't overwhelm
- Control
 - Actionable
 - Incremental
 - Reversible
 - Honour feedback





Study design

- 2 versions
 - Elucidebug with explanations and ability to control with feature feedback
 - Control with no explanations and only ability to label instances to correct AI
- 20 newsgroups: Hockey and baseball
- 77 participants
- Measurements
 - NASA TLX
 - Mental model score
 - Amount of feedback
 - Accuracy F1 score



Message Predictor 1.0.5.26968

Move message to folder... Only show predictions that just changed OFF Search Stanley Clear

Folders

- Unknown (1,180 messages)
- Baseball 4/8 correct predictions

Prediction totals

Hockey 278

Baseball 917

Messages containing "Stanley"

Baseball

Hockey

Unknown

Messages in the 'Unknown' folder

Original order	Subject	Predicted topic	Prediction confidence
9287	Re: Playoff Predictions	Hockey	99%
9294	Re: Schedule...	Baseball	60%
9306	Paul Kuryla and Canadian Work	Hockey	99%
9308	Re: My Predictions For 1993	Baseball	64%
9312	Re: NHL Team Captains	Baseball	64%
9316	Re: ugliest swing	Baseball	63%
9319	Re: Octopus in Detroit?	Hockey	67%
9339	Sparky Anderson Gets win #2000, Tigers beat A's	Baseball	99%
9347	Re: Goalie masks	Baseball	53%
9362	Re: Young Catchers	Baseball	82%
9371	Re: Winning Streaks	Baseball	53%
9379	Royals	Baseball	64%
9390	Phillies Mailing List?	Baseball	65%
9410	Reds snap 5-game losing streak: RedReport 4-18	Baseball	98%
9423	Re: Juggling Dodgers	Baseball	57%
9424	Re: Candlestick Park experience (long)	Baseball	99%
9433	Re: Notes on Jays vs. Indians Series	Baseball	53%
9434	Re: When did Dodgers move from NY to LA?	Baseball	53%
9439	Playoff pool	Hockey	96%
9441	Re: Hockey and the Hispanic community	Hockey	99%
9449	Re: Yooi-lims		

Re: Octopus in Detroit?
From: georgeh@ghsun (George H)
Harold Zazula <DLMQC@CUNYVM.BITN>
>I was watching the Detroit-Minnesota game and thought I saw an octopus on the ice after Yasebert scored the game at two. What gives? >(is there some custom to throw octopus on the ice in Detroit?)
It is a long standing good luck Redwing's tradition to throw an octopus on the ice during a Stanley Cup game. They say it dates back to '52 at the Olympia when the Wings became the 1st team (I think) to sweep the cup in 8 games. A lot harder to throw one from Joe Louis seats than from the old Olympia balcony, though.
Funniest I ever saw was when some Tiger fans threw one on the field during a Detroit/Toronto baseball game ... I was living in California and the folks I was watching with had never heard of hockey and were incredulous when I recognized the octopus BEFORE the camera closeup !!

Why Hockey?

Part 1: Important words
This message has more important words about Hockey than Baseball

baseball hockey stanley tiger

The difference makes the computer think this message is 2.3 times more likely to be about Hockey than Baseball.

AND

Part 2: Folder size
The Baseball folder has more messages than the Hockey folder

Hockey: 7

Baseball: 8

The difference makes the computer think each Unknown message is 1.1 times more likely to be about Baseball than Hockey.

Important words

These are all of the words the computer used to make its prediction.

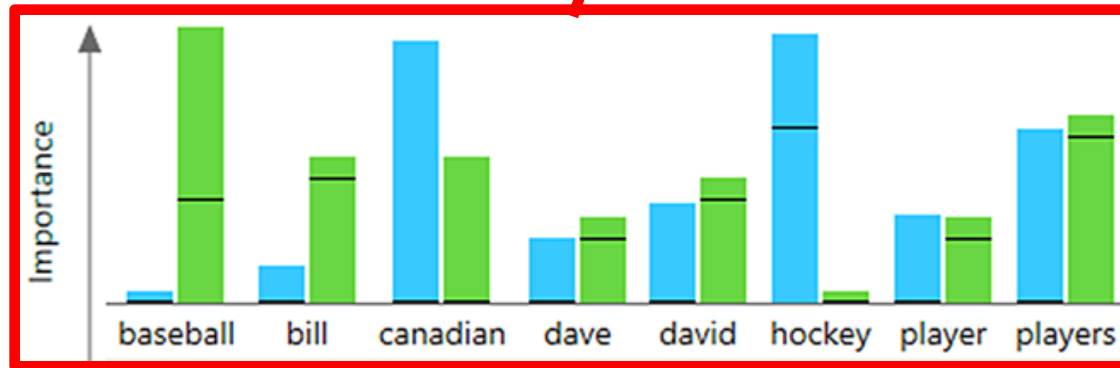
Importance

baseball bill canadian dave david hockey player players prime stanley stats tiger time

Add a new word or phrase

Remove word

Undo importance adjustment



Why Hockey?

Part 1: Important words

This message has more important words about Hockey than about Baseball

baseball hockey stanley tiger

The difference makes the computer think this message is 2.3 times more likely to be about Hockey than Baseball.

AND

Part 2: Folder size

The Baseball folder has more messages than the Hockey folder



The difference makes the computer think each Unknown message is 1.1 times more likely to be about Baseball than Hockey.

YIELDS

67% probability this message is about Hockey

Combining 'Important words' and 'Folder size' makes the computer think this message is 2.0 times more likely to be about Hockey than about Baseball.





Results

- More accurate system with less effort
 - 0.85 for our system versus 0.77% without explanations at end of study
 - Made adjustments to 47 messages while without explanations had to label 182 messages
- With better understanding
 - 15.8 mental model score versus 10.4
 - The more you understand, the better you can make the system
- Does not overwhelm
 - No difference in NASA TLX workload measures



Lots of work to make explanations ‘useable’

- What should be explained?
 - Global/local explanations, intelligibility types, etc.
- How should we explain?
 - Natural language dialogue, textual explanations, visualisations, etc.



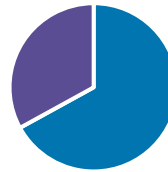
University
of Glasgow
Oilthigh Ghlaschu

Explanation content versus explanation presentation/style

- What information is transmitted in an explanation versus its form and presentation
- E.g. decision confidence

0.67341

67% Accept / 33% Reject



I think it's a little bit more likely that this application should be accepted.



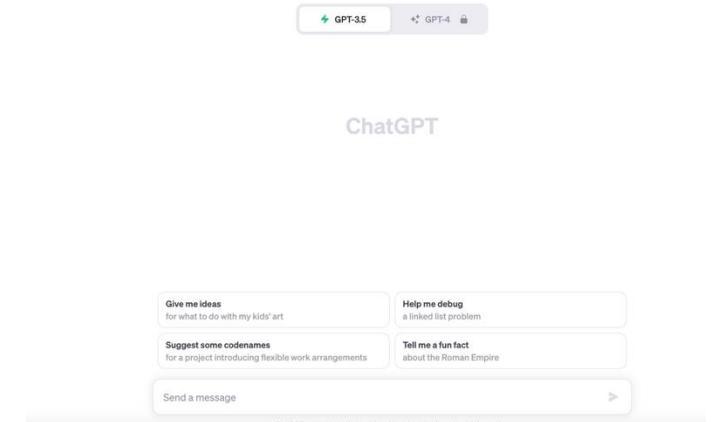
Different stakeholders = different explanations?

- End users / lay users (e.g. loan applicants, patients)
- Decision makers / domain experts (e.g. doctors, judges, loan officers)
- Regulatory agencies (e.g. FDA, European commission)
- Researchers, developers and engineers



University
of Glasgow
Oilthigh Ghlaschu

Transparency for other kinds of AI





Problems with current explanations for generative or autonomous AI

- Explanations are delivered in visual form – no good for certain situations or people
- Explanations are meant to be pondered – not sure how to integrate into real-time settings for human-AI collaboration
- Currently we have a narrow view of explanations – what do we mean by ‘explanations’ and what should be explained



Explanations for LLMs

- Chain-of-Thought Reasoning
 - Explains the steps the LLM goes through in creating the response
 - But is really what is going on?
- Attribution models
 - Show the attention weights in LLMs
 - E.g. BertViz: <https://github.com/jessevig/bertviz>

what is the difference between a tomato and a potato. Explain your reasoning.

Short answer: they're both in the nightshade family (Solanaceae) but are completely different plant parts and used differently — a tomato is a fleshy fruit (develops from the flower and contains seeds), while a potato is an underground storage tuber (a swollen stem/root-like organ) used to store energy.

How I reason this (step-by-step)

1. Botanical origin / organ

- **Tomato:** develops from the flower's ovary and contains seeds → **botanical fruit** (specifically a berry).
- **Potato:** is a **tuber**, an enlarged underground stem (a storage organ). It does not develop from a flower into a seed-containing structure.

2. Taxonomy

- Both belong to the same plant family (Solanaceae), so they're related, but they are different species: tomatoes are usually *Solanum lycopersicum* and common potatoes are *Solanum tuberosum*. Being in the same family explains some shared traits (leaf shape variety, flower structure), but different species explain their different organs and uses.

3. Reproduction

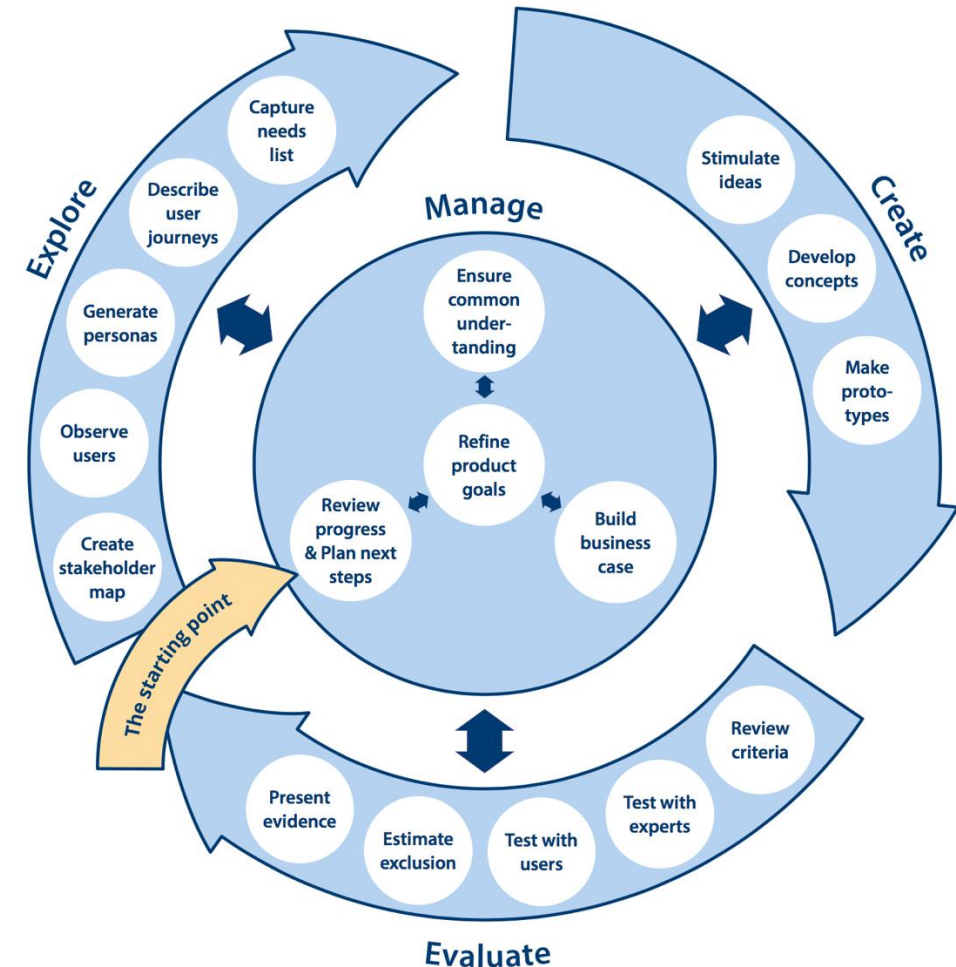
- **Tomato:** reproduces sexually — flowers → fruits → seeds. You can grow new plants from its seeds.
- **Potato:** commonly propagated asexually — plant pieces of tuber with "eyes" (buds) sprout new stems. Potatoes also produce flowers and true seeds (true potato seed) but that's not the usual method gardeners/farmers use.





Human-centric explainable AI (HCXAI) design

- Need to know who the user is and what they need to know
- What is the purpose of the explanations?
- Usually a combination of global and local explanations
- Measure explanations in terms of the purpose and other associated effects





HCXAI Challenges

- No explanations desired for certain tasks and contexts [Bunt et al. IUI 2012]
- Different people need different explanations [Gunning et al. Science Robotics 2019]; lay users neglected at the moment
- Explanations affect trust and reliance [Bussone et al. ICMI 2015, Holliday et al. IUI 2016, Nourani et al. HCOMP 2019]; “placebic” explanations [Eiband et al. CHI 2019]
- Explanations might come from outside of the ML [Ehsan et al. CHI 2021]
- Explanations, and then what? [Wang et al. 2022]
- Explanations for other purposes or modern technologies (e.g. fairness, LLMs)



Summary

- Transparency is required and XAI has made some strides towards opening the black box
- However, 'transparency' is a very vague term and 'explanations' can come in different forms
- Need for a human-centred approach to transparency and explanations
- Consider what explanations are used/useful for



Resources

- Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. *interactions* 4, 6: 42–61. <https://doi.org/10.1145/267505.267514>
- Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *Int. J. Hum.-Comput. Stud.* 67, 8: 639–662.
- Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*, 126–137. <https://doi.org/10.1145/2678025.2701399>
- Don Norman. 1983. *Some observations on mental models*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, US.
- David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37. <https://doi.org/10.1126/scirobotics.aay7120>
- Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. 2119–2128. <https://doi.org/10.1145/1518701.1519023>
- Christoph Molnar. *Interpretable Machine Learning*. Retrieved February 5, 2020 from <https://christophm.github.io/interpretable-ml-book/>
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (Un)reliability of Saliency Methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen and Klaus-Robert Müller (eds.). Springer International Publishing, Cham, 267–280. https://doi.org/10.1007/978-3-030-28954-6_14
- Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20)*, 607–617. <https://doi.org/10.1145/3351095.3372850>
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*, 2668–2677. Retrieved December 11, 2018 from <http://proceedings.mlr.press/v80/kim18d.html>
- Thanks to Hima Lakkaraju and her tutorial on XAI!
- Zijie J. Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E. Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. 2022. Interpretability, Then What? Editing Machine Learning Models to Reflect Human Knowledge and Values. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, 4132–4142. <https://doi.org/10.1145/3534678.3539074>