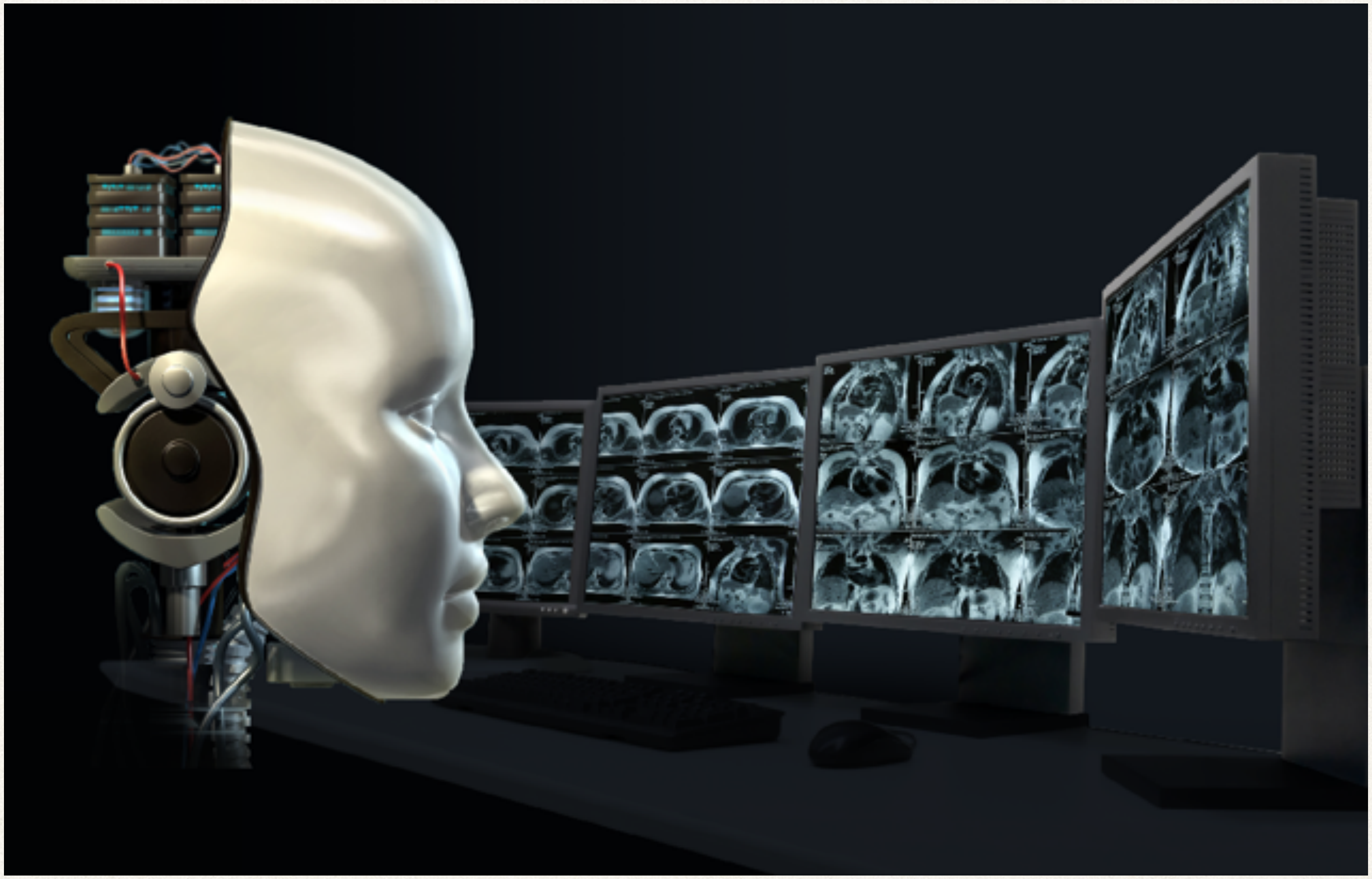
Explainable Deep Learning for High Risk AI Applications



Ulas Bagci, PhD., Center for Research in Computer Vision, University of Central Florida.

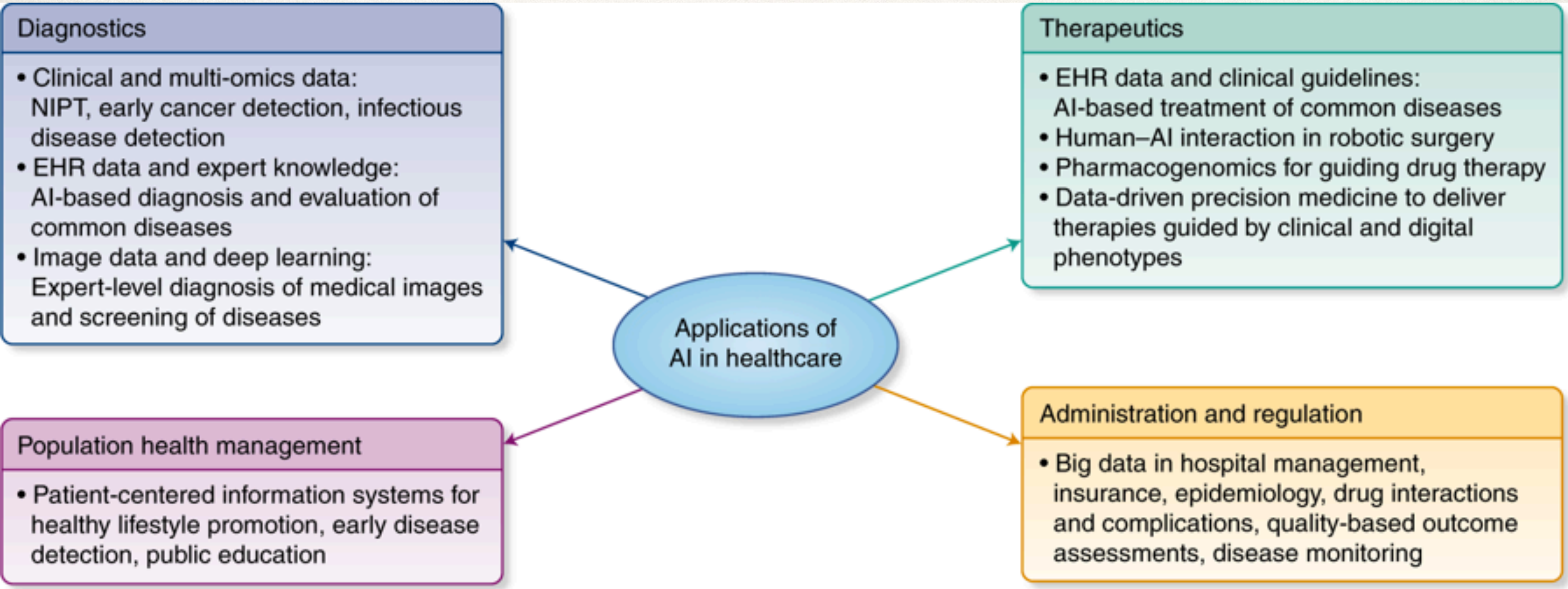
ICMLA-Special Session: Machine Learning in Advanced Machine Vision (AMV 2019)
19 December 2019 - Boca Raton, FL

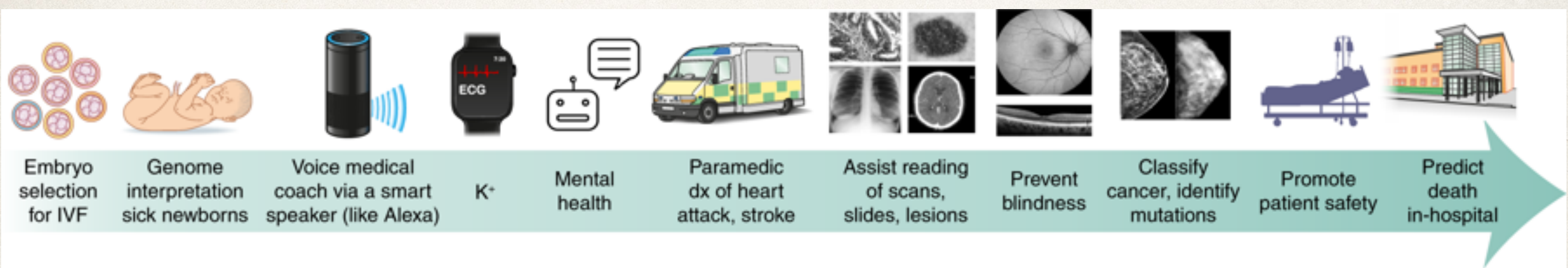
Picture Credit: "Aidoc"

Outline of today's talk

- Interpretability/Explainability in DL
- Image based diagnosis as a high-risk AI application
- Eye-Tracking for human in the loop DL systems
- Visual attribute learning for building the thrust

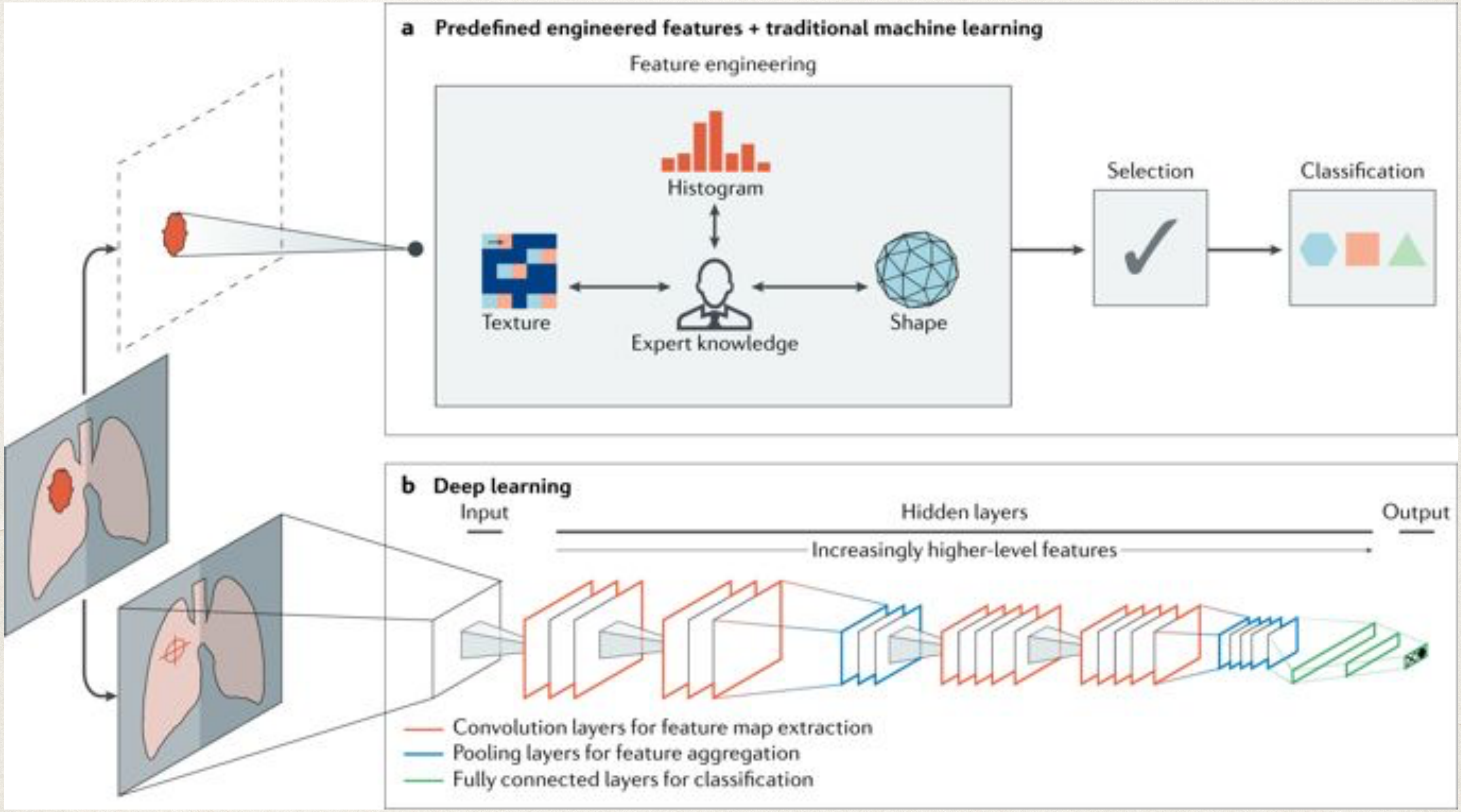
Deep Learning / AI revolutionizes Medicine!





Credit: "E.Topol"

Less Artificial - More Intelligent!



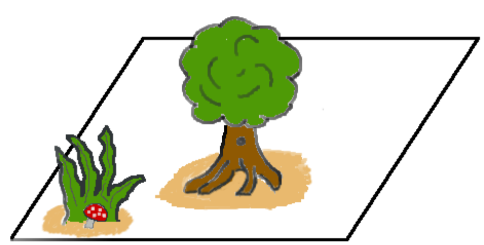
Nature reviews cancer 18, 500-510, 2018

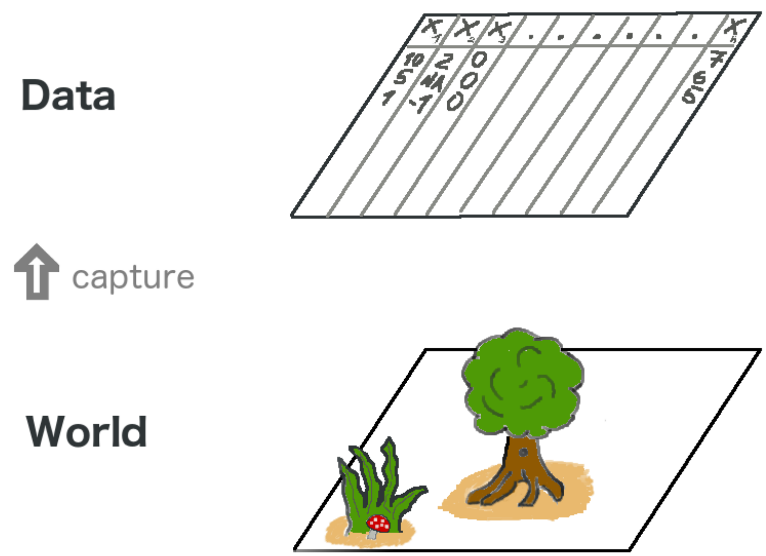
Mostly Black-Box Nature of DL

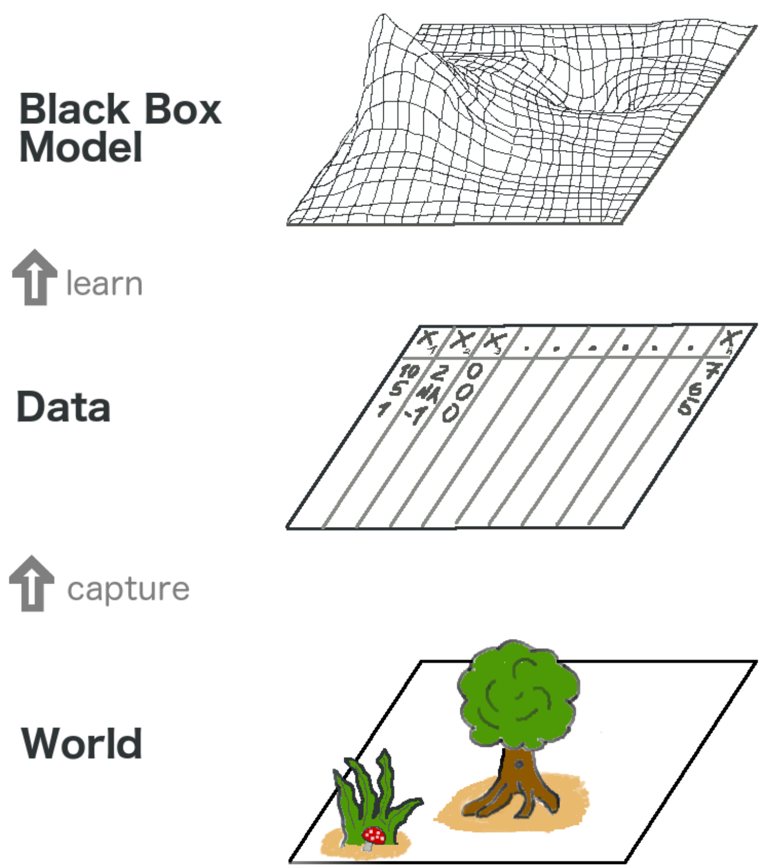
- Imagine a physician using a DNN to diagnose a patient.

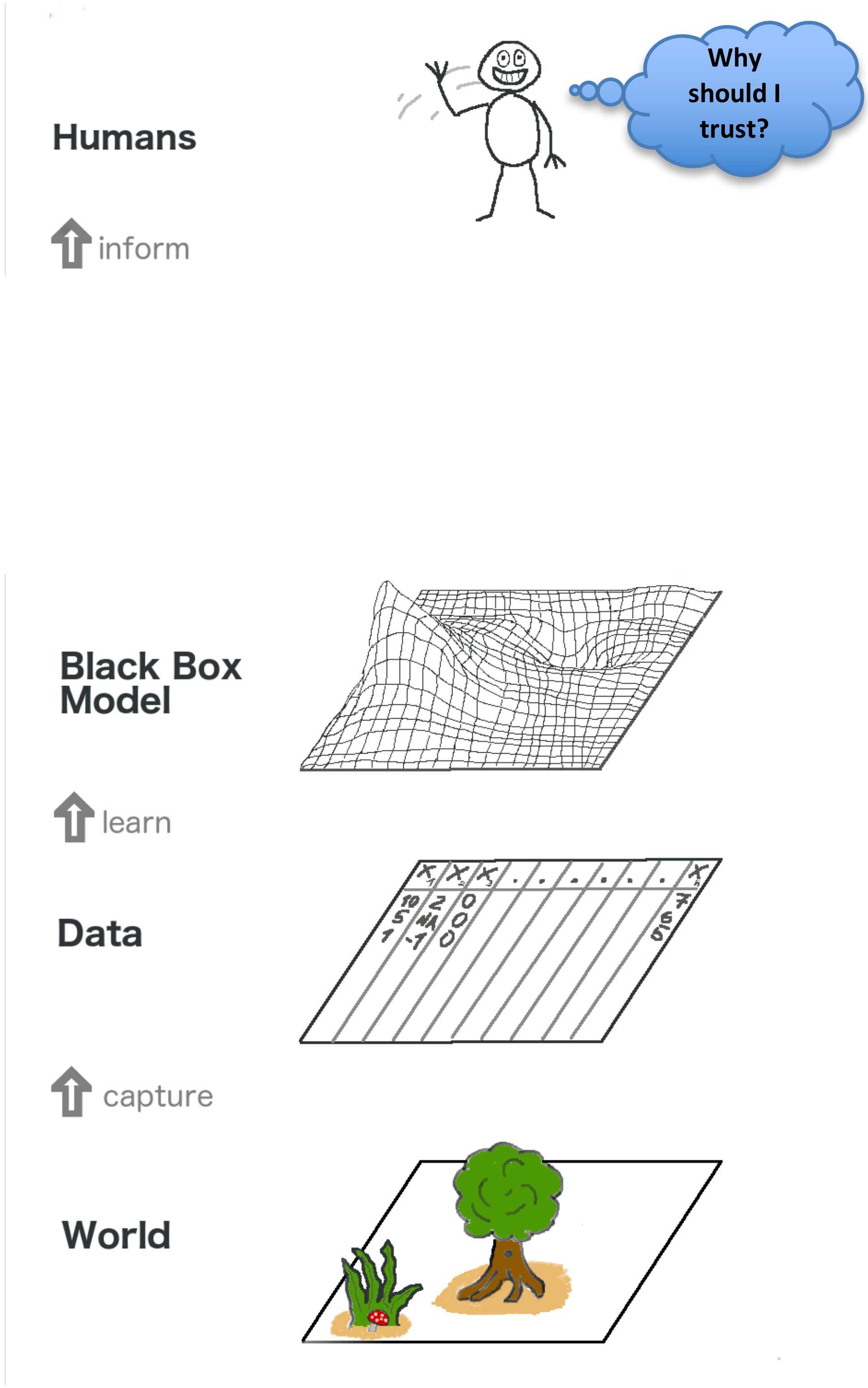
- S/he will most likely **not trust** an automated diagnosis unless s/he **understands the reason** behind a certain prediction (e.g. highlighted regions in the brain that differ from normal subjects) allowing him/her to verify the diagnosis and reason about it.

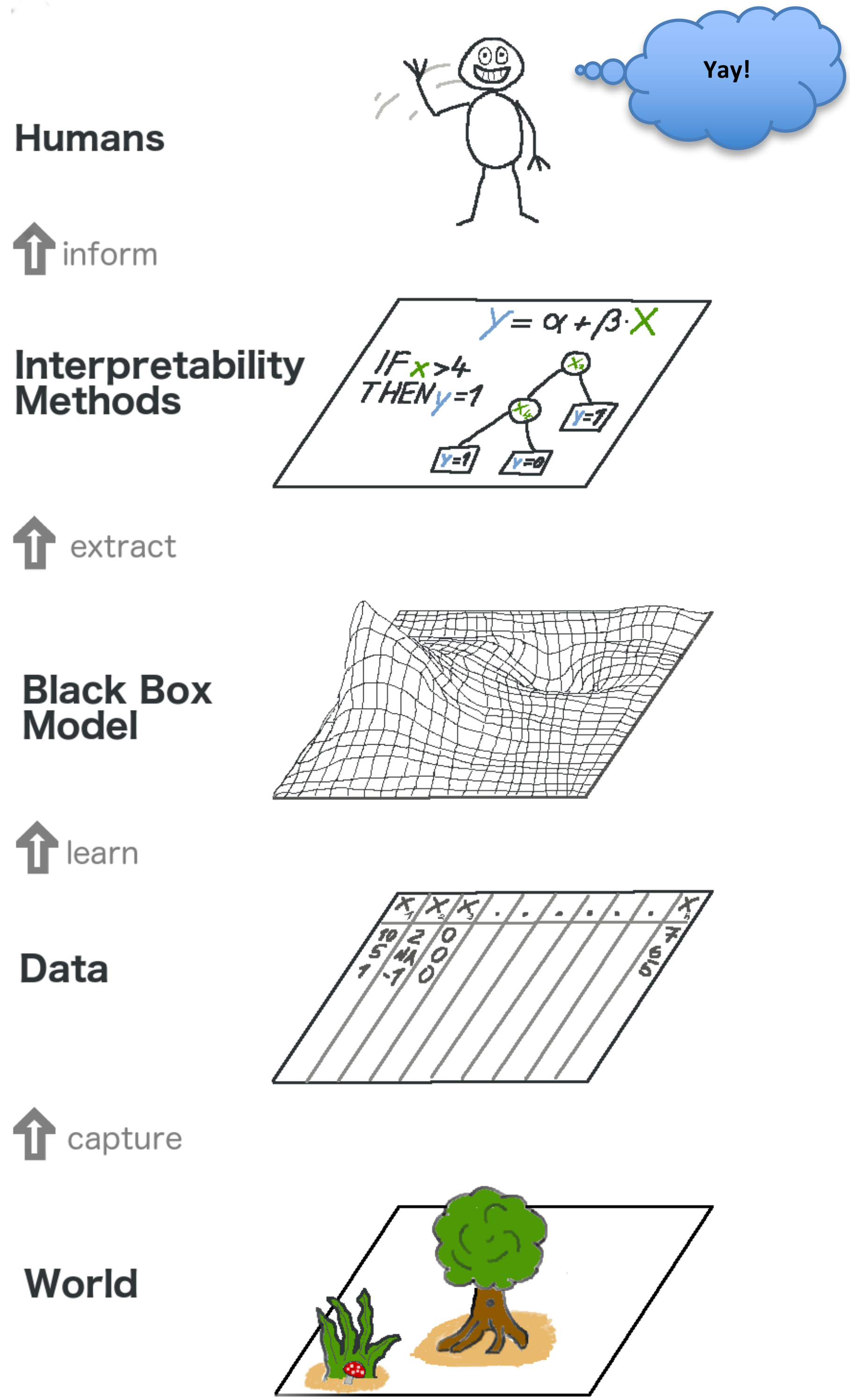
World



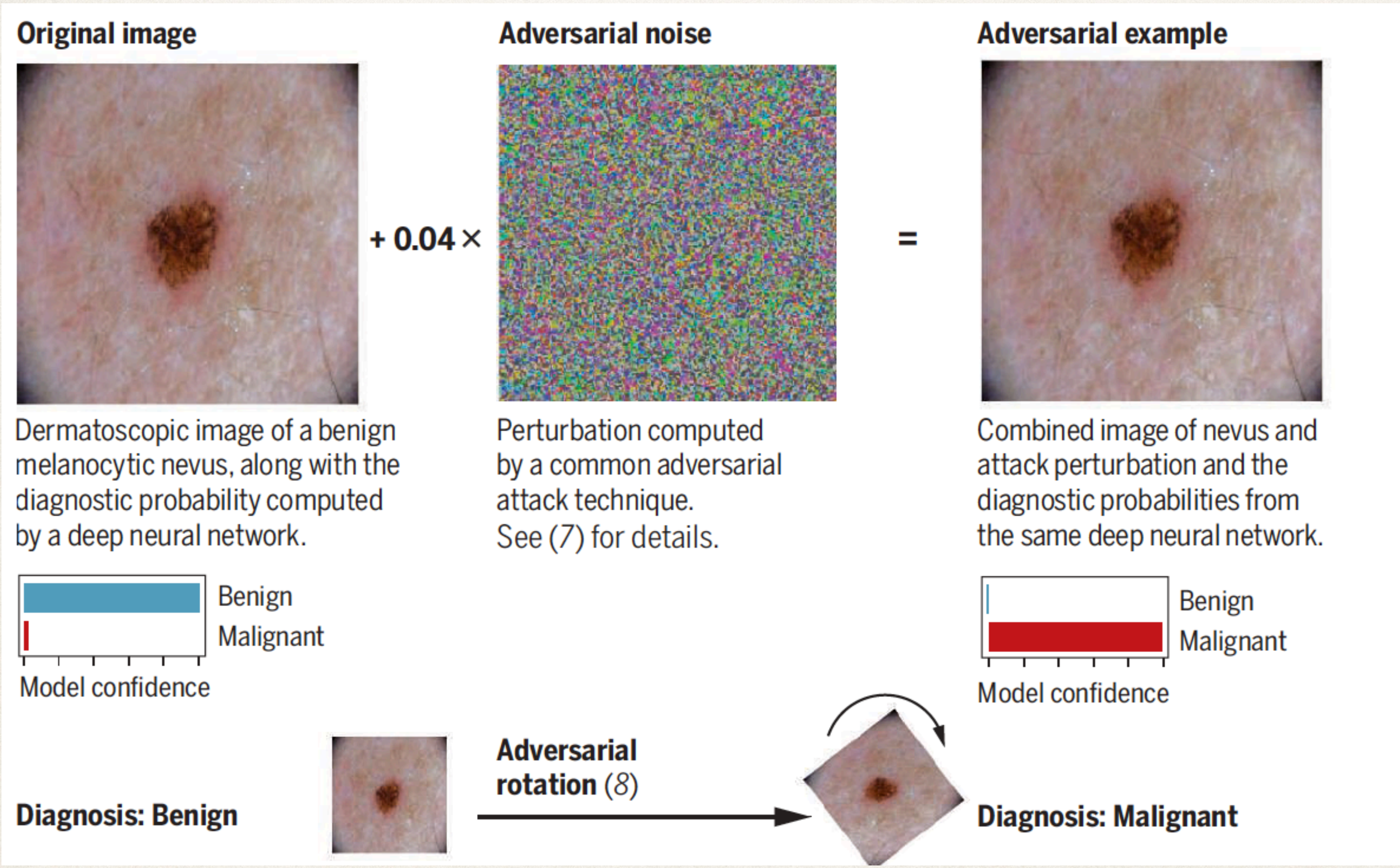

Data
capture
World


Black Box Model
learn
Data
capture
World


Humans
Why should I trust?
inform
Black Box Model
learn
Data
capture
World


Humans
Yay!
inform
Interpretability Methods
y = α + β·X
IF x>4 THEN y=1
extract
Black Box Model
learn
Data
capture
World

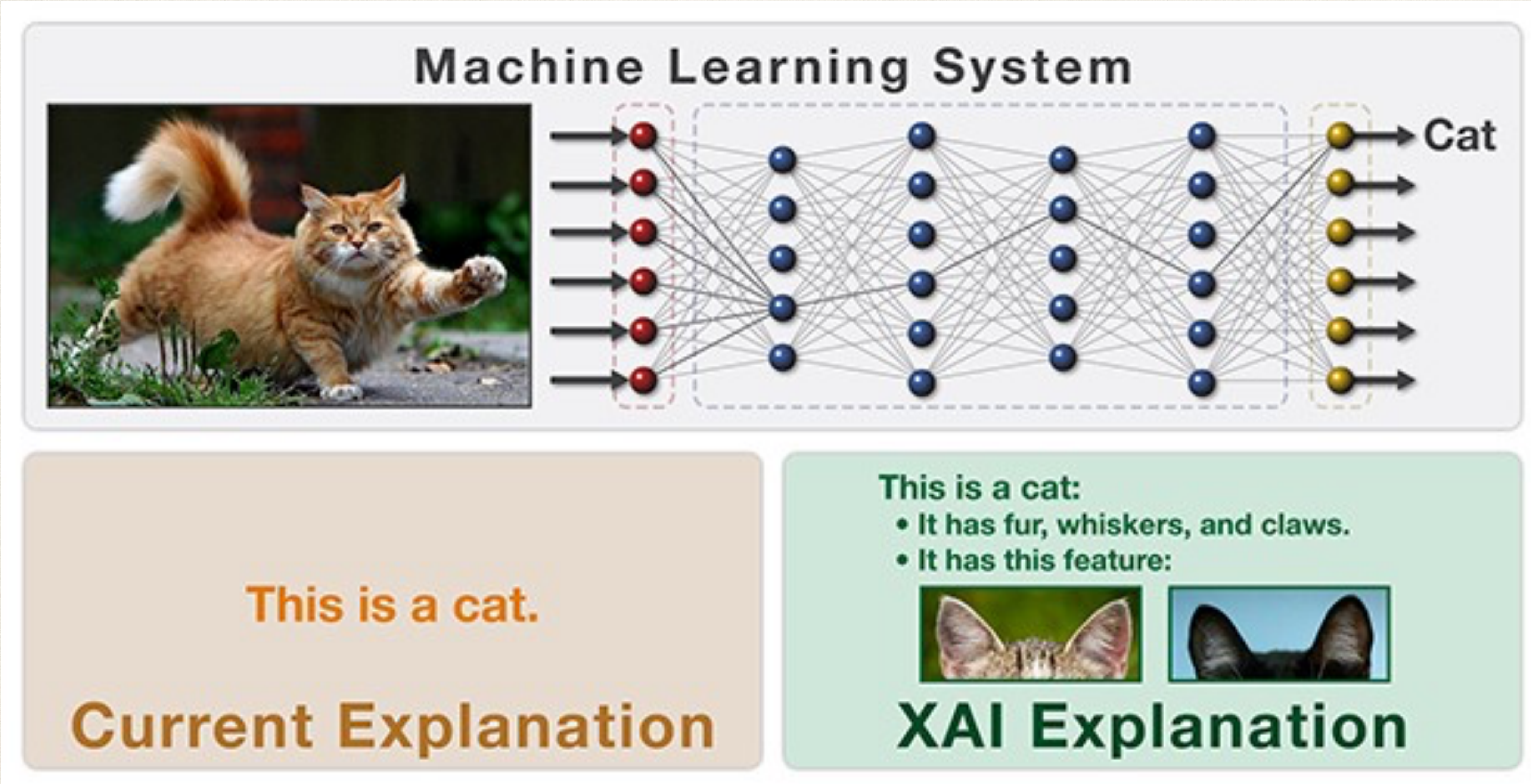
Why should I trust AI?



22 March 2019, Science

AI fails badly too!

- **Uber self-driving car kills a pedestrian**
- **Amazon AI recruiting tool is gender biased**
- **Google Allo suggested man in turban emoji as response to a gun emoji**
- **Google Translate shows gender bias in Turkish-English translations (doctors, hard-working —> he,   nurses, lazy —-> she)**
- **Facebook chatbots shut down after developing their own language**
- **AI misses the mark with Kentucky Derby predictions**
- **Google Home Minis spied on their owners**
- **…**

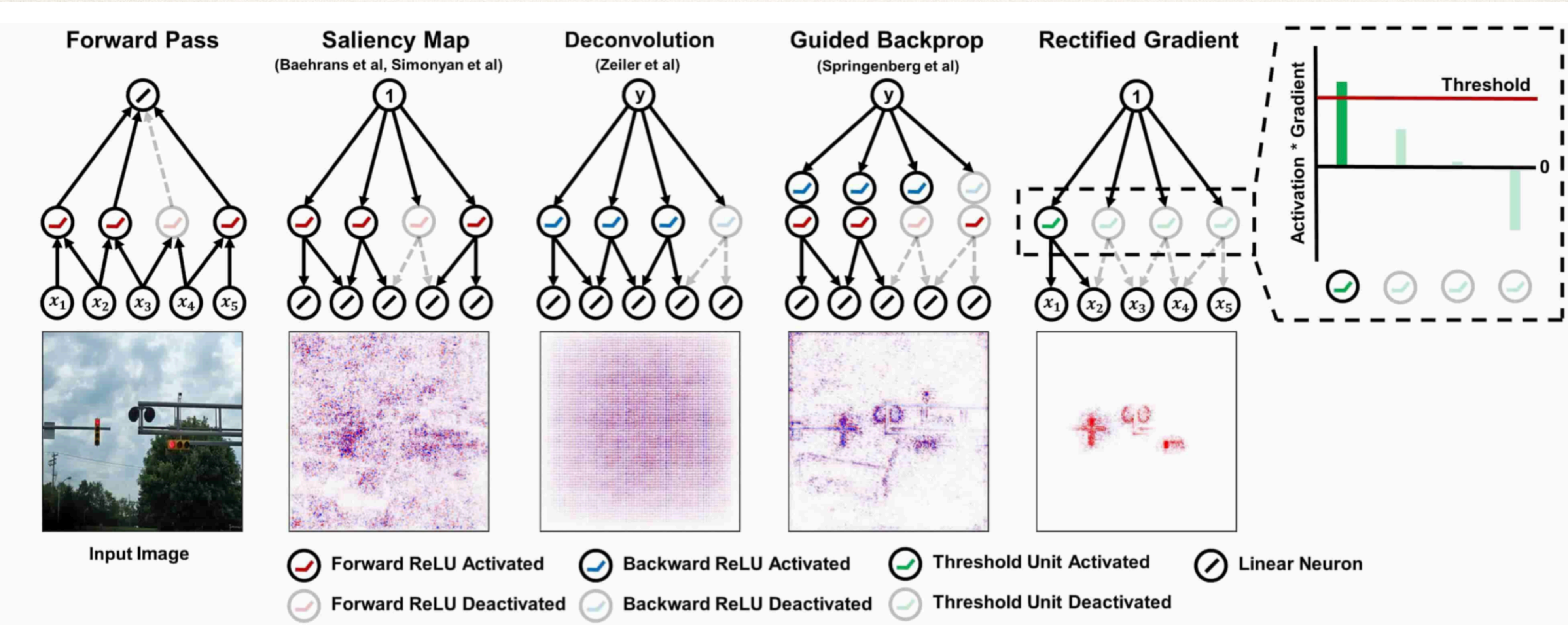


Source: https://www.darpa.mil/program/explainable-artificial-intelligence

Current XAI Approaches

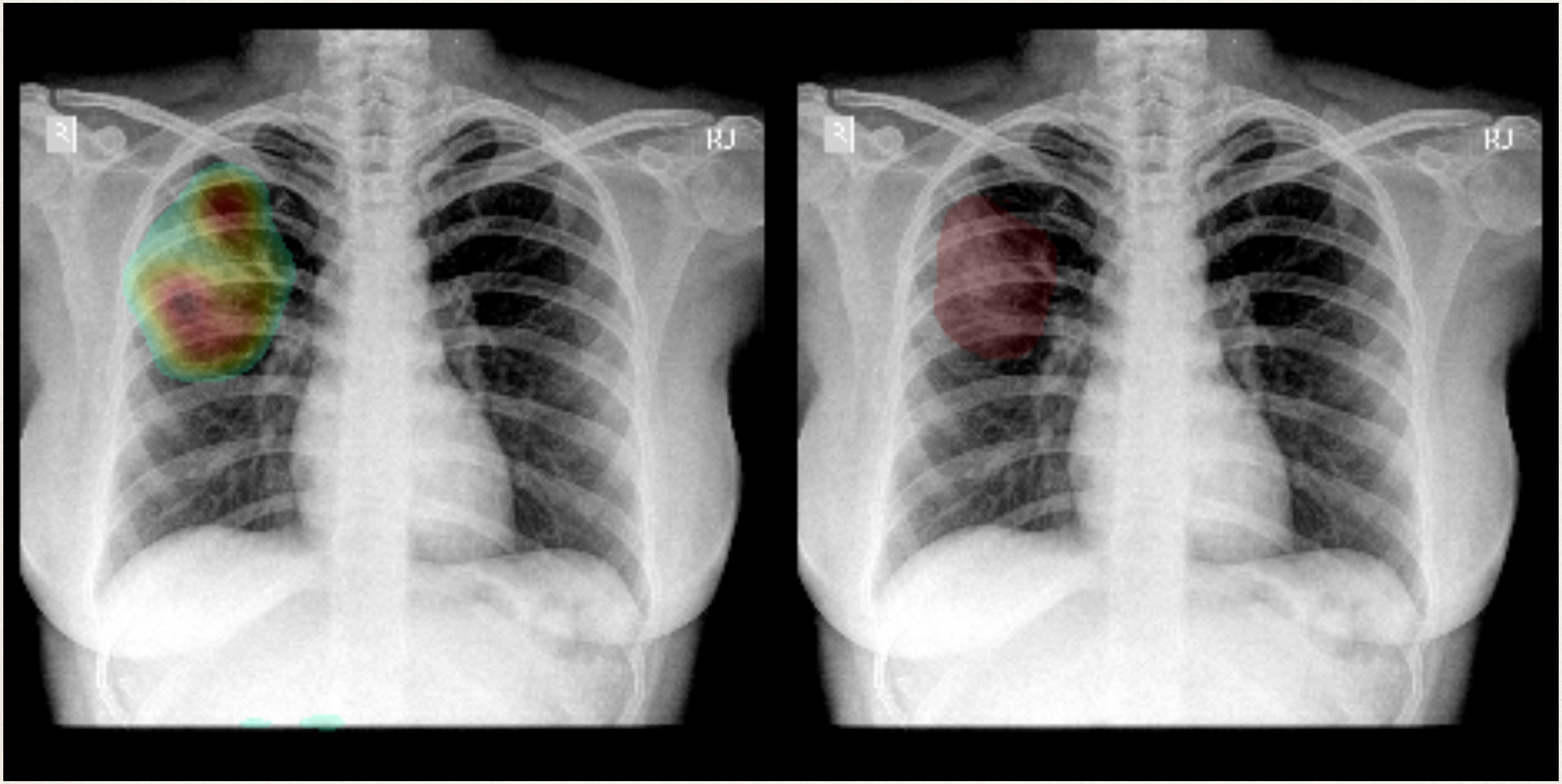
- One qualitative approach is to **highlight areas** that provide evidence in favor of, and against choosing a certain class.
  - Filtering/Visualization
  - Perturbation based methods (saliency maps, CAM, etc)

Current XAI Approaches



- The pixels which contribute maximally to the prediction, once altered, would drop the probability by the maximum amount.

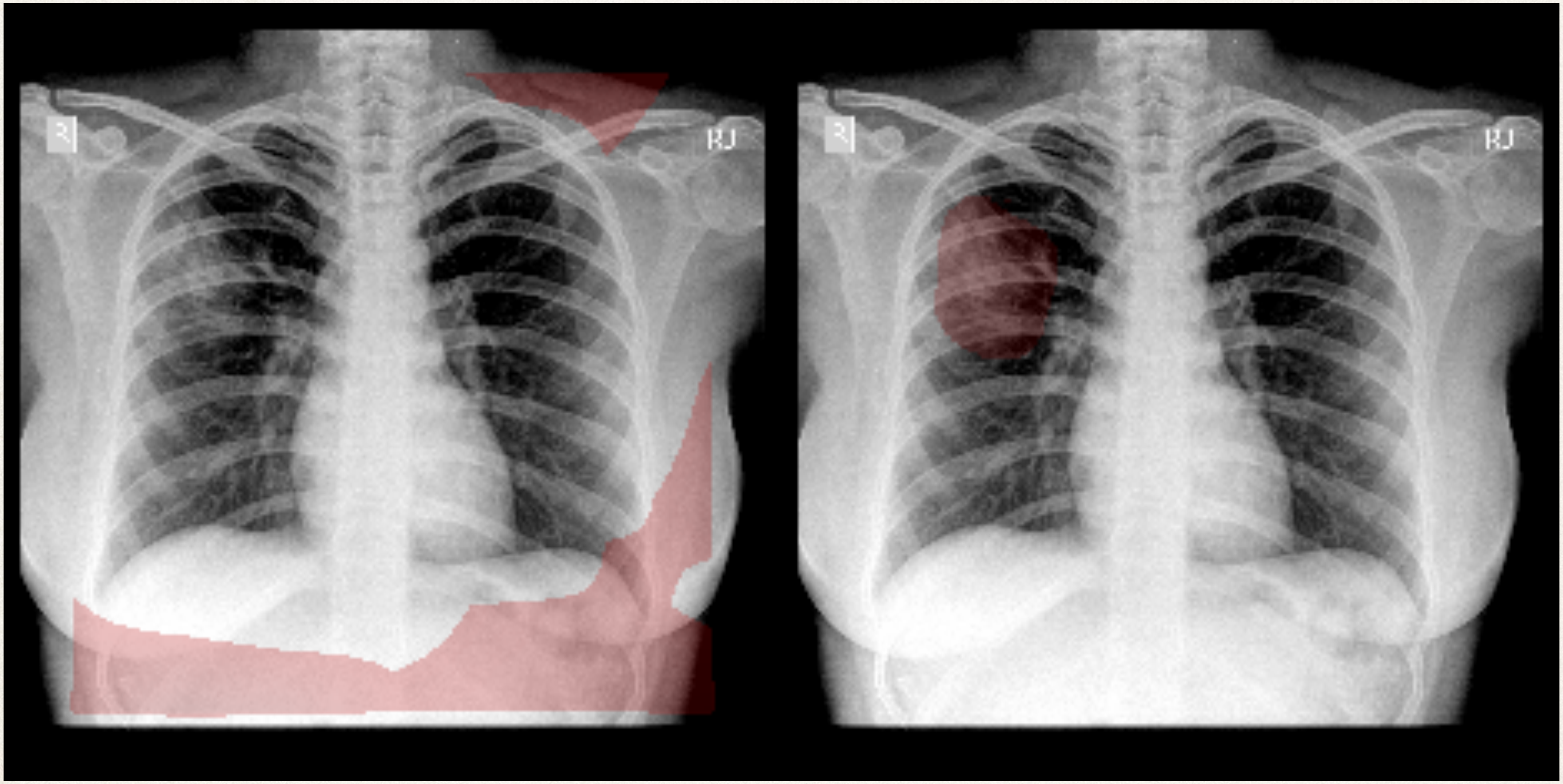
Current XAI Approaches



Qure.AI: Heatmap by GuidedBackprop against original annotation.

Drawbacks of Current XAI Algorithms

Qure.AI: Heatmap by GuidedBackprop against original annotation.



Not completely "true" explanation/reasoning

Artifact generation in visual maps

Limitation to specific architectures

....

What we propose

- Building-in-thrust (human in the loop)
-

What we propose

- Building-in-thrust (human in the loop)
- Explainable / interpretable DL system

What we propose

- Building-in-thrust (human in the loop)
- Explainable / interpretable DL system

Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems

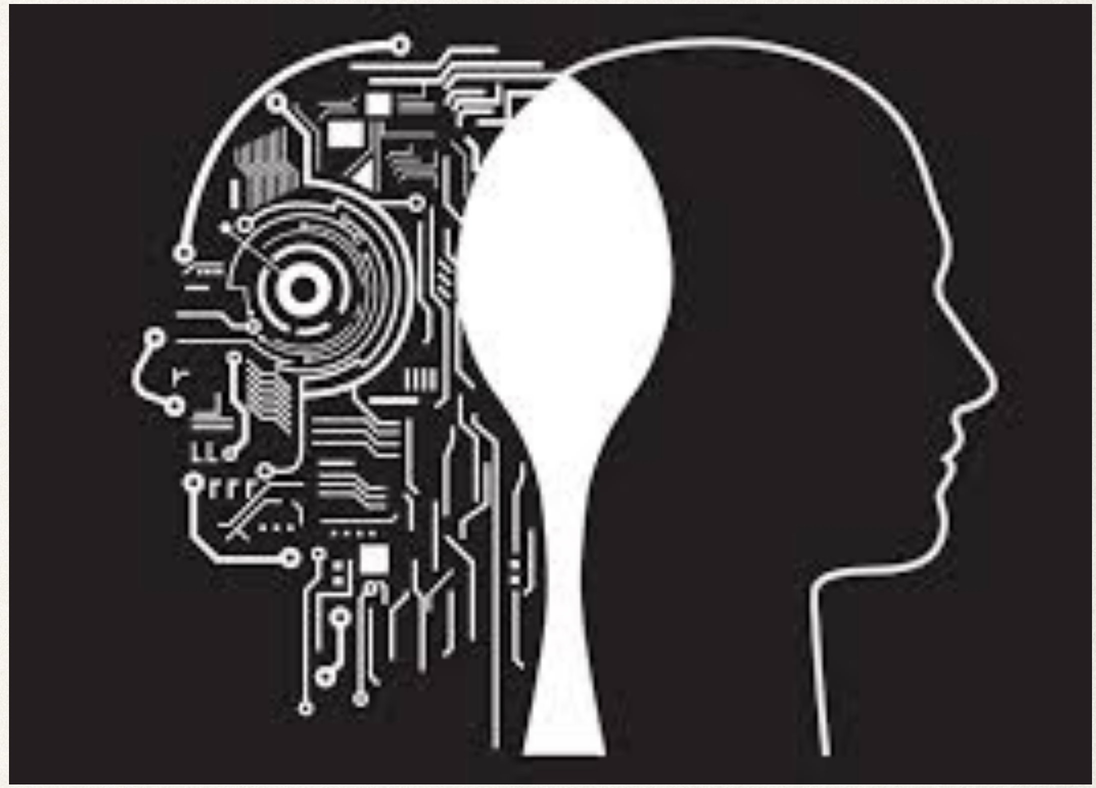
Richard Tomsett[1] **Dave Braines**[1 2] **Dan Harborne**[2] **Alun Preece**[2] **Supriyo Chakraborty**[3]

Defn. Interpretability is a domain-specific notion, so there cannot be all purpose definition.

Radiologist Centered AI Protocols

- Human + AI > AI
  - (human in the loop ML)

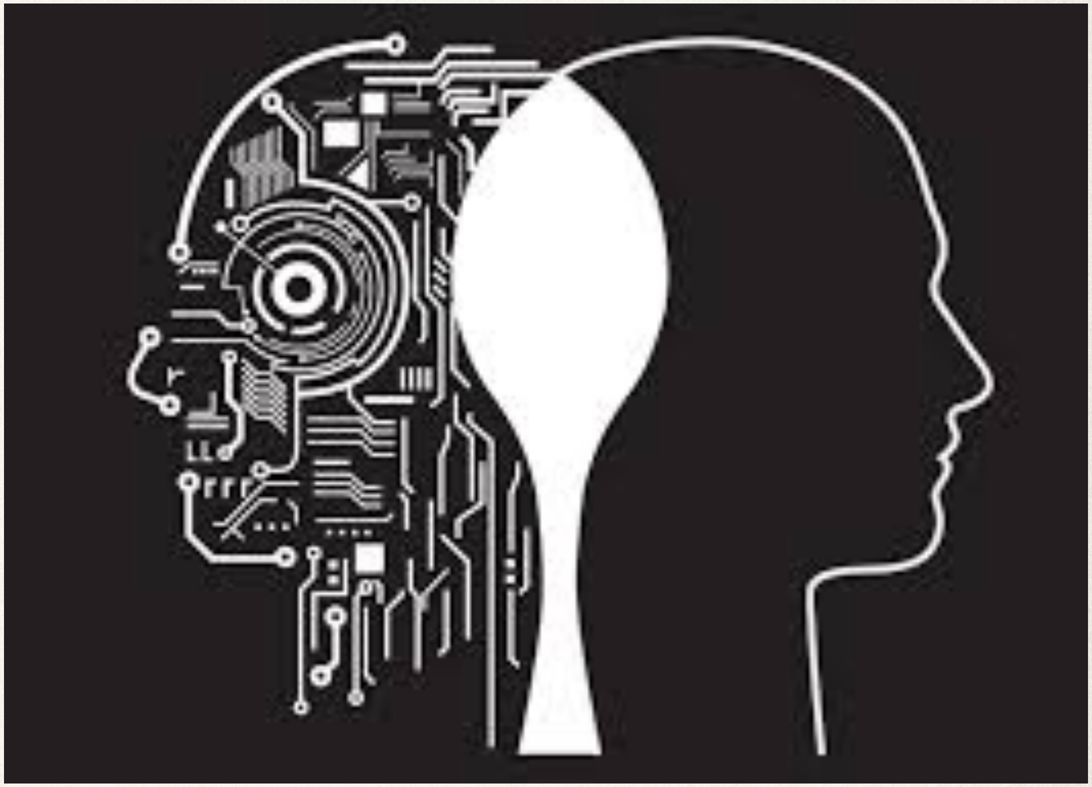
- get a computer system to learn some intelligence behavior by training it on <u>large</u> amount of data.



Radiologist Centered AI Protocols

- Human + AI > AI
  - (human in the loop ML)

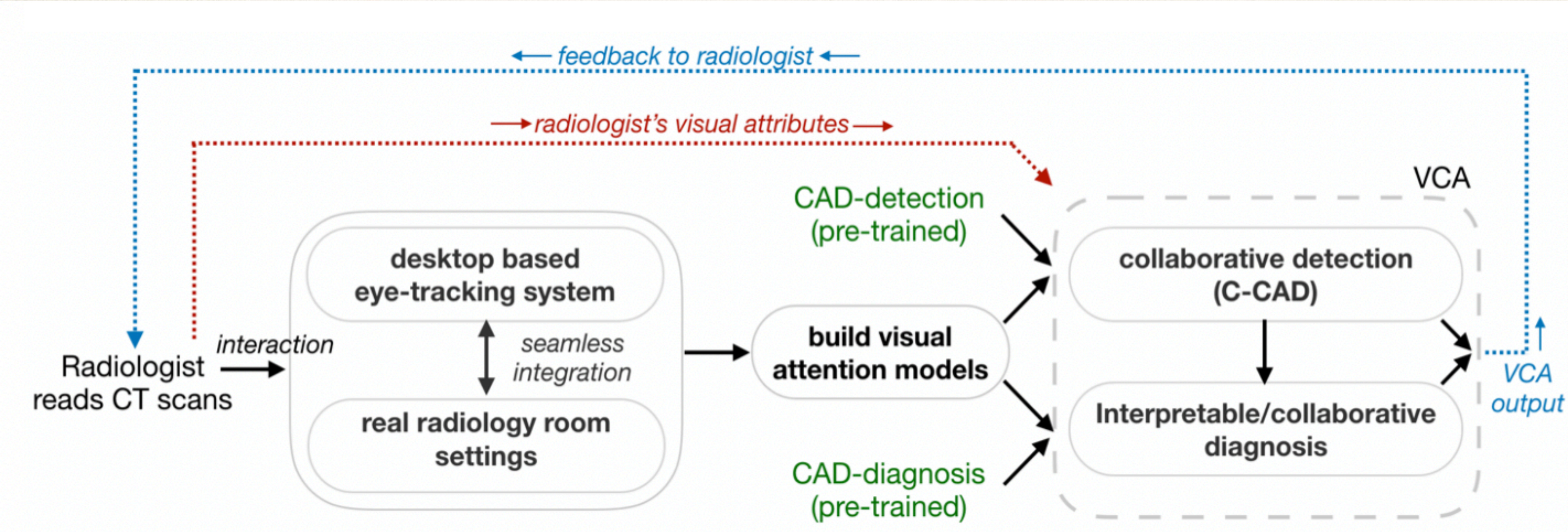
- get a computer system to learn some intelligence behavior by training it on <u>large</u> amount of data.

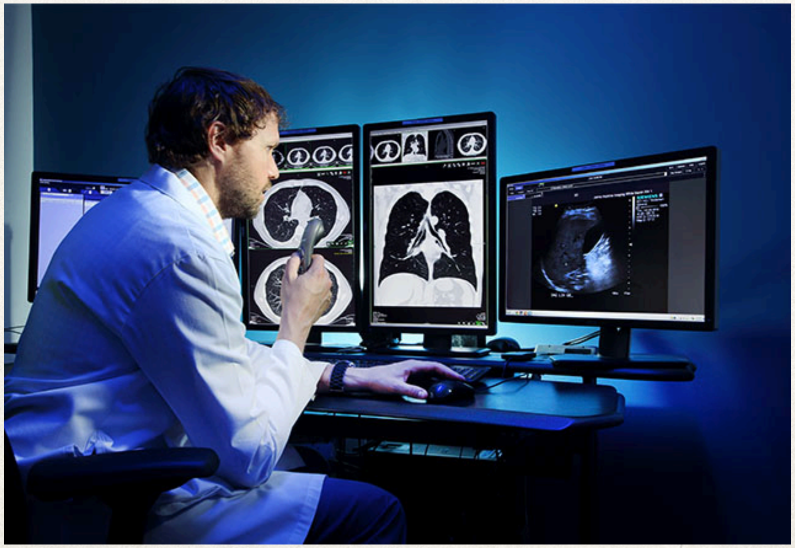


Example High Risk AI Application:

Detection and Malignancy characterization of lung nodule in CT images

RCAI (radiologist centered AI)



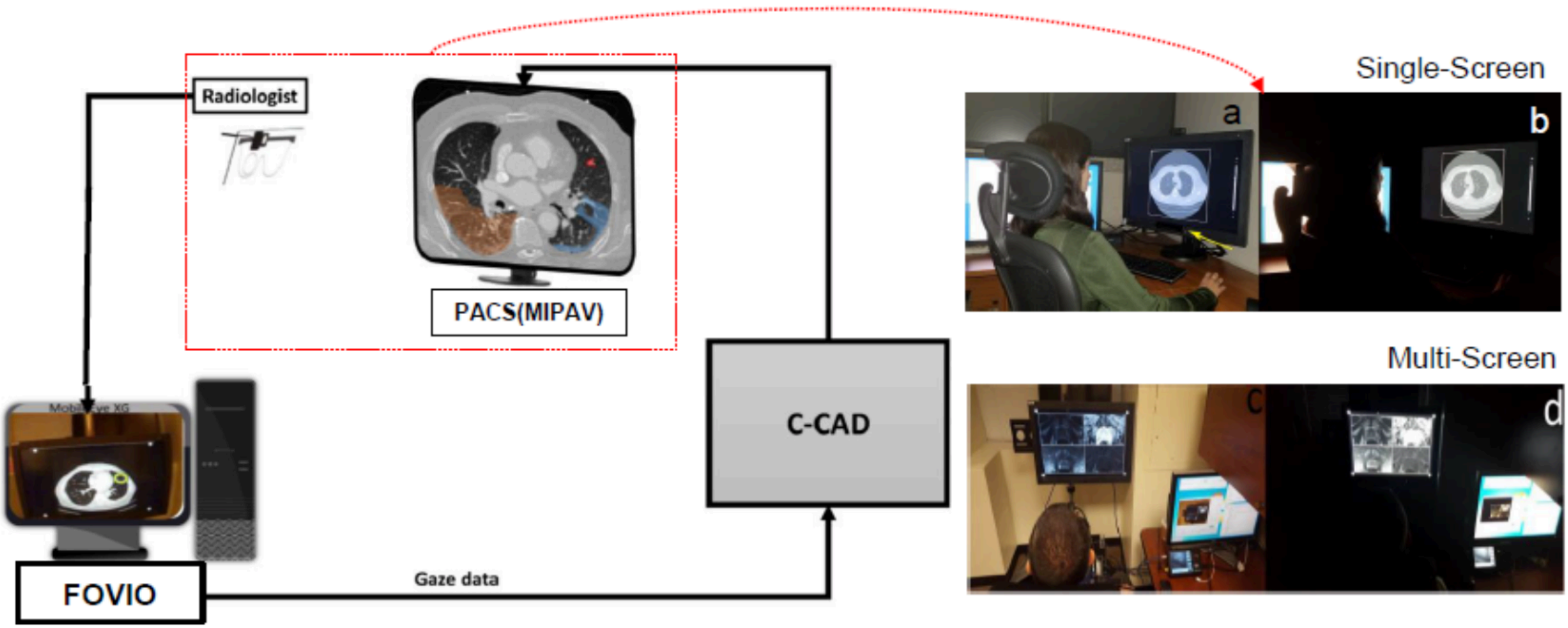


- Dedicated light source
- darkened environment
- limited distraction

Detection via Real-Time Eye-Tracking

Visual Search (Eye-Tracking) + AI Integration

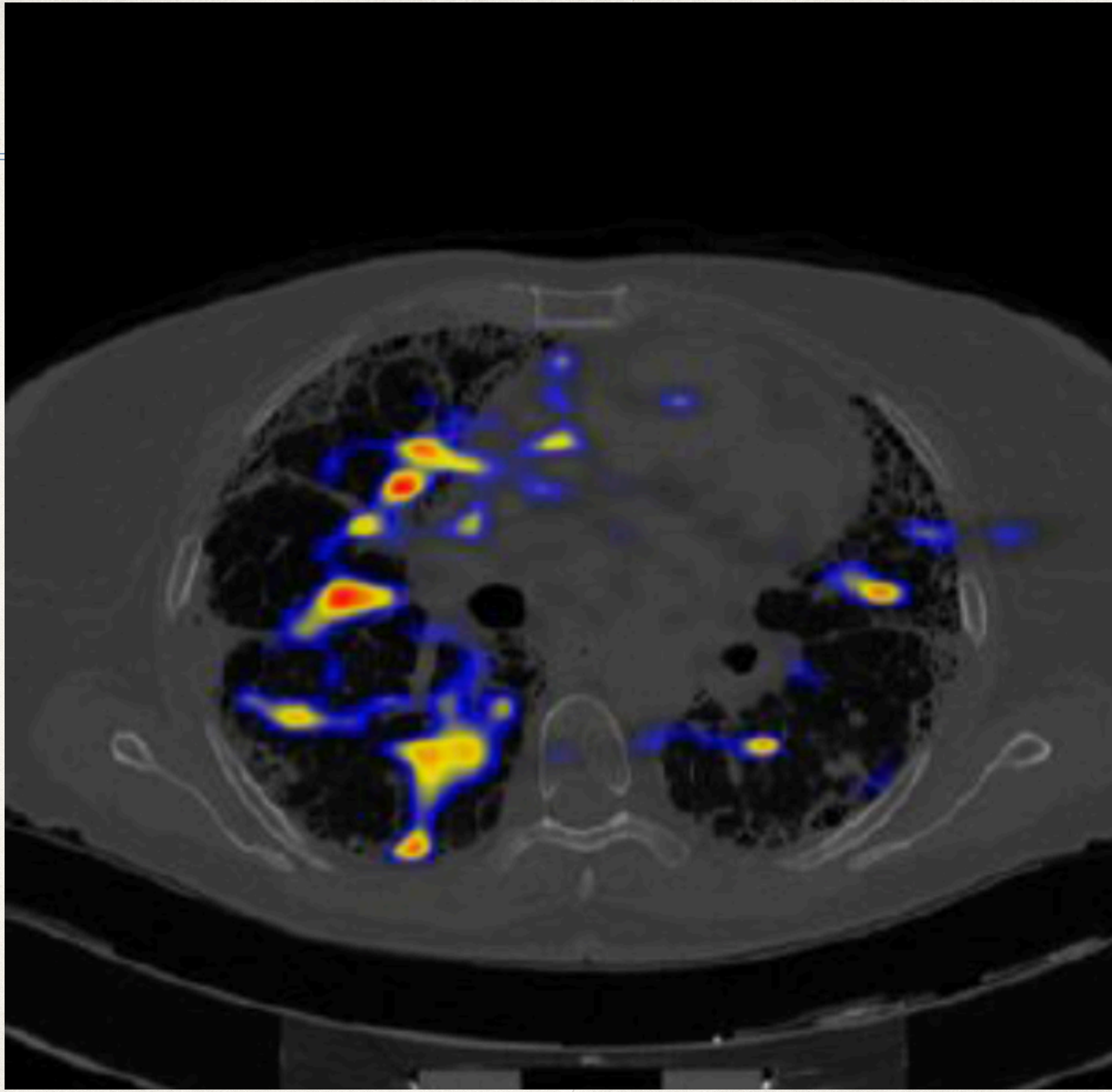
Soln: Combine complementary strengths of radiologists and AI



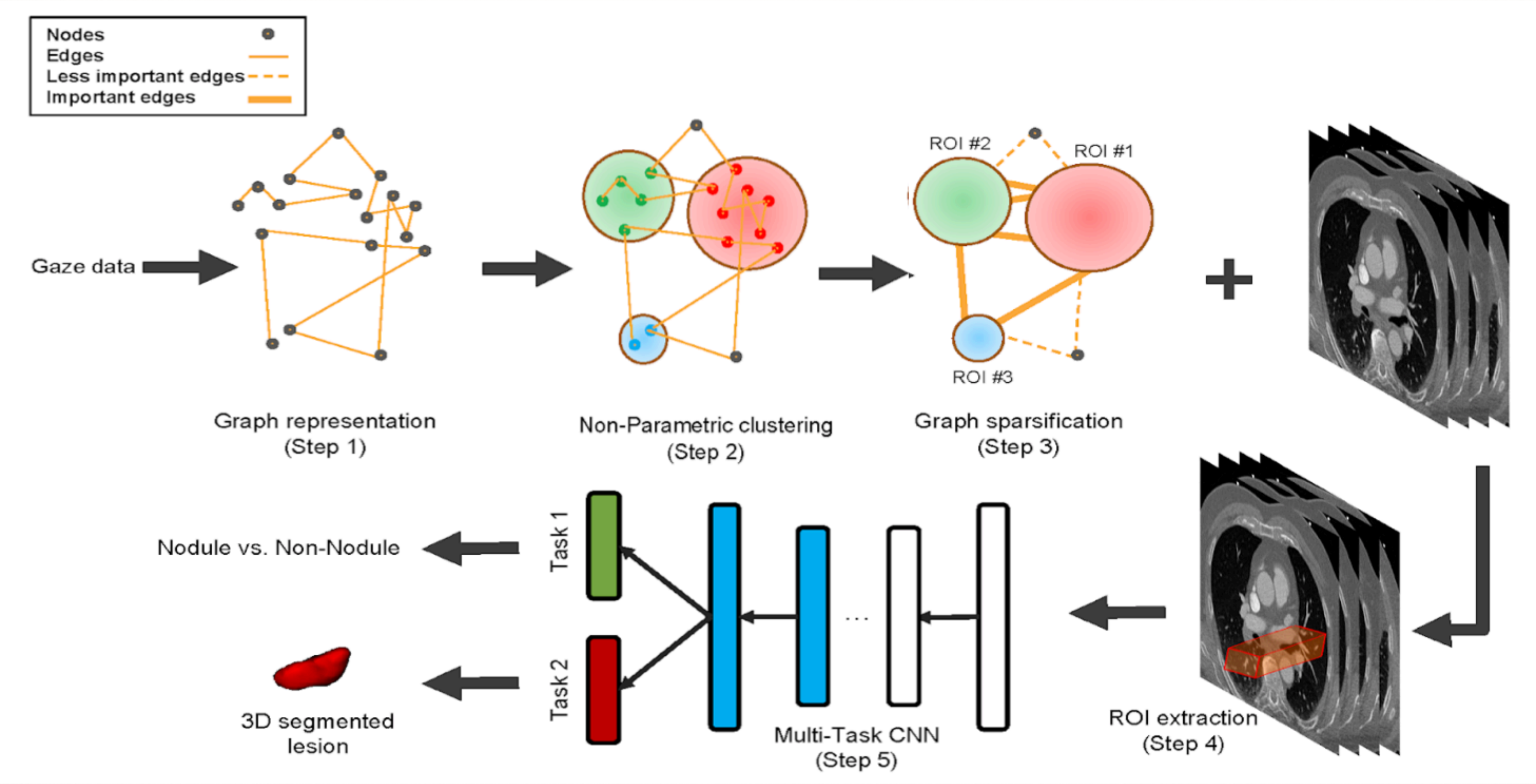
In real radiology rooms, in realistic settings!

Eye-Tracker / Device Info.

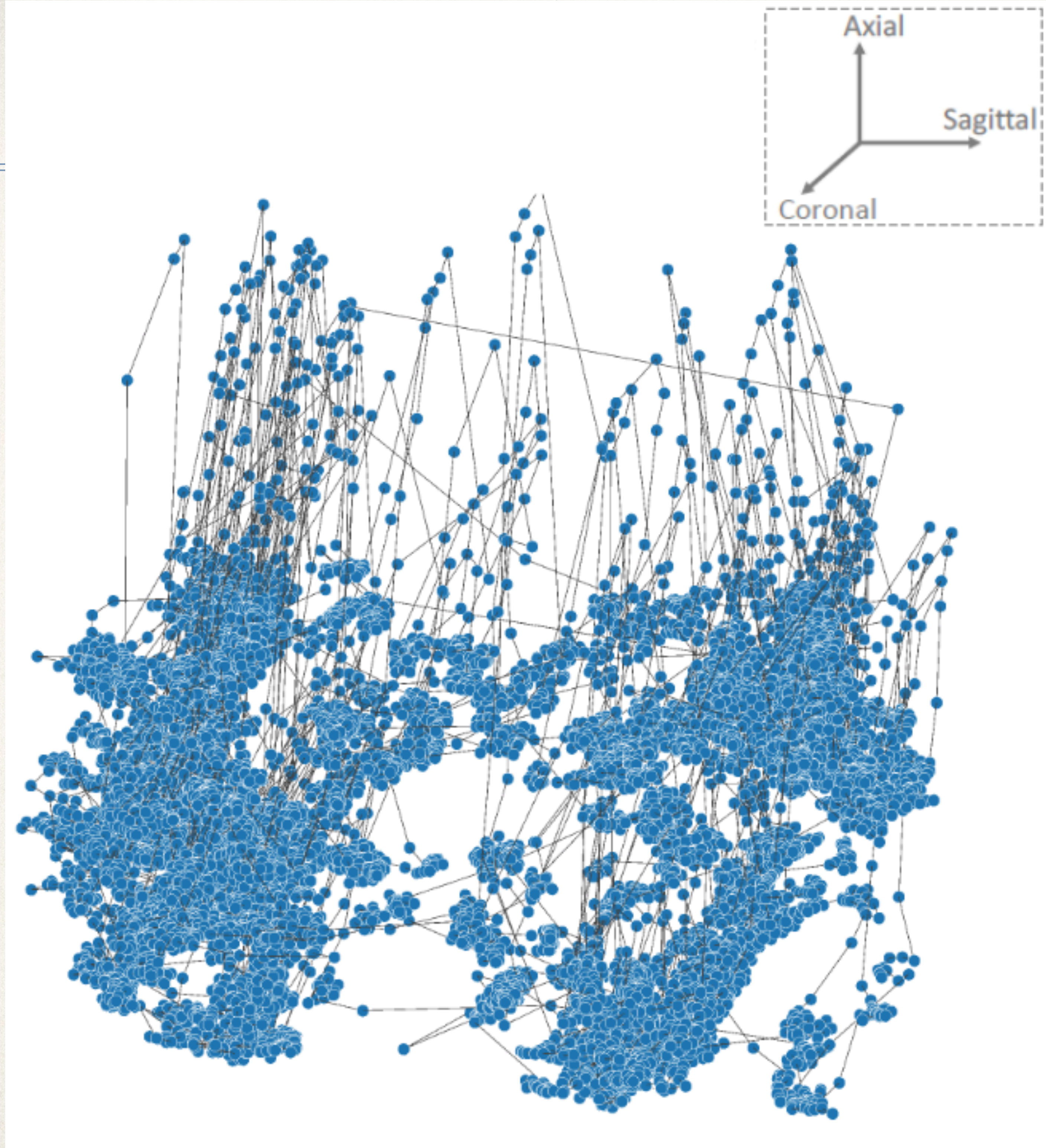
- **Fovio™ Eye Tracker** remote system.
- **System Type:** Remote (contact-free)
- **Sampling Rate:** 60Hz
- **Method:** Proprietary Algorithm
- **Binocular Tracking:** Yes
- **Accuracy:** 0.78 Degrees (Mean) 0.59 (Std. Dev.) angular error
- **Head Box:** 31cm x 40cm @ 65cm - range 40-80cm
- **Additional Details:** Large head box, robust to glasses and ambient light, multi-display tracking



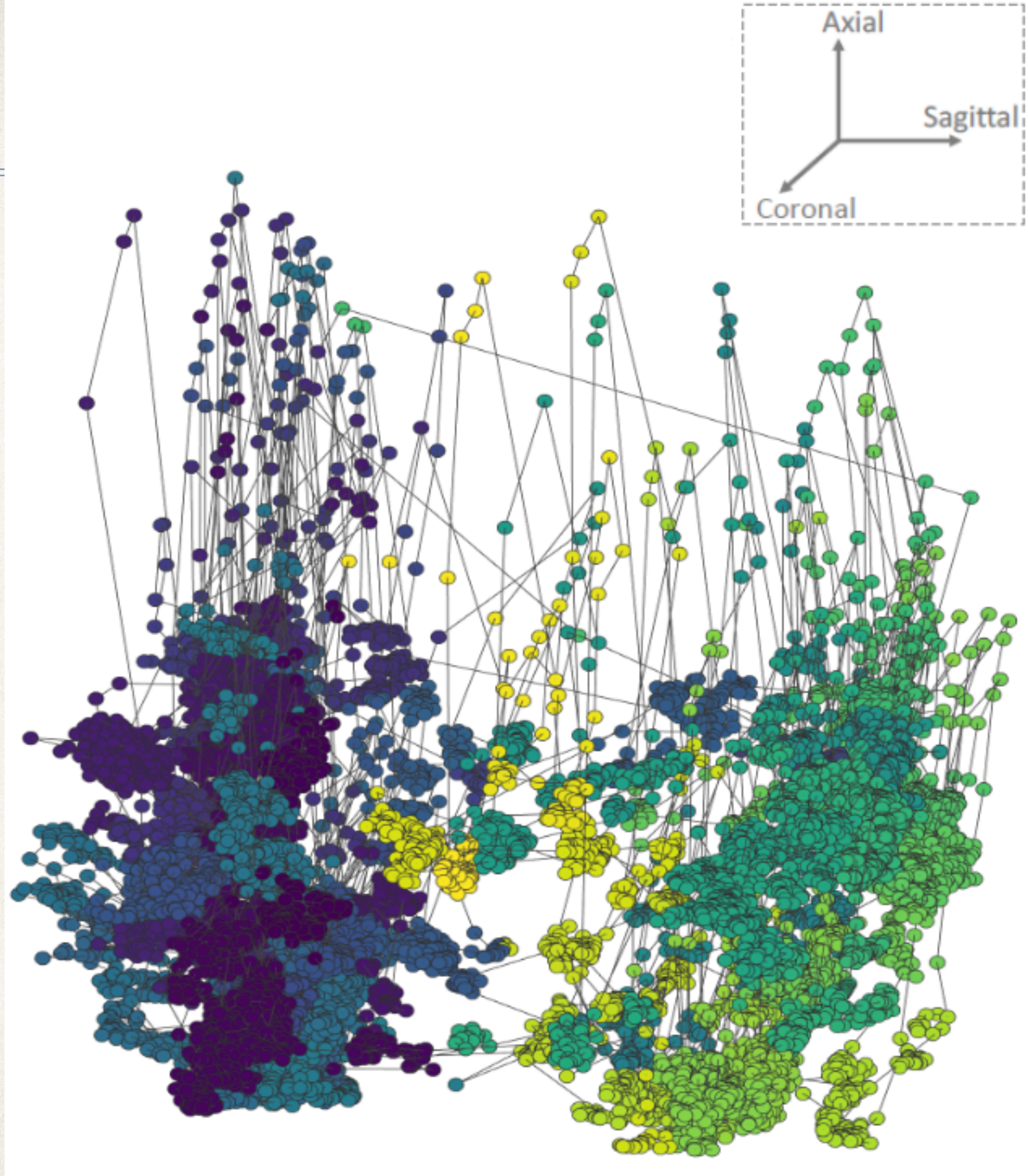
Human-AI Collaboration (Real Time) - Ex: Lung Cancer Screening

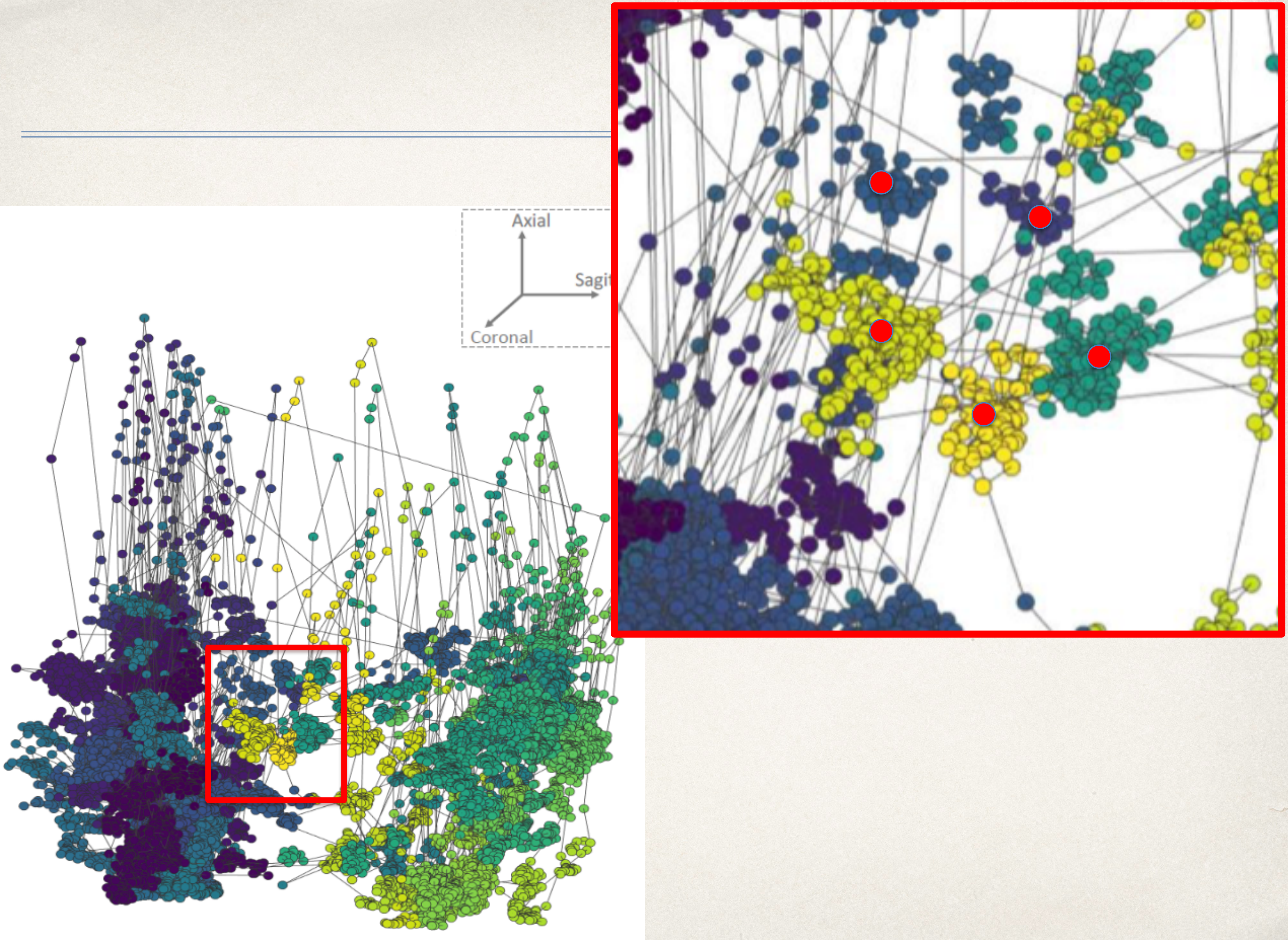


khosravan, et al. MedIA 2018

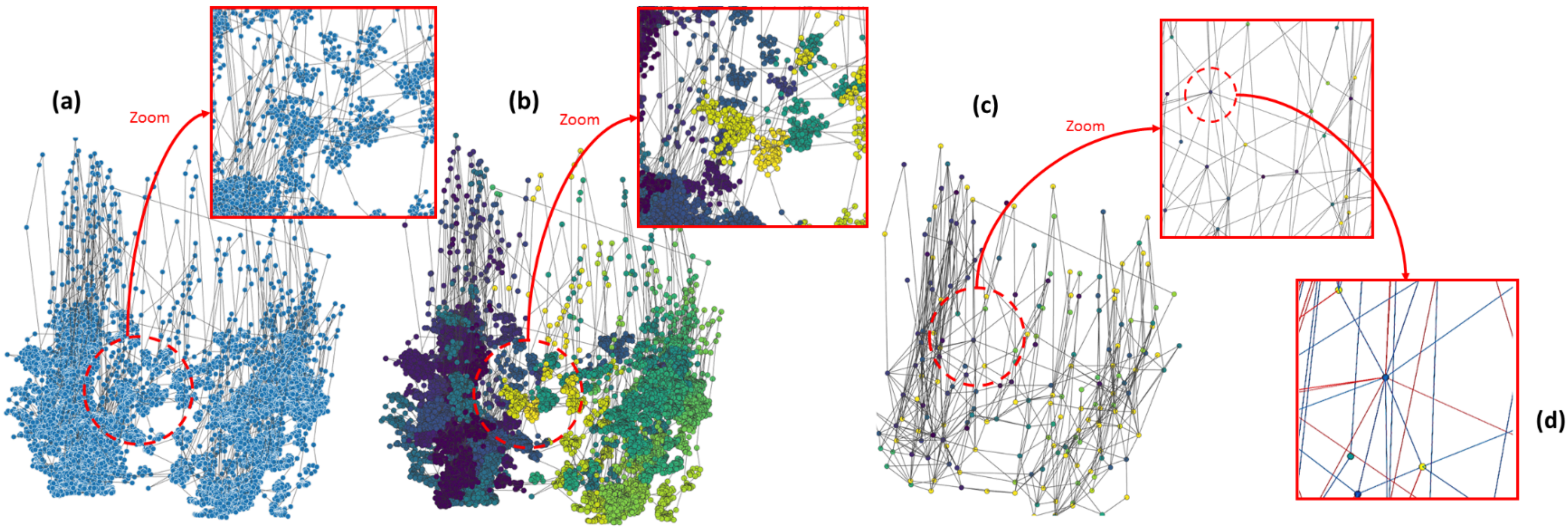


Axial
Sagittal
Coronal


Axial
Sagittal
Coronal


Axial
Coronal

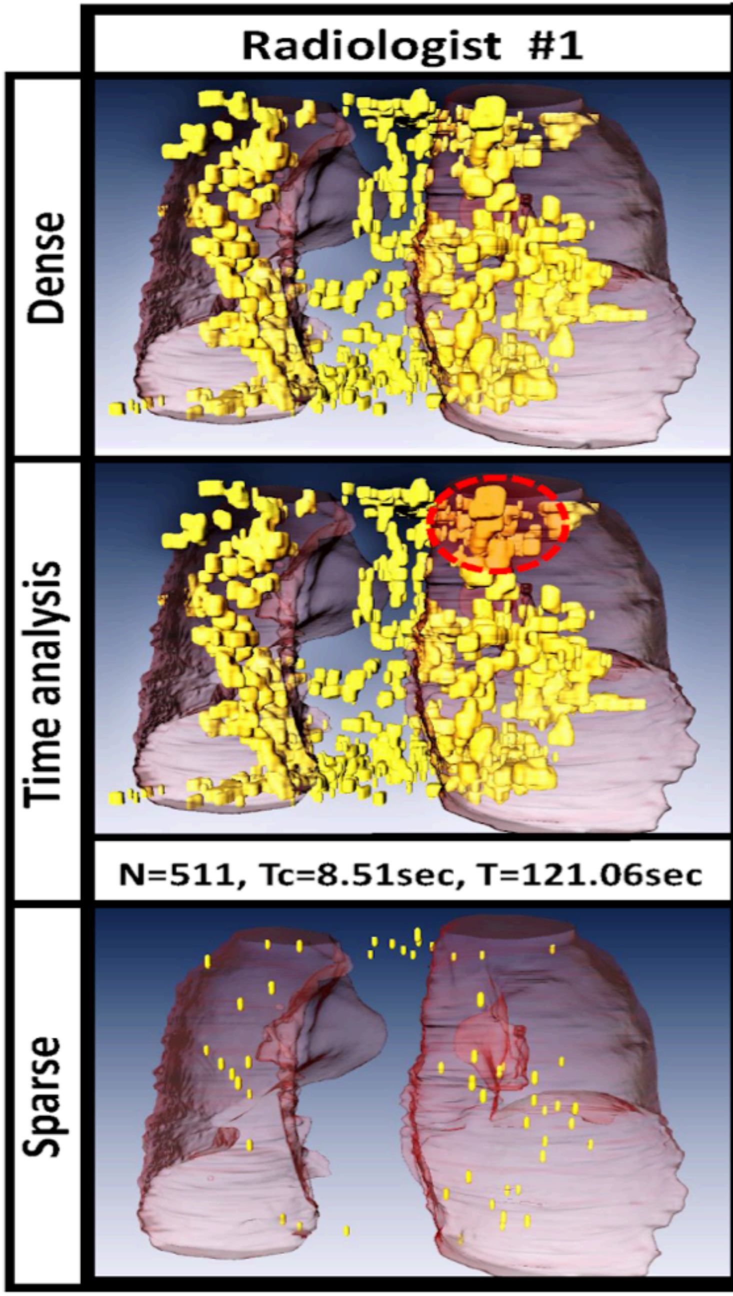
Reduce Gaze Points for Faster Analysis!



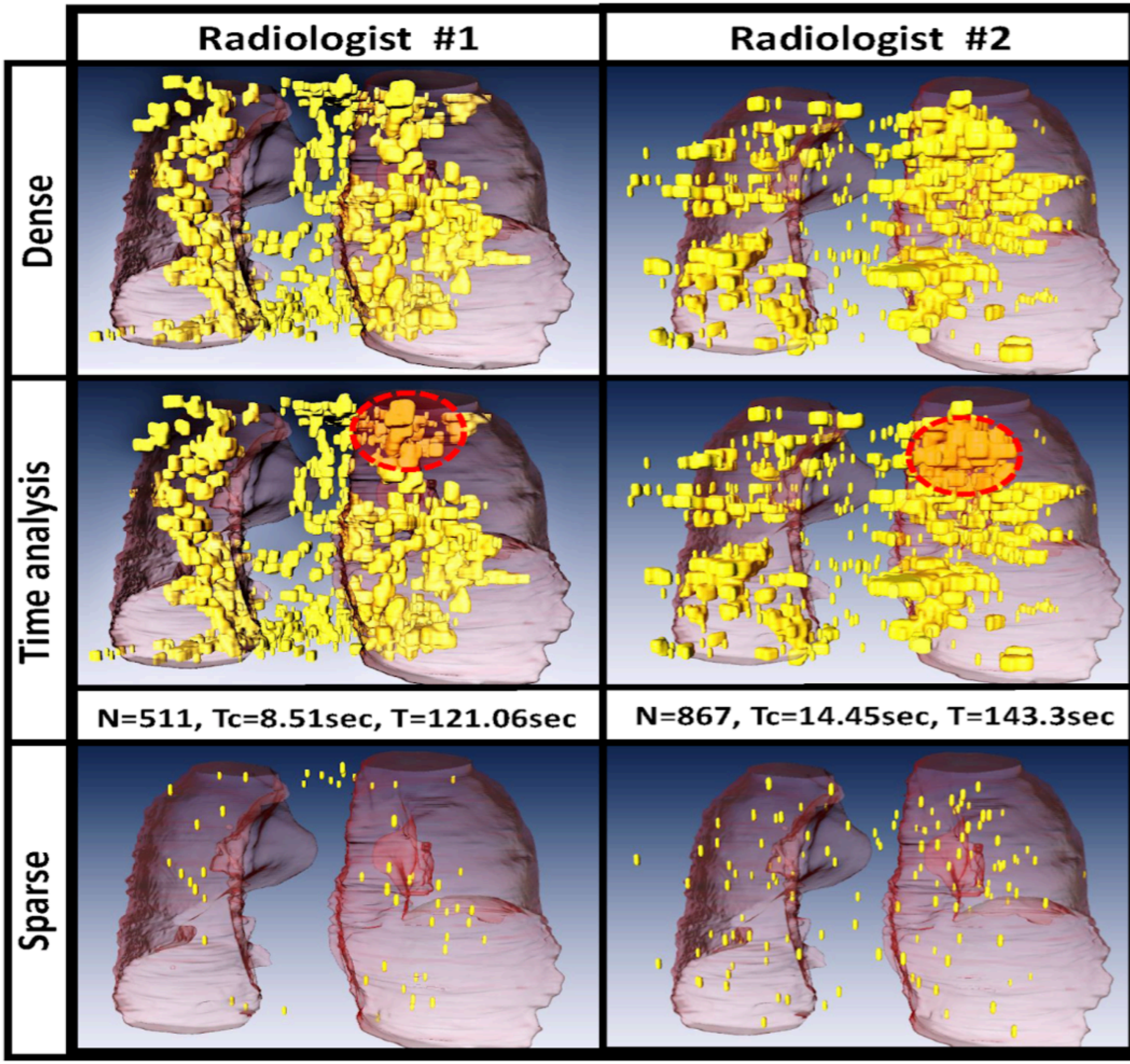
Raw data

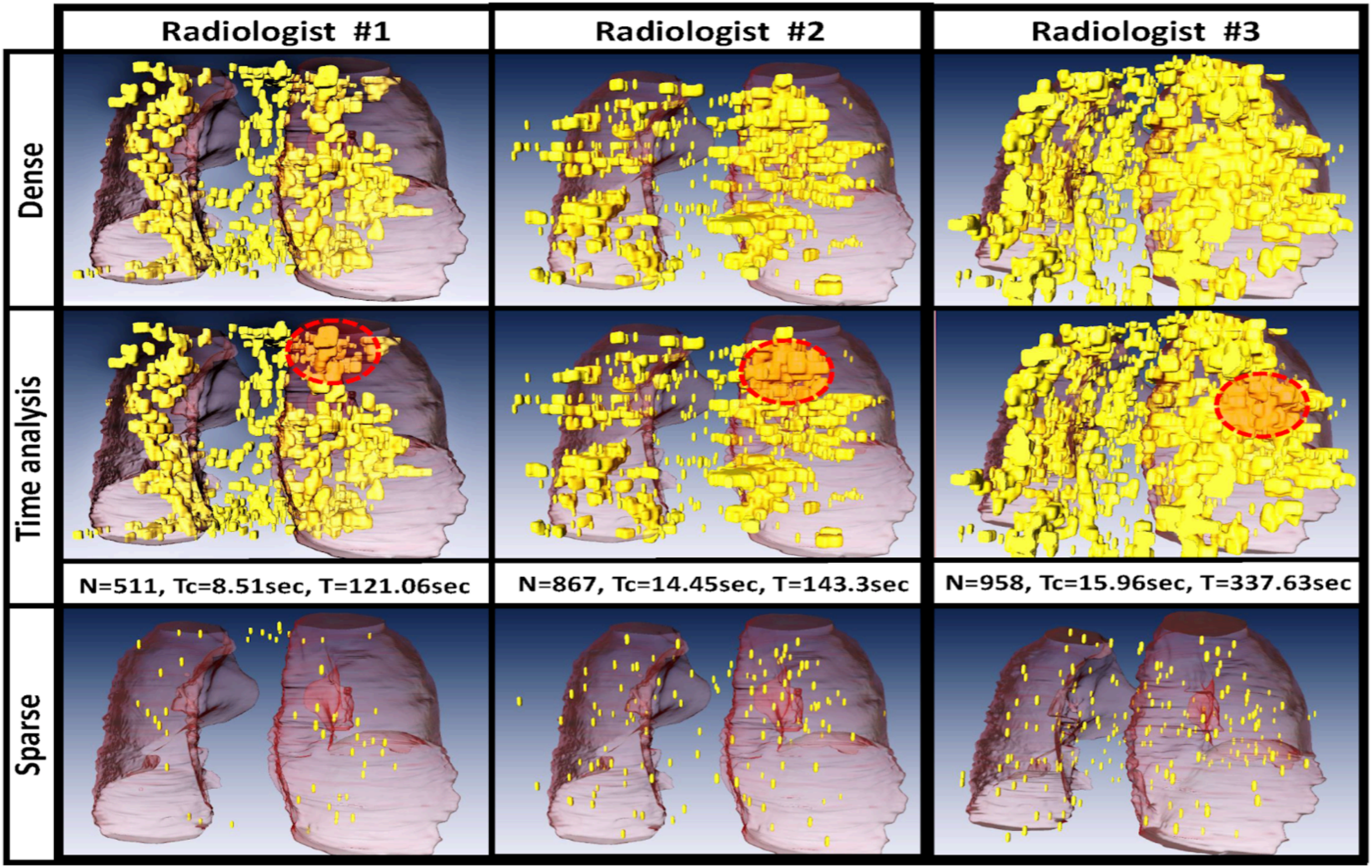
Clustered data

Graph Sparsification



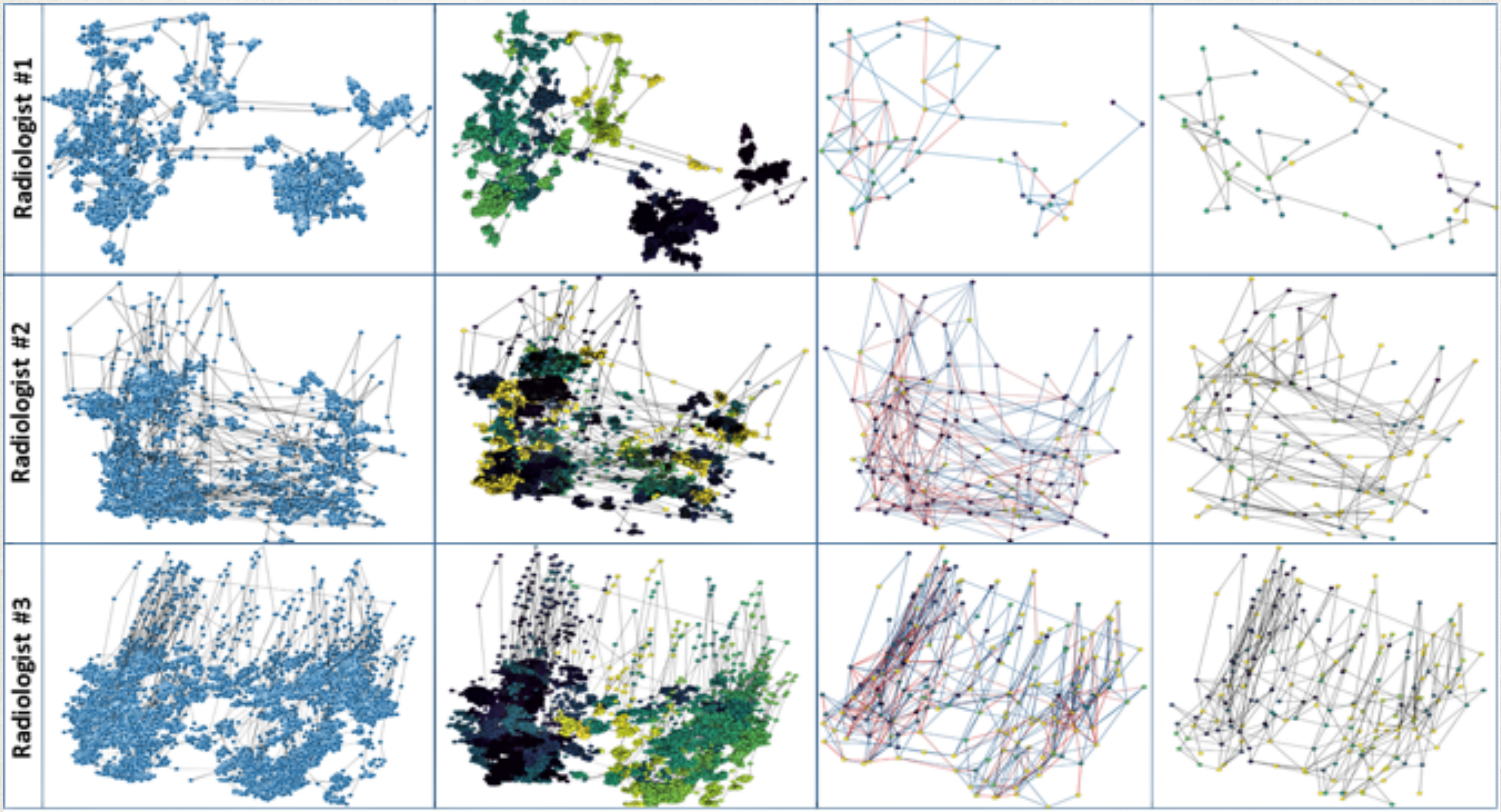
Radiologist #1
Dense
Time analysis
N=511, Tc=8.51sec, T=121.06sec
Sparse


Radiologist #1
Radiologist #2
Dense
Time analysis
N=511, Tc=8.51sec, T=121.06sec
N=867, Tc=14.45sec, T=143.3sec
Sparse



Radiologist #1
Radiologist #2
Radiologist #3
Dense
Time analysis
Sparse
N=511, Tc=8.51sec, T=121.06sec
N=867, Tc=14.45sec, T=143.3sec
N=958, Tc=15.96sec, T=337.63sec

Raw data	Clustered data	Graph build on cluster centers data	Sparsified data



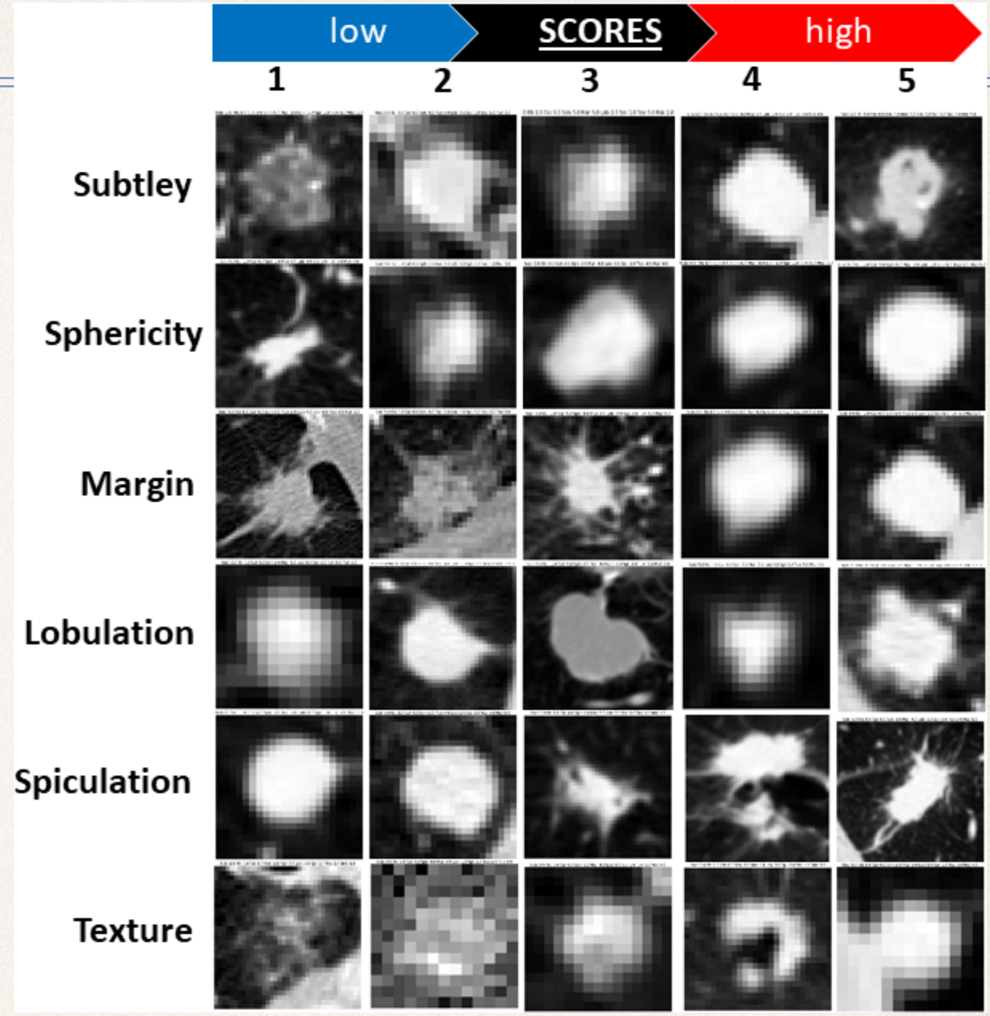
FPs/scan	0.125	0.25	0.5	1	2	4	8	Average
Sensitivity	0.773	0.870	0.924	0.941	0.962	0.980	0.986	**0.919**

Diagnosis via Visual Explanations

Visual Explanations via Attributes

Every lung nodule is associated with 6 attributes provided by the radiologist:

–Calcification
–Sphericity
–Margin
–Lobulation
–Spiculation
–Texture



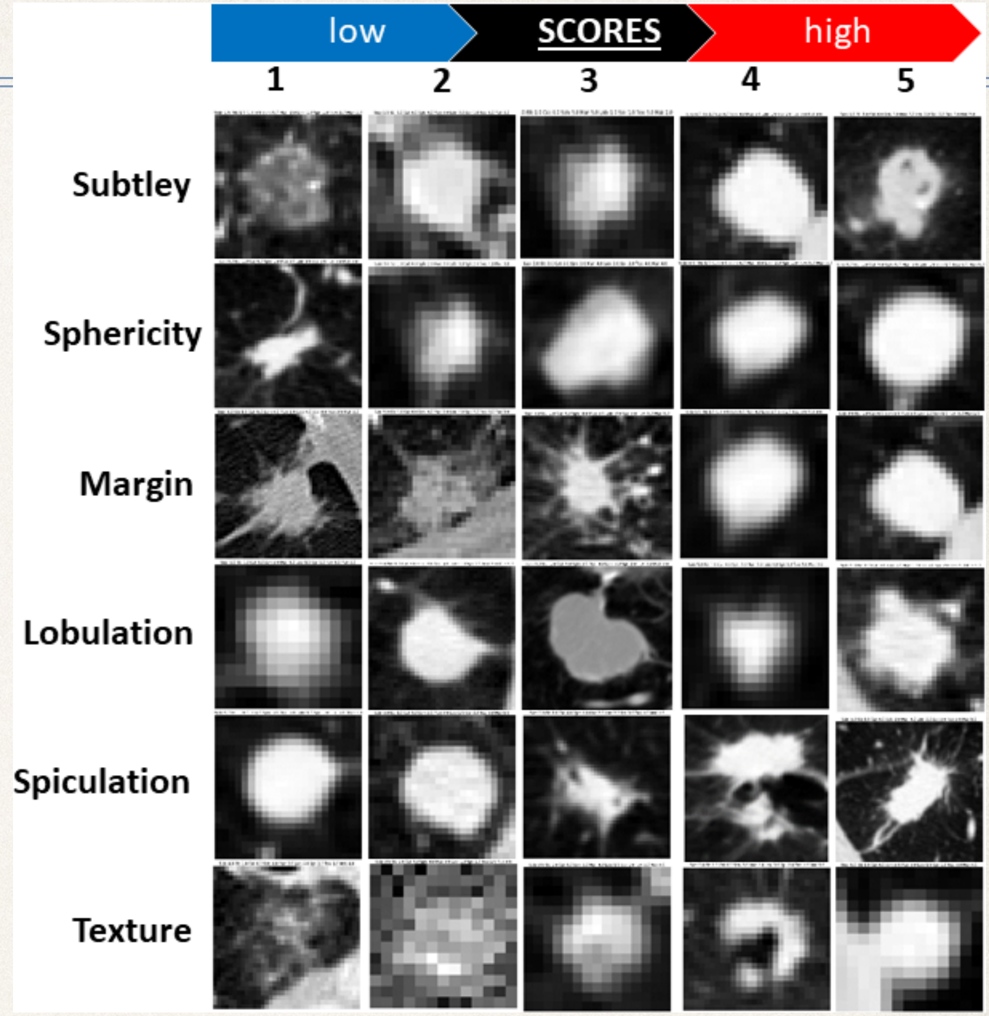
Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems

Richard Tomsett [1] **Dave Braines** [1 2] **Dan Harborne** [2] **Alun Preece** [2] **Supriyo Chakraborty** [3]

Visual Explanations via Attributes

Every lung nodule is associated with 6 attributes provided by the radiologist:

–Calcification
–Sphericity
–Margin
–Lobulation
–Spiculation
–Texture

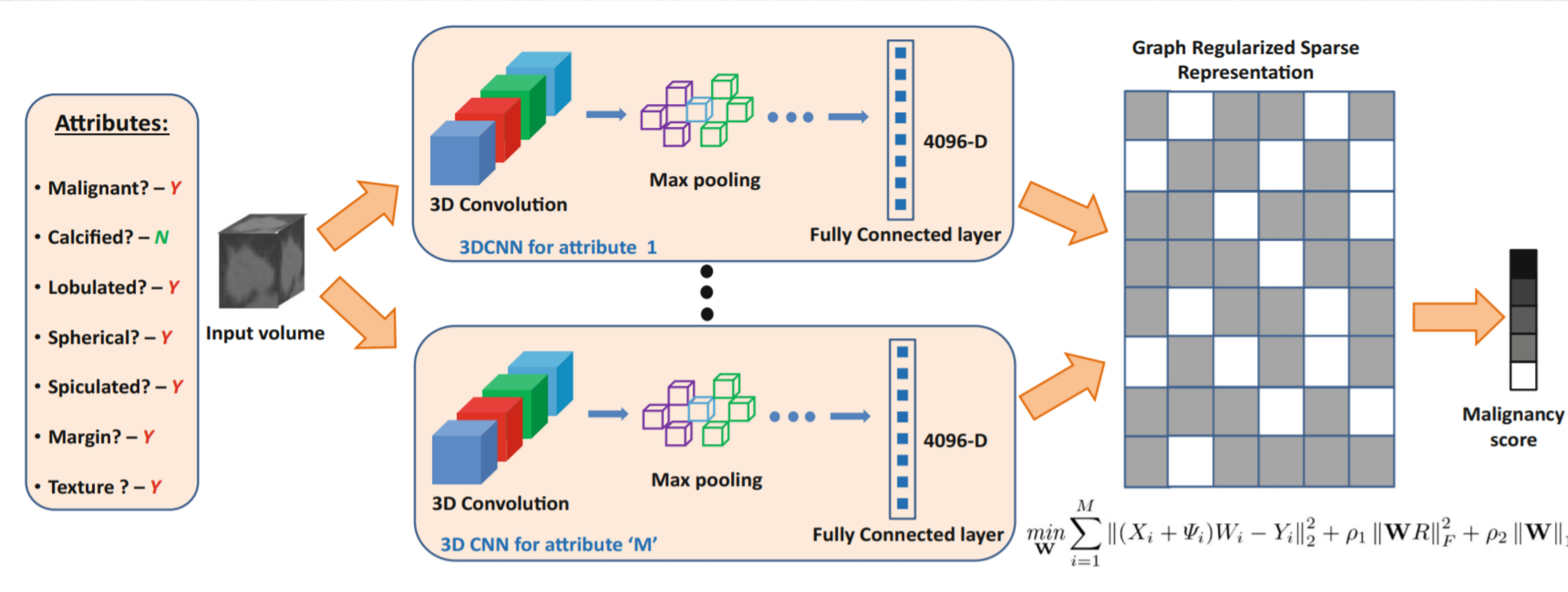


•*We explore the significance of these attributes to determine malignancy*
•*We concatenated these attribute score with 4096 dimension feature vector of CNN and perform Gaussian Progress regression*
•*LIDC-IDRI data base was used (1018 CT scans), multiple radiologists annotated the data sets.*

Multi-Task Learning of Visual Attributes

Hussein, et al, ISBI 2017

Hussein et al, IPMI, 2017



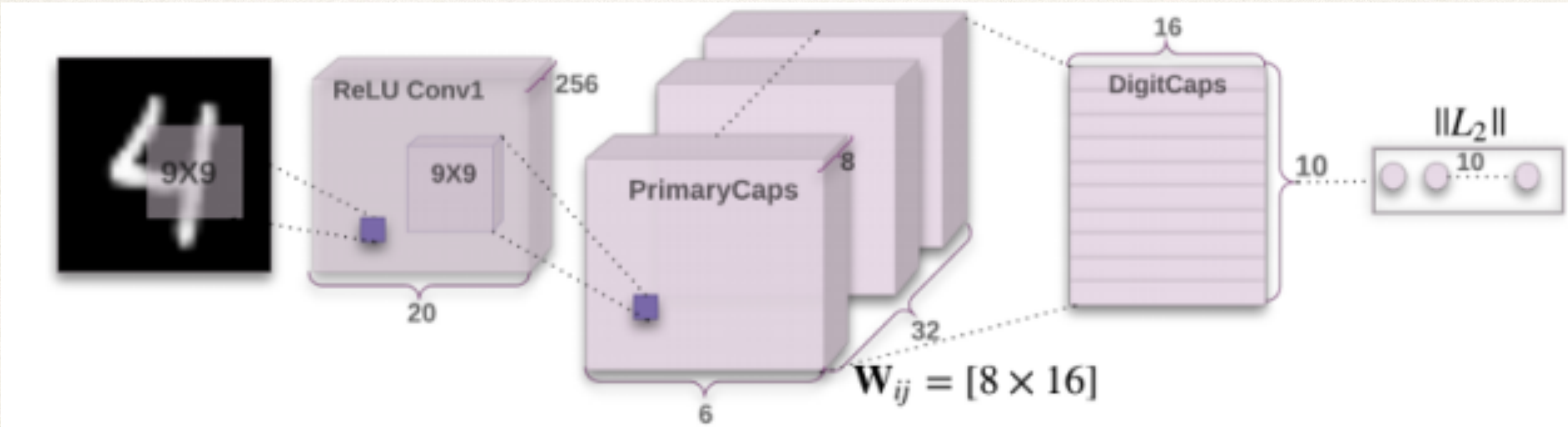
Methods	Accuracy	Mean Score Diff
GIST+LASSO	76.83%	0.6753
3D CNN MTL + Trace	80.08%	0.6259
Proposed approach	**91.26%**	**0.4593**

Drawbacks

- The requirement of large scale well annotated data
- Lack of object-part relationship with typical CNNs
- Fragile nature of the CNN systems (easily fooled!)

CapsNet (Capsule Networks)

Sabour, et al. NeurIPS 17

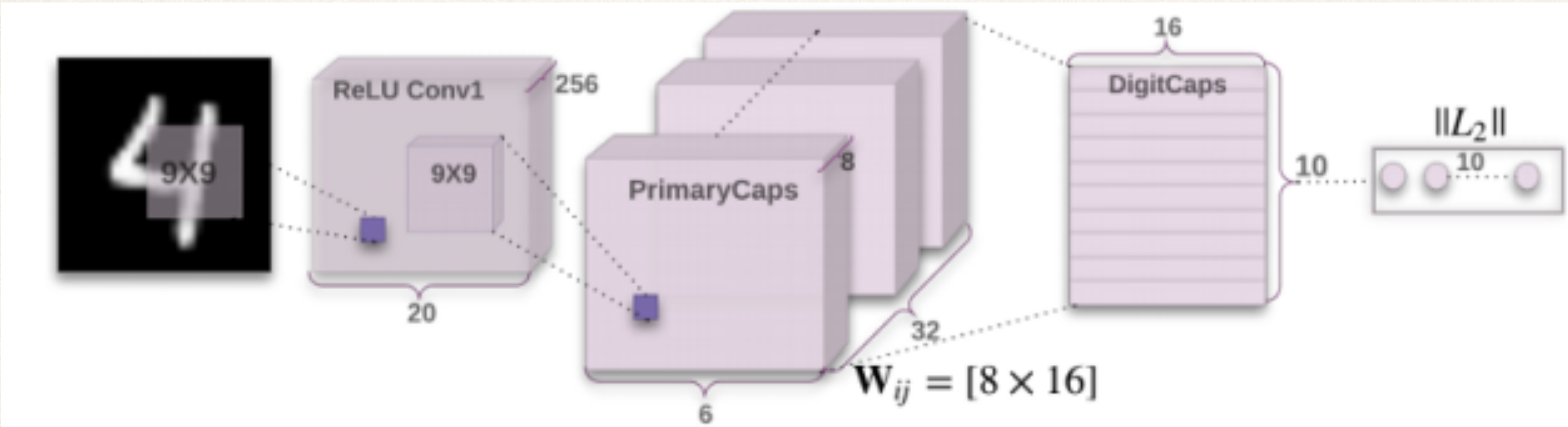


Two Simple Changes from CNNs:

- Features are now represented as **vectors** rather than scalars.
  - Vectors **store orientation information** about the input.
- Agreement between feature "predictions" is computed to **weight the presence and orientation of higher-level features**.

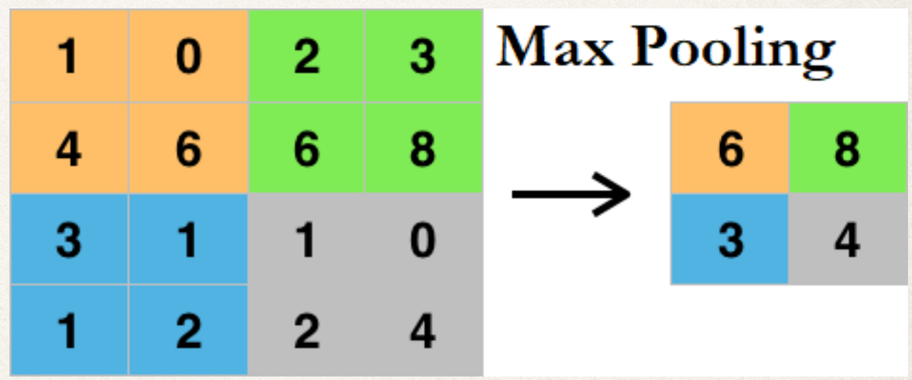
CapsNet (Capsule Networks)

Sabour, et al. NeurIPS 17



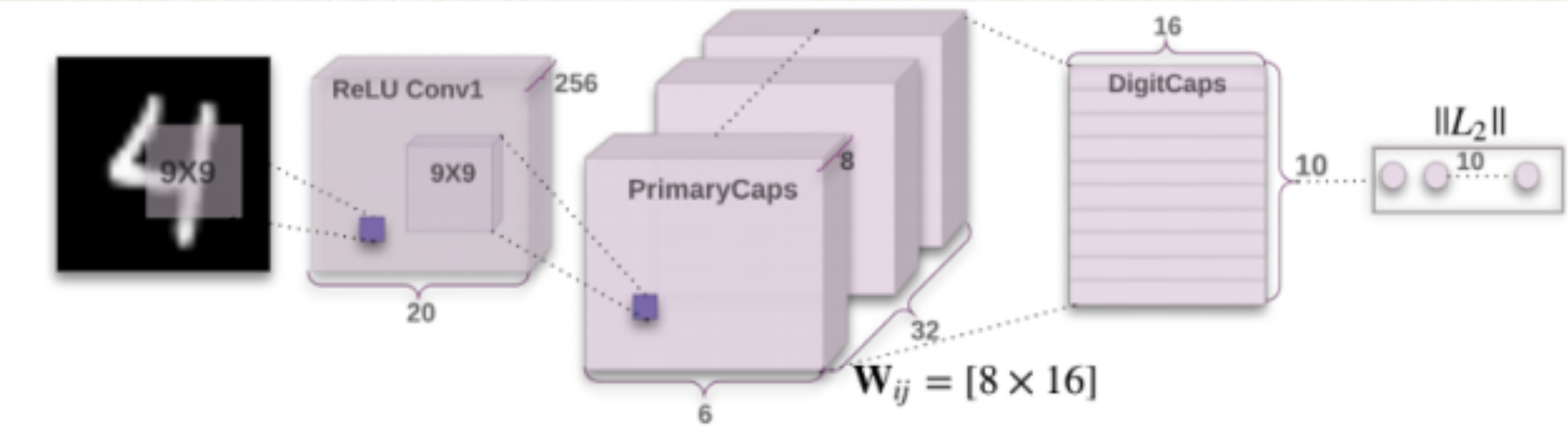
DRAWBACKS OF POOLING

- A form of routing, just an unintelligent one.
- **Pro:** Some spatial-invariance.
- **Pro:** Reduces memory burden.
- **Con:** Throws away information without regard to importance/ heterogeneity of the region.
- **Capsules** use strided-overlapping convolutions and dynamic routing.



CapsNet (Capsule Networks)

Sabour, et al. NeurIPS 17



- Requires less training data for good generalization.
- Preserves part-whole relationships and shape information.
- Capsule vectors encode information about input.

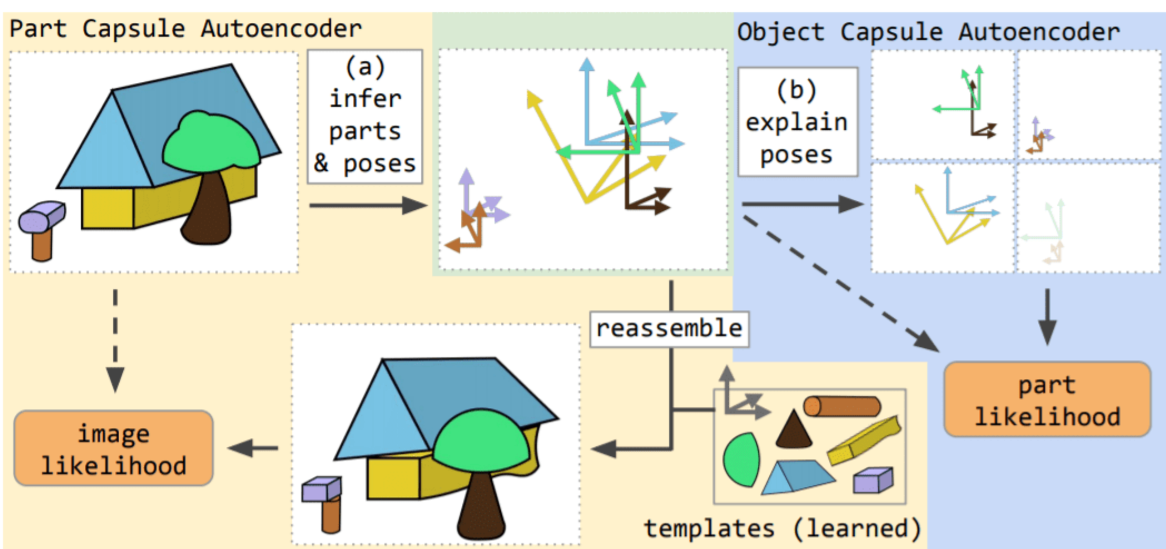
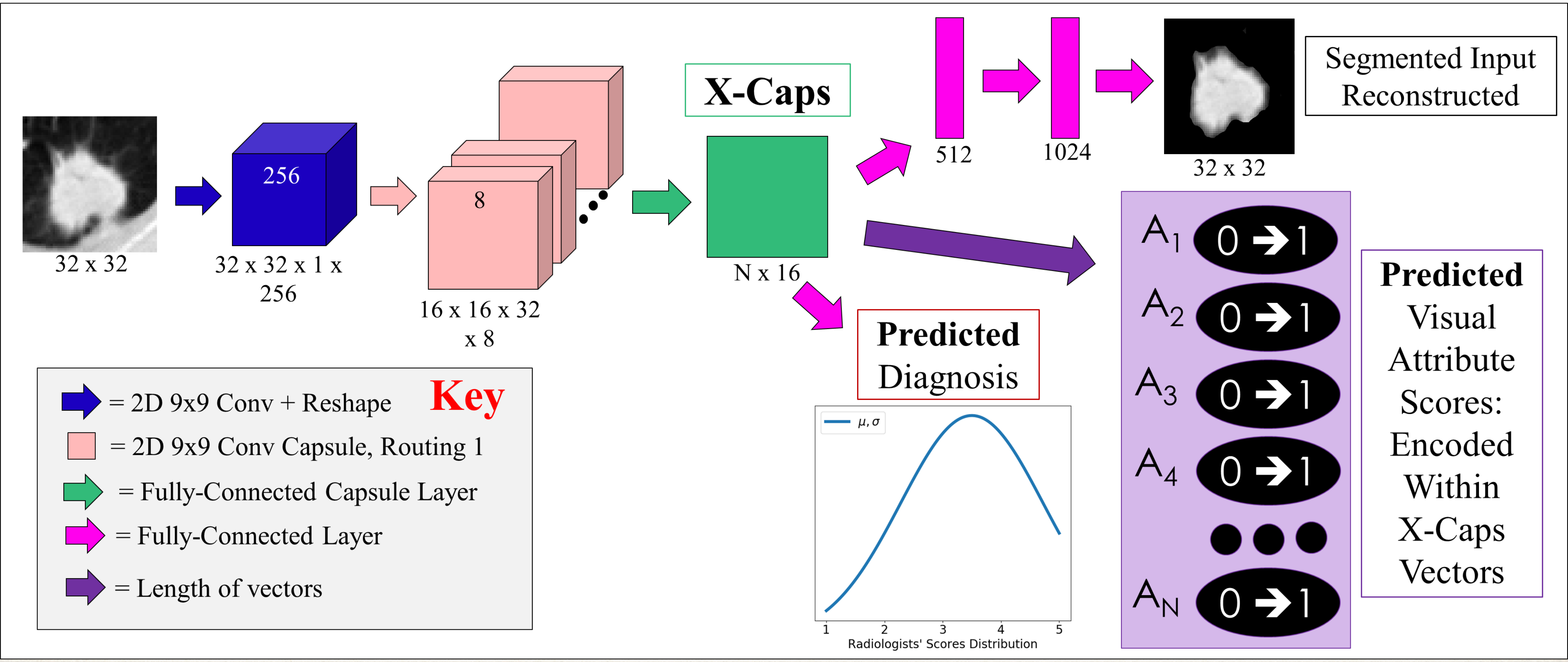


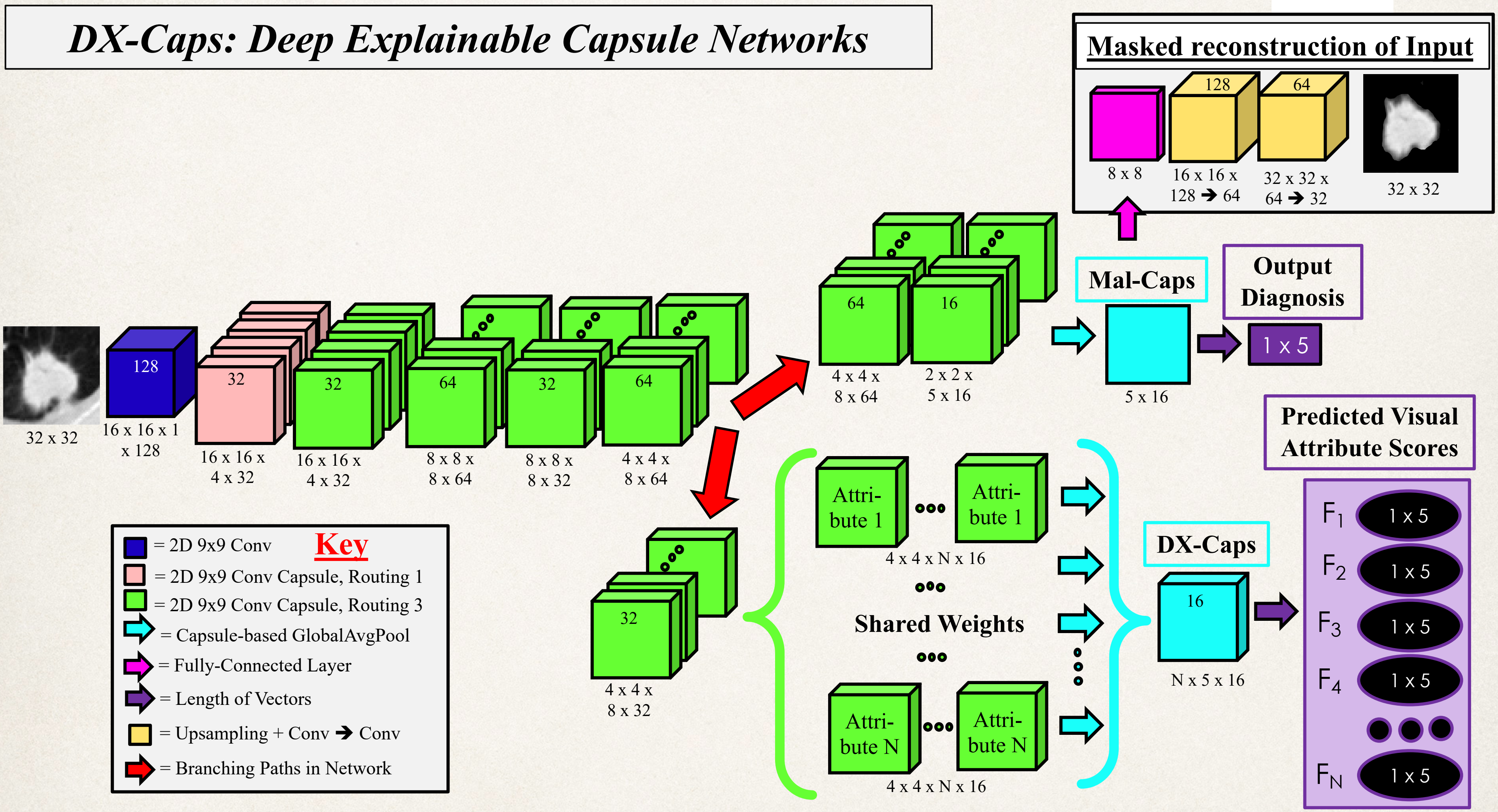
Figure 1: Stacked Capsule Autoencoder (SCAE): (a) *part* capsules segment the input into parts and their poses. The poses are then used to reconstruct the input by affine-transforming learned templates. (b) *object* capsules try to arrange inferred poses into objects, thereby discovering underlying structure. SCAE is trained by maximizing image and part log-likelihoods subject to sparsity constraints.

X-Caps

X-Caps: Explainable Capsule Networks



DX-Caps


DX-Caps: Deep Explainable Capsule Networks
Masked reconstruction of Input
128
64
8 x 8
16 x 16 x 128 ➔ 64
32 x 32 x 64 ➔ 32
32 x 32
Mal-Caps
Output Diagnosis
1 x 5
5 x 16
32 x 32
128
16 x 16 x 1 x 128
32
16 x 16 x 4 x 32
32
16 x 16 x 4 x 32
64
8 x 8 x 8 x 64
32
8 x 8 x 8 x 32
64
4 x 4 x 8 x 64
64
4 x 4 x 8 x 64
16
2 x 2 x 5 x 16
Predicted Visual Attribute Scores
Key
= 2D 9x9 Conv
= 2D 9x9 Conv Capsule, Routing 1
= 2D 9x9 Conv Capsule, Routing 3
= Capsule-based GlobalAvgPool
= Fully-Connected Layer
= Length of Vectors
= Upsampling + Conv ➔ Conv
= Branching Paths in Network
32
4 x 4 x 8 x 32
Attri-bute 1
Attri-bute 1
4 x 4 x N x 16
Shared Weights
Attri-bute N
Attri-bute N
4 x 4 x N x 16
DX-Caps
16
N x 5 x 16
F1 1 x 5
F2 1 x 5
F3 1 x 5
F4 1 x 5
FN 1 x 5

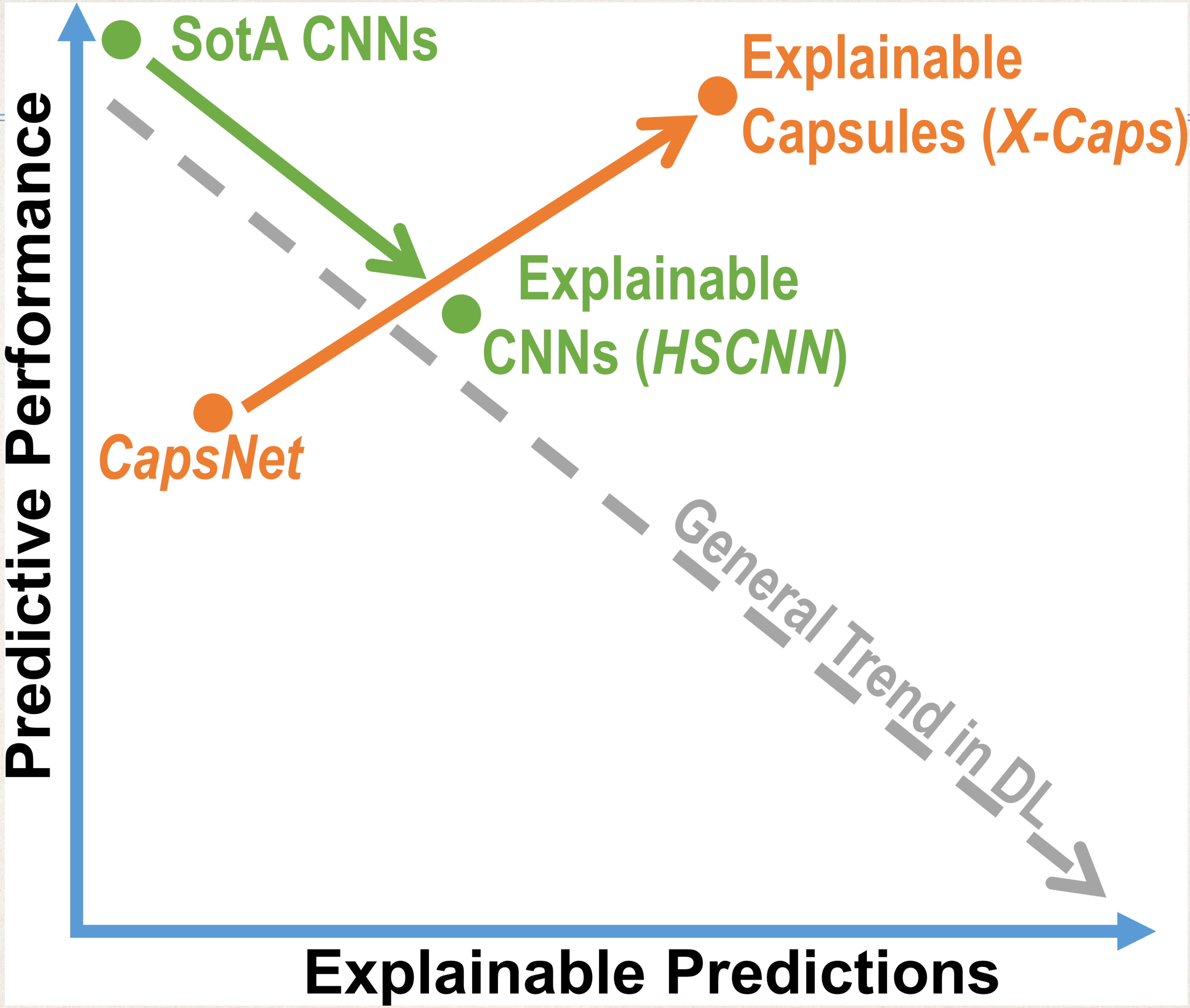

SotA CNNs
Explainable Capsules (*X-Caps*)
Explainable CNNs (*HSCNN*)
CapsNet
General Trend in DL
Predictive Performance
Explainable Predictions

Table 2: Prediction accuracy of visual attribute learning with capsule networks. Dashes (-) represent values which the given method could not produce. *X-Caps* significantly outperforms the state-of-the-art explainable method (*HSCNN*) at attribute modeling (the main goal of both studies), while also producing higher malignancy prediction scores, approaching state-of-the-art non-explainable methods performance.

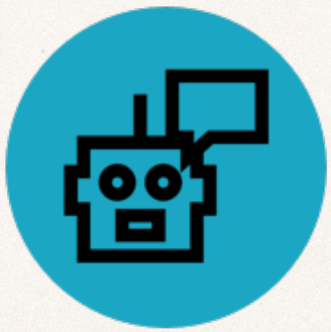
	Attribute Prediction Accuracy %						Malignancy Accuracy %
	subtlety	sphericity	margin	lobulation	spiculation	texture	
Non-Explainable Methods							
3D Multi-Scale + RF [31]	-	-	-	-	-	-	86.84
3D Multi-Crop [32]	-	-	-	-	-	-	87.14
3D Multi-Out-DenseNet [6]	-	-	-	-	-	-	90.40
3D Dual-Path GBM [38]	-	-	-	-	-	-	90.44
CapsNet [28]	-	-	-	-	-	-	77.04
Explainable Methods							
3D Dual-Path-Dense *HSCNN* [30]	71.9	55.2	72.5	-	-	83.4	84.20
Proposed *X-Caps*	**90.39**	**85.44**	**84.14**	**70.69**	**75.23**	**93.10**	**86.39**

Concluding Remarks

1 **STRENGTHEN TRUST AND TRANSPARENCE**

Without understanding the contribution of each explanatory variable to the outcome, we will have no guarantee that the model will make a relevant and fair recommendation.

Concluding Remarks



2 **EXPLAIN DECISIONS**

An interpretable Machine Learning model allows humans to understand the proposed outcome and establish the diagnosis.

Concluding Remarks

3 IMPROVE THE MODELS

Interpretability ensures data scientists that the model is good for the right reasons and wrong for the right reasons as well. Interpretability offers new possibilities for feature engineering and model debugging.

Concluding Remarks

Visual search is gold standard; however, prone to errors, malpractice/ perceptual error etc., time consuming, and sub-optimal

Concluding Remarks

Visual search is gold standard; however, prone to errors, malpractice/ perceptual error etc., time consuming, and sub-optimal

Computer (AI) helps radiologists to find pathologies that can be missed however, computer also depicts so many false positives that radiologists easily capture them with true labels!

Concluding Remarks

Visual search is gold standard; however, prone to errors, malpractice/ perceptual error etc., time consuming, and sub-optimal

Computer (AI) helps radiologists to find pathologies that can be missed however, computer also depicts so many false positives that radiologists easily capture them with true labels!

We can design new AI tools that are more intelligent and less artificial by collaborating with humans (experts)!

Concluding Remarks

Visual search is gold standard; however, prone to errors, malpractice/ perceptual error etc., time consuming, and sub-optimal

Computer (AI) helps radiologists to find pathologies that can be missed however, computer also depicts so many false positives that radiologists easily capture them with true labels!

We can design new AI tools that are more intelligent and less artificial by collaborating with humans (experts)!

Explainability plays an important role in building thrust and robust systems; hence, increasing the chance of deploying such system in real clinic

Thank you!