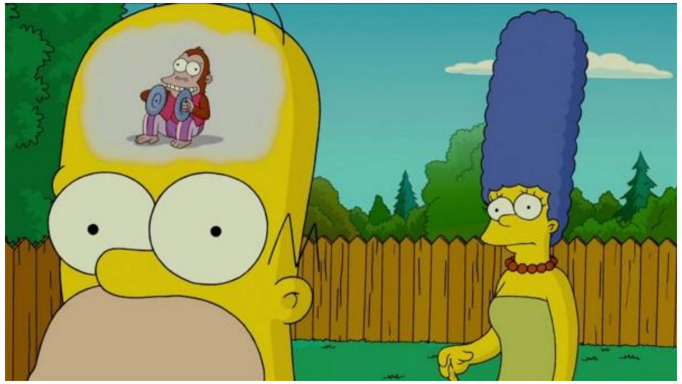
Interpretability of Deep Learning Models

Devinder Kumar
PhD Candidate, UWaterloo & Vector Inst. for AI
Lead AI Scientist in Residence, NextAI
July 12th, 2018

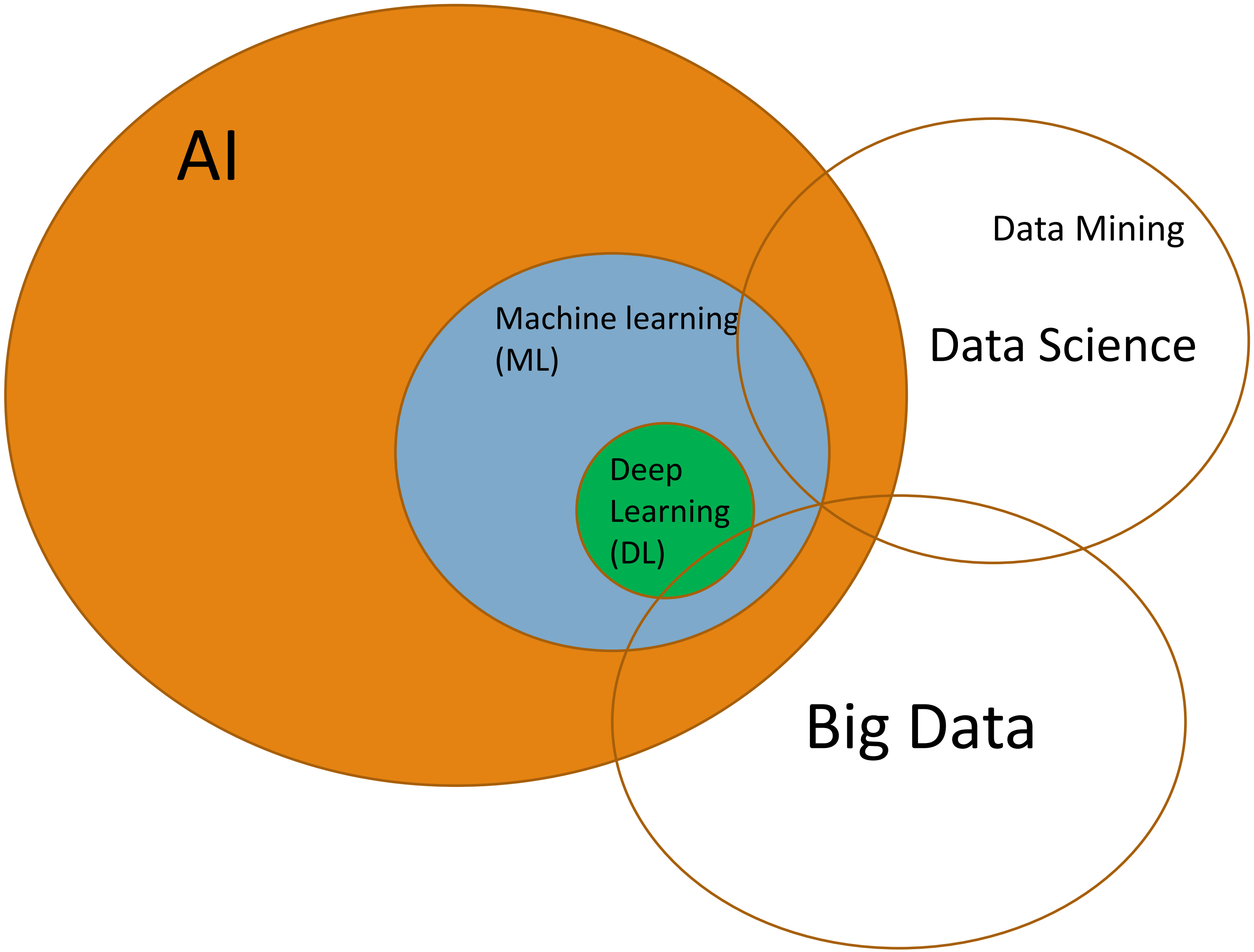
Today's Session Andrew Ng's words

- ~~Most of today's material is not very mathematical.~~

- Key ideas:
  - 1. What do we mean by Interpretability?
  - 2. Why do we need Interpretability?
  - 3. How it is useful?

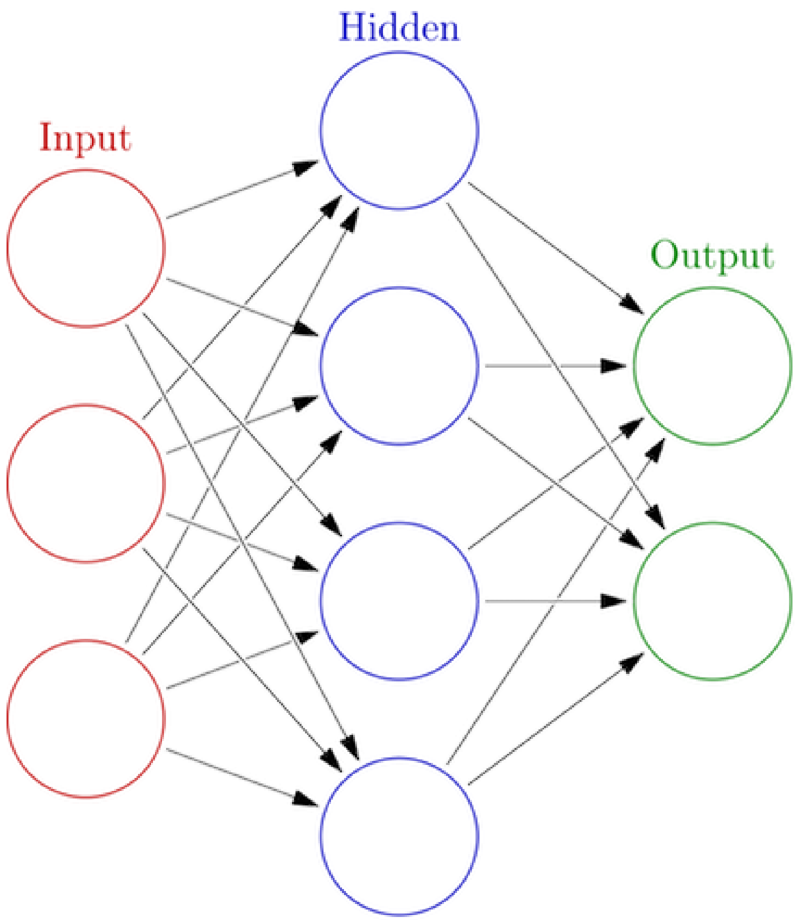


What is AI?


AI
Machine learning (ML)
Deep Learning (DL)
Data Mining
Data Science
Big Data

Evolution of Neural Networks

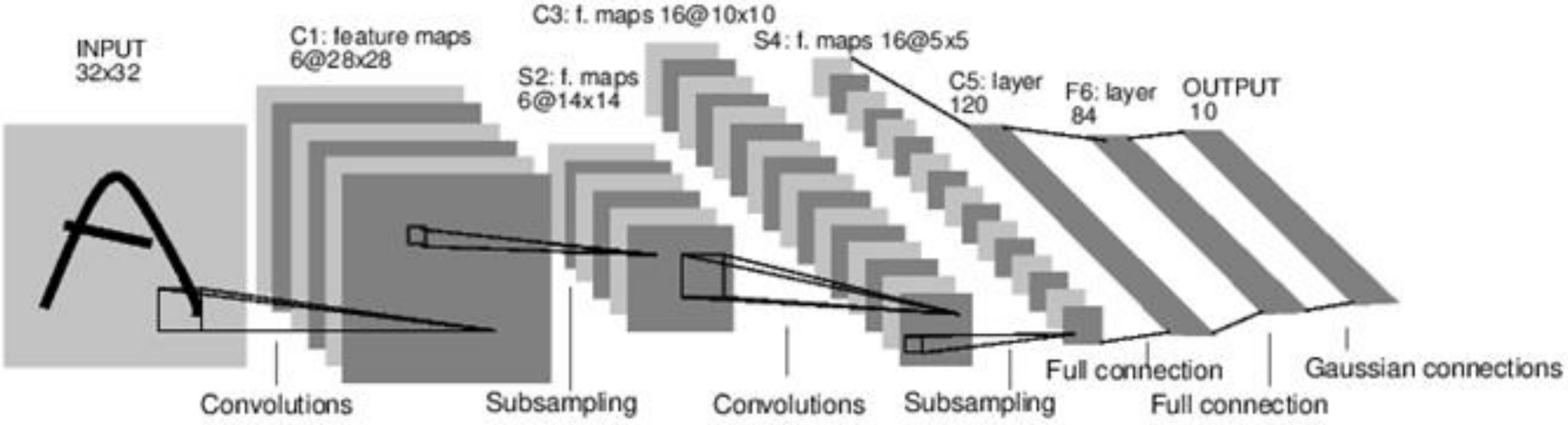
Simple Neural Net
1980s




KEEP
CALM
AND
GO
DEEP

Evolution of Convolutional Neural Networks!

LeNet 1989 (LIP-6, Paris)



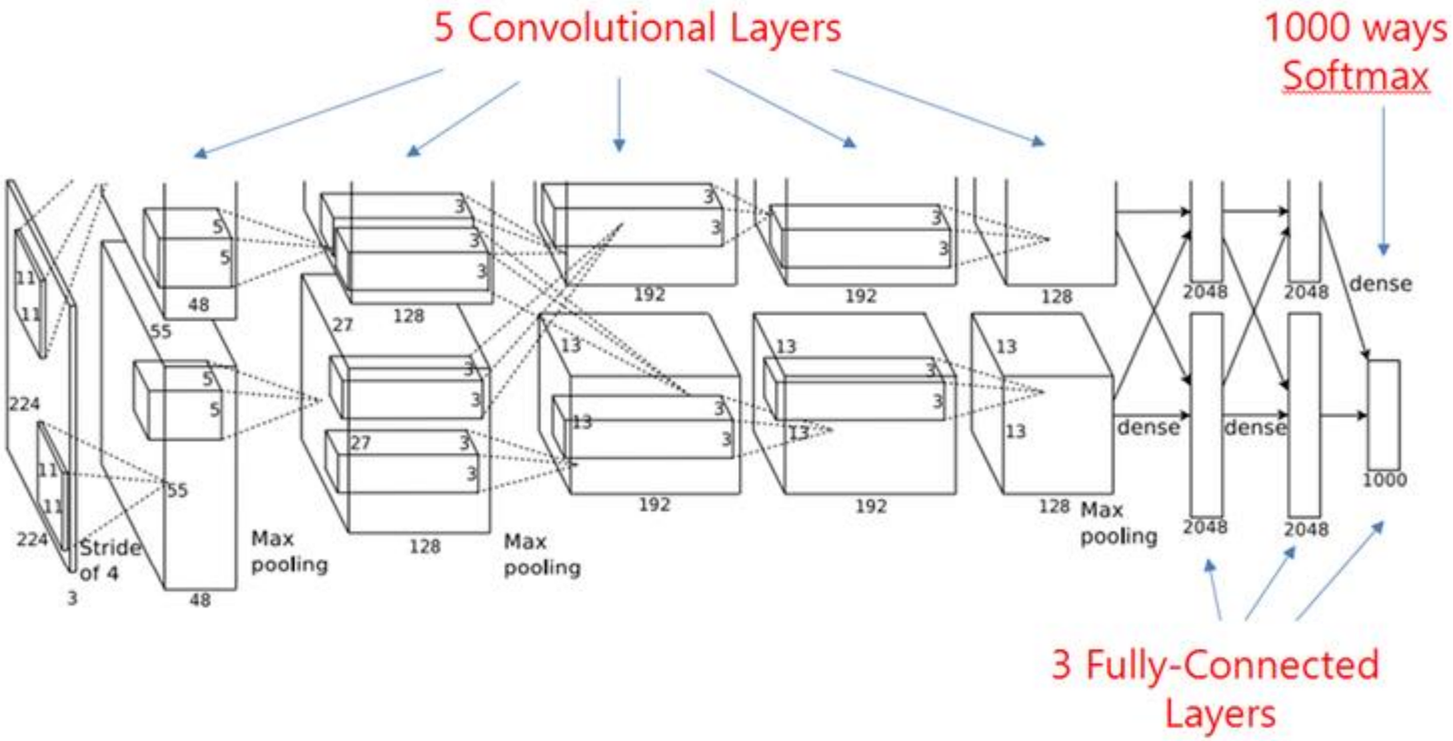
A Full Convolutional Neural Network (LeNet)


WE NEED TO GO
DEEPER

Evolution of Convolutional Neural Networks!

AlexNet 2012

60M
parameters

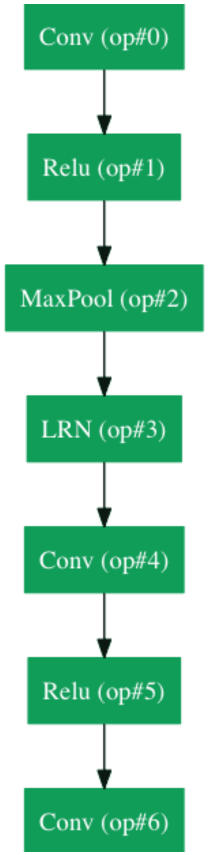



WE NEED TO GO DEEPER
CHALLENGE ACCEPTED
quickmeme.com

Current Models

GoogleNet / Inception, 2015/16

6.9 M parameters in the model



Future AI : Promises

Future AI : Reality

Scalable Oversight: How can we efficiently ensure that a given AI system respects aspects of the objective that are too expensive to be frequently evaluated during training?

- Google on its challenges for AI, Dailymail UK, June 2016

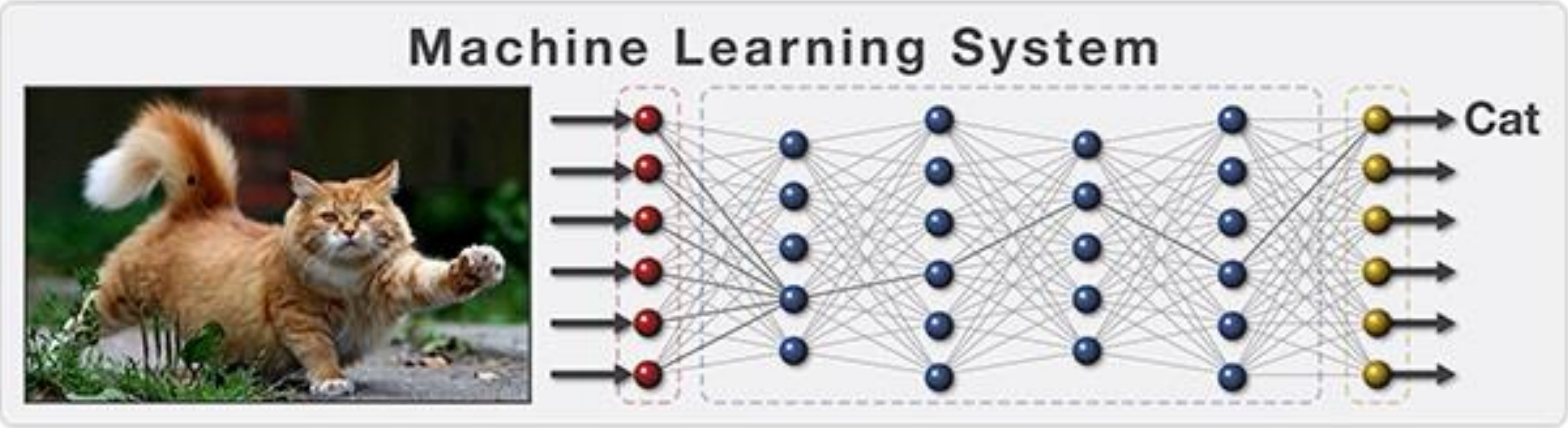
"There is no neural network in the world, and no method right now that can be trained to identify objects and images, play Space Invaders, and listen to music."

- R.Hadsel (Google DeepMind), The Verge, Oct 2016

"We can build these models, but we don't know how they work."

- Deep Patient , MIT Review (April,2017)

Future AI: Explainable



Input Data → Black box AI Model → Cat

Via XAI: DARPA

Why do we care about X-AI?

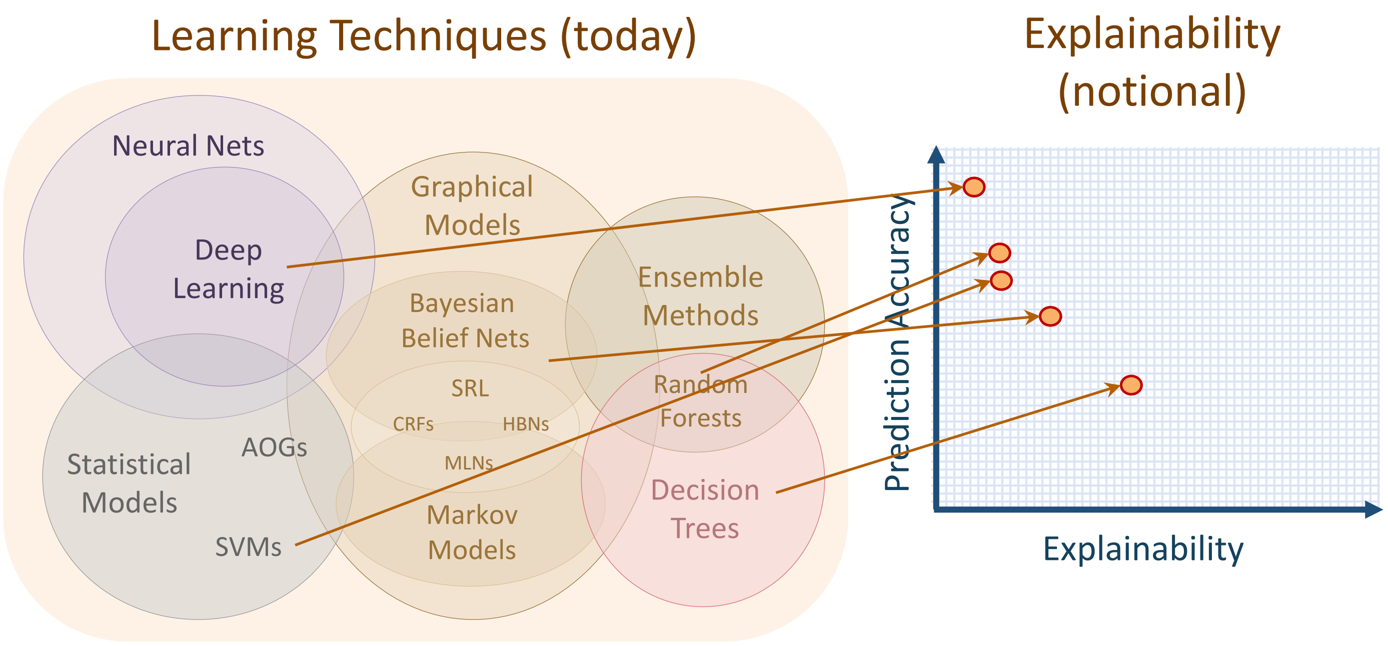
Explainability

Photo credit: oragesky3

Explainable Models

New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

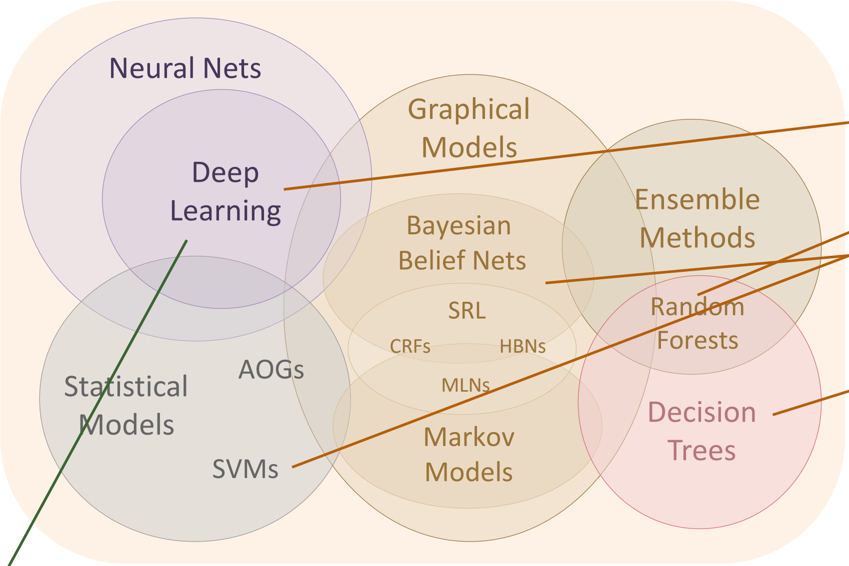


Explainable Models

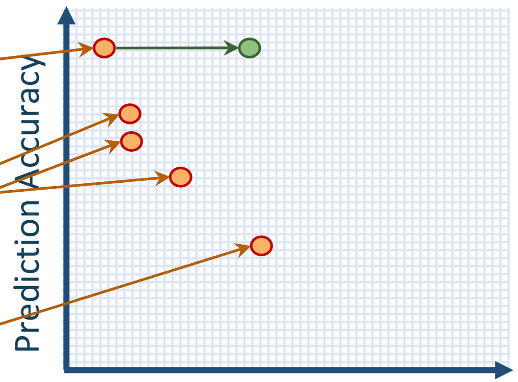
New Approach

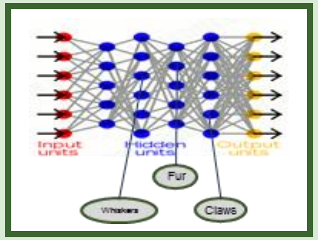
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)





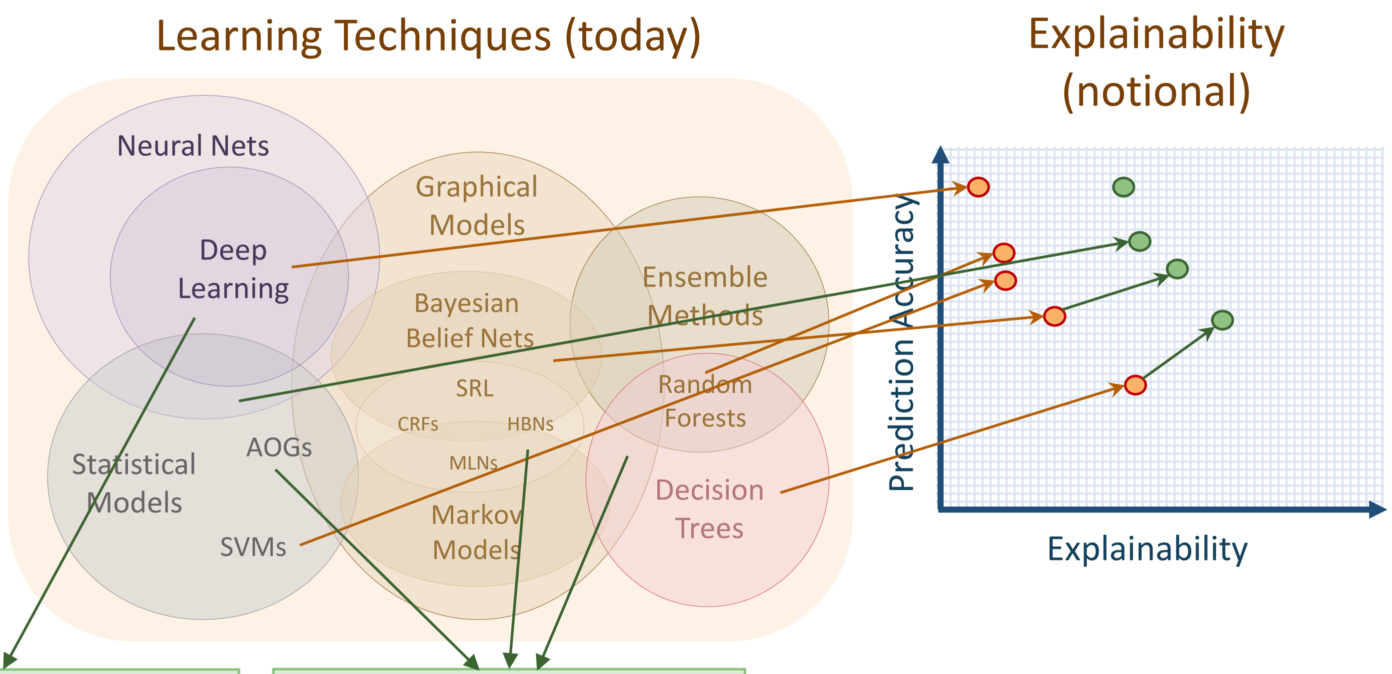
Deep Explanation

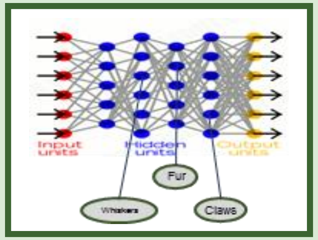
Modified deep learning techniques to learn explainable features

Explainable Models

New Approach

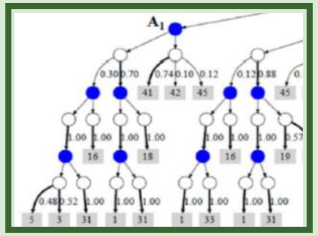
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance





Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models

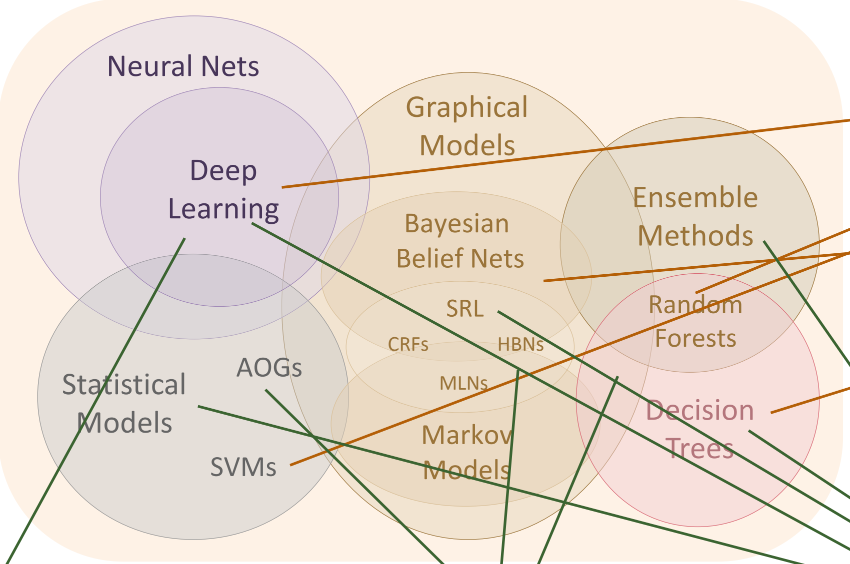
via DARPA

Explainable Models

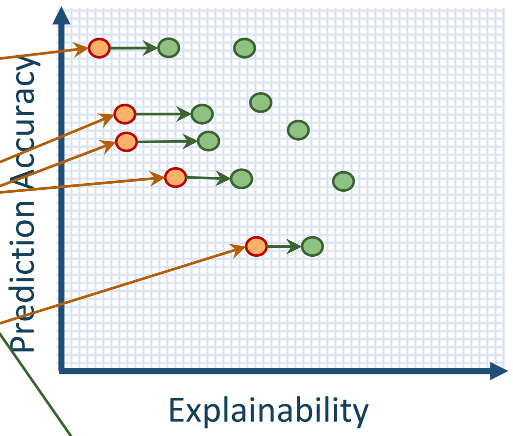
New Approach

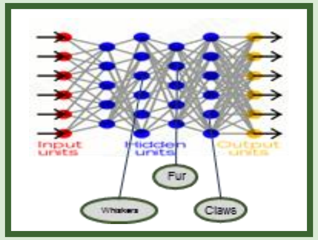
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



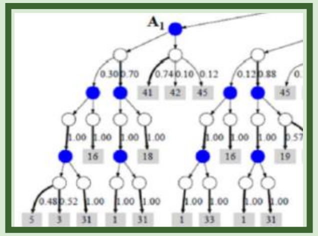
Explainability (notional)





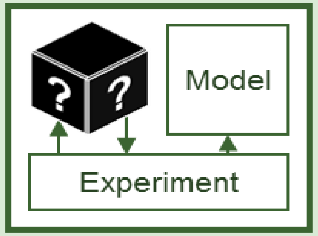
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

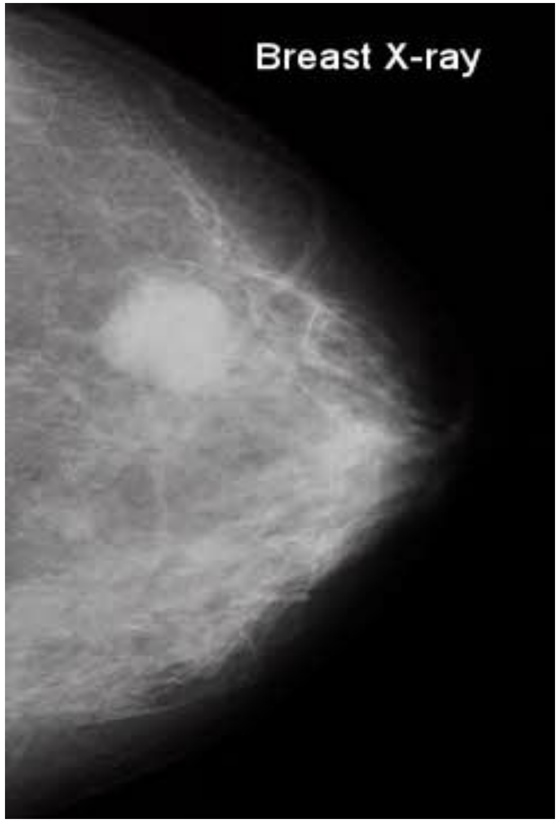
Techniques to learn more structured, interpretable, causal models

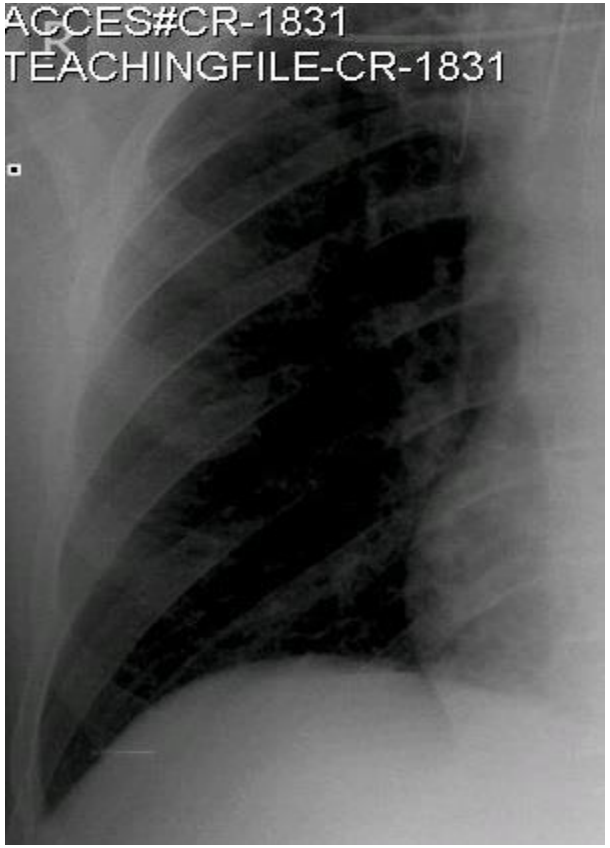


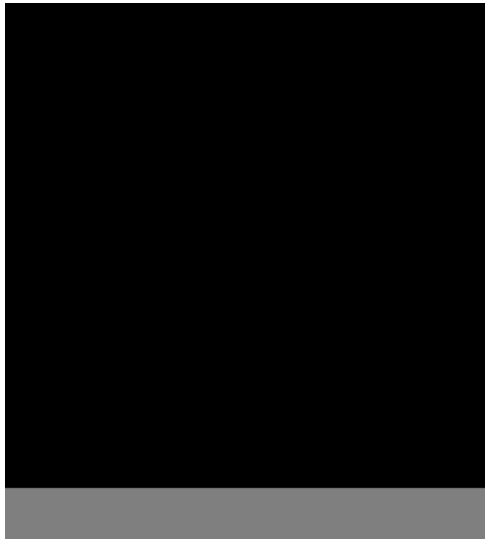
Model Induction

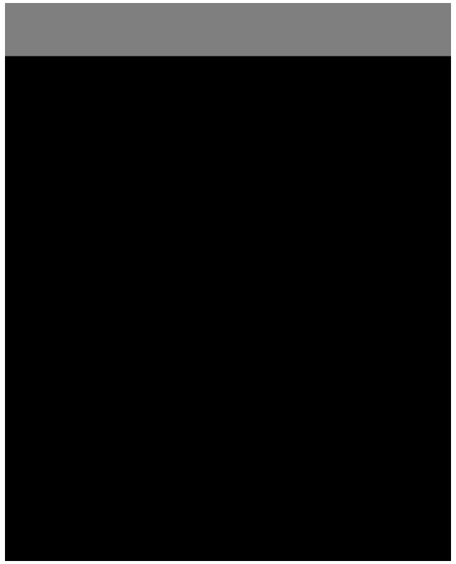
Techniques to infer an explainable model from any model as a black box

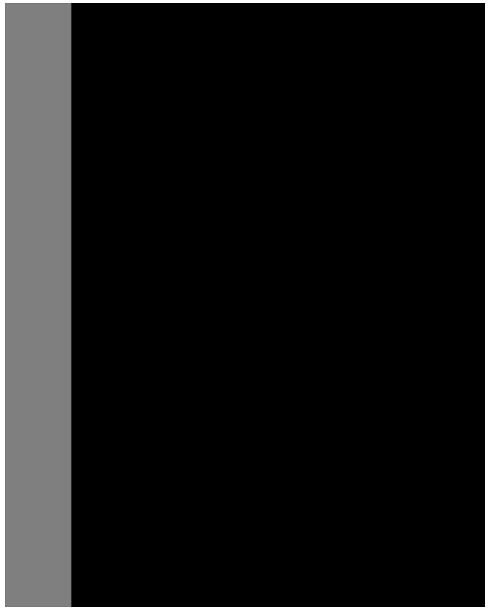
What is Happening?

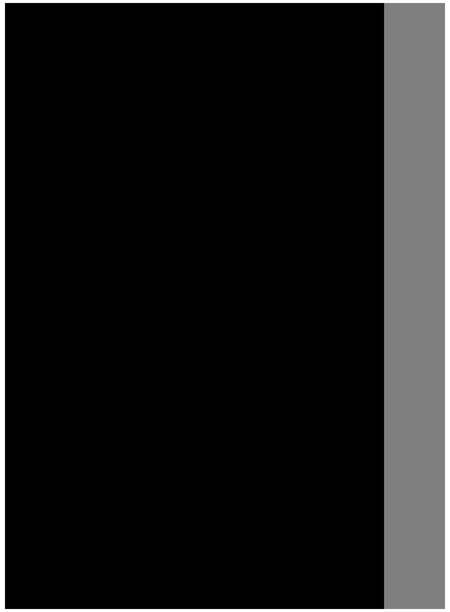








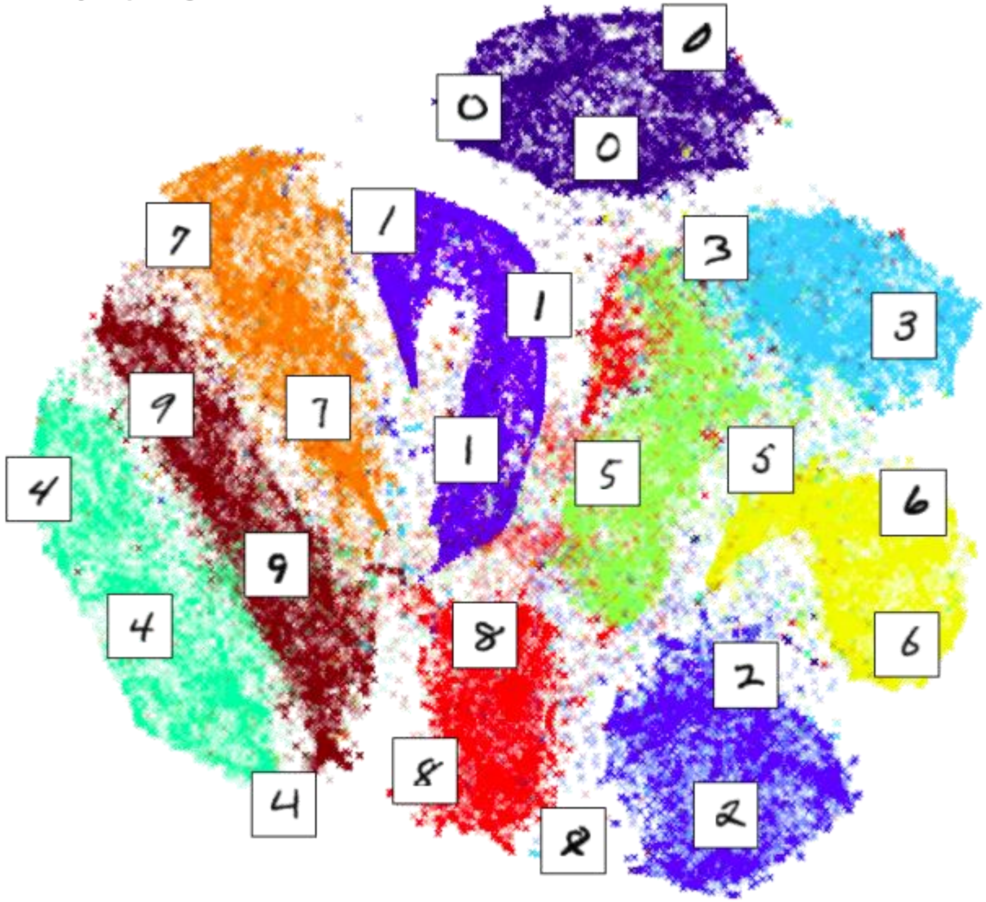




General Data Protection Regulation (GDPR)

“more and clearer information about processing”

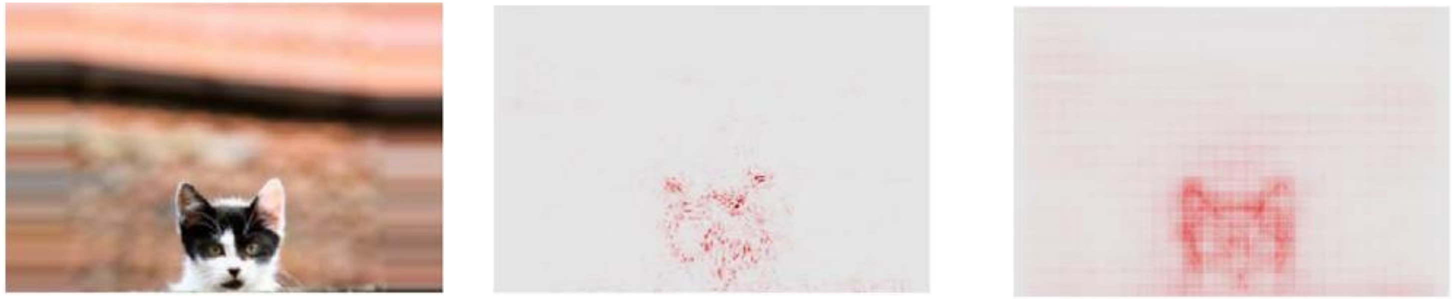
T-SNE: Visualization!



Input Data → Black box AI Model → Cat

Deconvolution: Zeiler et.al. ECCV 14

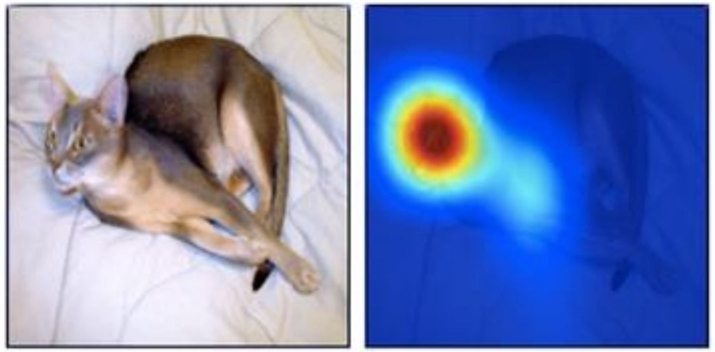
Guided backpropagation: ICLR 2015



Saliency: Simonyan et.al.
CVPR 2013

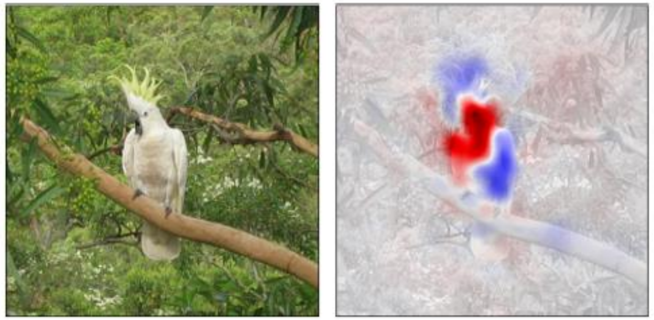
Deep Taylor Decomp.
Montavon et. al.
PR journal 2017

Input Data → Black box AI Model → Cat



Class Activation Maps (CAM) Bolei Zhou MIT, 2016

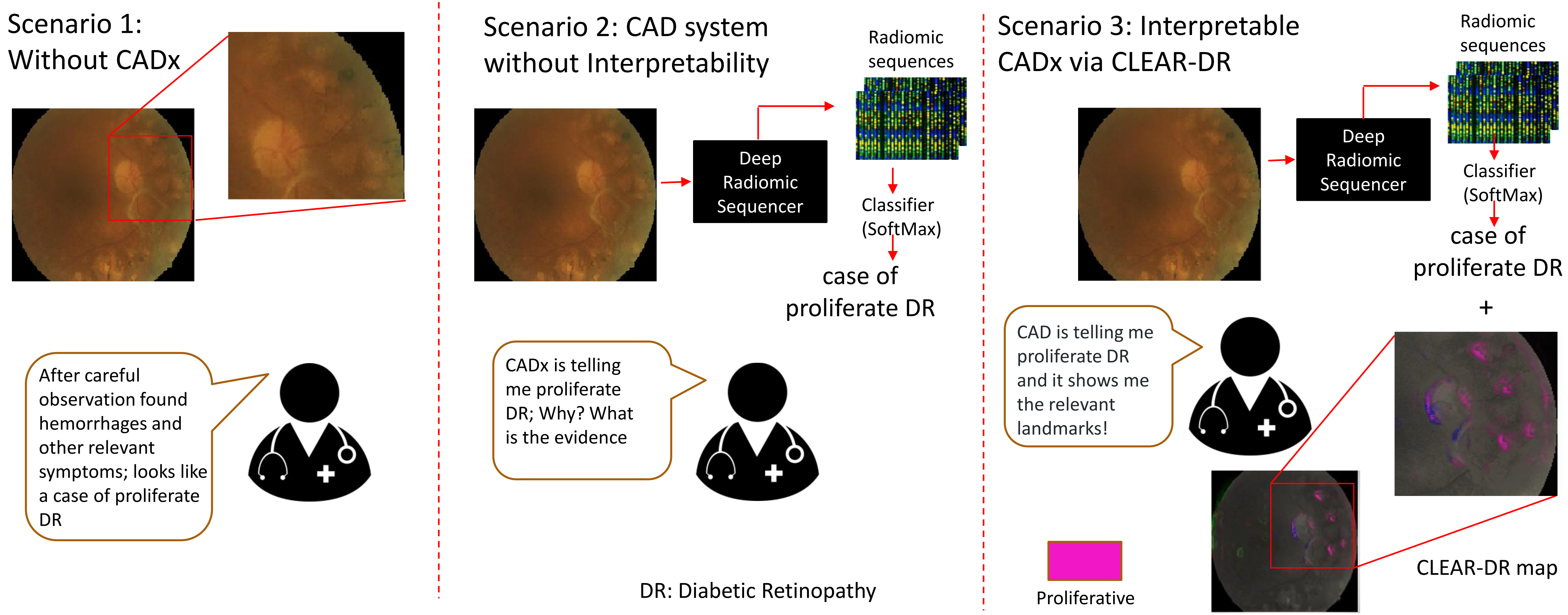
Input Data → Black box AI Model → Cat



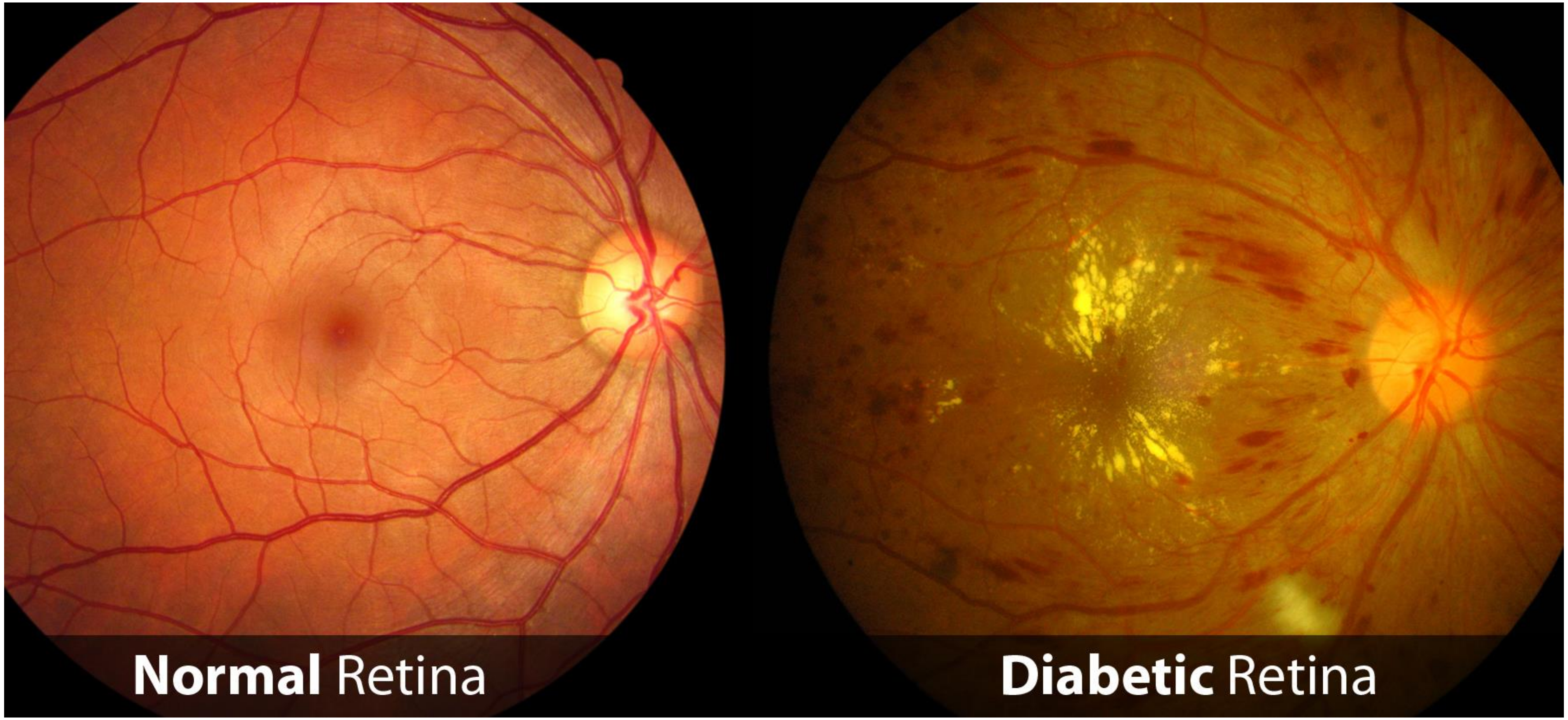
Prediction Difference:
Zintgraf et. al. ICLR 2017

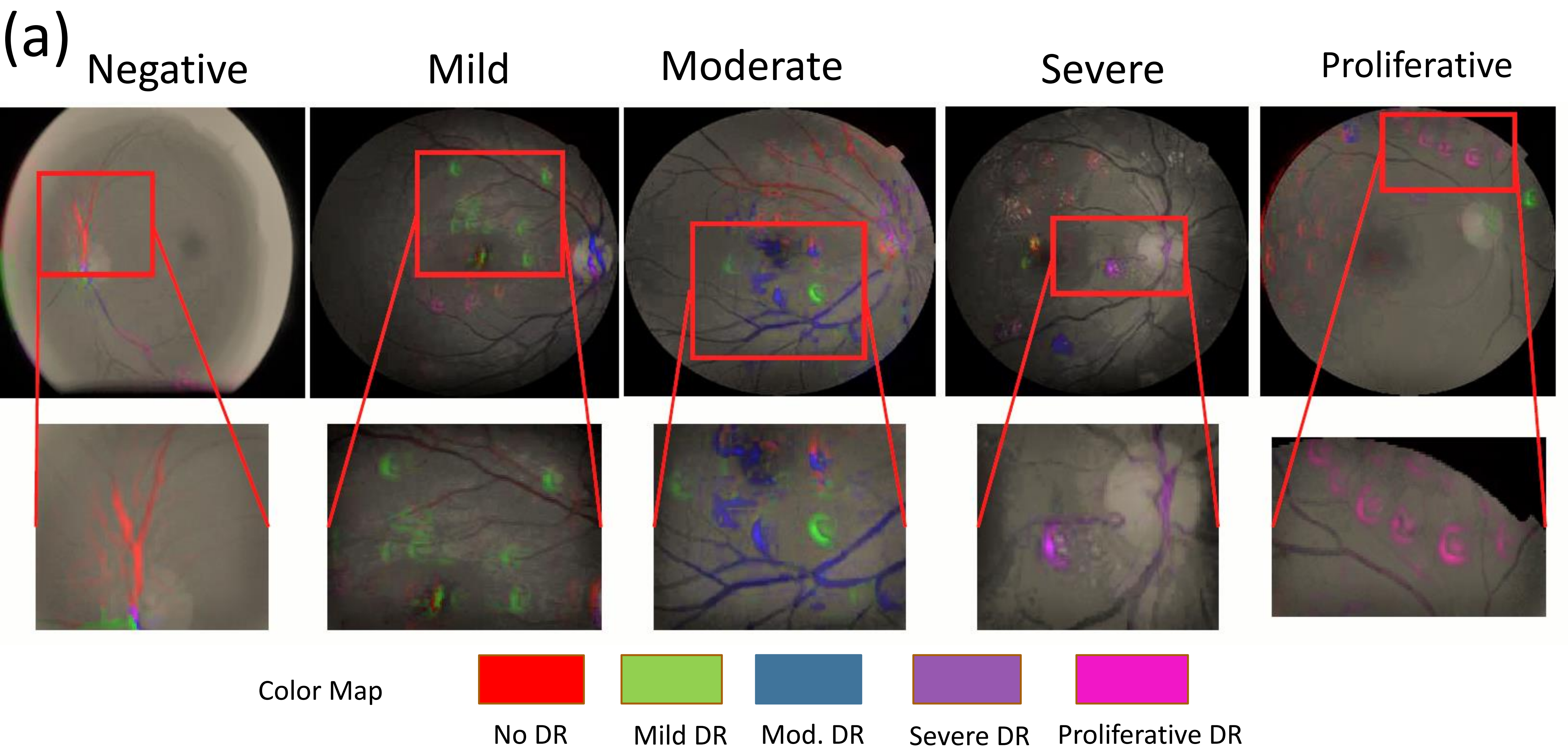
Stanford Dog Dataset Results

Application in Healthcare


Scenario 1: Without CADx
After careful observation found hemorrhages and other relevant symptoms; looks like a case of proliferate DR
Scenario 2: CAD system without Interpretability
Deep Radiomic Sequencer
Radiomic sequences
Classifier (SoftMax)
case of proliferate DR
CADx is telling me proliferate DR; Why? What is the evidence
DR: Diabetic Retinopathy
Scenario 3: Interpretable CADx via CLEAR-DR
Deep Radiomic Sequencer
Radiomic sequences
Classifier (SoftMax)
case of proliferate DR
+
CAD is telling me proliferate DR and it shows me the relevant landmarks!
Proliferative
CLEAR-DR map

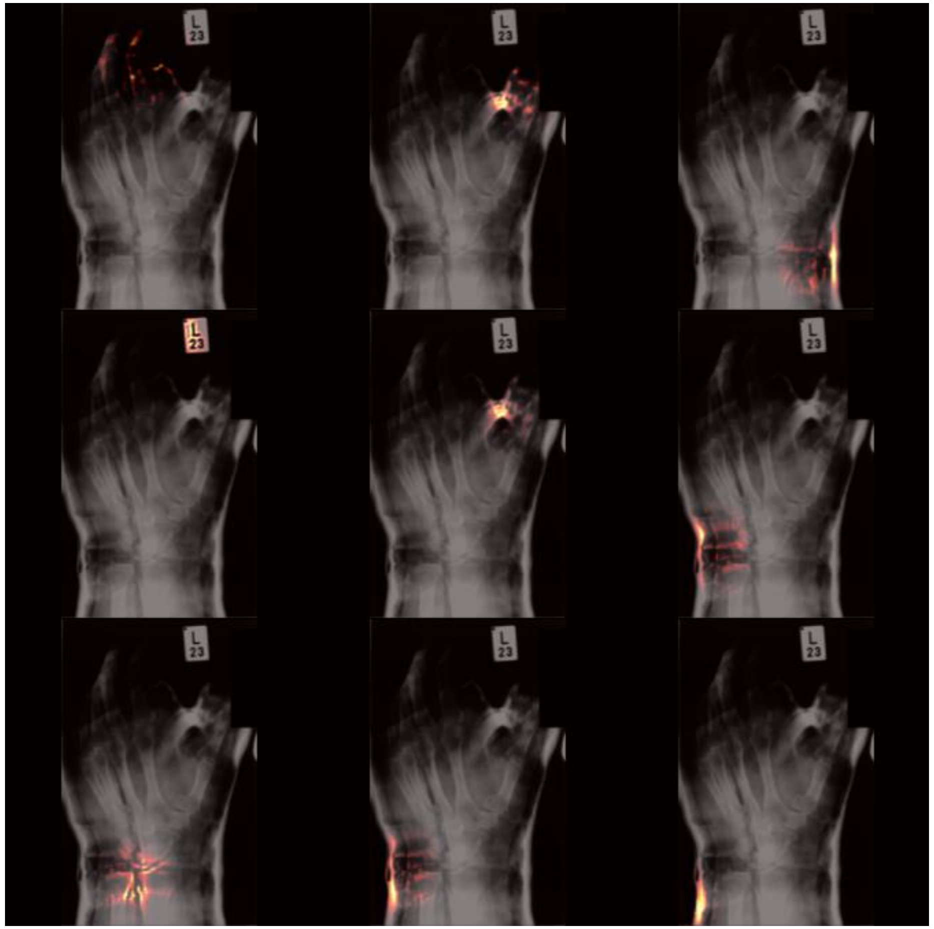
Diabetic Retinopathy:
Leading Cause of Blindness in the world



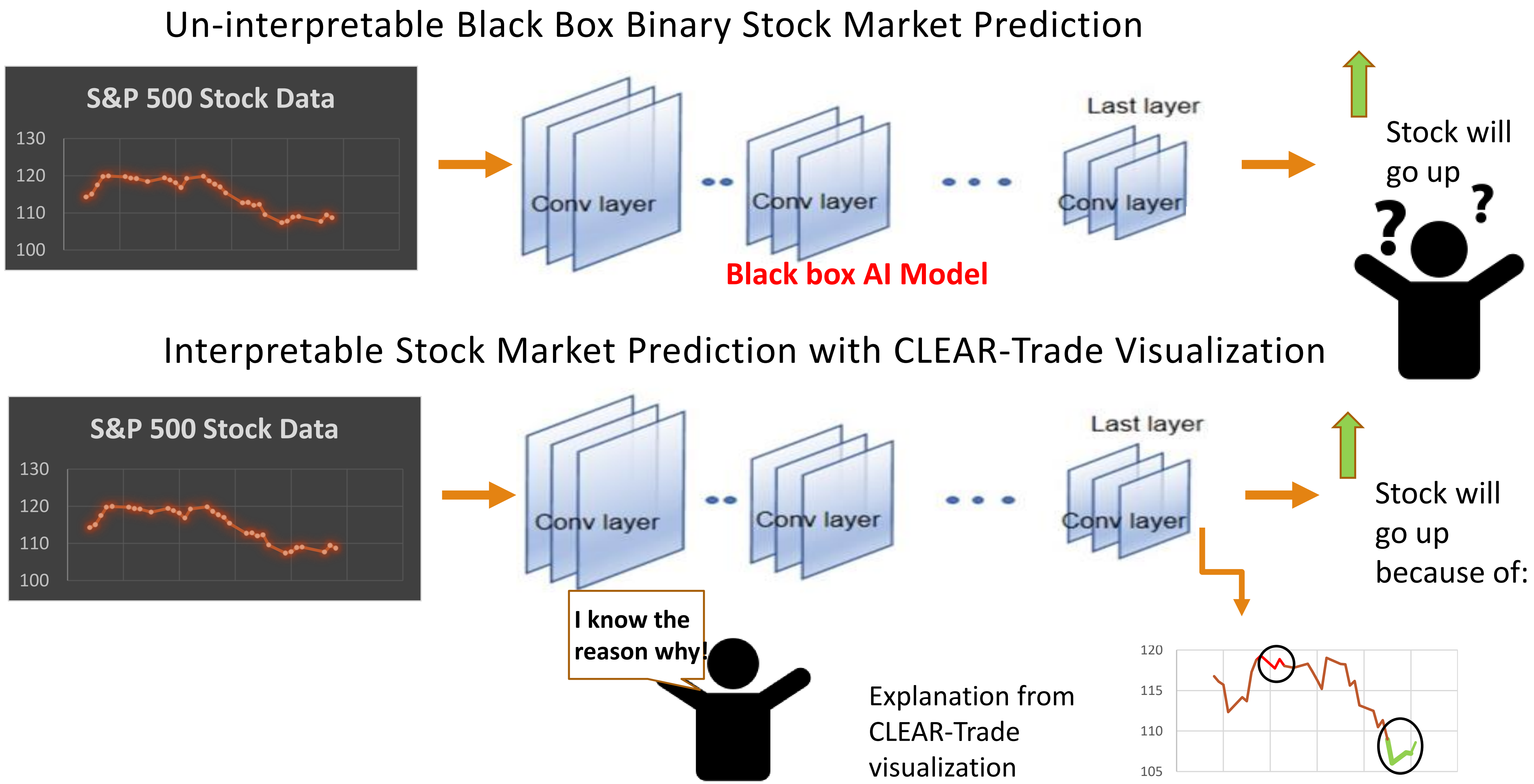

(a)
Negative
Mild
Moderate
Severe
Proliferative
Color Map
No DR
Mild DR
Mod. DR
Severe DR
Proliferative DR


Distal Phalanges
R
Interphalangeal Joint
Proximal Phalanges (1-5)
Metatarso-phalangeal Joints
Sesamoid Bones
Metatarsals (I-V)
Cuneiforms
- Medial
- Intermediate
- Lateral
Navicular
Talus
Tibia
Intermediate Phalanges (A-D)
Head of 5th Metatarsal
Shaft of 5th Metatarsal
Base of 5th Metatarsal
Cuboid
Calcaneum
Fibula

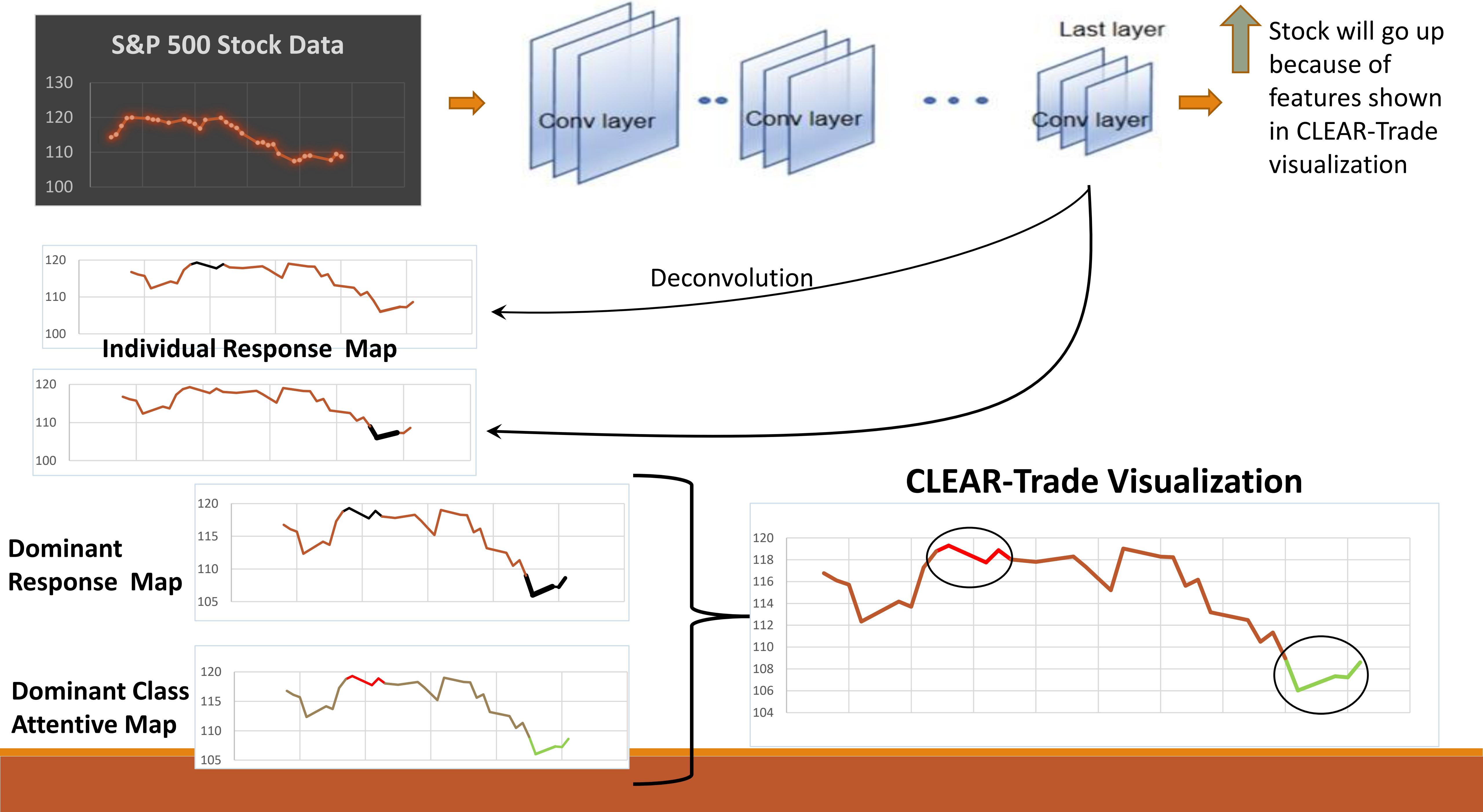

L
wiki
Distal Interphalangeal Joint
Proximal Interphalangeal Joint
Metacarpo-phalangeal Joint
Distal Phalanx
Middle Phalanx
Phalanges
Proximal Phalanx
Metacarpals
Hamate
Capitate
Triquetrium
Pisiform
Lunate
Ulna Styloid Process
Ulna
Trapezoid
Trapezium
Scaphoid
Radial Styloid Process
Radius
Carpals



Application in Finance

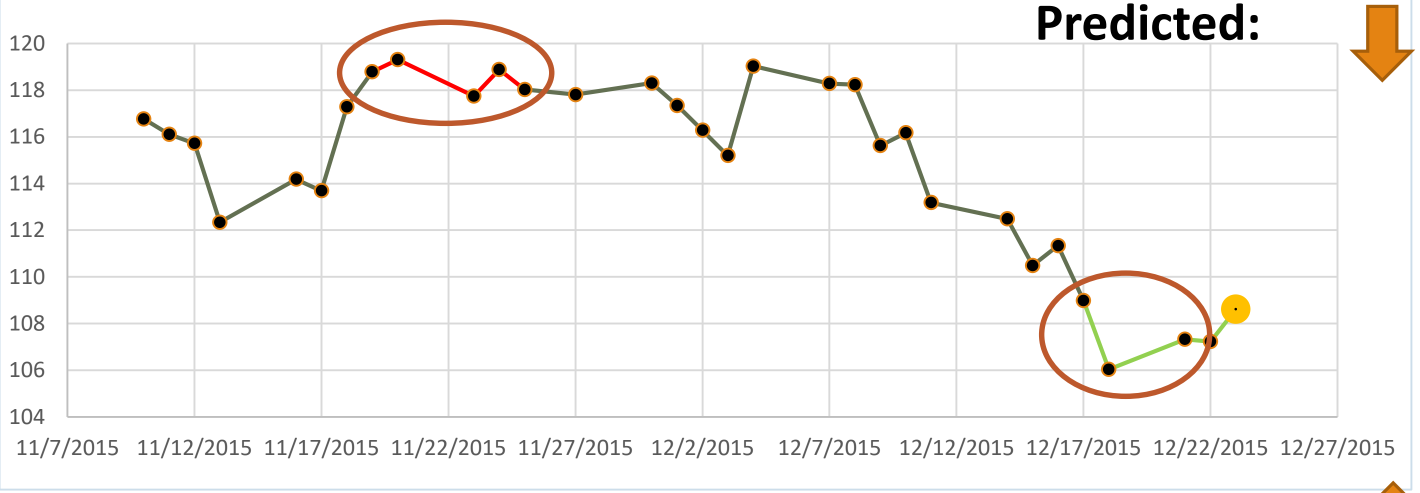


Un-interpretable Black Box Binary Stock Market Prediction
S&P 500 Stock Data
130
120
110
100
Conv layer
Conv layer
Last layer
Conv layer
Black box AI Model
Stock will go up
Interpretable Stock Market Prediction with CLEAR-Trade Visualization
S&P 500 Stock Data
130
120
110
100
Conv layer
Conv layer
Last layer
Conv layer
Stock will go up because of:
I know the reason why!
Explanation from CLEAR-Trade visualization
120
115
110
105

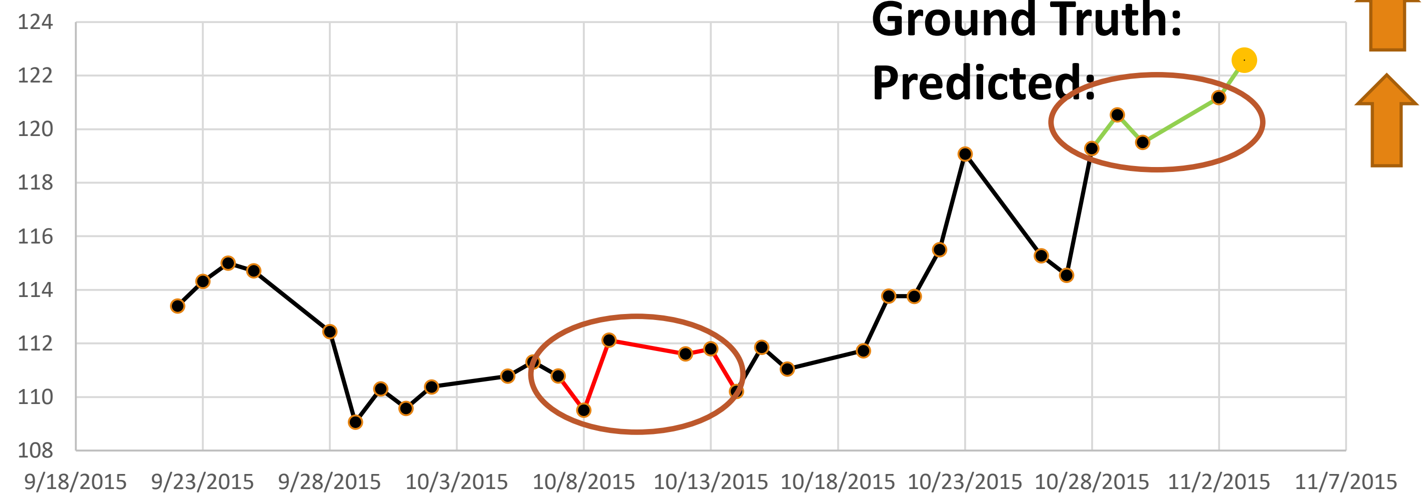

S&P 500 Stock Data
130
120
110
100
Conv layer
Conv layer
Last layer
Conv layer
Stock will go up because of features shown in CLEAR-Trade visualization
Deconvolution
120
110
100
Individual Response Map
120
110
100
Dominant Response Map
120
115
110
105
Dominant Class Attentive Map
120
115
110
105
CLEAR-Trade Visualization
120
118
116
114
112
110
108
106
104

Correct Binary Stock Prediction

Ground Truth:
Predicted:



Ground Truth:
Predicted:

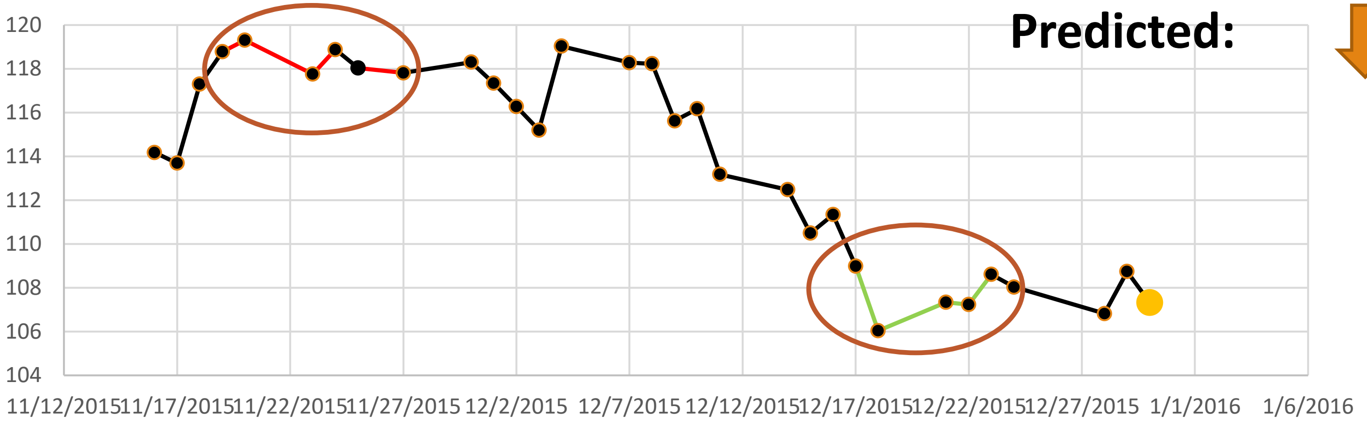


Date	Open	High	Low	Close	Volume
12/23/2015	107.27	108.85	107.2	108.61	3.3E+07
12/22/2015	107.4	107.72	106.45	107.23	3.3E+07
12/21/2015	107.28	107.37	105.57	107.33	4.8E+07
12/18/2015	108.91	109.52	105.81	106.03	9.6E+07
12/17/2015	112.02	112.25	108.98	108.98	4.5E+07
12/16/2015	111.07	111.99	108.8	111.34	5.6E+07
12/15/2015	111.94	112.8	110.35	110.49	5.3E+07
12/14/2015	112.18	112.68	109.79	112.48	6.4E+07
12/11/2015	115.19	115.39	112.85	113.18	4.7E+07
12/10/2015	116.04	116.94	115.51	116.17	2.9E+07
12/9/2015	117.64	117.69	115.08	115.62	4.6E+07
12/8/2015	117.52	118.6	116.86	118.23	3.4E+07
12/7/2015	118.98	119.86	117.81	118.28	3.2E+07
12/4/2015	115.29	119.25	115.11	119.03	5.8E+07
12/3/2015	116.55	116.79	114.22	115.2	4.2E+07
12/2/2015	117.34	118.11	116.08	116.28	3.3E+07
12/1/2015	118.75	118.81	116.86	117.34	3.5E+07
11/30/2015	117.99	119.41	117.75	118.3	3.9E+07
11/27/2015	118.29	118.41	117.6	117.81	1.3E+07
11/25/2015	119.21	119.23	117.92	118.03	2.1E+07
11/24/2015	117.33	119.35	117.12	118.88	4.3E+07
11/23/2015	119.27	119.73	117.34	117.75	3.2E+07
11/20/2015	119.2	119.92	118.85	119.3	3.4E+07
11/19/2015	117.64	119.75	116.76	118.78	4.3E+07
11/18/2015	115.76	117.49	115.5	117.29	4.7E+07
11/17/2015	114.92	115.05	113.32	113.69	2.8E+07
11/16/2015	111.38	114.24	111	114.18	3.8E+07
11/13/2015	115.2	115.57	112.27	112.34	4.6E+07
11/12/2015	116.26	116.82	115.65	115.72	3.3E+07
11/11/2015	116.37	117.42	115.21	116.11	4.5E+07
11/10/2015	116.9	118.07	116.06	116.77	5.9E+07

Wrong Binary Stock Prediction

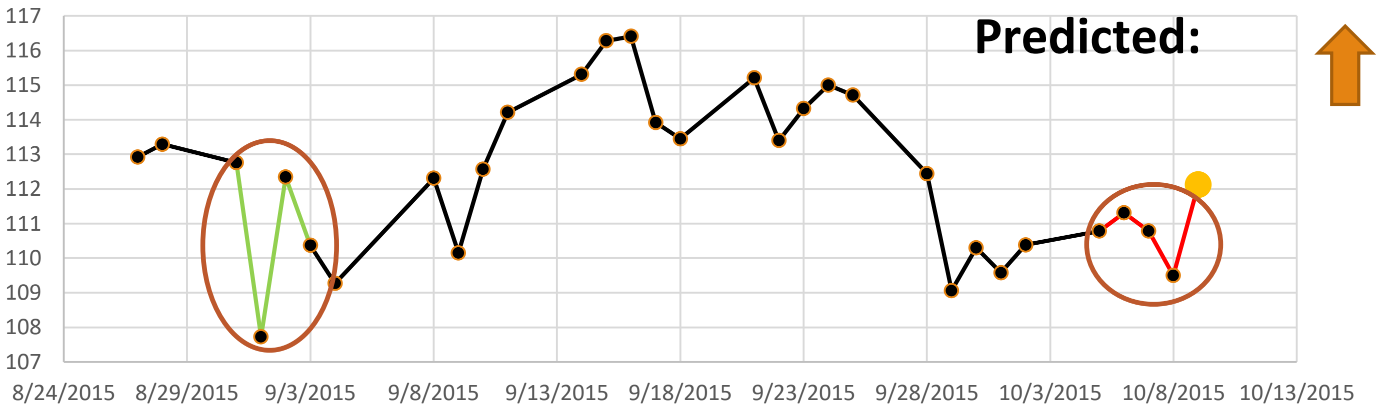
Ground Truth:

Predicted:



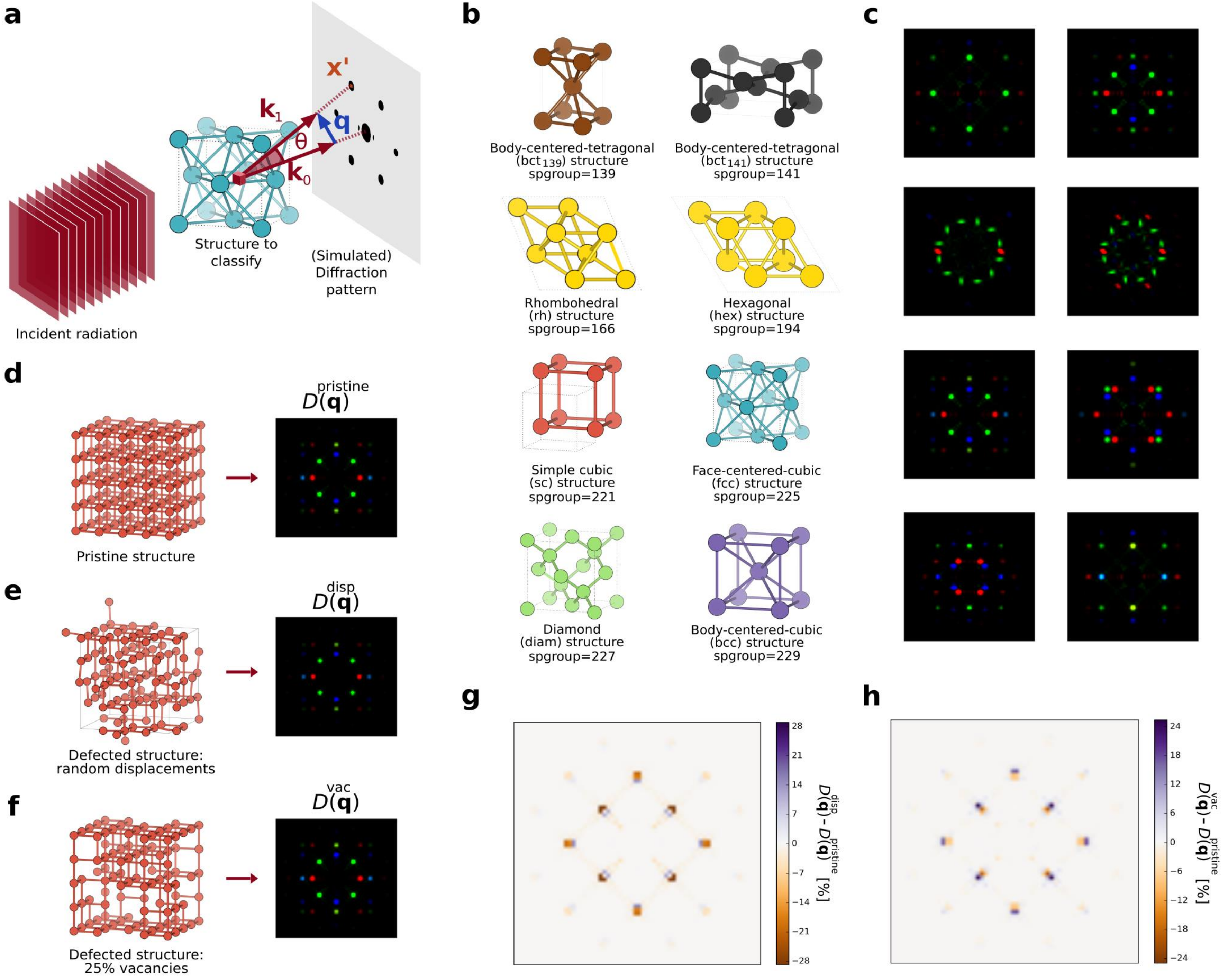
Ground Truth:

Predicted:



Date	Open	High	Low	Close	Volume
10/9/2015	110	112.28	109.49	112.12	5.3E+07
10/8/2015	110.19	110.19	108.21	109.5	6.2E+07
10/7/2015	111.74	111.77	109.41	110.78	4.7E+07
10/6/2015	110.63	111.74	109.77	111.31	4.8E+07
10/5/2015	109.88	111.37	109.07	110.78	5.2E+07
10/2/2015	108.01	111.01	107.55	110.38	5.8E+07
10/1/2015	109.07	109.62	107.31	109.58	6.4E+07
9/30/2015	110.17	111.54	108.73	110.3	6.6E+07
9/29/2015	112.83	113.51	107.86	109.06	7.3E+07
9/28/2015	113.85	114.57	112.44	112.44	5.2E+07
9/25/2015	116.44	116.69	114.02	114.71	5.6E+07
9/24/2015	113.25	115.5	112.37	115	5E+07
9/23/2015	113.63	114.72	113.3	114.32	3.6E+07
9/22/2015	113.38	114.18	112.52	113.4	5E+07
9/21/2015	113.67	115.37	113.66	115.21	5E+07
9/18/2015	112.21	114.3	111.87	113.45	7.4E+07
9/17/2015	115.66	116.49	113.72	113.92	6.4E+07
9/16/2015	116.25	116.54	115.44	116.41	3.7E+07
9/15/2015	115.93	116.53	114.42	116.28	4.3E+07
9/14/2015	116.58	116.89	114.86	115.31	5.8E+07
9/11/2015	111.79	114.21	111.76	114.21	5E+07
9/10/2015	110.27	113.28	109.9	112.57	6.3E+07
9/9/2015	113.76	114.02	109.77	110.15	8.5E+07
9/8/2015	111.75	112.56	110.32	112.31	5.5E+07
9/4/2015	108.97	110.45	108.51	109.27	5E+07
9/3/2015	112.49	112.78	110.04	110.37	5.3E+07
9/2/2015	110.23	112.34	109.13	112.34	6.2E+07
9/1/2015	110.15	111.88	107.36	107.72	7.7E+07
8/31/2015	112.03	114.53	112	112.76	5.6E+07
8/28/2015	112.17	113.31	111.54	113.29	5.3E+07
8/27/2015	112.23	113.24	110.02	112.92	8.5E+07

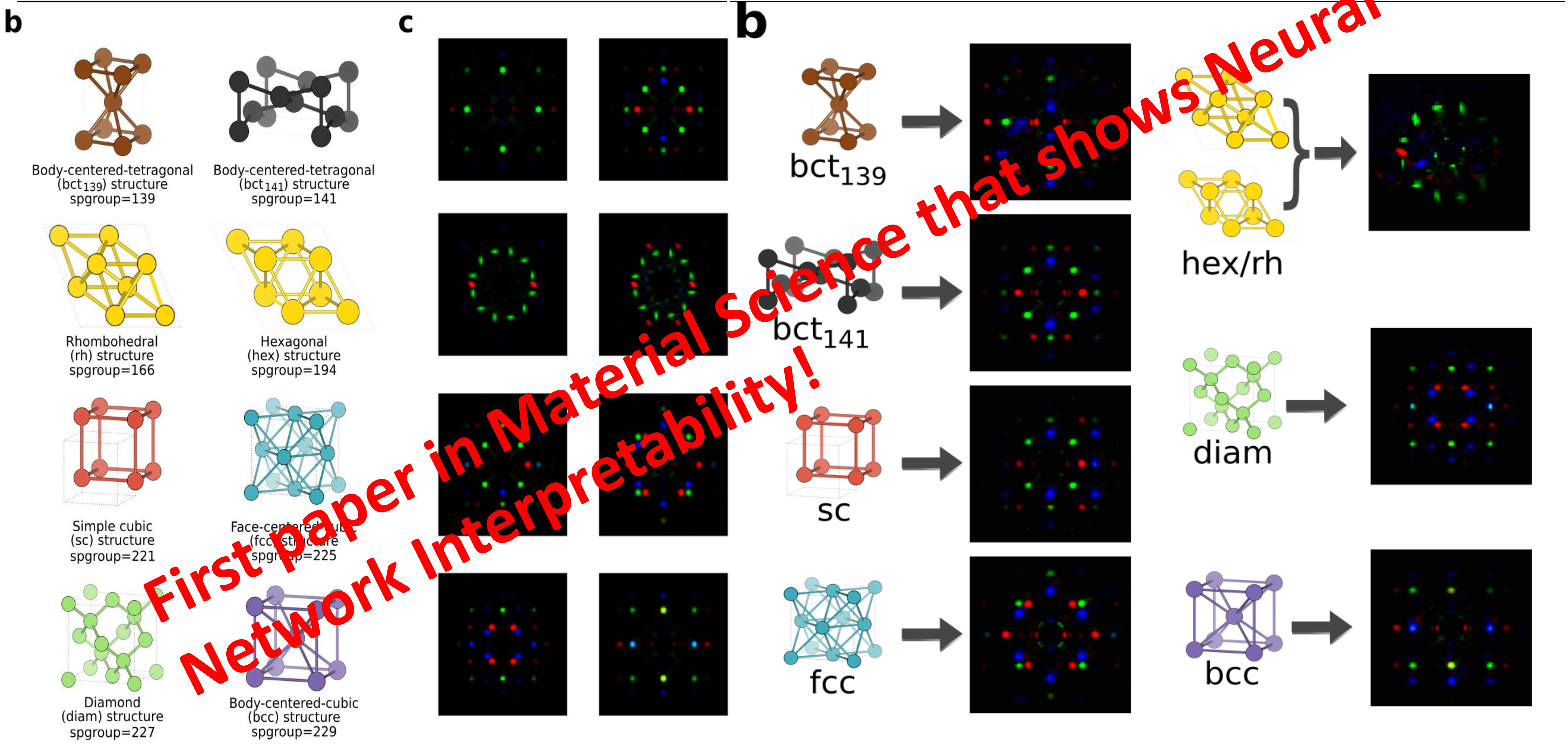
Application in Science (the real one!)



a
x'
k1
q
θ
k0
Incident radiation
Structure to classify
(Simulated) Diffraction pattern
b
Body-centered-tetragonal (bct139) structure spgroup=139
Body-centered-tetragonal (bct141) structure spgroup=141
Rhombohedral (rh) structure spgroup=166
Hexagonal (hex) structure spgroup=194
Simple cubic (sc) structure spgroup=221
Face-centered-cubic (fcc) structure spgroup=225
Diamond (diam) structure spgroup=227
Body-centered-cubic (bcc) structure spgroup=229
c
d
pristine D(q)
Pristine structure
e
disp D(q)
Defected structure: random displacements
f
vac D(q)
Defected structure: 25% vacancies
g
D(q)disp - D(q)pristine [%]
28
21
14
7
0
−7
−14
−21
−28
h
D(q)vac - D(q)pristine [%]
24
18
12
6
0
−6
−12
−18
−24

Ground Truth

Predicted



First paper in Material Science that shows Neural Network Interpretability!

Thank You!

devinder.kumar@uwaterloo.ca

http://devinderkumar.com