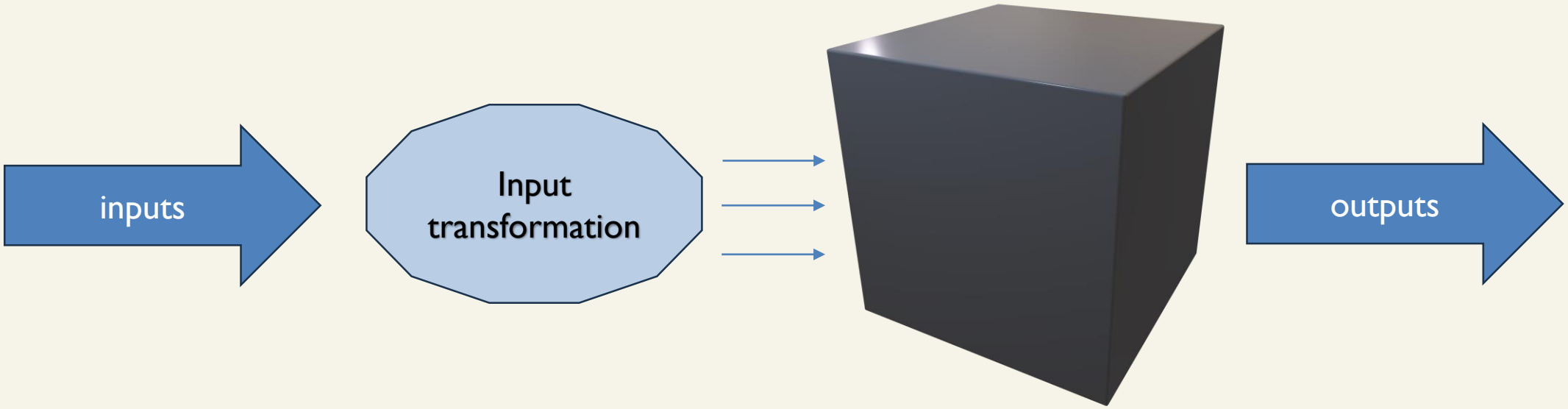
Attacking your black box classifier with

David Kelly
Informatics
King's College, London

Reasoning about black boxes

What can we say about a black box system?

Reasoning about black boxes



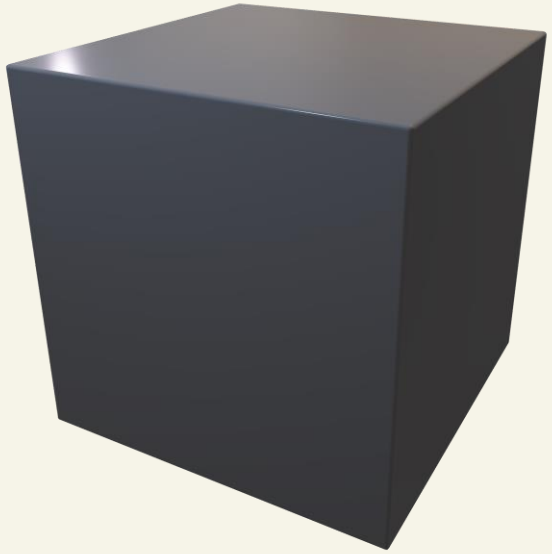
Intervene on inputs

Observe changes in output

Reason about the way DNN makes decisions

Explanations of DNN decisions

Input



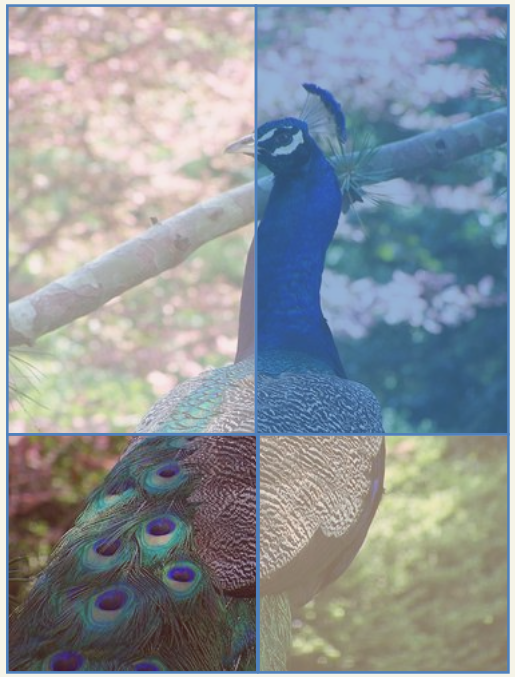
peacock

Causal Explanation:
Minimal, sufficient,
non-trivial subset of
the pixels of the image

Because of this part!

Explanations of DNN decisions

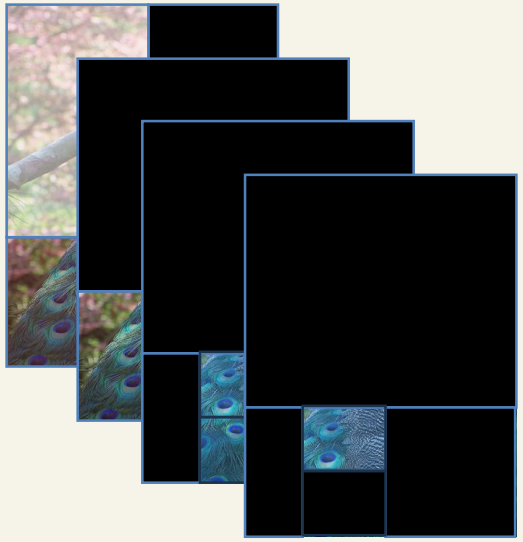
- Iterative partitions
- Flexible distribution
- Refinement
- Weighted responsibility
- Guided search
- Spatially aware explanations
- Disjoint explanations
- Multiple explanations

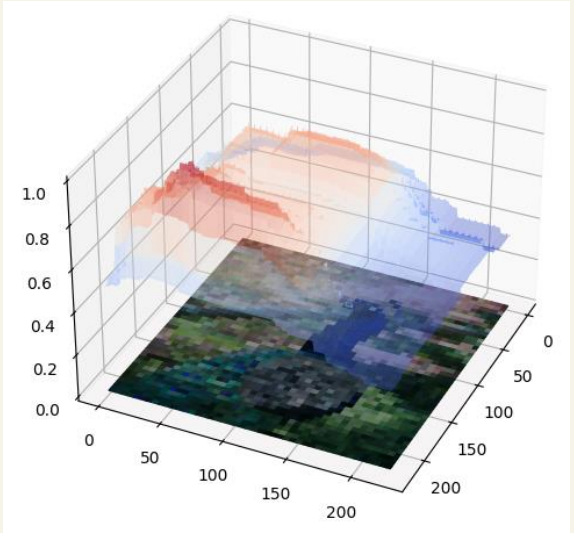


ReX – High Level View

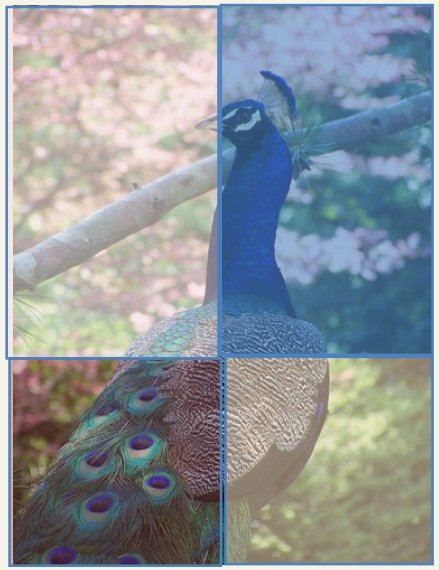
Causal Ranking

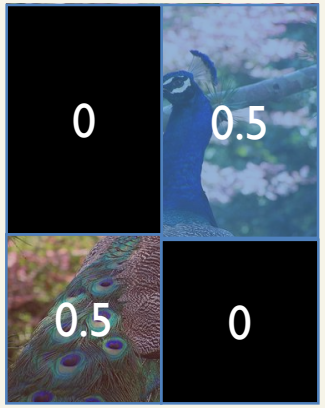
Extraction of Explanation(s)

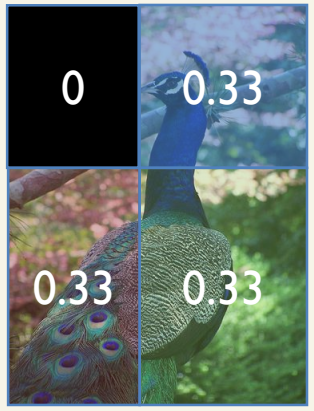




ReX – High Level View



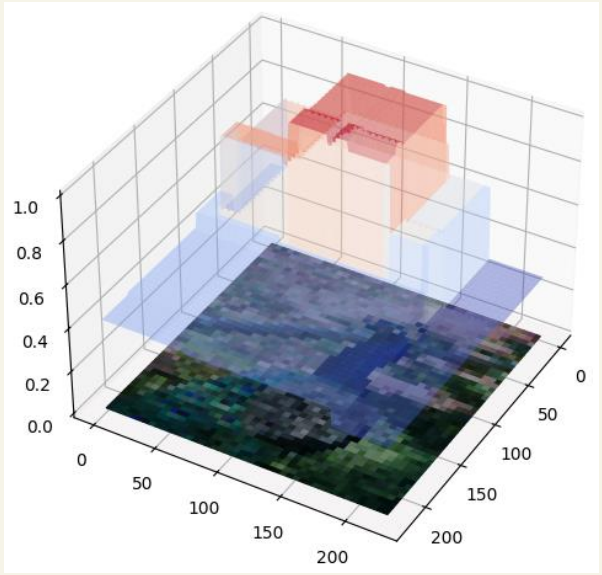




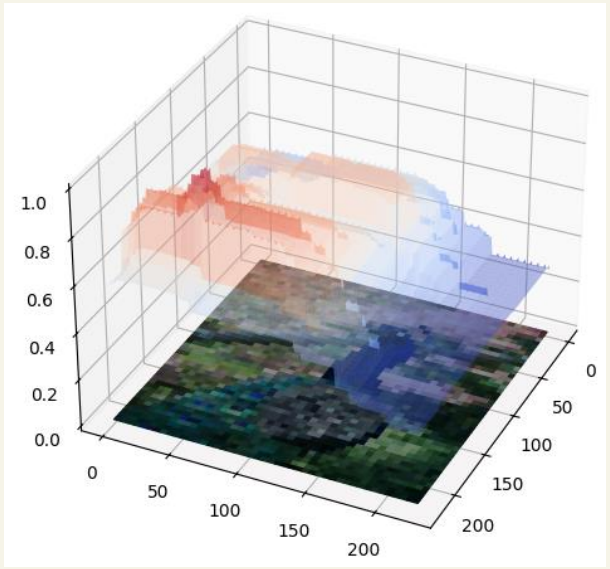
- Partition into regions
- Compute responsibility of each region
- Order and throw away irrelevant regions
- Continue with high-ranked regions
- Repeat many times and take the average

Explanations of DNN decisions

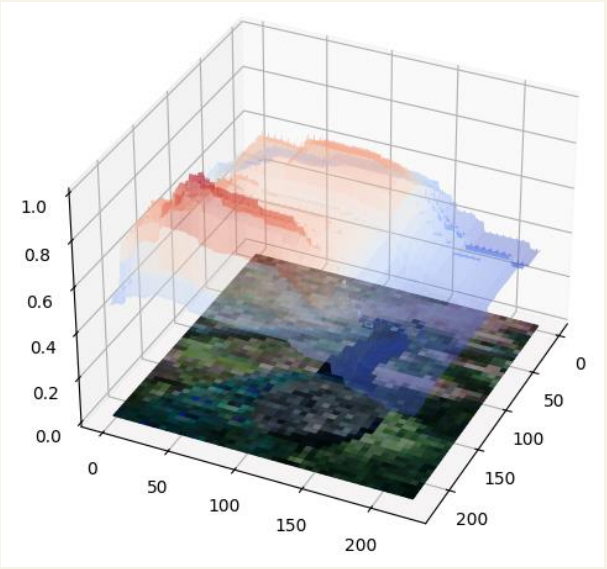
After 1 iteration



After 10 iterations



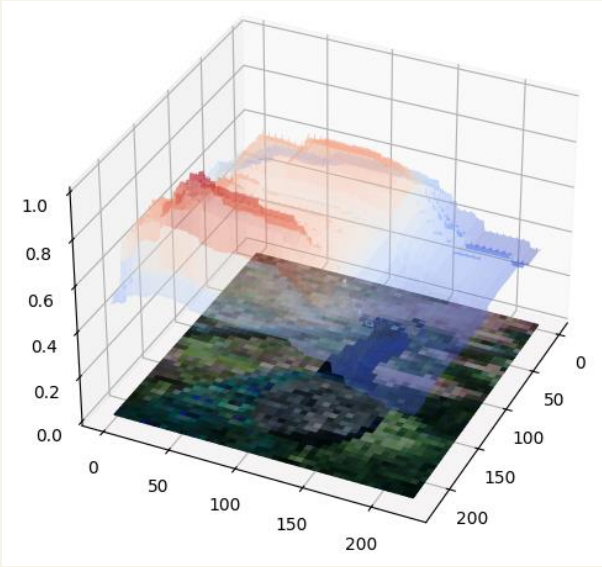
After 20 iterations



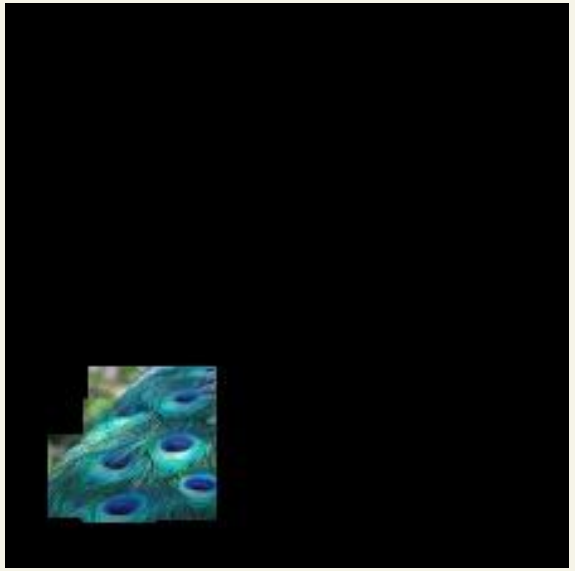
We end up with a saliency landscape from which we extract explanations

Explanations of DNN decisions

After 20 iterations



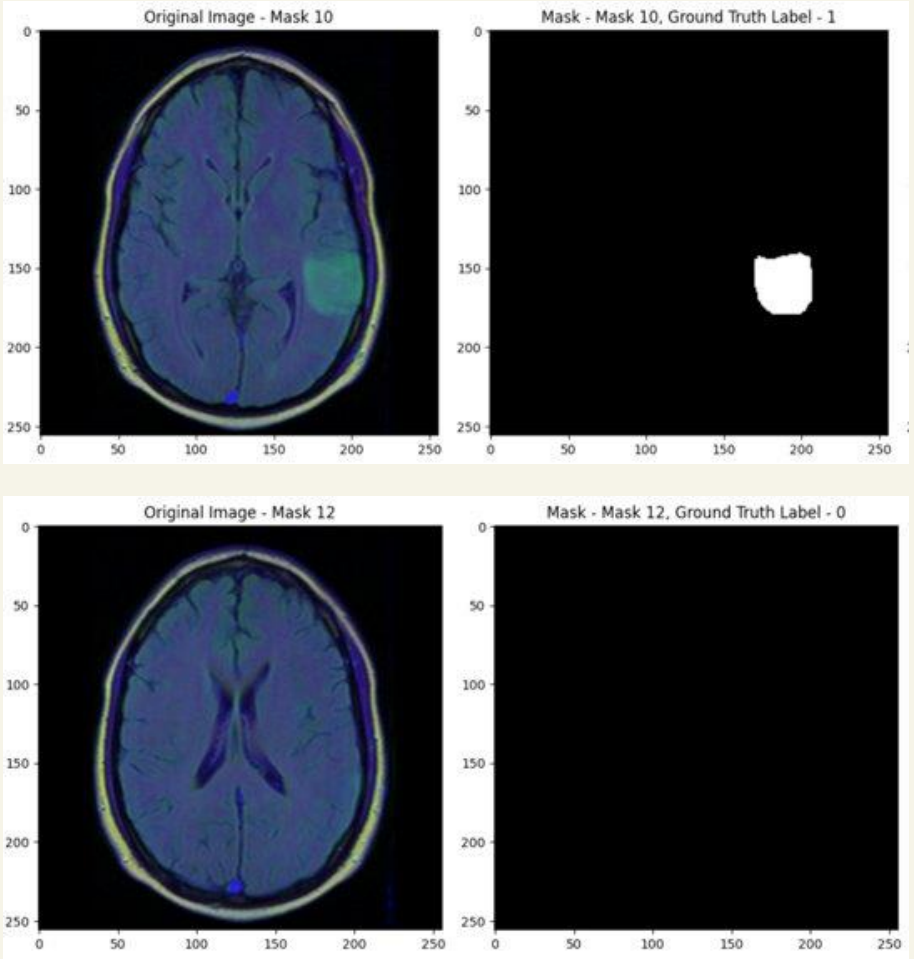
One explanation for peacock



But not the only explanation!

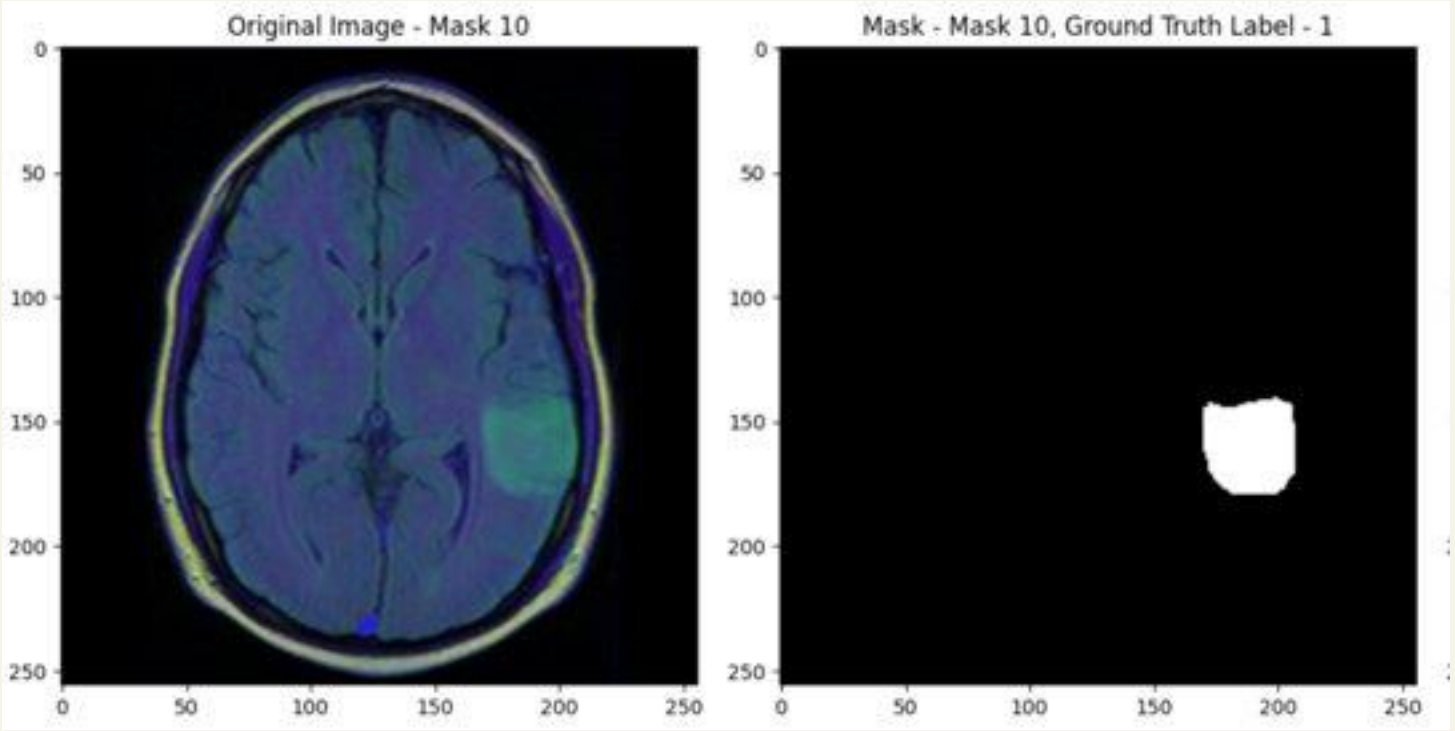
Medical Explanations Dataset

- 110 pre-operative patients with low grade gliomas
- Each patient had between 20 and 88 slices taken, a total of 3929 images
- All images are (256, 256, 3)
- The FLAIR images were annotated with binary masks as 0 (no tumour) or 1 (tumour)

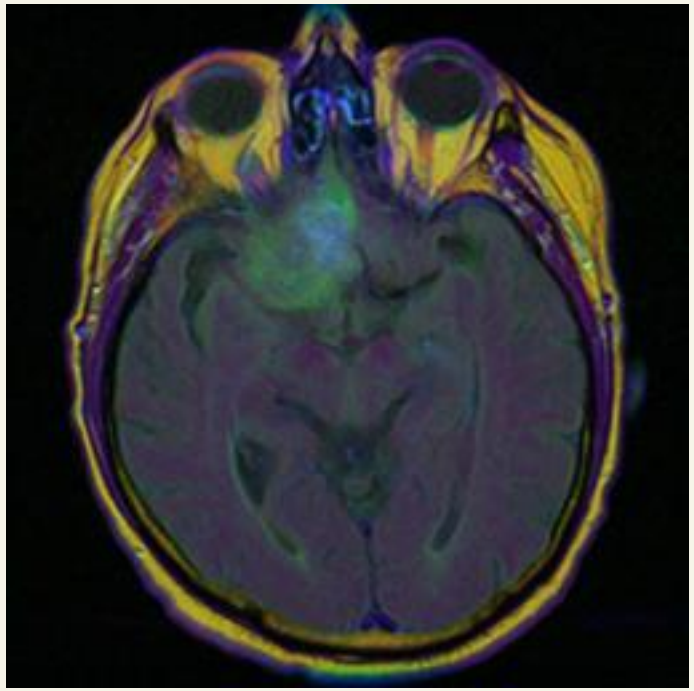


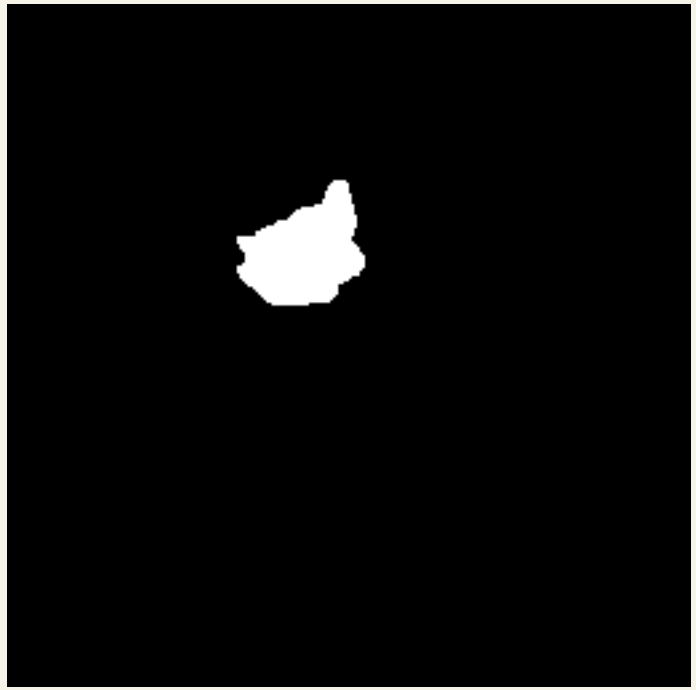
Medical Explanations

- White box tools
  - Grad-CAM
- Black box tools
  - ReX
  - Lime
  - Shap
  - RISE



Let's take a single brain…



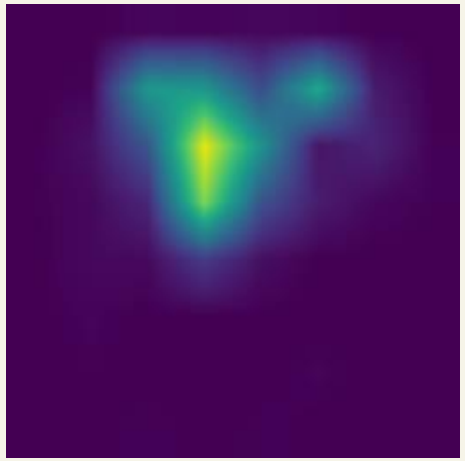


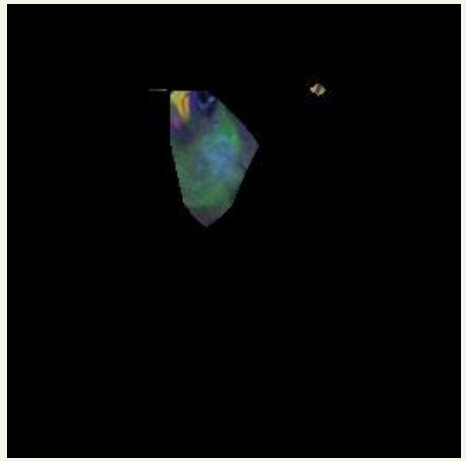
Same brain, same slice, same tumour, same model… same explanations?

What do the different tools do?

Grad-CAM: looking inside

- Looks at penultimate layer to create an attention map

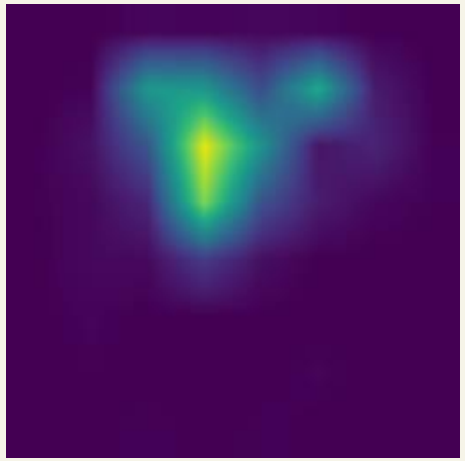


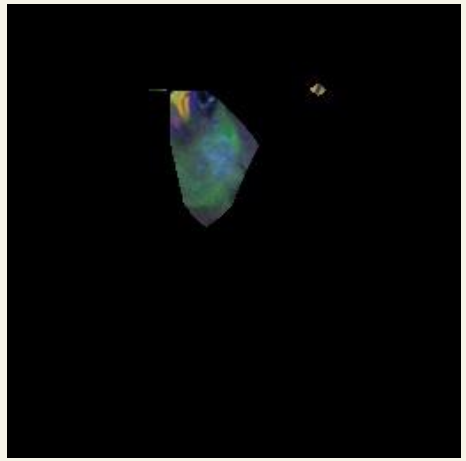


GradCam works well, but…

Grad-CAM: looking inside

- Looks at penultimate layer to create an attention map



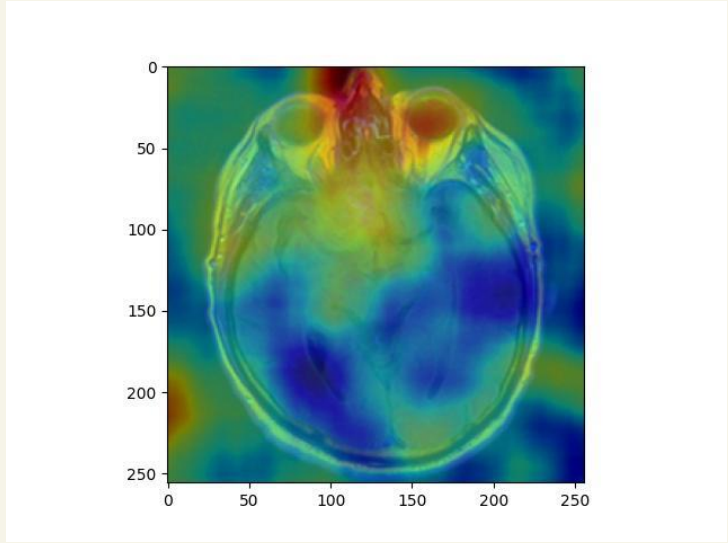


GradCam works well, but…

- Needs access to model
- Fragile
- Gharbani et al. (2019) small perturbations can highlight different pixels

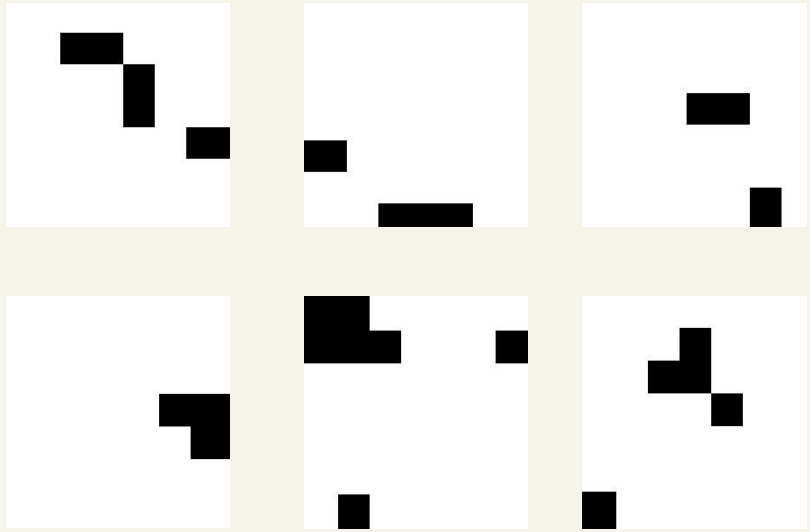
RISE: Naïve Perturbation

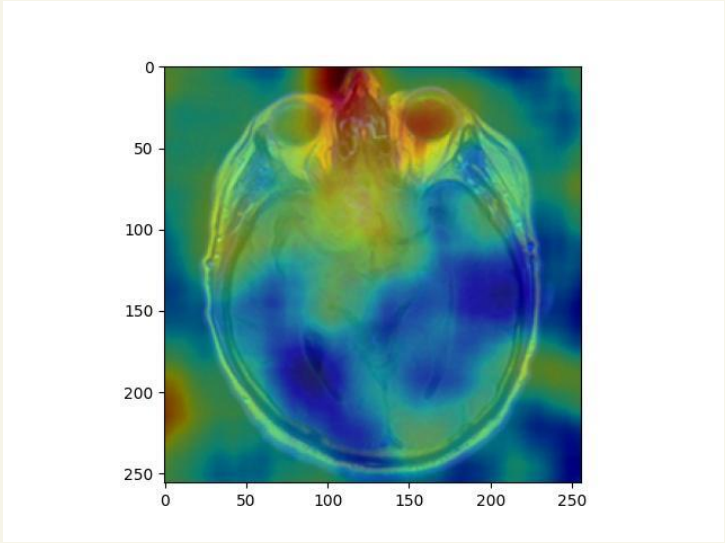
- RISE uses a simple form of SBFL
- We test with 2000 mutants



RISE: Naïve Perturbation

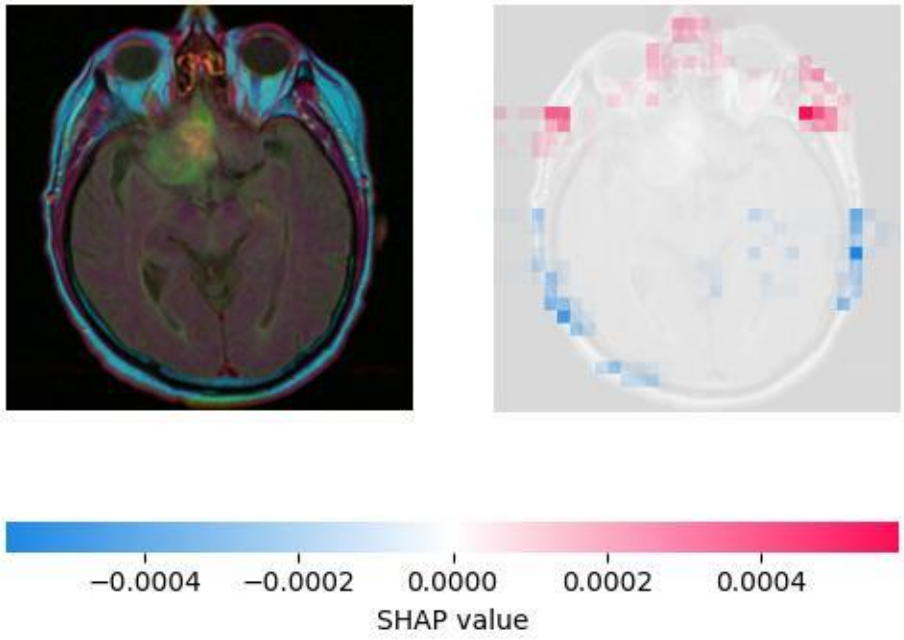
- RISE uses a simple form of SBFL
- We test with 2000 mutants





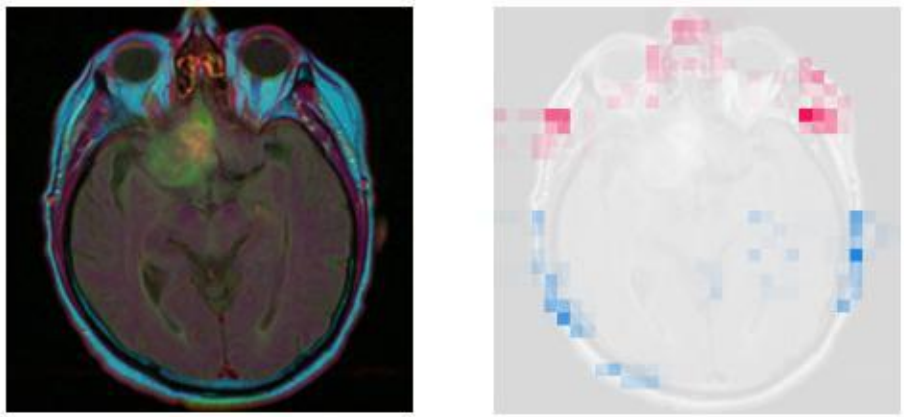
Shap

- Shap uses Shapely values from game theory
- Sample from all possible combinations of features to find average effect of feature to classification



Shap

- Shap uses Shapely values from game theory
- Sample from all possible combinations of features to find average effect of feature to classification



Why doesn't shap work?

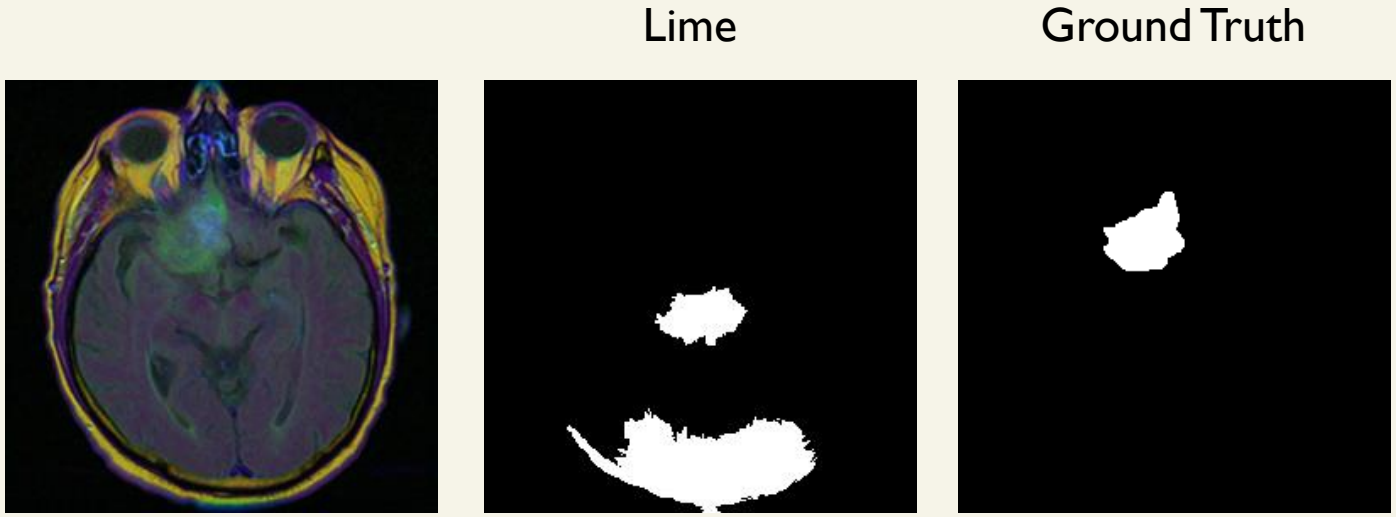
Lime

- Lime generates mutants and learns a locally explainable model
- It relies on prior image segmentation

Lime

Lime

- Lime generates mutants and learns a locally explainable model
- It relies on prior image segmentation



Why doesn't lime work?

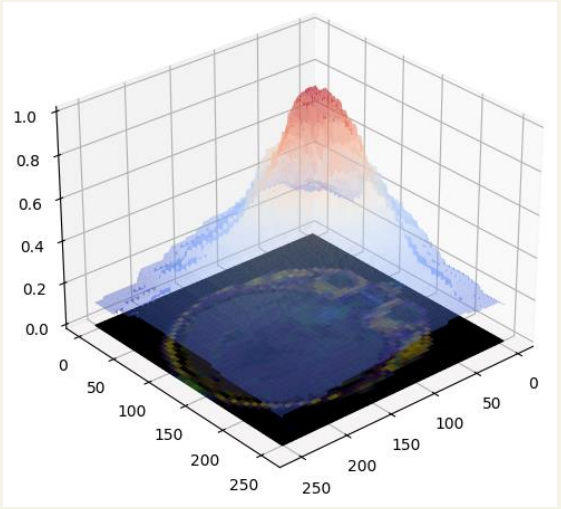
What is going on?

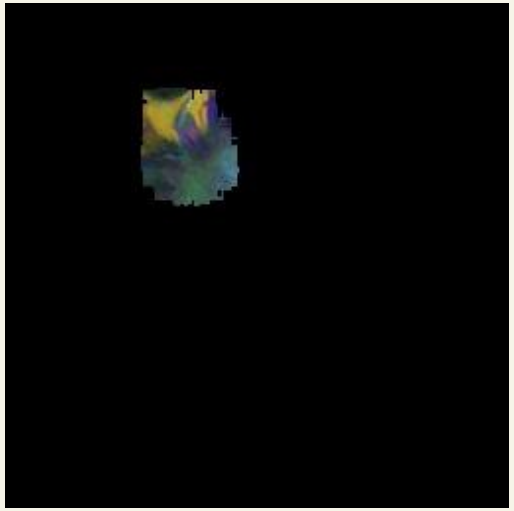
KING'S College LONDON

- These tools work on general images (imagenet)
- Model accepts too much
- Poor mutant generation leads to poor performance
  - Are the mutants sufficiently diverse?
  - Do the mutants make sense? Do they need to?
- Grad-CAM does not require mutant generation

Default ReX

- Partial tumour discovery
- No false positives
- Why is it not perfectly positioned over the tumour?
- Sensitive to configuration





We are closer than the other black box tools, but we still need improvement…