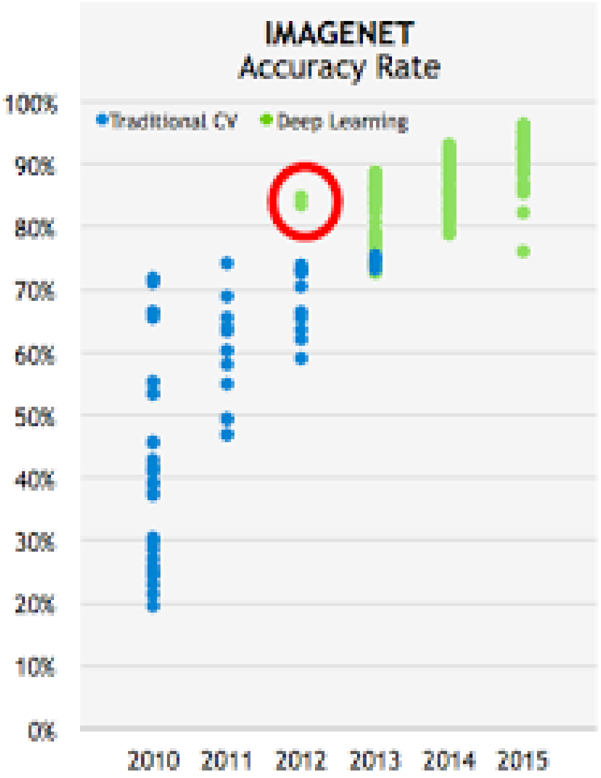
Creating Innovations that Matter

Deep Learning for Medical Imaging

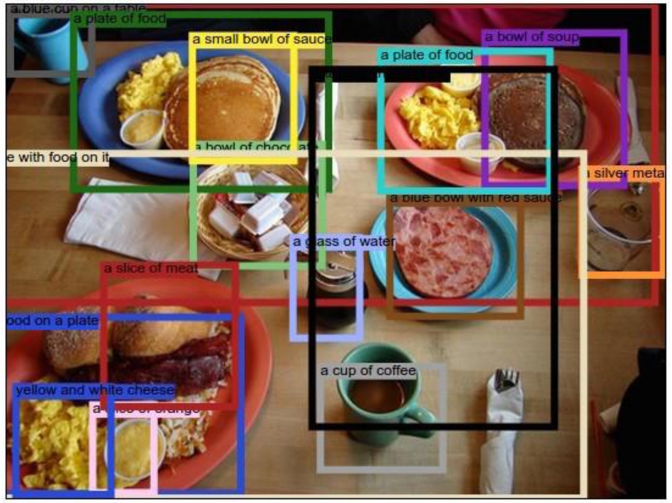
Christine Swisher, PhD

Guest Seminar, MIT Course 6.S897/HST.S53: Machine Learning for Healthcare Spring 2017

Philips Research North America

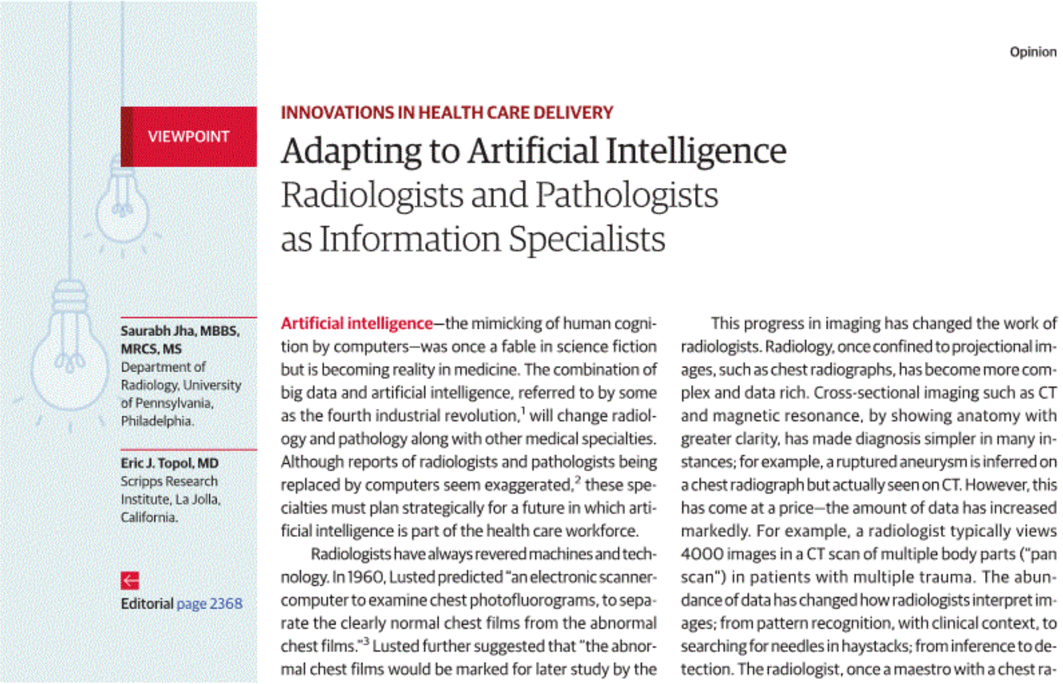


ML watches YouTube for three straight days!
(and learns to recognize cats?!)



a plate of food. food on a plate. a blue cup on a table. a plate of food. a blue bowl with red sauce. a bowl of soup. a cup of coffee. […]

PHILIPS



Opinion

VIEWPOINT

INNOVATIONS IN HEALTH CARE DELIVERY

Adapting to Artificial Intelligence
Radiologists and Pathologists as Information Specialists

Saurabh Jha, MBBS, MRCS, MS
Department of Radiology, University of Pennsylvania, Philadelphia.

Eric J. Topol, MD
Scripps Research Institute, La Jolla, California.

Editorial page 2368

Artificial intelligence—the mimicking of human cognition by computers—was once a fable in science fiction but is becoming reality in medicine. The combination of big data and artificial intelligence, referred to by some as the fourth industrial revolution,[1] will change radiology and pathology along with other medical specialties. Although reports of radiologists and pathologists being replaced by computers seem exaggerated,[2] these specialties must plan strategically for a future in which artificial intelligence is part of the health care workforce.

Radiologists have always revered machines and technology. In 1960, Lusted predicted "an electronic scanner-computer to examine chest photofluorograms, to separate the clearly normal chest films from the abnormal chest films."[3] Lusted further suggested that "the abnormal chest films would be marked for later study by the

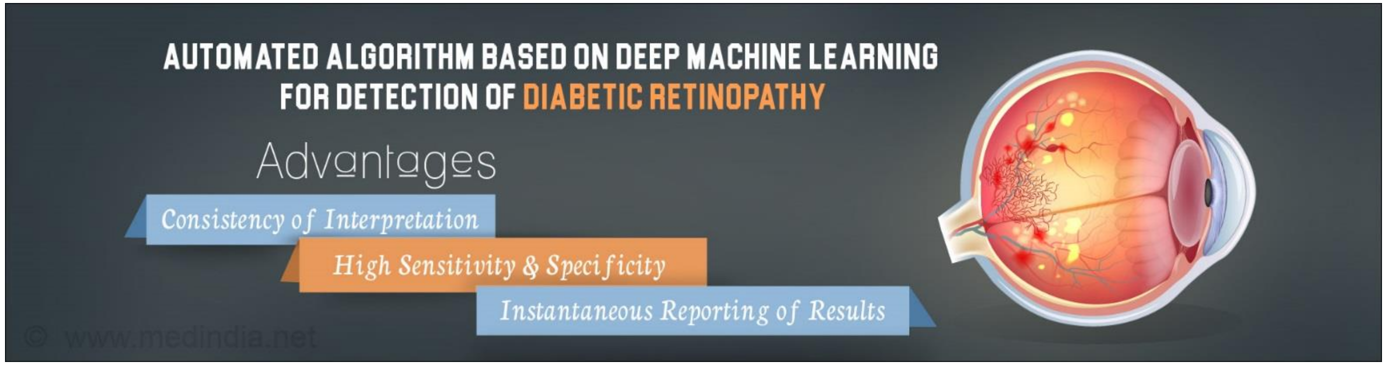
This progress in imaging has changed the work of radiologists. Radiology, once confined to projectional images, such as chest radiographs, has become more complex and data rich. Cross-sectional imaging such as CT and magnetic resonance, by showing anatomy with greater clarity, has made diagnosis simpler in many instances; for example, a ruptured aneurysm is inferred on a chest radiograph but actually seen on CT. However, this has come at a price—the amount of data has increased markedly. For example, a radiologist typically views 4000 images in a CT scan of multiple body parts ("pan scan") in patients with multiple trauma. The abundance of data has changed how radiologists interpret images; from pattern recognition, with clinical context, to searching for needles in haystacks; from inference to detection. The radiologist, once a maestro with a chest ra-

"Deep learning technology applied to medical imaging may become the most disruptive technology radiology has seen since the advent of digital imaging." –Nadim Daher

"Radiologists and pathologists need not fear artificial intelligence but rather must adapt incrementally to artificial intelligence, retaining their own services for cognitively challenging tasks." –Eric Topol


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
LESIONS LEARNT
Artificial intelligence powers detection of skin cancer from images PAGES 36 & 115


Chest Defense: Deep Learning Spots Disease Early Using Chest X-Rays


AUTOMATED ALGORITHM BASED ON DEEP MACHINE LEARNING FOR DETECTION OF DIABETIC RETINOPATHY
Advantages
Consistency of Interpretation
High Sensitivity & Specificity
Instantaneous Reporting of Results
© www.medindia.net

Deep Learning is Everywhere!

A Street Vendor in China

Deep Learning Service - System Development & Testing
Caffe installation: 10 Yuan = $1.5
CNN: 5 Yuan = $0.75 per layer
RNN: 8 Yuan = $1.2 per layer

Slide borrowed from Hua Xie, Philips Research North America

Machine Learning that Matters

Kiri L. Wagstaff KIRI.L.WAGSTAFF@JPL.NASA.GOV
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA

Link to paper

The three rules of meaningful ML innovation still apply

1. Eyes on the Prize
2. Involvement of the World Outside of ML
3. Meaningful Evaluation Methods

“With this positive trial result (NLST), we have the opportunity to realize the greatest single reduction of cancer mortality in the history of the war on cancer.”

– James Mulshine, MD

Three rules of meaningful ML innovation

1. Eyes on the Prize
    - How significant is the impact of a solution to the problem?
    - How many lives would it change? What is a severe unmet need we can overcome?
    - What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
    - Co-creation with clinicians
    - Feedback from hospital infrastructure and hospital administrator
    - Involve experts in business models, marketing & sales
    - ***Know your data!!!***

3. Meaningful Evaluation Methods
    - Performance in multisite clinical trails
    - Machine vs Human vs Machine + Human
    - Improvement of clinical outcome

Lung Screening at a Glance

IT CAUSES A LOT OF DEATHS

Lung cancer is the **number-one cancer killer**, taking more lives than colon, breast and prostate cancer combined.

Urgent need: Lung cancer **kills 450 people every day** in the US alone.

Source: Onco Iss 2014

EARLY DIAGNOSIS IS CRITICAL

Reduced Mortality:
Generally, early detection can increase five-year survival by nearly 90%.

Source: NEJM 2006

EXPECTED WIDESPREAD ADOPTION

In 2015, the CMS added annual screening for lung cancer with LDCT ensuring that **3-4 million** high-risk patients could get lifesaving intervention regardless of income level.
Source: NYTimes 2014.

Recommendation by NCCN and USPSTF .

Failure to screen lawsuits favor patients
Ex: DC jury awards $5M for failure to screen for cancer

Lung Screening at a Glance

IT CAUSES A LOT OF DEATHS

Lung cancer is the **number-one cancer killer**, taking more lives than colon, breast and prostate cancer combined.

Urgent need: Lung cancer **kills 450 people every day** in the US alone.

Source: Onco Iss 2014

EARLY DIAGNOSIS IS CRITICAL

Reduced Mortality:
Generally, early detection can increase five-year survival by nearly 90%.

Source: NEJM 2006

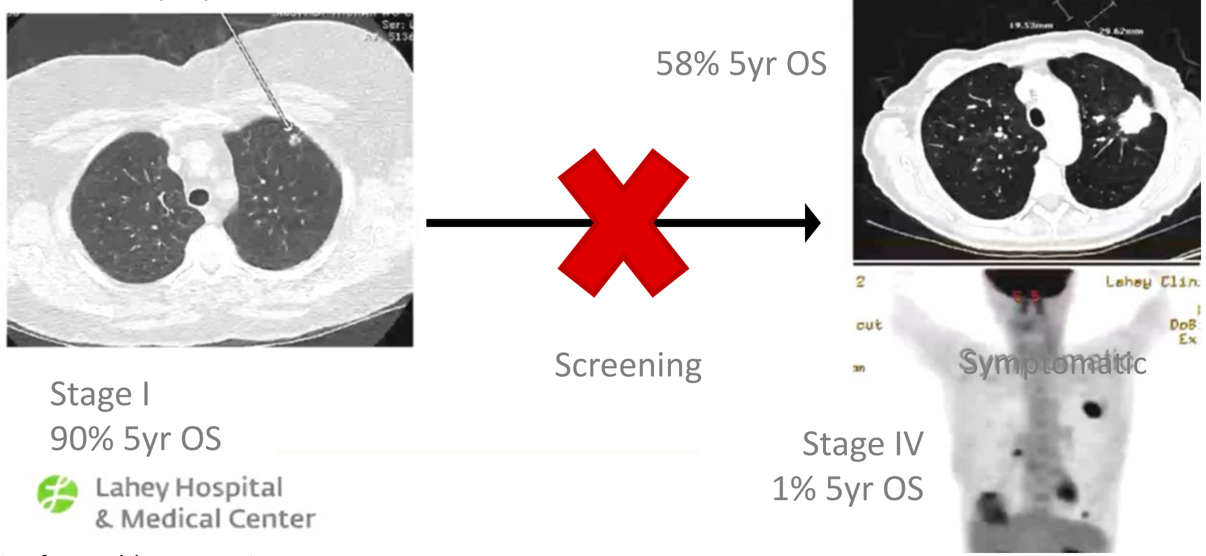
EXPECTED WIDESPREAD ADOPTION

In 2015, the CMS added annual screening for lung cancer with LDCT ensuring that **3-4 million** high-risk patients could get lifesaving intervention regardless of income level.
Source: NYTimes 2014.

Recommendation by NCCN and USPSTF .

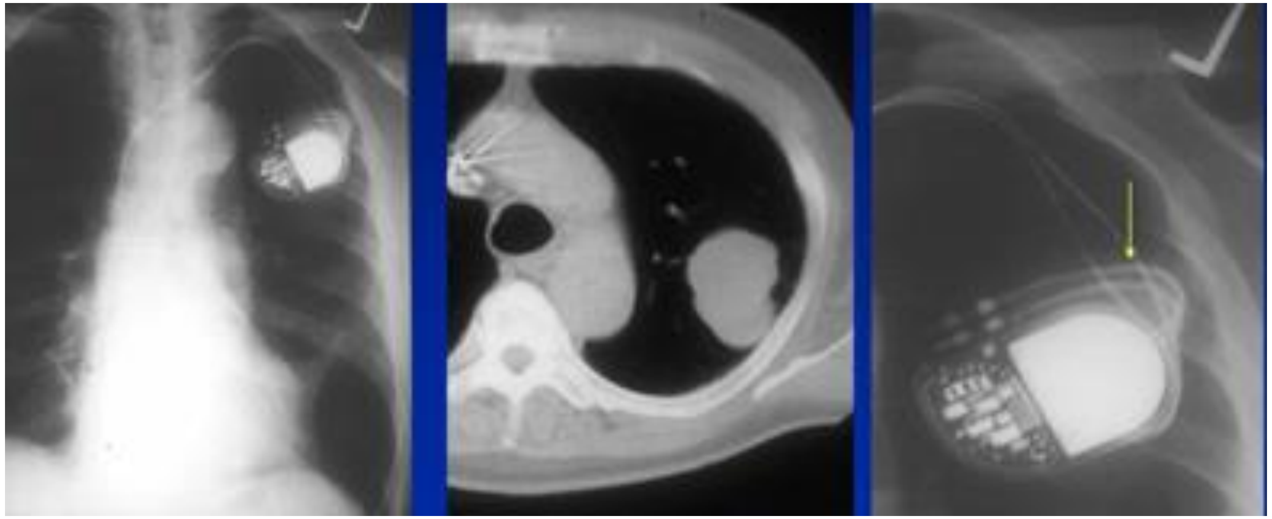
Failure to screen lawsuits favor patients
Ex: DC jury awards $5M for failure to screen for cancer

Asymptomatic



PHILIPS

Challenges for Adoption of LDCT



Cognitive Challenges:
- Vast majority are negative ~89.4%
- Satisfaction of search
- Volume and complexity of information

False Positives
- 96.4% FP of positive readings by LDCT
- Most have noninvasive imaging follow-up
- Invasive diagnosis procedure : 2.6%
- Complication rate: 1.4% (0.06% Major)

Overdiagnosis: More than 18% seem to be indolent.
- Bronchioloalveolar carcinoma 79% ; NSCLC 22% are overdiagnosed
- Risk: 11% by LDCT vs no screening and 9% vs CXR (lifetime follow-up)

Challenges for Adoption of LDCT

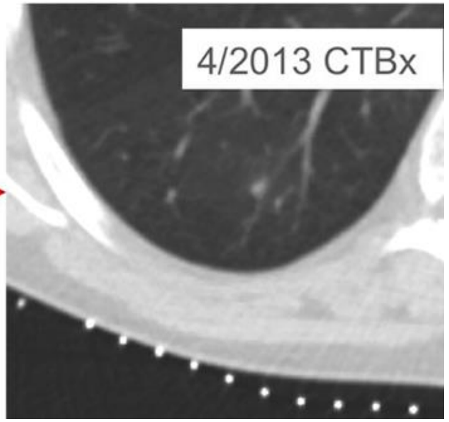
Cognitive Challenges:

- Vast majority are negative ~89.4%
- Satisfaction of search
- Volume and complexity of information

False Positives

- 96.4% FP of positive readings by LDCT
- Most have noninvasive imaging follow-up
- Invasive diagnosis procedure : 2.6%
- Complication rate: 1.4% (0.06% Major)

LDCT screen



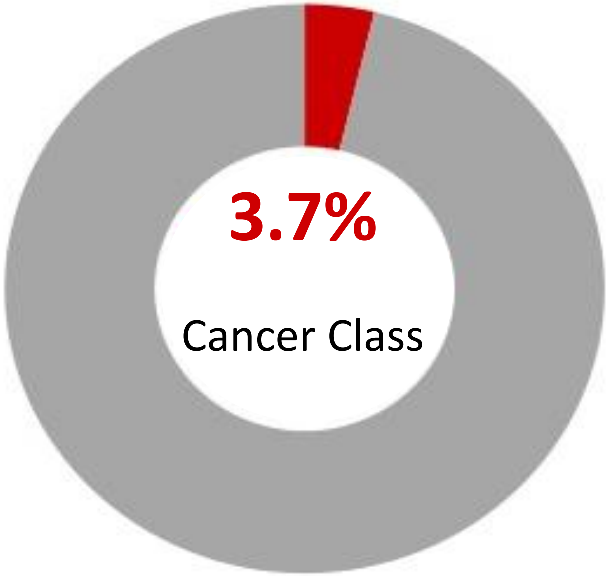
FP at pre-biopsy CT

Overdiagnosis: More than 18% seem to be indolent.

- Bronchioloalveolar carcinoma 79% ; NSCLC 22% are overdiagnosed
- Risk: 11% by LDCT vs no screening and 9% vs CXR (lifetime follow-up)

Class Imbalance

True positives and rare incidental findings, by virtue of being rare, are underrepresented. If not accounted for properly, the class imbalance will occur biasing the a model to predict the healthy-label.

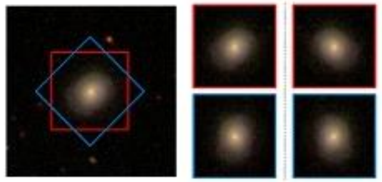


- 1000 samples (963 Negative; 37 positives)
- Network learns that all are negative
- Accuracy of 96.3% and PPV = 0

Class Imbalance

True positives and rare incidental findings, by virtue of being rare, are underrepresented. If not accounted for properly, the class imbalance will occur biasing the a model to predict the healthy-label.

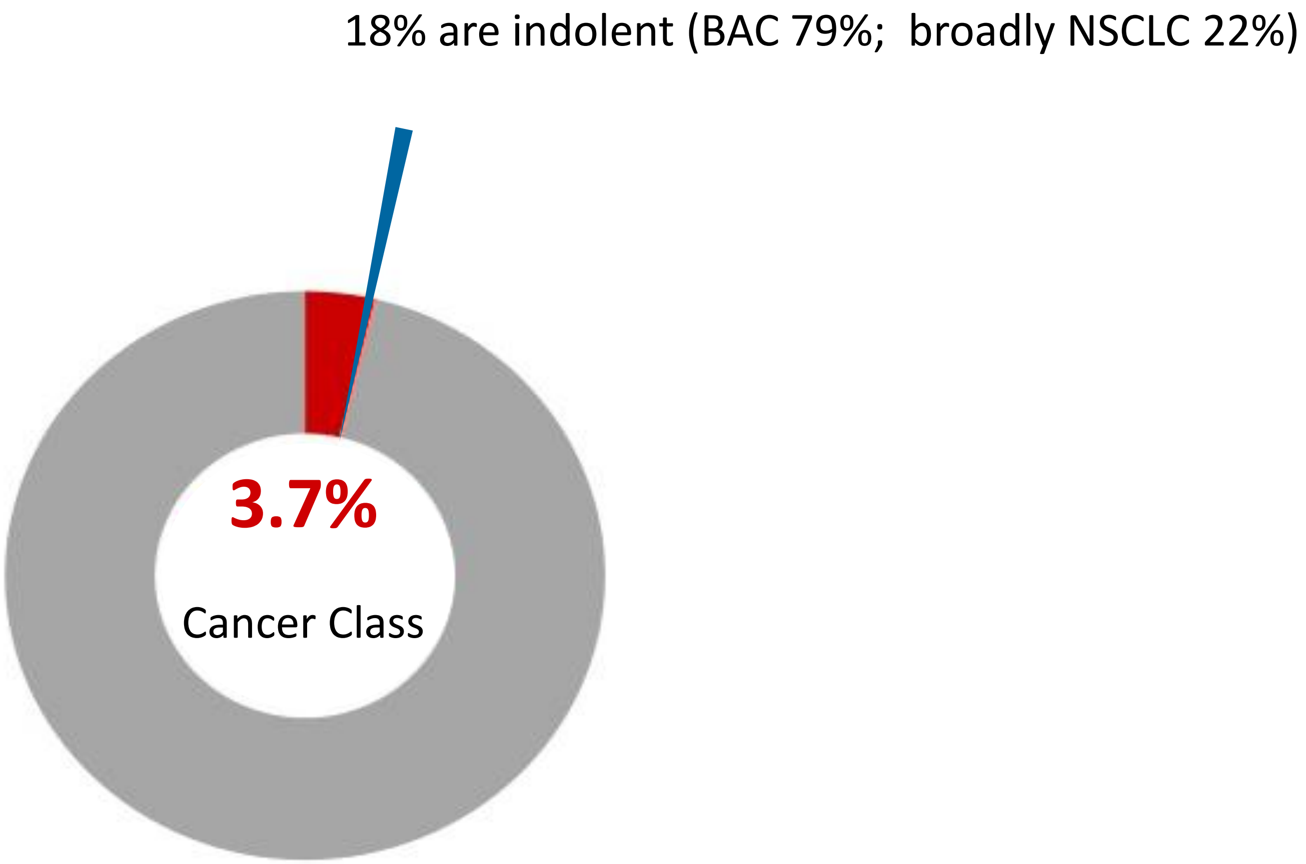
- Augmentation of underrepresented class*
- Train on an easier problem
- Weight the loss function
- Pre-training for lower level features



Source: Rotation-invariant convolutional neural networks for galaxy morphology prediction. Sander Dieleman et. al.

- Small rotations
- Small translation
- Scaling
- Flipping
- Brightness
- Noise

*Underrepresented class should have examples of various ways rare class can present.


18% are indolent (BAC 79%; broadly NSCLC 22%)
3.7%
Cancer Class

Goals

1. ***Reduce time*** and ***cognitive load*** for radiologists reading LDCT images
2. Reduce unnecessary escalation and resultant complications due to ***false positives*** reads

Three rules of meaningful ML innovation

1. Eyes on the Prize
   - How significant is the impact of a solution to the problem?
   - How many lives would it change? What is a severe unmet need we can overcome?
   - What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
   - Co-creation with clinicians
   - Feedback from hospital infrastructure and hospital administrator
   - Involve experts in business models, marketing & sales
   - ***Know your data!!!***

3. Meaningful Evaluation Methods
   - Performance in multisite clinical trails
   - Machine vs Human vs Machine + Human
   - Improvement of clinical outcome

PHILIPS

Value

Hospital:

- Reduce costs associated with unnecessary care escalation (10BE/yr on US health system)
- Reduced mis-diagnoses and resultant resource utilization
- Identify high risk patients for follow-up

Patient:

- Improved outcomes (quality of life, mortality, cost)

Staff:

- Increase staff efficiency (improve throughput/reduce radiologist man hours)

Health System:

- Estimates of total health expenditures for a national screening program range from $1B to $3B annually, constituting a 20% increase in expenditure for lung cancer overall.

What is the FDA approval process?

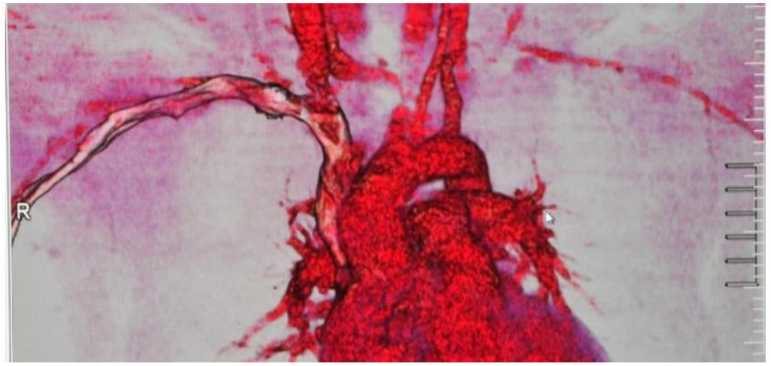
“Soft” Use-Case

- Current regulatory situation is reminiscent of the early days of computer-aided detection (CADe) devices.
- Cleared under the 510[k] process

“Hard” Use-case:

- Likely regulated as Class 2, even Class 3
- Requires a large randomized clinical trial
- Similar to computer-aided diagnosis (CADx) applications, which required premarket approval (PMA) process.

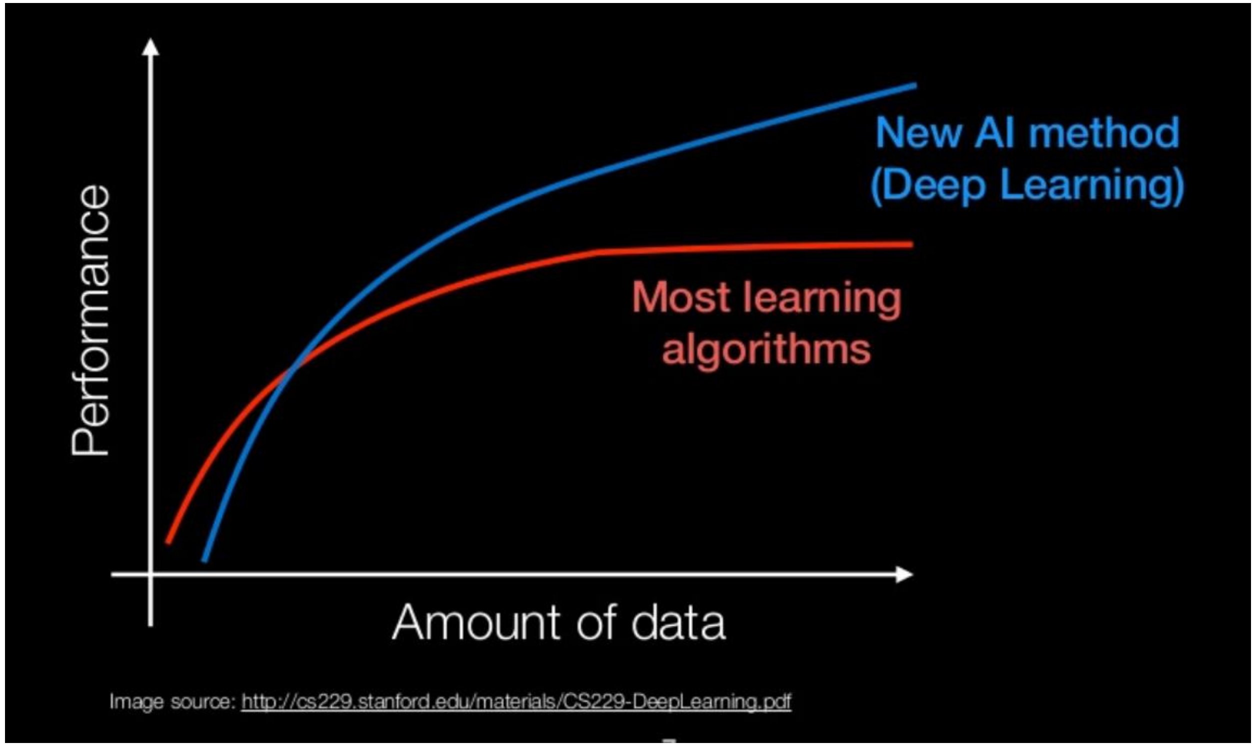
F FORBES BERNARD MARR



First FDA Approval For Clinical Cloud-Based Deep Learning In Healthcare

The FDA approval of a cloud based machine learning application to be used in a clinical setting to help physicians understand how a heart is functioning signals a major breakthrough. Cutting examination time from up to an hour to just 15 seconds paves the way for more AI algorithms in healthcare.

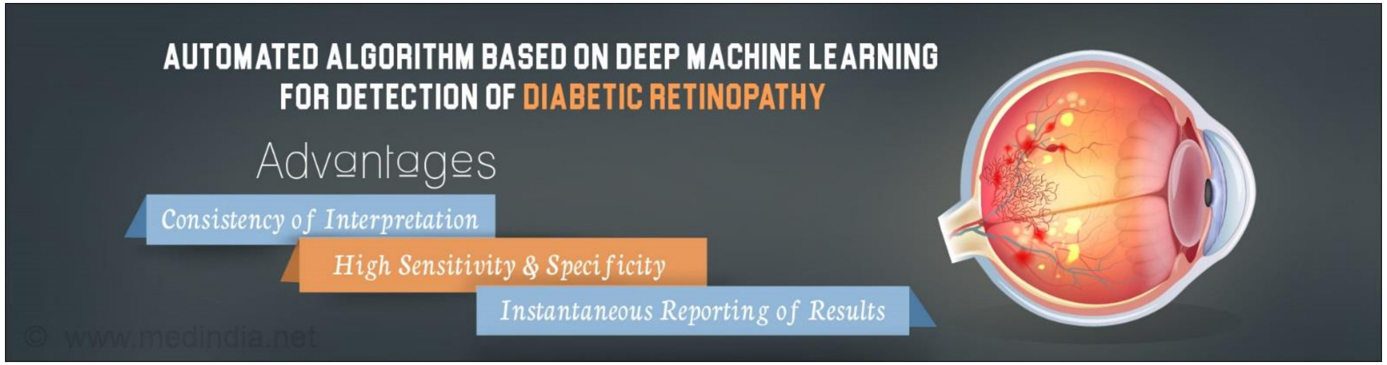
Data



New AI method
(Deep Learning)
Most learning
algorithms
Performance
Amount of data
Image source: http://cs229.stanford.edu/materials/CS229-DeepLearning.pdf

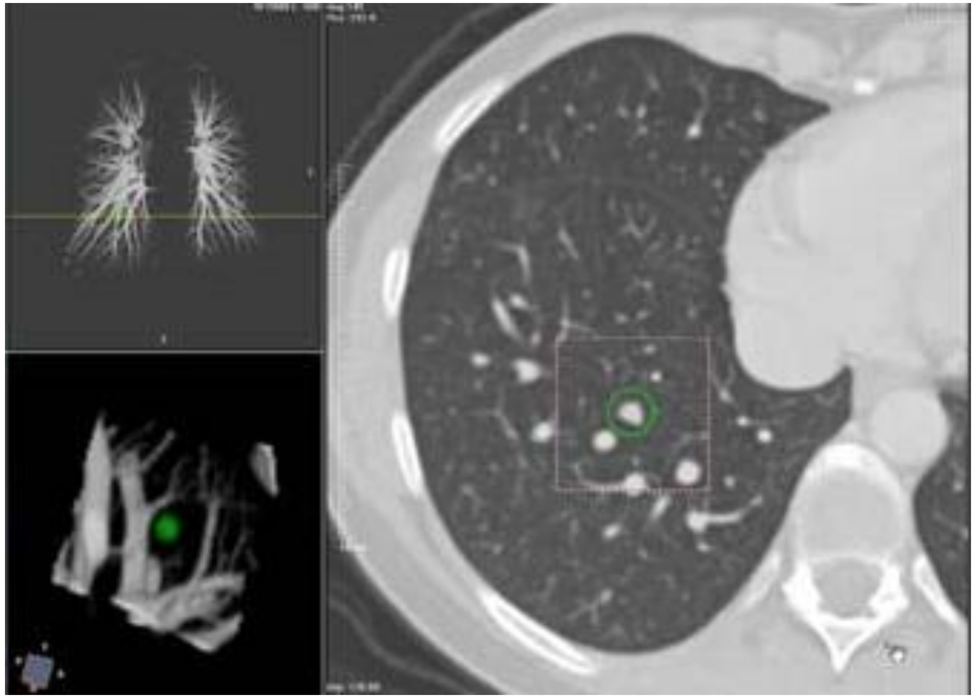

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
LESIONS LEARNT
Artificial intelligence powers detection of skin cancer from images PAGES 36 & 115


Chest Defense: Deep Learning Spots Disease Early Using Chest X-Rays

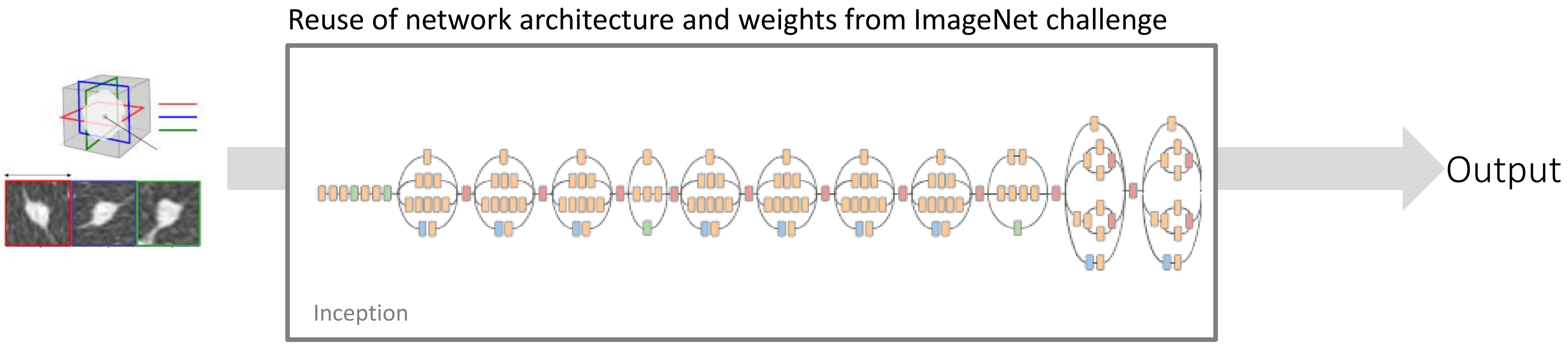

AUTOMATED ALGORITHM BASED ON DEEP MACHINE LEARNING FOR DETECTION OF DIABETIC RETINOPATHY
Advantages
Consistency of Interpretation
High Sensitivity & Specificity
Instantaneous Reporting of Results
© www.medindia.net

Unique challenges for medical images

Image characteristics are 3+ dimensional



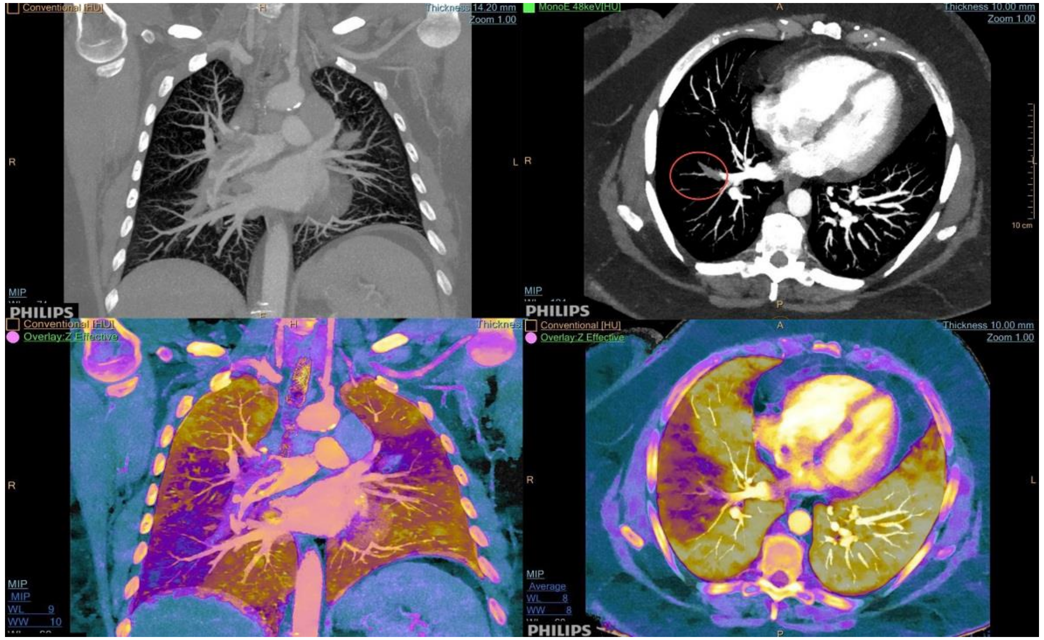
Commonly used transfer learning input that leverages the 3D structures



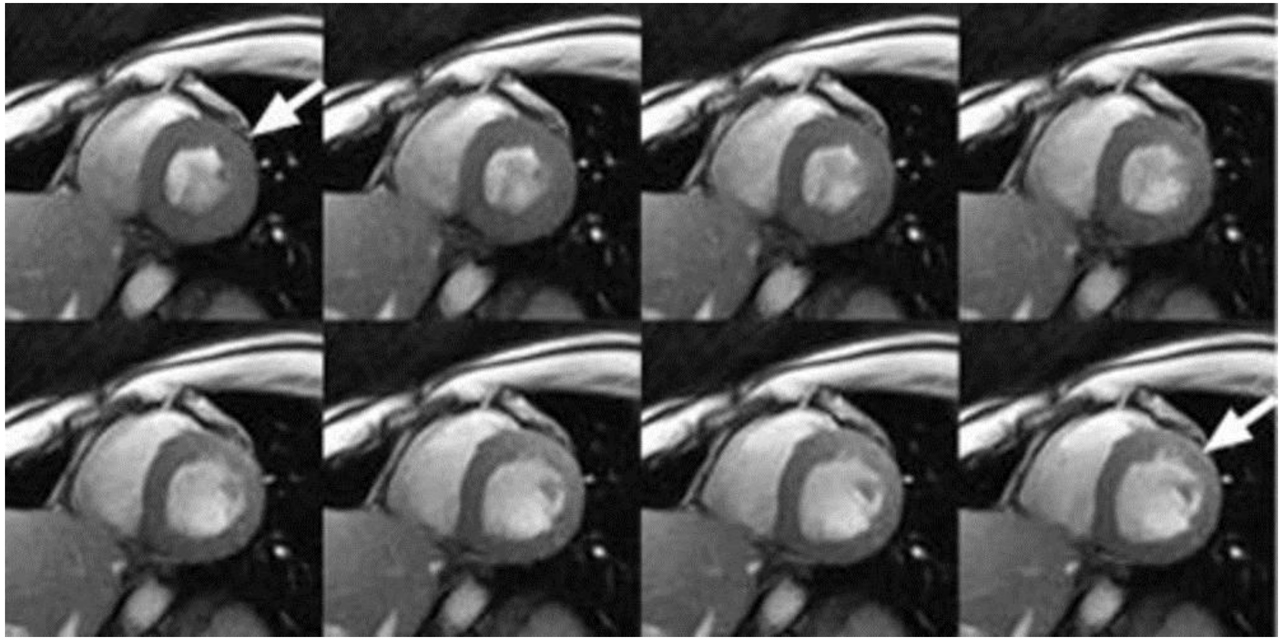
This is just one simple example. There are many approaches to take 3D structures into account. There are obvious limitations to this approach.

Image characteristics are 3+ dimensional

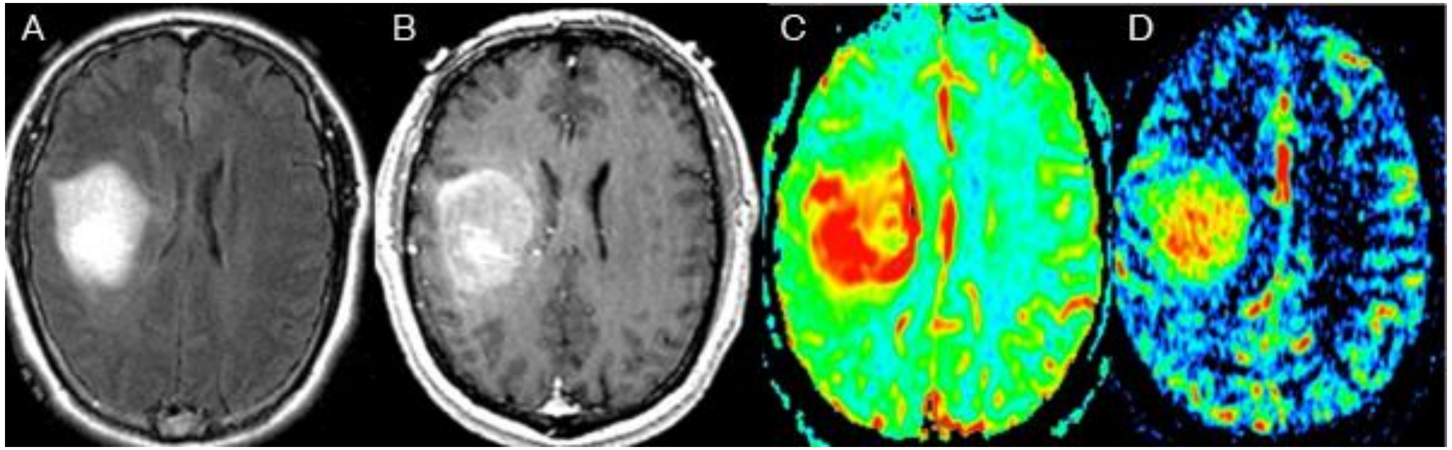
Volume and Chemistry



Volume and Time



Multimodal, Multiple Reconstructions, Registration Challenges



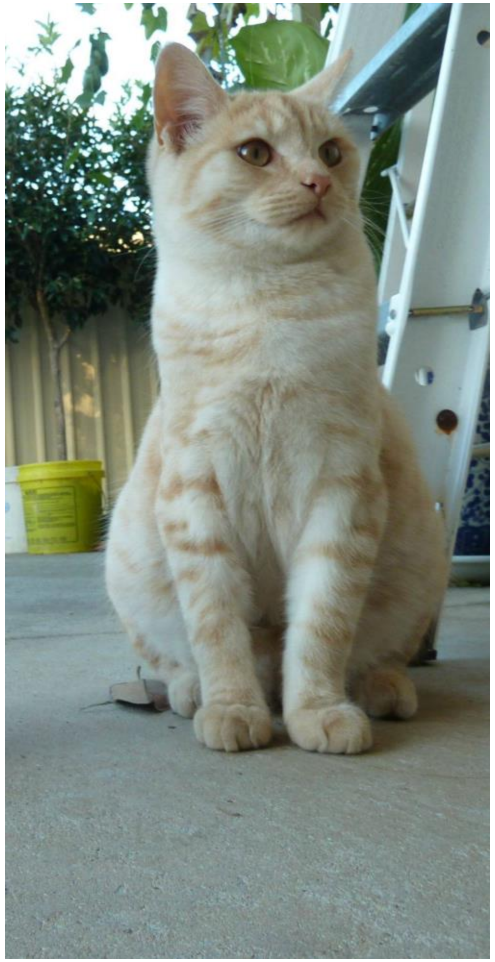
Sarah Nelson. UCSF's Neuroradiology Research Laboratory.

Soft Tissue - Axial | **Lung - Axial** | Multiplanar MIPs - Axial | Coronal | Sagittal

Scale Variance

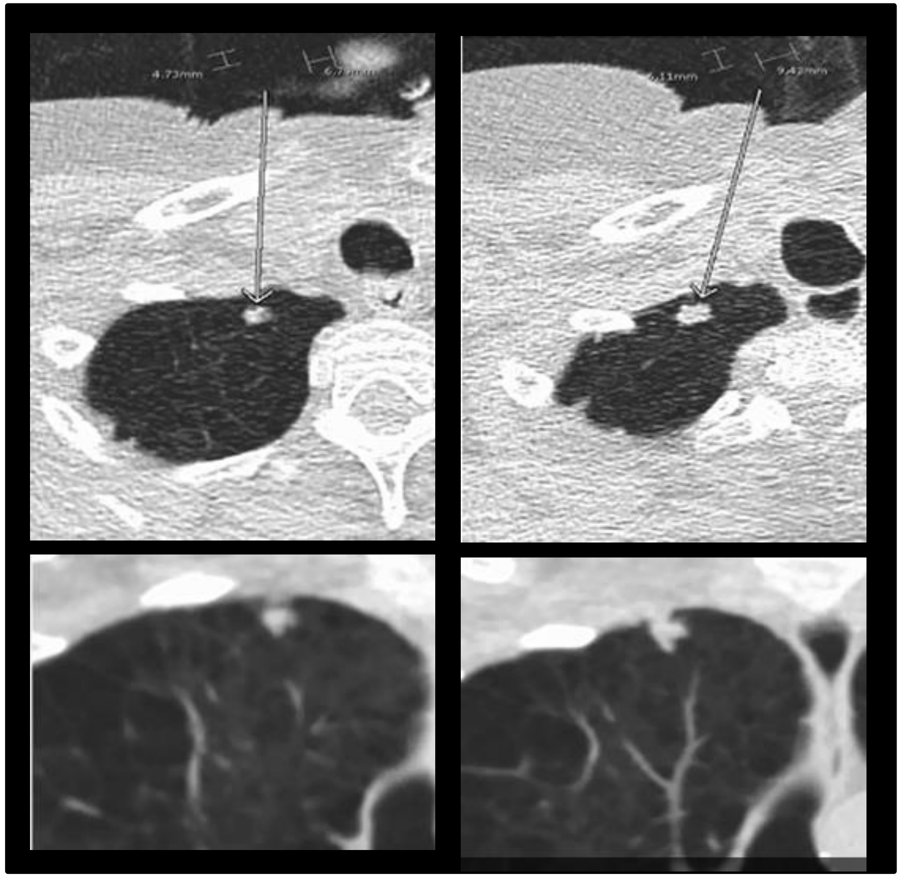
Cat

Also a Cat



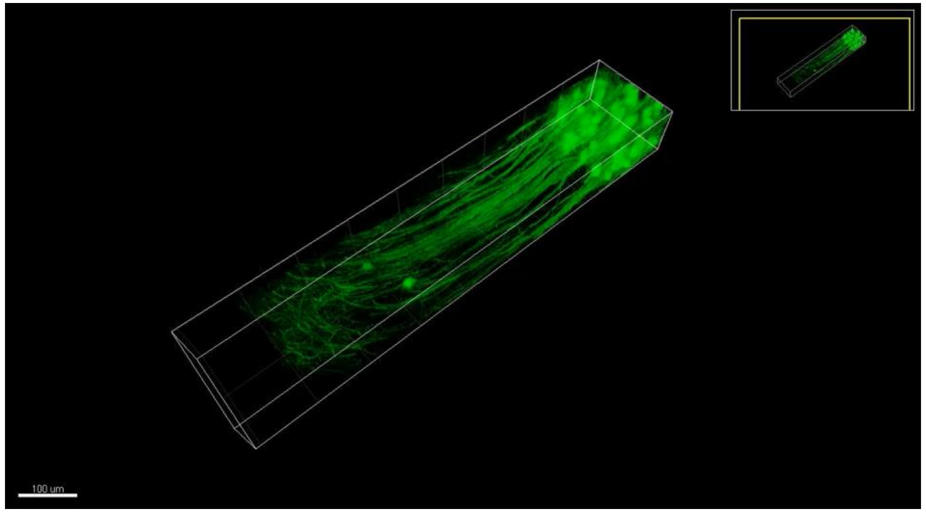
Negative Finding

Positive Finding
Follow-up Diagnostic Tests

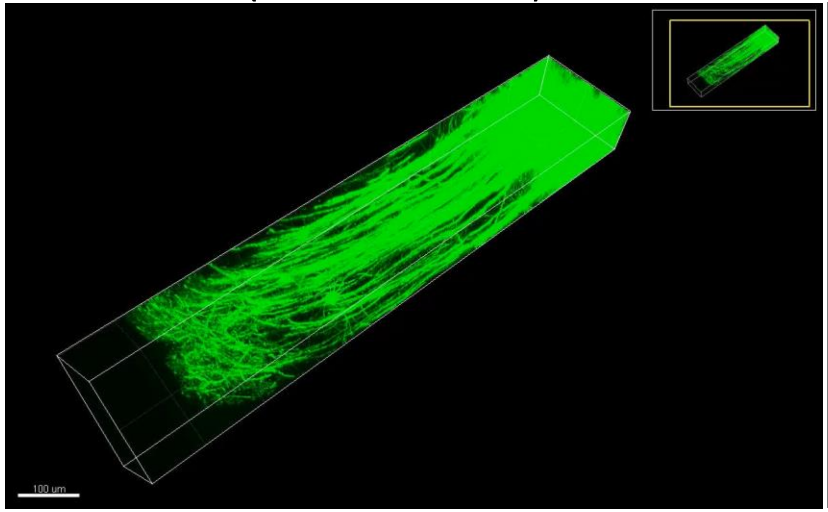


High Dynamic Range

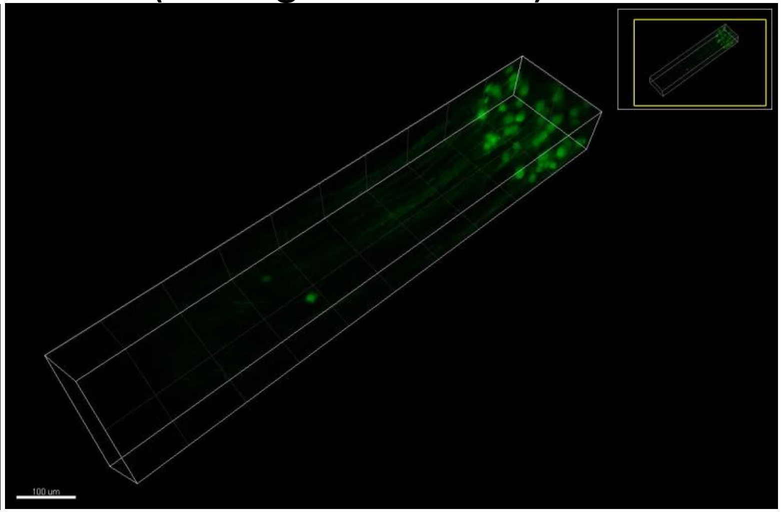
High Dynamic Range



Low Signal
(oversaturated)



High Signal
(low signal-to-noise)

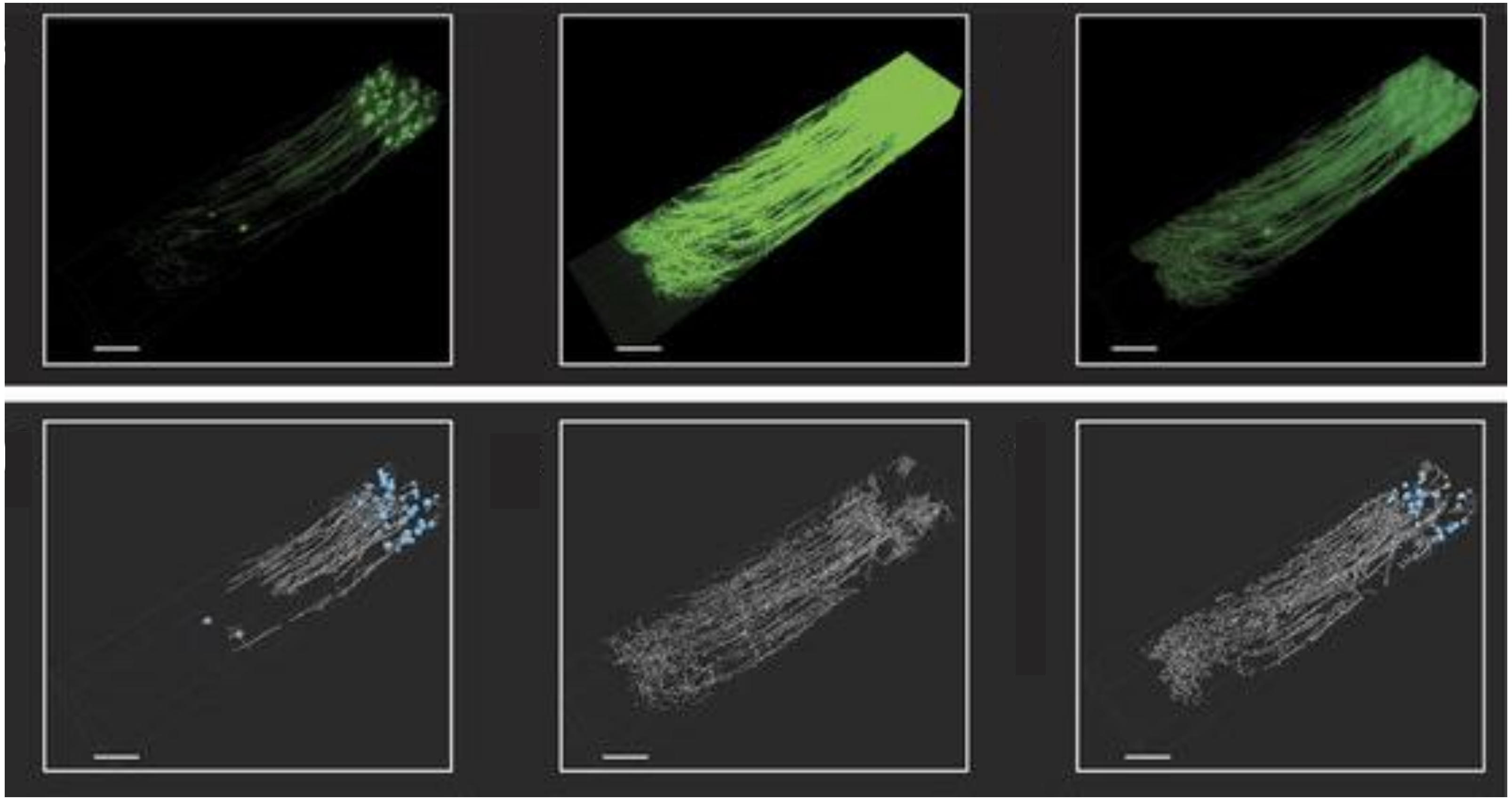


The scope of this paper is more about the value of HDR. Here, we are highlighting the insight that going from a HDR to LDR (e.g. 16-bit to 8-bit image) will destroy important image characteristics and reduce performance in computer vision tasks. This is particularly important in radiology and pathology, where images tend to have a higher dynamic range than natural images.
Swisher* & Vinegoni*. Nature Communications (2016); *Contributed equally.

PHILIPS

High Dynamic Range

High Signal | Low Signal | High Dynamic Range

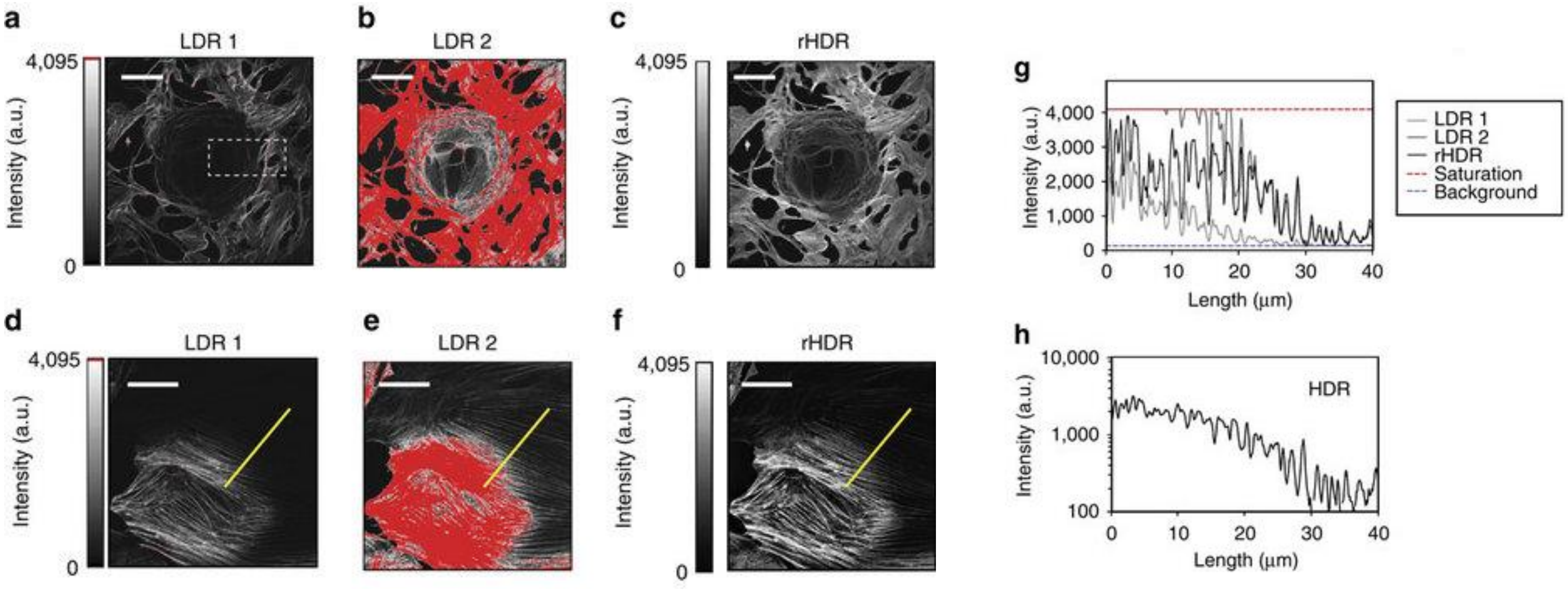


Images with high dynamic range do better in computer vision tasks

The scope of this paper is more about the value of HDR. Here, we are highlighting the insight that going from a HDR to LDR (e.g. 16-bit to 8-bit image) will destroy important image characteristics and reduce performance in computer vision tasks. This is particularly important in radiology and pathology, where images tend to have a higher dynamic range than natural images.
Swisher* & Vinegoni*. Nature Communications (2016); *Contributed equally.

PHILIPS

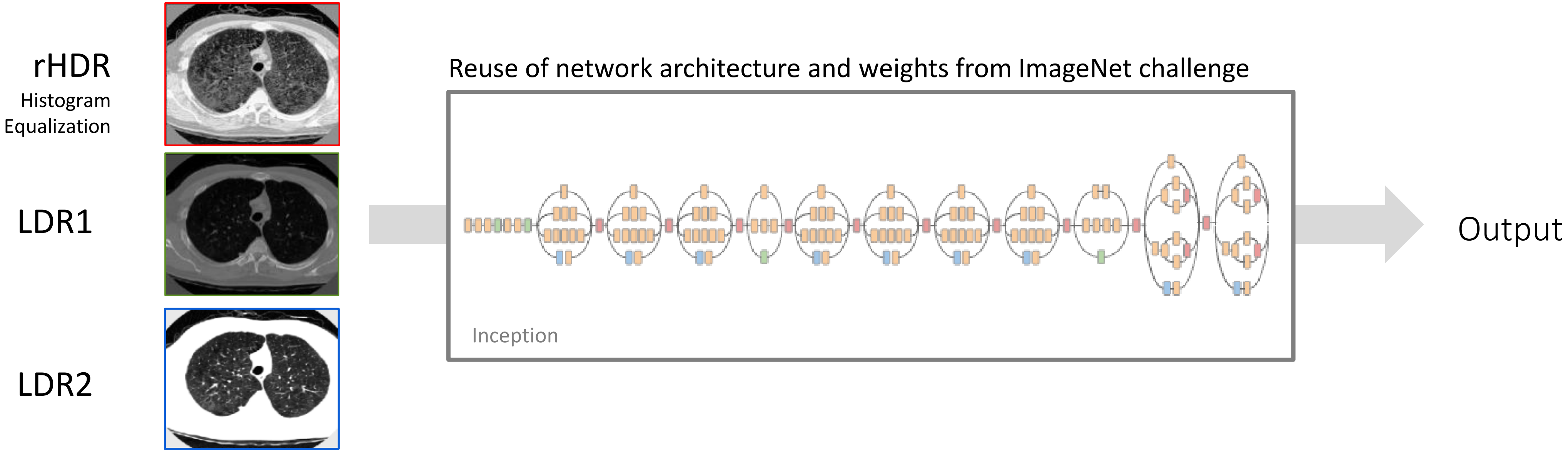
High Dynamic Range



The scope of this paper is more about the value of HDR. Here, we are highlighting the insight that going from a HDR to LDR (e.g. 16-bit to 8-bit image) will destroy important image characteristics and reduce performance in computer vision tasks. This is particularly important in radiology and pathology, where images tend to have a higher dynamic range than natural images.
Swisher* & Vinegoni*. Nature Communications (2016); *Contributed equally.

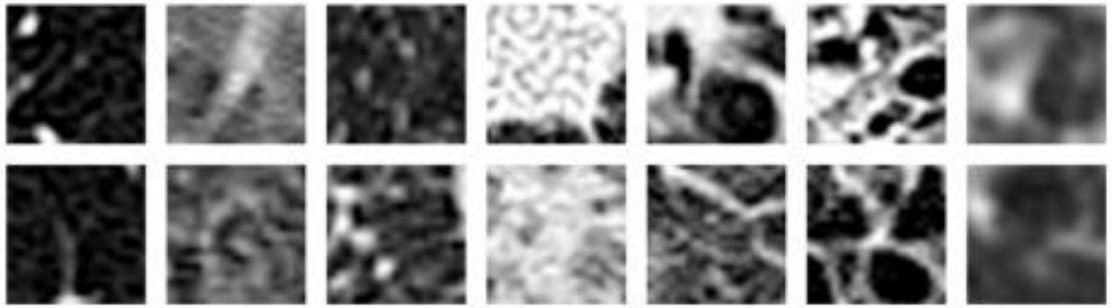
PHILIPS

Commonly used transfer learning input that leverages the full dynamic range

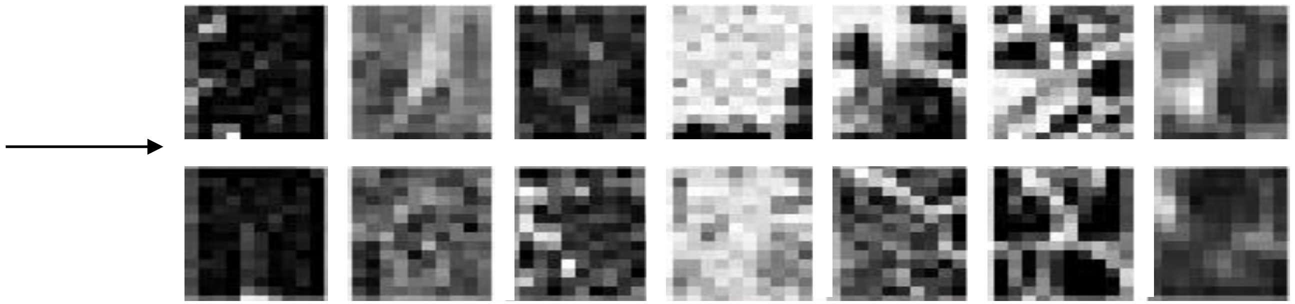


LDR = Low Dynamic Range; rHDR = reconstructed High Dynamic Range image at a Low dynamic range
This is just one simple example. There are many approaches to utilize HDR characteristics. There are obvious limitations to this approach.

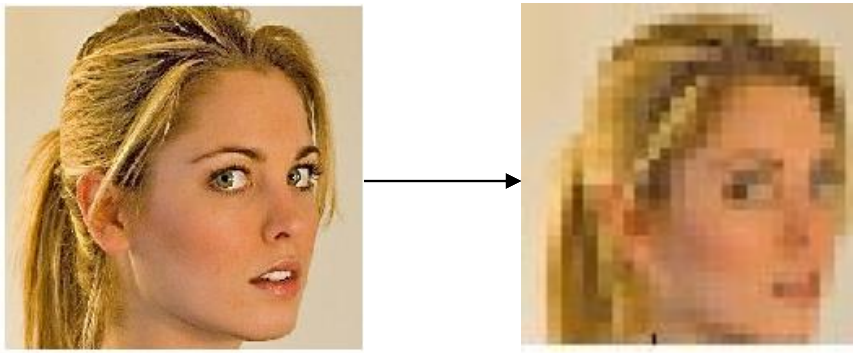
Downsampling - We must be creative in how we tackle dimensionality



Examples of healthy tissue and typical interstitial lung disease patterns (link to paper). Left to right: Healthy, ground glass opacity, micronodules, consolidation, reticulation, honeycombing, combination of ground glass and reticulation).



Clinical significant features look like noise.



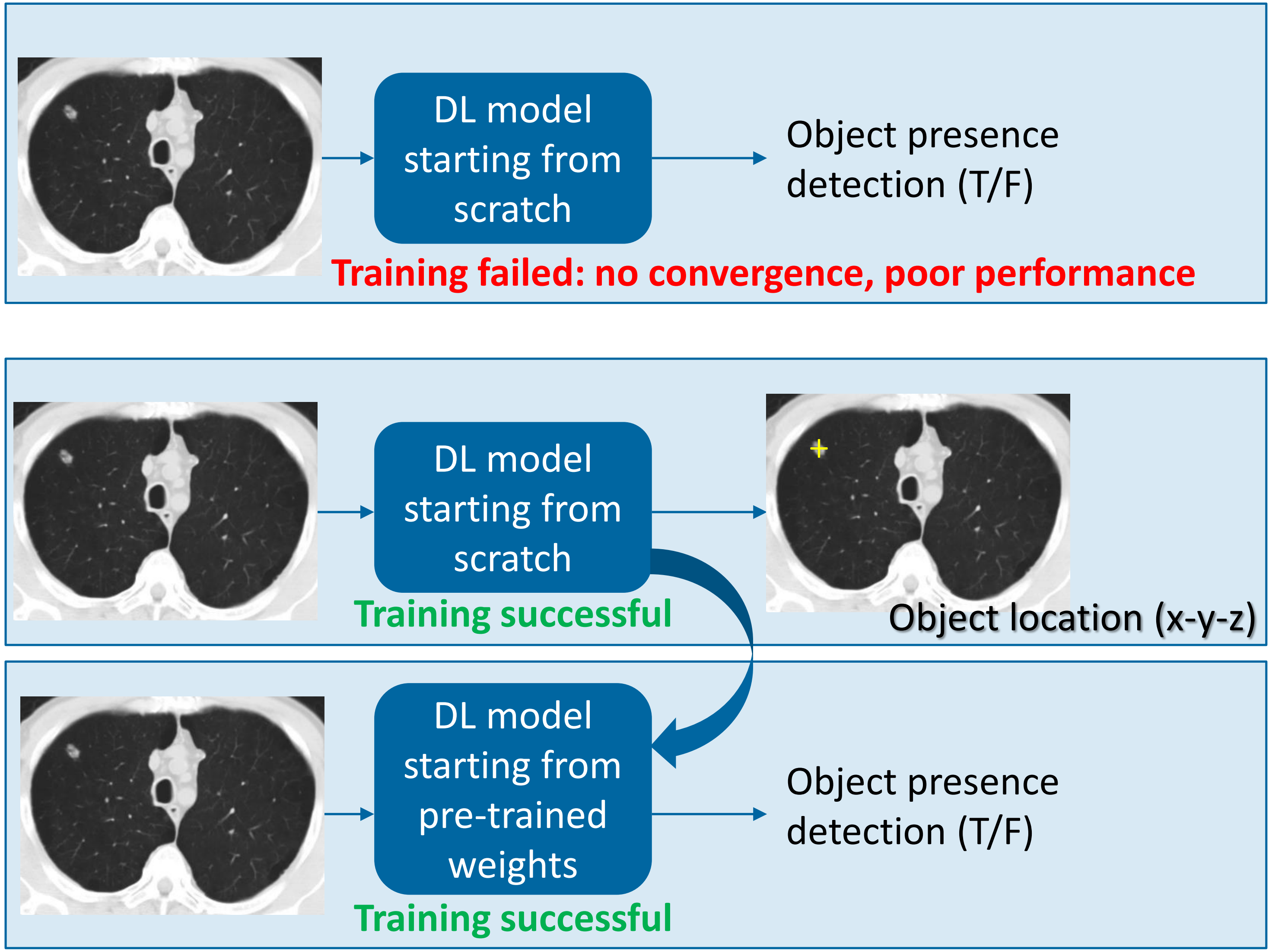
Still looks like a woman

Transfer Learning

	Very Similar Dataset	Very Different Dataset
Small dataset	Use Linear classifier on top layer	This is going to be challenging!
Large dataset	Fine-tune a few layers	Fine-tune a large number of layers

PHILIPS

Value of pre-training for DL tasks:



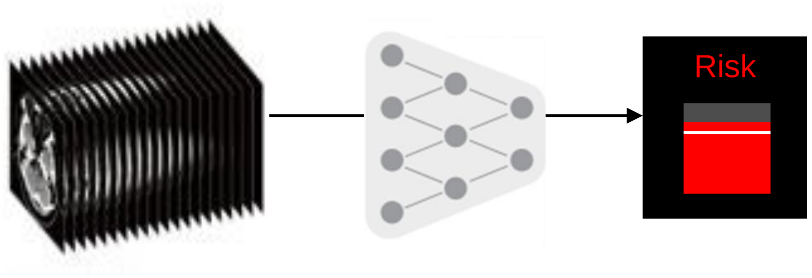
Aids in ambitious DL tasks:
Learning the 'easier' localization (regression) task served as 'stepping stone' for learning the detection task: the weights learned for localization were close enough to what was needed for detection to allow convergence.

Multitask Capability:
Network detects and localizes

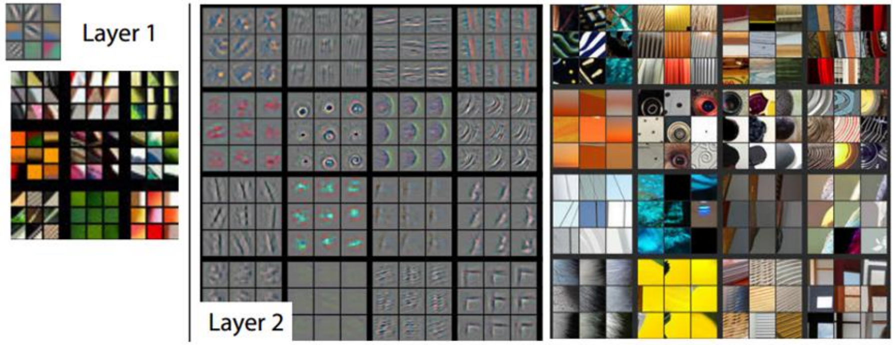
Transparency:
Easier to understand and justify the output of DNNs

Re-use:
Re-use successful DNNs for new tasks

“Deep Learning is a black box” – most physicians

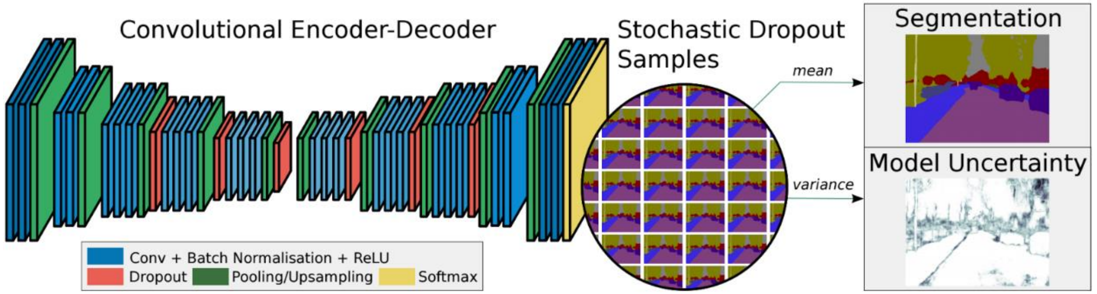


Feature understanding



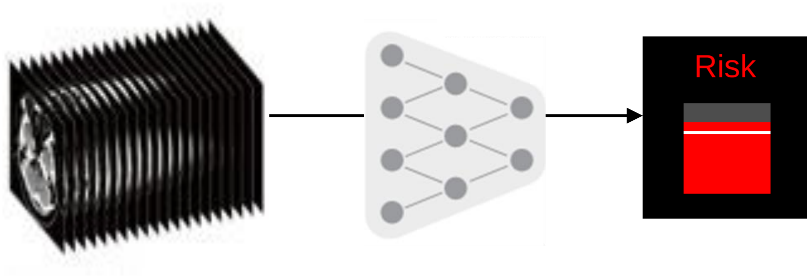
http://www.matthewzeiler.com/pubs/arxive2013/arxive2013.pdf

Uncertainty

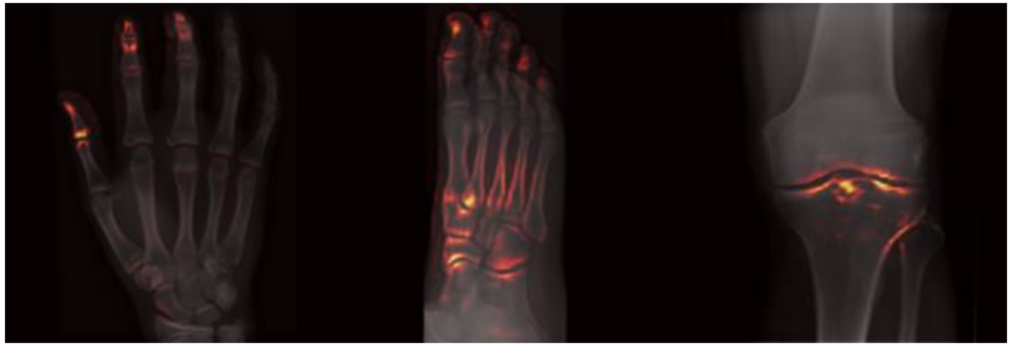


http://www.computervisionblog.com/2016/06/making-deep-networks-probabilistic-via.html

“Deep Learning is a black box” – most physicians

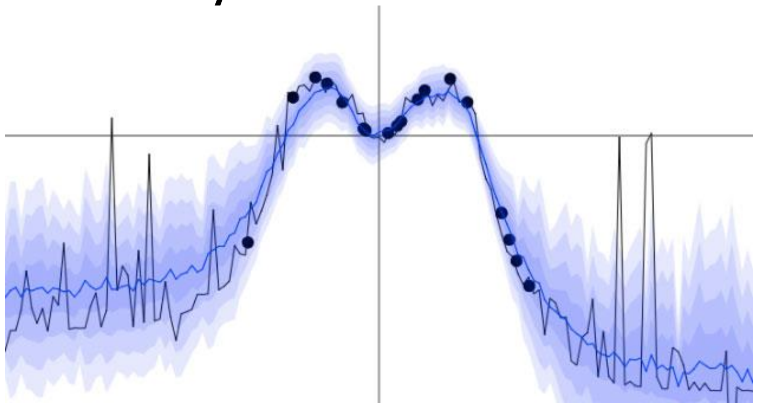


Feature understanding



Paper from Philips Research ([link]())

Uncertainty



Must read blog ([Link]())

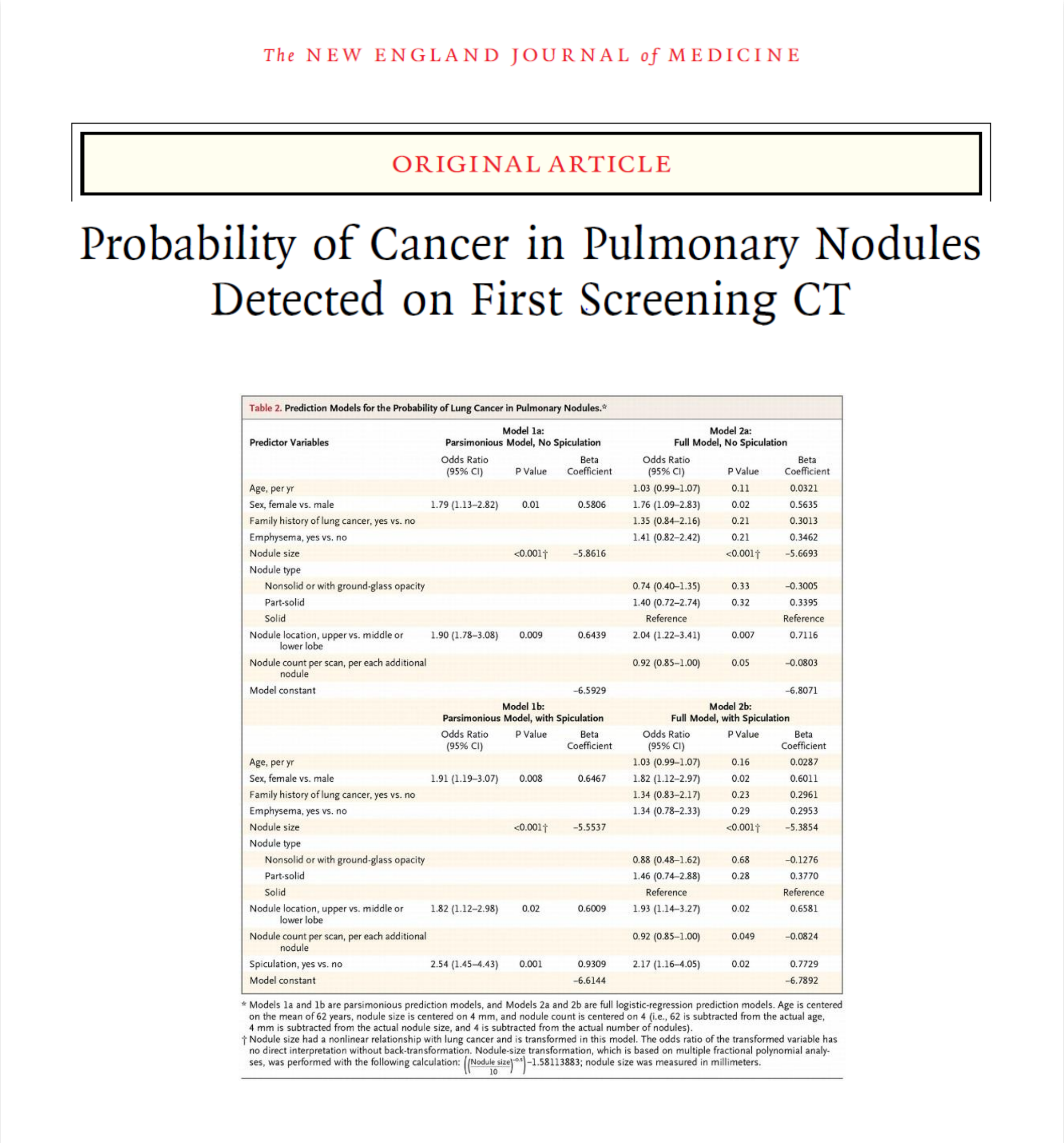
Three rules of meaningful ML innovation

1. Eyes on the Prize
   - How significant is the impact of a solution to the problem?
   - How many lives would it change? What is the severe unmet need we can overcome?
   - What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
   - Co-creation with clinicians
   - Feedback from hospital infrastructure and hospital administrator
   - Involve experts in business models, marketing & sales
   - ***Know your data!!!***

3. Meaningful Evaluation Methods
   - Generalization - multisite clinical trails, sustainability to changes in technology
   - Machine vs Human vs Machine + Human
   - Improvement of clinical outcome

PHILIPS



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

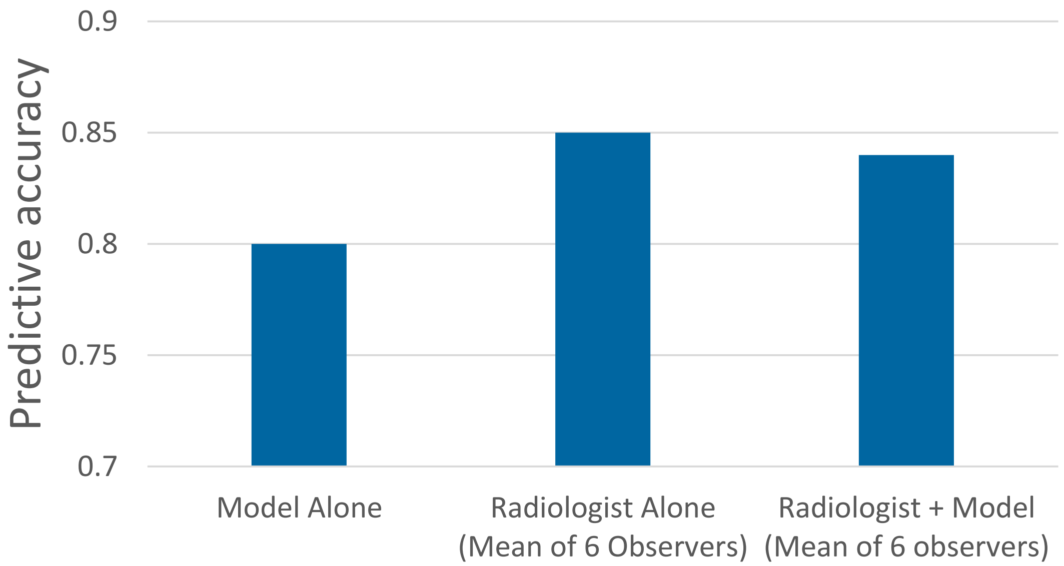
Probability of Cancer in Pulmonary Nodules Detected on First Screening CT

Table 2. Prediction Models for the Probability of Lung Cancer in Pulmonary Nodules.*

Predictor Variables	Model 1a: Parsimonious Model, No Spiculation			Model 2a: Full Model, No Spiculation		
	Odds Ratio (95% CI)	P Value	Beta Coefficient	Odds Ratio (95% CI)	P Value	Beta Coefficient
Age, per yr				1.03 (0.99–1.07)	0.11	0.0321
Sex, female vs. male	1.79 (1.13–2.82)	0.01	0.5806	1.76 (1.09–2.83)	0.02	0.5635
Family history of lung cancer, yes vs. no				1.35 (0.84–2.16)	0.21	0.3013
Emphysema, yes vs. no				1.41 (0.82–2.42)	0.21	0.3462
Nodule size		<0.001†	–5.8616		<0.001†	–5.6693
Nodule type						
Nonsolid or with ground-glass opacity				0.74 (0.40–1.35)	0.33	–0.3005
Part-solid				1.40 (0.72–2.74)	0.32	0.3395
Solid				Reference		Reference
Nodule location, upper vs. middle or lower lobe	1.90 (1.78–3.08)	0.009	0.6439	2.04 (1.22–3.41)	0.007	0.7116
Nodule count per scan, per each additional nodule				0.92 (0.85–1.00)	0.05	–0.0803
Model constant			–6.5929			–6.8071
	Model 1b: Parsimonious Model, with Spiculation			**Model 2b: Full Model, with Spiculation**		
	Odds Ratio (95% CI)	P Value	Beta Coefficient	Odds Ratio (95% CI)	P Value	Beta Coefficient
Age, per yr				1.03 (0.99–1.07)	0.16	0.0287
Sex, female vs. male	1.91 (1.19–3.07)	0.008	0.6467	1.82 (1.12–2.97)	0.02	0.6011
Family history of lung cancer, yes vs. no				1.34 (0.83–2.17)	0.23	0.2961
Emphysema, yes vs. no				1.34 (0.78–2.33)	0.29	0.2953
Nodule size		<0.001†	–5.5537		<0.001†	–5.3854
Nodule type						
Nonsolid or with ground-glass opacity				0.88 (0.48–1.62)	0.68	–0.1276
Part-solid				1.46 (0.74–2.88)	0.28	0.3770
Solid				Reference		Reference
Nodule location, upper vs. middle or lower lobe	1.82 (1.12–2.98)	0.02	0.6009	1.93 (1.14–3.27)	0.02	0.6581
Nodule count per scan, per each additional nodule				0.92 (0.85–1.00)	0.049	–0.0824
Spiculation, yes vs. no	2.54 (1.45–4.43)	0.001	0.9309	2.17 (1.16–4.05)	0.02	0.7729
Model constant			–6.6144			–6.7892

* Models 1a and 1b are parsimonious prediction models, and Models 2a and 2b are full logistic-regression prediction models. Age is centered on the mean of 62 years, nodule size is centered on 4 mm, and nodule count is centered on 4 (i.e., 62 is subtracted from the actual age, 4 mm is subtracted from the actual nodule size, and 4 is subtracted from the actual number of nodules).

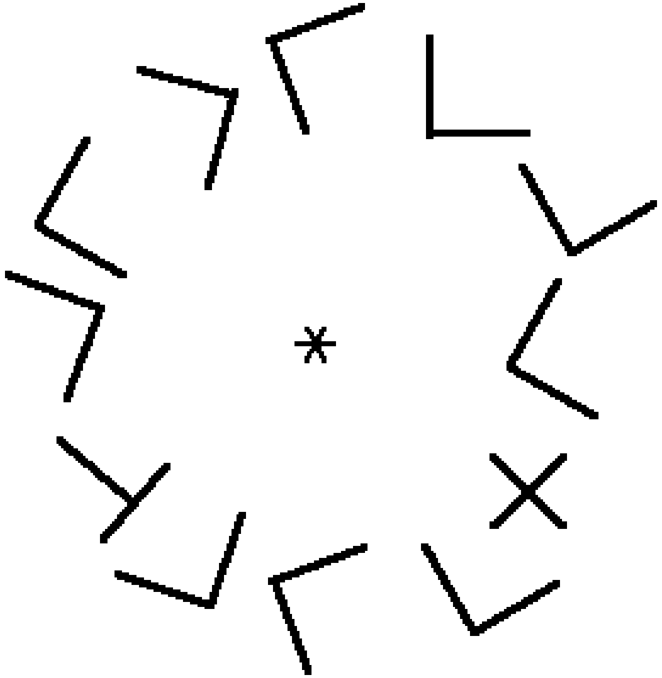
† Nodule size had a nonlinear relationship with lung cancer and is transformed in this model. The odds ratio of the transformed variable has no direct interpretation without back-transformation. Nodule-size transformation, which is based on multiple fractional polynomial analyses, was performed with the following calculation: $\left(\left(\frac{\text{Nodule size}}{10}\right)^{-0.5}\right)-1.58113883$; nodule size was measured in millimeters.



[Link to article]

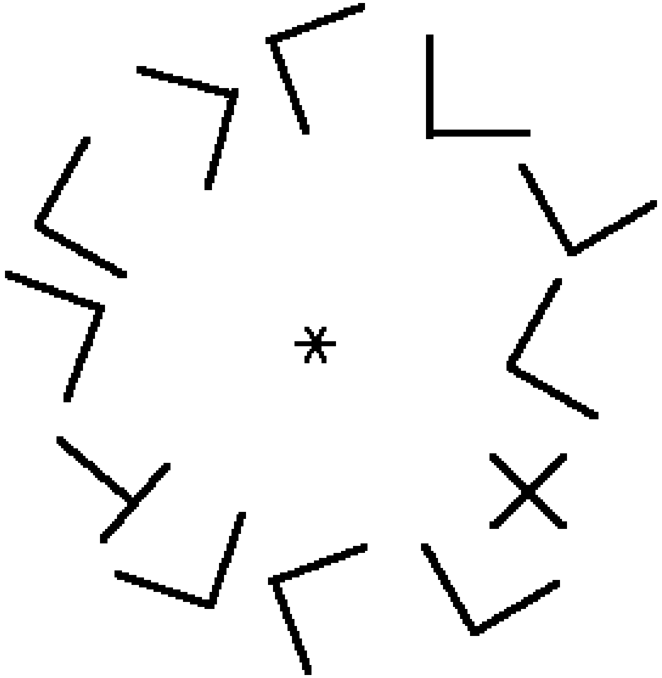
PHILIPS

Look at the star in the center



PHILIPS

There is an X in this image, can you find it?

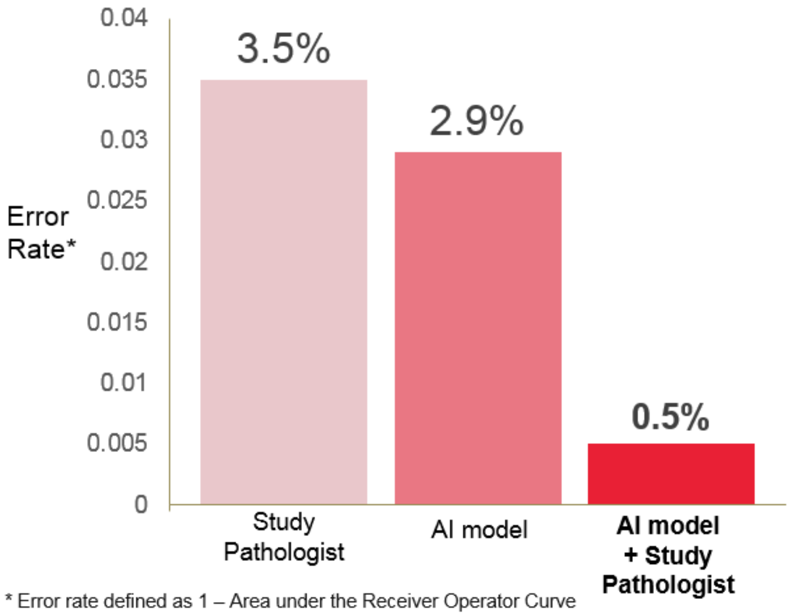


How many people noticed the T?

Think about where and when the algorithm will be used so that it will actually deliver improved clinical outcomes.

Example of meaningful evaluation metric

(AI + Pathologist) > Pathologist



* Error rate defined as 1 – Area under the Receiver Operator Curve
** A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.

Speaker for next week

Three rules of meaningful ML innovation

1. Eyes on the Prize
   - How significant is the impact of a solution to the problem?
   - How many lives would it change? What is a severe unmet need we can overcome?
   - What would constitute a meaningful improvement over the status quo?

2. Involvement of the World Outside
   - Co-creation with clinicians
   - Feedback from hospital infrastructure and hospital administrator
   - Involve experts in business models, marketing & sales
   - ***Know your data!!!***

3. Meaningful Evaluation Methods
   - Generalization - multisite clinical trails, sustainability to changes in technology
   - Machine vs Human vs Machine + Human
   - Improvement of clinical outcome

PHILIPS

Acknowledgements

Philips Research - Eindhoven
Dimitrios Mavroeidis
Stojan Trajanovski
Jack He
Ulf Grossekathofer
Erik Bresch
Binyam Gebre
Teun van Den Heuvel
Bas Veeling
Devinder Kumar
Vlado Menkovski

Philips HealthCare
Homer Pien

Philips Research - Hamburg
Tobias Klinder
Rafael Wiemker

Philips Research – North America
Sadid Hasan
Jonathan Rubin
Cristhian Potes
Yuan Ling
Joey Liu
Nikhil Galagali
Eric Carlson
Sophia Zhou
Amir Tahmasebi
Sandeep Dalal

Lahey Medical Center
Sebastian Flacke
Christoph Wald
Brady Mckee
Ali Ardestani

MGH
Anthony Samir
John Gilbertson

Lahey Hospital & Medical Center
NIH NATIONAL CANCER INSTITUTE
MASSACHUSETTS GENERAL HOSPITAL
HARVARD MEDICAL SCHOOL

PHILIPS