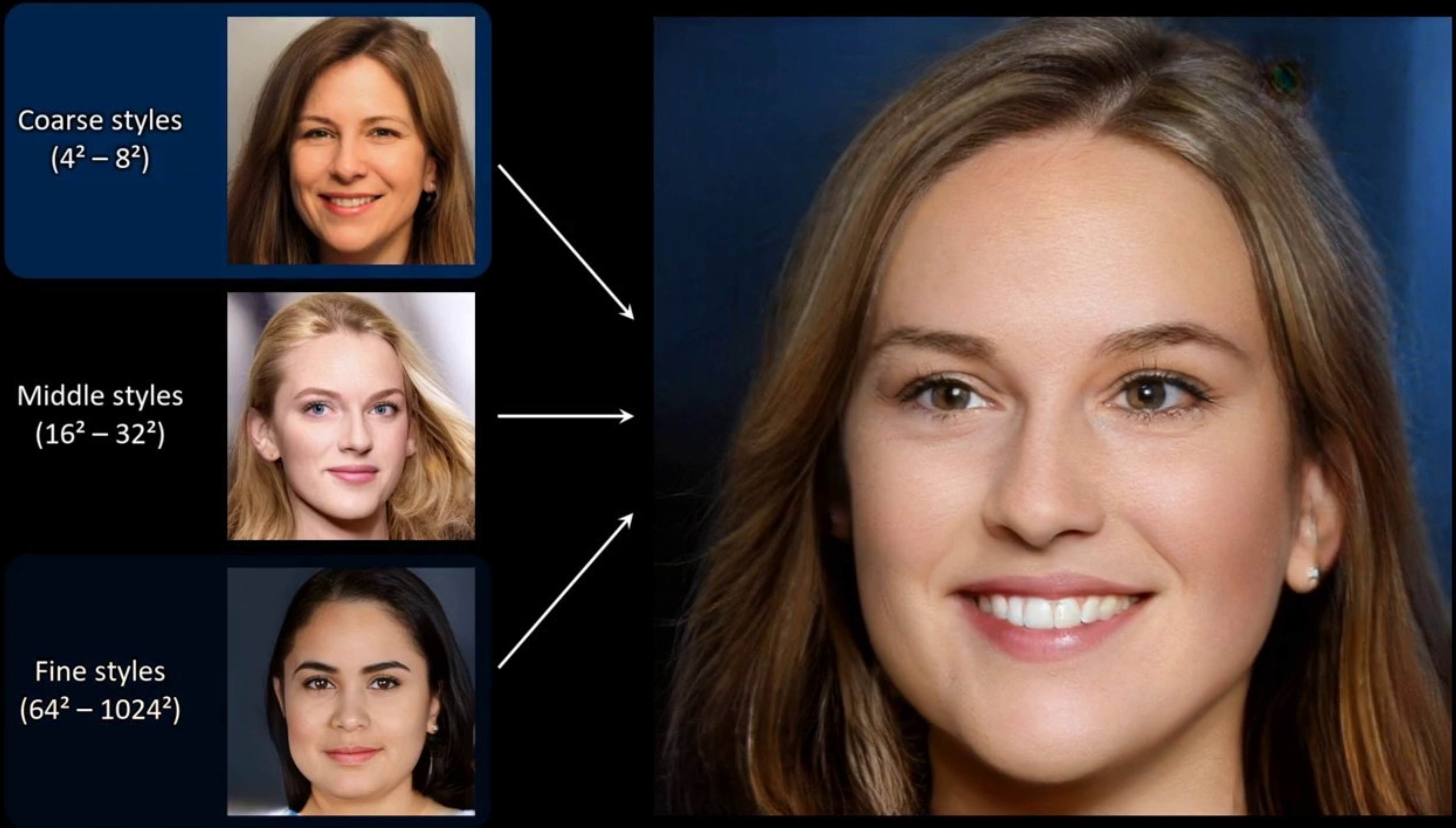
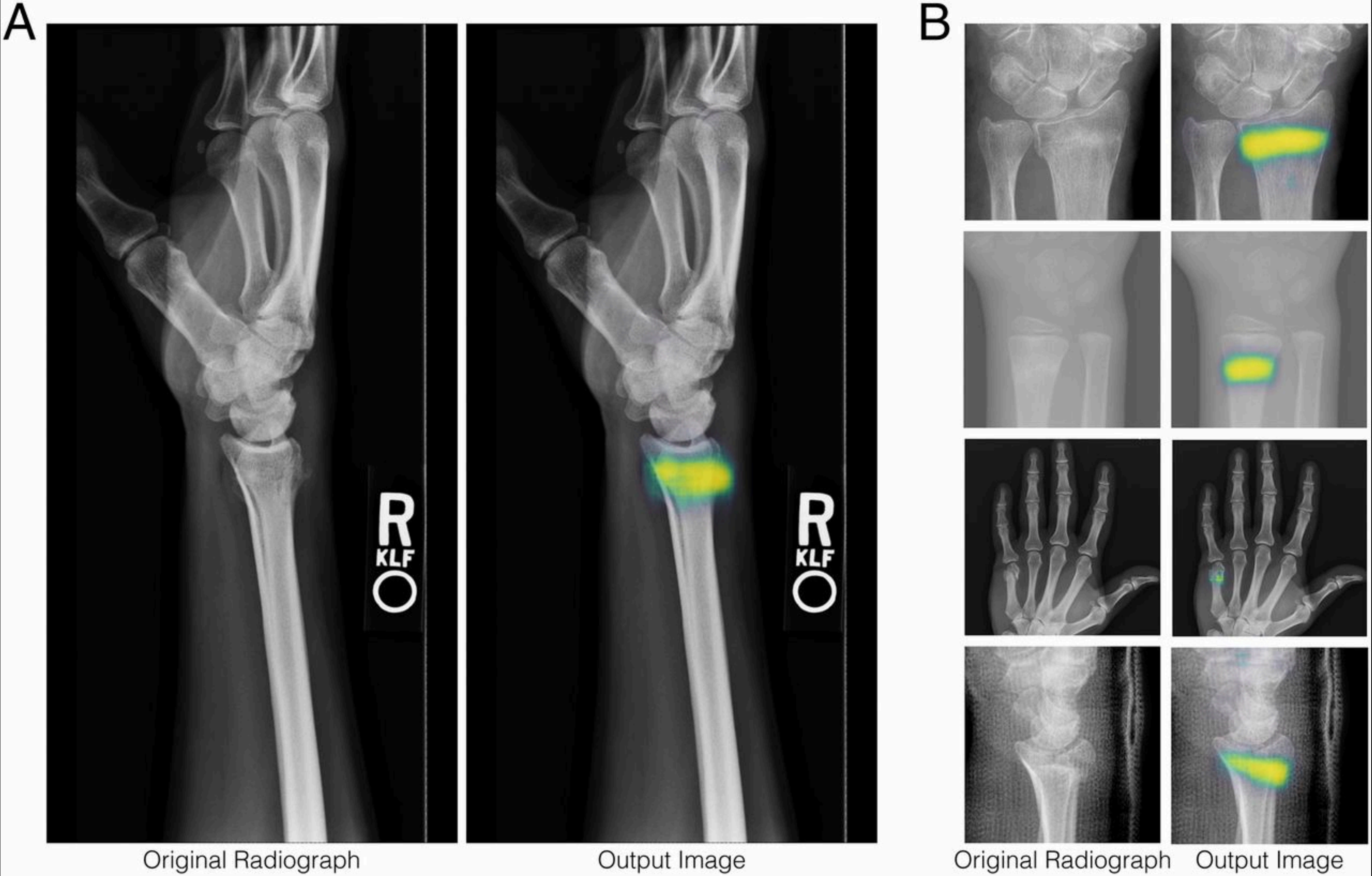
INTERPRETABLE DEEP LEARNING

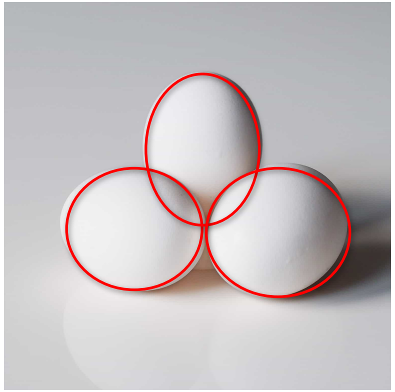
Cengiz Öztireli


Coarse styles
(4² – 8²)
Middle styles
(16² – 32²)
Fine styles
(64² – 1024²)


A
R
KLF
Original Radiograph
Output Image
B
Original Radiograph
Output Image

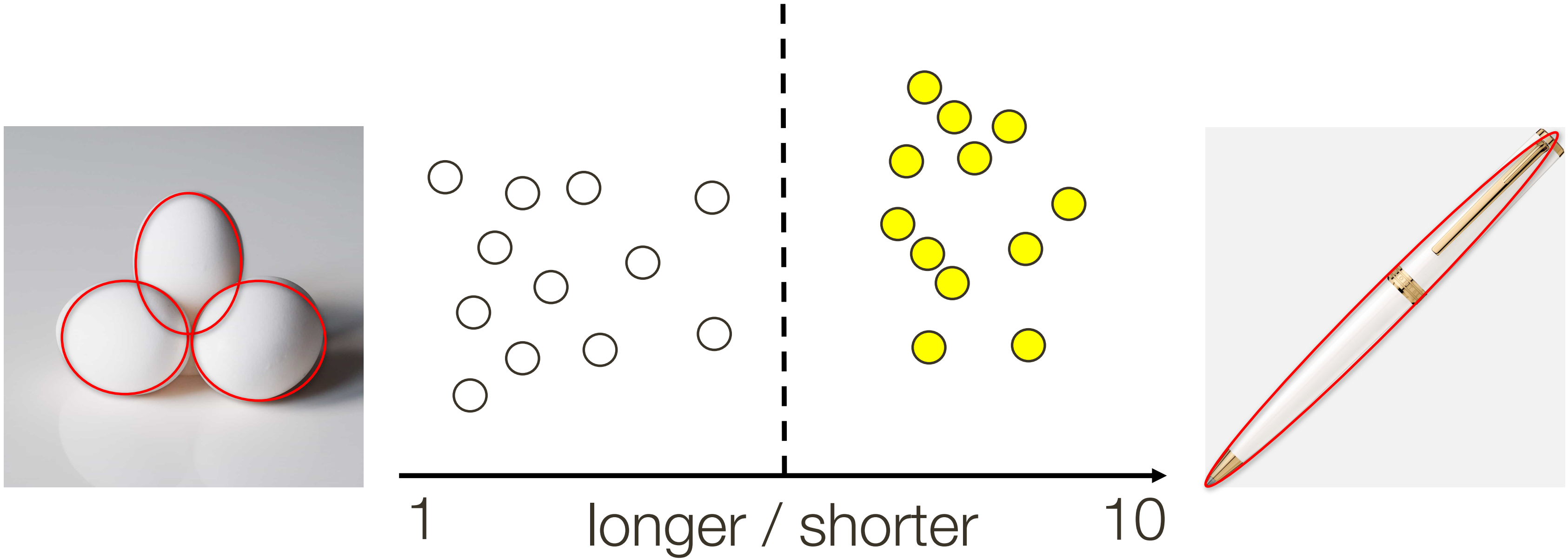
WHAT IS INTERPRETABLITY?

WHAT IS INTERPRETABLITY?

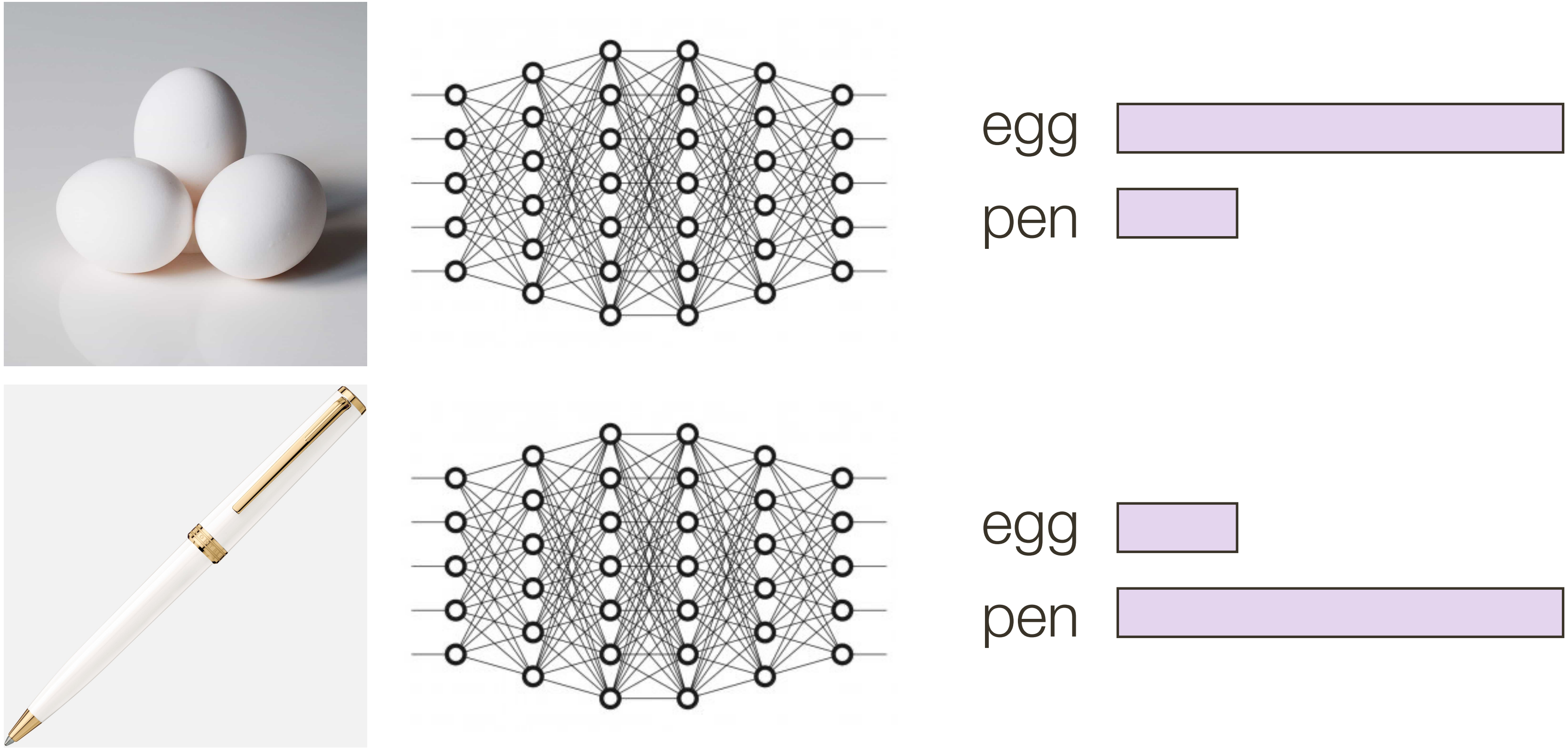


1 longer / shorter 10

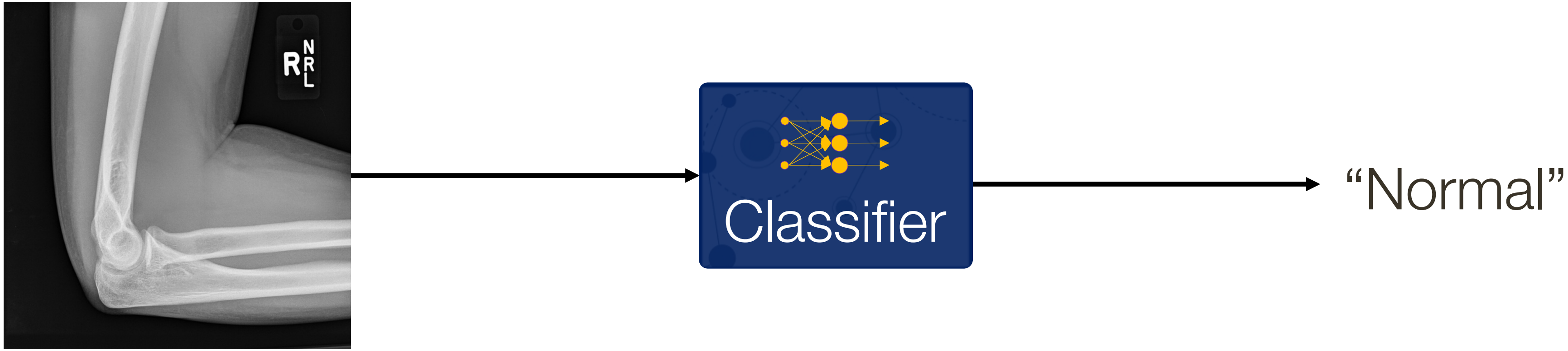
WHAT IS INTERPRETABLITY?



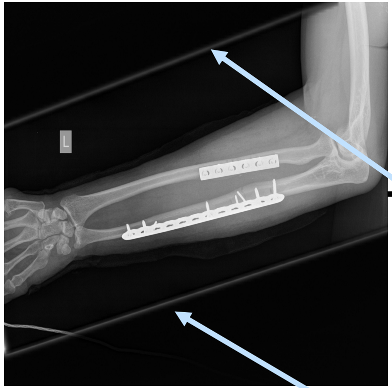
WHAT IS INTERPRETABLITY?

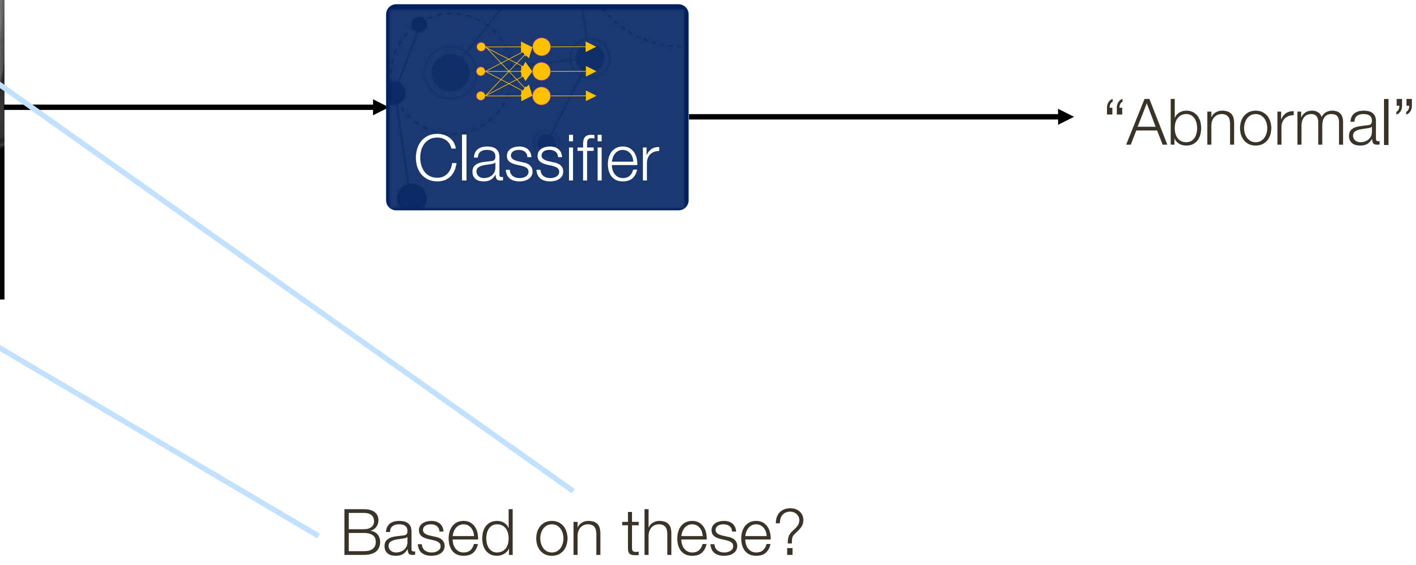


WHY INTERPRETABLE ML/ AI



WHY INTERPRETABLE ML/ AI





WHY INTERPRETABLE ML/ AI

Transparency

"Right to explanation"

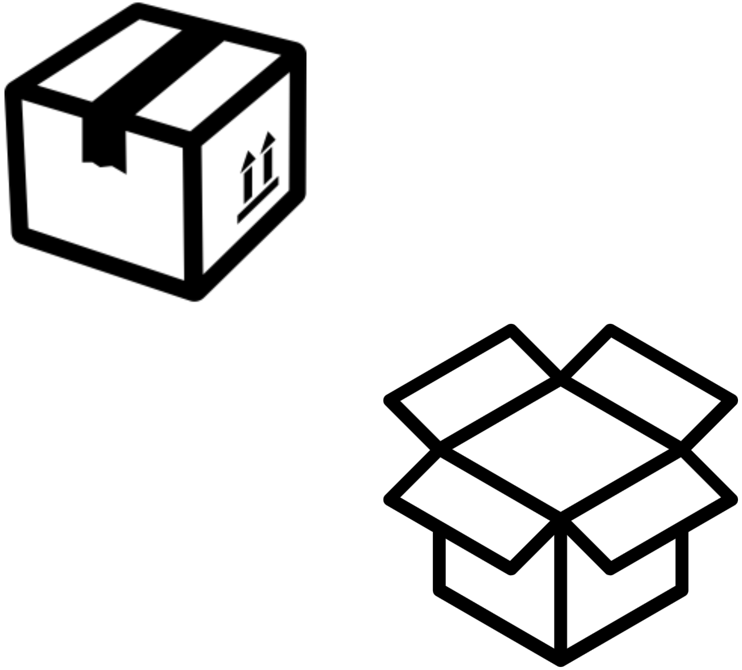
The data subject should have the right not to be subject to a decision […] which is based solely on automated processing *[…] such as automatic refusal of an online credit application without any human intervention.*
[…]
In any case, such processing should be subject to suitable safeguards, which should include ***the right to obtain human intervention,*** *to express his or her point of view,* ***to obtain an explanation of the decision reached***

(EU General Data Protection Regulation, Recital 71)

WHY INTERPRETABLE ML/ AI


Transparency


Understanding

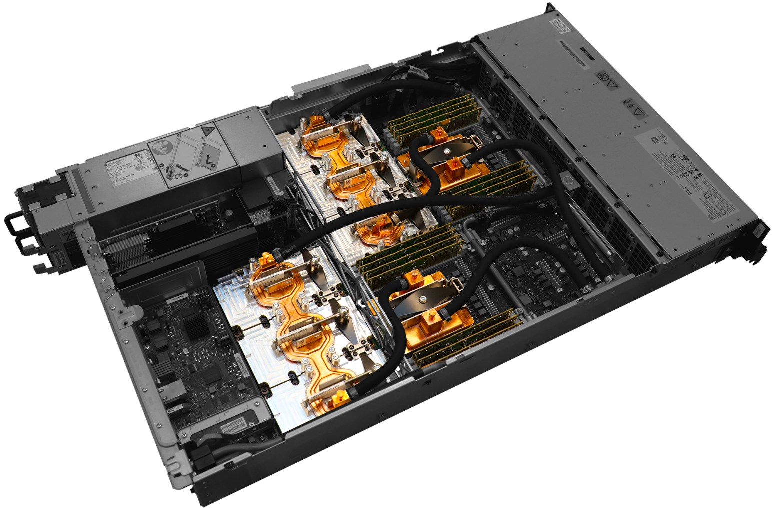


WHY INTERPRETABLE ML/ AI

Transparency

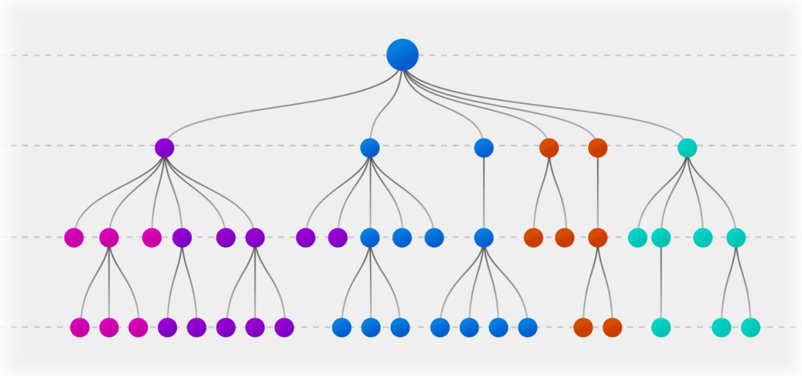
Understanding

Efficiency

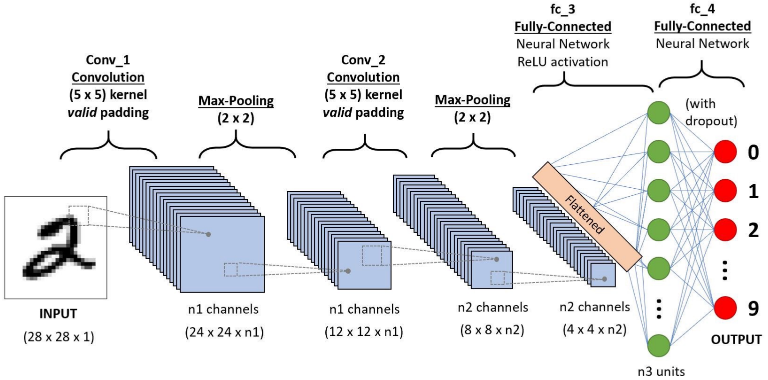


TYPES OF INTERPRETABILITY

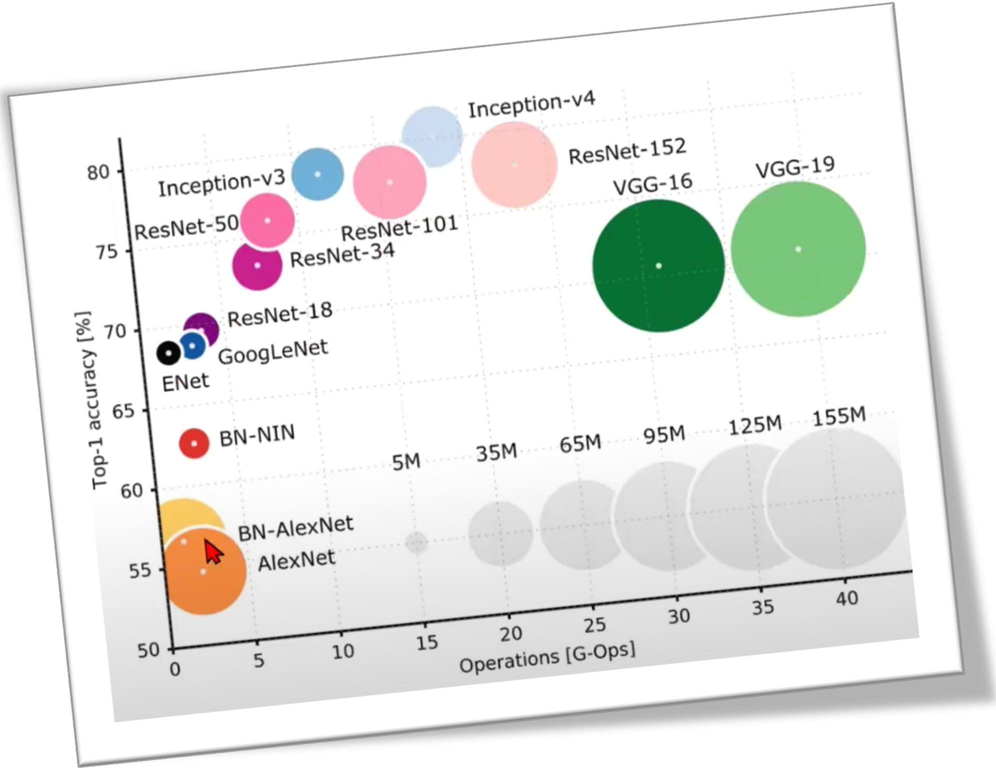
Interpretable by construction



Not directly interpretable



TYPES OF INTERPRETABILITY



How do you interpret millions of parameters?

TYPES OF INTERPRETABILITY

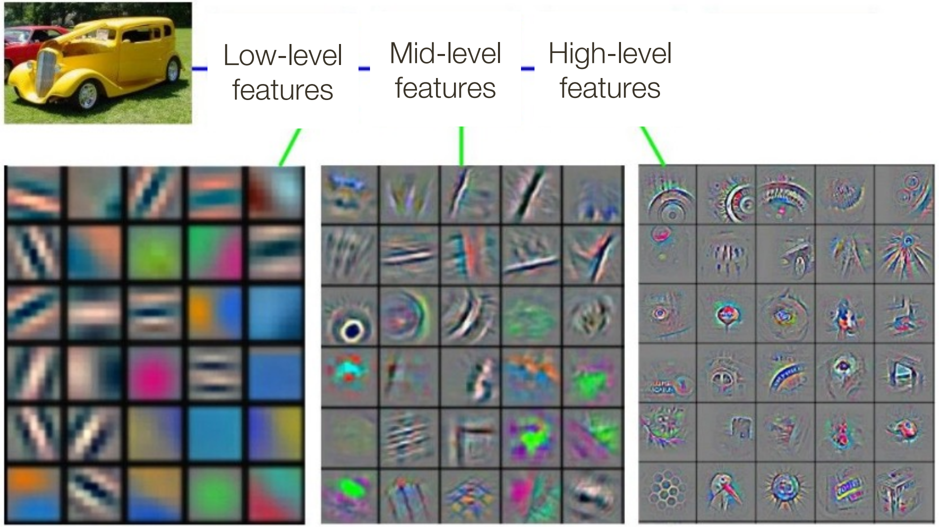
Model ⟷ Input/Output

TYPES OF INTERPRETABILITY

Model

Input/Output

What structures
are learned?



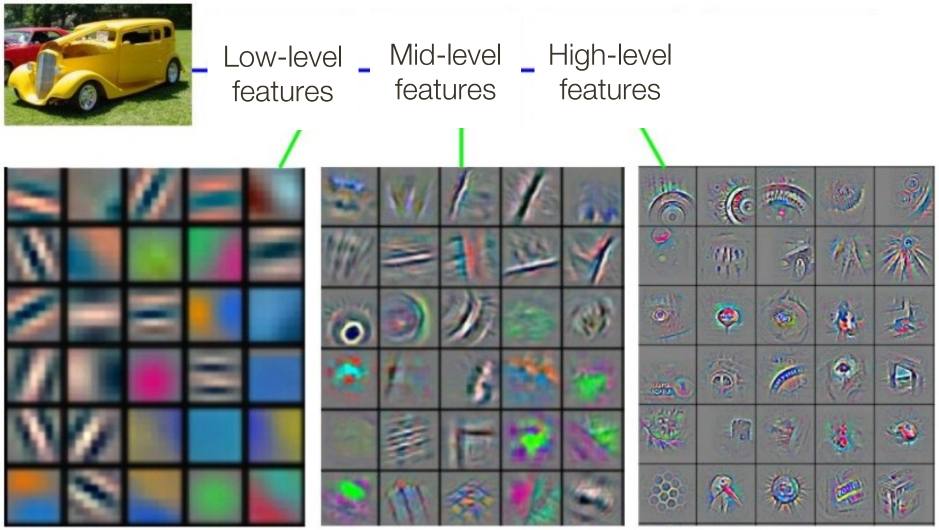
TYPES OF INTERPRETABILITY

Model

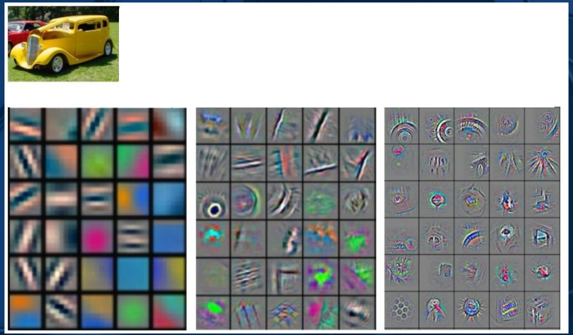
Input/Output

What structures are learned?

What parts are important for a given input/output?



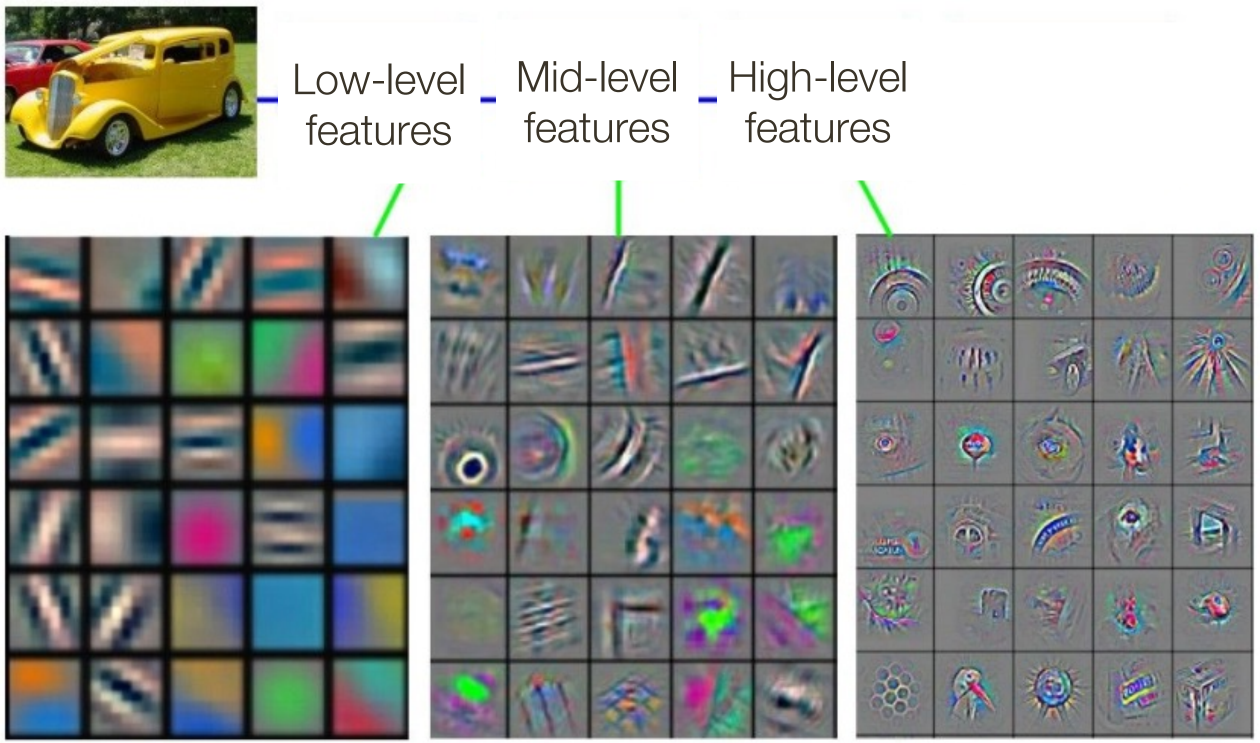
MODEL INTERPRETABILITY
GLOBAL UNDERSTANDING



INTERPRETING DEEP MODELS

Visualizing weights

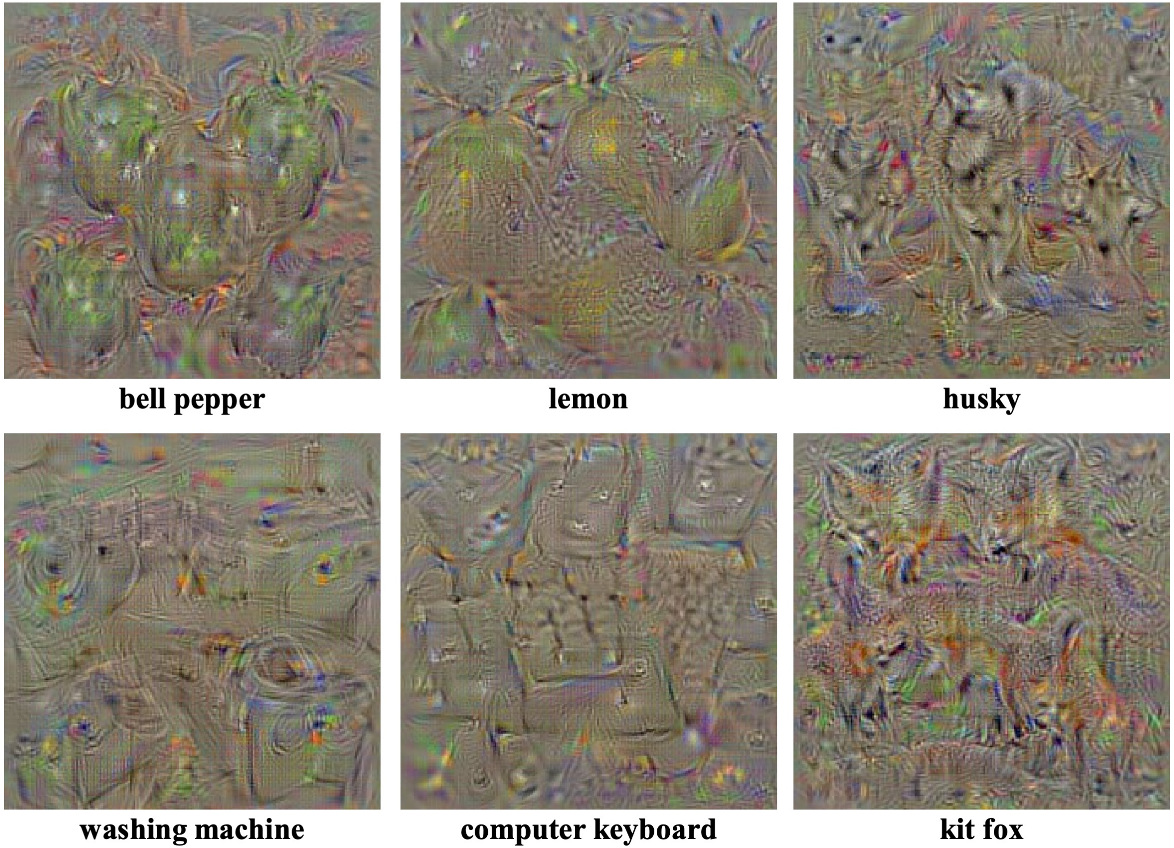
What weights/filters do the networks learn?



INTERPRETING DEEP MODELS

Activation patterns

Which patterns activate certain neurons most?

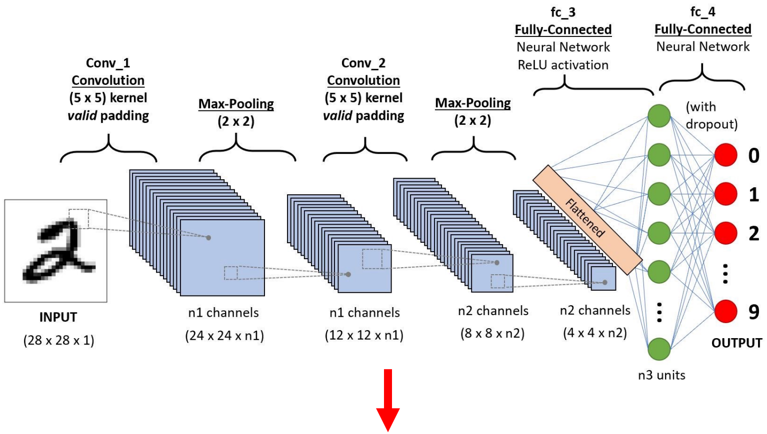


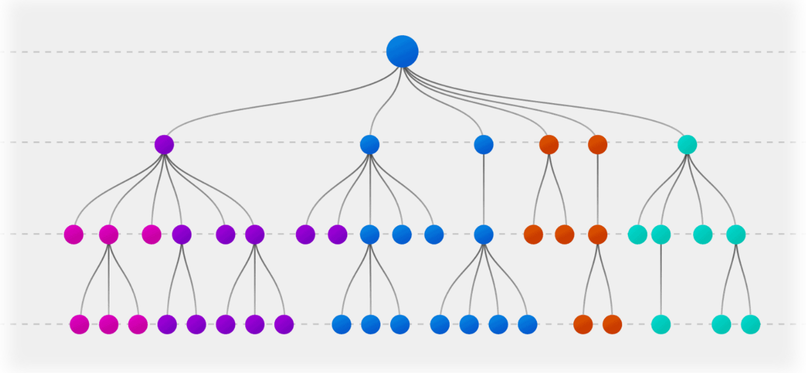
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

INTERPRETING DEEP MODELS

Surrogate models

Which is an interpretable model that generates similar results?





INTERPRETING DEEP MODELS

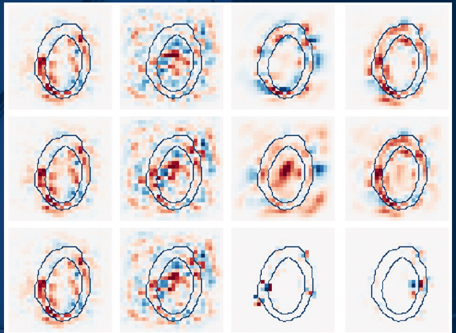
Influential data

Which data in the training set has influenced the decision most?

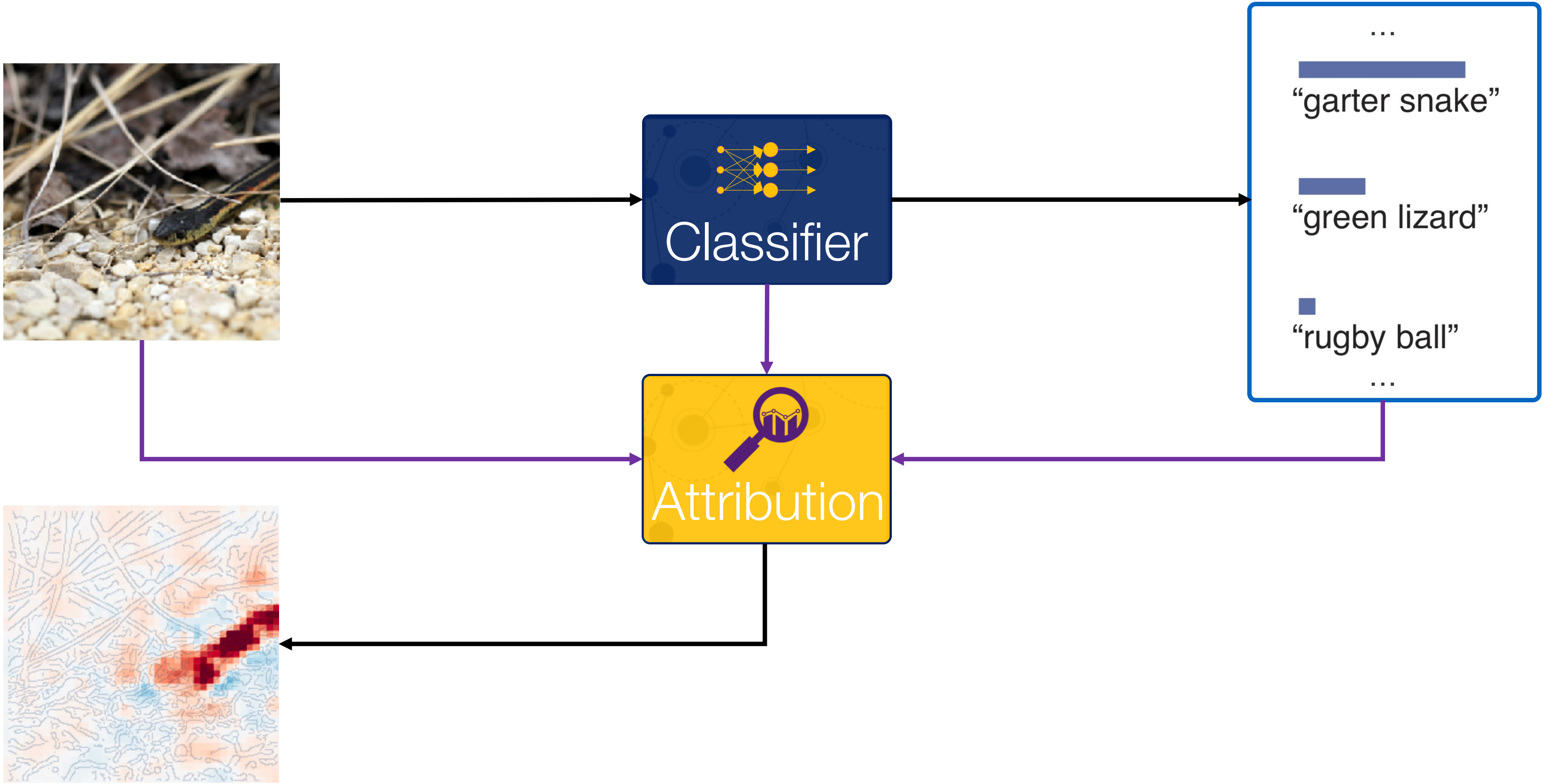
Test data

Training data

INPUT/OUTPUT INTERPRETABILITY
LOCAL UNDERSTANDING

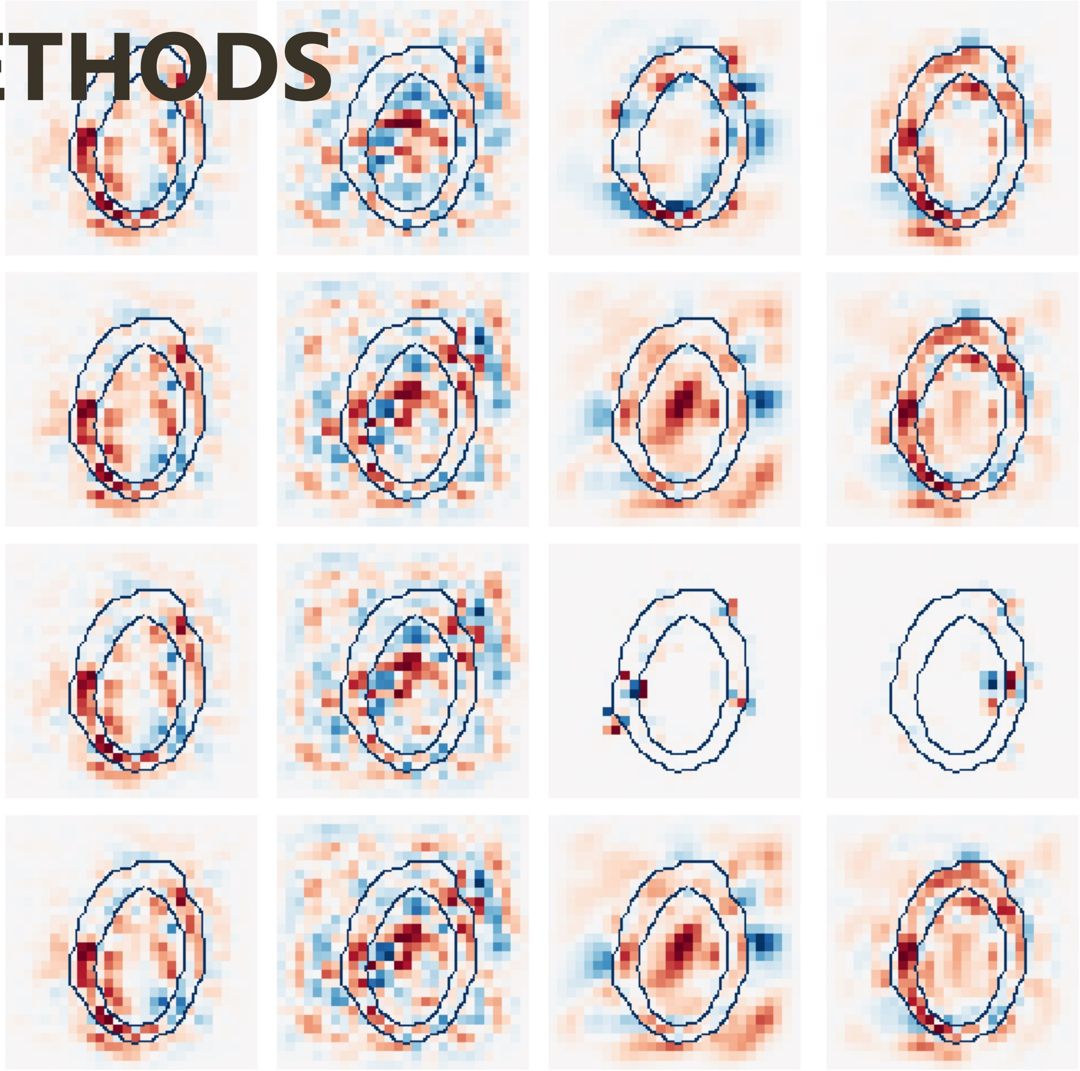


ATTRIBUTION METHODS



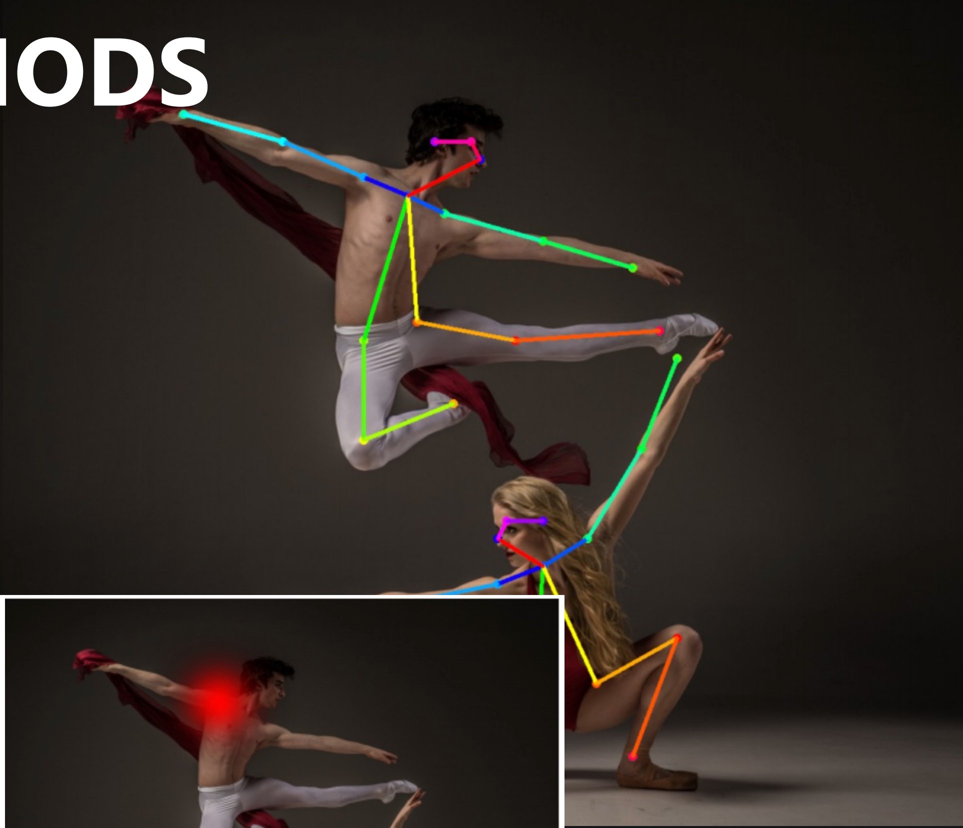
ATTRIBUTION METHODS

How to measure how much each pixel is important for a given input/output pair?



ATTRIBUTION METHODS

ATTRIBUTION METHODS



ATTRIBUTION METHODS

Desired properties

- Theoretically well-founded
- Implementation invariant
- Efficient to compute

ATTRIBUTION METHODS

Saliency Maps
Simonyan et al. 2015

Integrated Gradients
Sundararajan et al. 2017

DeepLIFT
Shrikumar et al. 2017

Deconvolutional Networks
Zeiler et al. 2014

Gradient * Input
Shrikumar et al. 2016

Layer-wise Relevance Propagation (LRP)
Bach et al. 2015

Guided Backpropagation
Springenberg et al. 2014

Grad-CAM
Selvaraju et al. 2016

Simple occlusion
Zeiler et al. 2014

Meaningful Perturbation
Fong et al. 2017

Prediction Difference Analysis
Zintgraf et al. 2017

...

ATTRIBUTION METHODS

Saliency Maps
Simonyan et al. 2015

Gradient * Input
Shrikumar et al. 2016

Simple occlusion
Zeiler et al. 2014

Unified framework

Integrated Gradients
Sundararajan et al. 2017

Layer-wise Relevance Propagation (LRP)
Bach et al. 2015

Meaningful Perturbation
Fong et al. 2017

DeepLIFT
Shrikumar et al. 2017

Guided Backpropagation
Springenberg et al. 2014

Shapley values

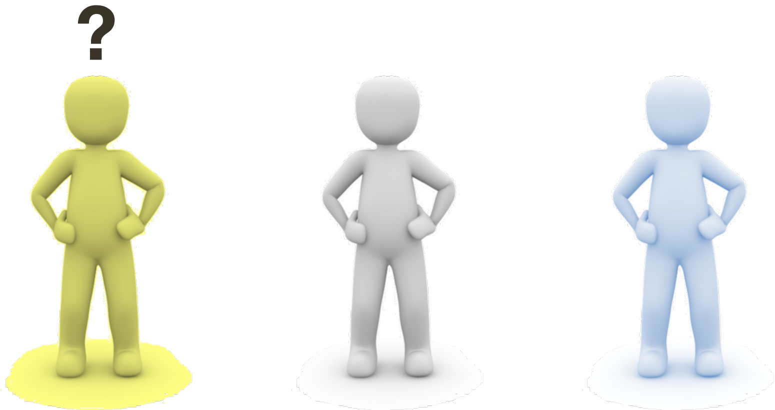
Prediction Difference Analysis
Zintgraf et al. 2017

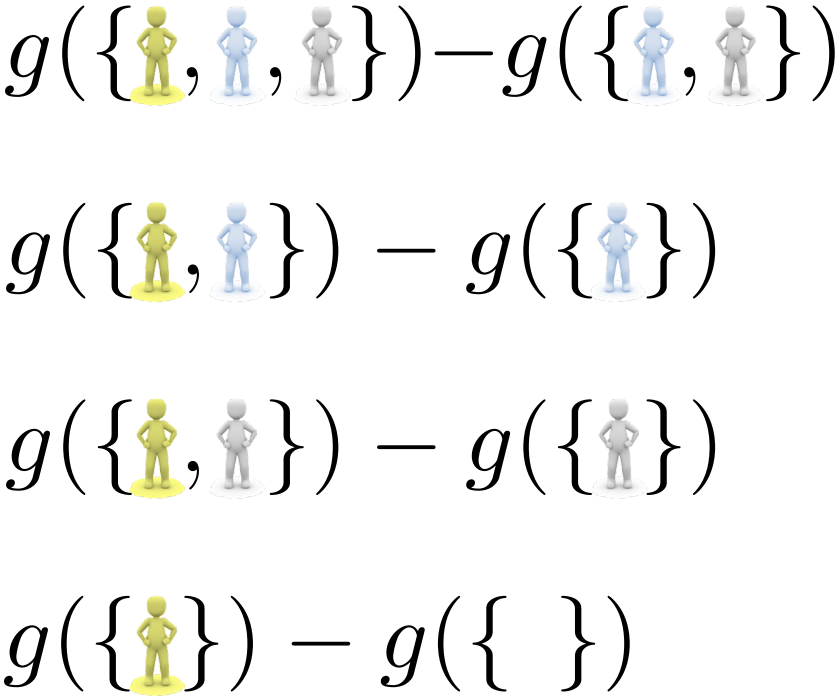
Deconvolutional Networks
Zeiler et al. 2014

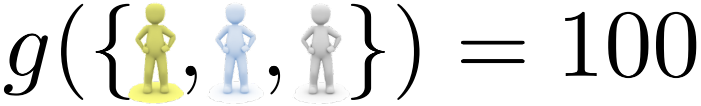
Grad-CAM
Selvaraju et al. 2016

…

ATTRIBUTION: SHAPLEY VALUES







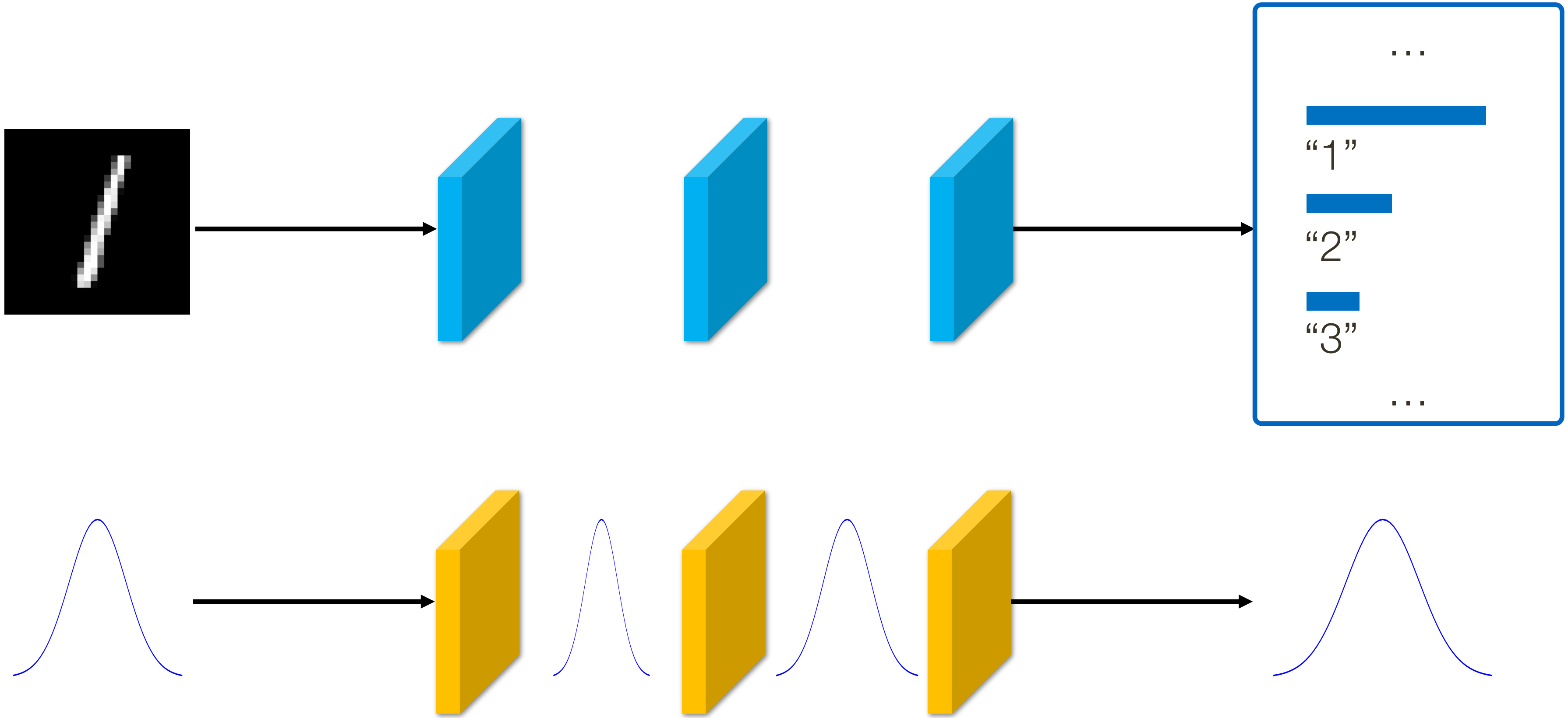
SHAPLEY VALUES

Desired properties

- Theoretically well-founded
- Implementation invariant
- ~~Efficient to compute~~

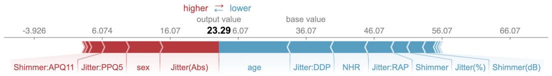
$$O(2^N)$$

ATTRIBUTION: DEEP SHAPLEY VALUES



ATTRIBUTION: DEEP SHAPLEY VALUES

DNA sequence classification



Parkinson’s disease factors

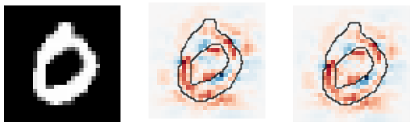


Image classification

APPLICATIONS OF INTERPRETABILITY

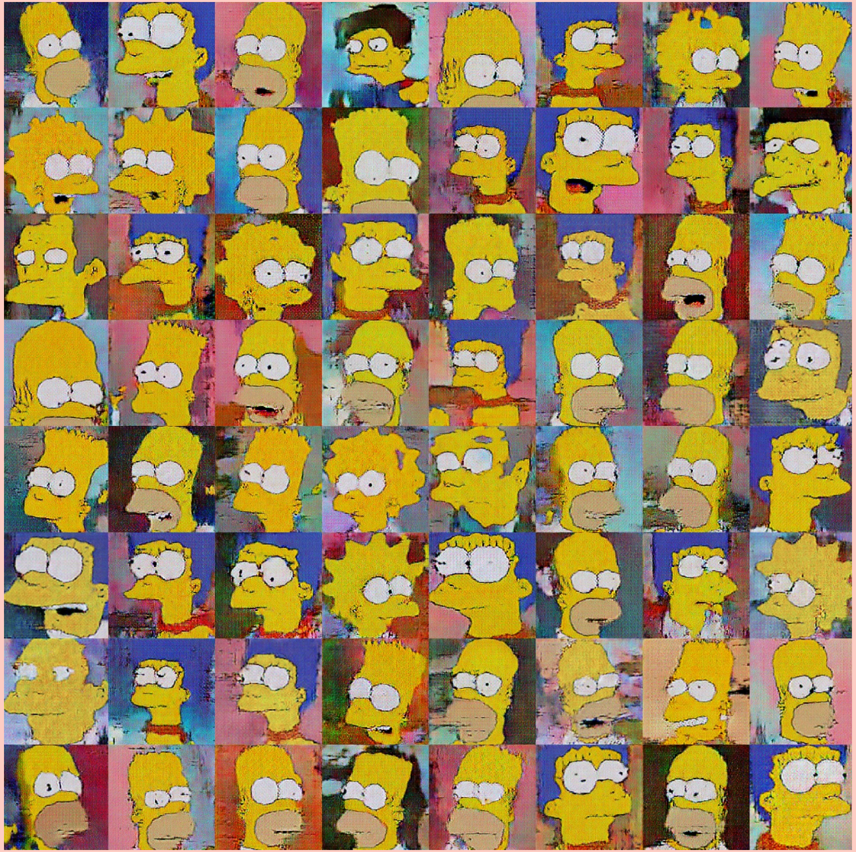
LIFE DECISIONS

ATTACKING DEEP SYSTEMS

Optimize for a t-shirt that makes you undetectable

DEEP SYSTEMS GONE WRONG

DEEP SYSTEMS GONE WRONG




DEEP ART

AVOID BIAS
UNDERSTAND WEAKNESSES
FAIL GRACEFULLY
SCIENCE NOT MAGIC
ENCOURAGE RIGOR
KEEP SANITY

DEEP ART

INTERPRETABLE DEEP LEARNING

Cengiz Öztireli