Colin Jacobs, Swinburne University of Technology
Unimelb
26 August 2020

Probing Neural Networks in Astronomy

ASTRO 3D

ARC CENTRE OF EXCELLENCE FOR ALL SKY ASTROPHYSICS IN 3D

Deep learning - and its failures

More and more applications in **science** (and real life!)

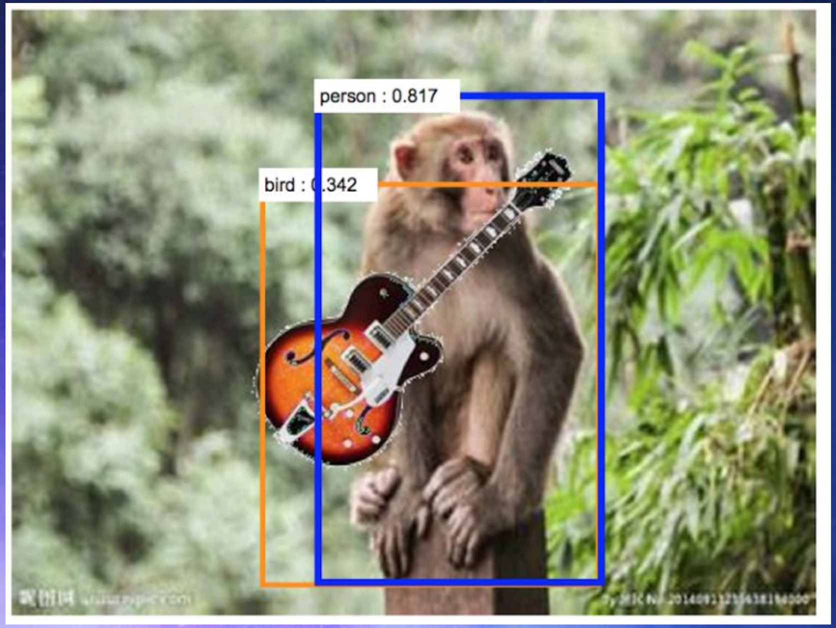
How can we find its weaknesses and know how it might fail?

- Can only know how well it will do on the data we already have, may not be real world
- More sensitive to changes that would not fool a human
- We might be blind to biases in the training set

These issues have consequences.

For **science**:

- Hard to understand biases
- Hasd to quantify errors



Source: Wang 2017

AI in science and society

AI coming soon to your life:

Hiring and firing

Financial access

University admission

School rankings

Legal system

Advertising

"The best minds of my generation are thinking about how to make people click ads. (That sucks.)"

- Jeff Hammerbacher

Fairness, transparency, accountability

Tech Policy / AI Ethics

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** Jan 21, 2019

REUTERS Business Markets World Politics TV Mo

BUSINESS NEWS OCTOBER 10, 2018 / 2:12 PM / 10 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

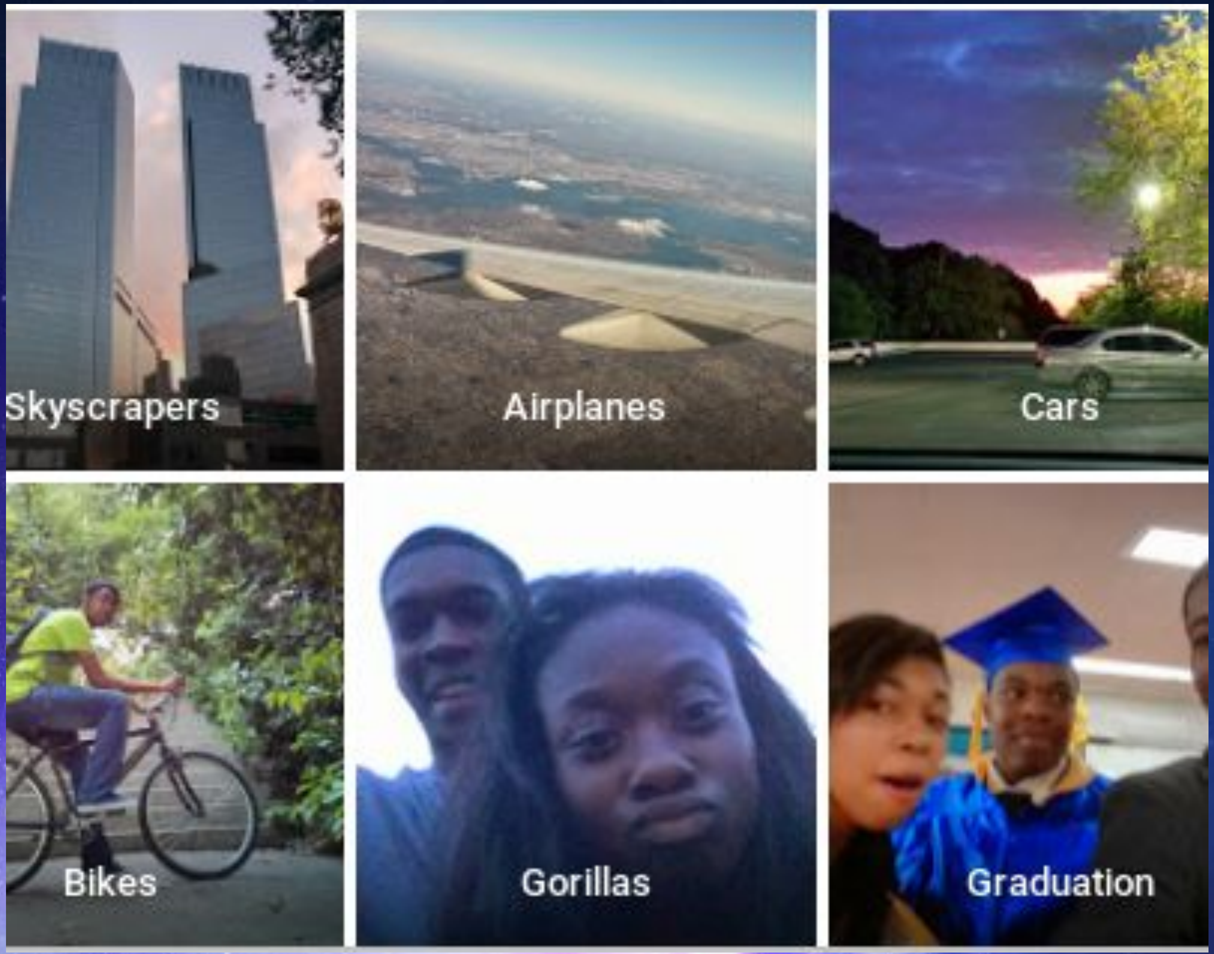
Bias in AI

Challenges:

Framing the problem

Training data biased

Lack of social context



AI ethics

Artificial intelligence (AI)

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

Alex Hern

@alexhern

Thu 14 Feb 2019 17.00 GMT

6,621 572

Mathematics

Maths and tech specialists need Hippocratic oath, says academic

Exclusive: Hannah Fry says ethical pledge needed in tech fields that will shape future

▲ Hannah Fry: 'The future doesn't just happen. We are building it all the time.' Photograph: Paul Wilkinson

Ian Sample *Science editor*

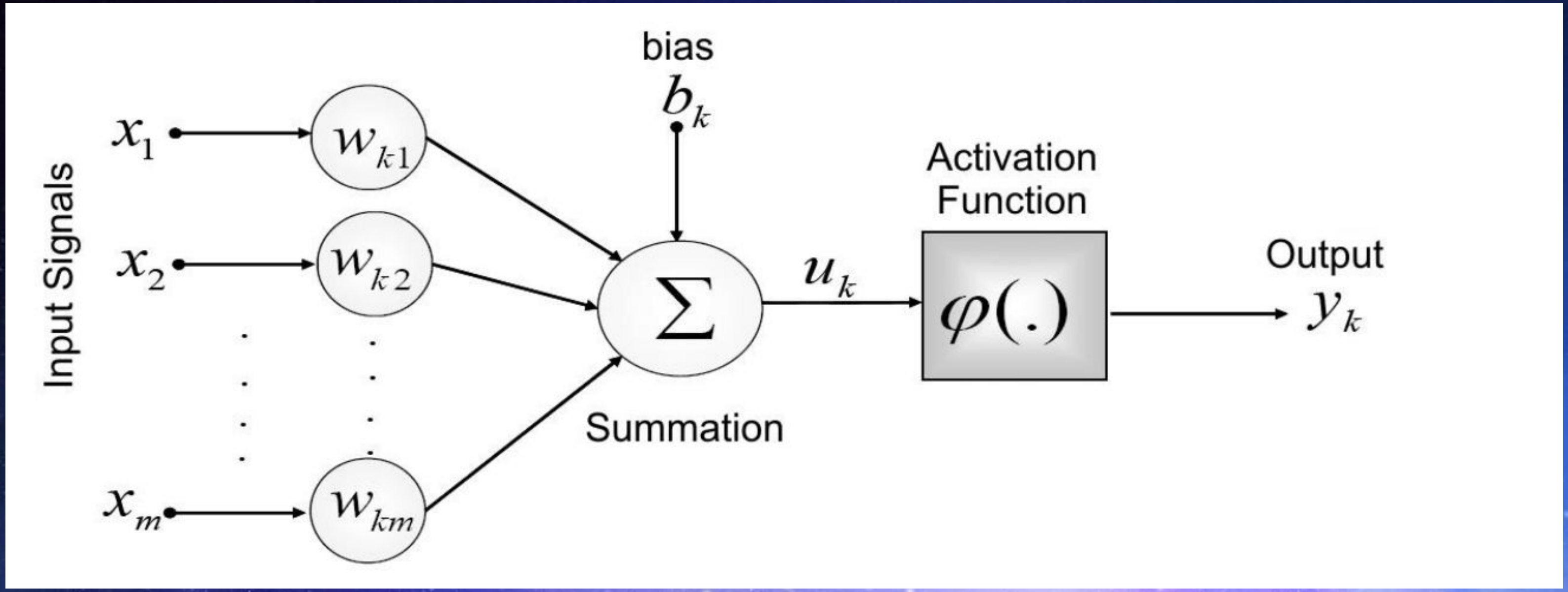
Interpreting neural networks

- Interpreting a trained ML model is vital to validate that the representation has accurately captured the general features of the data and not overfit.
- High performance is mediated by generalisability.
- An important step in ensuring the reproducibility of results.
- Cars, medicine, courts, finance… urgent!

Need something ***Explanatory*** and ***Interpretable***

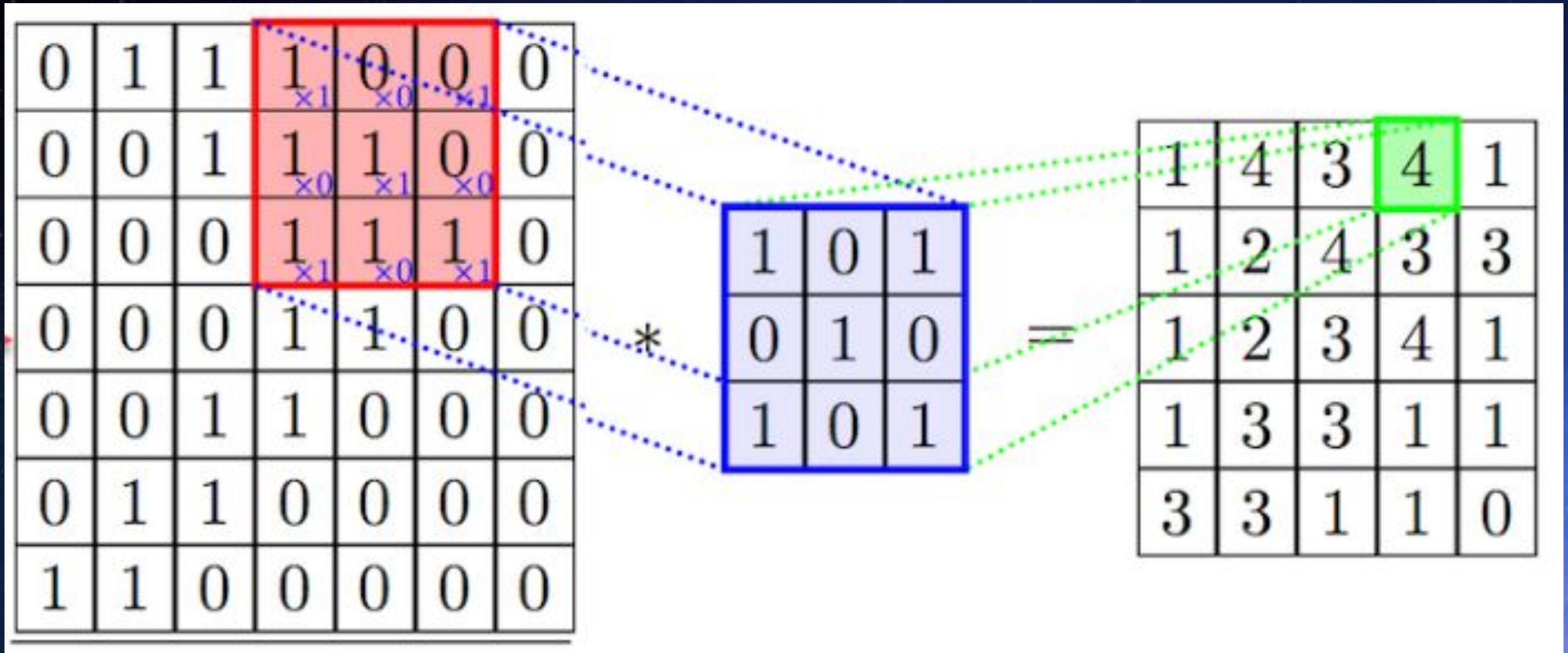
SEE: Montavon, Samek and Muller (2018) and Lipton (2016)

Neural networks - simple but complex

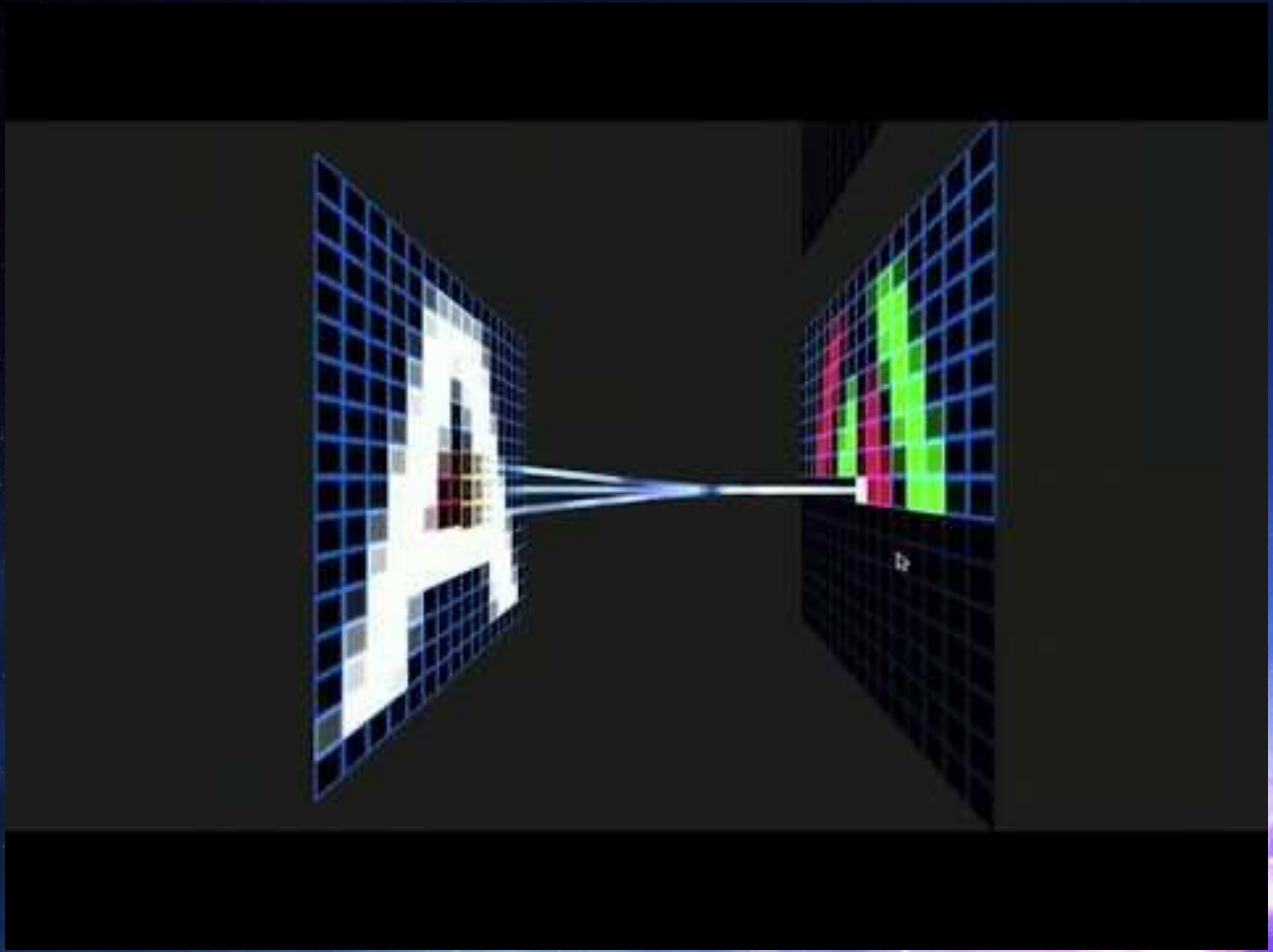


Source: Veronez 2011

Convolutional neural networks - less simple but not too complex



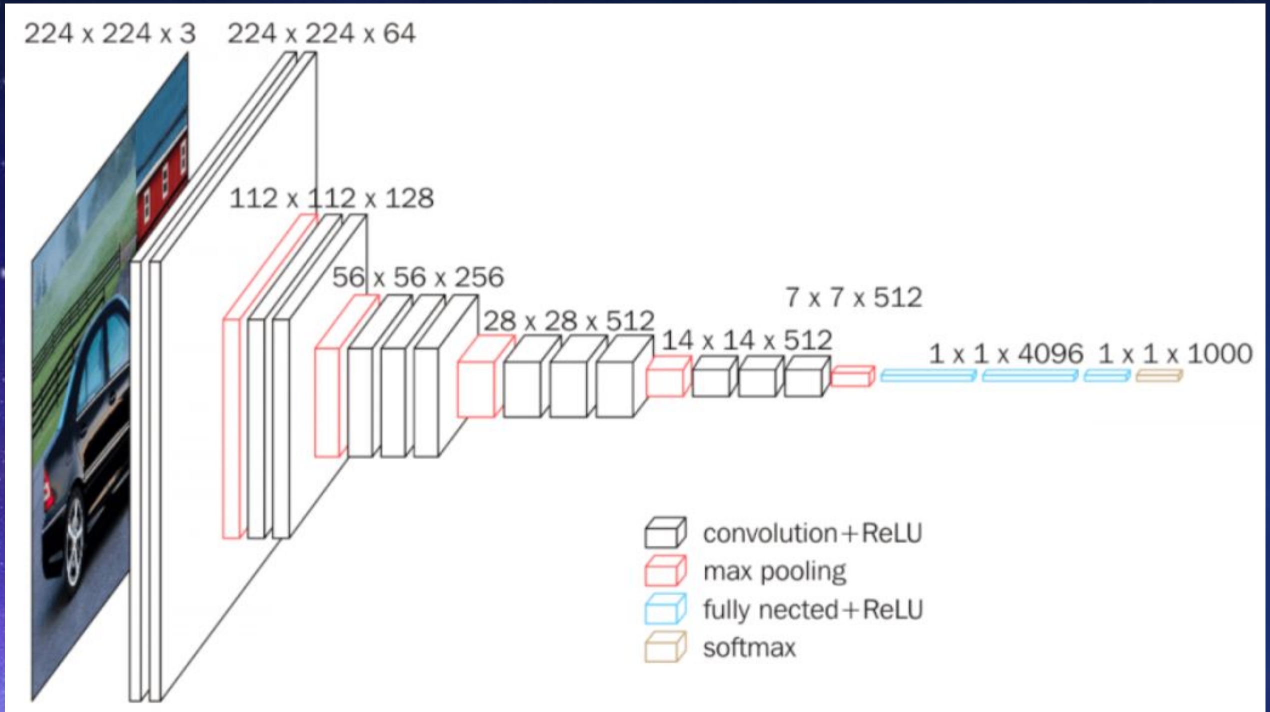
Source: Micheal Lanham 2018



What's going on?

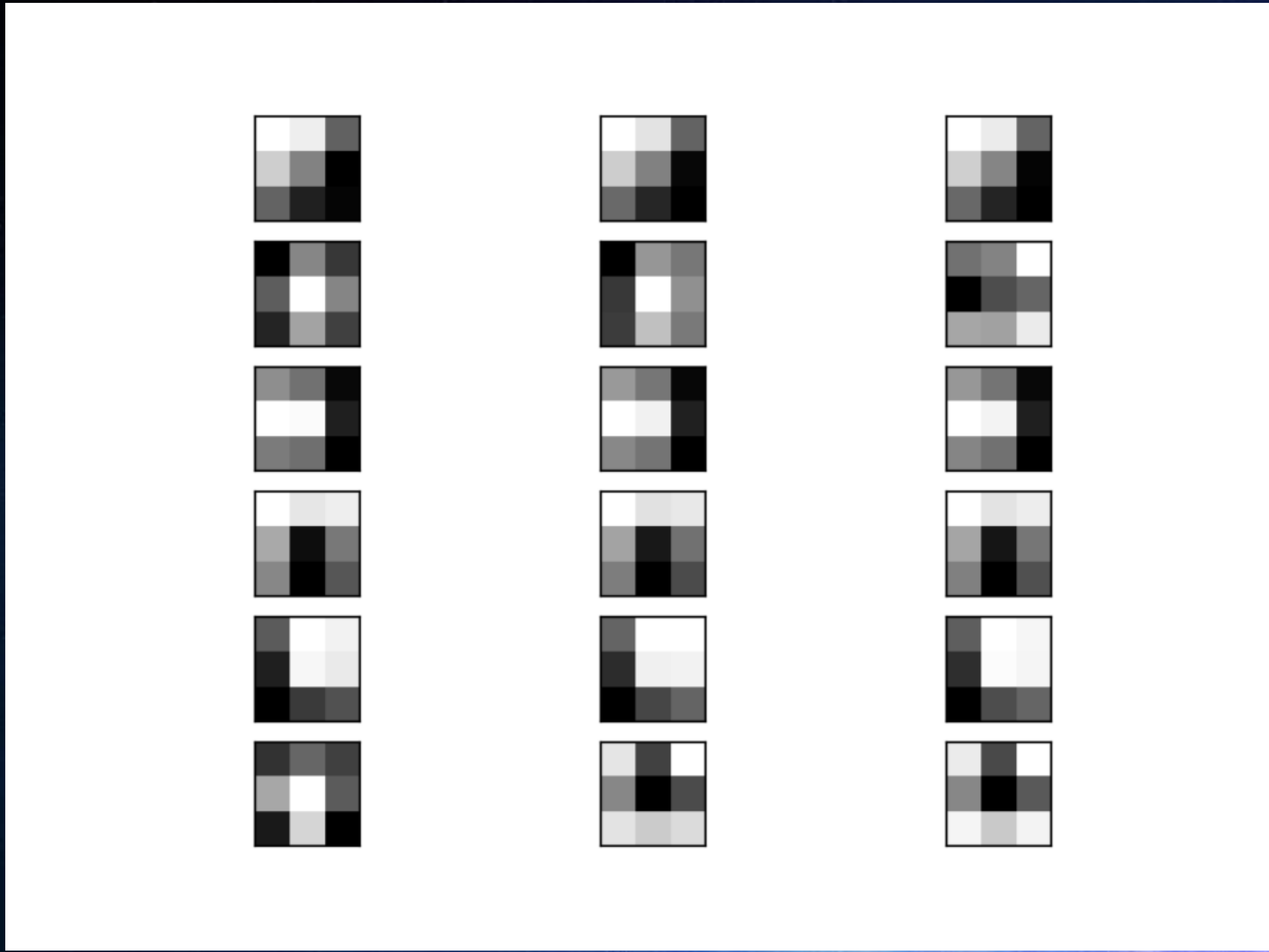
Challenges with ANNs:

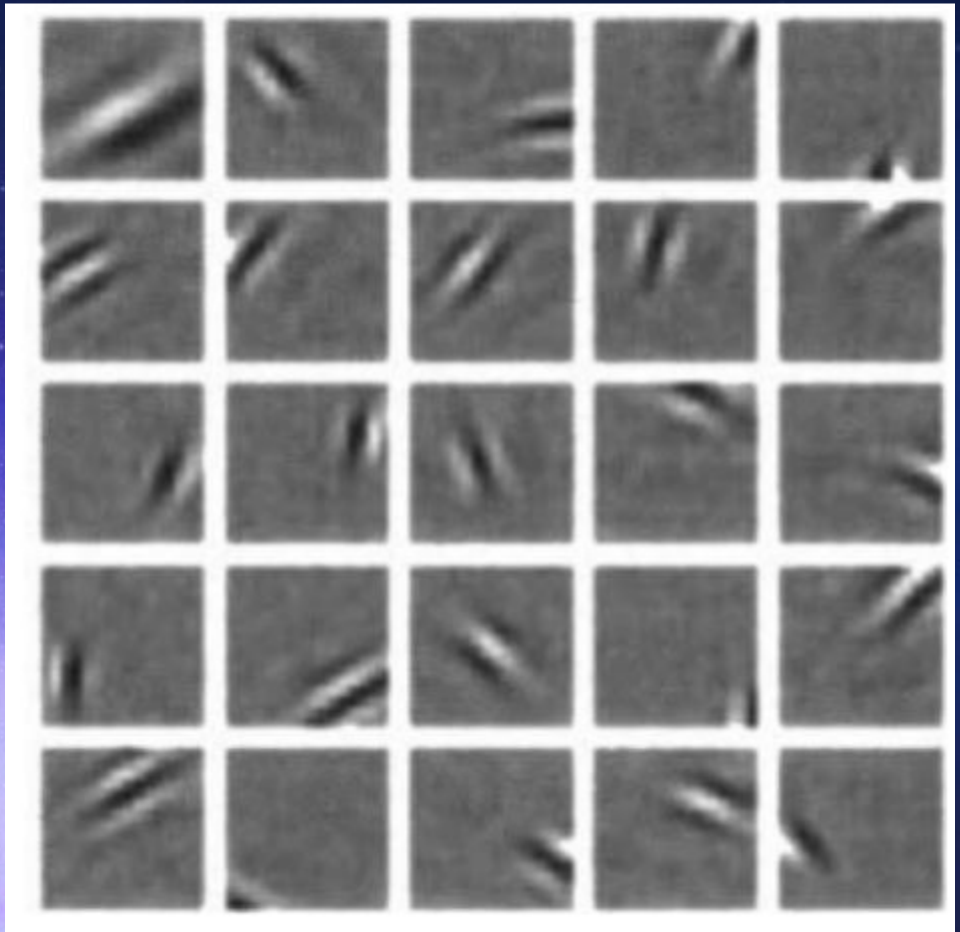
- Dimensionality of inputs enormous
- Trainable weights ~10^6 - 10^9
- Hundreds of feature maps
- Highly abstract and non-linear
- Distribution of inputs, and gaps, hard to comprehend



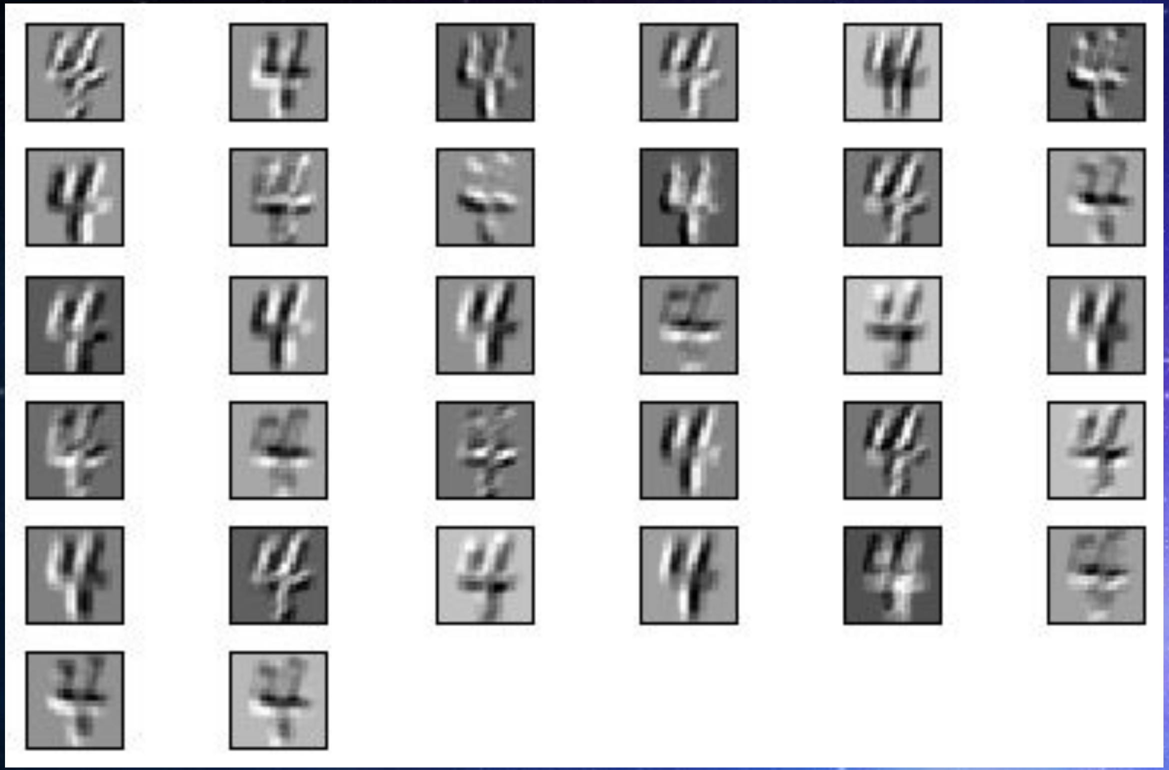
Simonyan and Zisserman (2014)

First attempt: Convolutional kernels





Feature maps



Input optimisation

Take a trained model and train the ***inputs*** to maximise the activation for a particular class (maximise the output of a particular neuron).

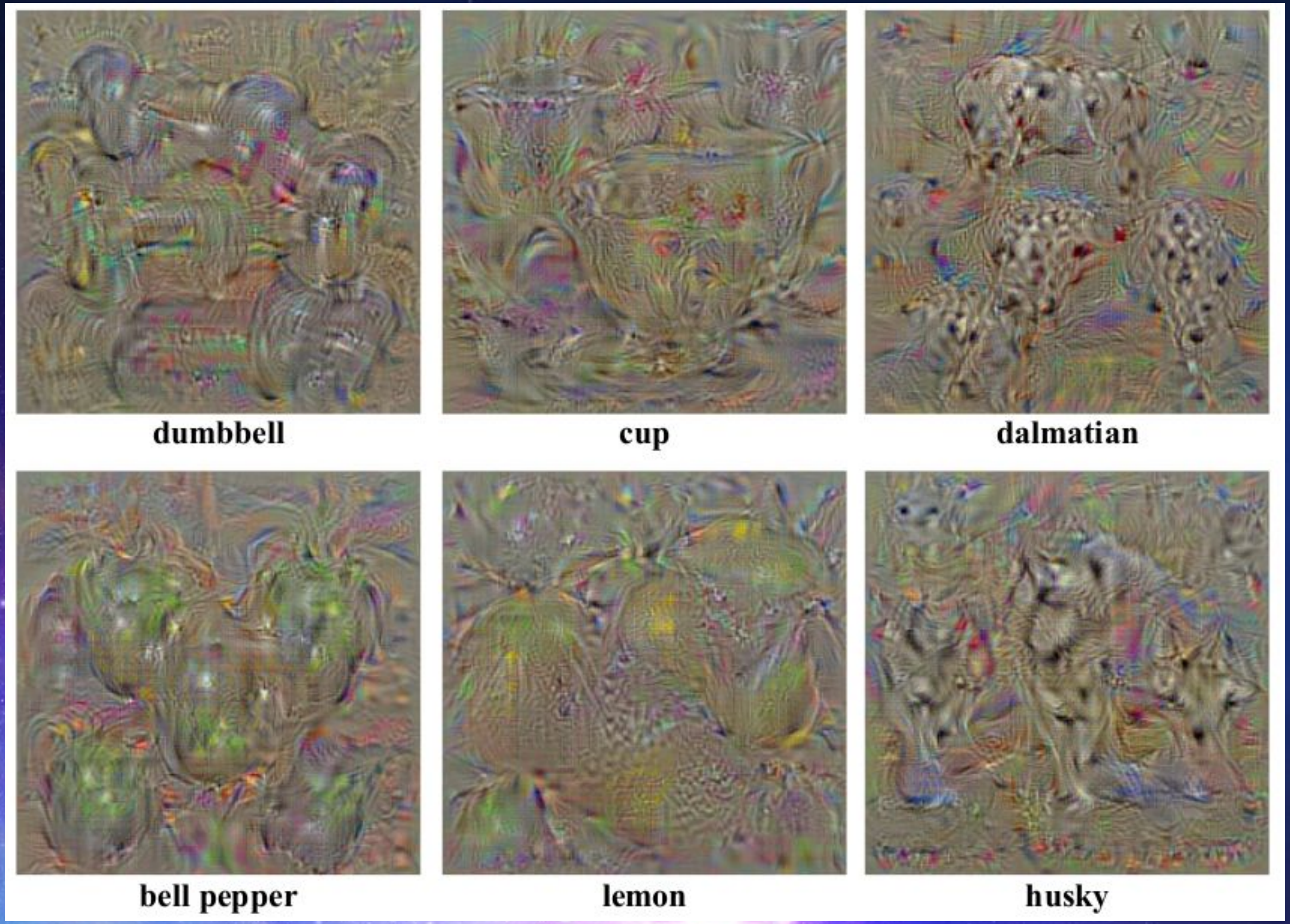


Image: Varma and Das 2018

Deep Dream

Pouff: https://www.youtube.com/watch?v=DgPaCWJL7XI

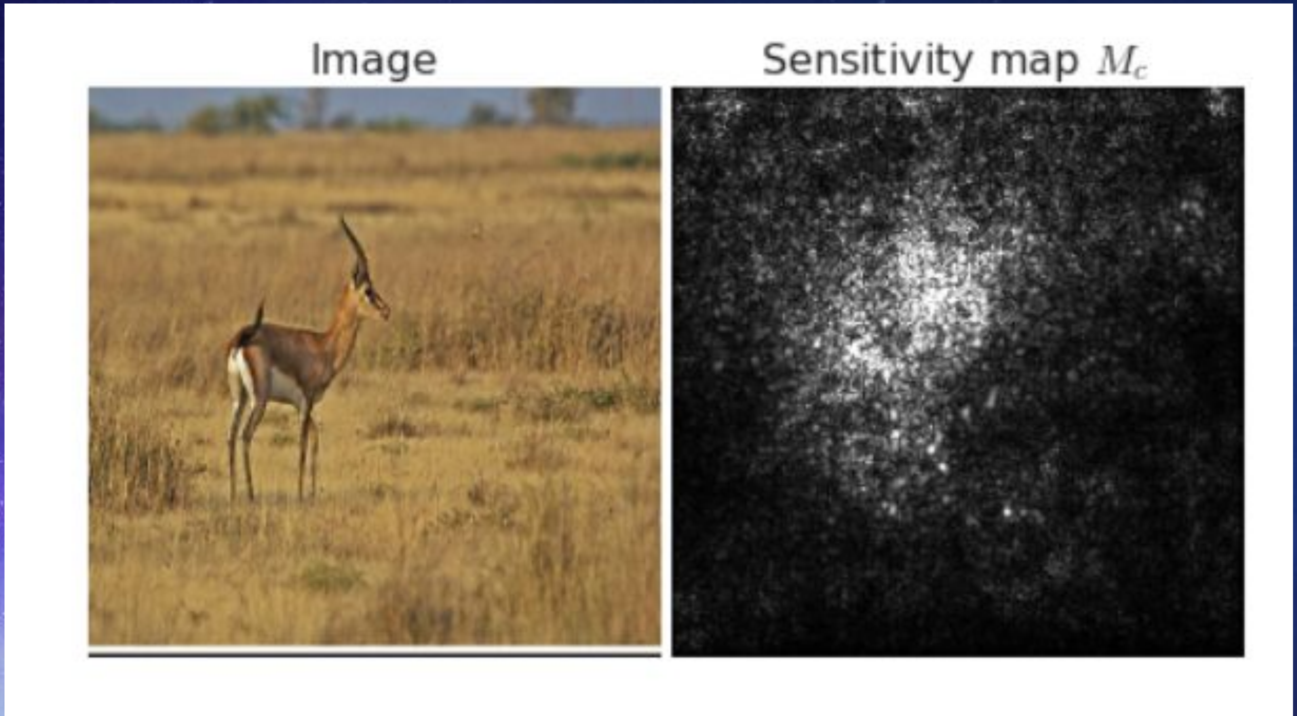
ASTRO 3D

Occlusion sensitivity

Calculate the sensitivity to a particular pixel: i.e.

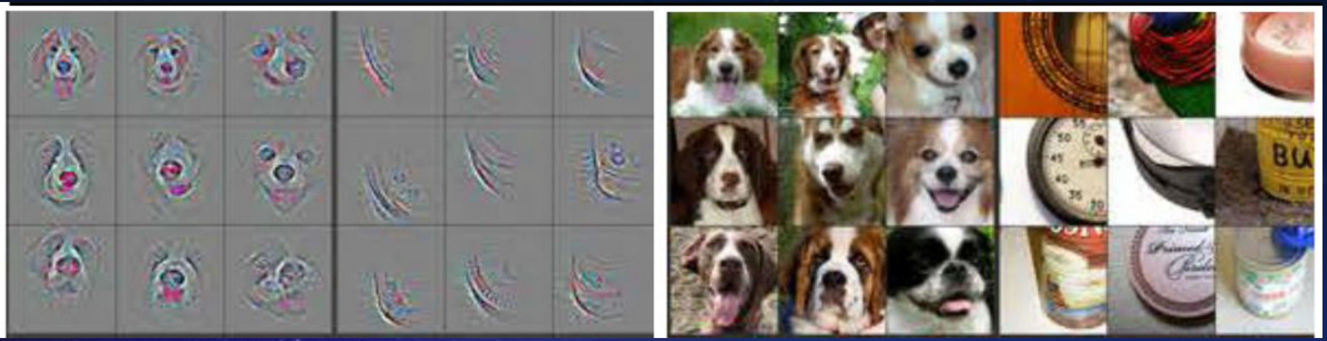
d neuron/d pixel_i

Very noisy!



Smilkov et al 2017

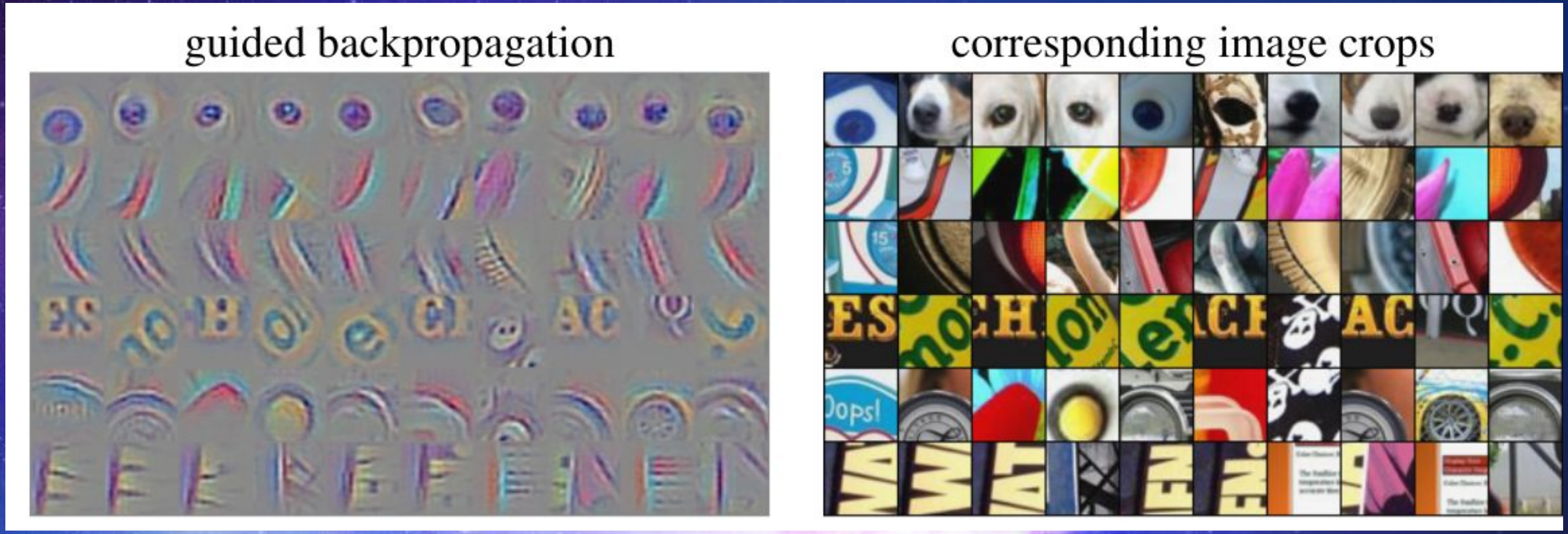
Other attempts



Deconvnet: Zeiler and Fergus 2014

Deconvolution: Zeiler and Fergus 2014

Guided backprop: Gradient of a particular neuron, through a ReLU. (Springenberg et al 2015).

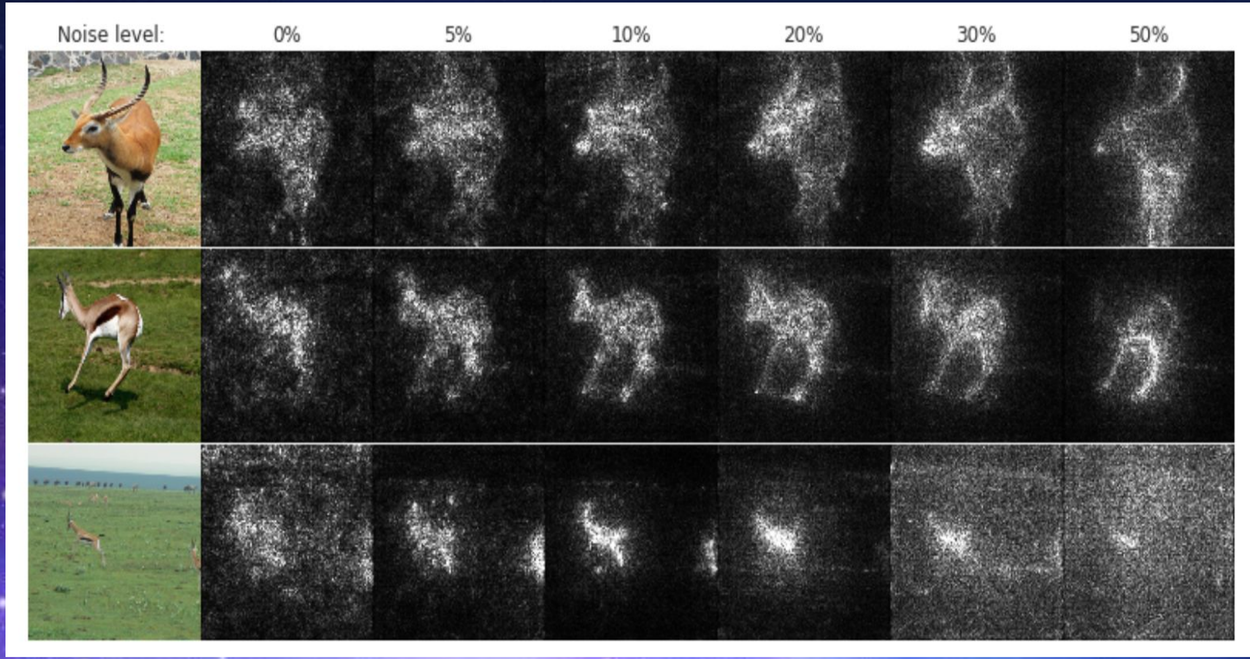


Springenberg et al 2015

Occlusion sensitivity

Smoothgrad: Smilkov 2017

Adding noise to get more signal - sample an image many times (with added noise) and display the mean sensitivity map



Smilkov et al 2017

Saliency mapping

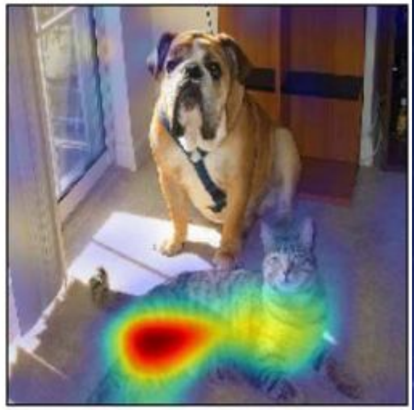
E.g. Grad-CAM (Selvaraju 2017)

Take activations at last convolutional layer, determine importance to score

Pool over feature maps -> importance

Sum maps weighted by importance

Upscale and project back onto input image.



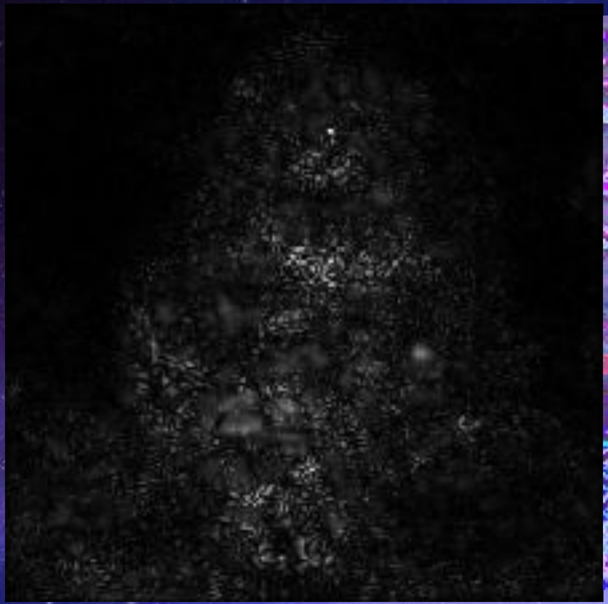
(c) Grad-CAM 'Cat'

(i) Grad-CAM 'Dog'

Selvaraju et al 2017

Saliency mapping: State of the art

Input



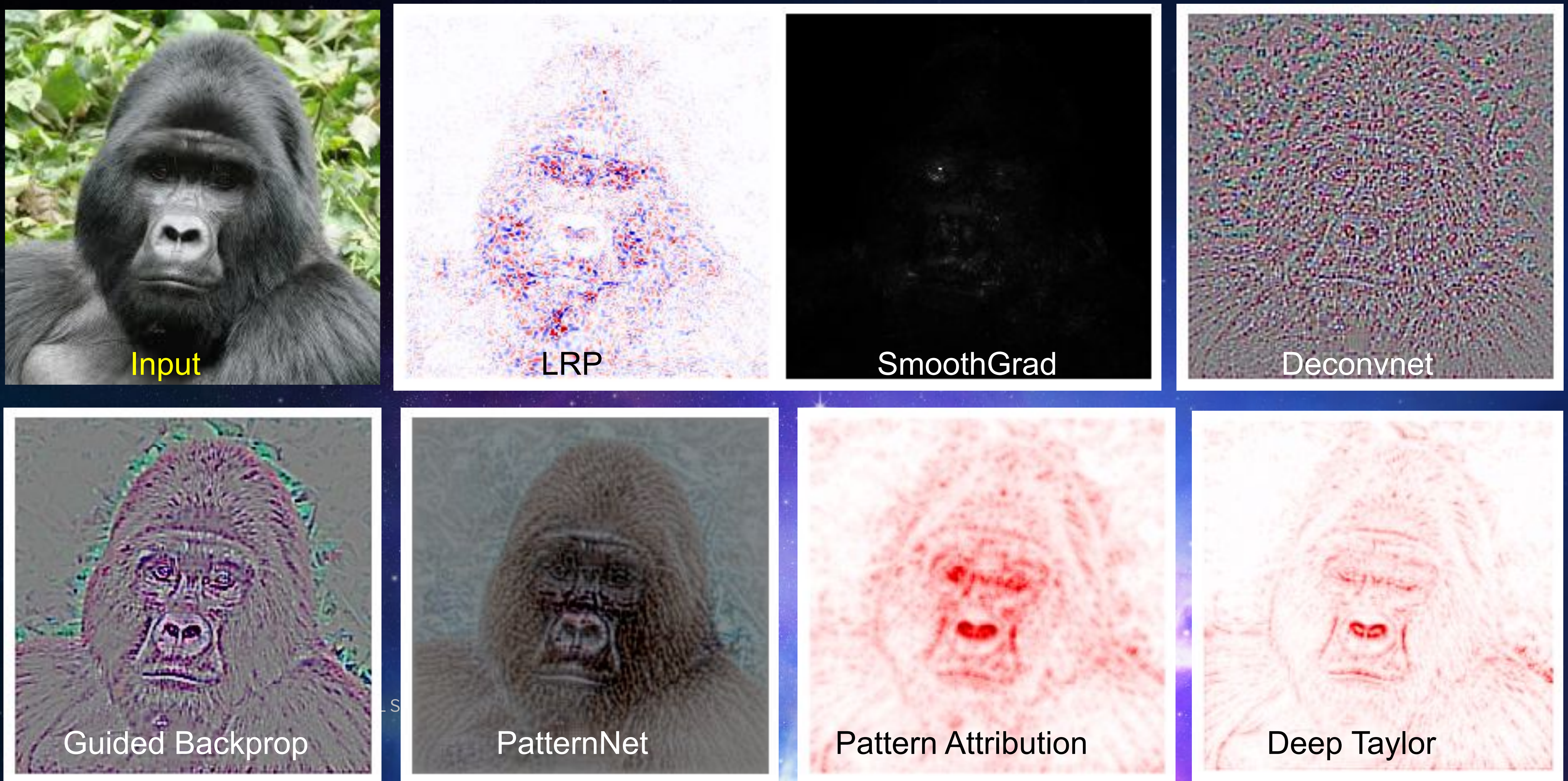
Integrated Gradients

Occlusion

Grad-CAM

ASTRO 3D

Saliency mapping



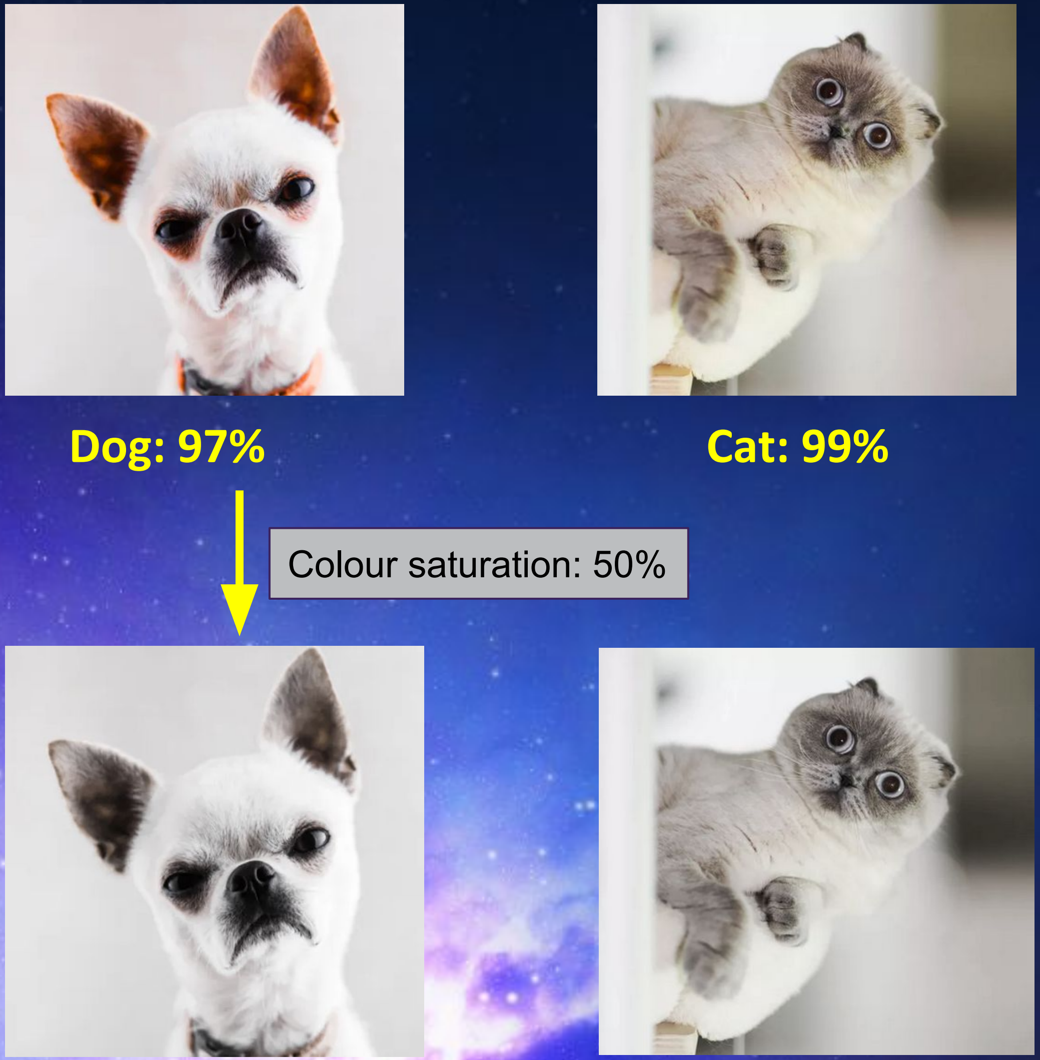
Sensitivity Analysis

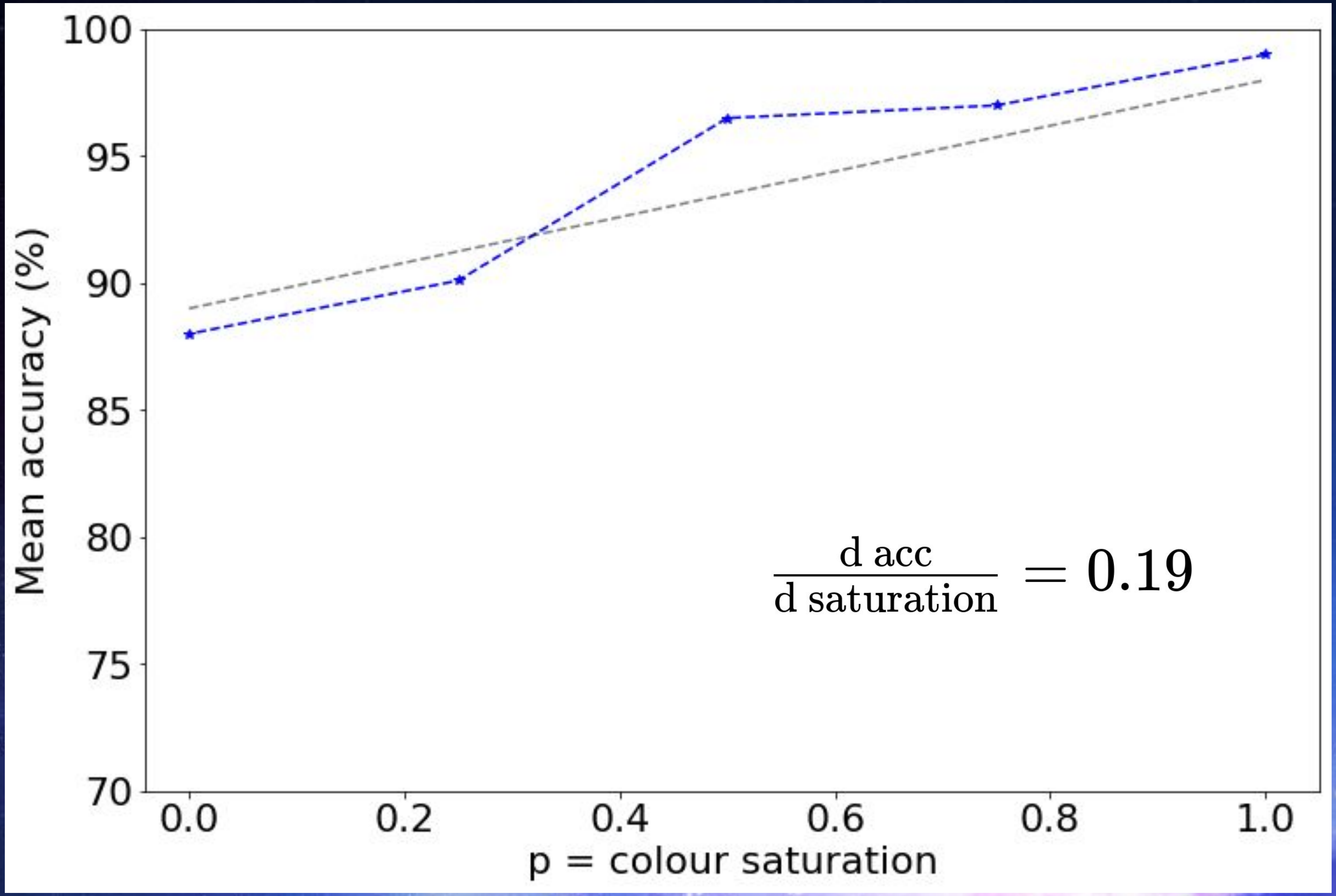
How *sensitive* is the network to:

- A transformation of the data?
- Some inherent property of the data?

Can we use this to identify weaknesses?

Consider the correct-class probability as the key metric; could use another key measure.




100
95
90
85
80
75
70
Mean accuracy (%)
0.0
0.2
0.4
0.6
0.8
1.0
p = colour saturation
$\frac{\mathrm{d\ acc}}{\mathrm{d\ saturation}} = 0.19$

Sensie

Automates sensitivity analysis -

if you know what questions to ask!

Available on Github

The Journal of Open Source Software

Sensie: Probing the sensitivity of neural networks

Colin Jacobs[1]

1 Center for Astrophysics and Supercomputing, Swinburne University of Technology

DOI: 10.21105/joss.02180

Software

- Review
- Repository
- Archive

Editor: George K. Thiruvathukal

Reviewers:

Introduction

Deep neural networks (DNNs) are finding increasing application across a wide variety of fields, including in industry and scientific research. Although DNNs are able to successfully tackle data problems that proved intractable to other methods, for instance in computer vision, they suffer from a lack of interpretability. Some well-known methods for visualising and interpreting the outputs of DNNs include directly inspecting the learned features of a

Jacobs 2020

Sensie: Use case (MNIST)

MNIST - sensitivity to input orientation

In [37]:
```
mnist = tf.keras.datasets.mnist

(X_train, y_train), (X_test, y_test) = mnist.load_data()
X_train, X_test = X_train / 255.0, X_test / 255.0
X_train = X_train[:, :, :, np.newaxis]
X_test = X_test[:, :, :, np.newaxis]
```

In [38]:
```
def make_model_mnist():
    input_shape = (28, 28, 1)
    model = keras.models.Sequential()
    model.add(layers.Conv2D(32, kernel_size=(3, 3),
                     activation='relu',
                     input_shape=input_shape))
    model.add(layers.Conv2D(64, (3, 3), activation='relu'))
    model.add(layers.MaxPooling2D(pool_size=(2, 2)))
    model.add(layers.Dropout(0.25))
    model.add(layers.Flatten())
    model.add(layers.Dense(128, activation='relu'))
    model.add(layers.Dropout(0.5))
    model.add(layers.Dense(10, activation='softmax'))

    model.compile(optimizer='adam',
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])
    return model
```

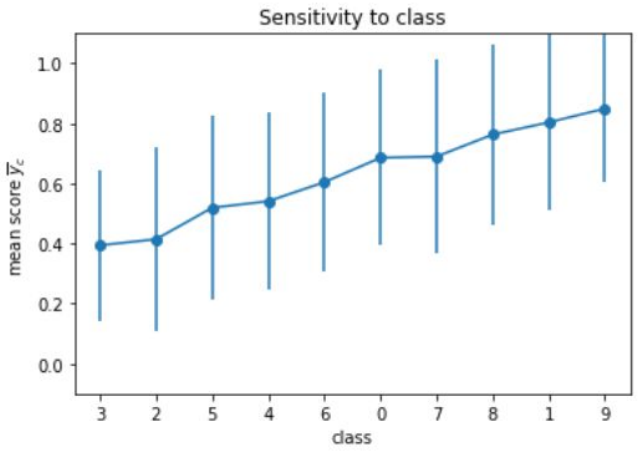
Sensie: Use case (CIFAR10)

```
In [10]: (_, y_train_c), (_, y_test_c) = cifar10.load_data()
         y_test = y_test_c[:, 0]
```

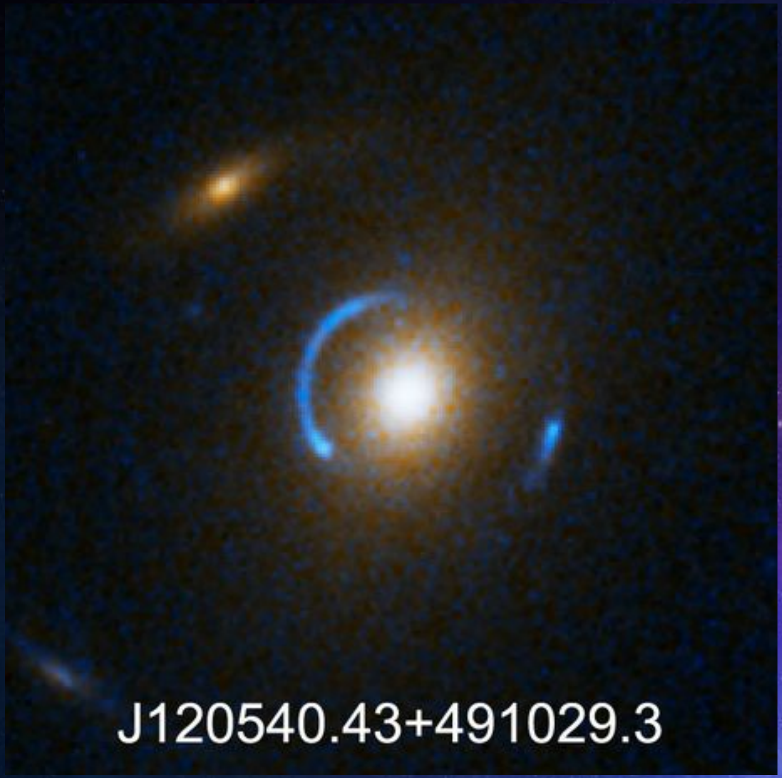
Are some classes more difficult for the network than others? This method is a quick way to visualize the confusion. This may be relevant if class and other tested properties are highly correlated.

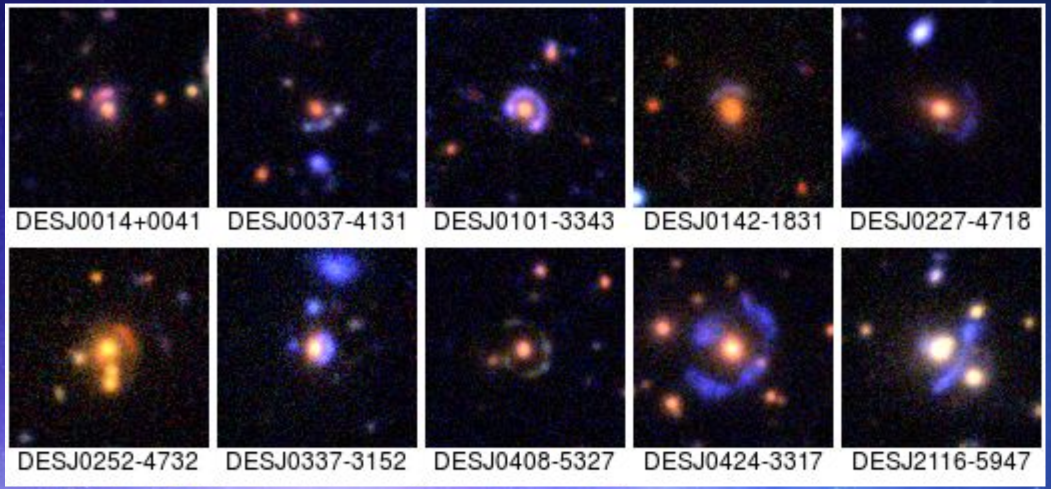
```
In [11]: cifar_probe = sensie.Probe(model)
         class_test = cifar_probe.test_class_sensitivity(X_test, y_test, plot=True)
```

```
[########################################] 100% (1/1)     class
```



Querying an AI astronomer

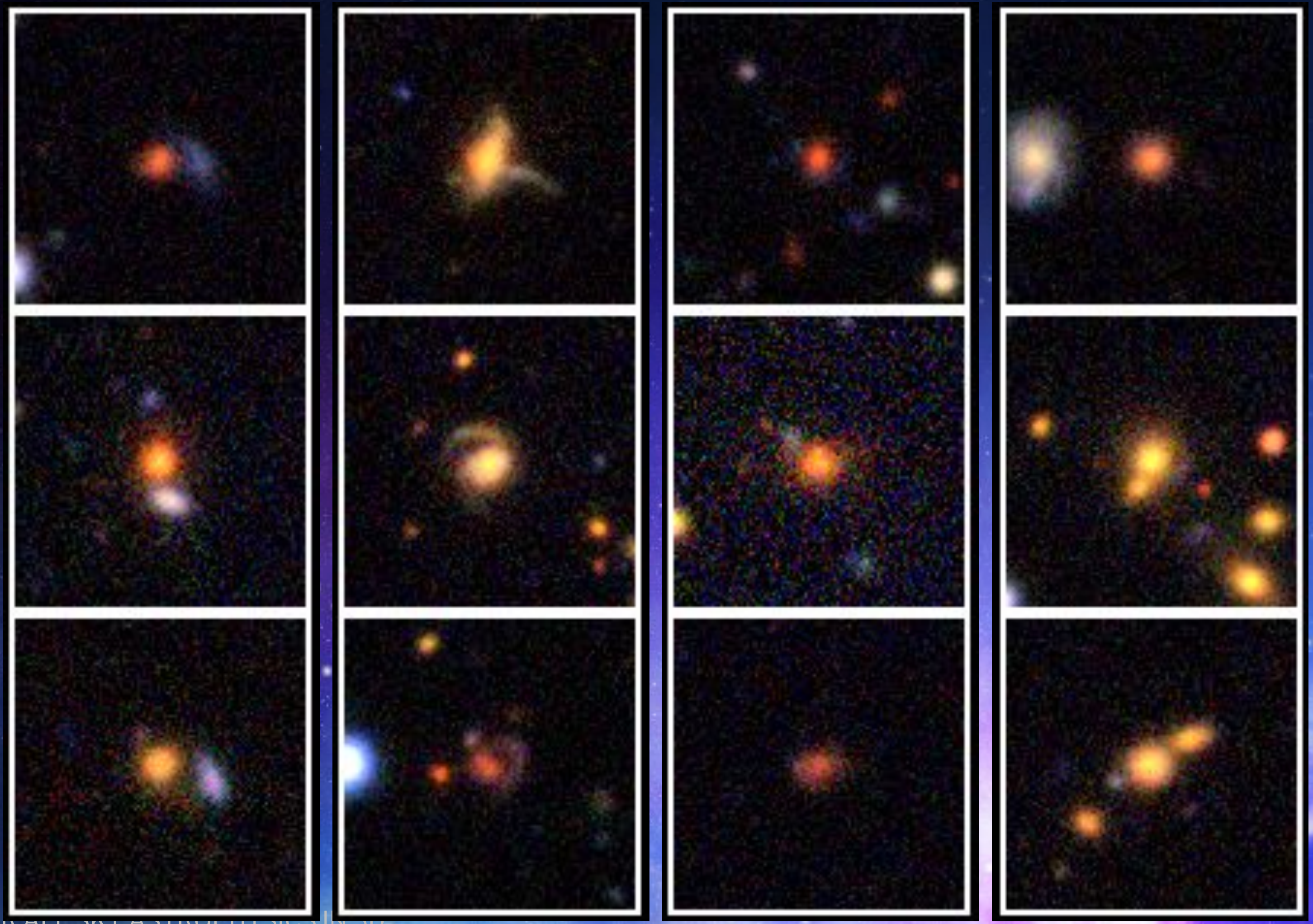




Jacobs+ 2019b

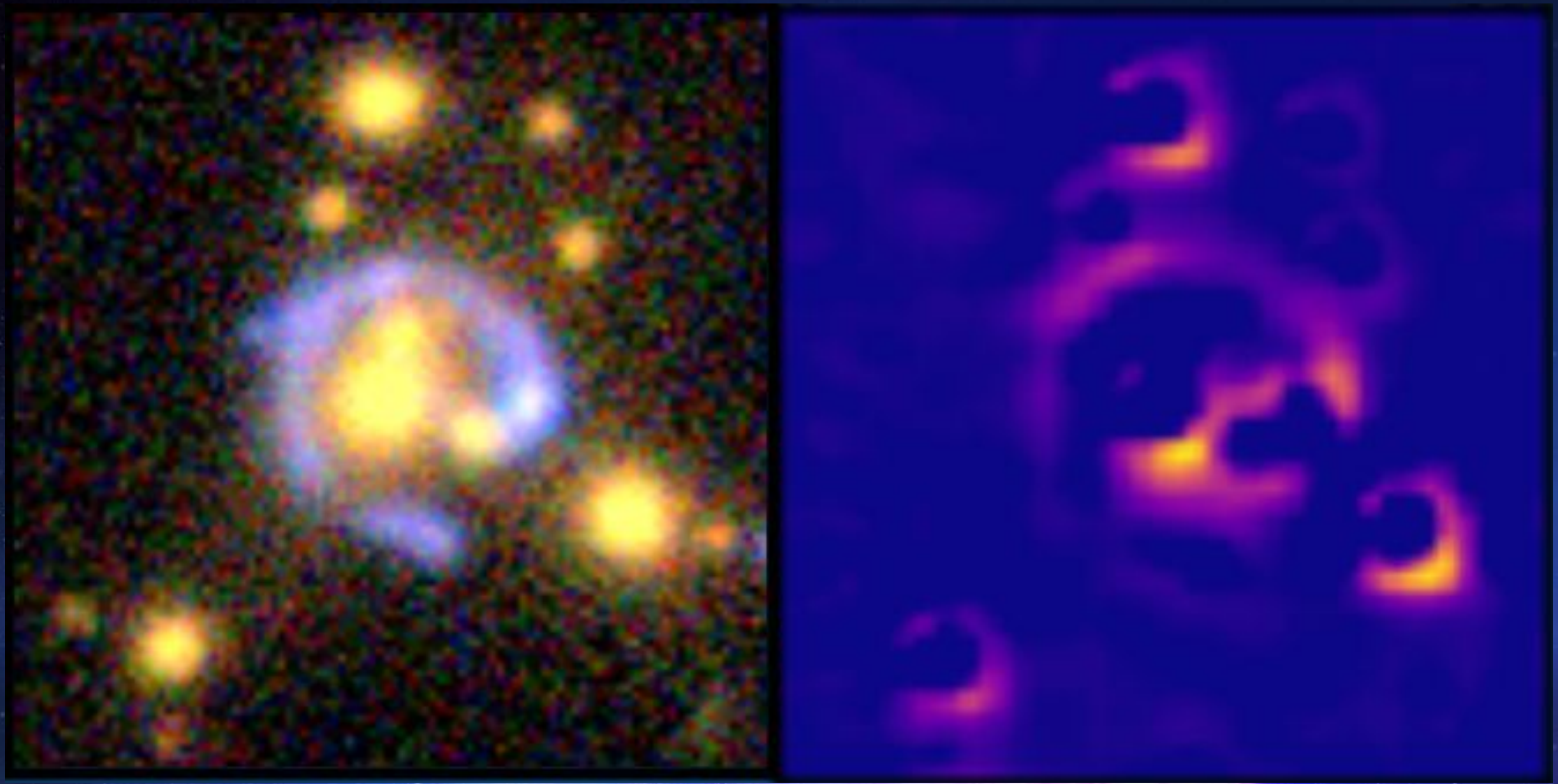
Querying an AI astronomer

False positives - Why?

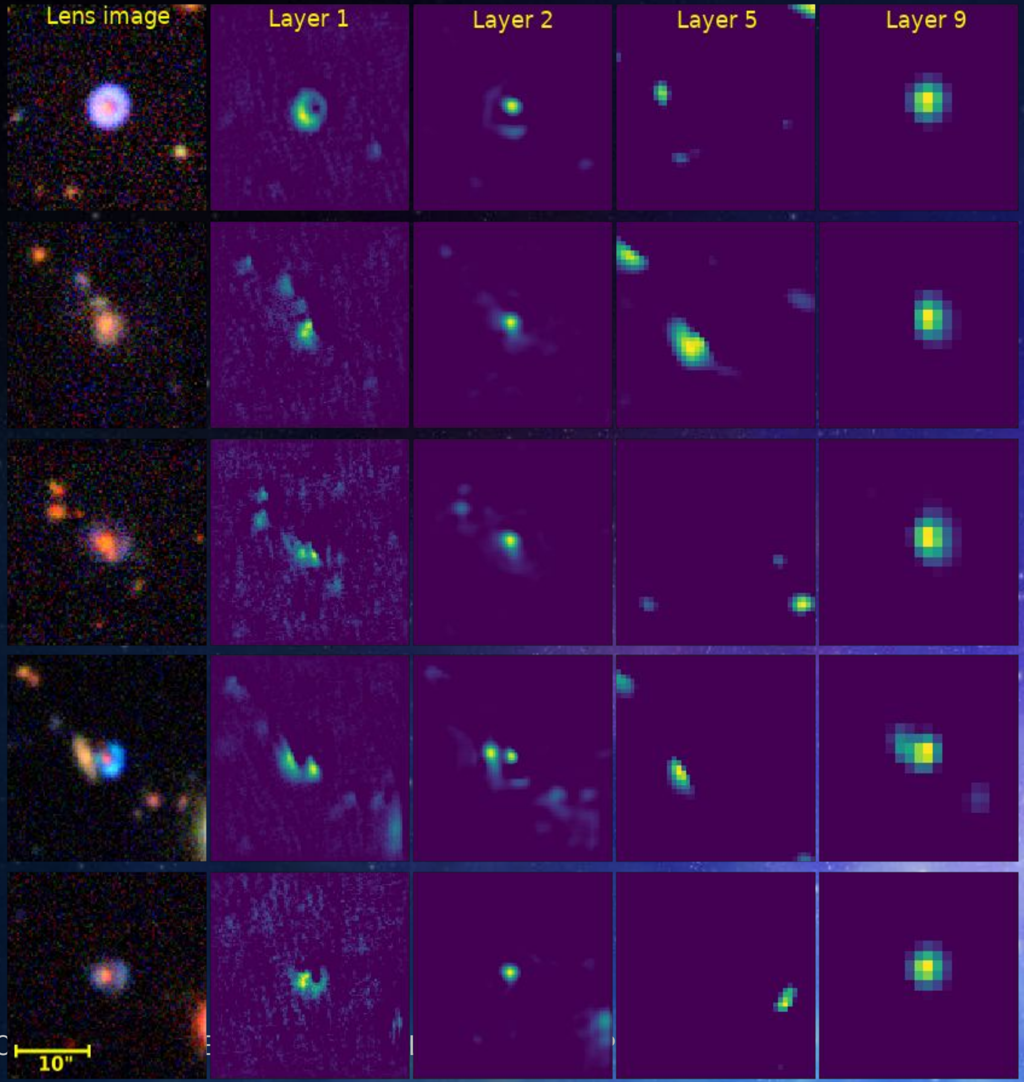


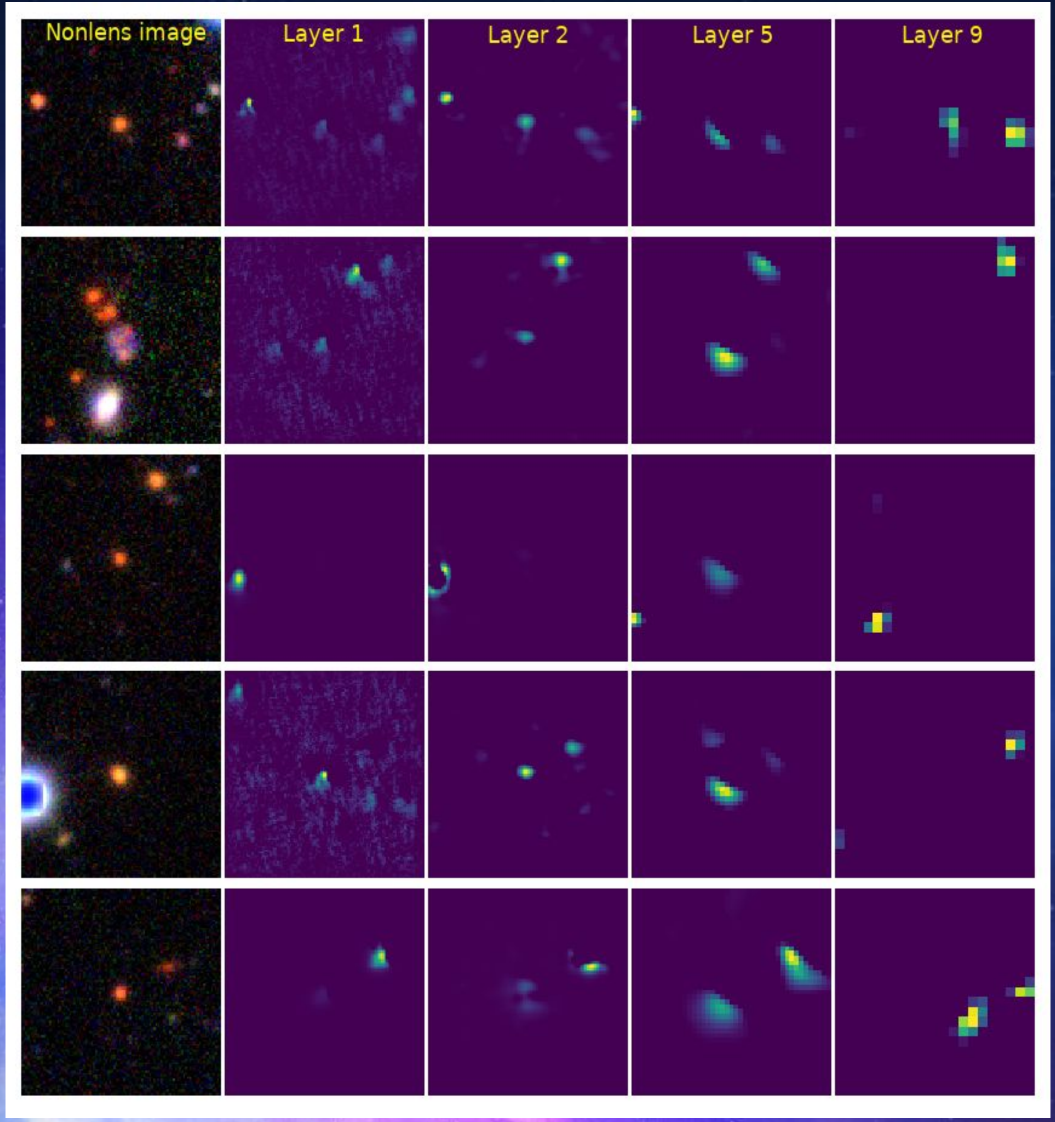
ASTRO 3D

Feature activations

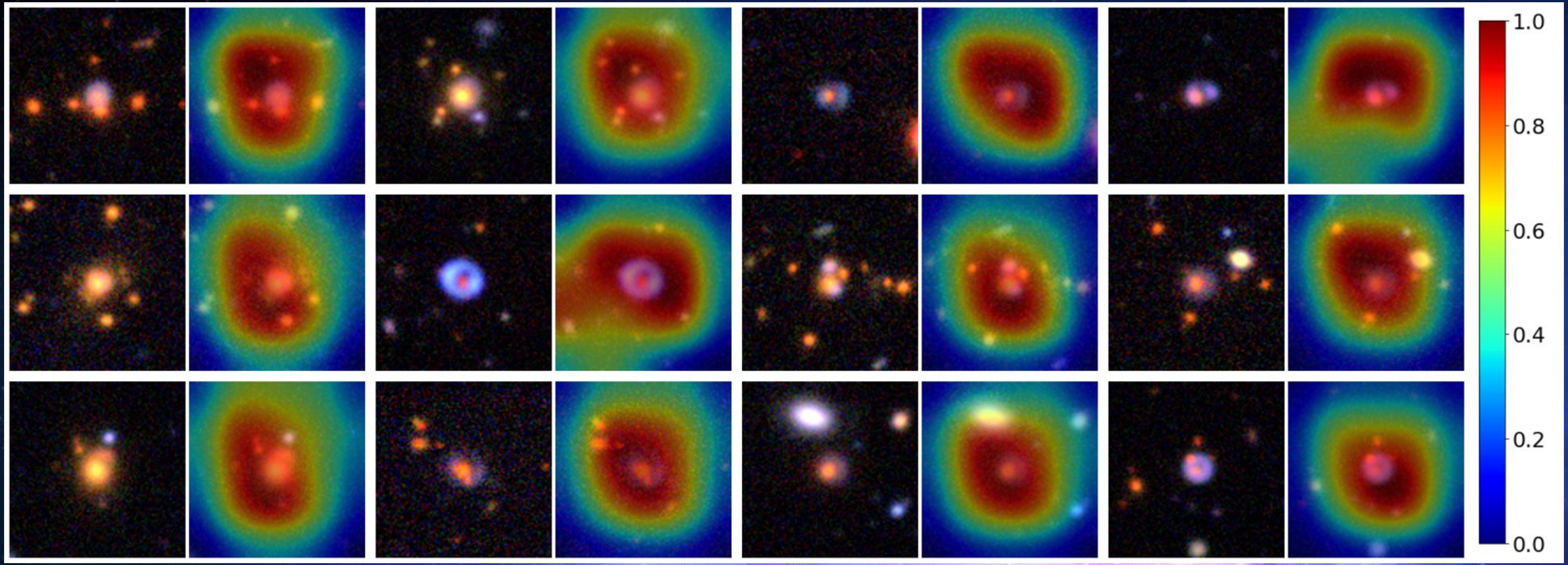


ASTRO 3D

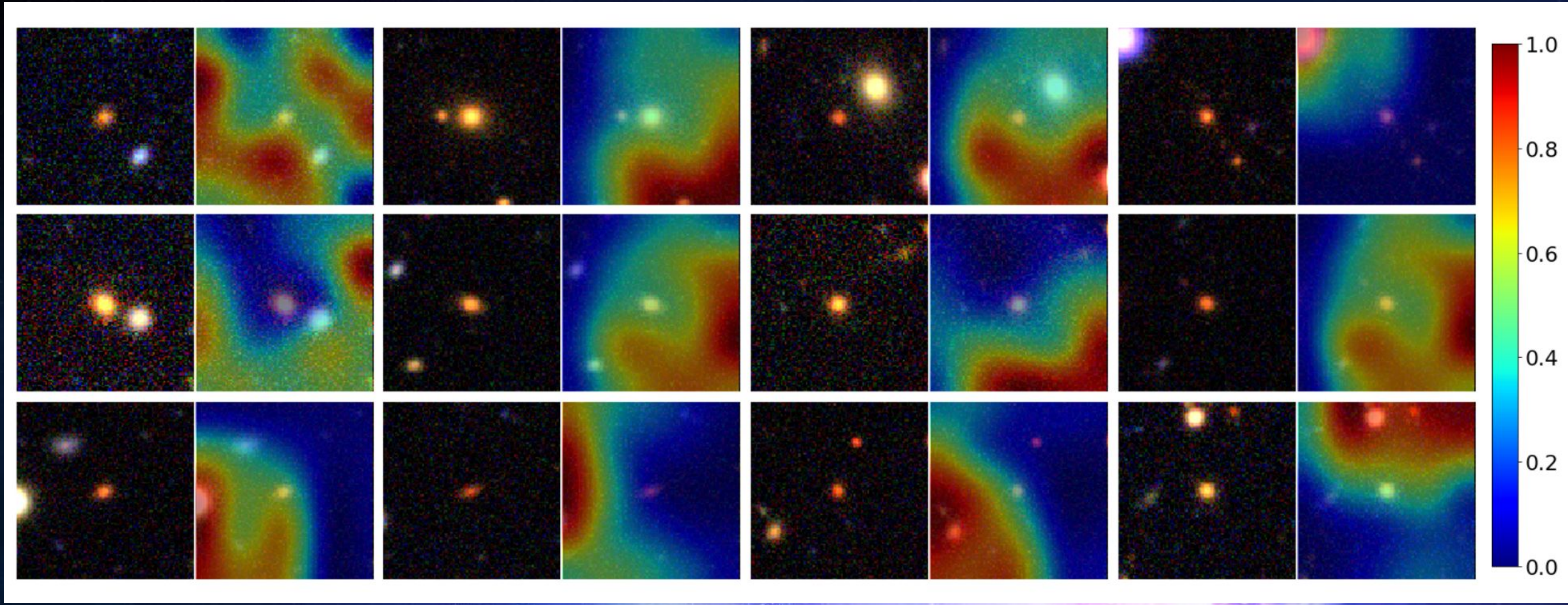

Lens image
Layer 1
Layer 2
Layer 5
Layer 9
10"


Nonlens image
Layer 1
Layer 2
Layer 5
Layer 9

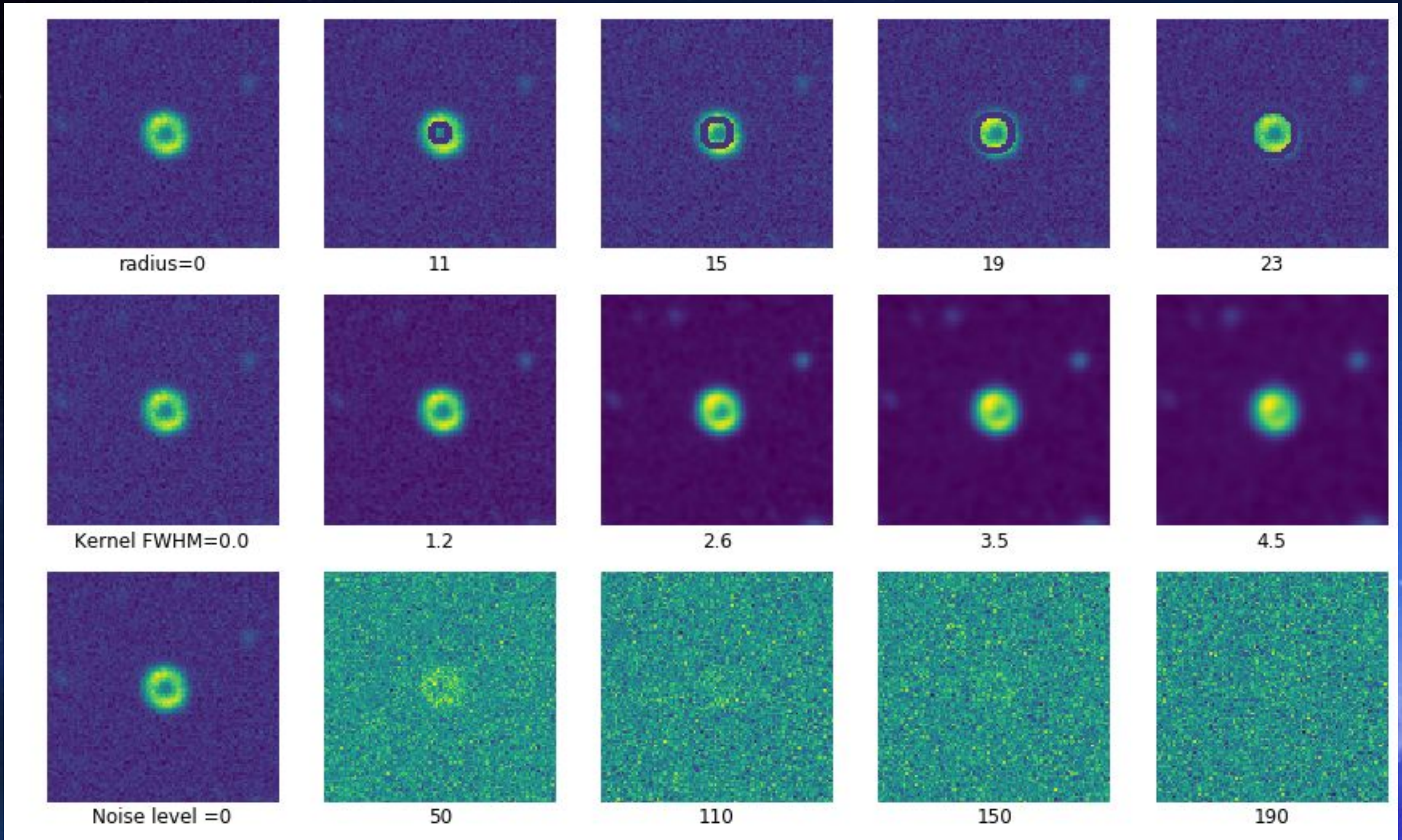
Saliency mapping: Grad-CAM



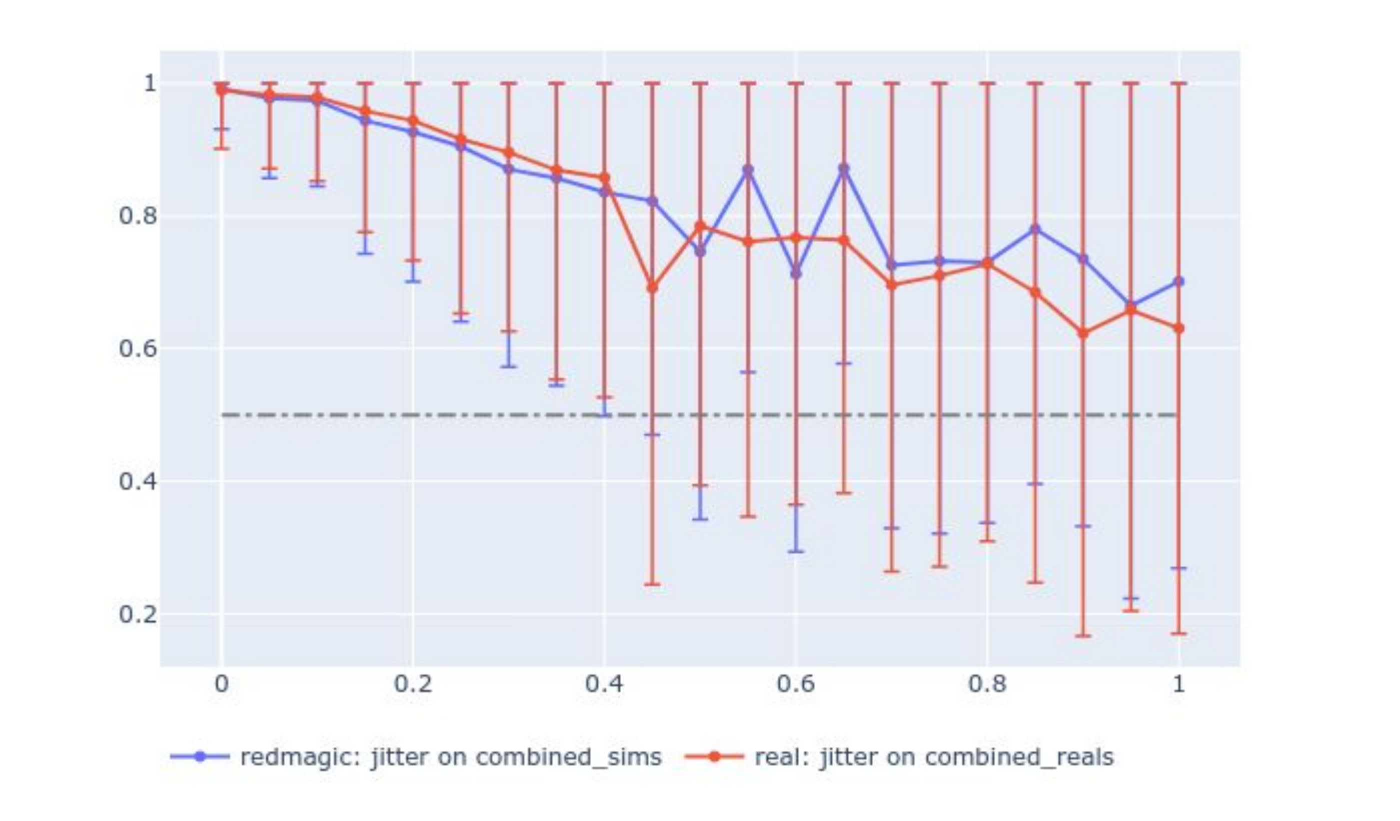
Grad-CAM - negative



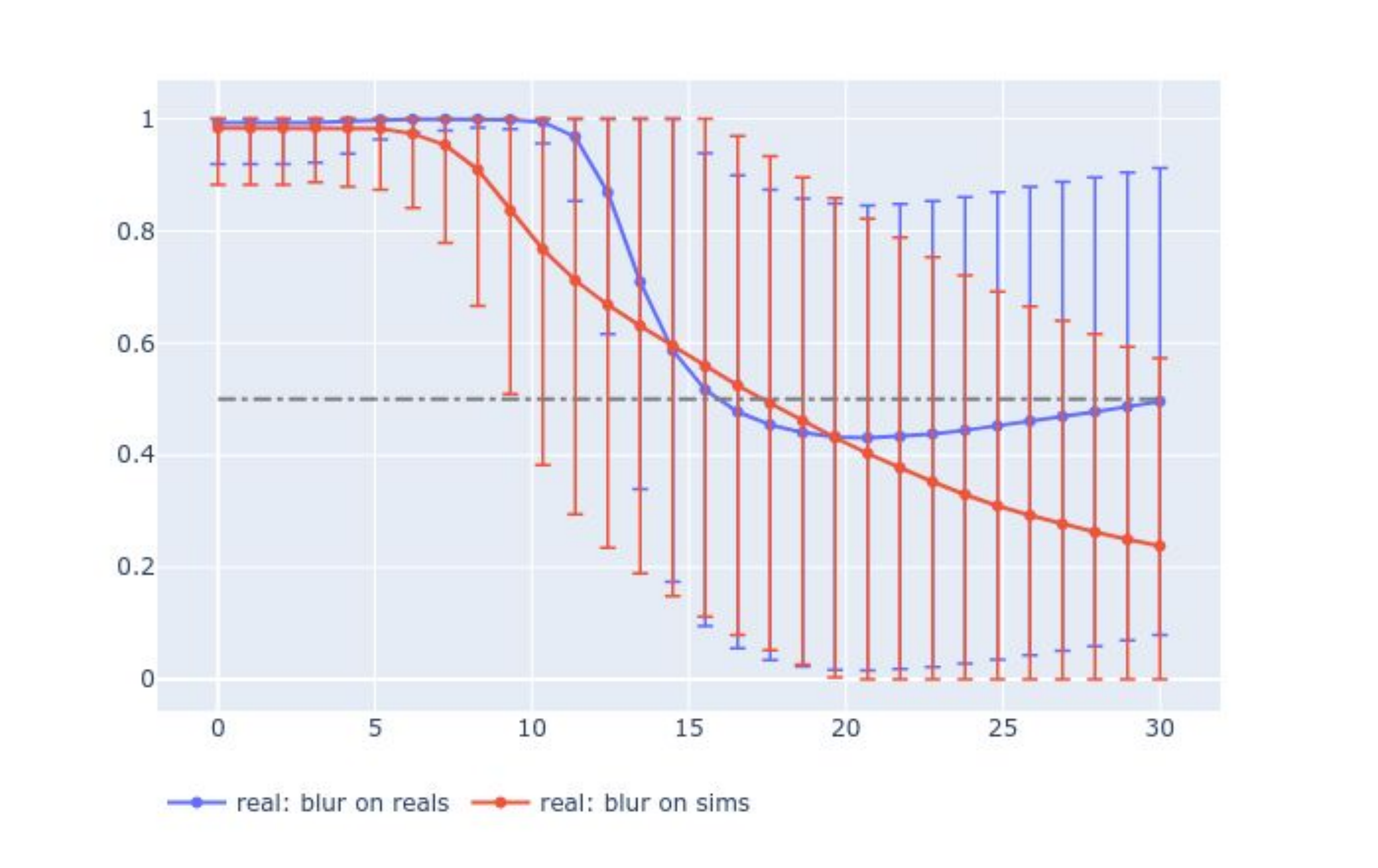
Probing with Sensie: Perturb test set



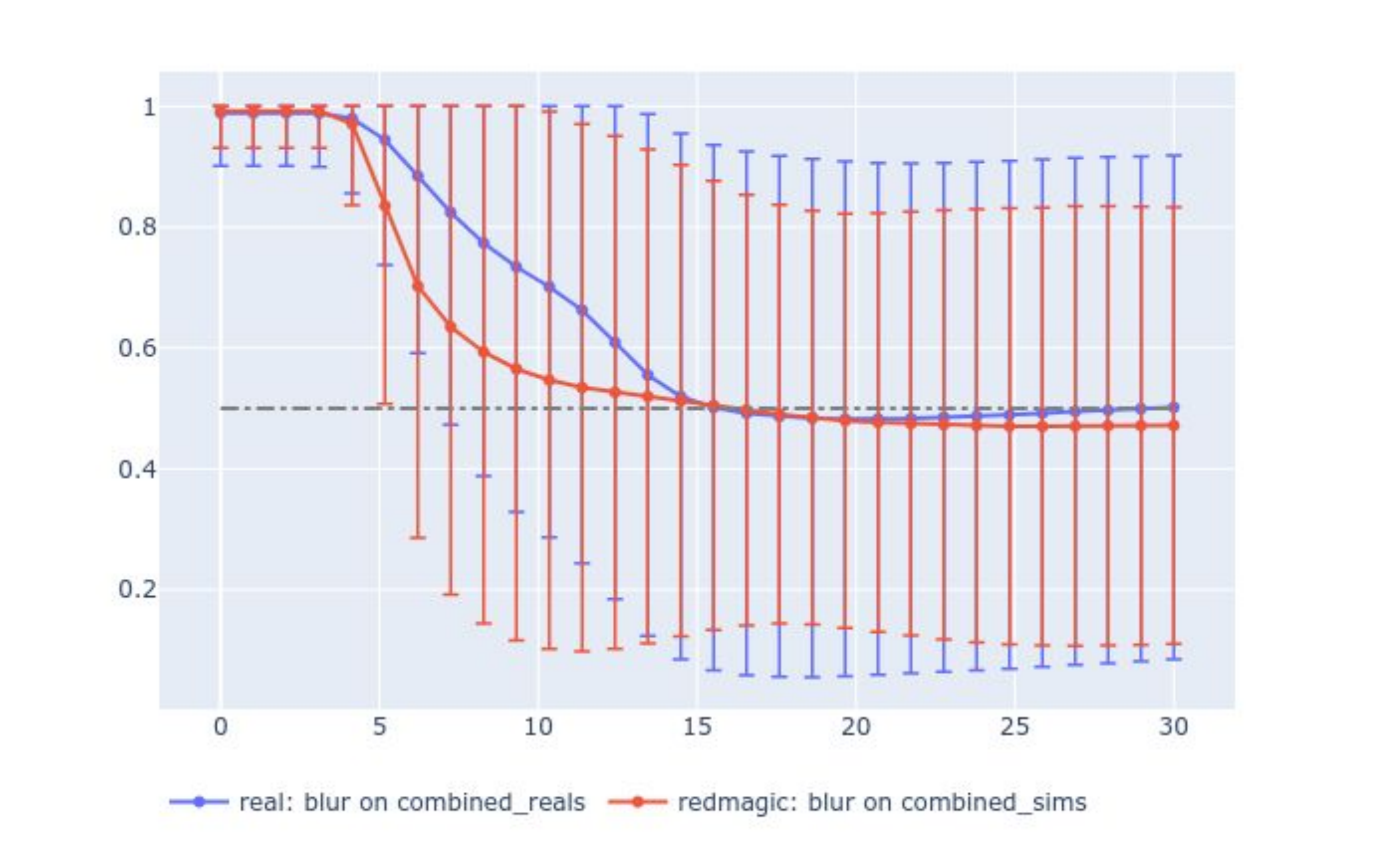
Results: Colour



Results: Blur (seeing)

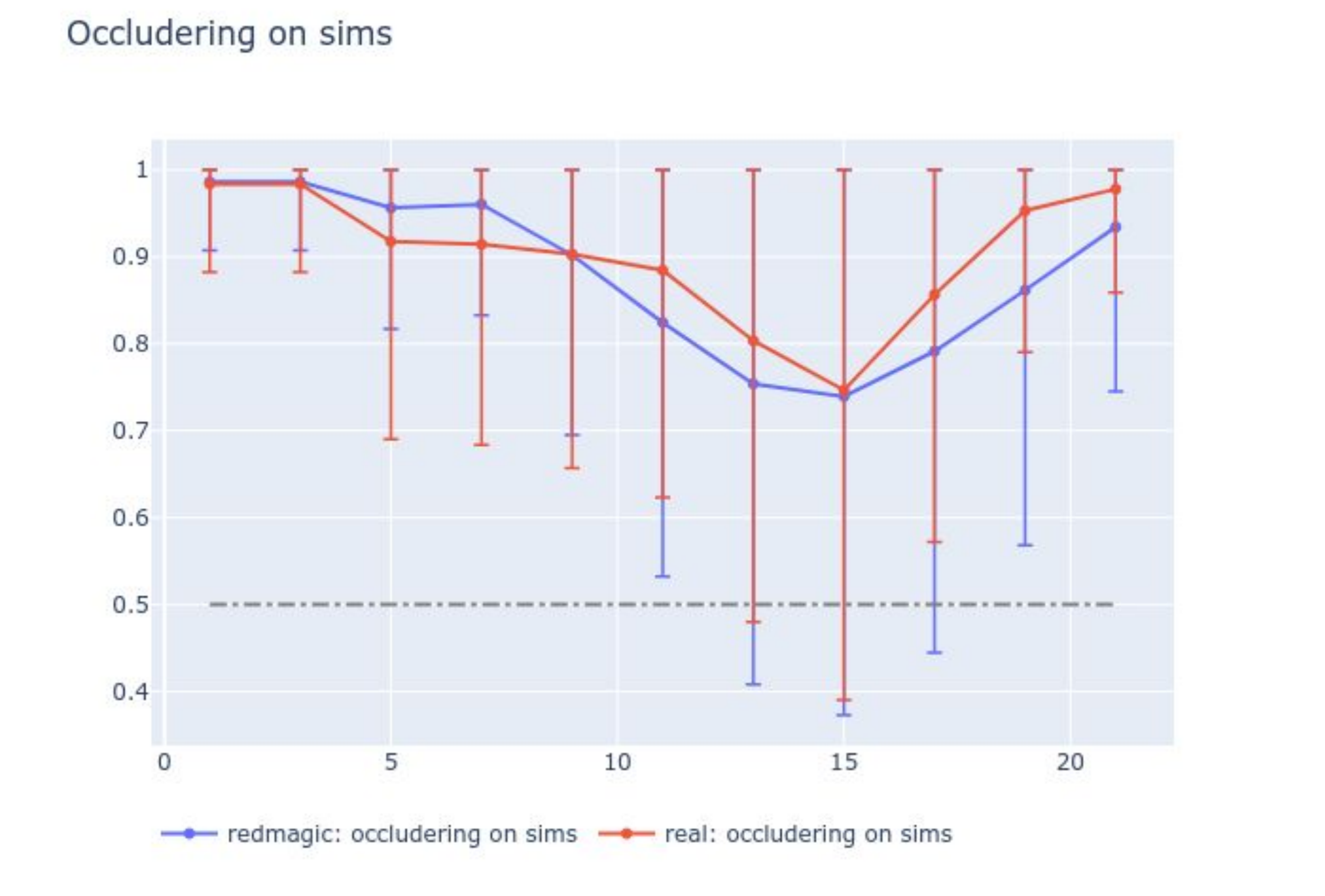


Effect on sims



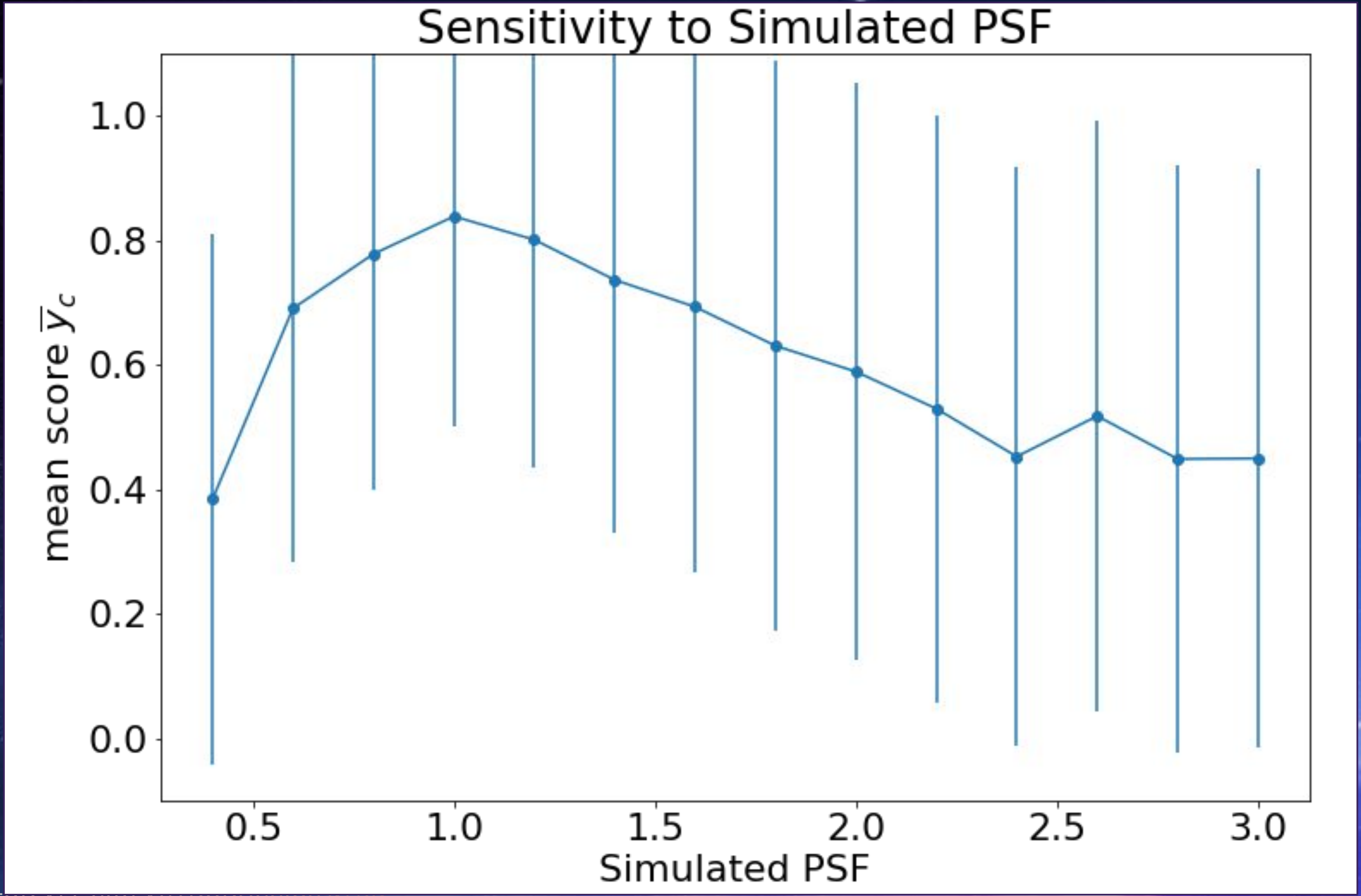
Effect on accuracy

Results: Occlusion

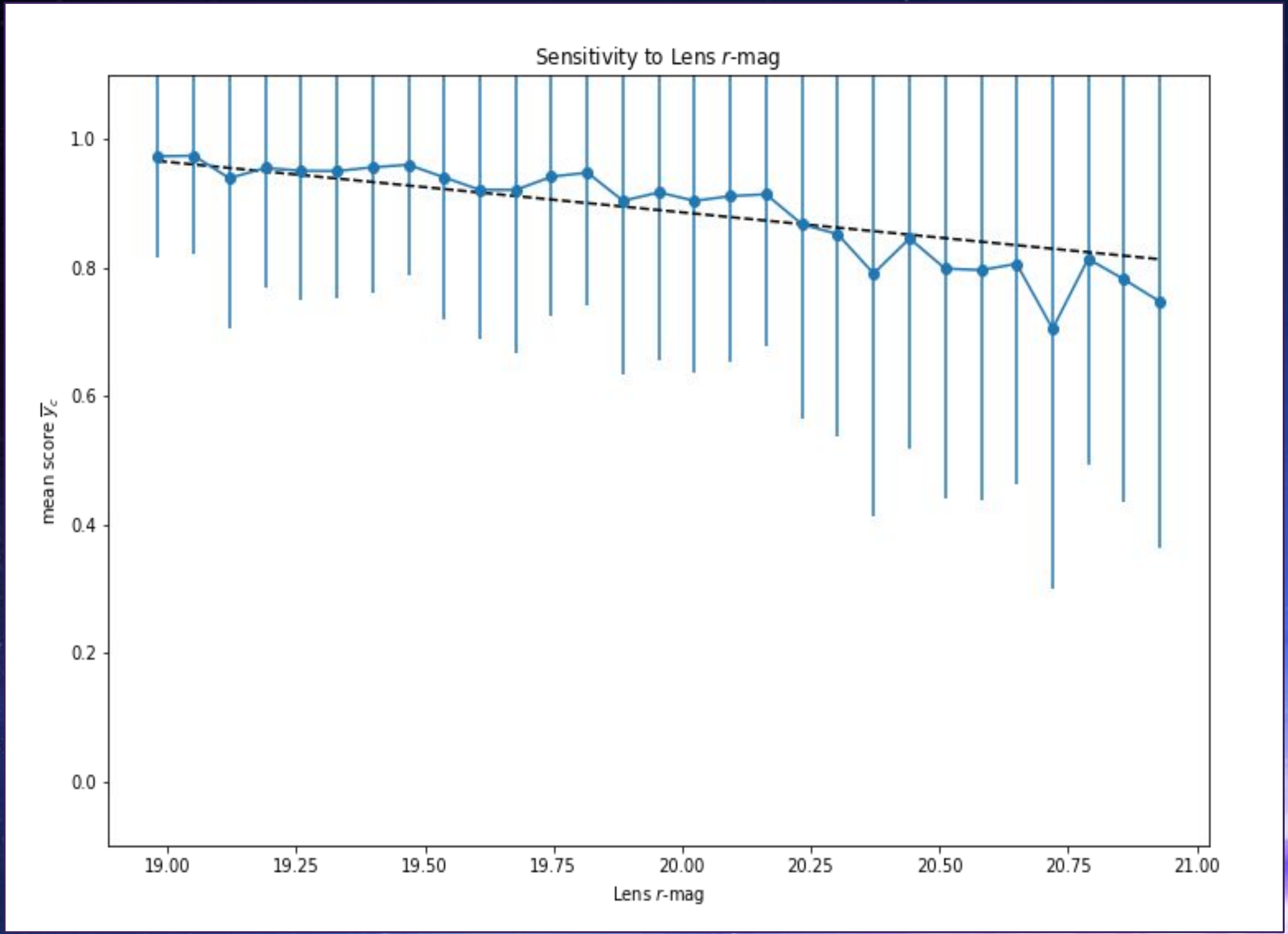


ASTRO 3D

Results: PSF

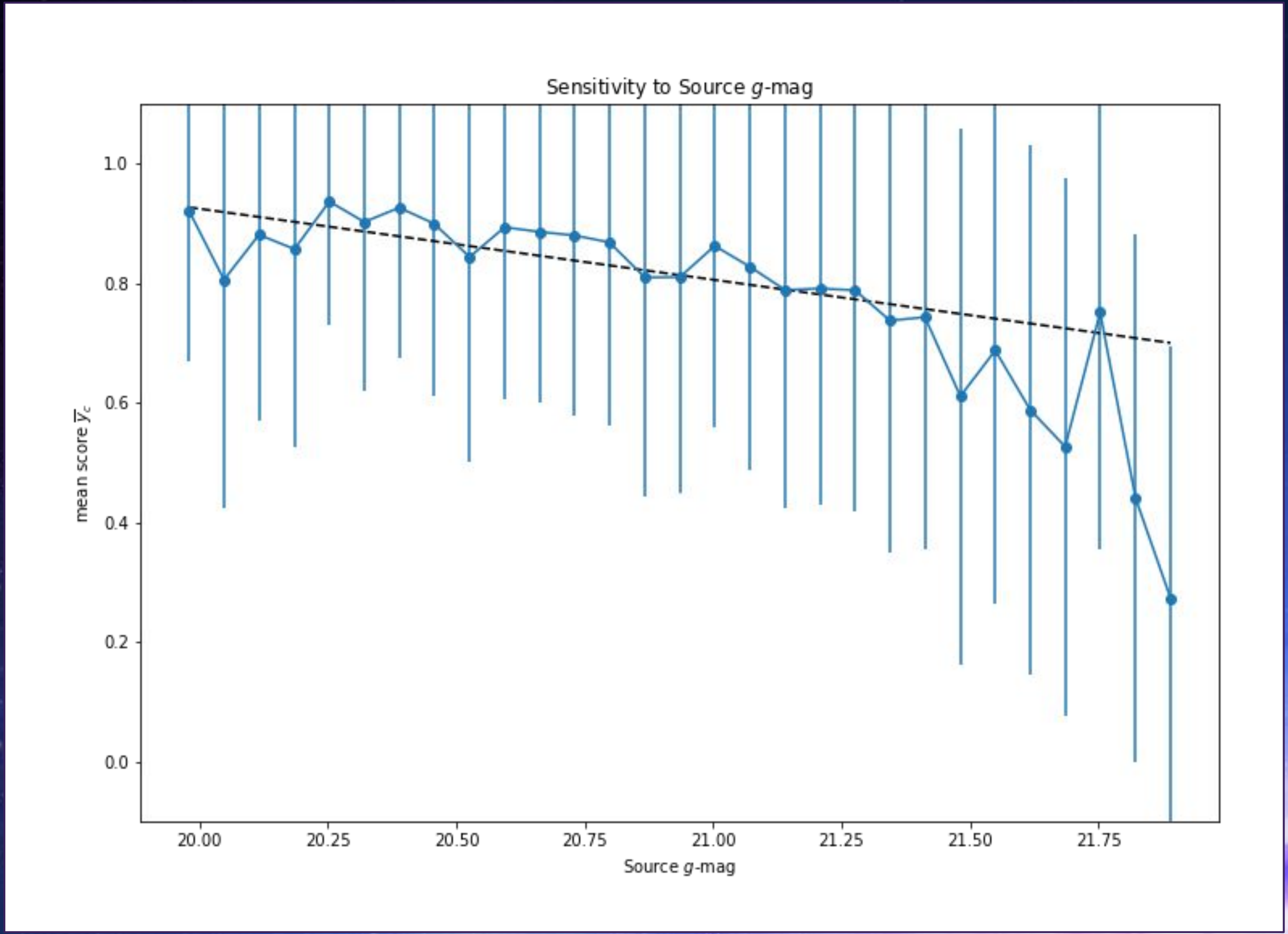


Results: Magnitude

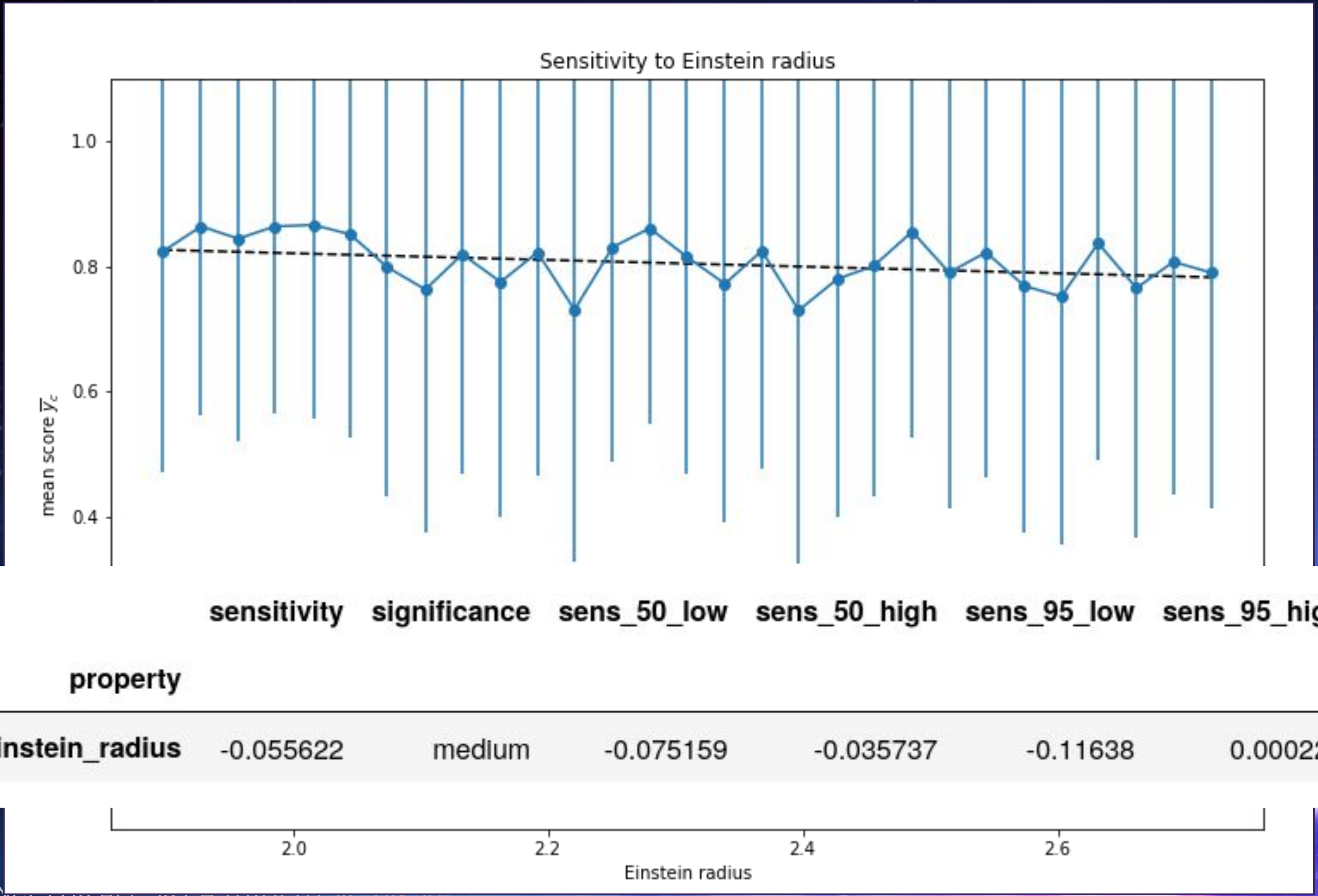


ASTRO 3D

Results: Magnitude

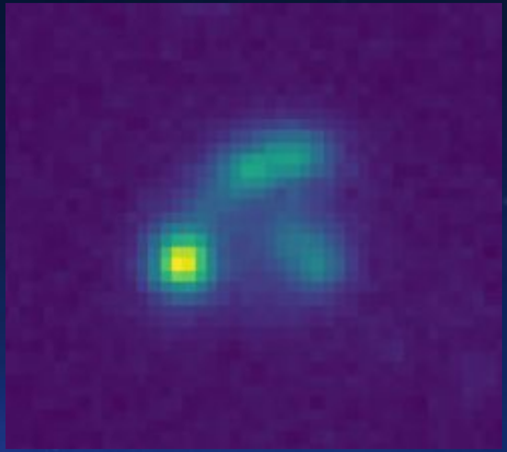


Results: Einstein Radius



	sensitivity	significance	sens_50_low	sens_50_high	sens_95_low	sens_95_high
property						
einstein_radius	-0.055622	medium	-0.075159	-0.035737	-0.11638	0.000224

Conclusions



Learned a few things:

Good/expected:

- Not sensitive to Einstein radius
- Robust to faint sources
- Sensitive to colour - physics?
- Some idea of a selection function

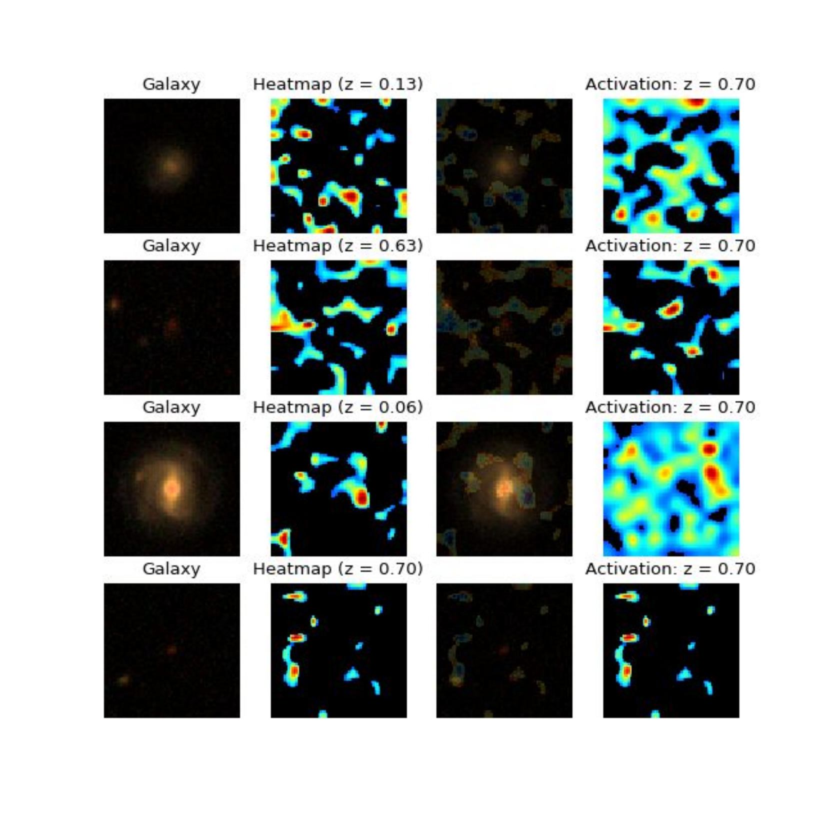
Bad:

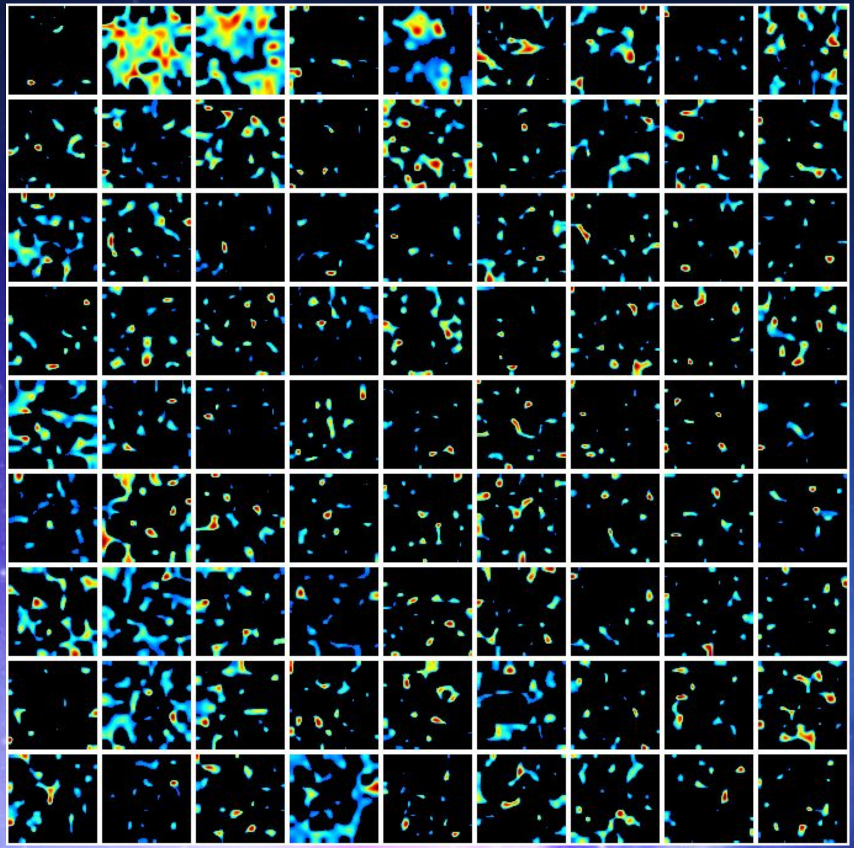
- Sensitive to simulated PSF

Need to improve training set!

github.com/coljac/sensie

Further application: Redshifts





REFERENCES

- Montavon, G., Samek, W. and Müller, K.R., 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, pp.1-15.
- Lipton, Z.C., 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.
- Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Greydanus, S., Kaul, A., Dodge, J. and Fern, A., 2017. Visualising and understanding atari agents. *arXiv preprint arXiv: 1711.00138.*
- Zeiler, M. D., & Fergus, R. 2014, in Computer Vision – ECCV 2014, ed. D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars, Vol. 8689 (Cham: Springer International Publishing), 818–833
- Selvaraju, R. R., Cogswell, M., Das, A., et al. 2017, in Proceedings of the IEEE International Conference on Computer Vision, 618–626
- Binder, A., Bach, S., Montavon, G., Müller, K.-R., & Samek, W. 2016, in Information Science and Applications (ICISA) 2016, ed. K. J. Kim & N. Joukov, Lecture Notes in Electrical Engineering (Springer Singapore), 913–922
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017), arXiv e-prints, arXiv:1706.03825.