Interpretable AI in Medicine: Opening the Black Box for Patient Safety, Trustworthiness, and Improved AI

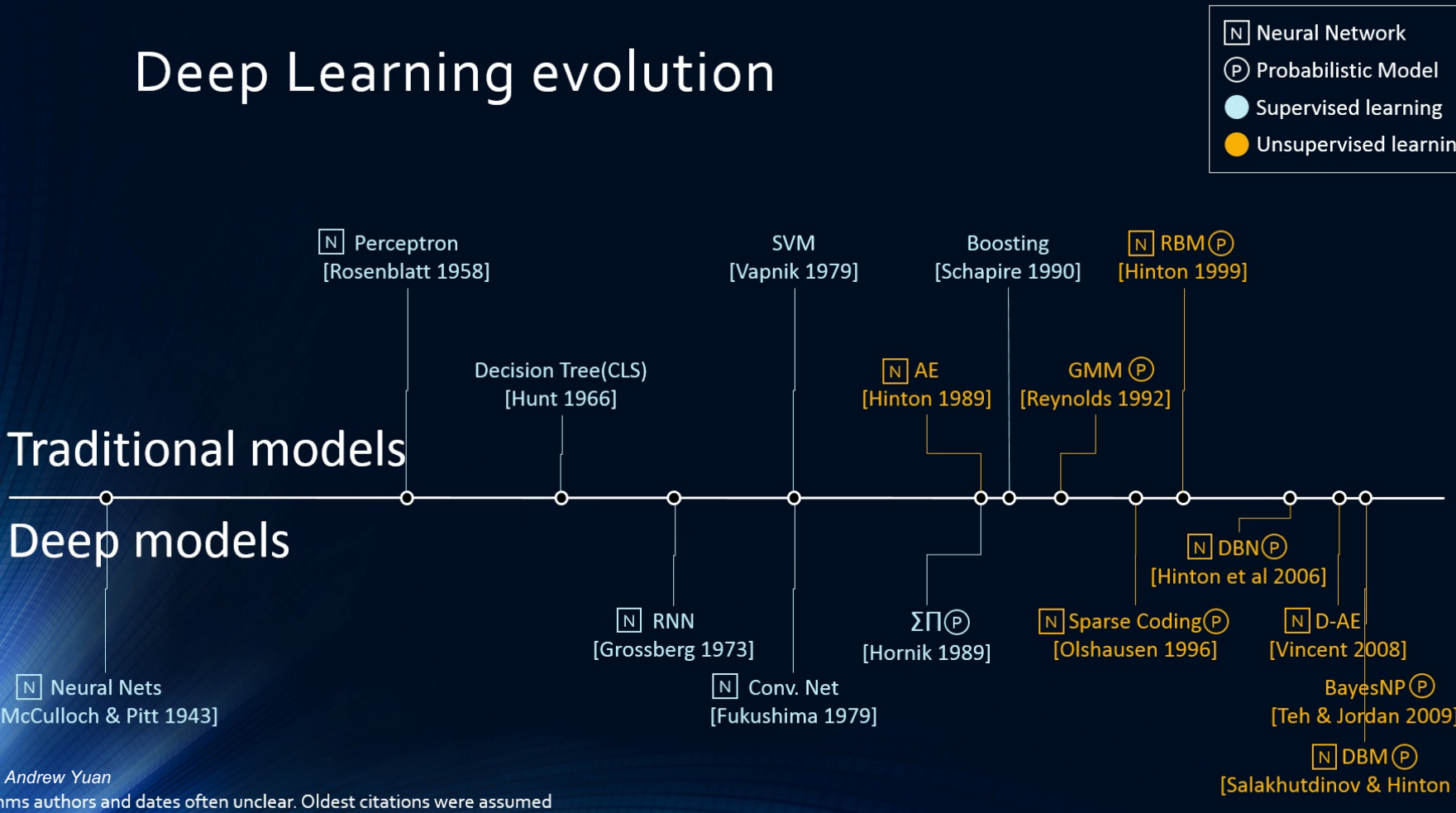
Prof. Mauricio Reyes, PhD
mauricio.reyes@unibe.ch




From this…

To this…

Deep Learning evolution



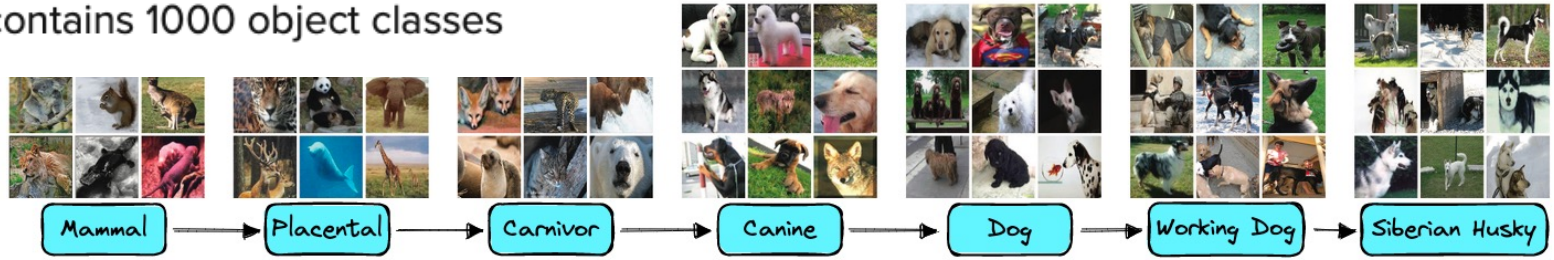
Source: Andrew Yuan

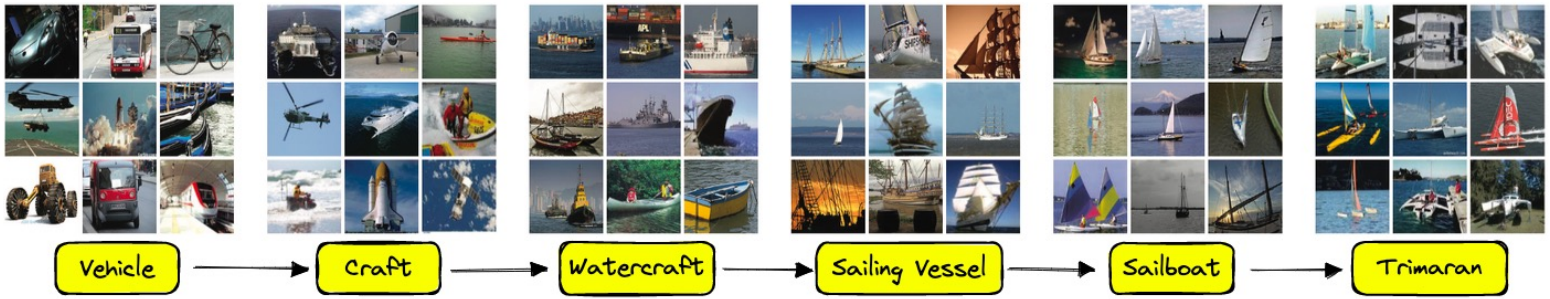
Algorithms authors and dates often unclear. Oldest citations were assumed

Classifications based on Yann LeCun's Deep Learning class at NYU – spring 2014

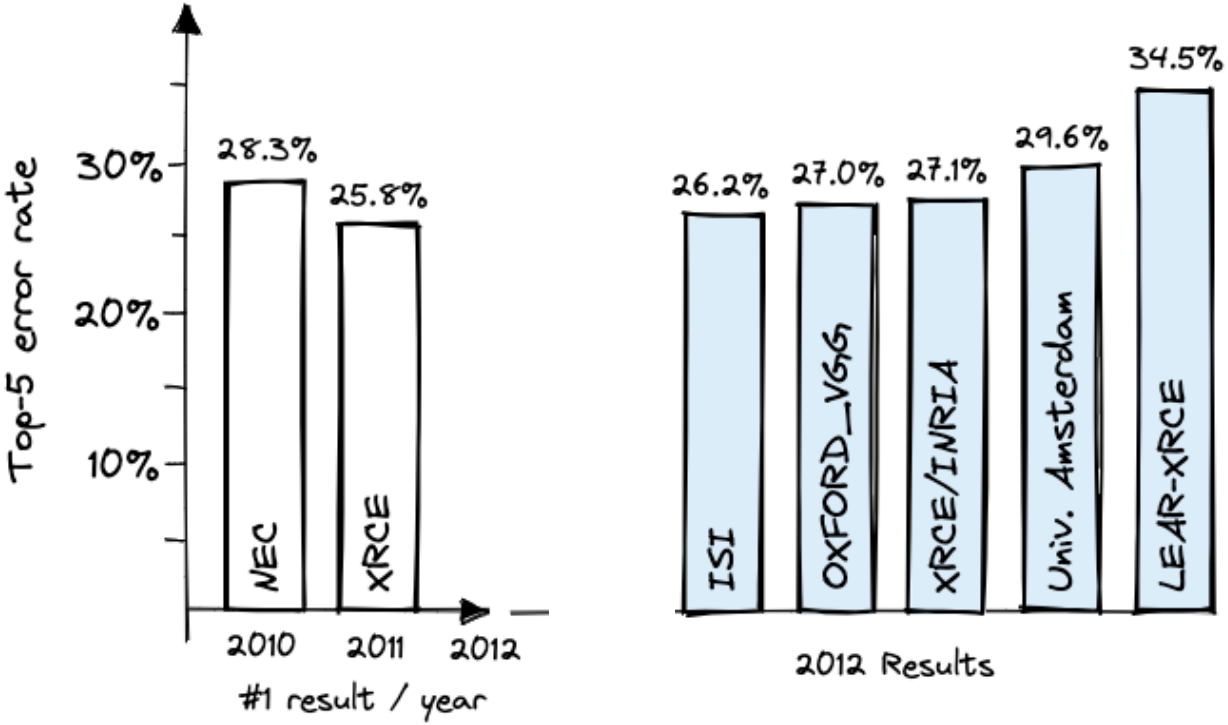
Deep Learning – The ImageNet Benchmark

- ImageNet contains 1,281,167 training images
- ImageNet contains 50,000 validation images
- ImageNet contains 100,000 test images
- ImageNet contains 1000 object classes



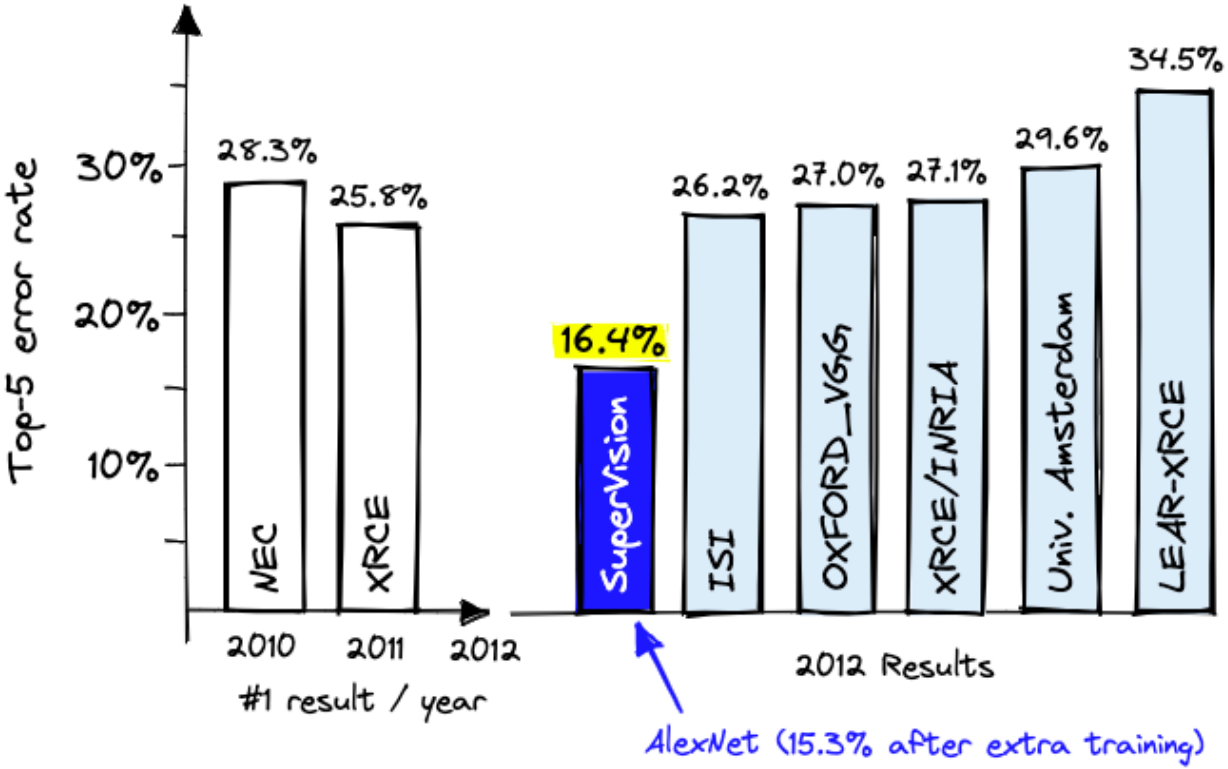


Deep Learning – The ImageNet Benchmark



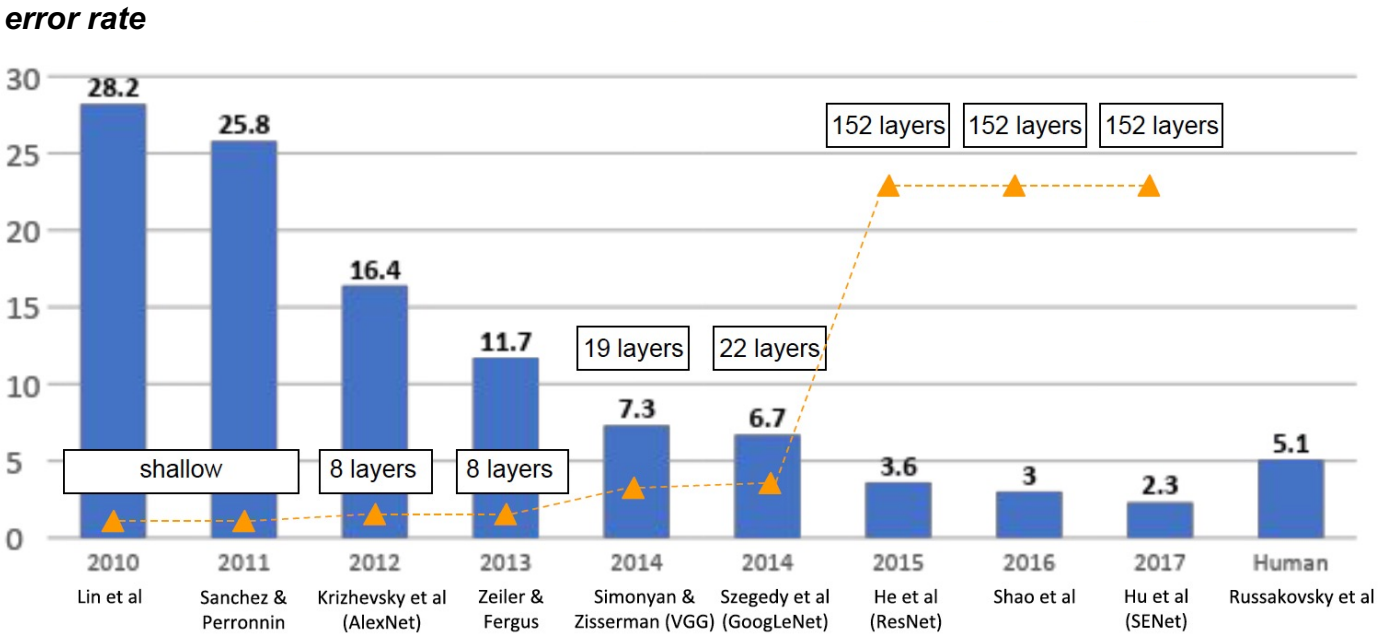
Source: Pinecone

Deep Learning – The ImageNet Benchmark



Source: Pinecone

Deep Learning – The ImageNet Benchmark



Source: ITT's tech-blog

Prof. Geoff Hinton – Nov. 2016

"Artificial intelligence will not replace radiologists. Yet, those radiologists who use AI will replace the ones who don't."

Curtis Langlotz, Professor of Radiology and Biomedical Informatics at Stanford University, GPU Tech Conference in San Jose, May 2017

Curtis P. Langlotz, MD, PhD, is the RSNA president. Dr. Langlotz is professor of radiology, medicine and biomedical data science, director of the Center for Artificial Intelligence in Medicine and Imaging, and associate chair for information systems in the Department of Radiology at Stanford University in California.

Brain Tumors

Glioblastoma

- Most common, complex, and treatment-resistant primary brain tumor[1]
- Currently no effective curative treatment
- Median survival of ~16 months

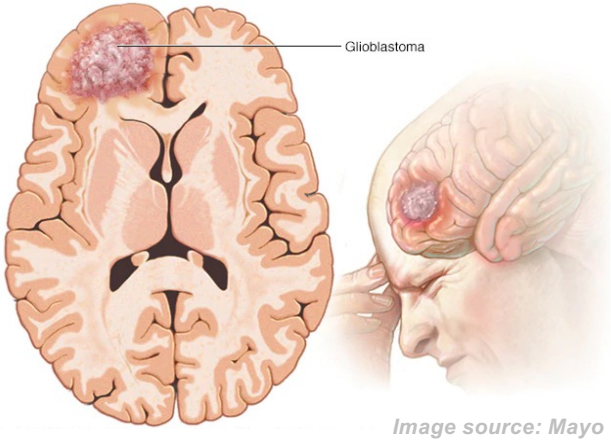


Image source: Mayo Clinic

Clinical workflow

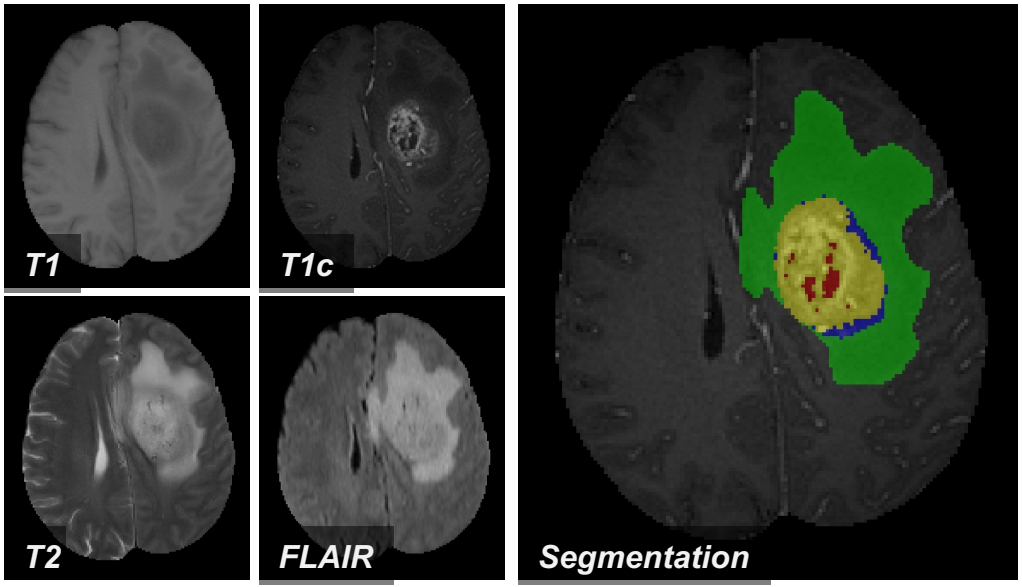
Diagnosis → ***Initial treatment*** → ***Follow-up cycle*** (↻)

- ***Diagnosis***
  - > *Radiological assessment*
- ***Initial treatment***
  - > *Surgical resection*
  - > *Radio- and/or chemotherapy*
- ***Follow-up cycle***
  - > *Response assessment*

MR imaging used throughout the process

[1]Ellingson et al., Neuro-oncology, 2015. | [2]Suchorska et al., Neuro-oncology, 2016. | Icons from the Noun Project

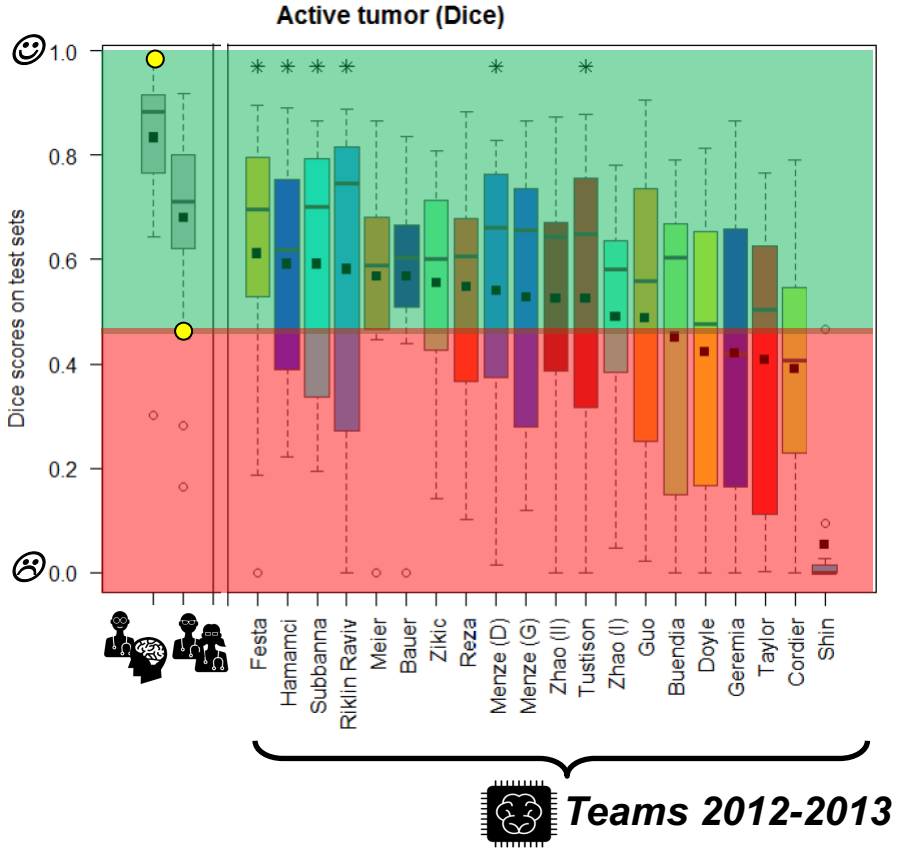
Brain Tumor Segmentation

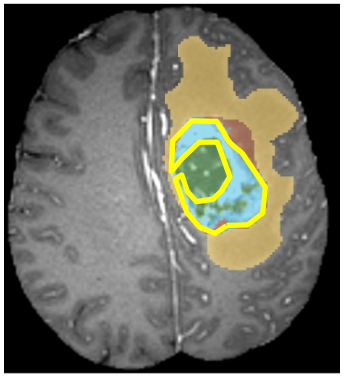


4 sub-compartments

- *Necrotic tissue*
- *Enhancing tumor*
- *Non-enhancing tumor*
- *Edema*

Brain Tumor Segmentation Challenge (BRATS) *Happy 10-year Anniversary BRATS!!*



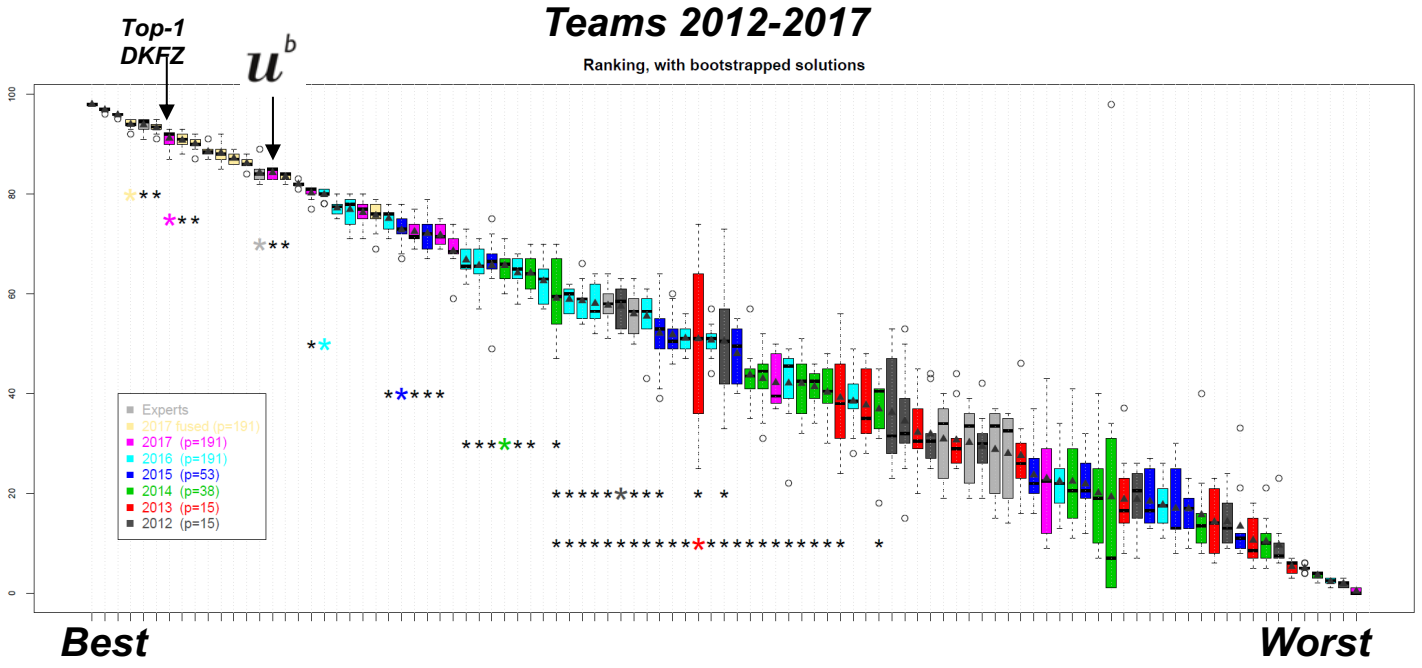


Active tumor

Menze et al. TMI 2015

Brain tumor Segmentation Challenge -

- *Top entries 2012, 2013, 2017, 2018*
- *MICCAI Young Scientist Impact Award 2016*
- *Ypsomed Innovation Award 2016*



Source: Bjoern Menze - BRATS 2018

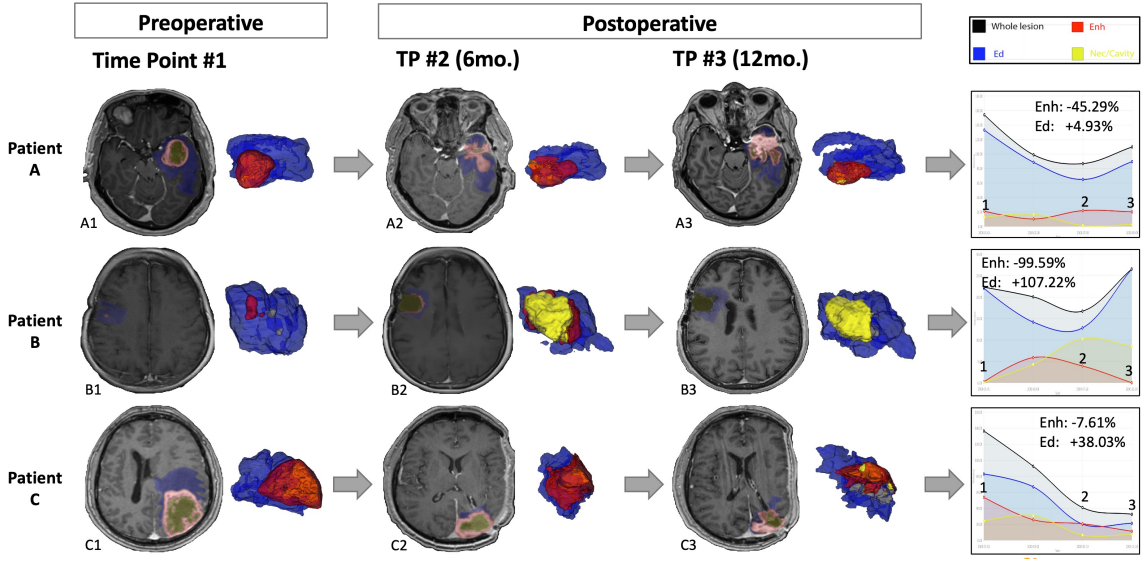
Translated AI as FDA-approved - Collaboration with Neosoma Inc.

Neuro-Oncology Advances
5(1), 1–10, 2022 | https://doi.org/10.1093/noajnl/vdac184 | Advance Access date 20 December 2022

NS-HGlio: A generalizable and repeatable HGG segmentation and volumetric measurement AI algorithm for the longitudinal MRI assessment to inform RANO in trials and clinics

Aly H. Abayazeed, Ahmed Abbassy, Michael Müeller, Michael Hill, Mohamed Qayati, Shady Mohamed, Mahmoud Mekhaimar, Catalina Raymond, Prachi Dubey, Kambiz Nael, Saurabh Rohatgi, Vaishali Kapare, Ashwini Kulkarni, Tina Shiang, Atul Kumar, Nicolaus Andratschke, Jonas Willmann, Alexander Brawanski, Reordan De Jesus, Ibrahim Tuna, Steve H. Fung, Joseph C. Landolfi, Benjamin M. Ellingson, and Mauricio Reyes

- ***The algorithm was trained on a large, heterogeneous dataset of more than 3,000 subjects using preoperative and postoperative MRIs***
- ***The data set underwent an extensive ground truthing process - by multiple, highly experienced neuroradiologists (double over-read design)***
- ***The technology was validated internally and externally and was tested under an FDA approved performance testing protocol***



Aly H Abayazeed, Ahmed Abbassy, Michael Müeller, Michael Hill, Mohamed Qayati, Shady Mohamed, Mahmoud Mekhaimar, Catalina Raymond, Prachi Dubey, Kambiz Nael, Saurabh Rohatgi, Vaishali Kapare, Ashwini Kulkarni, Tina Shiang, Atul Kumar, Nicolaus Andratschke, Jonas Willmann, Alexander Brawanski, Reordan De Jesus, Ibrahim Tuna, Steve H Fung, Joseph C Landolfi, Benjamin M Ellingson, Mauricio Reyes, NS-HGlio: A generalizable and repeatable HGG segmentation and volumetric measurement AI algorithm for the longitudinal MRI assessment to inform RANO in trials and clinics, *Neuro-Oncology Advances*, Volume 5, Issue 1, January-December 2023, vdac184, https://doi.org/10.1093/noajnl/vdac184

INTERPRETABLE AI

Why do we need interpretability/explainability?

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman, Seth Flaxman

(Submitted on 28 Jun 2016 (v1), last revised 31 Aug 2016 (this version, v3))

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which "significantly affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation.

Comments: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY

Ethics of AI in Radiology: European and North American Multi-society Statement

Transparency, interpretability, and explainability

Transparency, interpretability, and explainability are necessary to build patient and provider trust. When a radiologist makes a mistake, we want to know why, in part because we want to know whether the mistake is excusable. We want to know whether the mistake reflects malintent or negligence, or occurred due to other factors.

Similarly, if an algorithm fails or contributes to an adverse clinical event or malpractice, radiologists need to be able to understand why it produced the result that it did, and how it reached a decision.

There is a need for further research on the interrelated areas of medical AI to address the current clinical, socio-ethical and technical limitations. Examples of areas for future research include explainability and interpretability, bias estimation and mitigation, and secure and privacy-preserving AI.

European Parliament

Artificial intelligence in healthcare

Applications, risks, and ethical and societal impacts

Shortcut Learning in Deep Neural Networks

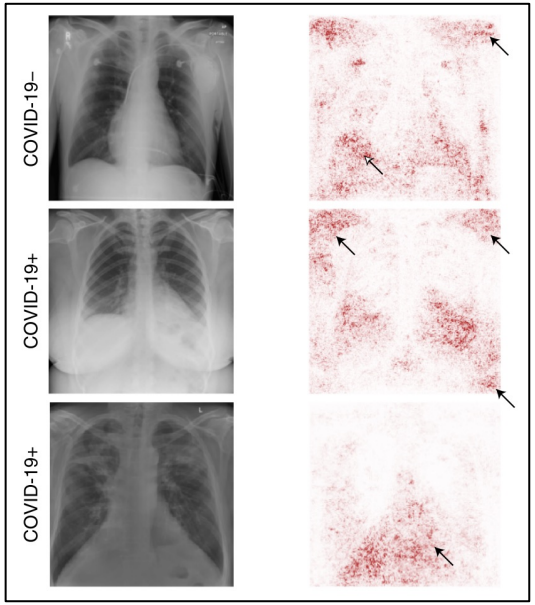
Robert Geirhos[1,2,*,§], Jörn-Henrik Jacobsen[3,*], Claudio Michaelis[1,2,*],
Richard Zemel[†,3], Wieland Brendel[†,1], Matthias Bethge[†,1] & Felix A. Wichmann[†,1]

nature machine intelligence
ARTICLES
https://doi.org/10.1038/s42256-021-00338-7
Check for updates

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave[1,2,3], Joseph D. Janizek[1,2,3] and Su-In Lee[1]✉

- ***Principle of "least effort"***
- ***Inductive bias:***
  - ***Model architecture***
  - ***Loss***
  - ***Optimization***
  - ***Training data***

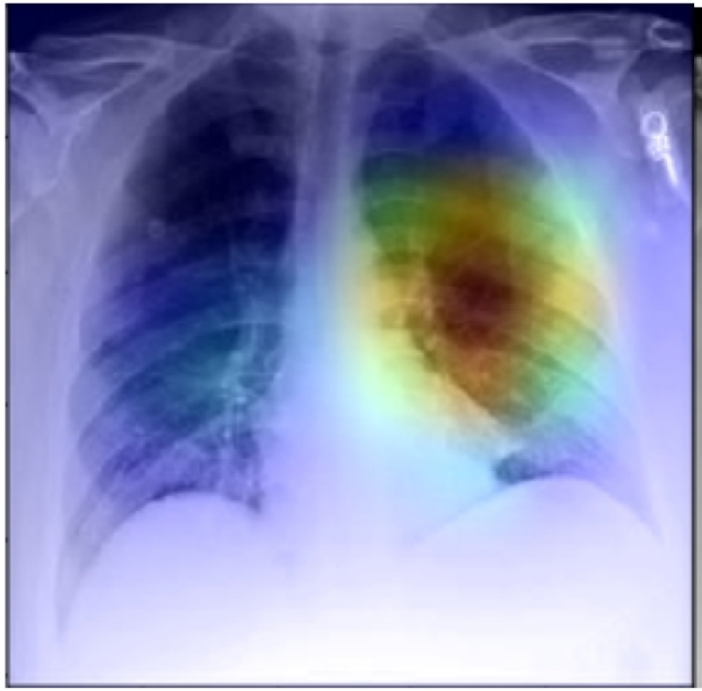


Laterality markers

Attention out of the region of interest

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira G



MXR Dense		(0.930-
MXR Dense		(0.823-

imagi

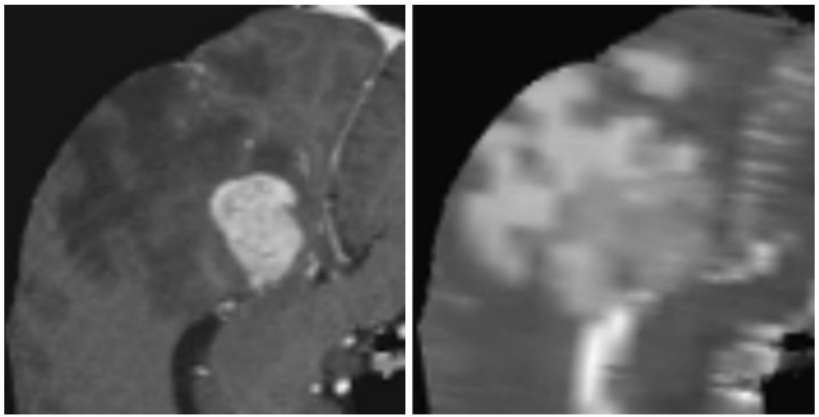
- ***Detec*** ***ntes***
- ***Patter*** ***persists across all anatomical regions***

Interpretability in AI Medical Imaging

Automated classification of Low and High-grade gliomas (LGG vs. HGG)

Q: Are there biases stemming from the data preparation process?

A: Bias of learned patterns detected via interpretability



T1c ***T2***

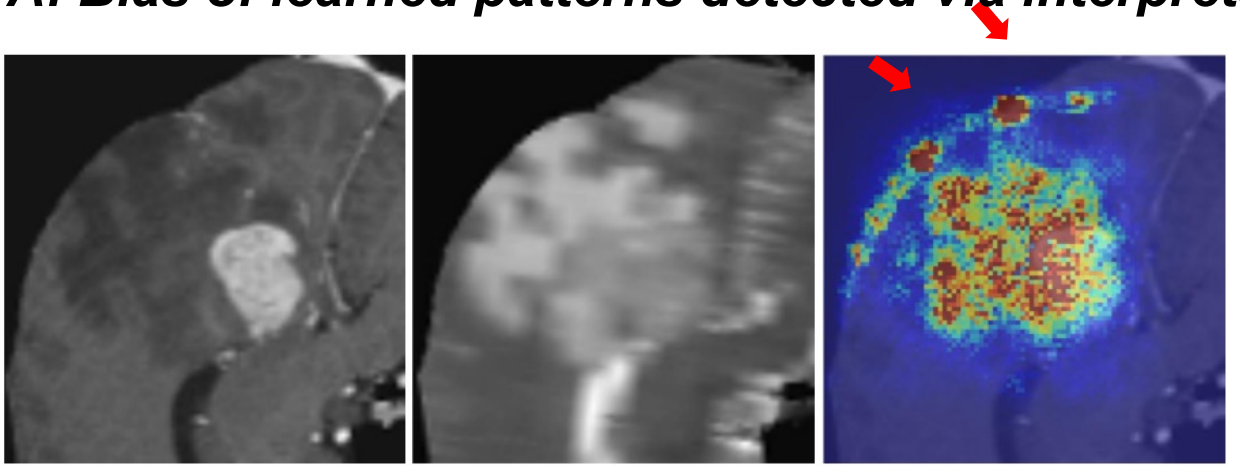
Pereira et al. MICCAI-iMIMIC 2018
Reyes et al. Radiology: Artificial Intelligence 2021

Interpretability in AI Medical Imaging

Automated classification of Low and High-grade gliomas (LGG vs. HGG)

Q: Are there biases stemming from the data preparation process?

A: Bias of learned patterns detected via interpretability



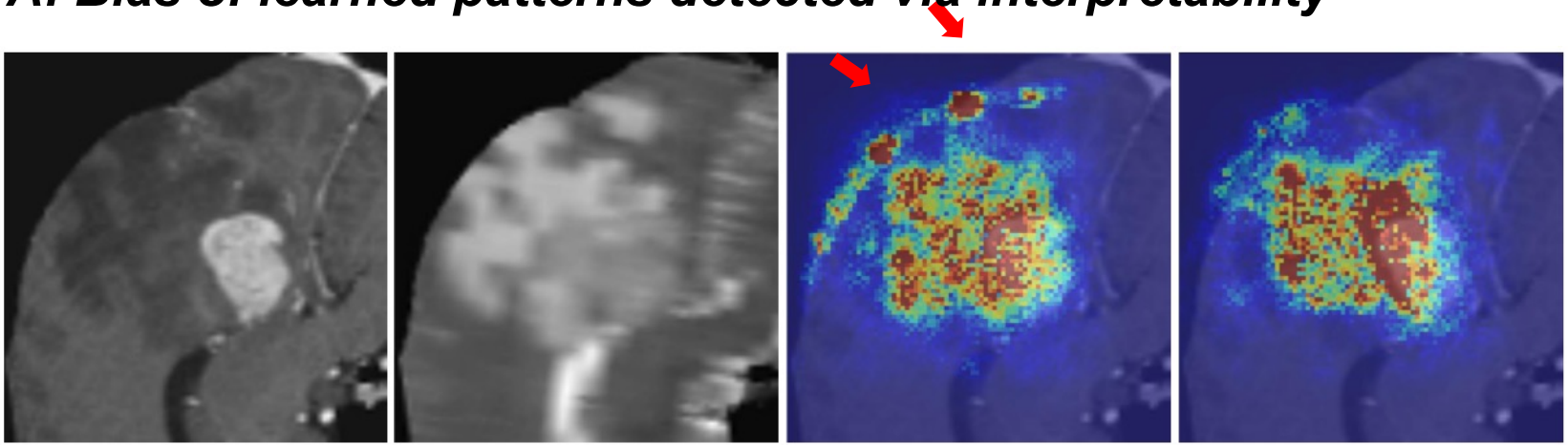
Pereira et al. MICCAI-iMIMIC 2018
Reyes et al. Radiology: Artificial Intelligence 2021

Interpretability in AI Medical Imaging

Automated classification of Low and High-grade gliomas (LGG vs. HGG)

Q: Are there biases stemming from the data preparation process?

A: Bias of learned patterns detected via interpretability

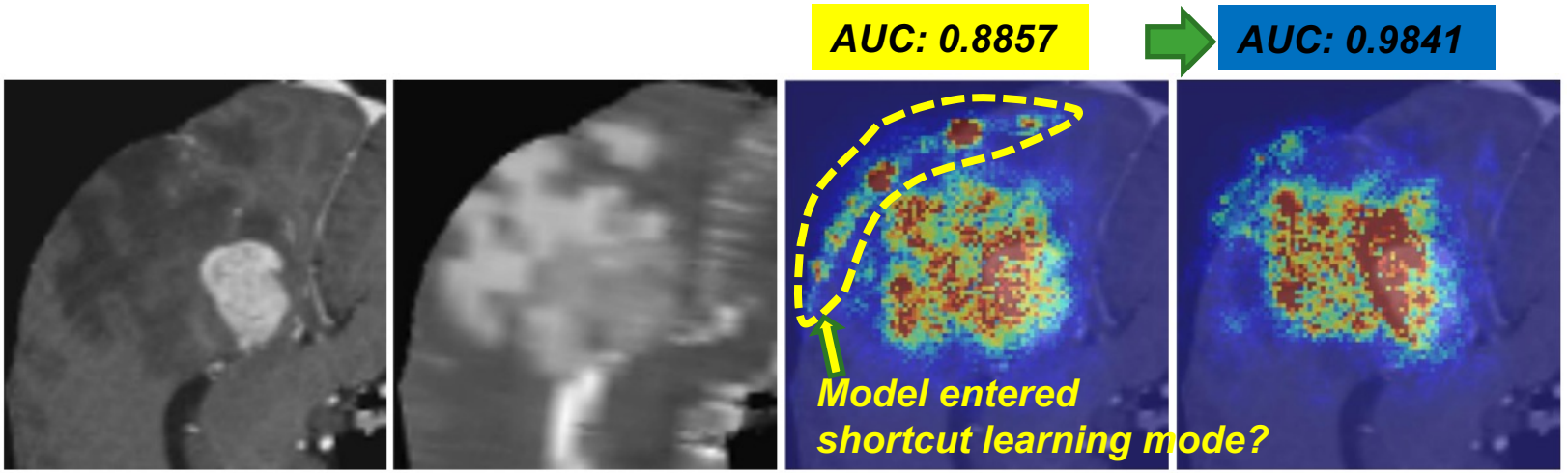


T1c ***T2*** ***Before*** ***After***

Saliency Maps

Interpretability in AI Medical Imaging

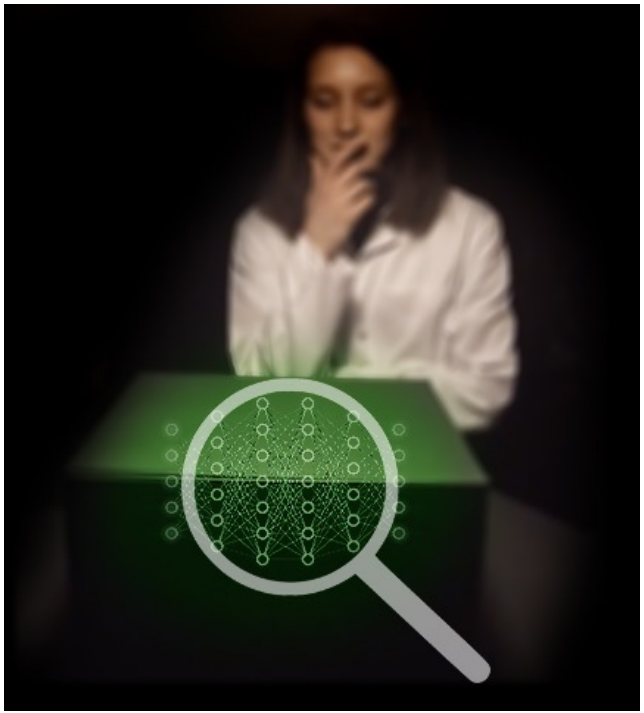
Key finding: interpretability enhances data preparation and AI-performance



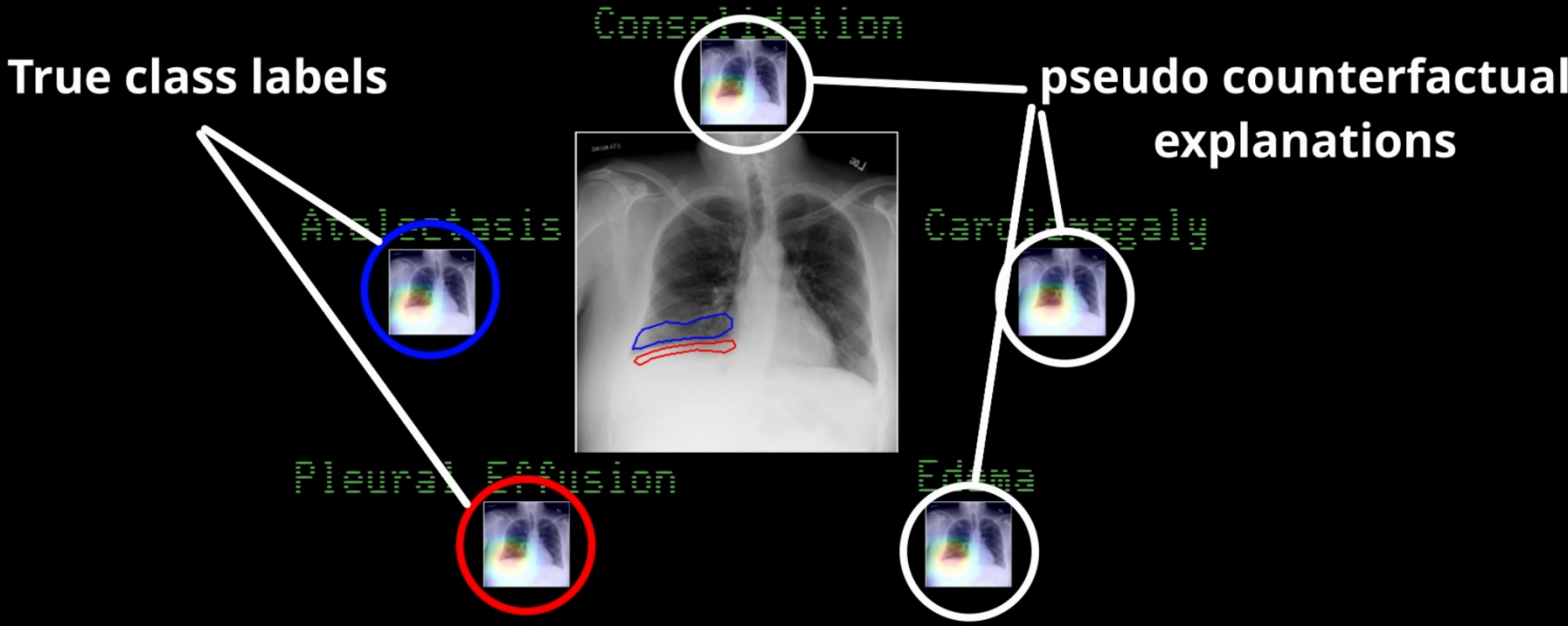
Pereira et al. MICCAI-iMIMIC 2018
Reyes et al. Radiology: Artificial Intelligence 2021

Beyond Interpretability?

Can we use this information for other purposes? (a.k.a. XXAI)

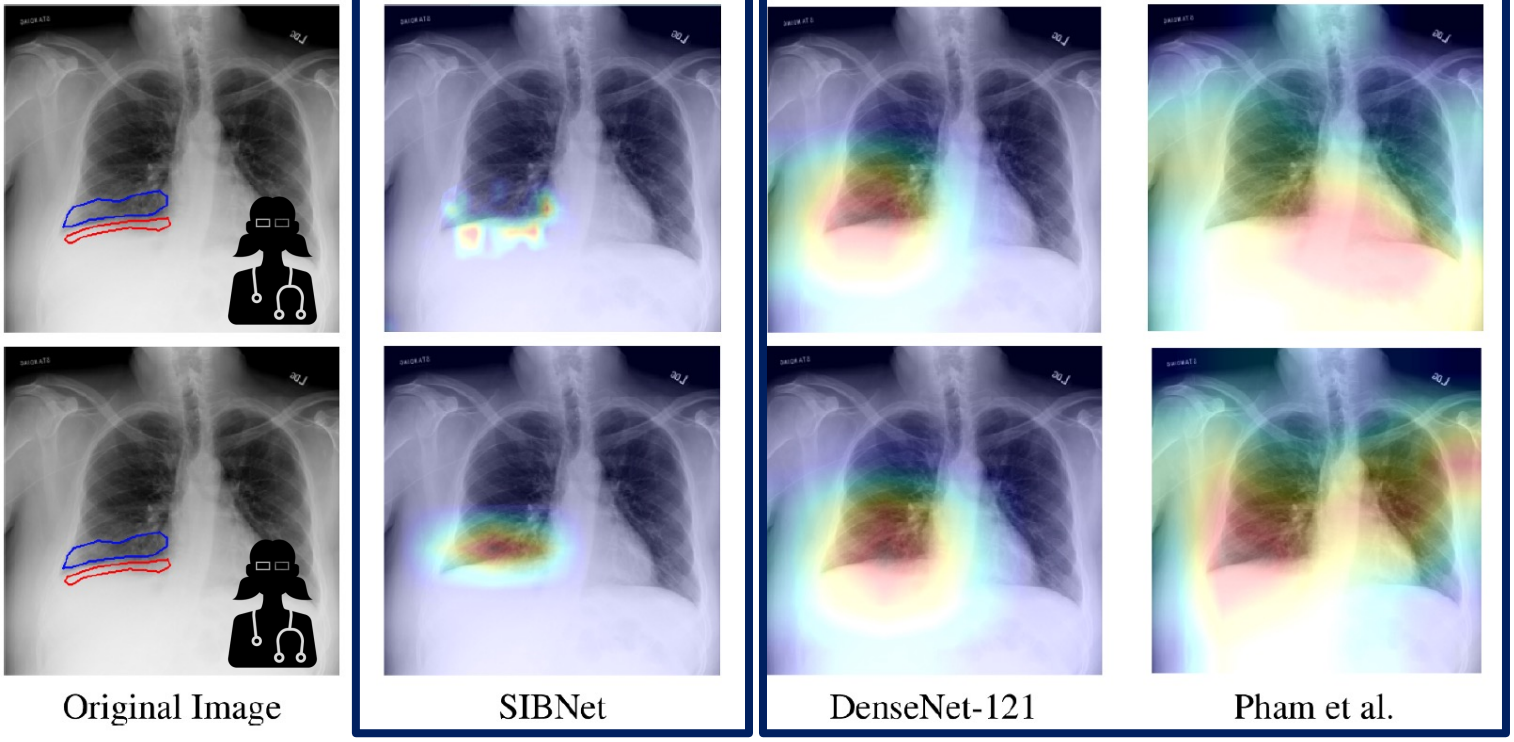


Intra-sample Saliency Maps



Results

- Qualitative comparison of saliency maps to expert-drawn maps



What is next?...incoming larger black box?

Omics Data

Genomics

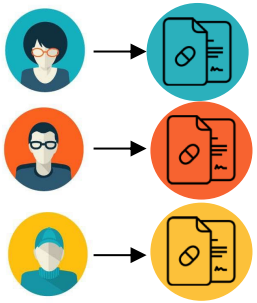
Transcriptomics

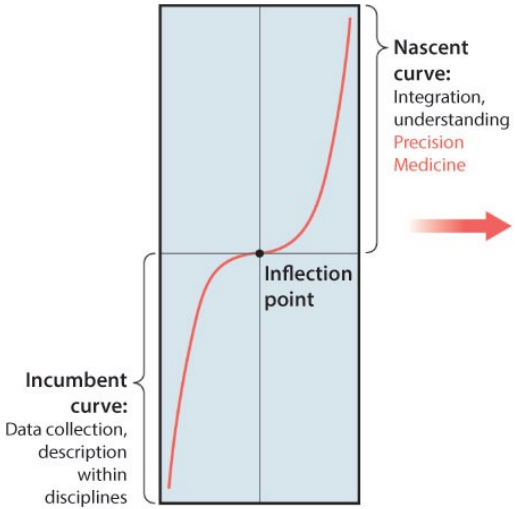
Epigenomics

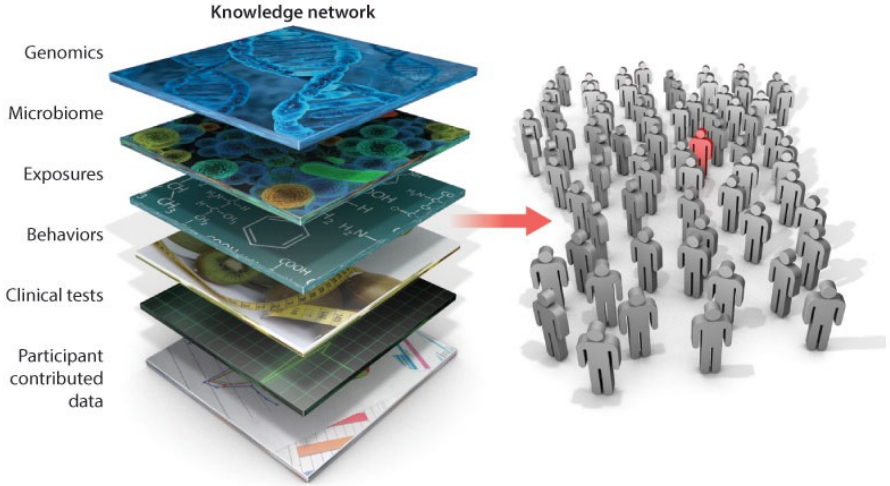
Proteomics

Metabolomics

And more ...

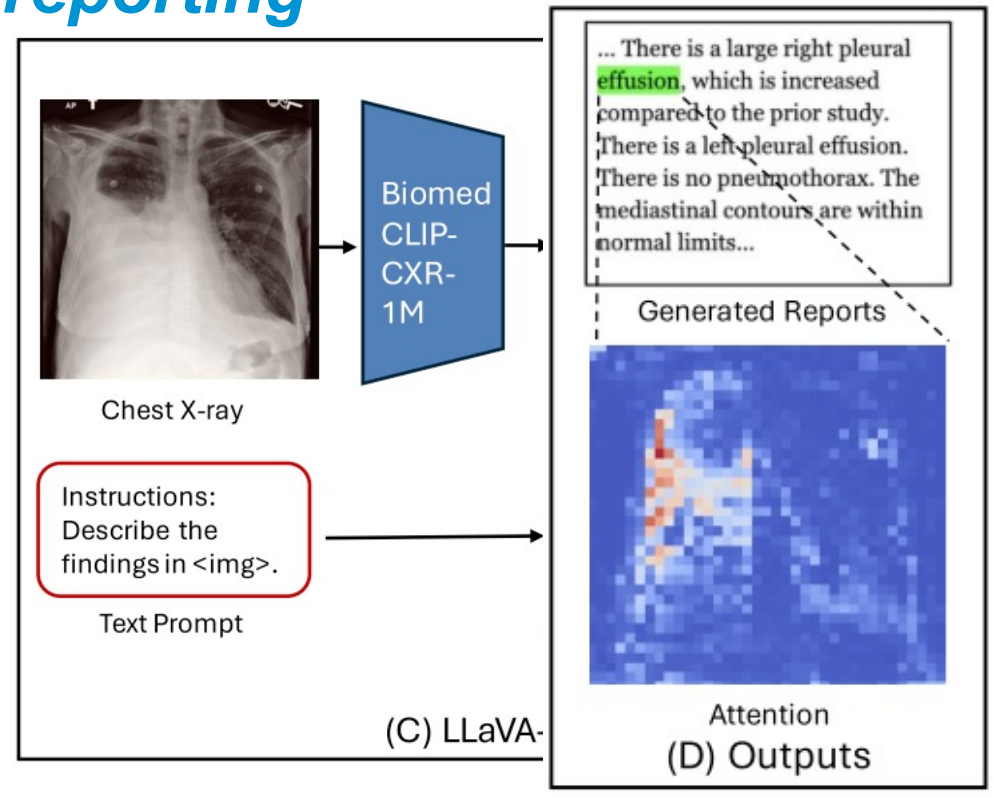






Source: Hawgood et al. 2015 Science Translational Medicine

Combining text and imaging: Automated radiology reporting



Training Small Multimodal Models to Bridge Biomedical Competency Gap: A Case Study in Radiology Imaging

Juan Manuel Zambrano Chaves[1⋄*] Shih-Cheng Huang[1⋄*], Yanbo Xu[*], Hanwen Xu[4*], Naoto Usuyama[*], Sheng Zhang[*], Fei Wang[2], Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen[3], Matthew P. Lungren, Serena Yeung-Levy[1], Curtis P. Langlotz[1], Sheng Wang[4,‡], Hoifung Poon[‡]

[1]Stanford University [2]University of Southern California

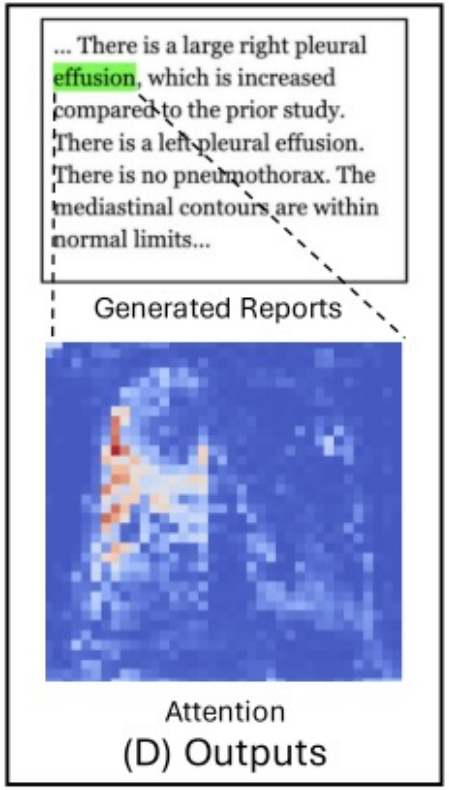
[3]University of California, Davis [4]University of Washington

Microsoft Research

Here "small" → 7 Billion param. Model

Strategy:
Fine-tune foundational models

Combining text and imaging: Automated radiology reporting



Training Small Multimodal Models to Bridge Biomedical Competency Gap: A Case Study in Radiology Imaging

Juan Manuel Zambrano Chaves[1⋄*] Shih-Cheng Huang[1⋄*], Yanbo Xu[*], Hanwen Xu[4*], Naoto Usuyama[*], Sheng Zhang[*], Fei Wang[2], Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen[3], Matthew P. Lungren, Serena Yeung-Levy[1], Curtis P. Langlotz[1], Sheng Wang[4,‡], Hoifung Poon[‡]

[1]Stanford University [2]University of Southern California

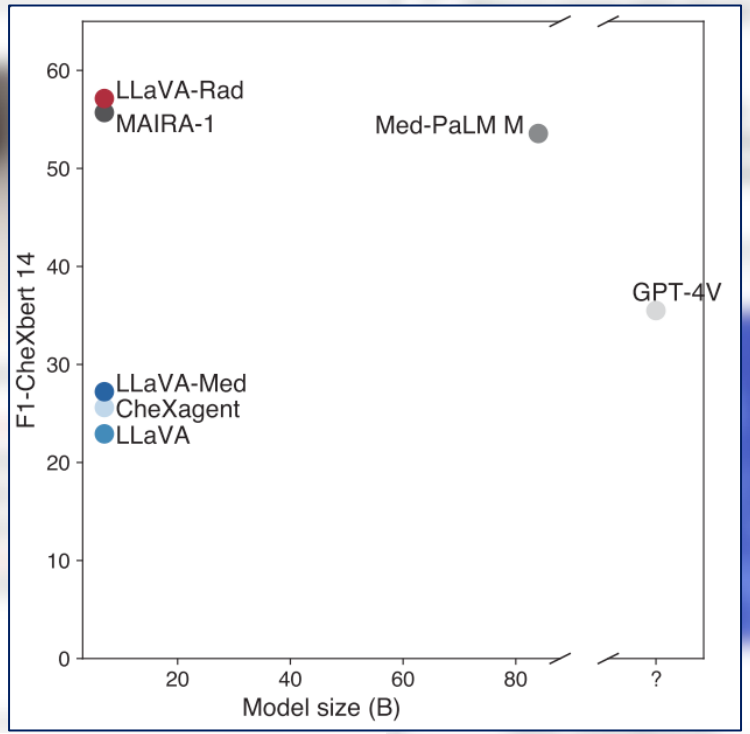
[3]University of California, Davis [4]University of Washington

Microsoft Research

Here "small" → 7 Billion param. Model

Strategy:
Fine-tune foundational models

Combining text and imaging: Automated radiology reporting



Training Small Multimodal Models to Bridge Biomedical Competency Gap: A Case Study in Radiology Imaging

Juan Manuel Zambrano Chaves[1⋄*] Shih-Cheng Huang[1⋄*], Yanbo Xu[*], Hanwen Xu[4*], Naoto Usuyama[*], Sheng Zhang[*], Fei Wang[2], Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen[3], Matthew P. Lungren, Serena Yeung-Levy[1], Curtis P. Langlotz[1], Sheng Wang[4,‡], Hoifung Poon[‡]

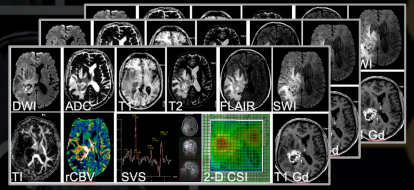
[1]Stanford University [2]University of Southern California

[3]University of California, Davis [4]University of Washington

Microsoft Research

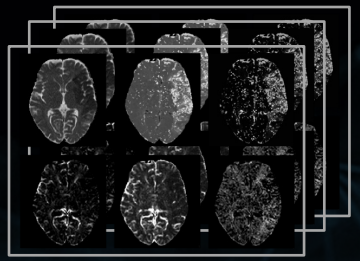
Here "small" → 7 Billion param. Model

Strategy:
Fine-tune foundational models

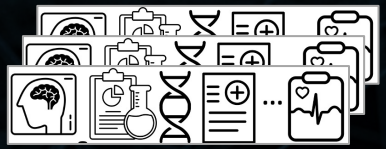


Task-specific Model - 1

Task-specific Model -2



Task-specific Model - 3



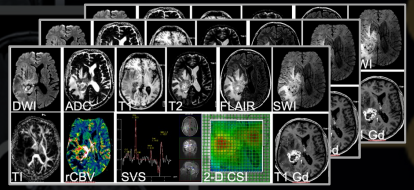
Task-specific Model - N

Task-specific Model - 1

Task-specific Model - 2

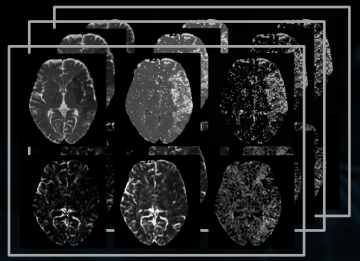
Task-specific Model - 3

Task-specific Model - N

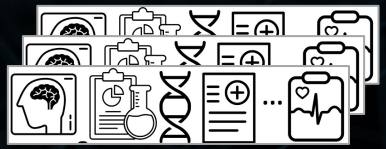


Task-specific Model - 1

Task-specific Model -2



Task-specific Model - 3



Task-specific Model - N

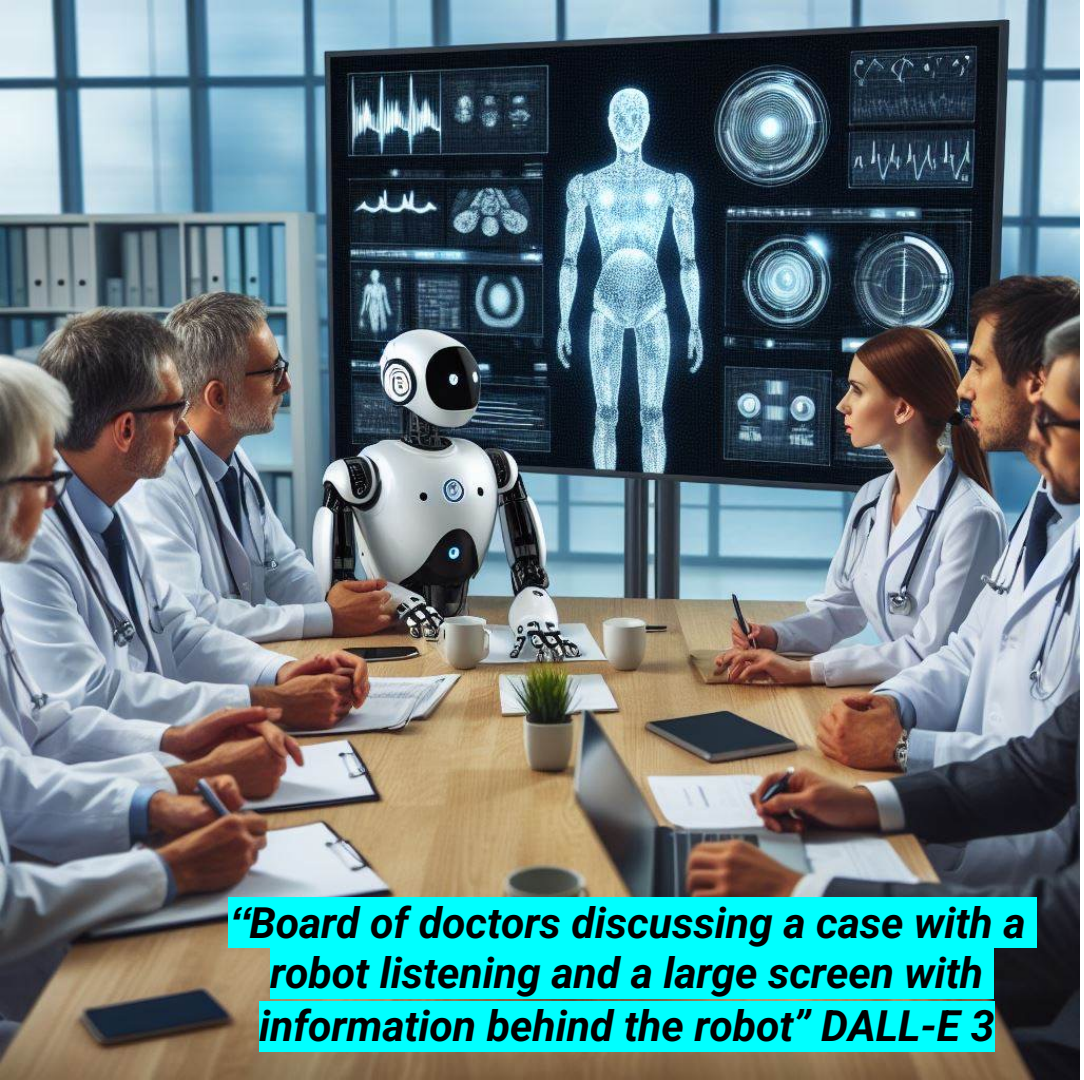
Task-specific Model - 1

Task-specific Model - 2

Task-specific Model - 3

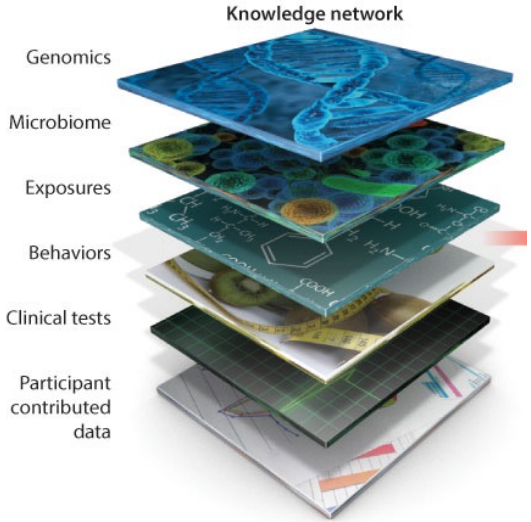
Task-specific Model - N

Orchestrator/ Smart assistant



"Board of doctors discussing a case with a robot listening and a large screen with information behind the robot" DALL-E 3

AI as **counselor/assistant** as opposed to "I tell you what to do" type of entity



Home-take message

- *Embracing **data-driven & human-centered** AI approaches in Medicine!*
- ***XAI** technologies to enhance the **trustworthiness and verification** of AI systems.*
- ***Clinically-oriented AI training/guidance** becomes more essential than ever.*
- ***Interconnected** and **orchestrated** AI for Medicine. Enlarged black box? What about curation and QC of multimodal data?*