IS4246
Smart Systems and AI Governance

Lecture 4

NUS
National University of Singapore
National University of Singapore

Agenda

- Review From Last Time
- Explainable AI

Learning Objectives

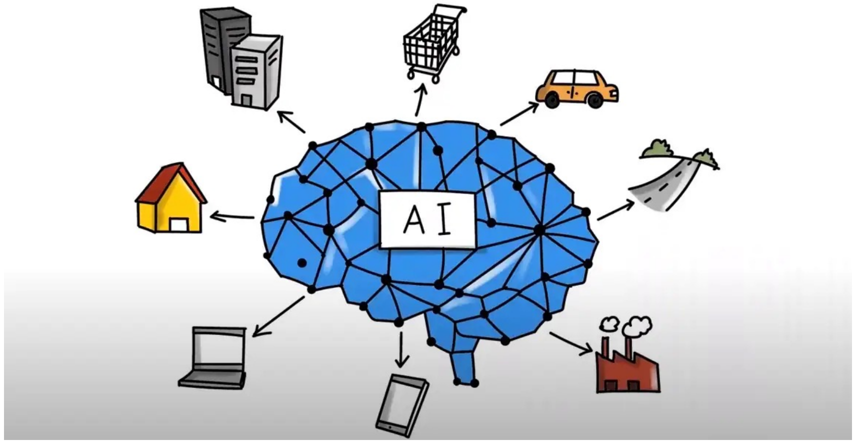
1. Understand the importance of XAI.
2. Explain Interpretability vs Explainability.
3. Understand how to explain black-box model decisions.
4. Appreciate differences between explainability methods.
5. Understand current/potential use cases for explainability in critical applications.

Explainable AI (XAI)

- XAI aims to help humans understand why a machine decision has been reached and whether or not it is trustworthy
- Its goal is to enable and widen acceptance of AI systems by humans
- XAI bridges the gap between machine intelligence and human intelligence

Critical importance of XAI

- Intelligent systems offer great possibility
- XAI raises concern of giving systems too much power
- Explanations of decision-makings processes must be understandable to domain experts
- XAI encourages creating human-like solutions and studying the brain
- User rights must be protected when machines take over decision process



Reasons to Explain

- Explain to justify
- Explain to control
- Explain to improve
- Explain to discover

Explain to justify

Controversies over AI/ML enabled systems yielding biased or discriminatory results

Need for explanations to ensure AI based decisions were not made erroneously

Explanation for a decision = need for reasons or justifications

Need for explanations to ensure compliance with legislation (e.g. GDPR)

European legislation that allows people to interrogate AI

Explain to control

- Need to understand more about system behavior
- Greater visibility over unknown vulnerabilities and flaws
- Rapidly identify and correct errors in low criticality situations (debugging)

Explain to Improve

- Need to be able to explain and understand the model for it to be more easily improved
- Knowing why the system produced specific outputs will know how to make it smarter

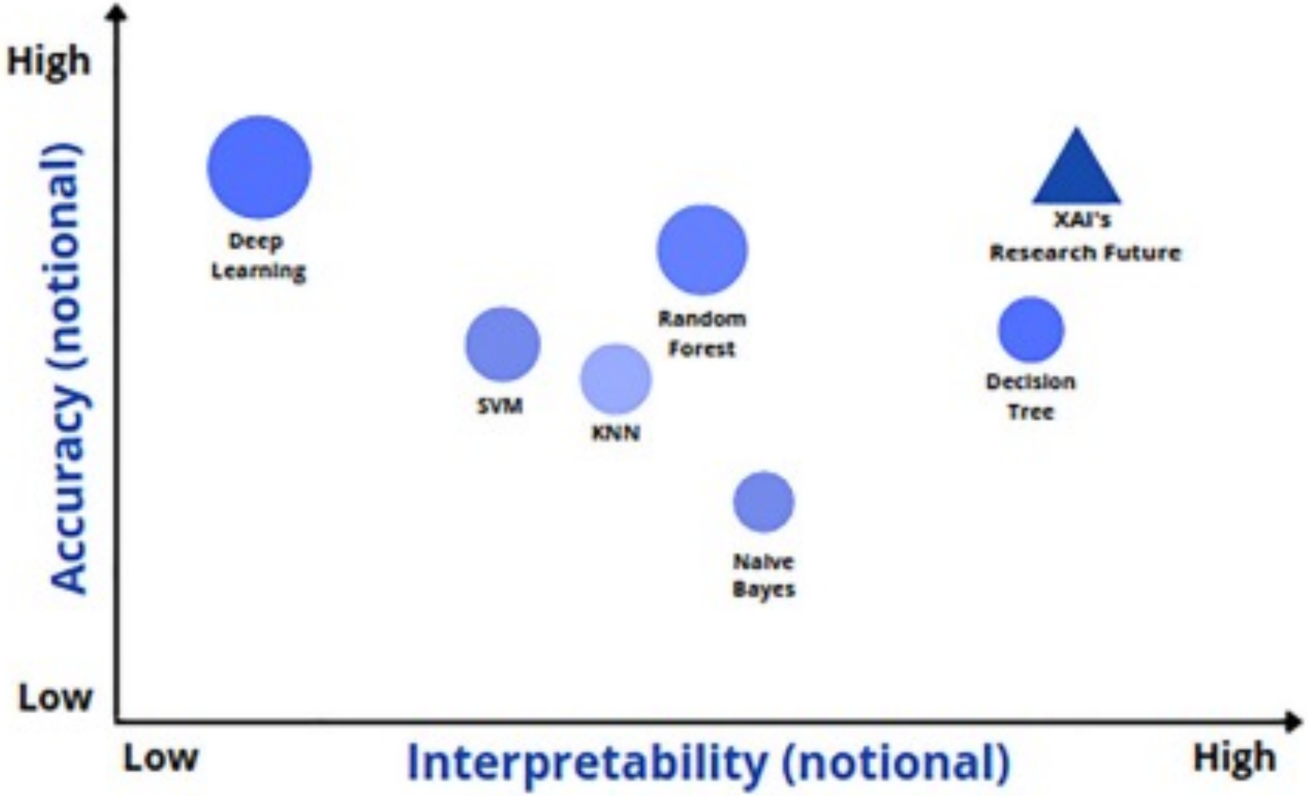
Explain to Discover

like for machines like stockfish etc it doesn't always explain how stockfish knows to make this move etc, but with further changes in AI, we might be able to have some kind of english explanations to why Stockfish make their decisions

Asking for explanations to learn new facts, to gather information and thus to gain knowledge

XAI models to teach us about new and hidden laws in science

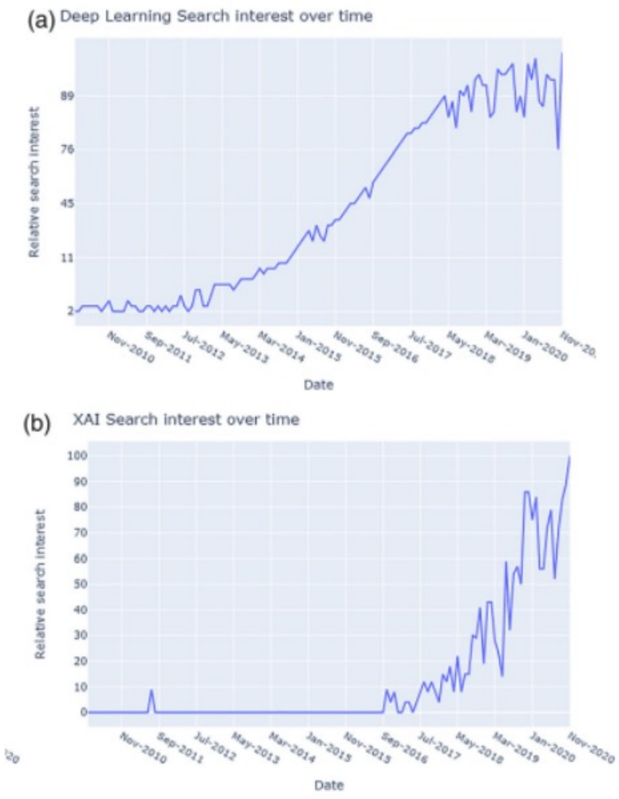
Trade-offs between Accuracy and Interpretability?



Accuracy vs. interpretability for different machine learning models

Historical Significance

- Early forms of AI & ML were interpretable & self-explanatory
- Increase in data complexity led to a focus on accuracy, forgetting explainability
- Recently, explainability has regained importance, necessary for acceptance by society & regulatory authorities
- Still an open research area for SVMs and Deep Learning, Neural Networks



Transparency, Interpretability and Explainability

if our model is so complex that we are unsure how the model arrived at xxx output, it's a black box

- **Transparent**: the *model's* potential to be understandable, *opposite to "black-box"*
- **Interpretable**: capacity to *provide interpretations* understandable by humans

saying in a "human way" whats going on

- **Explainable**: provides explanations as an interface between humans and an AI system. Must be both accurate and comprehensible

the accuracy of how well the explanation provided by the model matches against what the model really is doing
(i.e. the model just comes up with a random explanation to smoke you)

Principles to Strive For

- **Explanation**: An AI system must supply evidence, support; or reasoning for each decision made by the system.

try to map real world explanations
(you have fever + cough -> you may have flu)

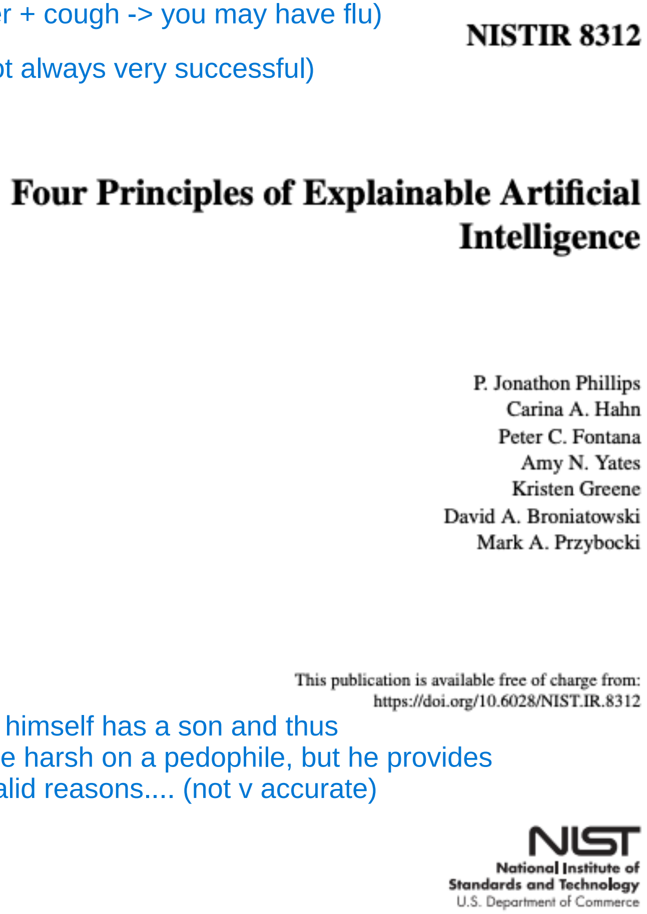
(not always very successful)

- **Meaningful**: The explanation provided by the AI system must be understandable by, and meaningful to, its users. As different groups of users may have different necessities and experiences, the explanation provided by the AI system must be fine-tuned to meet the various characteristics and needs of each group.

- **Accuracy**: The explanation provided by the AI system must reflect accurately the system's processes.

i.e. the judge himself has a son and thus
might be more harsh on a pedophile, but he provides
logical and valid reasons.... (not v accurate)

- **Knowledge limits**: AI systems must identify cases that they were not designed to operate in and, therefore, their answers may not be reliable.



NISTIR 8312

Four Principles of Explainable Artificial Intelligence

P. Jonathon Phillips
Carina A. Hahn
Peter C. Fontana
Amy N. Yates
Kristen Greene
David A. Broniatowski
Mark A. Przybocki

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8312

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

Additional Goals of Explainable Systems

- Causal
- Counterfactual
- Social
- Selective
- Transparent
- Semantic
- Interactive

Causal Explanations

lots of economics models are very accurate but not able to explain why they are accurate, and the more "causal" they are the less accurate they are

- Knowing what relationship there is between input and output, or between input features
- Causal explanations are largely lacking in the machine learning literature
- How to measure the causal understanding of an explanation (causability)
- Measuring the causal understanding of an explanation of a machine statement has to be based on a causal model

lots of MLs are very bad at giving us causal explanations

how can we check the causality of a system? - like doing experiments with controlled variables and seeing what changes

Counterfactuals

- Empirical evidence indicates that humans psychologically prefer counterfactual or contrastive explanations
- People asking why event P happened, instead of some event Q
- Issues related to the diversity and proximity of counterfactuals arise in designing counterfactual explanations

have to ensure that the definitions fed to the AI might not overlap so that the counterfactual explanations can make sense

Social

- Interactive transfer of knowledge tailored for the recipient's background and level of expertise
- Explanations involving one or more explainers and explainees engaging in information transfer
- Conversational or argumentative processes can enhance user's inspection of explanations and increase trust in the system

interesting that since gpt isn't an expert but it's good at explaining it's process

i.e. ask it to explain step by step, or to argue with it on some of it's points

Are you able to get it to give you an explanation or reason, that it REALLY believes in? I.e. is it telling you not to skip lunch because it doesn't want you to be anorexic or is it because they want you to perform well for your competition

Selective Explanations

- Explanations do not always need to be complex representations of the real world
- Informational content of explanations must be selected according to user's background and needs
- Explanations can be tailored to doctor's level of technicality or lay user's need for simplicity

Transparency

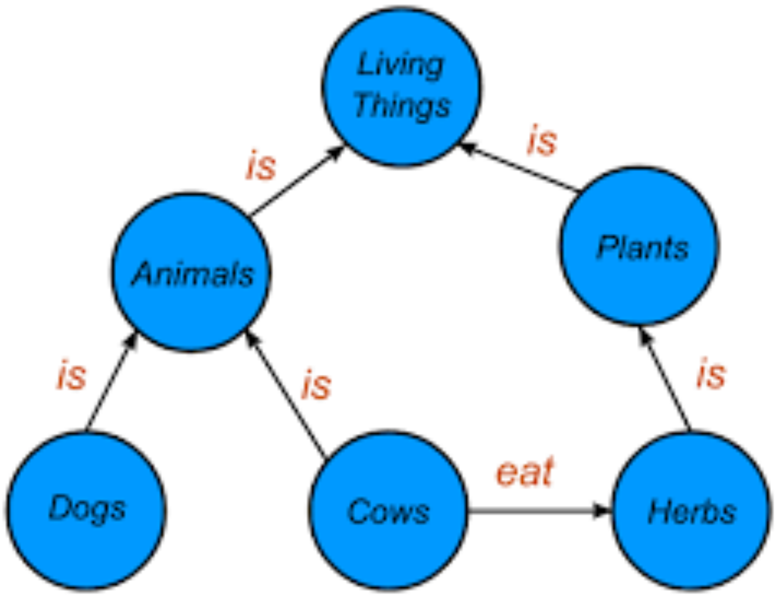
- Explanations should help in understanding the underlying logic and identify wrong system behaviour
- Trade-off between transparency and privacy must be found when generating explanations
- Differentially private model should be used to generate local and global explanations

like if the company's AI is trained to identify fraud, you can't really afford to be too transparent, otherwise the fraudsters will know how to beat your AI

Semantic

- Symbolic grounding by means of ontologies, conceptual networks, or knowledge graphs
- Formal representation and reasoning for knowledge manipulation
- Manner in which to provide personalized explanations for different stakeholders

microsoft and google have been trying to create these logical maps that make coherent sense in like what is related to what and how.



ensemble
- by combining output from many different models together

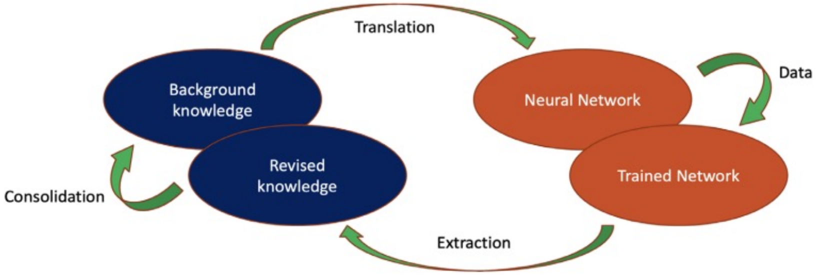
for example fraud detection
- there could be 1 model to check IP, another to check name, another to check previous transactions
- this makes it so that we can very easily tell why the overall output is xxx based on the individual ouputs of each more specific model (i.e. flagged for suspicious I.P)

Interactive Explanations

- Explanations should be interactive and allow explainee to revise and consolidate knowledge Background knowledge can be used for meaningful semantics of explanations
- Background knowledge injected back to improve model performance

social is like asking the model to elaborate, explain and argue etc.

interactive is like you are feeding more data in the model and it will "grow"



Taxonomy of approaches

Model types:

- Transparent
- Opaque

Explanation methods:

- Model-agnostic
- Model-specific
- Explanation by simplification
- Explanation by feature relevance
- Visual explanation
- Local explanation

Transparent Models

k Nearest Neighbor

- kNN, decision trees, rule-based learning, and Bayesian networks
- Transparent decisions, but transparency in itself doesn't guarantee explainability

baysian model - have an initial guess that constantly updates its guesses based on the inputs

Model-Agnostic XAI Approaches

- Designed to be general
- Relate input of a model to its outputs without depending on the intrinsic architecture

Model-Specific XAI Approaches

- Bring transparency to a particular type of model by taking advantage of its features

Explanation by Simplification

- Alternate model such as a linear model or decision tree to explain a more complex model

Explanation by Feature Relevance

- Evaluate a feature based on expected marginal contribution to the model's decision after considering all combinations

doesnt really tell you how the model is working in the back end but might tell us more info that is more relevant to us

Visual Explanation

- Data visualization approaches to interpret the prediction or decision over the input data

Local Explanations

- Approximate the model in a narrow area around a specific instance
- Explain how the model operates when encountering similar inputs

we can give people more relevant information, that is localized to you and more useable

also can conceal information regarding the larger model

Opaque Models

- Random forest, neural networks, and SVMs
- High accuracy, but not transparent

Overview of Taxonomy

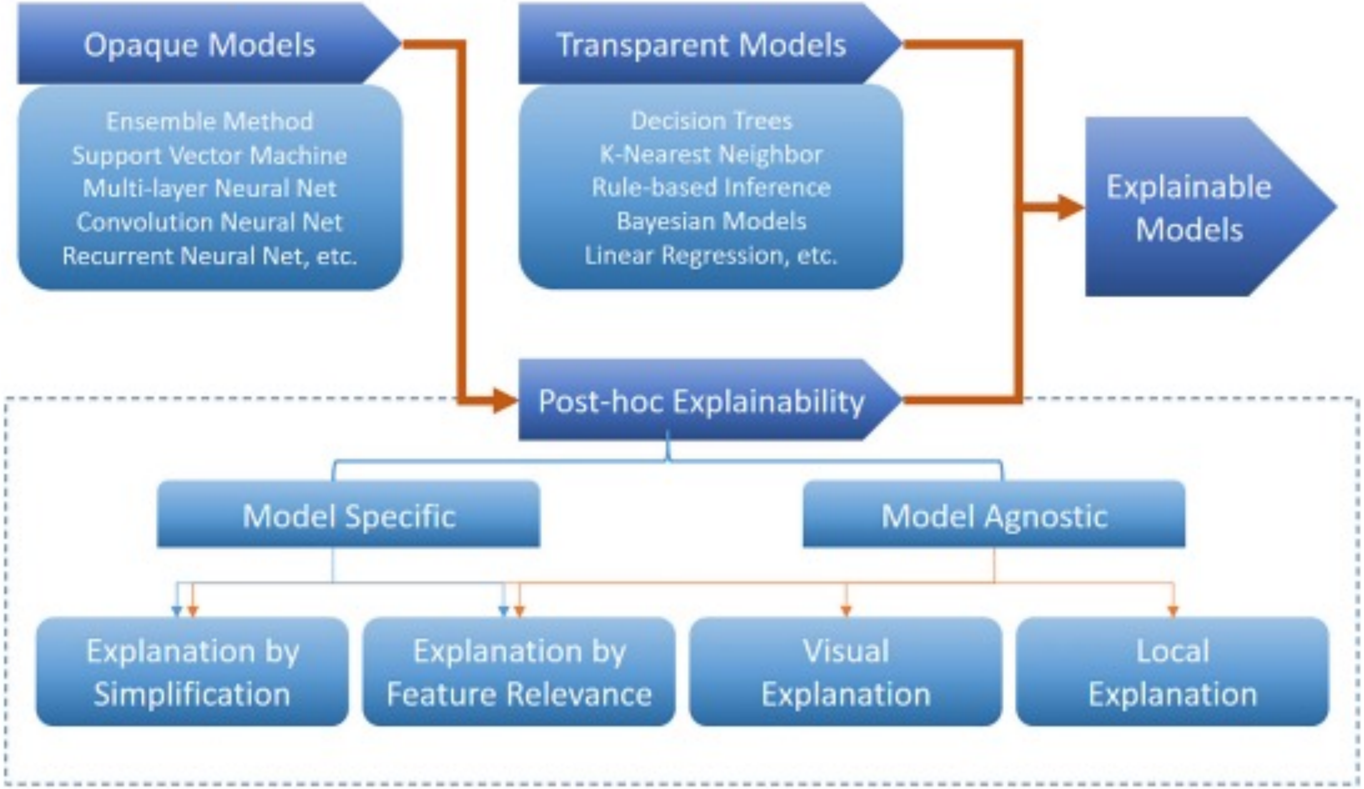


FIGURE 3 The high-level ontology of explainable artificial intelligence approaches

State of the Art

Widely Used Methods

Features-oriented methods (e.g. SHAP)

Class activation maps for CNNs

congrigutional net?

Global methods (GAMs)

Concept Models

Surrogate Models

Local Explanations (e.g., LIME)

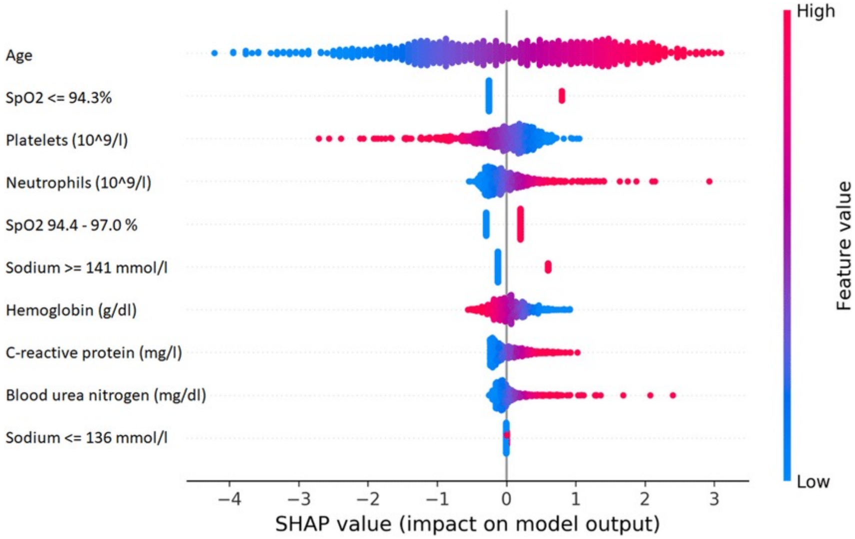
Local Pixel-Based

Human-Centric Methods

Features-oriented methods (e.g. SHAP)

- SHapley Additive exPlanation (SHAP) is a method to explain the contribution of the features to a prediction
- SHAP is based on the concepts of game theory and the Shapley Value which assigns an individual contribution to each factor in the model

helps to identify the "weight" of each attribute or feature that helps to tell you how an outcome is reached
like "race" or "sports achivements" for entering a school



Features-oriented methods (e.g. SHAP)

- SHAP deconstructs the ML model into individual features and then determines the contribution of each of those features to a given output • SHAP can work with any machine learning model and is model-agnostic
- It works by calculating the importance of each feature by comparing it to a "reference" value, which is often the average output of the model
- SHAP is designed to be able to work with both linear and non-linear models

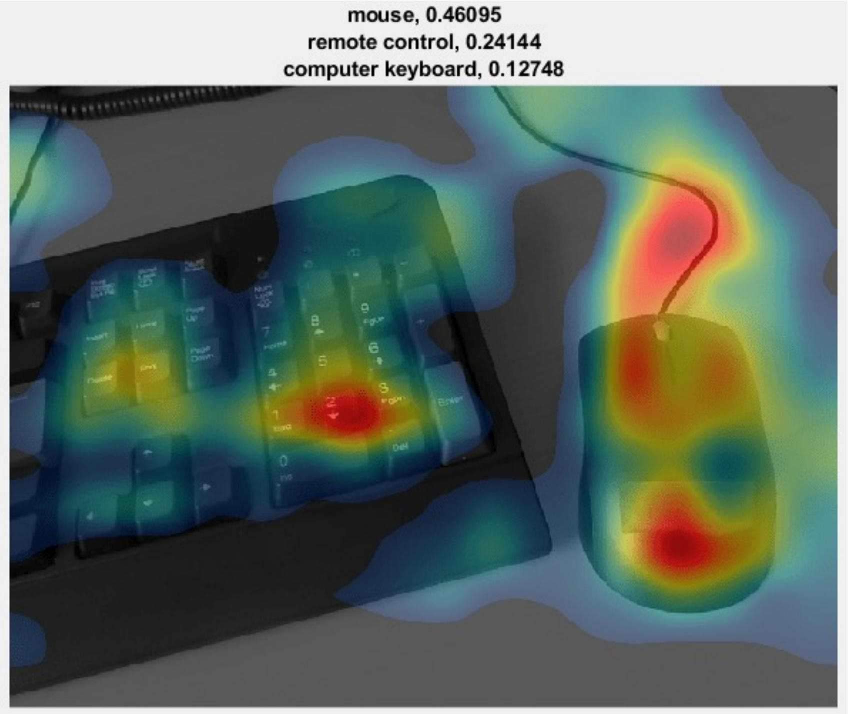
Class activation maps (CAMs) specific to CNNs

- CAMs represent the per-class weighted linear sum of visual patterns
- Applied to final convolutional feature map prior to output layer
- Highlight areas in input image most influential over CNN decisions
- Cannot be applied to pre-trained networks

(might have changed due to advancement of technology)

- Map scaling may lead to loss of spatial info

identify what the AI is "looking" at the most?



identifying features that are used to identify an object as the object, then it will kind of count the number of 'positive' features there are.

this is useful as we can check where the positive values are found on the picture for the CAM

Global Methods

Examples: Craven & Shavlik (1995), Frosst & Hinton (2017), Odense & Garcez (2017), Zhou, Jiang& Chen (2003), Lou et al. (2012, 2013)

Goal: Generate general representations of black-box models and the features it has been trained on

Strategies: Extract decision trees, decision rules and feature importance vectors

Surrogate Models: LIME

pros: lime is widely cited, easy to understand and easy to implement

cons: assumes local linearly, requires a large number of samples around explained instance, and is very computationally expensive

to deal with this linear problem, people have been breaking the datasets into clusters and making multiple LIME analysis

Local Interpretable Model-Agnostic Explanations (LIME)

it is possible to sneak in biases that will slip past LIME

A model-agnostic technique to create explanations of ML models by training surrogate models on a set of perturbed instances of the original data.

lime's perturbed instances are random and can sometimes create unstable output, where everytime you run it you might get some different outputs

Image classification involves perturbing superpixels in an image

Local model is not always informative or reliable at a human level if parameters are chosen based on heuristics

constructs local linear model based on inputs and outputs

- helps you to understand what types of changes affect the output the most

to perturb the model - given an input of interest, LIME makes small changes to the input to see how the output would change

Use Cases

- Medical
- Criminal Justice
- Autonomous Vehicles

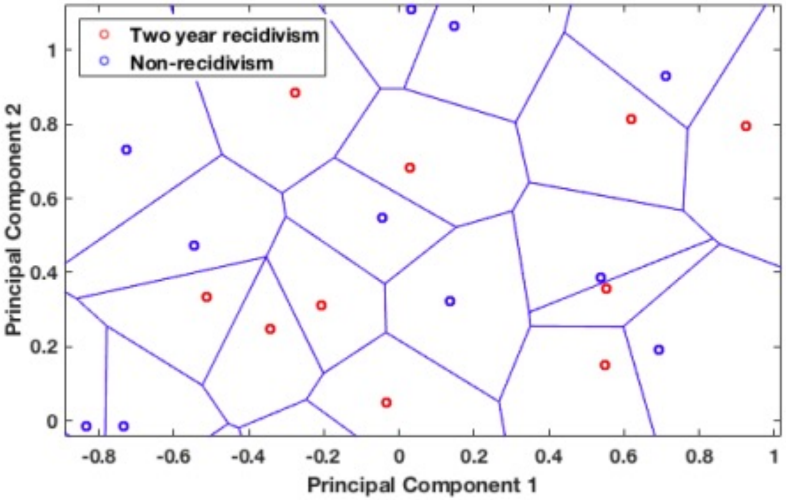
Medical AI Applications

- Growing demand during COVID-19 pandemic • To be trustworthy, transparent, interpretable, and explainable
- Example: Employing DL to Identify COVID-19 via CT scans
- Outperforms GoogleNet, Resnet and VGG-16
- Explanable Architecture for Decision Visualization
- Expandable to include more classes

some researchers created a new AI that outperforms other things to identify cases of covid that were serious, and they were useable to doctors

Criminal Justice

- In some countries, automated algorithms are used to predict criminal behavior
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a widely used criminal risk assessment tool
- There is potential for racial bias to be introduced into predictive models



Autonomous Systems

- XAI and Autonomous Systems
  - Self-driving vehicles - Crash of an Autonomous car owned by Uber (Stilgoe, 2020) -

- Approaches
  - Prototype-based approaches used for understanding visual scene (Soares et al., 2019)
  - Can provide explainable rules

Thank You!