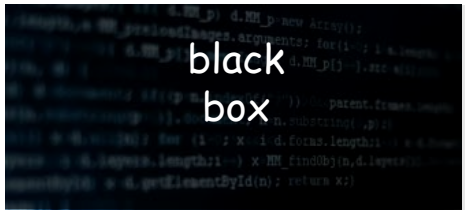
Causal Explanations For Image Classifiers

Hana Chockler

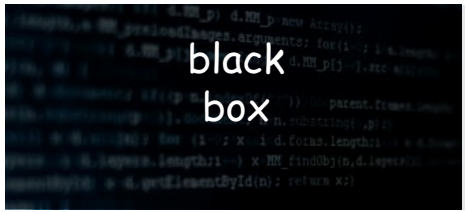
KING'S College LONDON

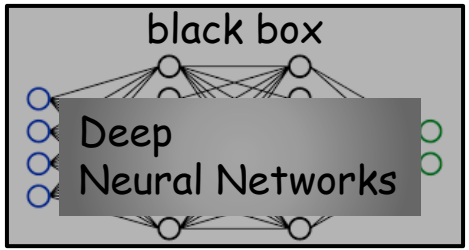
Modern computerized systems are **huge and difficult to understand**



Modern computerized systems are huge and difficult or even impossible to understand

How to explain the system's output?





©Halpern & Pearl, 2001
+
©Halpern – many papers

Actual Causality

A theoretical concept from AI

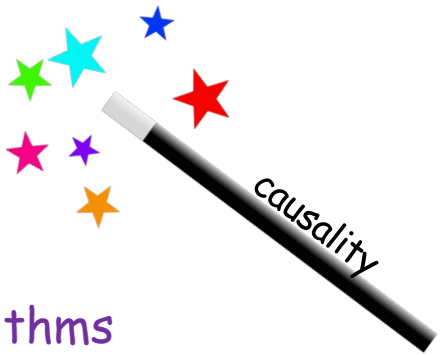
Extends causal counterfactual reasoning

+

Quantification of causality,

allowing to rank causes by importance

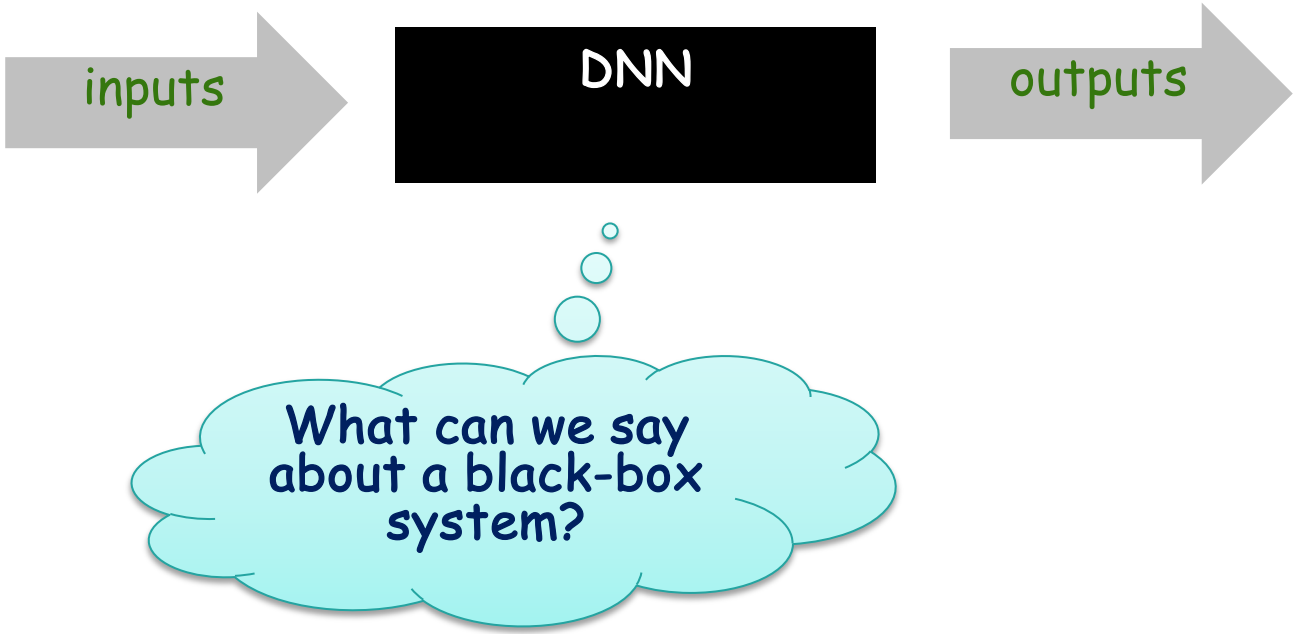
©Chockler & Halpern, 2003



Turns out to be very useful!

Intractable – but there are efficient approximation algorithms and sufficient partial solutions

Reasoning about black-boxes



Reasoning about DNN's decisions

inputs

Input transfor mation

DNN

(causal model)

outputs

Intervene on inputs

Reason about the way DNN makes its decisions

Observe the outputs

We can reason about various properties of the system without opening the black box

Explanations for Deep Neural Network's decisions

DNN for classifying animals

red panda

Explanation: minimal subset of the pixels of the image sufficient to get the same label

Because of this part:

Subtle misclassification – uncovered by explanations

DNN for classifying images

seems ok

cowboy hat

Because of this part:

Explanation uncovered misclassification!

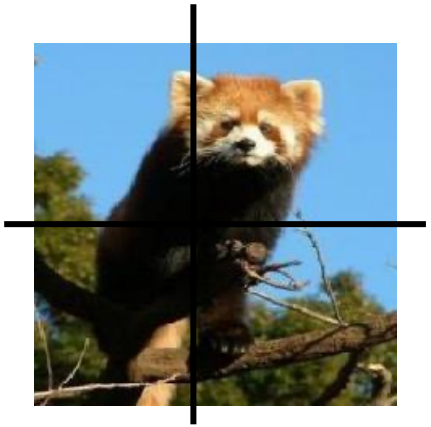
Refine training

High-level structure of

red panda

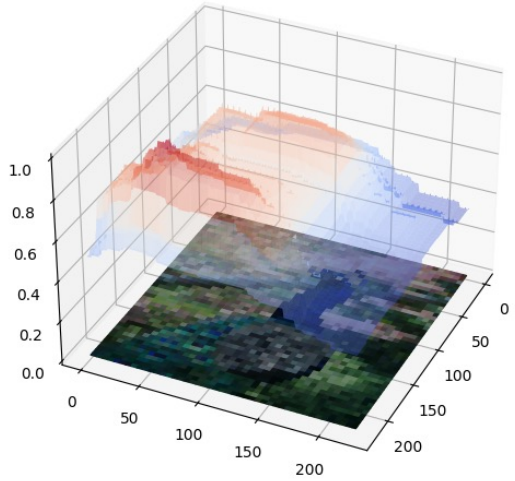
Ranking algorithm

Explanation extraction algorithm



Ranking algorithm

1. Partition into regions
2. Compute responsibility (rank) of each region
3. Order and throw away irrelevant regions
4. Continue with high-ranked regions
5. Repeat with different partitions and take the average

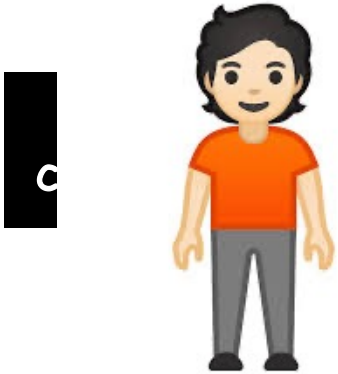


Explanation extraction algorithm

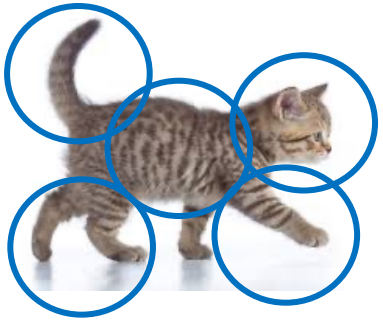
1. Input: a ranked list OR a saliency landscape
2. From the highest ranked pixels, add pixels greedily.
3. Can be spatially-aware or agnostic.
4. Stop when the resulting area(s) get the same label as the input.

Works for non-continuous explanations
and for multiple explanations

Multiple different explanations

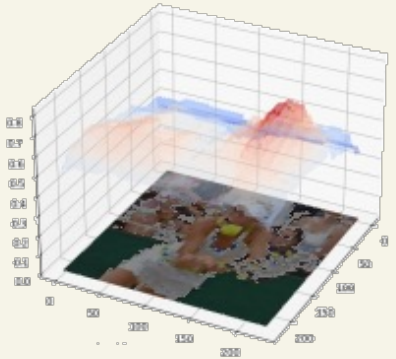


cat



Explanations

[illegible]



B[illegible]

[illegible]

[illegible]

[illegible]

How is trust gained?

BBC: Robin Brant doesn't trust self-driving cars

"Trust Me? (I'm an Autonomous Machine)" **Project**

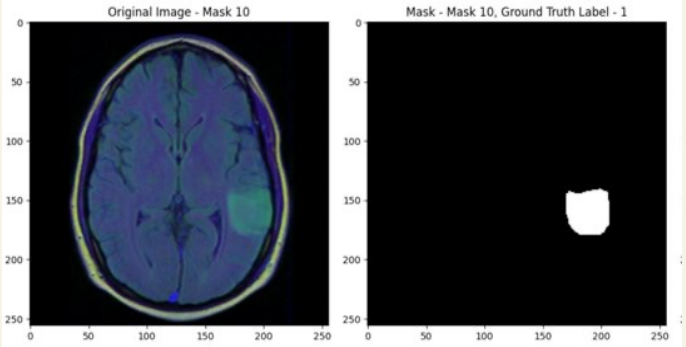
https://trustme-liart.vercel.app/shapes

How is trust gained?
Do explanations increase trust?

Are these explanations causal?
Do they need to be causal?
Do they need to be causal for domain experts?

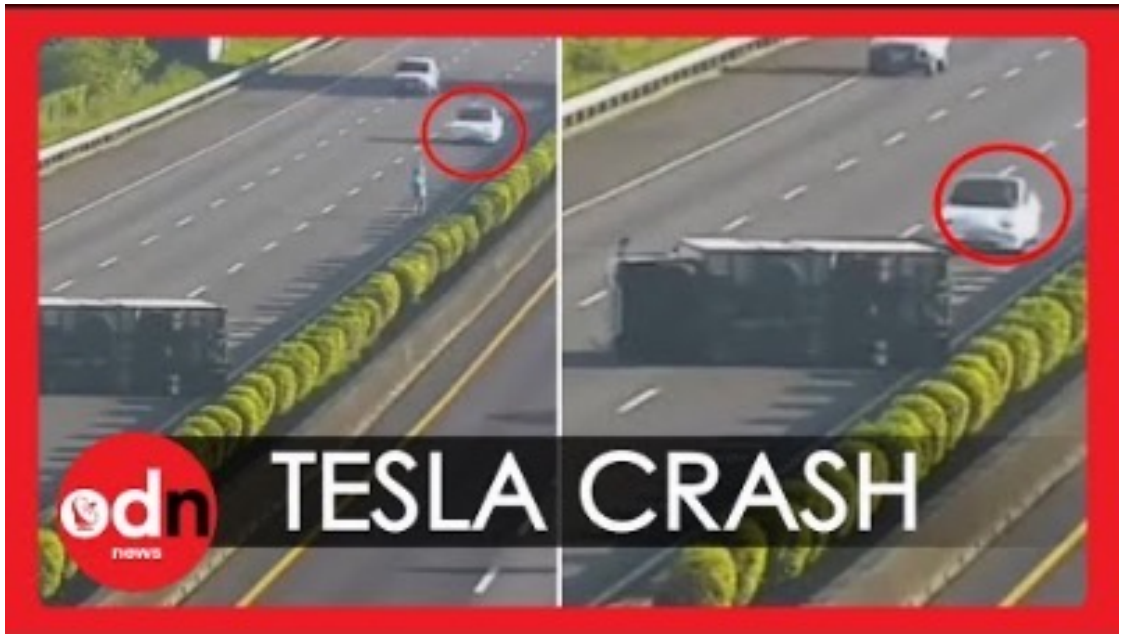
[illegible]

- D[illegible]
- E[illegible] 20 [illegible]88 [illegible], [illegible]4 [illegible]
- A[illegible](256, 256, 3)
- [illegible]F[illegible]AI[illegible]* [illegible] [illegible]0 ([illegible]) [illegible] ([illegible])



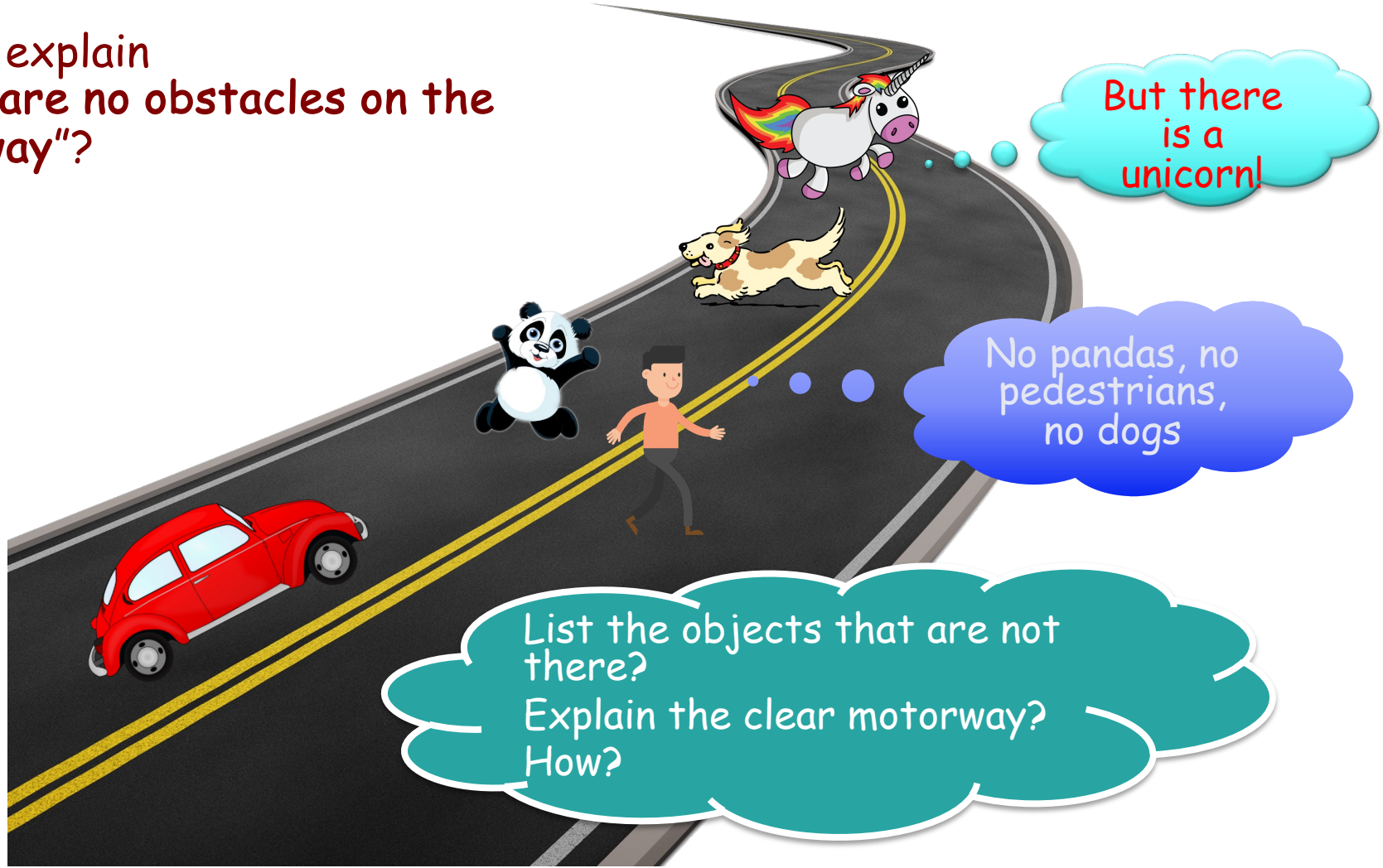
* F[illegible]A[illegible]

Explanation of absence



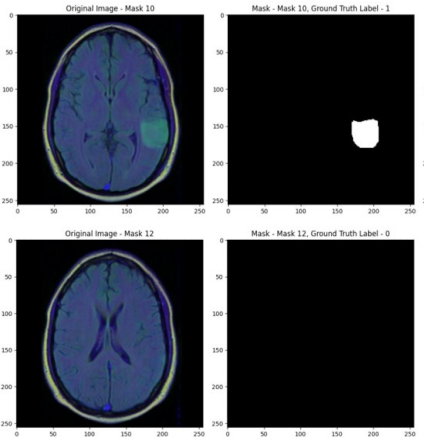
Why did Tesla think there is nothing there?

How to explain "there are no obstacles on the motorway"?



Explanations in the healthcare domain

clinicians



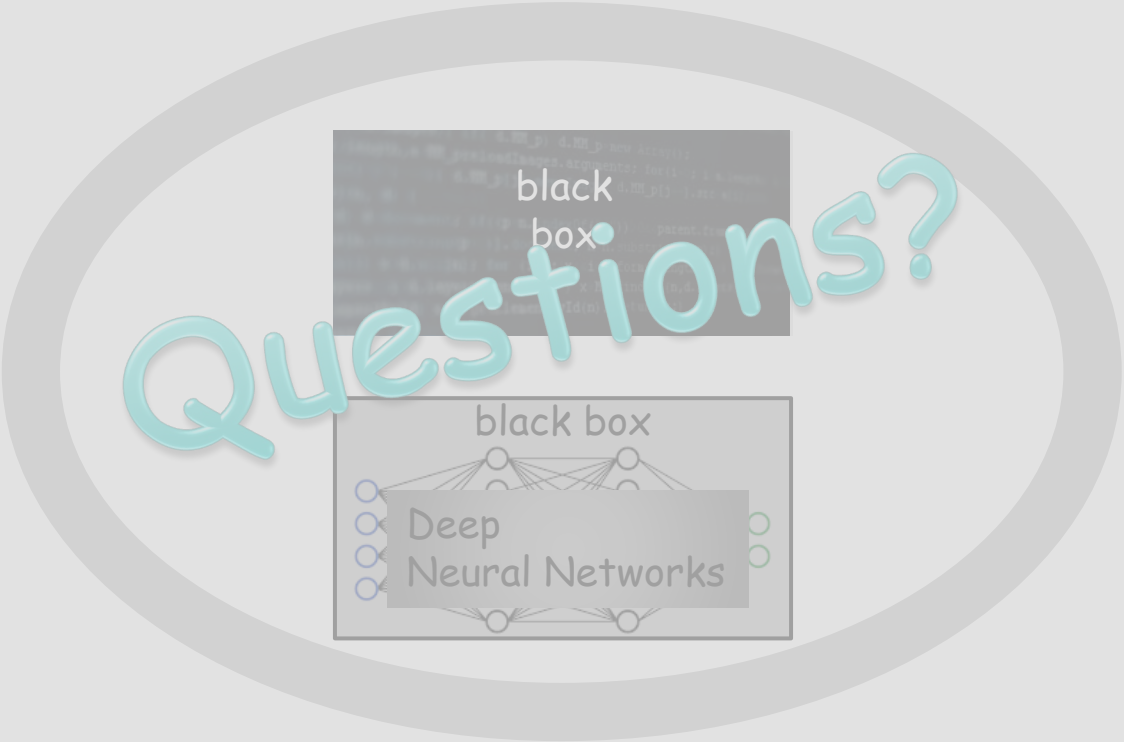
DNN for detecting brain tumours → Yes/No

Input from clinicians:

- If the answer is No, it needs to be explained as well
- More complex scenario: if the clinician thinks there is a tumour, but the classifier's label is "no tumour", the clinician needs an explanation of the negative classification

Open questions / Current work

- Explanations of absence / negative classification
- Really fast explanations
- Explanations for medical professionals
- Explanations of videos
- Explanations of detected deep-fake images
- Explanations of a class of images (“what are the characteristics of pandas?”)


black
box
Questions?
black box
Deep
Neural Networks