MODEL EXPLAINABILITY

WHY IS IT IMPORTANT?

Fiona Chow

WHOAMI

born and raised

KUALA LUMPUR, MALAYSIA (FOOD PARADISE <3)

studied

ACCOUNTING AND FINANCE - PROF. QUALIFICATION **ACCA**

MSc in BIG DATA, UNIVERSITY OF STIRLING

works

BIG DATA ENGINEER

bird.i

BIRD.I

what is Bird.i?

A Satellite Image & Intelligence platform.

what do we do?

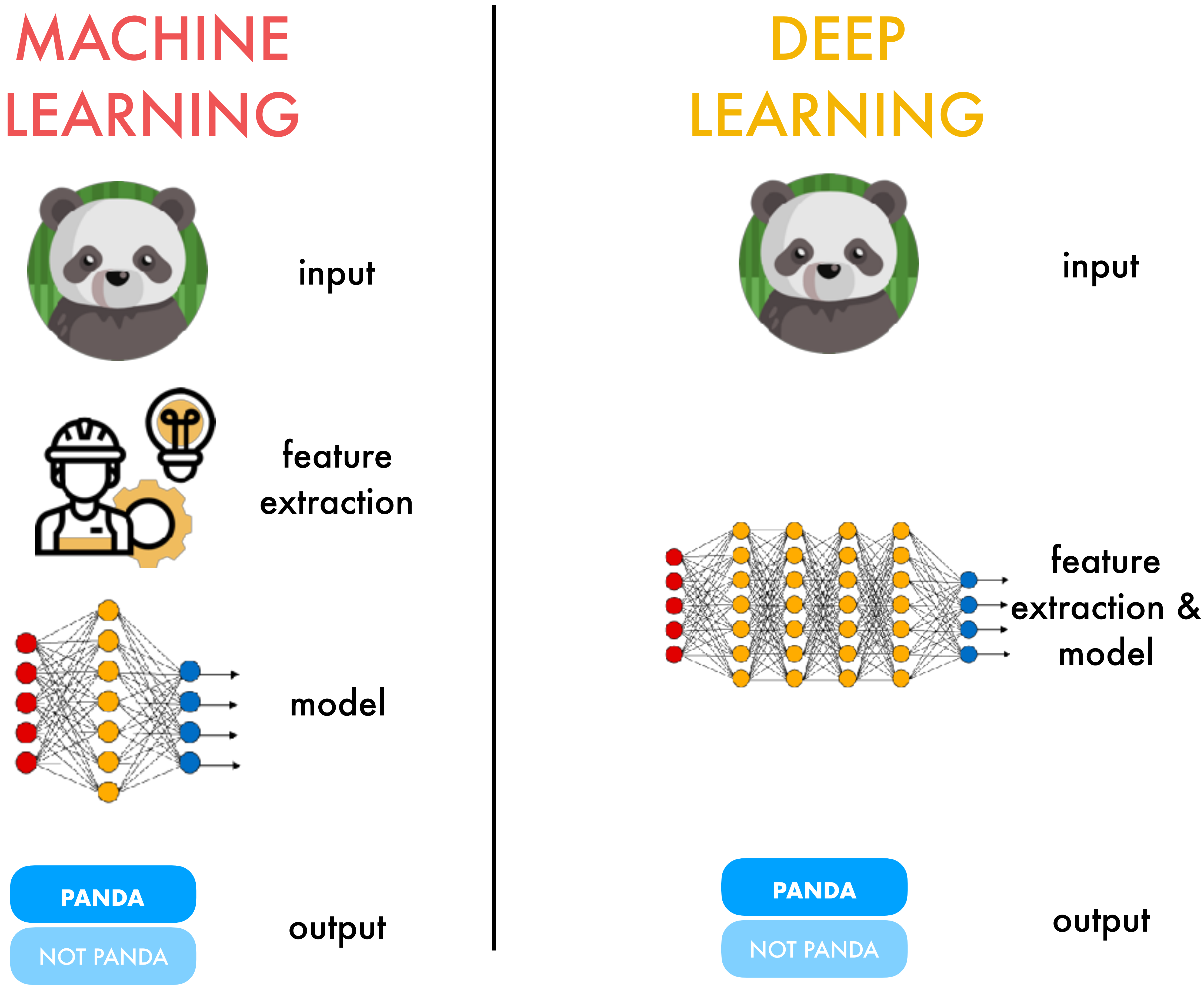
Integrate images from world leading satellite operators

Extract intelligence from satellite images using computer vision and machine learning techniques

Bird.i Portal

USER FAQ

"What did the model see to make that decision?"



MACHINE LEARNING
input
feature extraction
model
PANDA
NOT PANDA
output
DEEP LEARNING
input
feature extraction & model
PANDA
NOT PANDA
output

MACHINE LEARNING	DEEP LEARNING
have control over what features are used for training	more complex features learnt by model itself
a lot of time spent on manually selecting, extracting, engineering features	
requires a reasonable amount of data	requires a significant amount of data
lower training time	higher training time
lower computation overheads	higher computation overheads (**needs GPU!**)
accuracy plateaus	accuracy performance is more superior

WHAT ARE FEATURES?

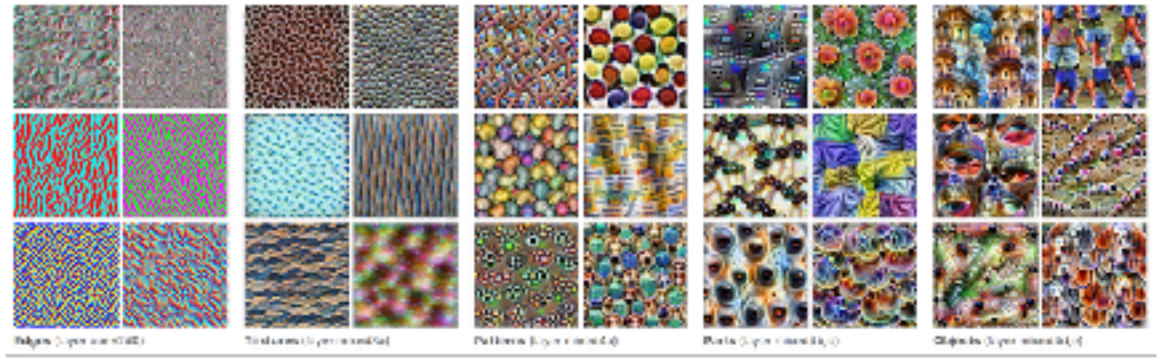
TEXT

"IT FEELS GREAT BEING HERE"

TABULAR

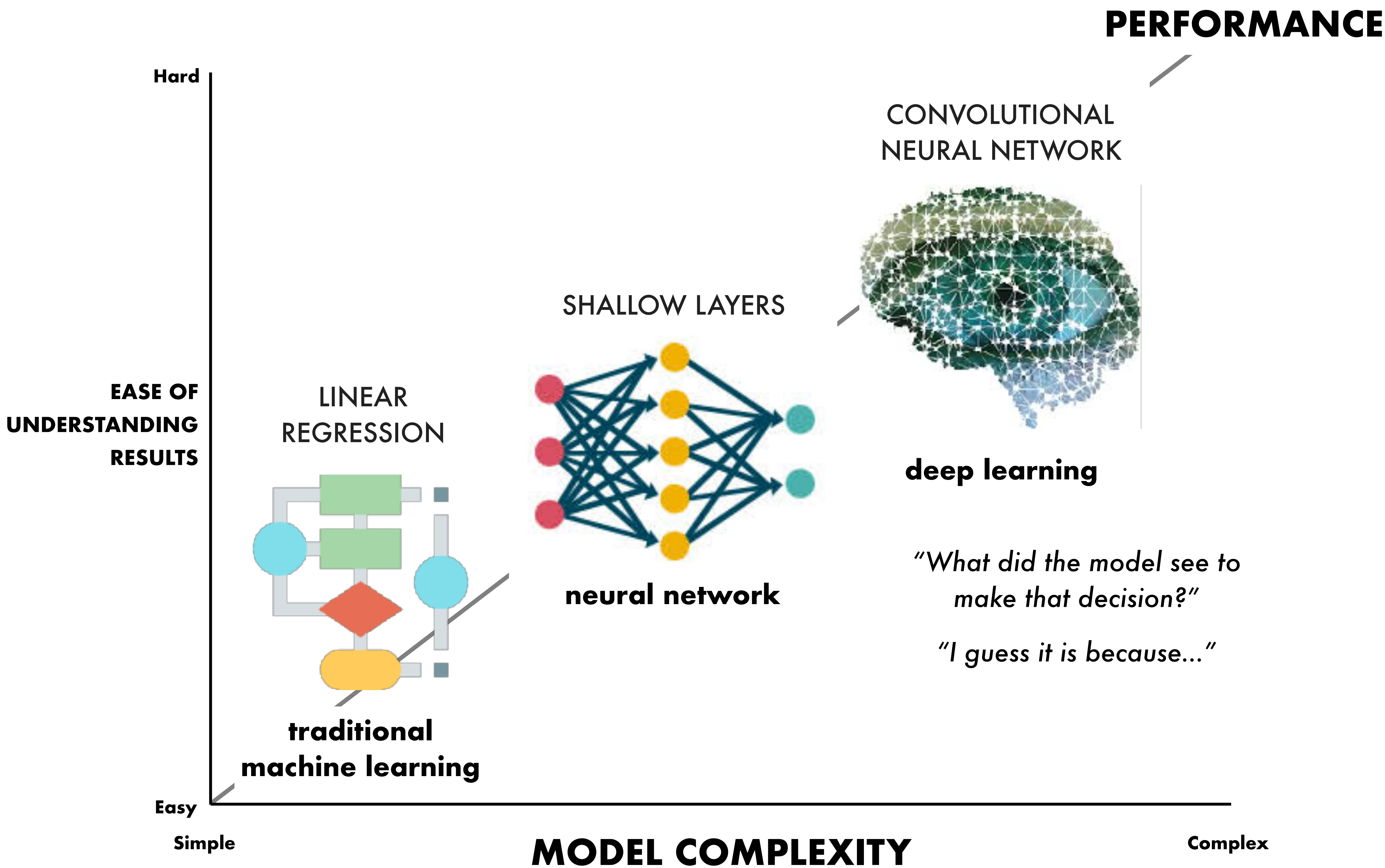
FEMALE, 20-30y/o, 152cm, EMPLOYED

IMAGE



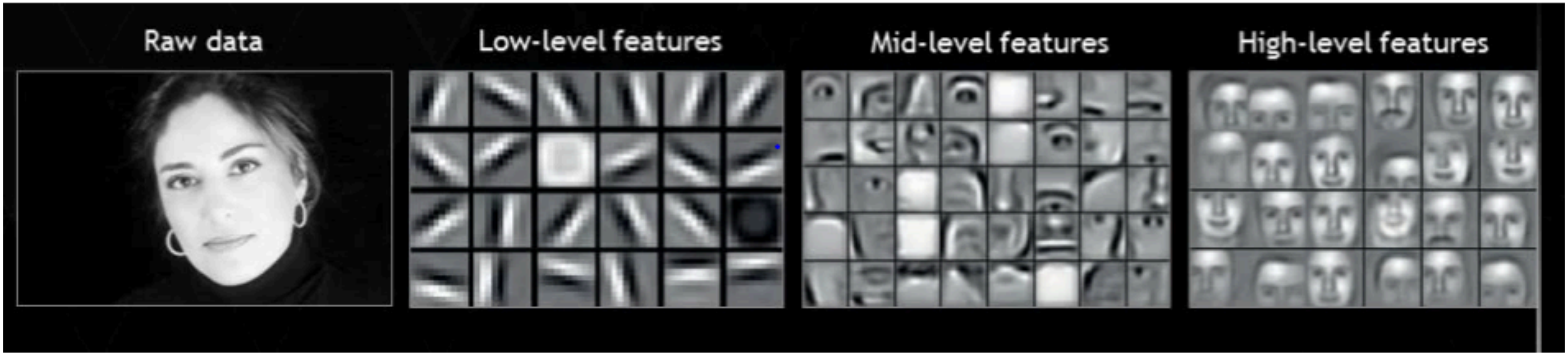
AUDIO

ML PROBLEM



PERFORMANCE
Hard
EASE OF
UNDERSTANDING
RESULTS
Easy
Simple
MODEL COMPLEXITY
Complex
CONVOLUTIONAL
NEURAL NETWORK
deep learning
SHALLOW LAYERS
neural network
LINEAR
REGRESSION
traditional
machine learning
"What did the model see to
make that decision?"
"I guess it is because..."

COMPLEXITY IN IMAGES



https://www.analyticsvidhya.com/blog/2017/04/comparison-between-deep-learning-machine-learning/

MODEL EXPLAINABILITY

"The extent to which the internal mechanics of a machine or deep learning system can be explained in **human terms**"

"Why is this happening?"

IMPORTANCE OF MODEL EXPLAINABILITY

DATA SCIENTISTS

understand what features are important

fix dataset if wrong features are learnt

point to the right direction for future data collection

provide a better explanation than "I think"

IMPORTANCE OF MODEL EXPLAINABILITY

USERS

build <u>**TRUST & ACCOUNTABILITY**</u>
between users and model

NO TRUST

↓

RELIED ON

↓

VALUE NOT DELIVERED

FORENSIC

MEDICAL **LEGAL**

FINANCE

EXPLAINABILITY ON IMAGES

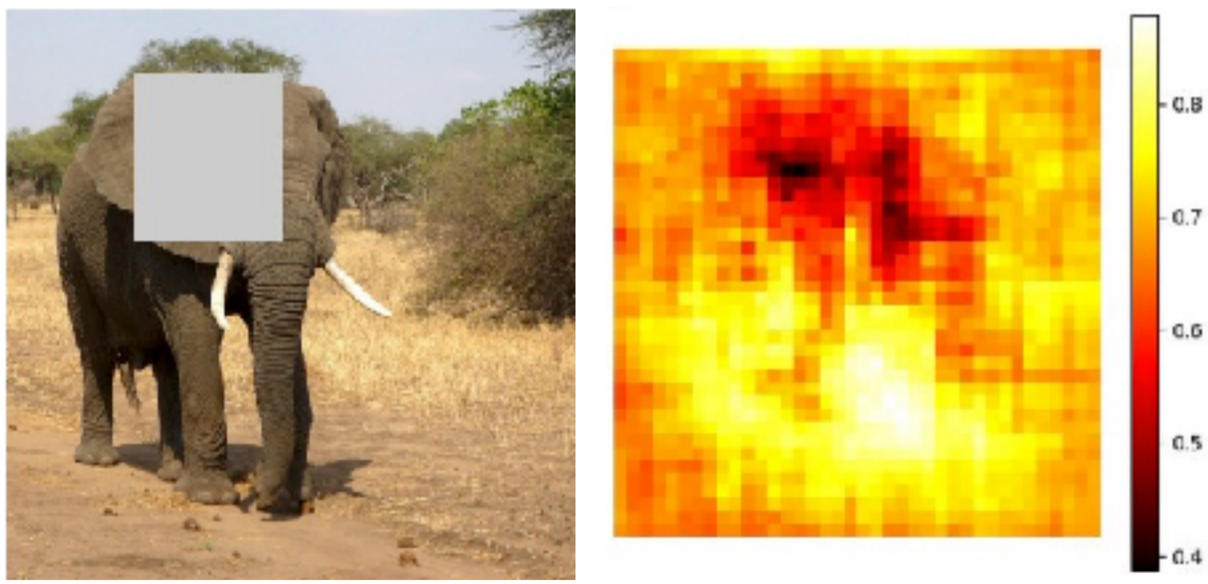


IMAGE OCCLUSION

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture12.pdf

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

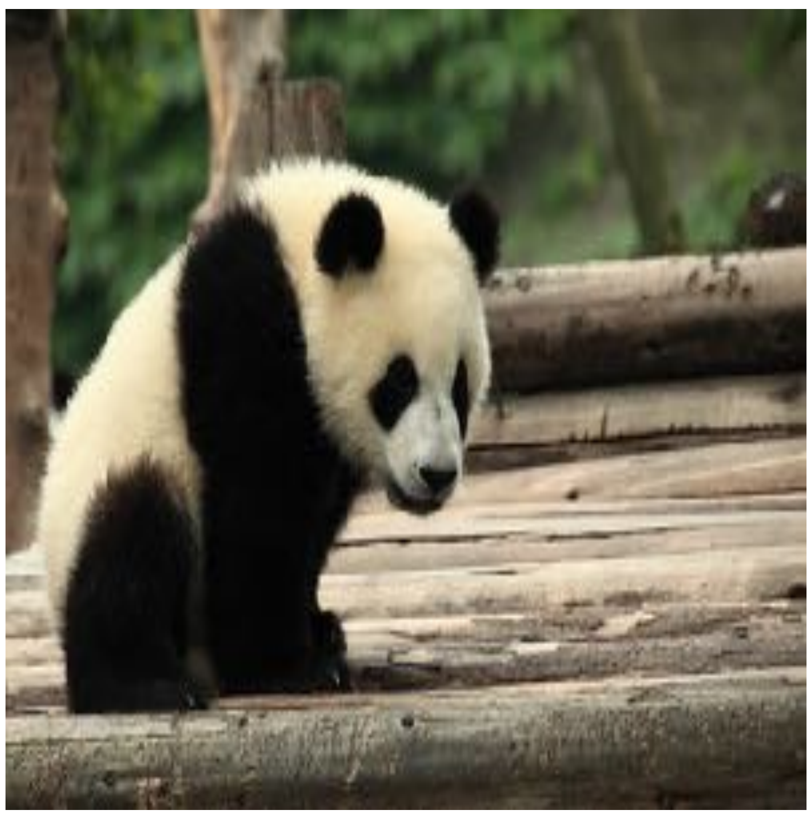
https://arxiv.org/pdf/1602.04938v1.pdf

CLASS ACTIVATION MAP

http://cnnlocalization.csail.mit.edu/Zhou_Learning_Deep_Features_CVPR_2016_paper.pdf

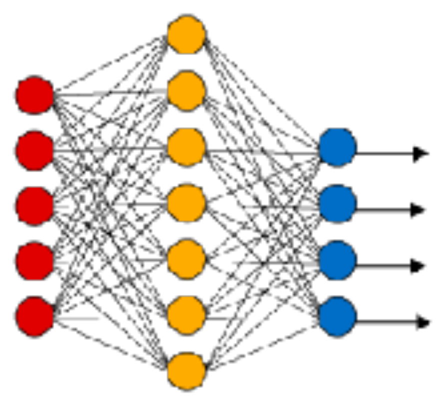
EXAMPLE

input



model

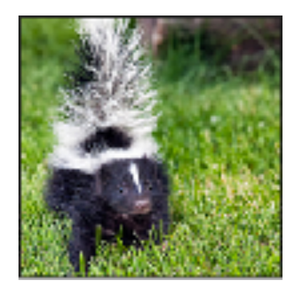
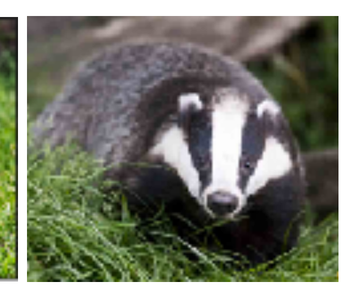
(VGG16)



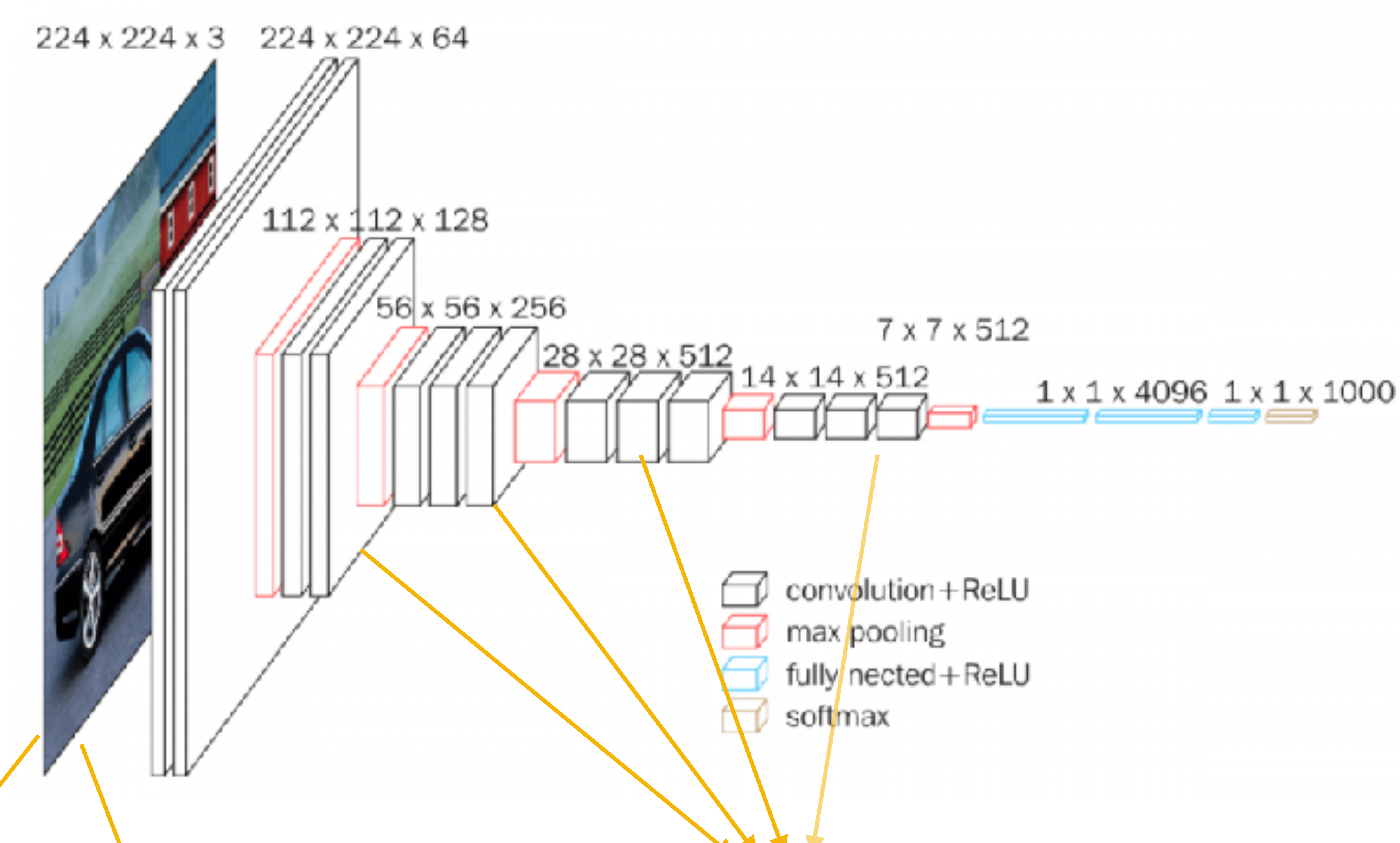
output

label_id, label_name, probability

```
[('n02510455', 'giant_panda', 0.99959975),
 ('n02445715', 'skunk', 0.0001676529),
 ('n02447366', 'badger', 0.00013132524),
 ('n02443114', 'polecat', 6.1703563e-06)]
```


EXAMPLE MODEL (VGG16)



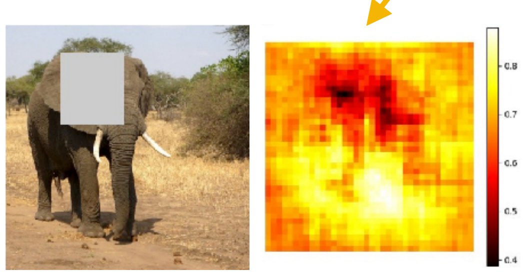
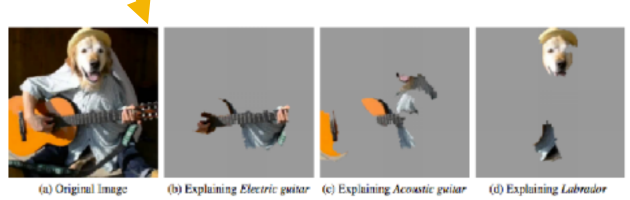
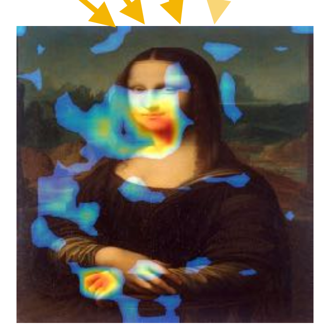


IMAGE OCCLUSION



LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

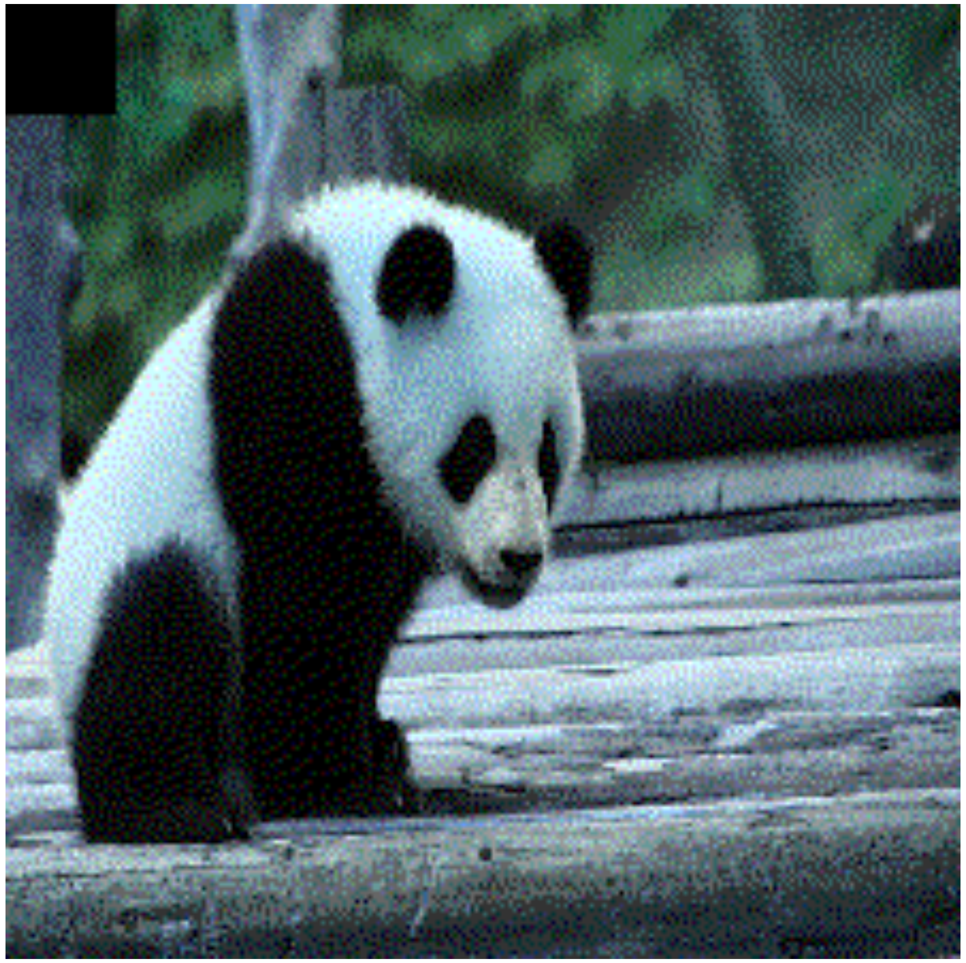


CLASS ACTIVATION MAP

https://arxiv.org/pdf/1409.1556.pdf

IMAGE OCCLUSION

OCCLUSION



HEAT MAP

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

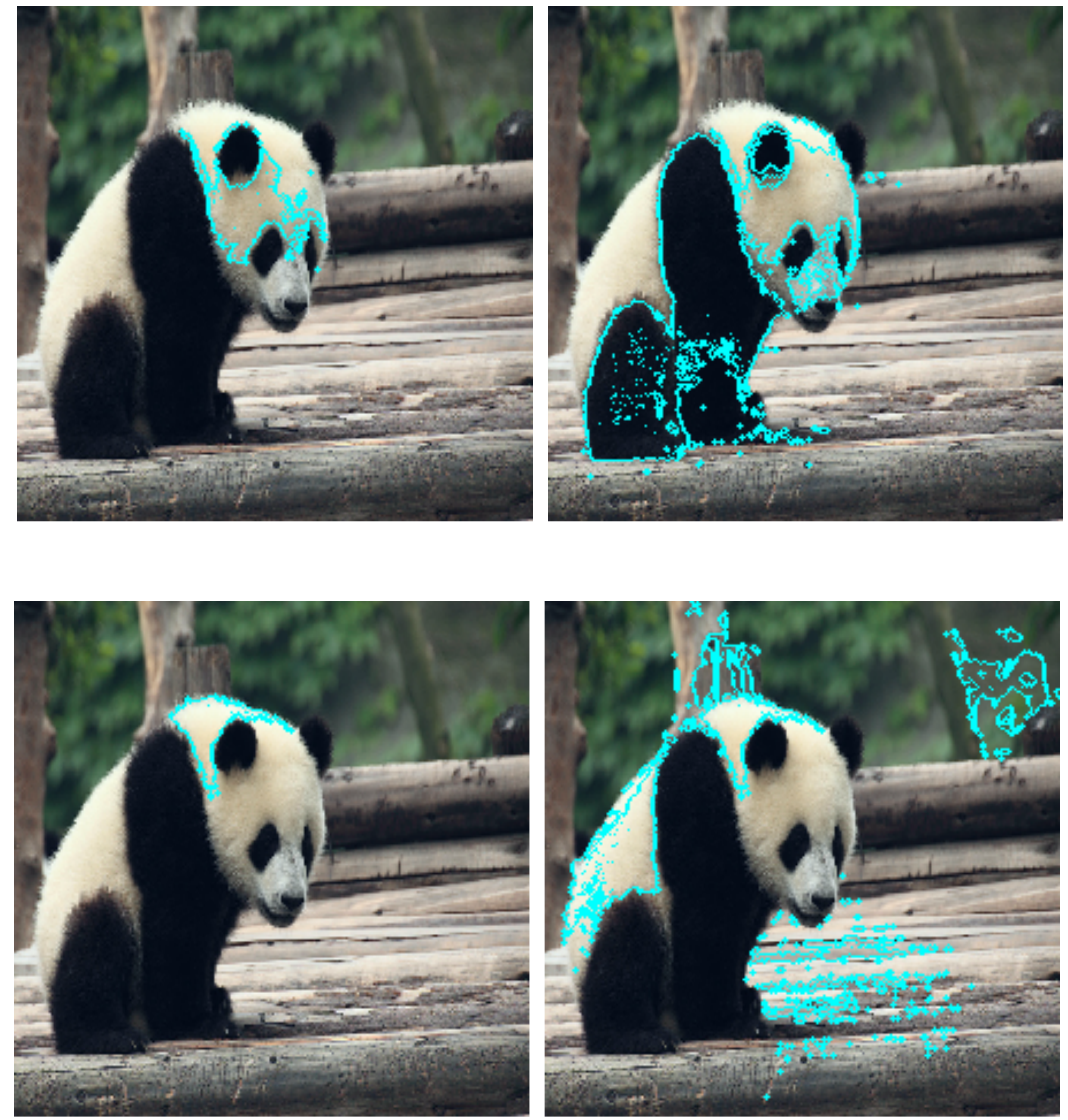
https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME

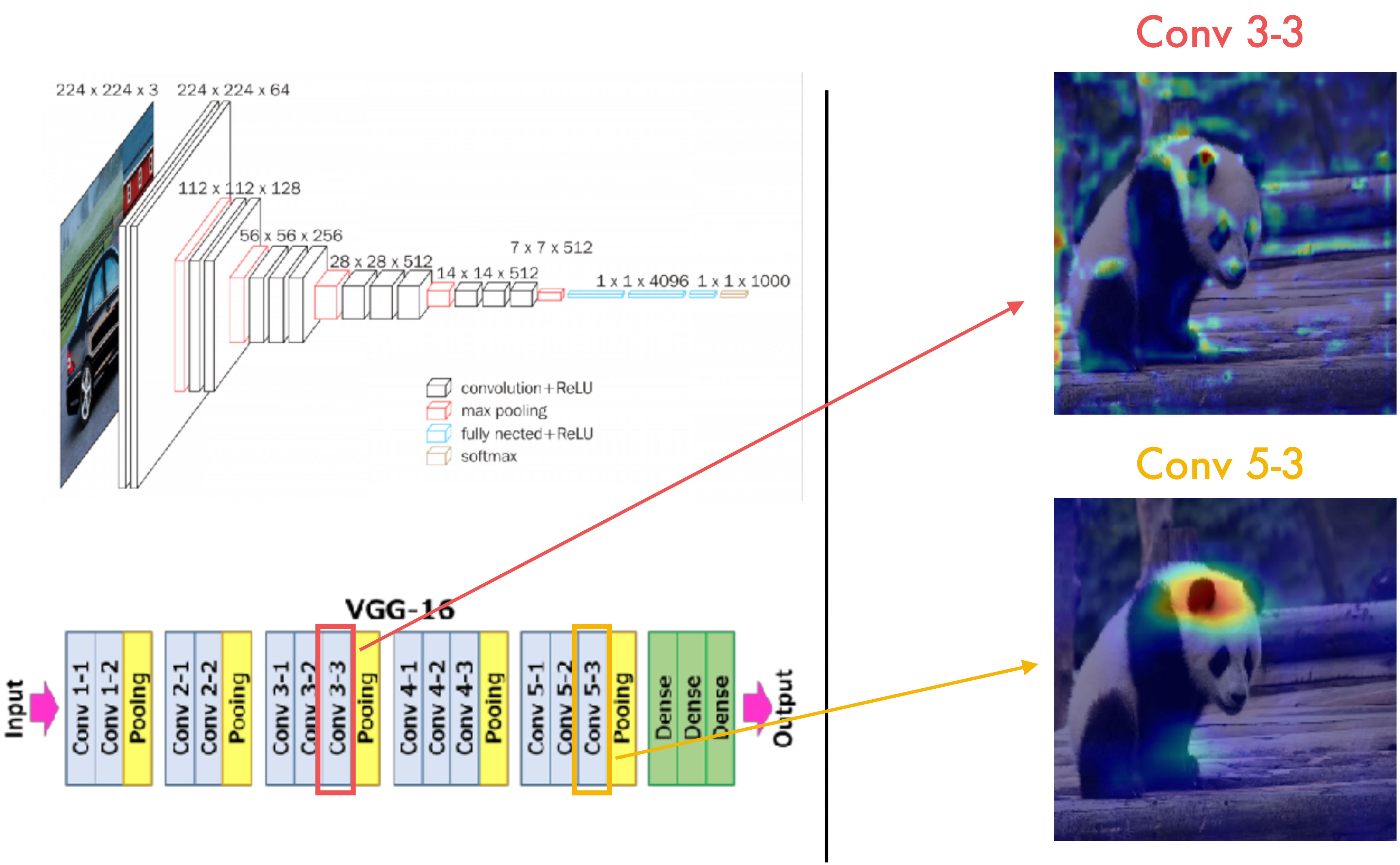
Top Feature | Top 5 Features

GIANT PANDA BECAUSE

SKUNK BECAUSE...



CLASS ACTIVATION MAP



THIS IS A GIANT PANDA

BECAUSE...

IMAGE OCCLUSION

LIME

CLASS ACTIVATION MAP

SUMMARY

- More complex the model, more difficult to explain the results
- Users who don’t understand tend to not trust the results and as a result not see the value of automation
- More emphasis on model explainability by experts in the industry. We can expect more research and techniques in future.

THANK YOU!