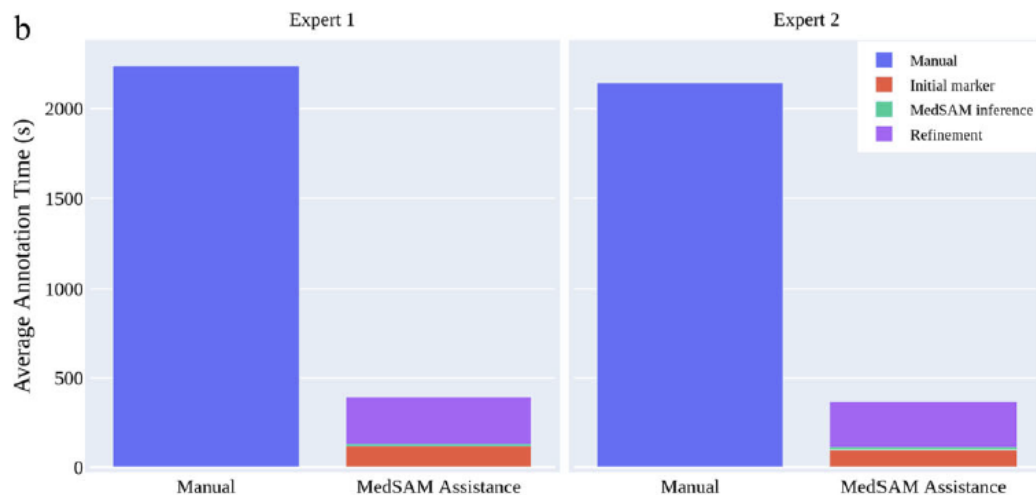


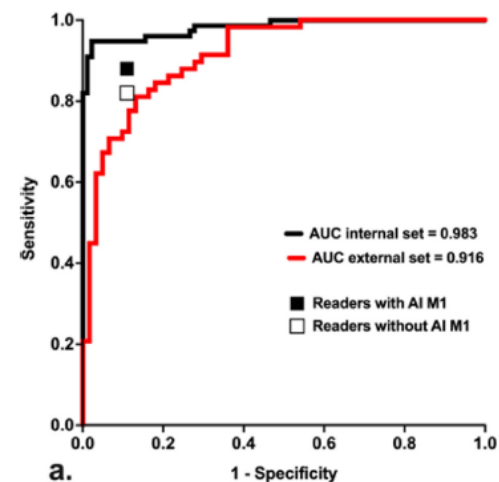
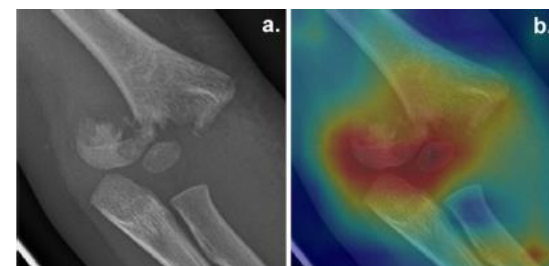
Human in the loop

Large purpose segmentation



Ma et al Nat Comm 2024

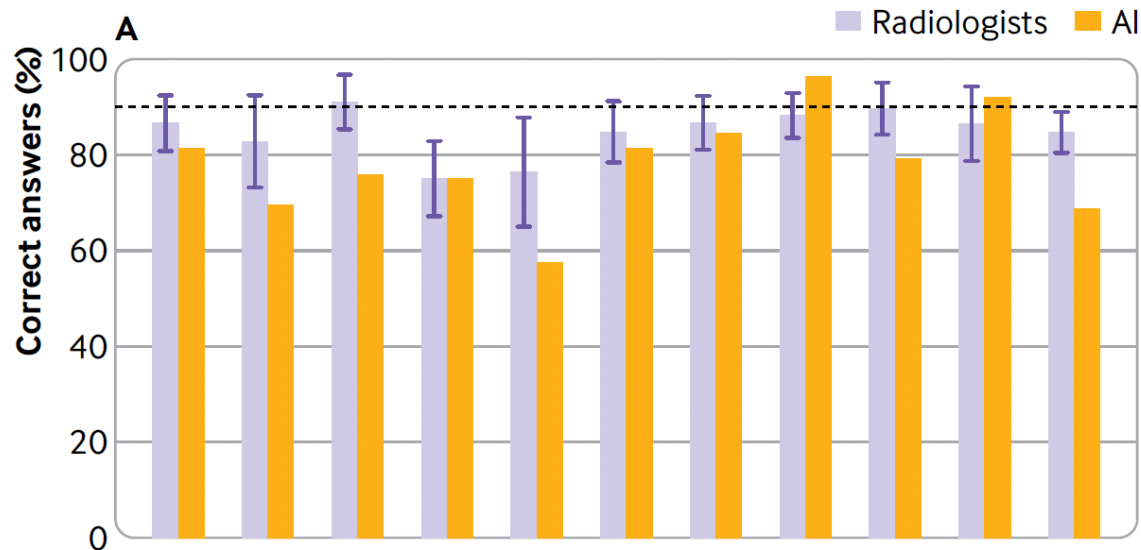
Elbow trauma detection



Rozwag et al Res Dia Int Imaging 2023

Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination?

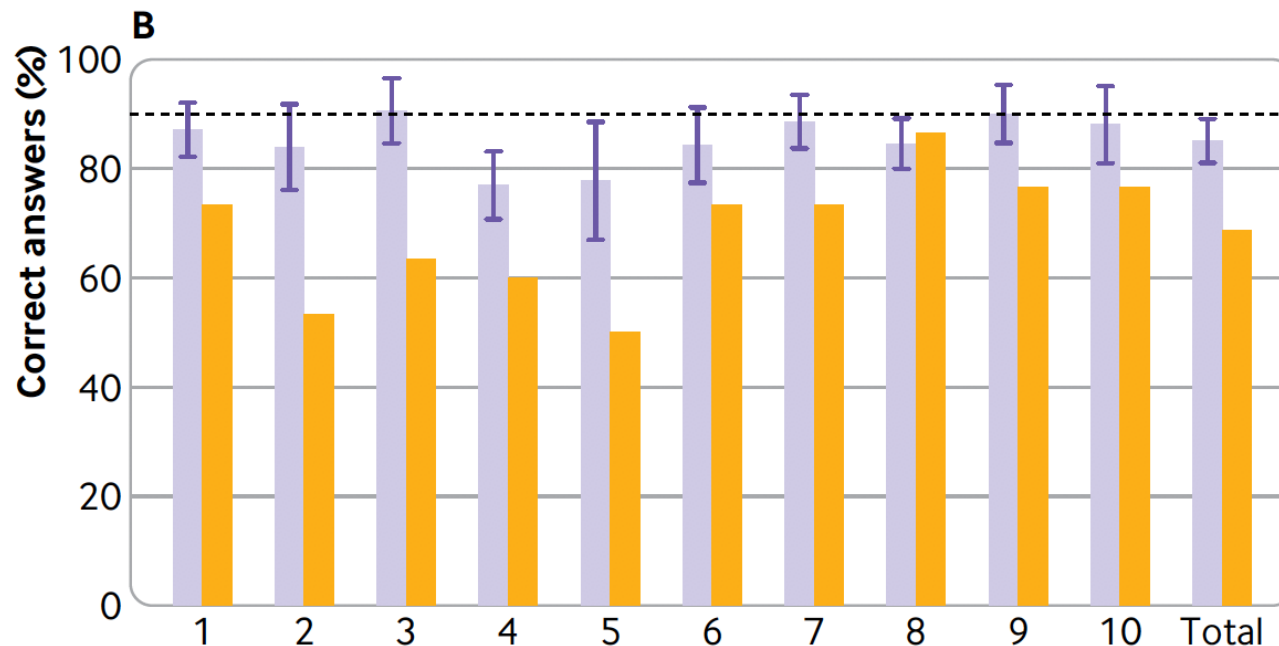
26 radiologists vs Milvue
(certified to analyse musculoskeletal and chest radiographs)



<http://dx.doi.org/10.1136/bmj-2022-072826> Chen et al. BMJ 2022

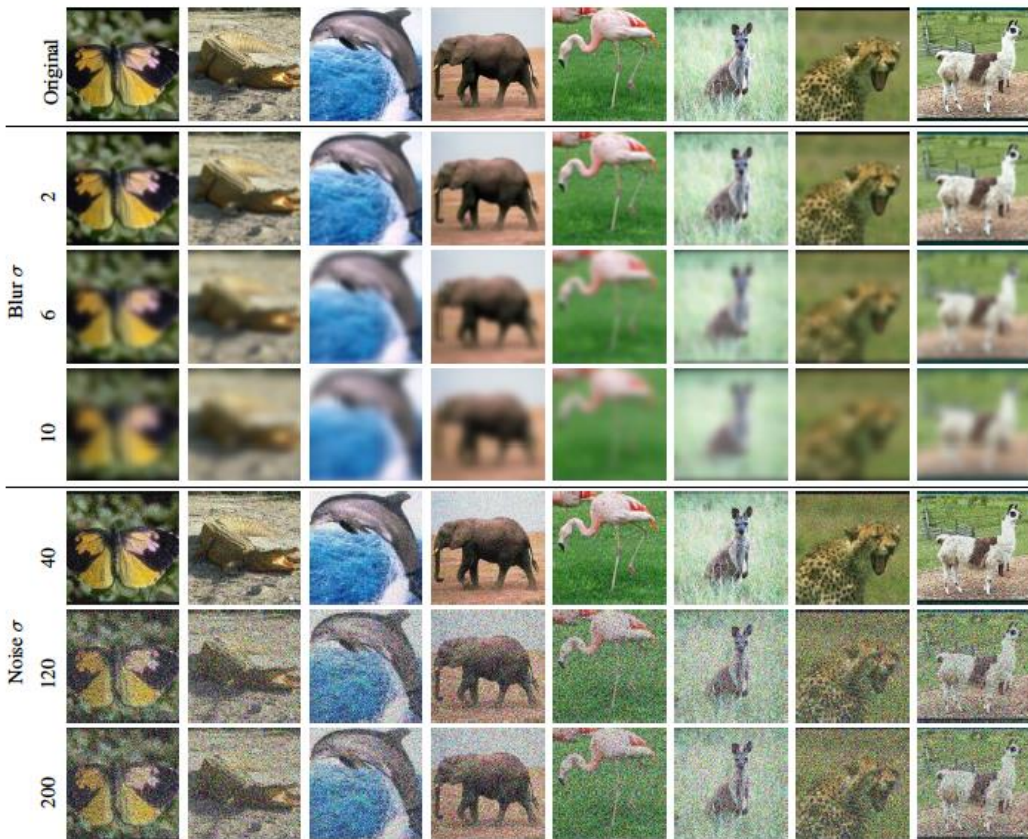
Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination?

axial skeleton (skull, spine, dental views) or abdominal radiographs

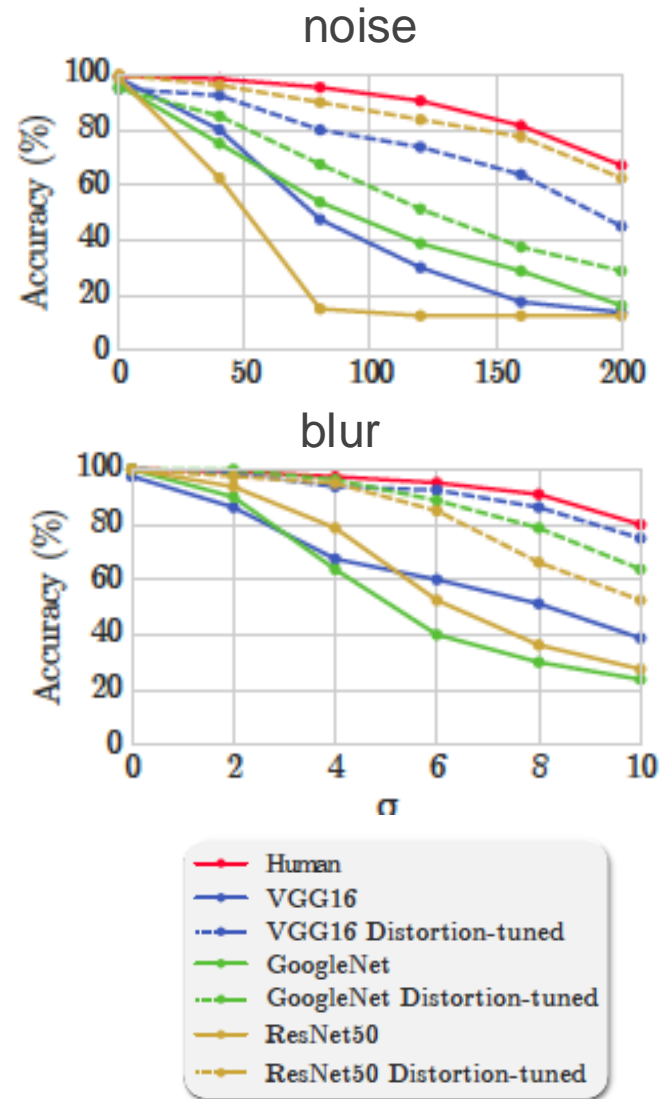


<http://dx.doi.org/10.1136/bmj-2022-072826> Chen et al. BMJ 2022

Nobody is perfect ...



[Dodge and Karam ICCV 2017]]





[Geirhos ICLR 2019]

Bias toward texture ...



(a) Texture image

81.4%	Indian elephant
10.3%	indri
8.2%	black swan



(b) Content image

71.1%	tabby cat
17.3%	grey fox
3.3%	Siamese cat

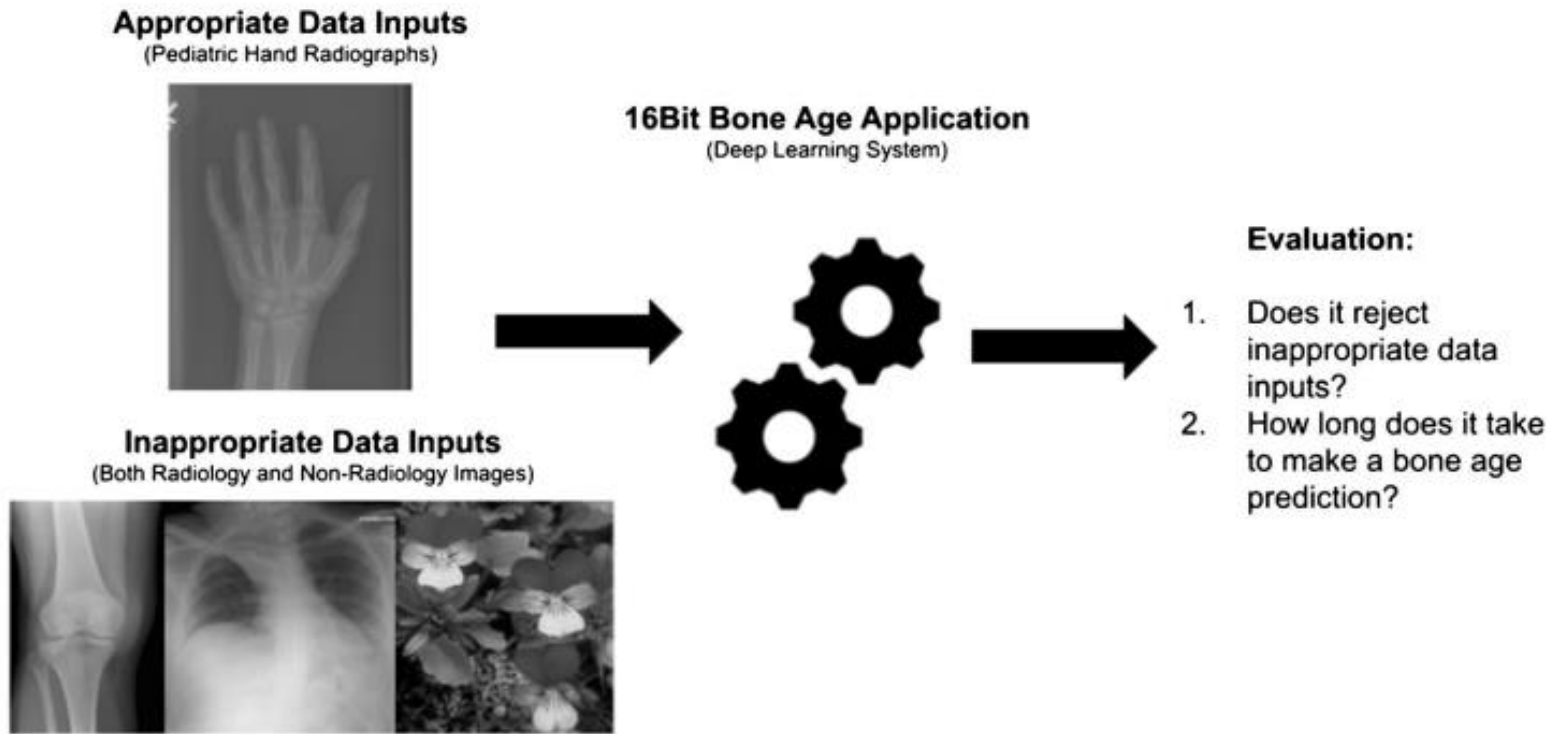


(c) Texture-shape cue conflict

63.9%	Indian elephant
26.4%	indri
9.6%	black swan

[Geirhos ICLR 2019]

bone age application (<https://www.16bit.ai/bone-age>) [14], which analyzes pediatric left-hand posterior–anterior (PA) radiographs and automatically returns the predicted bone age



Yi et al Skeletal Radiology (2022) 51:401–406

No control ...



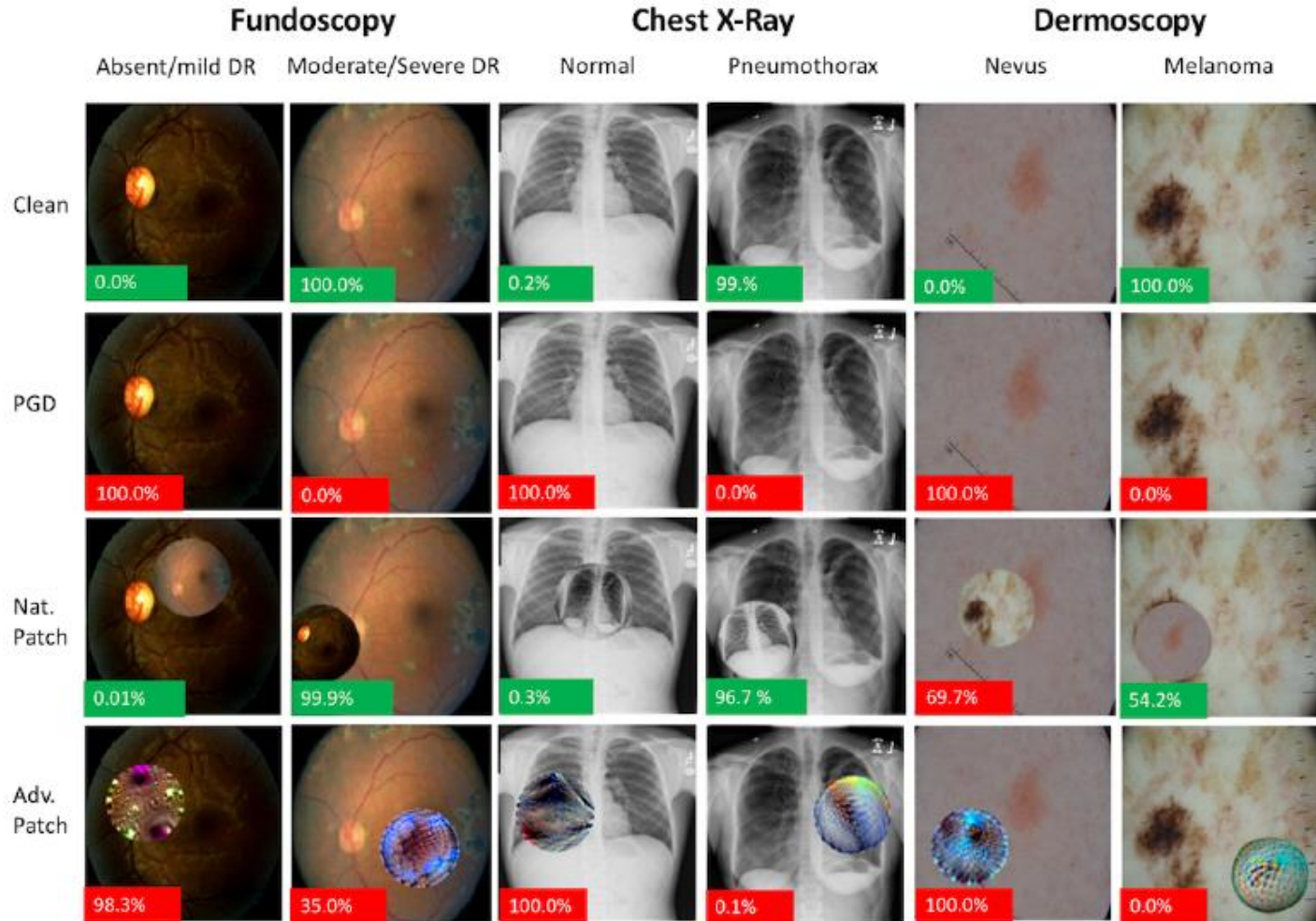
Predicted Bone Age:
13 years, 9 months
Inference Time:
1.1783 seconds

Predicted Bone Age:
1 year, 1 month
Inference Time:
1.0324 seconds

Predicted Bone Age:
15 years, 11 months
Inference Time:
1.0376 seconds

<https://doi.org/10.1007/s00256-021-03880-y>

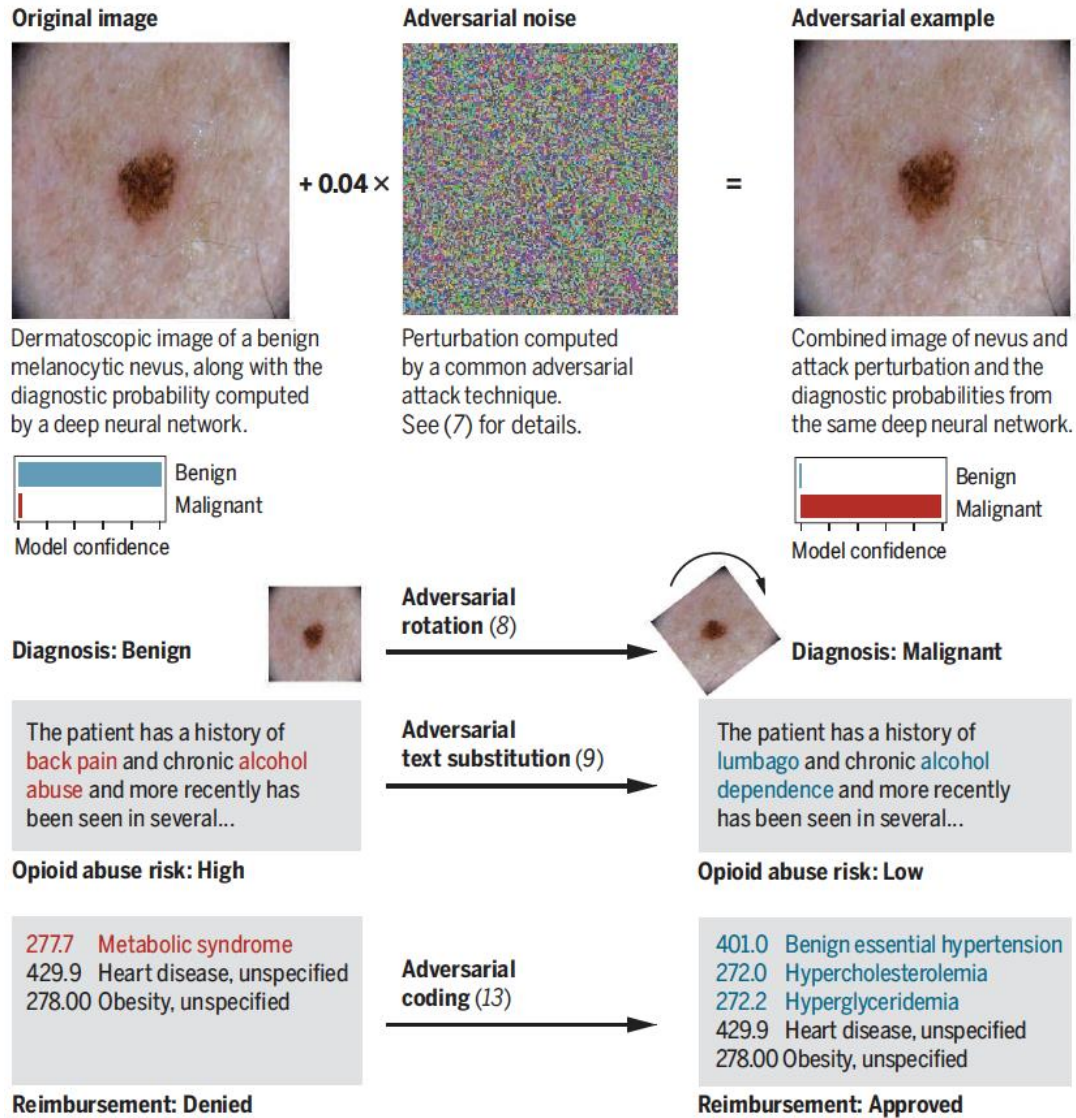
Vulnerable



[Finlayson et al ACM 2019]

Vulnerable

[Finlayson et al Science 2019]



Explicability/ Interpretability / Trustable

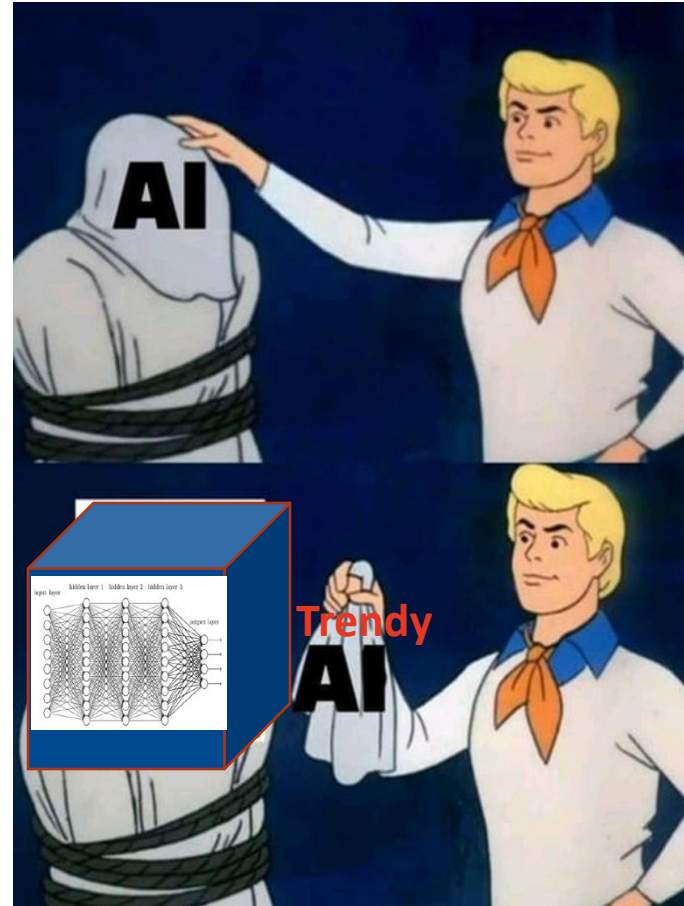
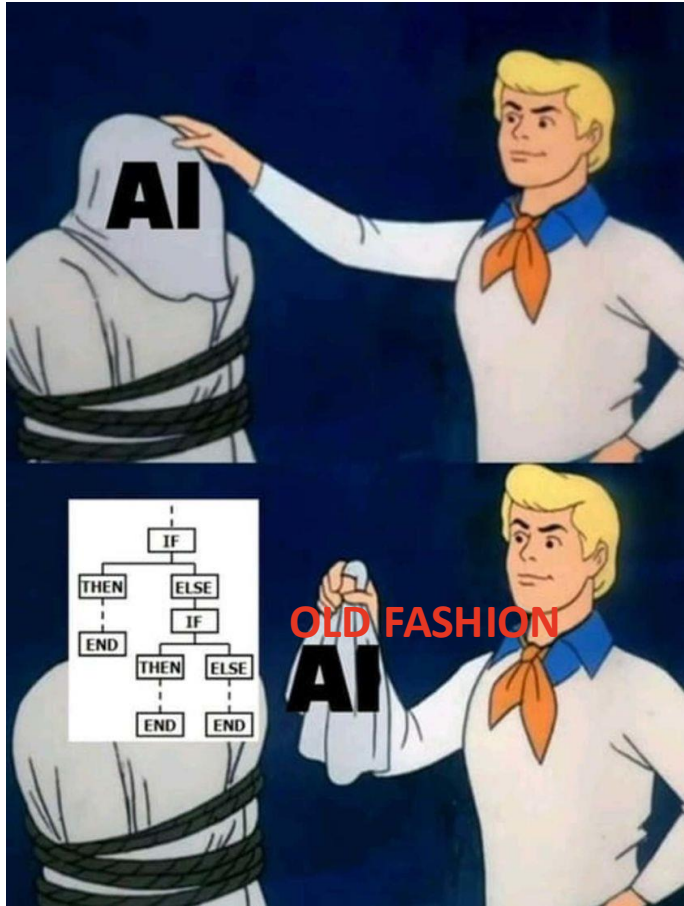


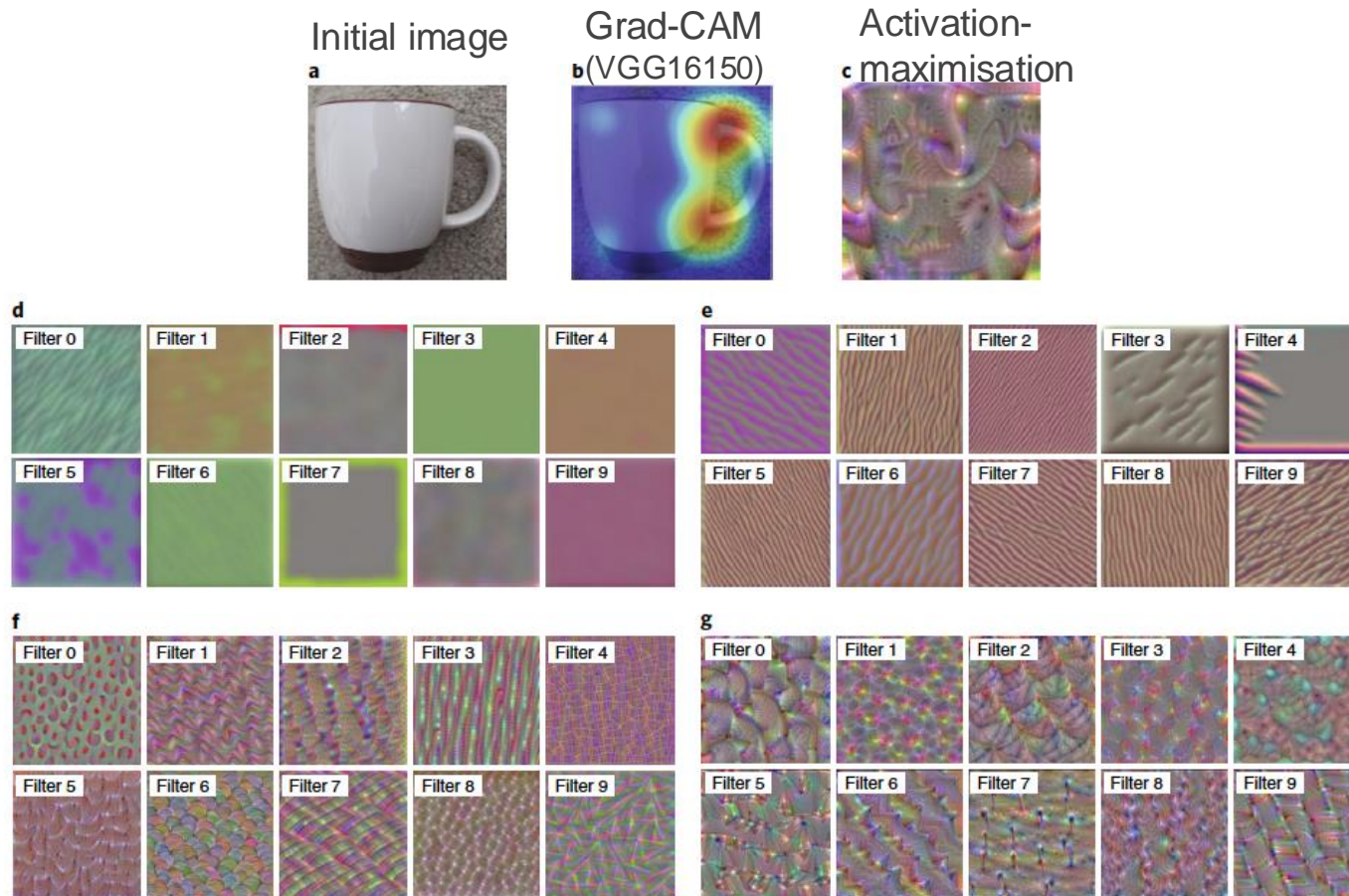
Table 3 | Comparisons between human evaluations and different types of AI approaches

Approaches	Model comprehensibility	Performance	Reproducibility	Dependency on prior knowledge	Development and training costs ^a	Running costs	Around-the-clock availability	Update costs
Human evaluation	High	Moderate or high	Moderate	High	High	High	Low	High
Rule-based algorithms	High or Moderate	Moderate or high	High	High	Moderate or high	Low	High	High
Feature-based machine-learning methods	Moderate or high	Moderate or high	High	Moderate ^b	Moderate	Low	High	Moderate ^c
Deep artificial neural networks	Low or moderate	High	High	Low	Moderate	Low	High	Low

^aThe estimated cost of training professionals that carry out the clinical tasks (human evaluation) or of developing the automated system (rule-based, feature-based or deep-artificial-neural-network-based) that performs the tasks. ^bFor feature-based machine-learning methods, prior knowledge may facilitate the derivation of useful features from the raw data. ^cWhen the update requires encoding new features, the update cost of feature-based machine-learning methods includes feature engineering and model retraining.

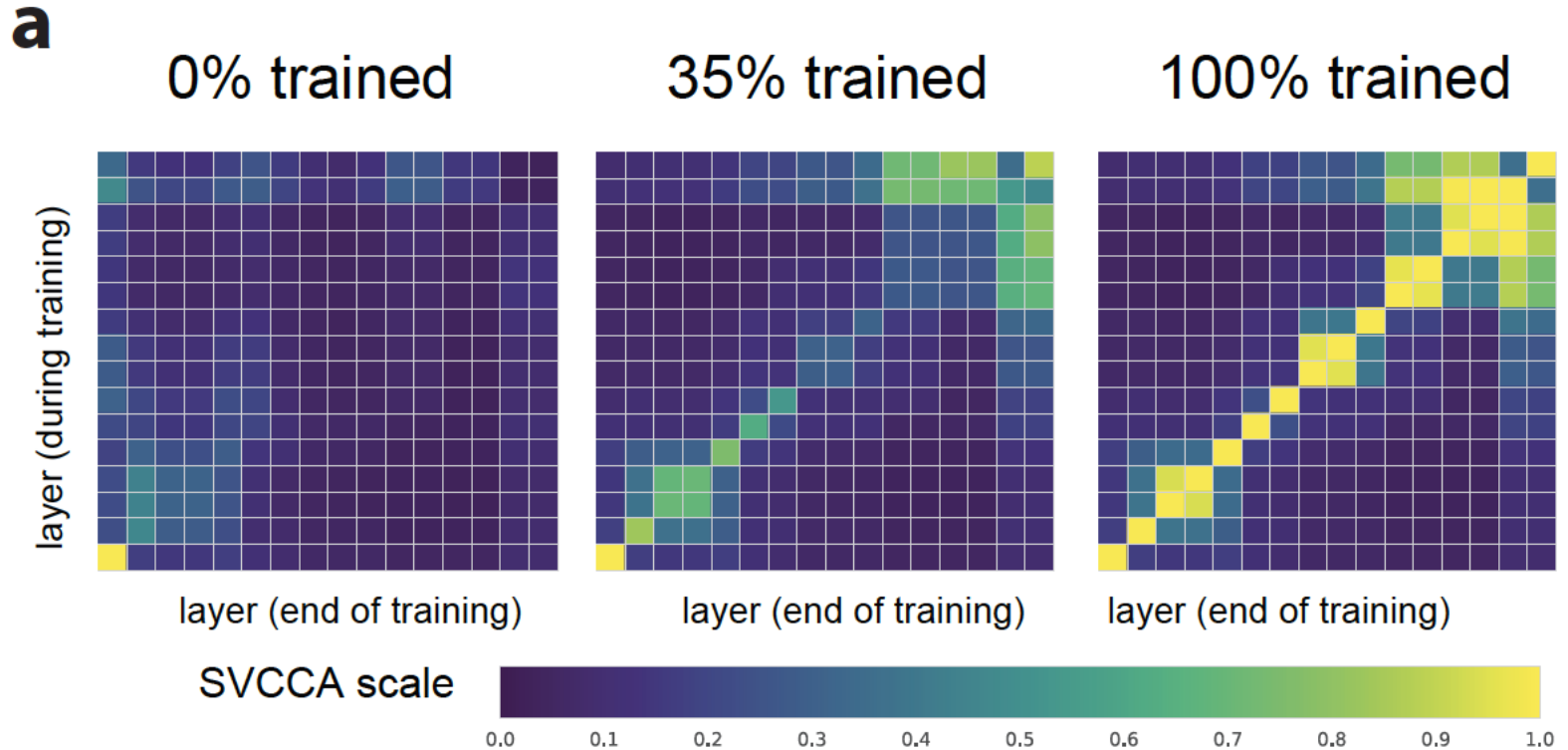
[Yu et al Nat Bio Eng 2018]

Interpretation



[Yu et al Nat Bio Eng 2018]

What has been learnt?



[Raghu et al Ad Neur Inf Proc Sys 2017]

What has been learnt?

Kamnitsas et al Media 2017

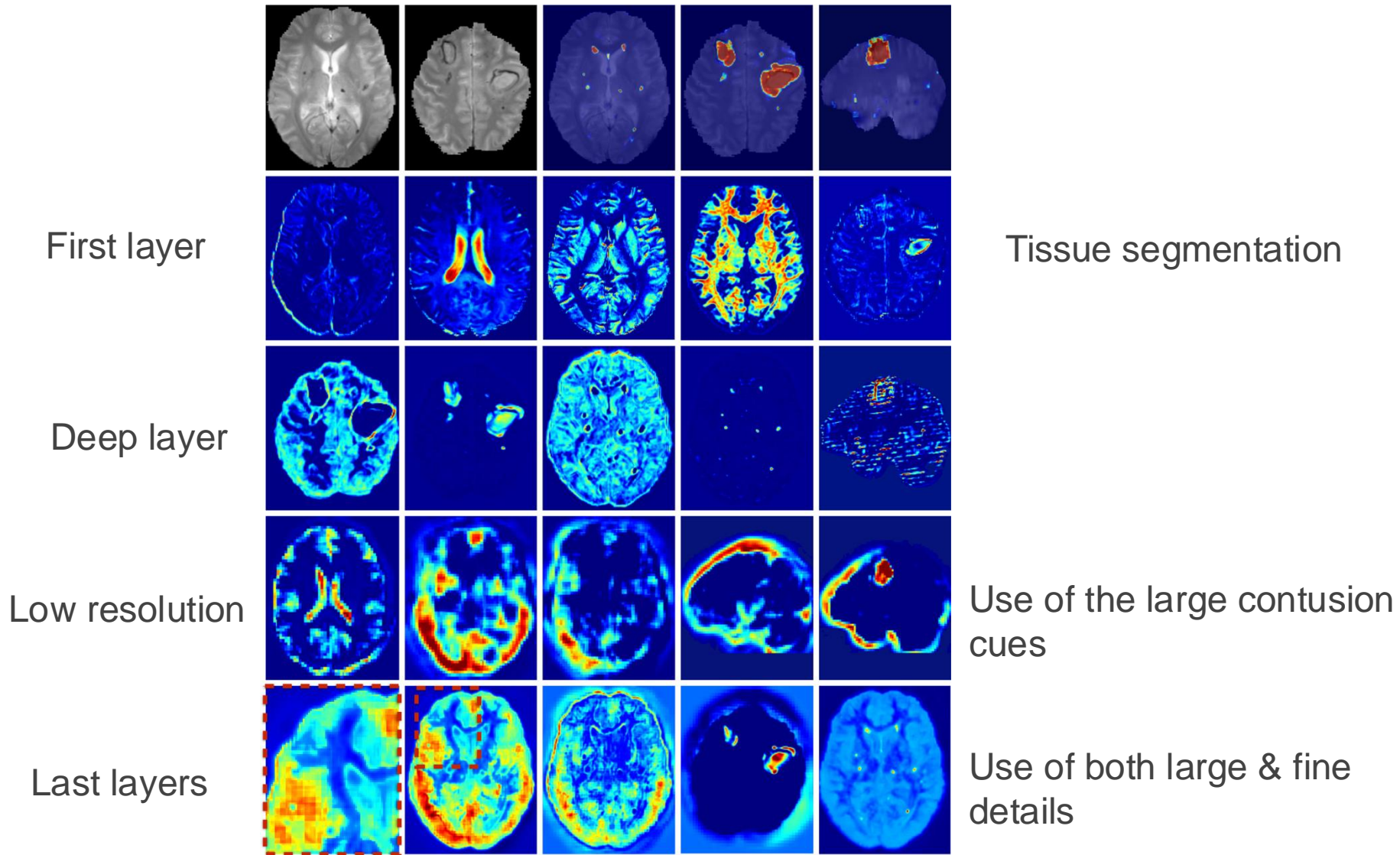
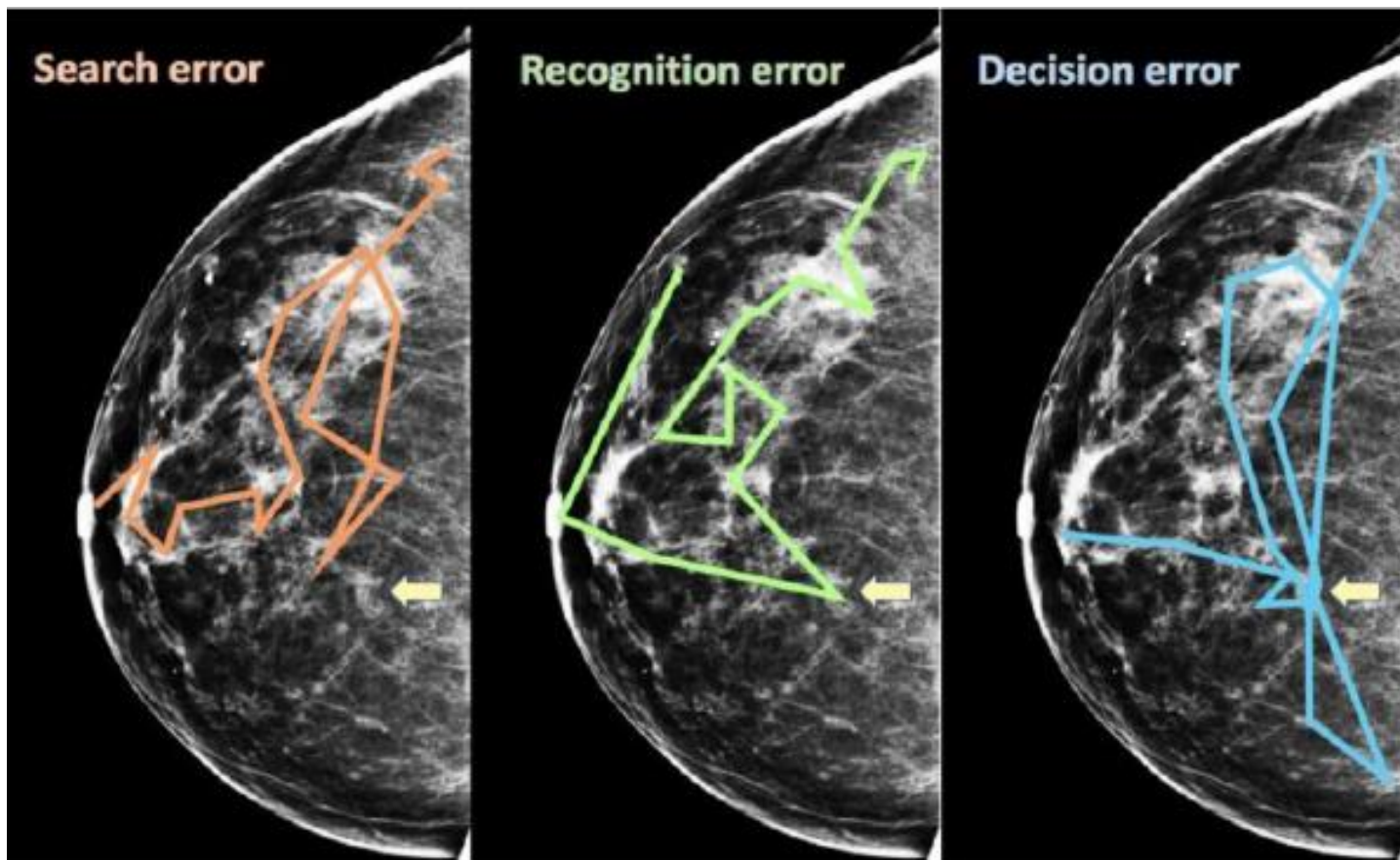


Image exploration: Man vs Machine



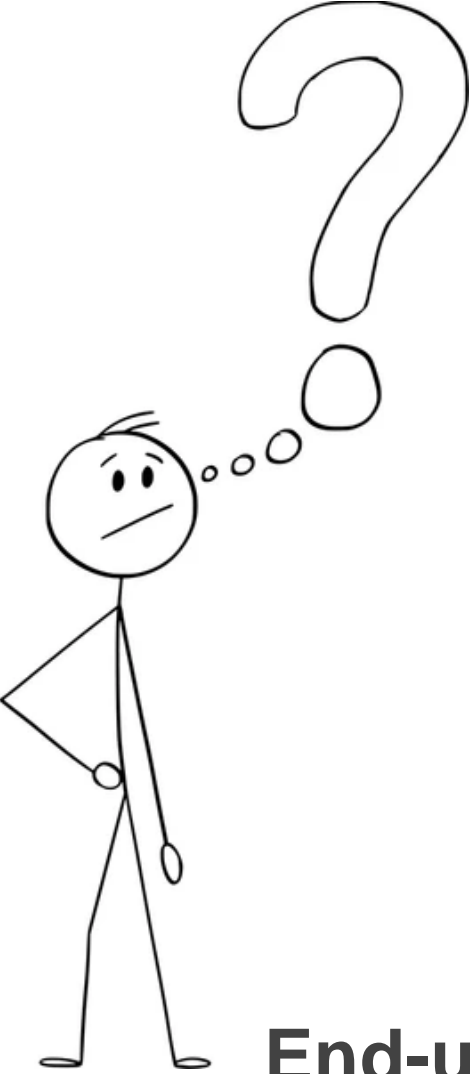
Wu et al. 2019 Vision

Image exploration: Man vs Machine

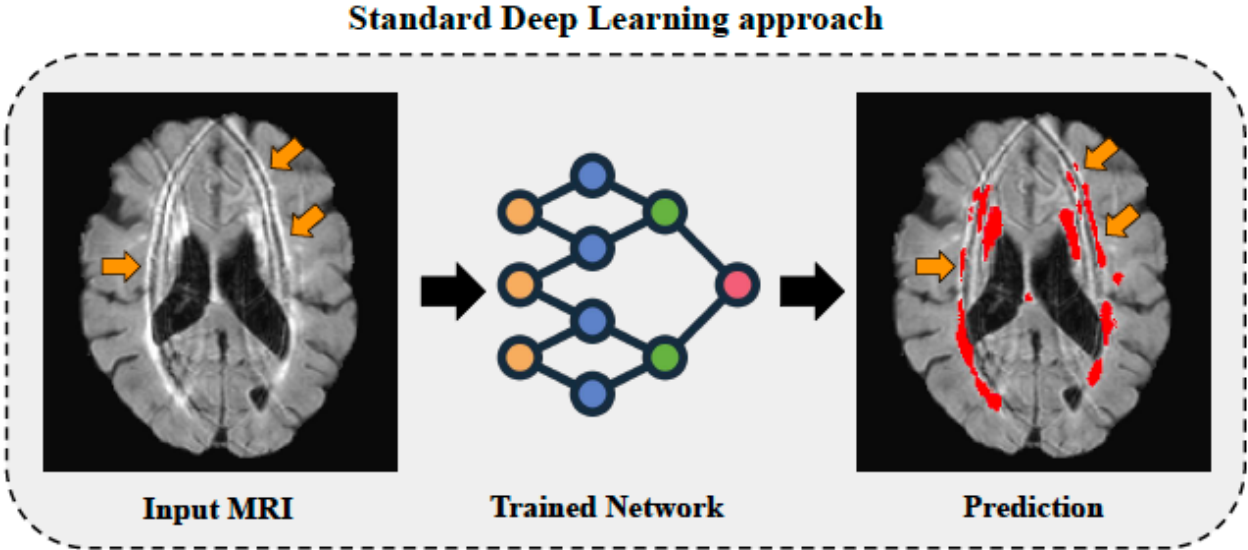
Matrice de confusion	Originale	ResNet34	Mean 11	Mean 11
			Experts	Novices
Vrai positif				
Vrai Négatif				
Faux positif				
Faux Négatif				

Vallée R, PhD thesis 2022

The curse of the Black-Box



End-user

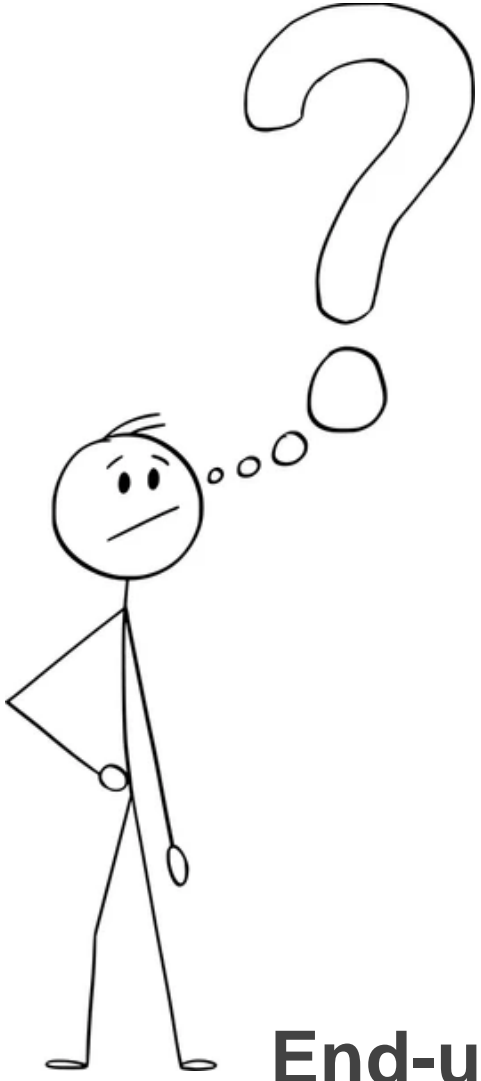


The curse of the Black-Box

Explainability (XAI)
Interpretability
Understandability

*For who?
About what?
At which level?
...*

[Erasmus et al 2021 Philosophy & Technology]



End-user

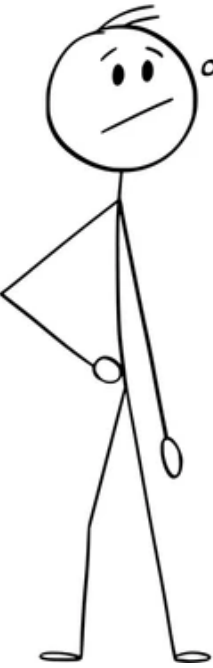


The curse of the Black-Box



~~Explainability
Interpretability~~

*For who?
About what?
At which level?
...*



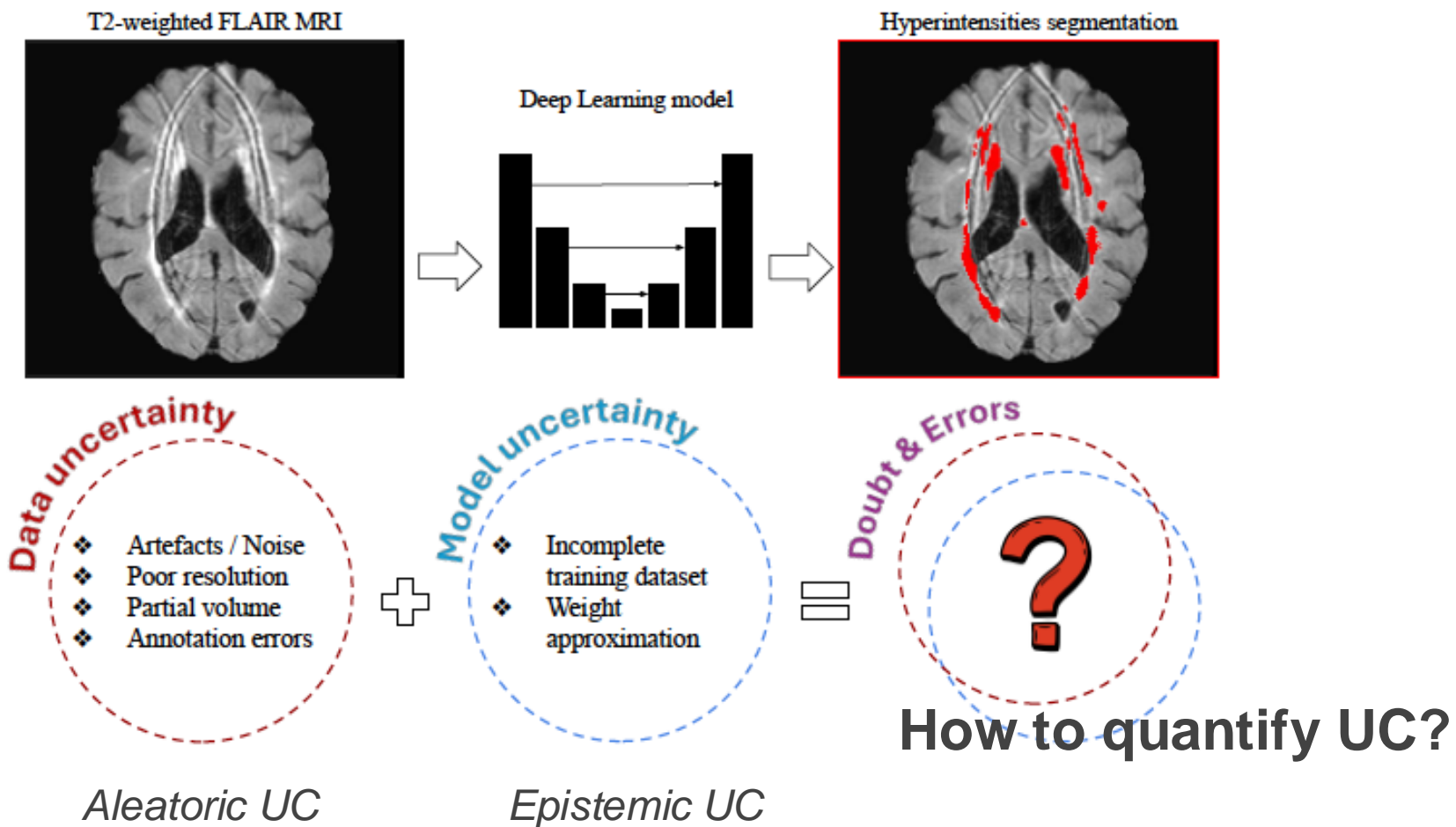
To Improve Confidence

To trust in AI

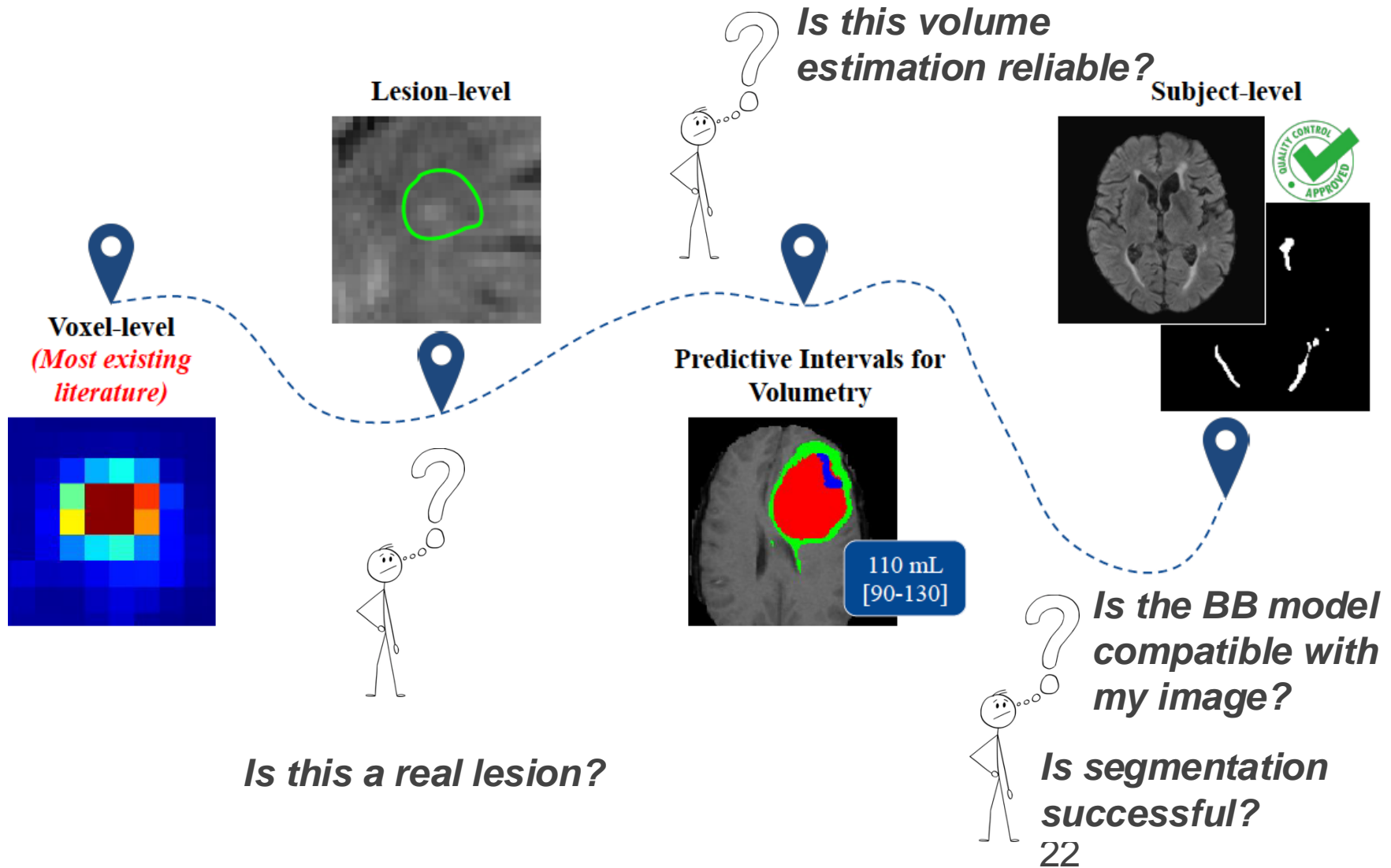


*Validation
Usage conditions
Adverse effects
.....*

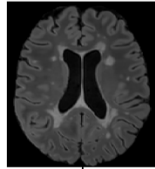
Where is uncertainty (UC) hidden?



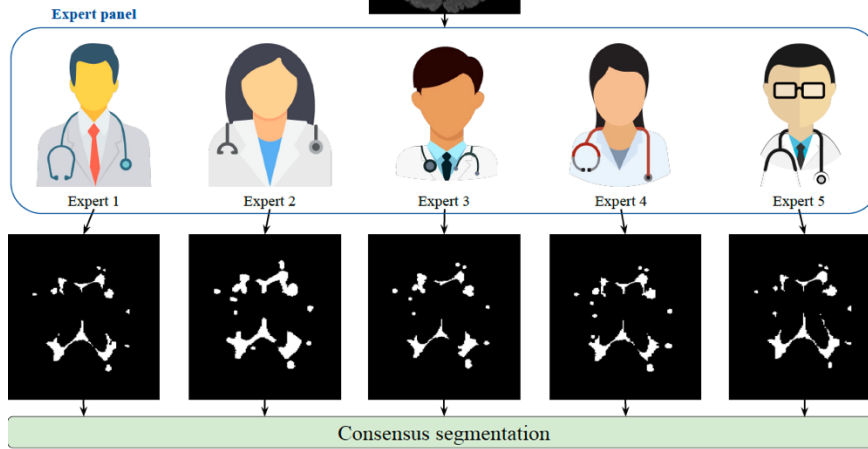
The multi-dimensional aspect of uncertainty



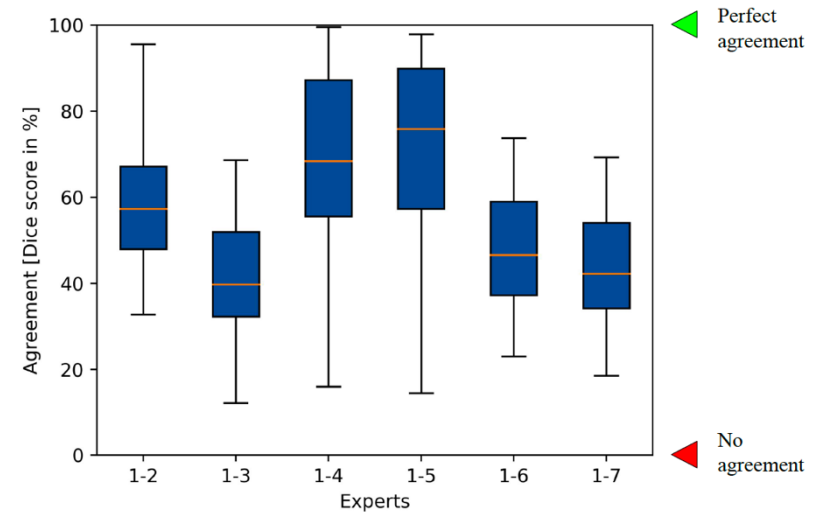
Consensus meeting



Definition of ground truth

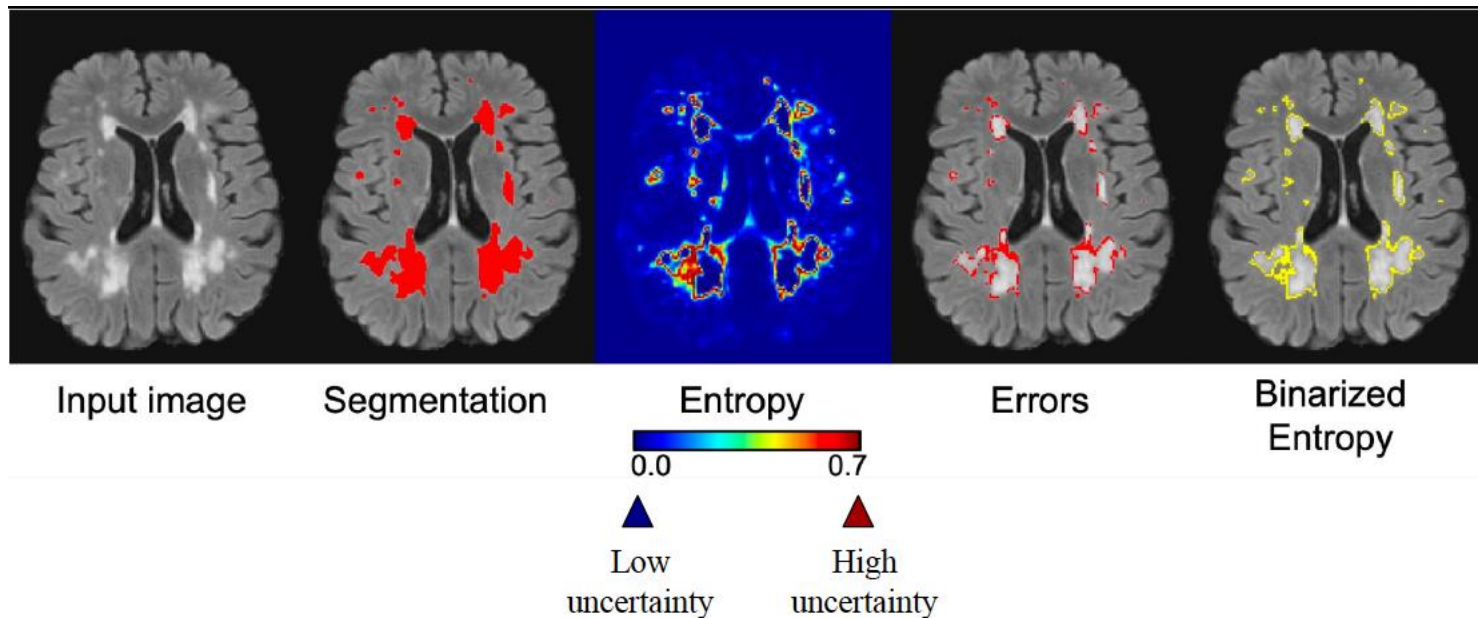
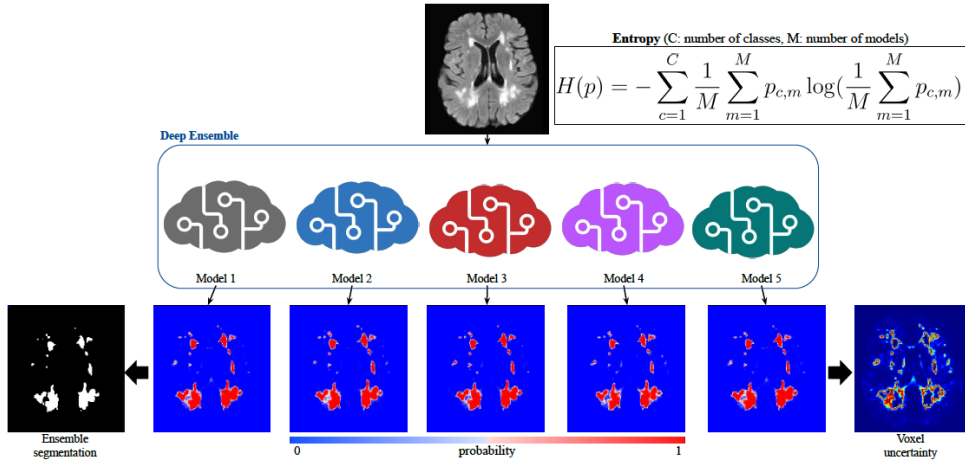


Inter-expert variability



[Commowick et al NeuroImage 2021]

Voxel level UQ: Ensembling

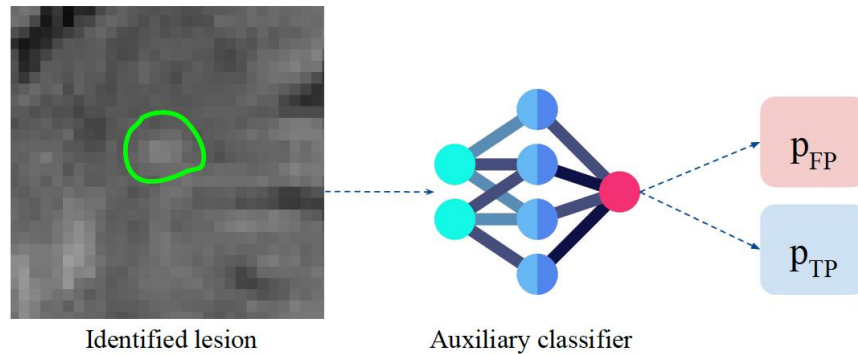


Lesion level UQ

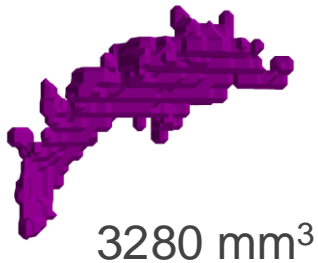
Goal: Identify FP

Estimate P_{FP} : proba that the lesion is FP using an auxillary classifier

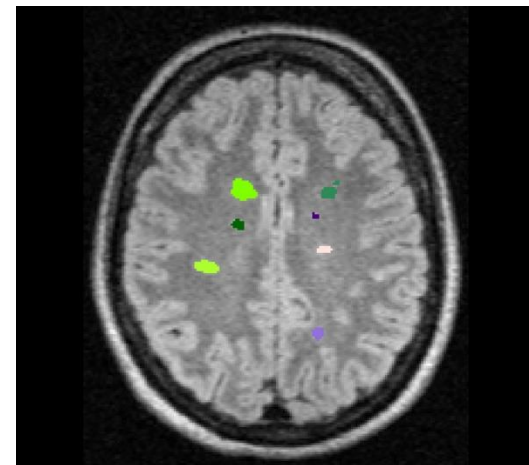
Lesion level UQ



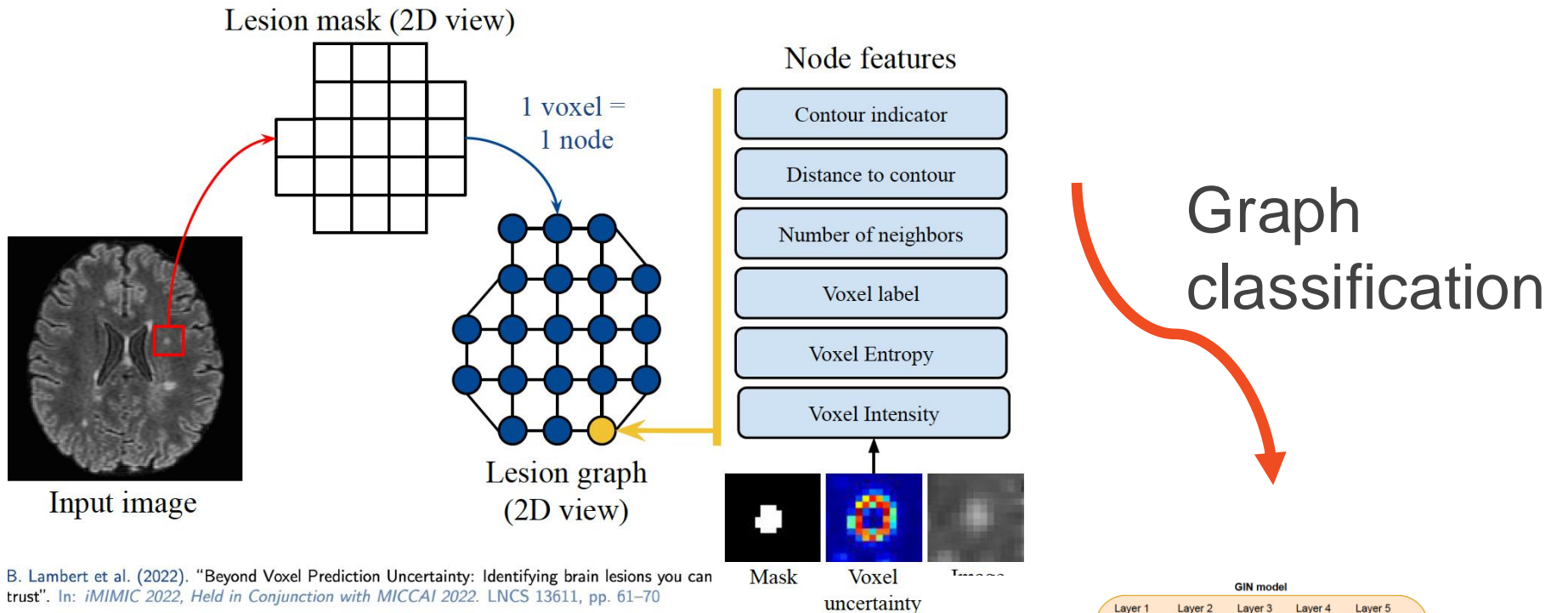
Note: Lesion are highly variable **in shape**



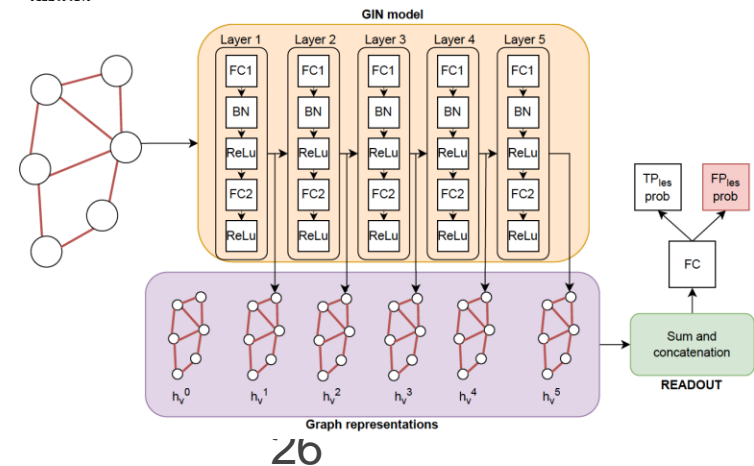
and **position**



Lesions as Graphs



Training to separate
True and False Positive lesions



The Graph Isomorphism Network

Deep Ensemble (5 models)



Multicentric, 3 Tesla
MSSEG 2016 - ISBI 2015 - WMH 2017
(N = 219)



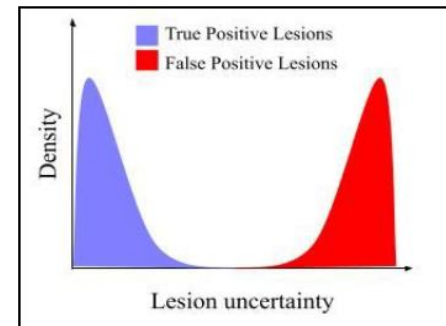
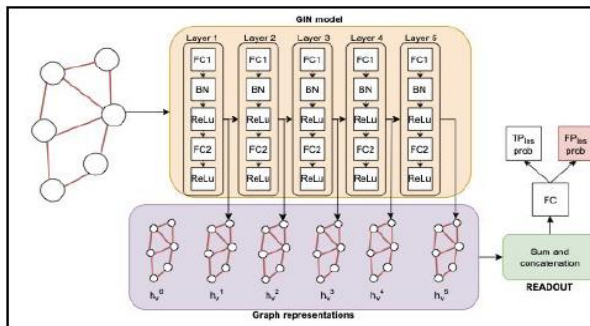
10775 identified lesions (9050 TP / 1725 FP)

Fit

6854 TP
1197 FP

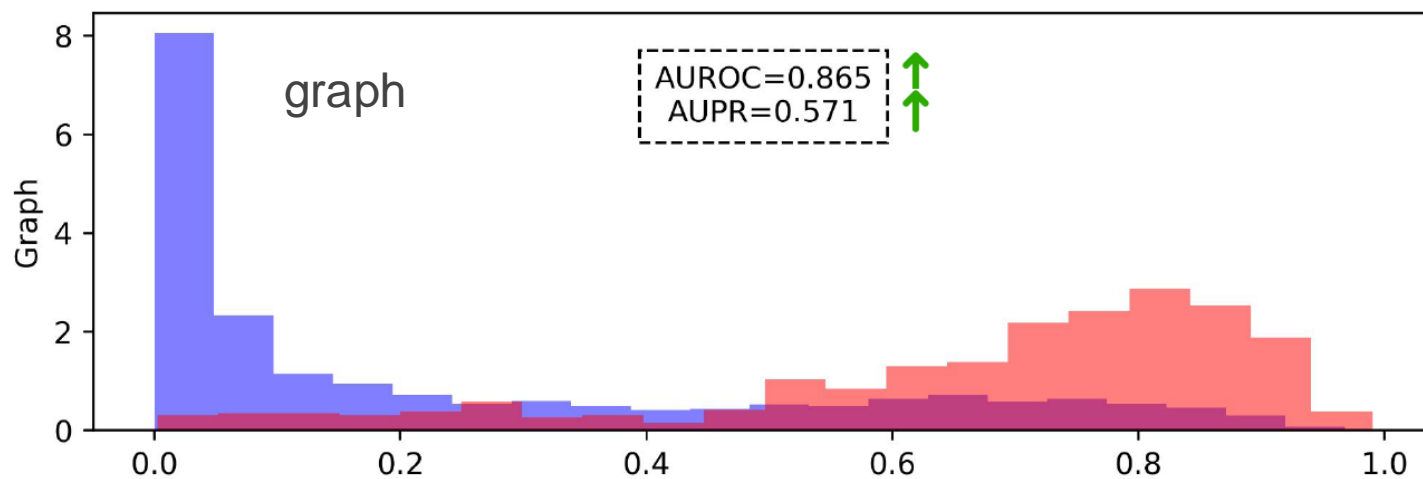
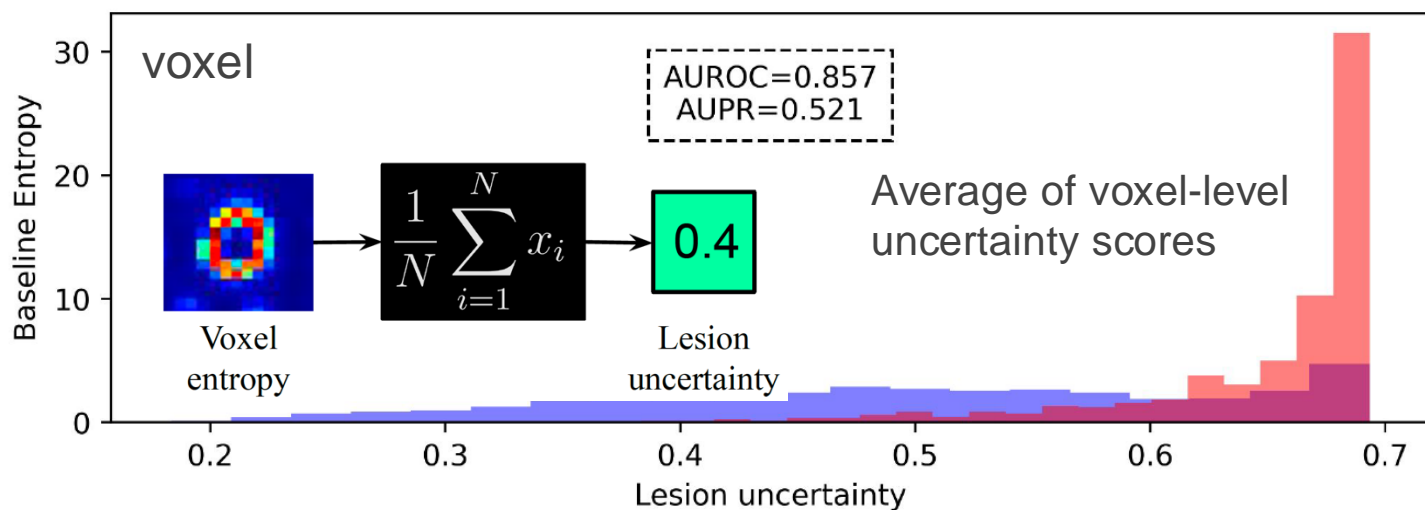
Test

2196 TP
528 FP

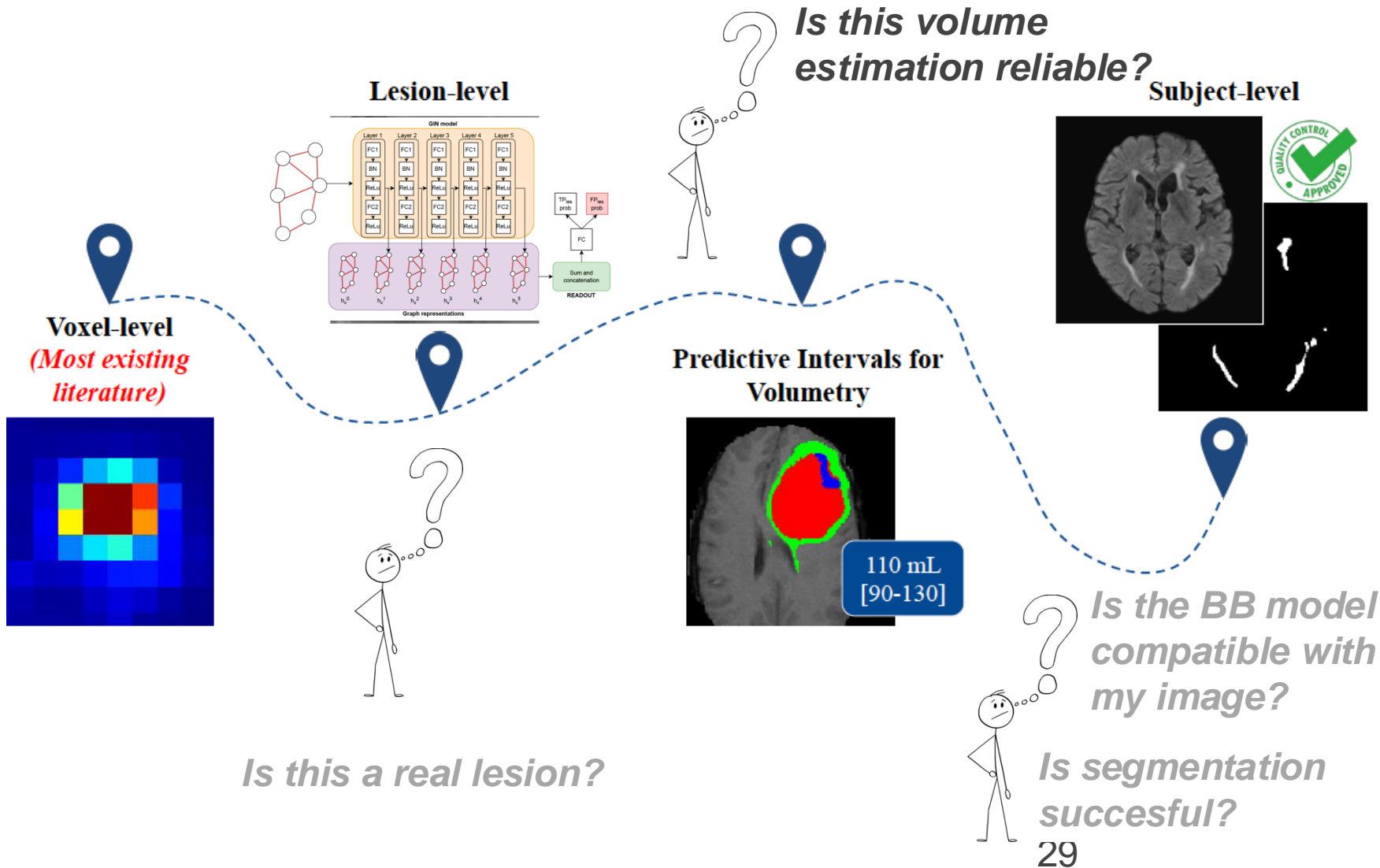


FC: Fully-connected layer. BN: Batch Normalization. ReLU: Rectified Linear Unit. ~~27~~ Parameters: 26 700

The Graph Isomorphism Network



The multi-dimensional aspect of uncertainty



Predictive intervals in volumetry

Definition

- $X \in \mathbb{R}^{N-1}$ are estimates of the true volumes $Y \in \mathbb{R}^{N-1}$, obtained from the segmentation.
- A predictive interval $\Gamma_\alpha(X)$ is a range of values intended to encompass Y with a specified degree of confidence $1 - \alpha$ (e.g. 90%, 95%), so that $P(Y \in \Gamma_\alpha(X)) \geq 1 - \alpha$

Sampling-based approaches

- Sample a set of estimated volumes X_1, \dots, X_K for the given image.
 - Estimate the mean $\mu(X)$ and standard deviation $\sigma(X)$.
 - Assuming $Y|X \sim \mathcal{N}(\mu(X), \sigma(X))$
- $$\Gamma_\alpha(X) = [\mu(X) - z\sigma(X), \mu(X) + z\sigma(X)]$$

Limitations

- **Inference time**, due to the sampling procedure.
- The **normality assumption**, which may not always hold.
- **Lack of flexibility**, as intervals are symmetrical by design.

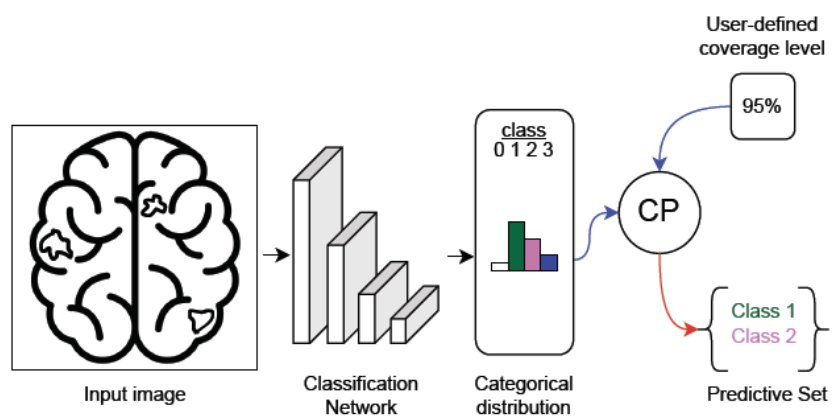
Predictive intervals in volumetry

Definition

- $X \in \mathbb{R}^{N-1}$ are estimates of the true volumes $Y \in \mathbb{R}^{N-1}$, obtained from the segmentation.
- A predictive interval $\Gamma_\alpha(X)$ is a range of values intended to encompass Y with a specified degree of confidence $1 - \alpha$ (e.g. 90%, 95%), so that $P(Y \in \Gamma_\alpha(X)) \geq 1 - \alpha$

Sampling-based approaches

- Sample a set of estimated volumes X_1, \dots, X_K for the given image.
- Estimate the mean $\mu(X)$ and standard deviation $\sigma(X)$.



Direct approaches

- Directly estimate the quantiles $\hat{t}_{\alpha/2}(X)$ and $\hat{t}_{1-\alpha/2}(X)$.
- The PI is computed as:

$$\Gamma_\alpha(X) = [\hat{t}_{\alpha/2}(X), \hat{t}_{1-\alpha/2}(X)] \quad (2)$$

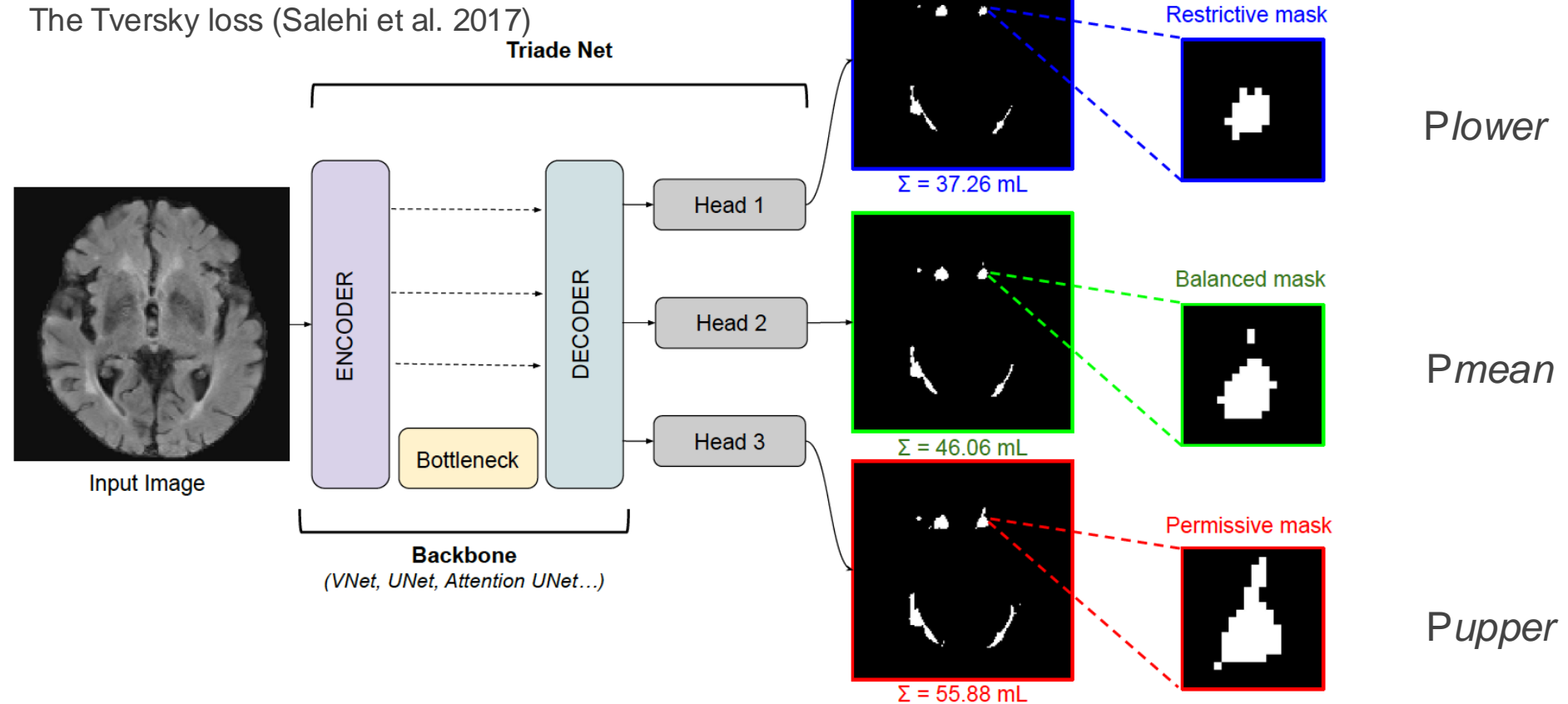
Conformal prediction

The TriadNet approach

$$\mathcal{T}_{\gamma,\beta} \approx 1 - \frac{TP}{TP + \gamma FP + \beta FN}$$

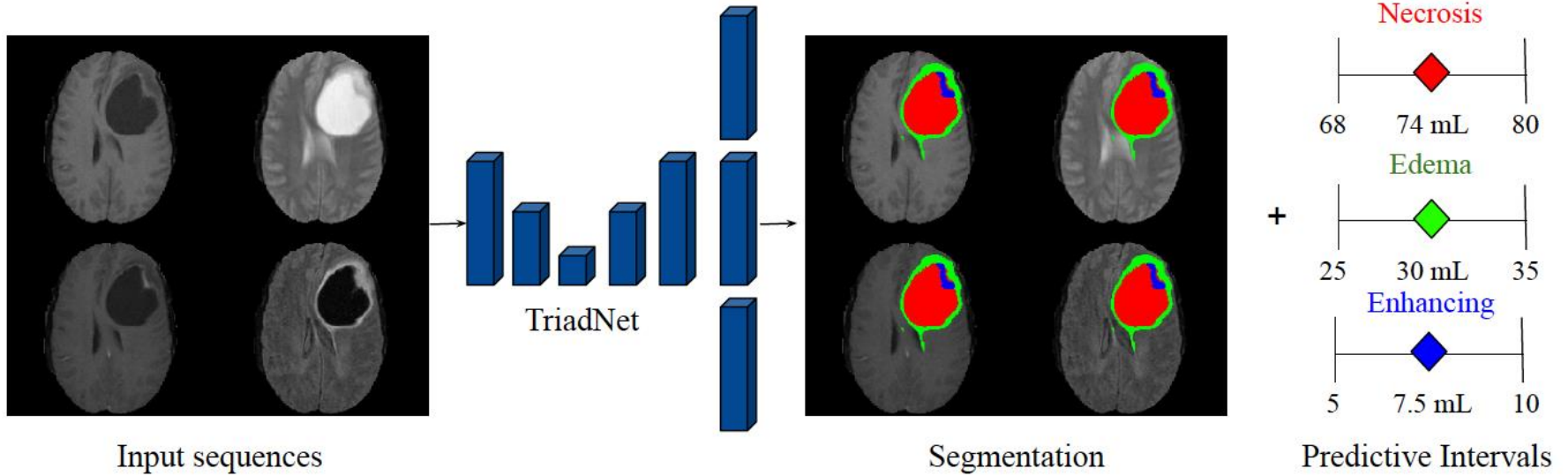
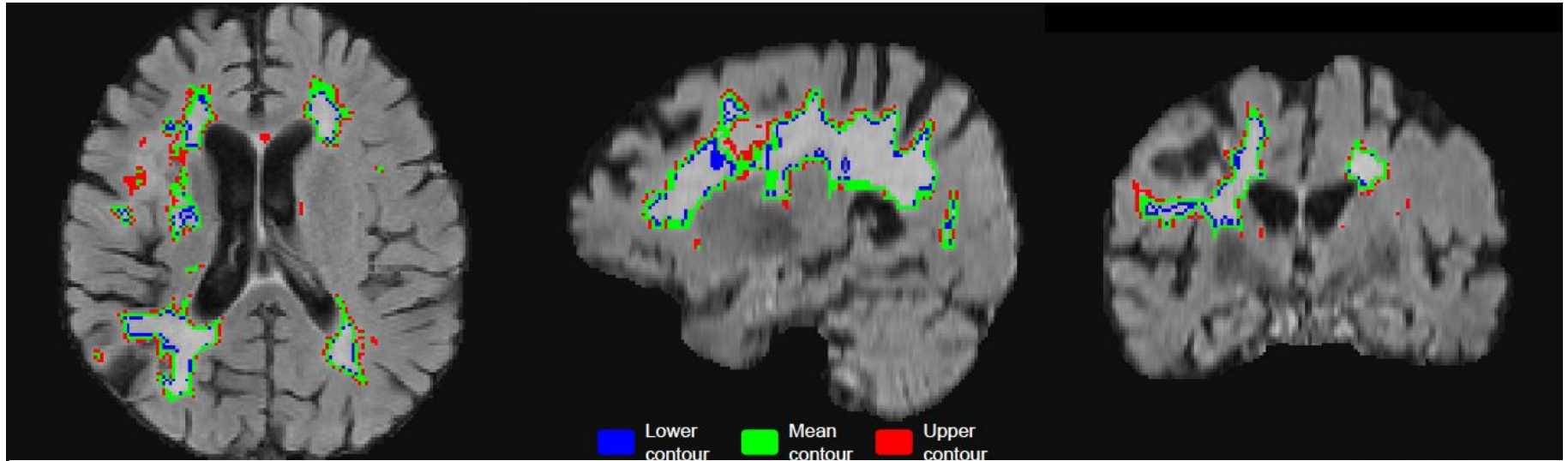
The Tversky loss (Salehi et al. 2017)

$$\text{TriadLoss} = \mathcal{T}_{0.8,0.2}(p_{lower,y}) + \mathcal{T}_{0.5,0.5}(p_{mean,y}) + \mathcal{T}_{0.2,0.8}(p_{upper,y})$$



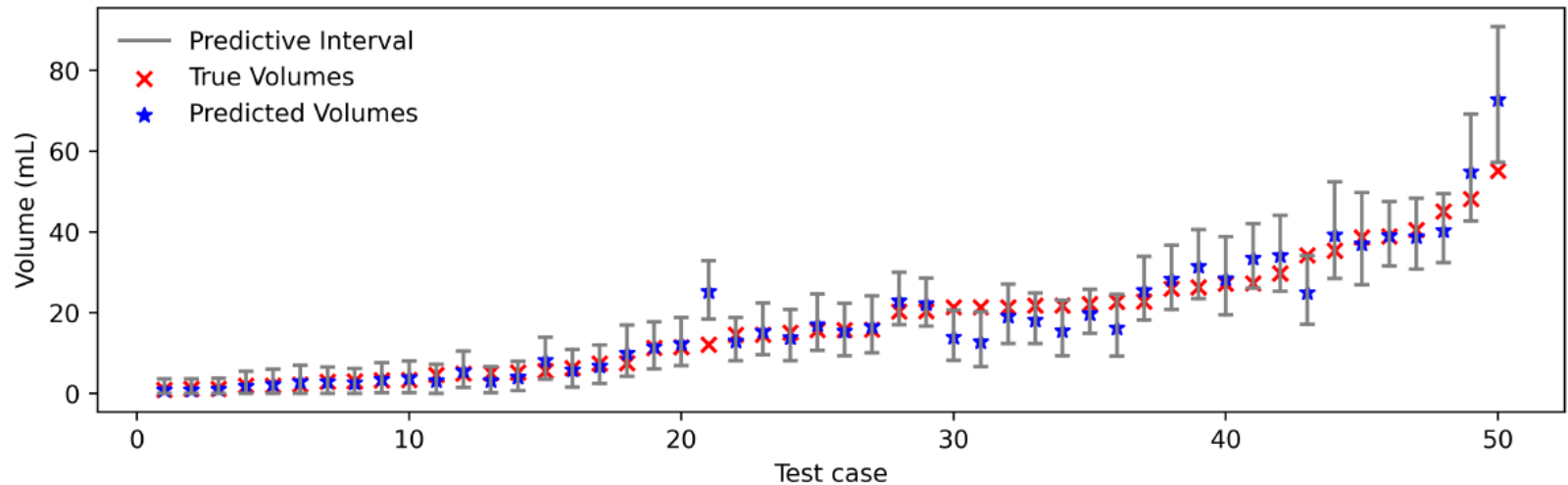
B. Lambert et al. (2023). "TriadNet: Sampling-Free Predictive Intervals for Lesional Volume in 3D Brain MR Images". In: UNSURE 2023, LNCS 14291, pp. 32–41

The TriadNet approach

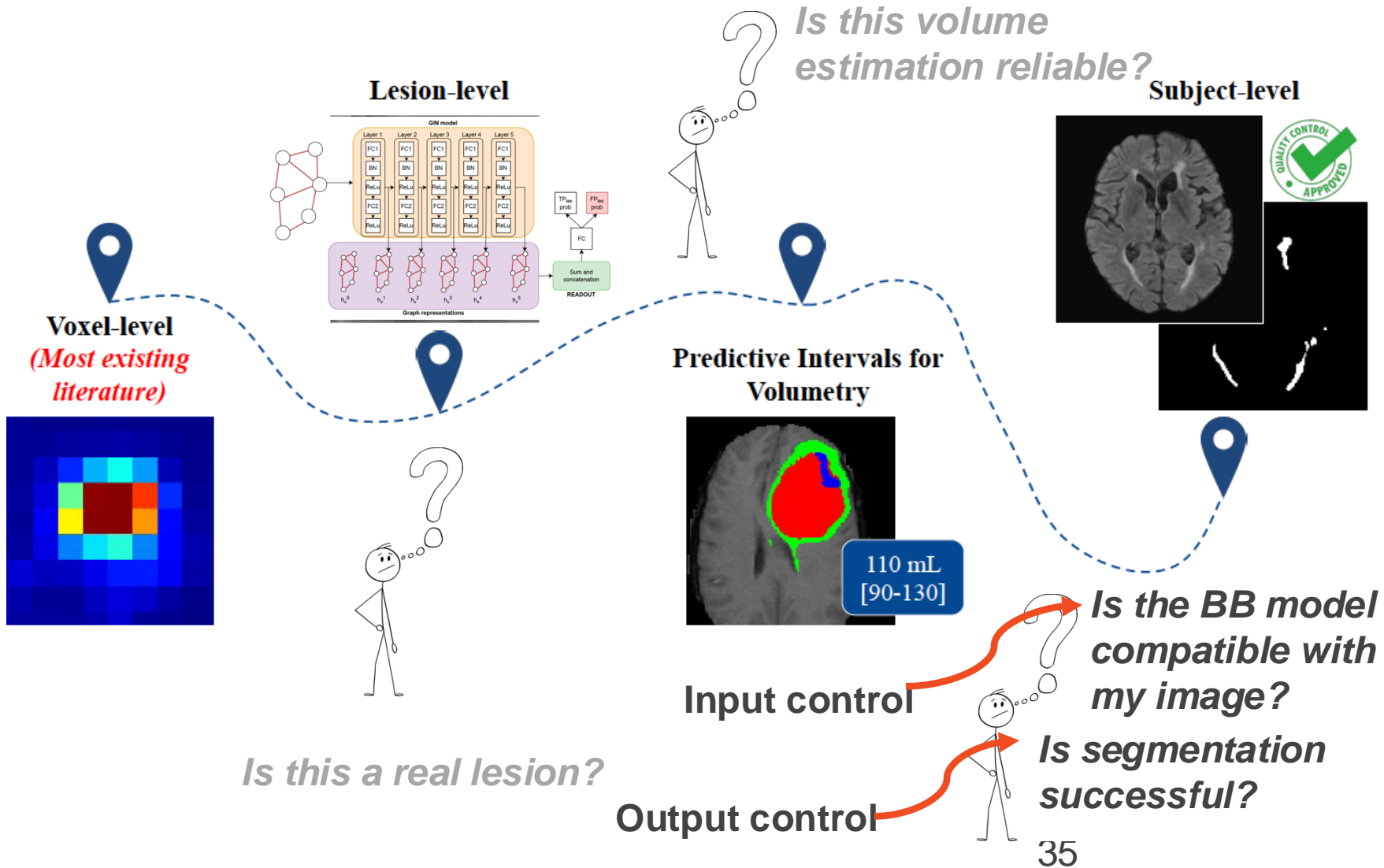


Need for calibration

- 120 subjects for training, 40 for calibration and 50 for in-distribution testing. (Multicentric - 3 Tesla: MSSEG 2016 / WMH 2017 / ISBI 2015)
- Intervals calibrated for a target coverage of 90%.
- Metrics (bootstrapping, $M = 15000$):
 - Mean Average Error: 3.08 ± 0.46 mL
 - Empirical Coverage: $92.06 \pm 5.34\%$

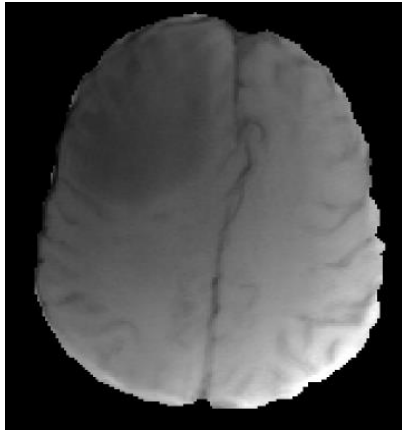


The multi-dimensional aspect of uncertainty

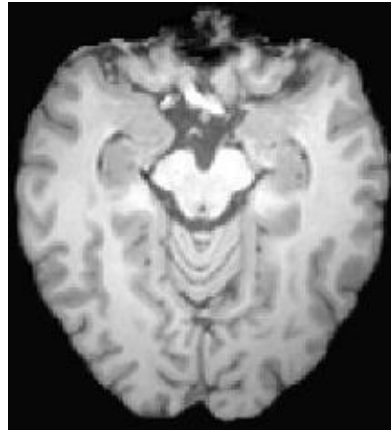


Answer to anything

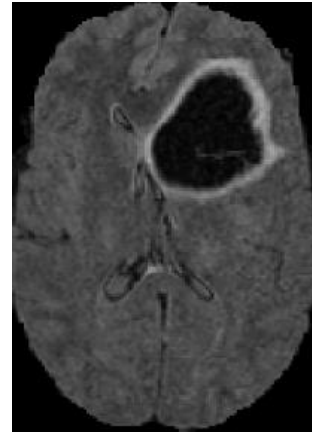
DLL trained for Glioblastoma detection



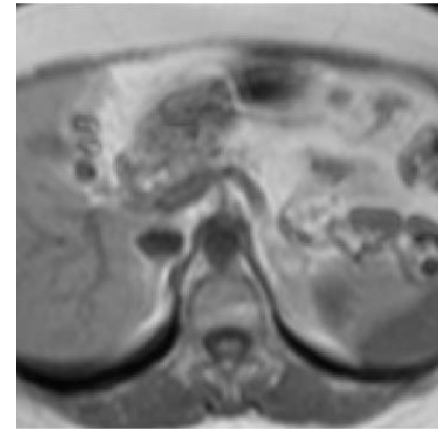
Artefacted T1w



Healthy subject



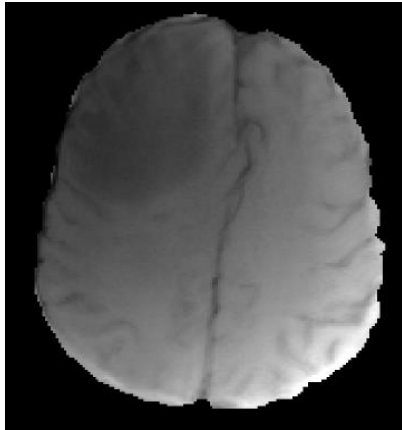
FLAIR



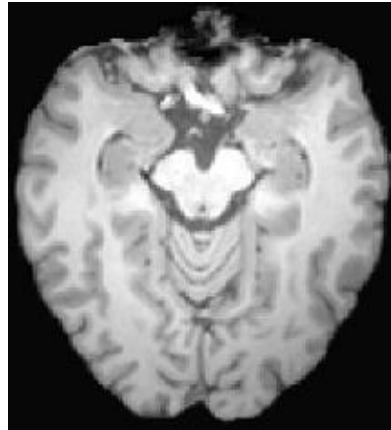
Abdominal T1w

Know-it-all

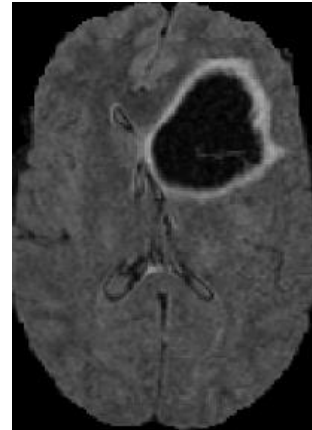
DLL trained for Glioblastoma segmentation on T1w



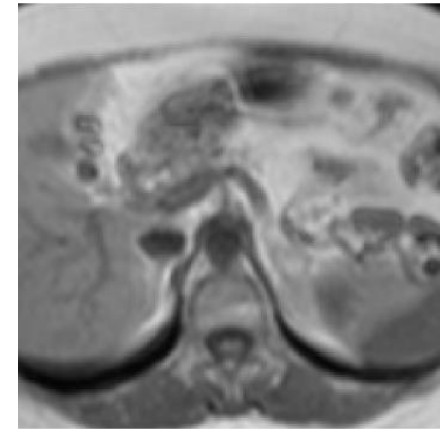
Artefacted T1w



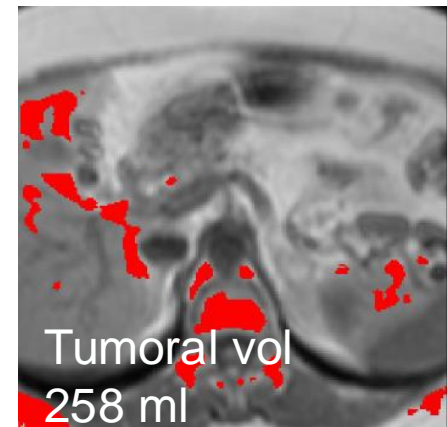
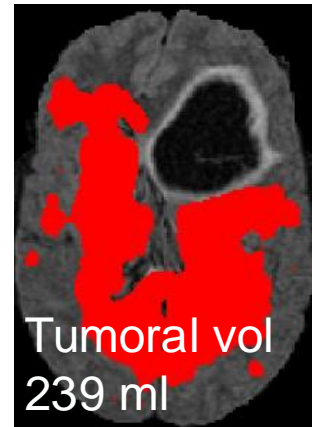
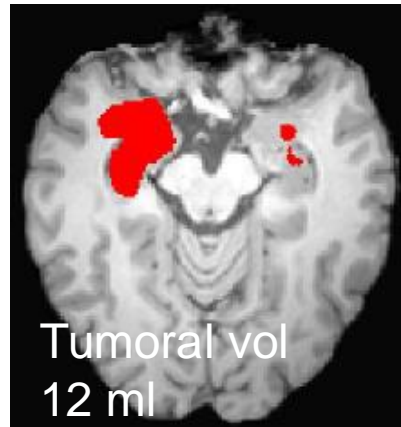
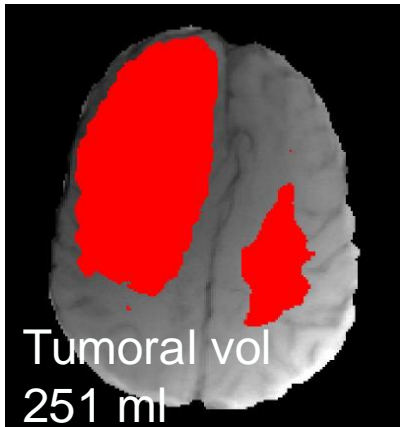
Healthy subject



FLAIR

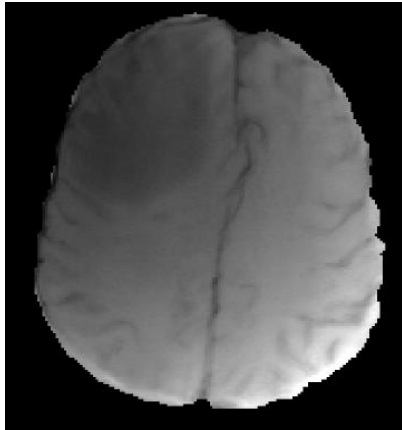


Abdominal T1w

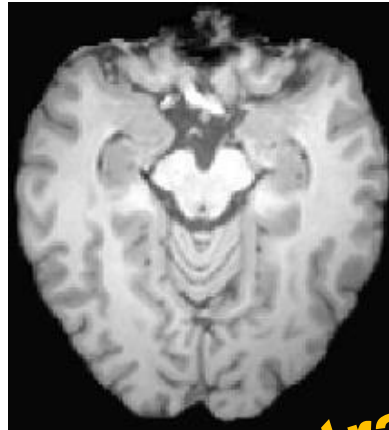


Input Quality Control

DLL trained for Glioblastoma segmentation on T1w



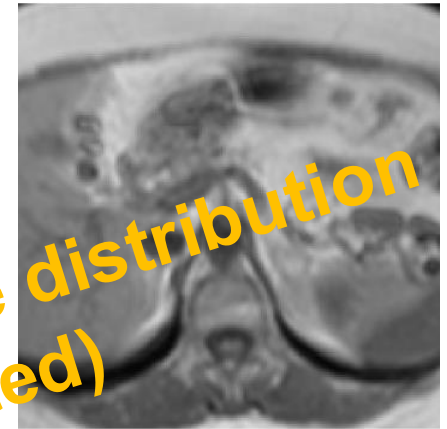
Artefacted T1w



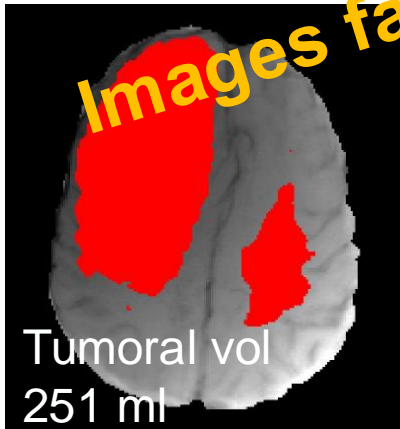
Healthy subject



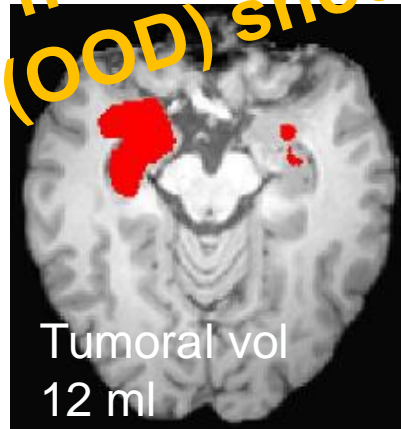
T1w



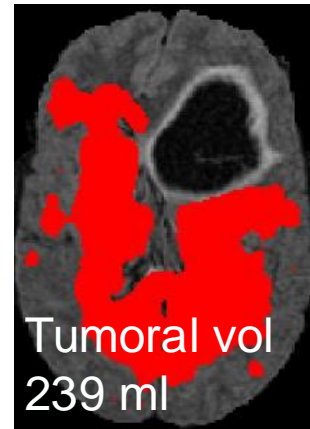
Abdominal T1w



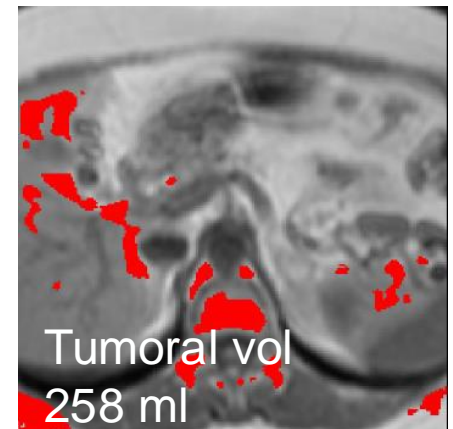
Tumoral vol
251 ml



Tumoral vol
12 ml



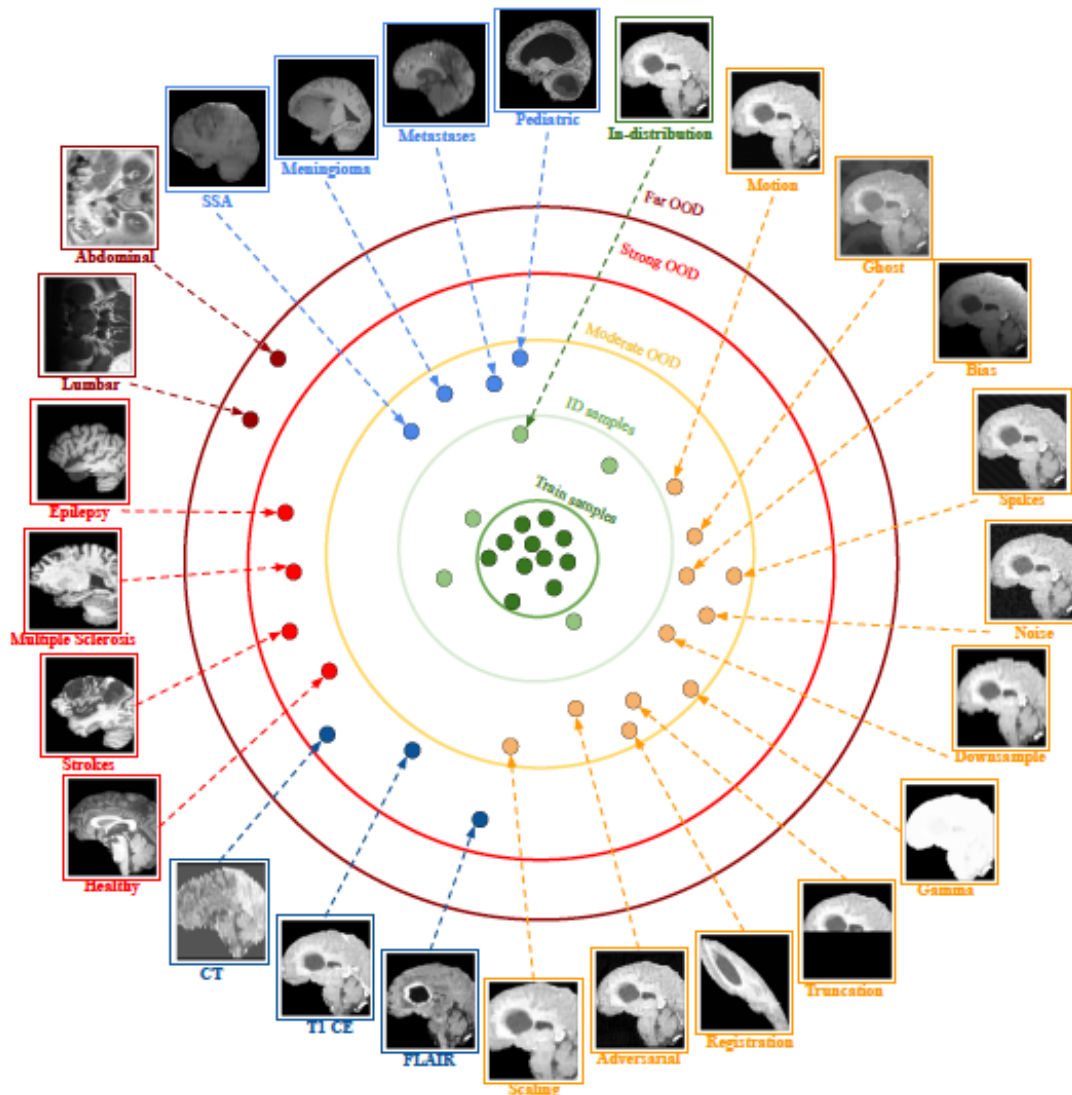
Tumoral vol
239 ml



Tumoral vol
258 ml

Images far from the training image distribution (OOD) should be detected

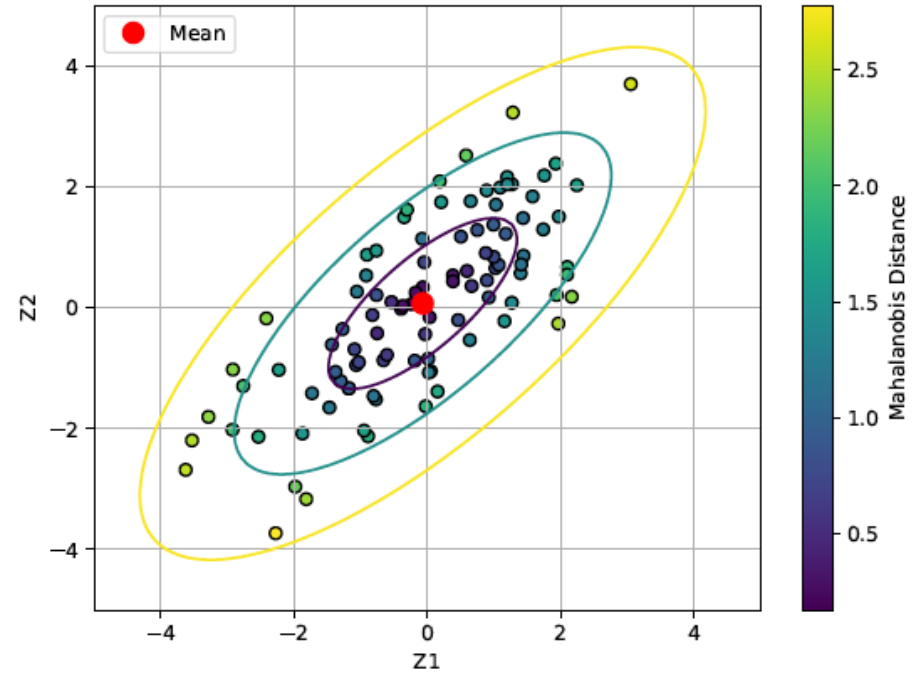
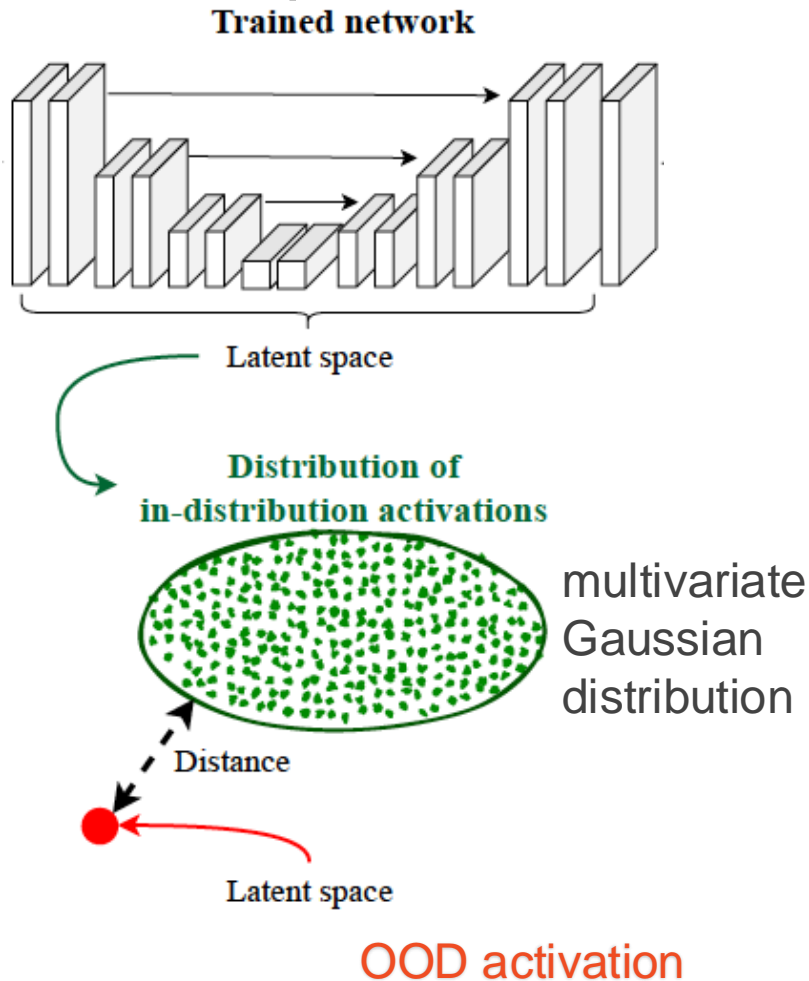
Why an image is OOD?



- In-distribution \leftrightarrow training distribution (T1 MRI of Adult glioblastoma patients)
- Are out-of-distribution:
 - Images corrupted with artifacts.
 - Shifts in the imaged population.
 - Shifts in image modality.
 - Diseases not present in the training set.
 - Incorrect organs.

Input Quality Control

The latent-space distance



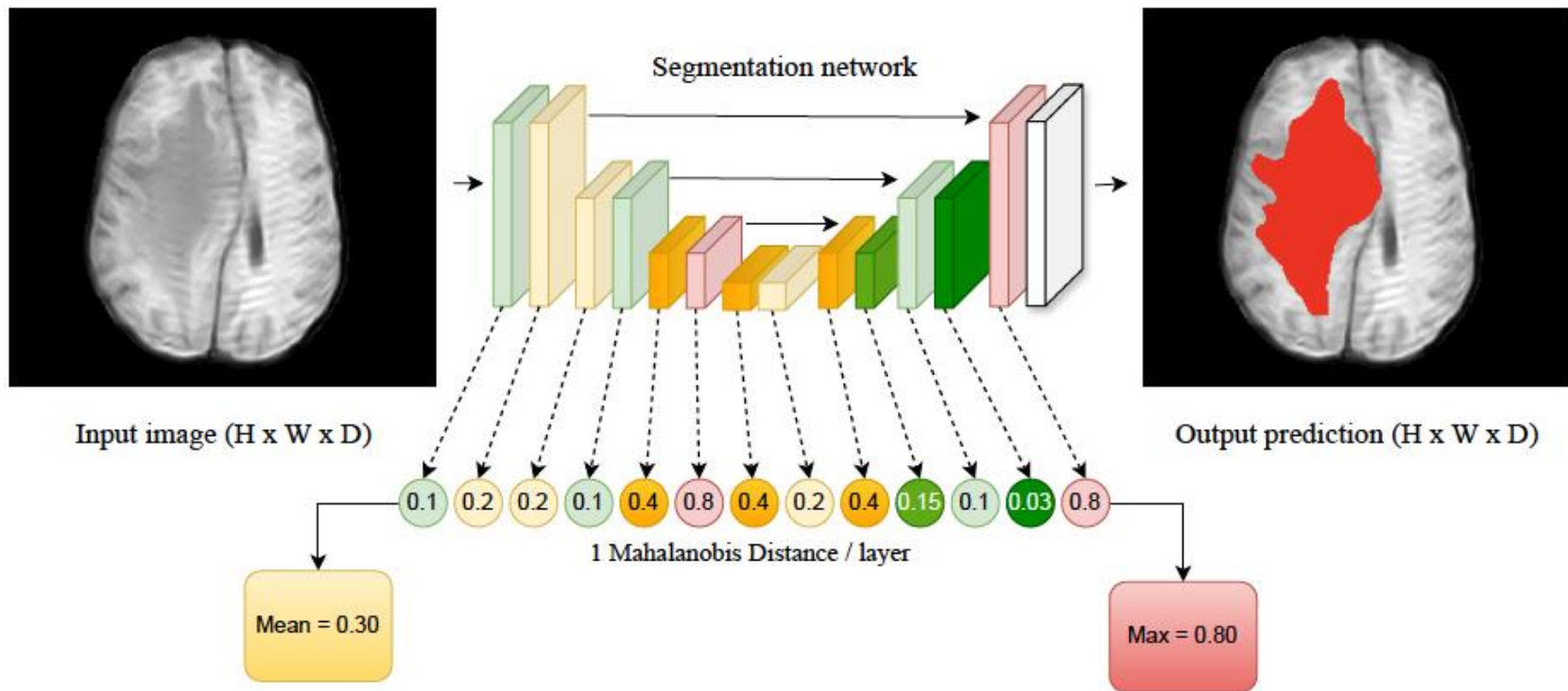
Computed with Mahalanobis distance (MD)

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\Sigma = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)(x_i - \mu)^T$$

$$MD(x_{test}; \mu, \Sigma) = (x_{test} - \mu)^T \Sigma^{-1} (x_{test} - \mu)$$

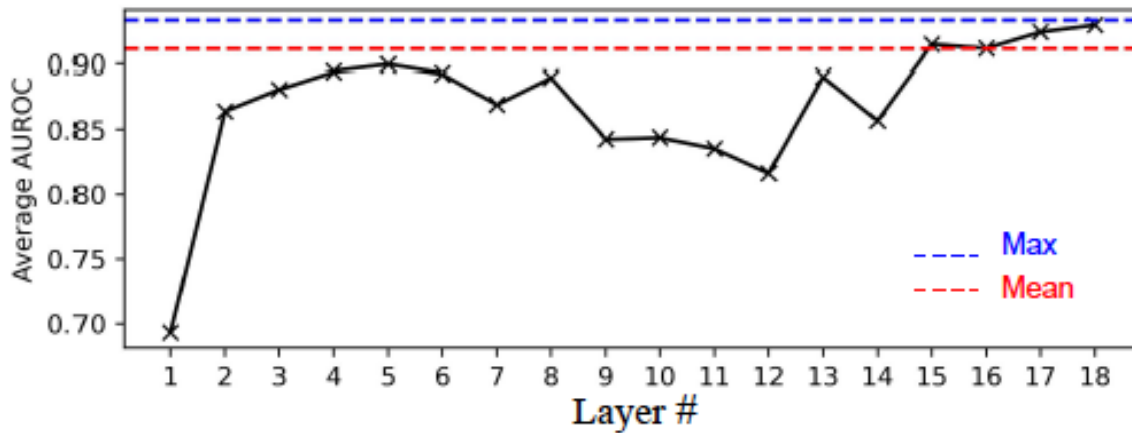
A multi-layer aggregation of MD



[B. Lambert et al. (2023). "Multi-layer Aggregation as a key to feature-based OOD detection". In: UNSURE 2023, Held in Conjunction with MICCAI 2023. LNCS 14291, pp. 104–114]

Brats: 876 subjects for training,
30 for validation,
227 for in-distribution testing

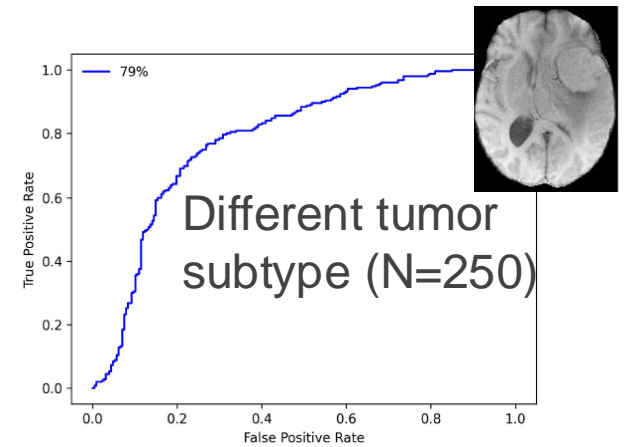
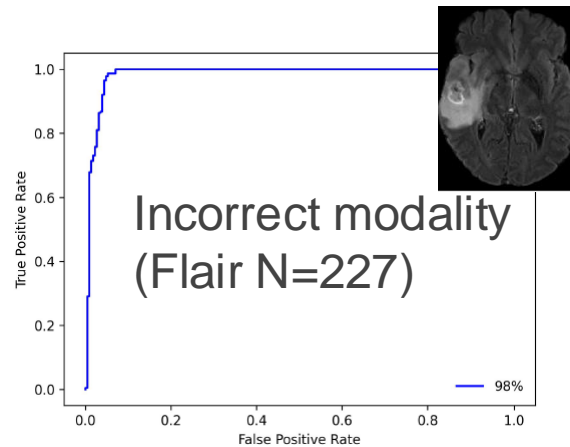
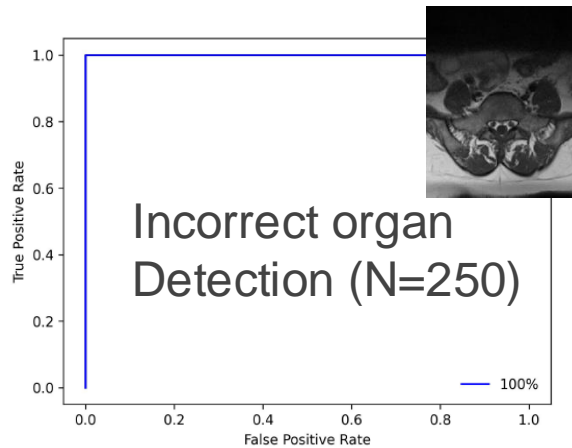
Input Quality Control Dynamic U-Net



The optimal layer for OOD detection depends on the segmentation architecture.

The multi-layer scores (**Mean** and **Max**) provides high detection accuracy.

Input Quality Control

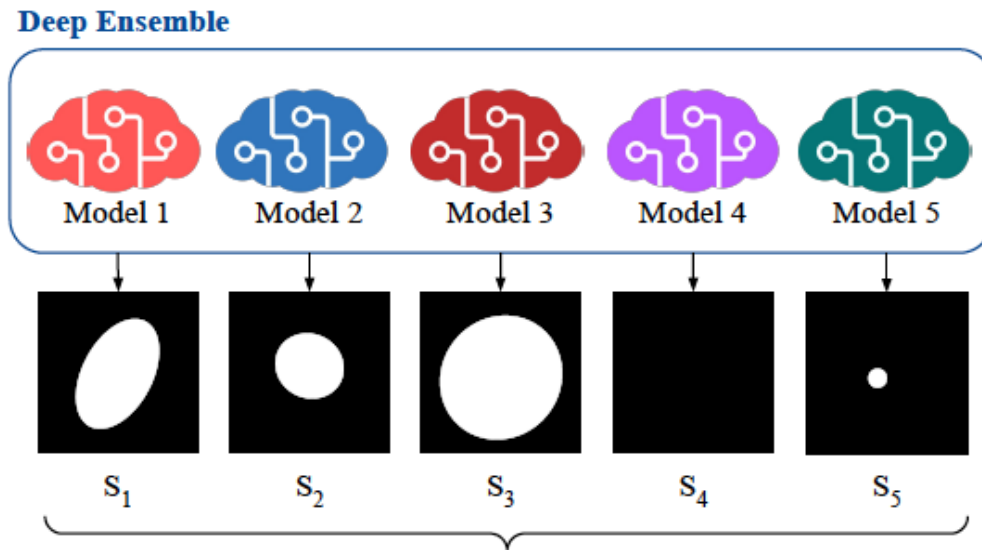


Latent-space distances efficient in detecting images far from the training distribution.

Output Quality Control

Goal: estimate the true segmentation accuracy.

How: Measure the segmentation variability among models.

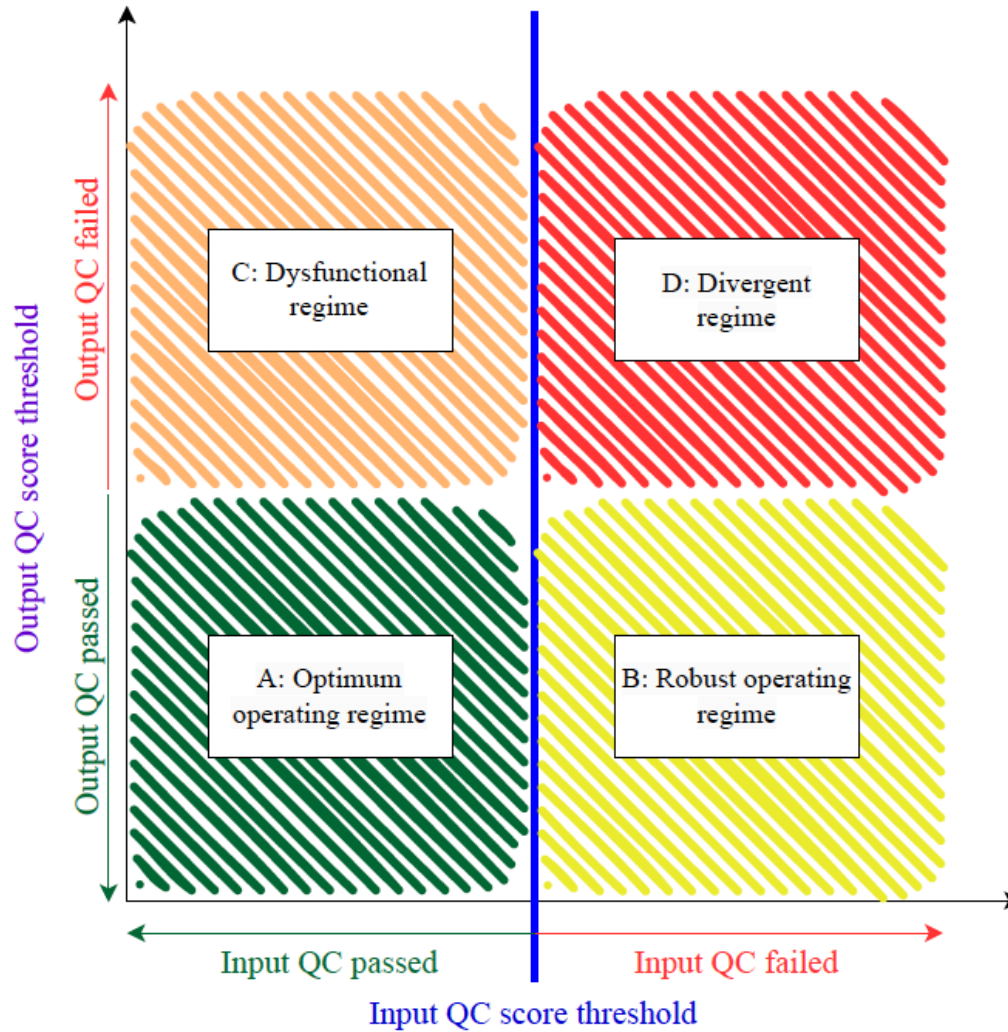


We note S_k the individual segmentations and MV the majority vote segmentation

Ensemble Prediction Agreement (EPA):
$$EPA = \frac{1}{K} \sum_{i=1}^K \text{Dice}(S_k, MV)$$

B. Lambert et al. (2024) “From Out-of-distribution detection to Quality Control”.
In: Trustworthy AI in Medical Imaging, MICCAI book series

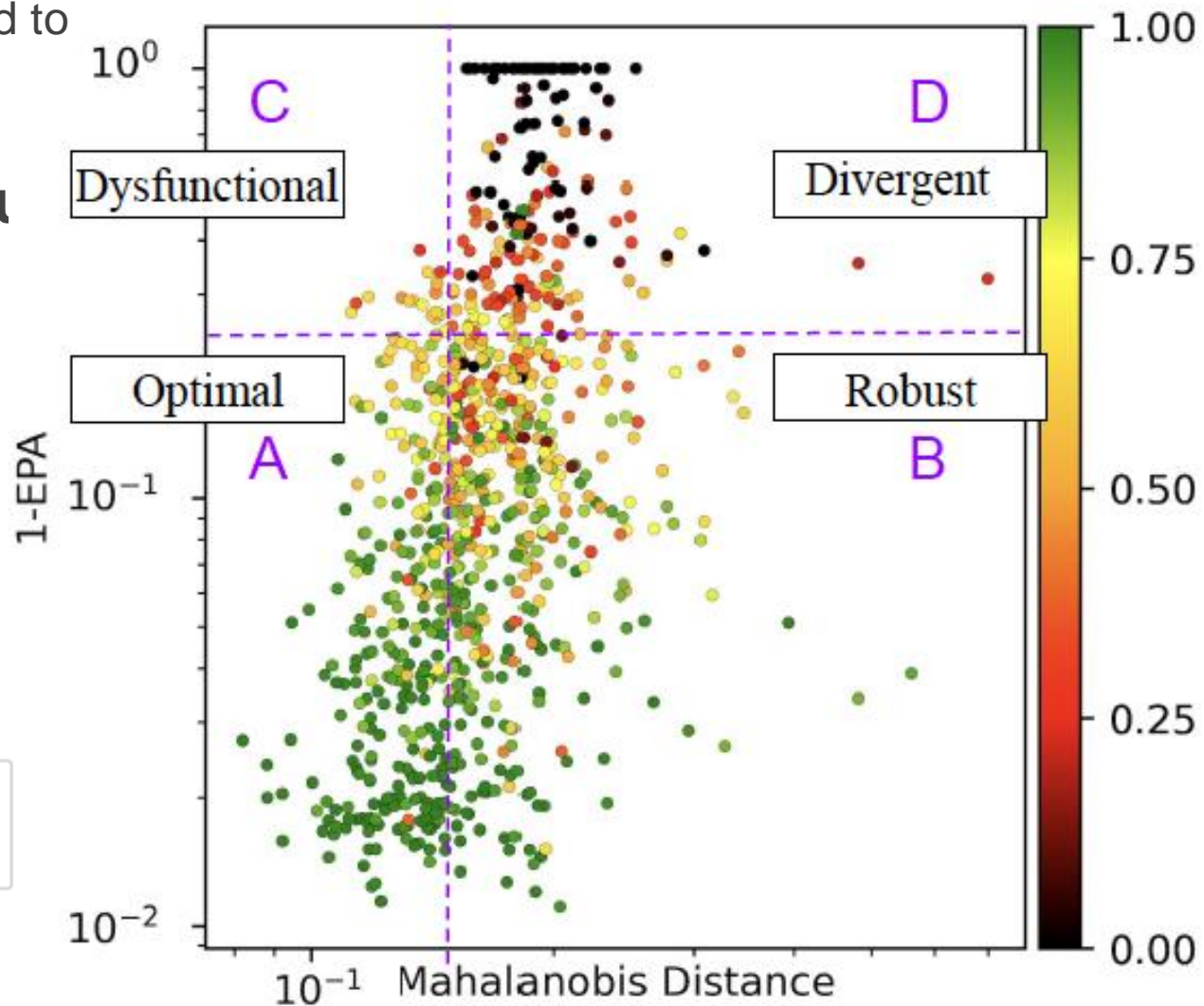
Unified Input-Output control



5 Dynamic U-Nets are trained to segment gliomas.
QC scores computed for 874 test subjects with variable difficulty.

Unified Input-Output

thresholds fixed on a validation dataset (N=30).



- | | |
|-------------------|--------------|
| • In Distribution | • Pediatric |
| • Synthetic | • Metastases |
| • Africa | • Meningioma |

- ✓ Confidence is central for AI deployment
- ✓ Uncertainty quantification improves user's confidence
- ✓ Uncertainty is multidimensional

Take home messages

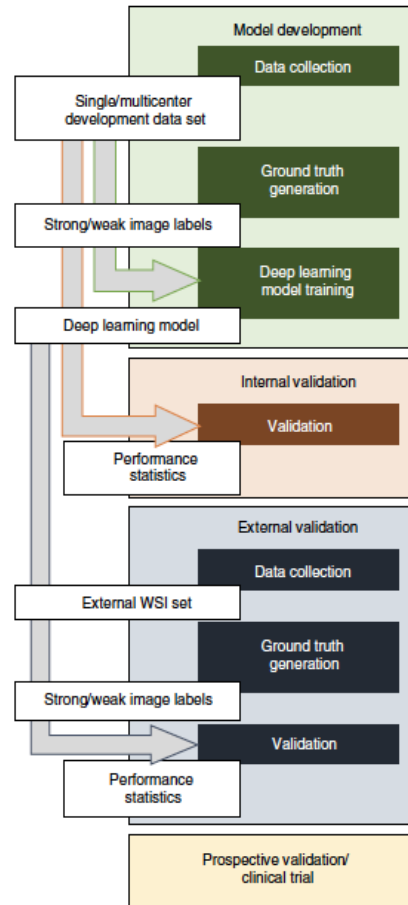
Lesion uncertainty scores
Predictive volume intervals
Unified input & output controls

} uncertainty quantification
for trusted AI



for AI penetration in clinical
routine

Validation process



[van des Laak et al Nat Med 2021]

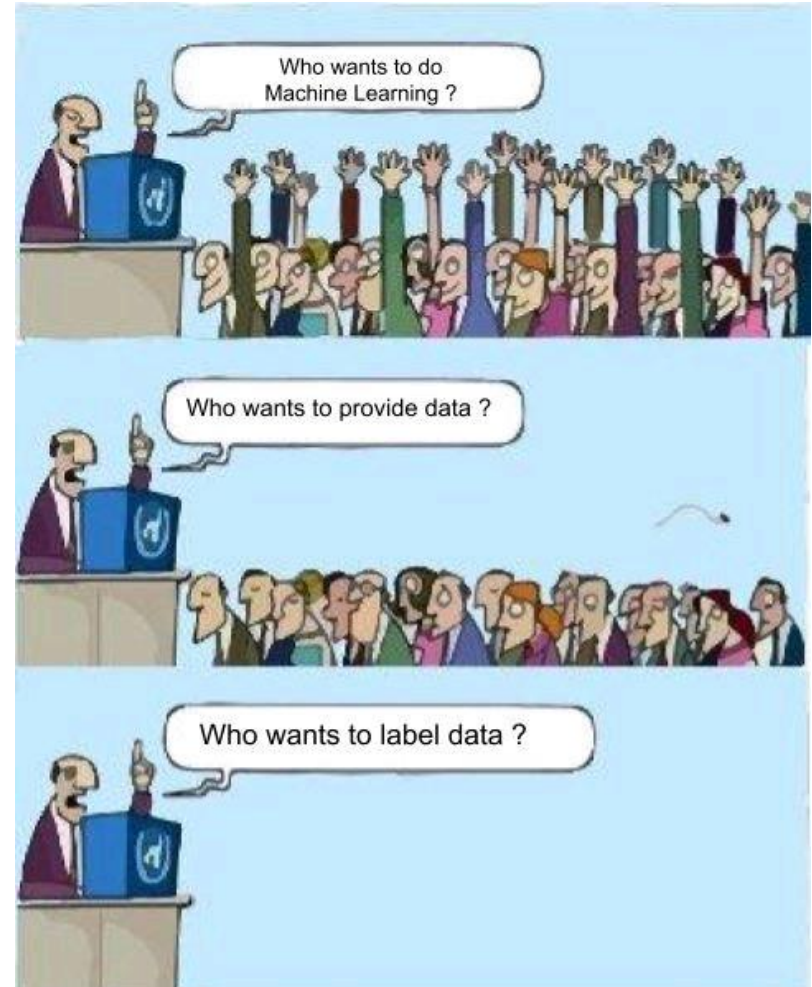
Conclusion: AI for Medical Imaging

- Pros:

- Excellent performances
- Automatic feature learning
- Knowledge emergence
- On the shelves tools
- Discharge Expert
- Automatic Quantification

- Cons:

- Importance of Image Quality
- Annotation
- Data hungry
- Computational cost
- Black box / trustability
- Specific to one problem
- Adversarial attack
- Catastrophic forgetting
- Ethic, social and law
- Needs for specific tools & infra



Future

- Insert *a priori* knowledge
- Improve training: interactive, transfert, unsupervised, self-supervised
- Transfert learning
- Appropriate tools /infra for specific applications (e.g. biomedical)
- Clinical Trial for Testing
 - the genericity
 - the robustness to noise (e.g. multicenter studies)
 - Preprocessing / Image quality influence
 - User acceptance
- Evolving models => regulation & validation

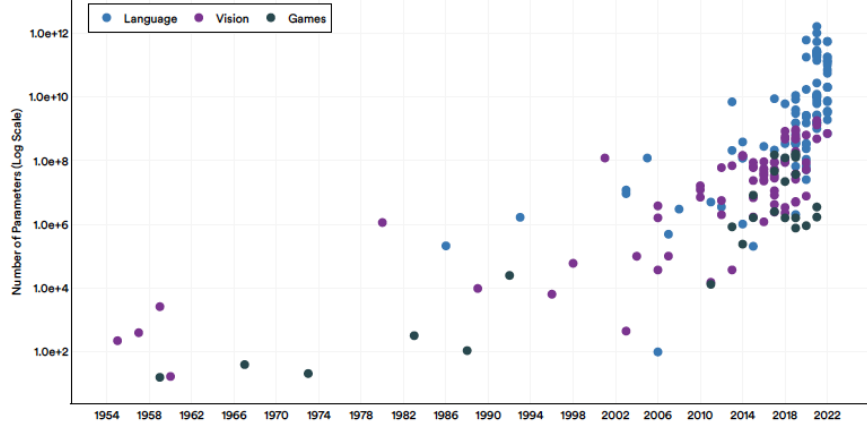
Some key points

- Industry races ahead academia
 - Mass of data, computer power and money
- AI is both helping and harming the environment
- The world's best new scientist ...AI?
 - Generate new antibodies, drugs, ...
- The number of incidents with the misuse of AI rises
- The demand for AI-related skills increases in all sectors.

More power ...

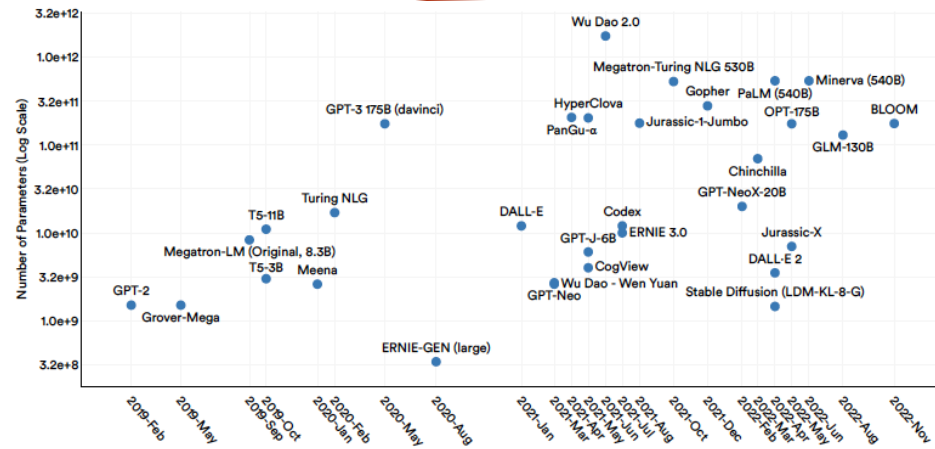
Number of Parameters of Significant Machine Learning Systems by Domain, 1950–2022

Source: Epoch, 2022 | Chart: 2023 AI Index Report



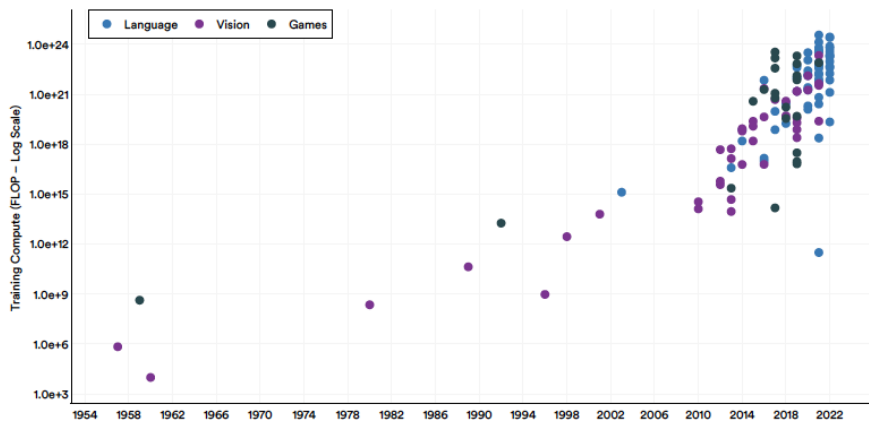
Number of Parameters of Select Large Language and Multimodal Models, 2019–22

Source: Epoch, 2022 | Chart: 2023 AI Index Report



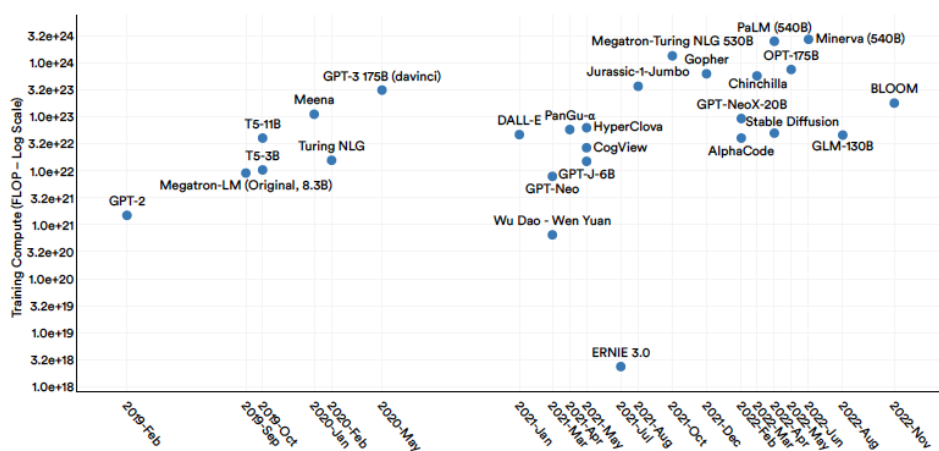
Training Compute (FLOP) of Significant Machine Learning Systems by Domain, 1950–2022

Source: Epoch, 2022 | Chart: 2023 AI Index Report



Training Compute (FLOP) of Select Large Language and Multimodal Models, 2019–22

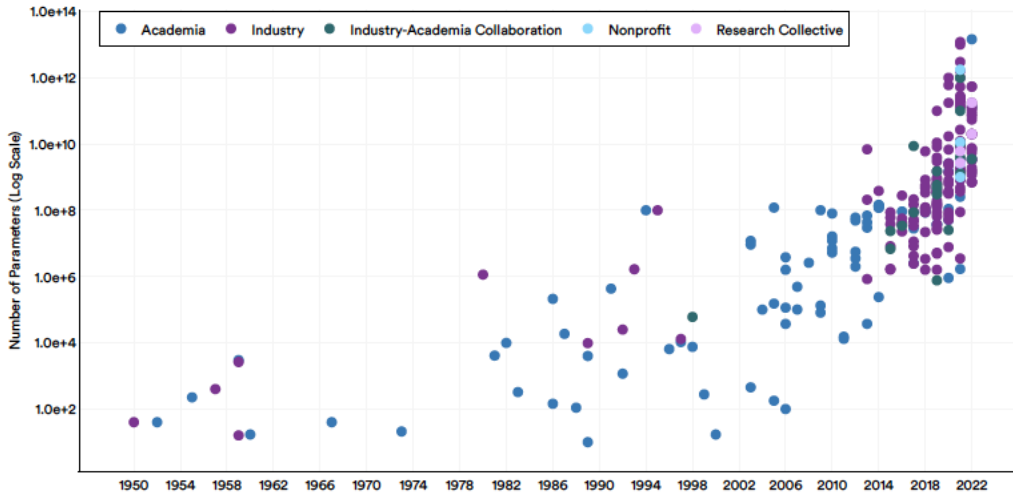
Source: Epoch, 2022 | Chart: 2023 AI Index Report



More industrials ...

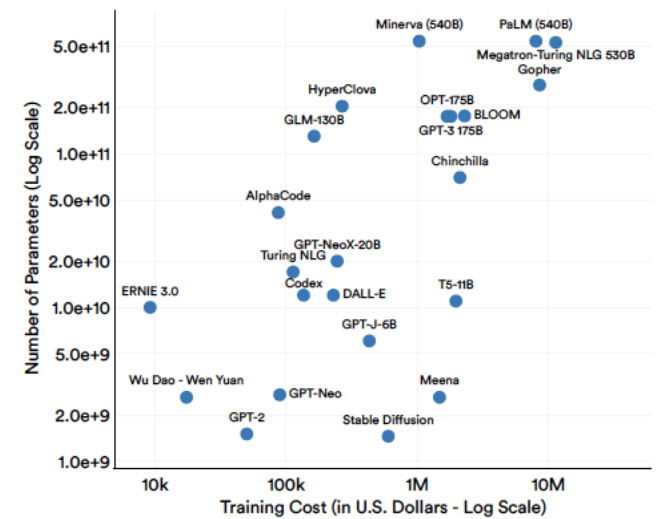
Number of Parameters of Significant Machine Learning Systems by Sector, 1950–2022

Source: Epoch, 2022 | Chart: 2023 AI Index Report



Estimated Training Cost of Select Large Language and Multimodal Models and Number of Parameters

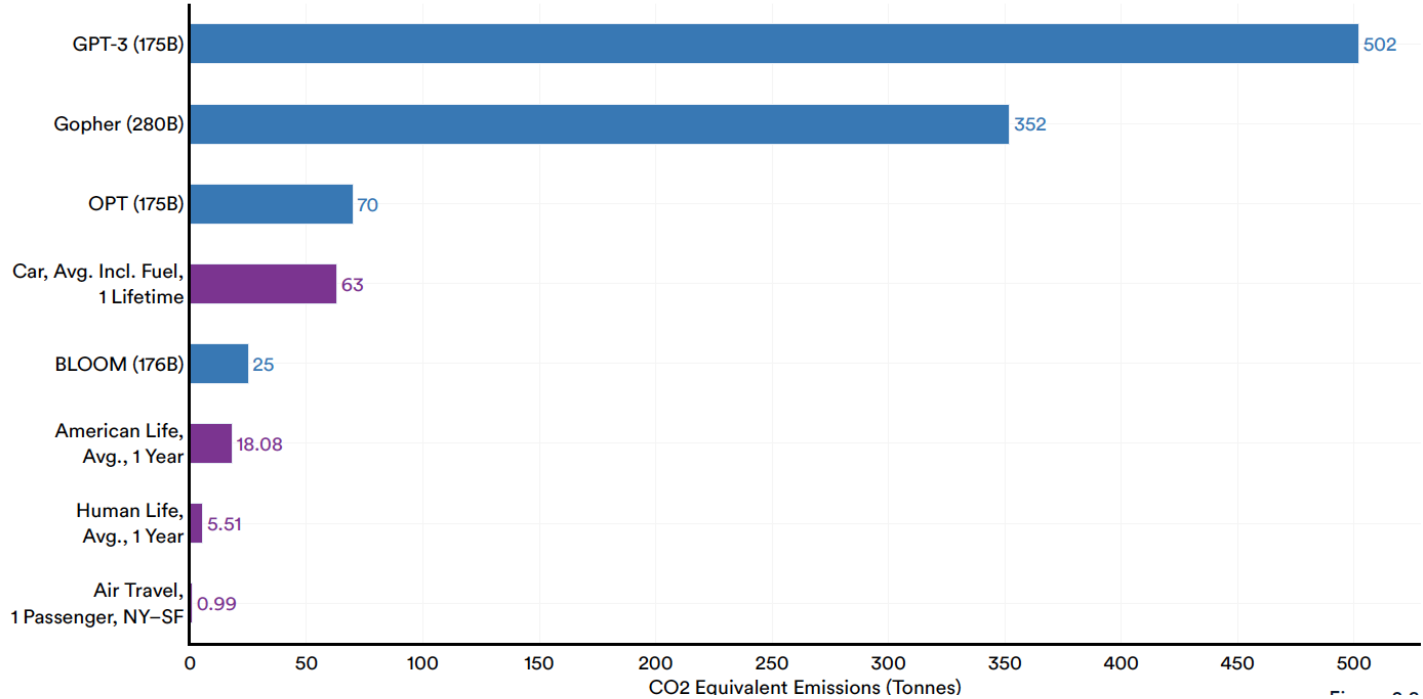
Source: AI Index, 2022 | Chart: 2023 AI Index Report



Environmental impact

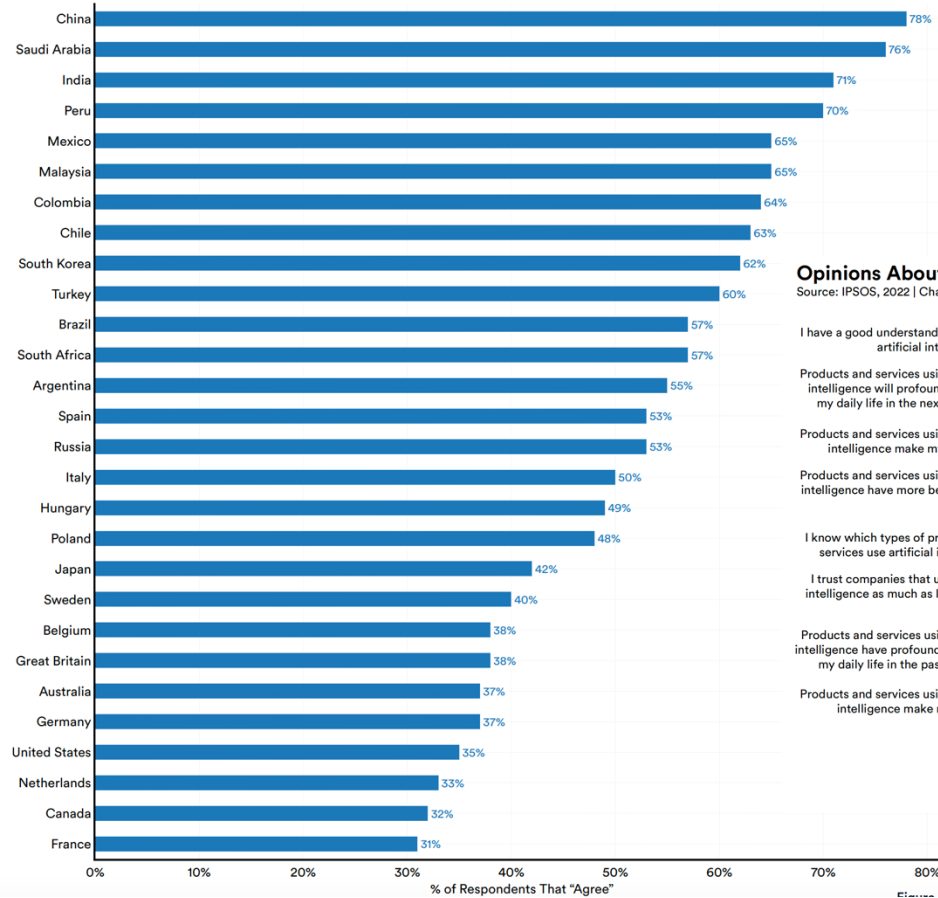
CO2 Equivalent Emissions (Tonnes) by Selected Machine Learning Models and Real Life Examples, 2022

Source: Luccioni et al., 2022; Strubell et al., 2019 | Chart: 2023 AI Index Report



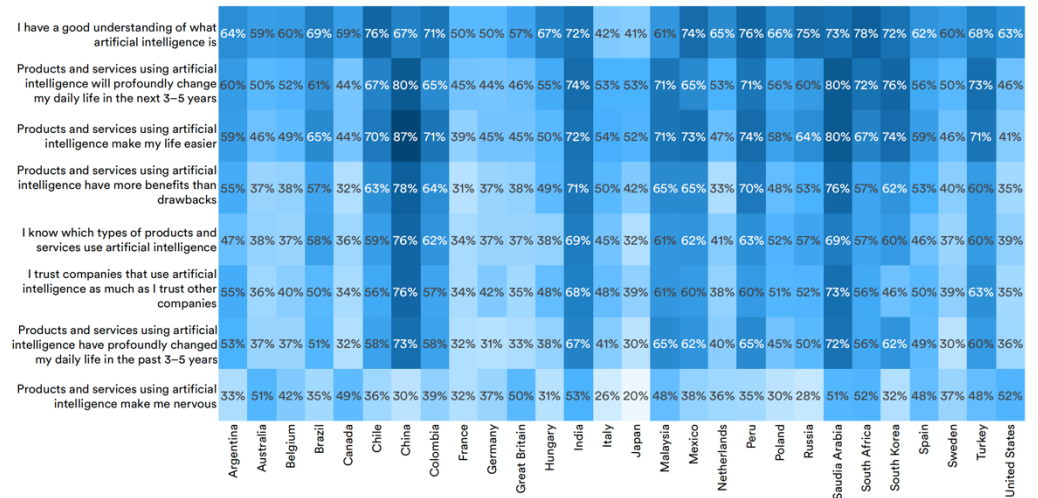
'Products and services using AI have more benefits than drawbacks,' by Country (% of Total), 2022

Source: IPSOS, 2022 | Chart: 2023 AI Index Report



Opinions About AI by Country (% Agreeing With Statement), 2022

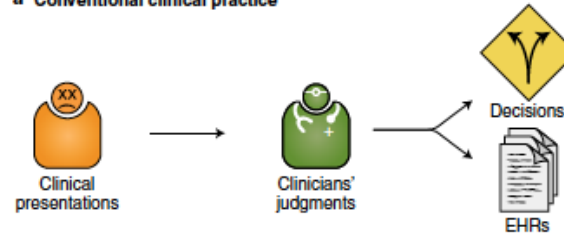
Source: IPSOS, 2022 | Chart: 2023 AI Index Report



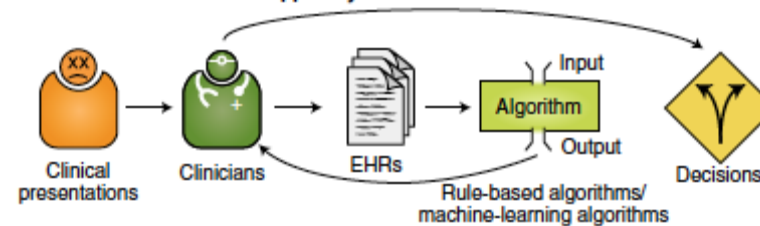
Different models of Information Flow

Yu Nat Bio Eng 2018]

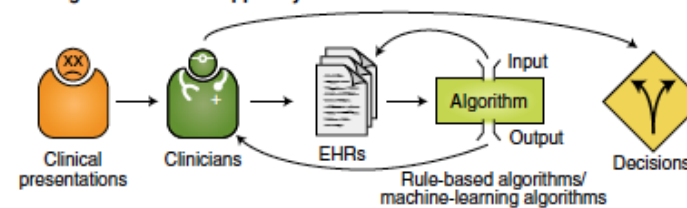
a Conventional clinical practice



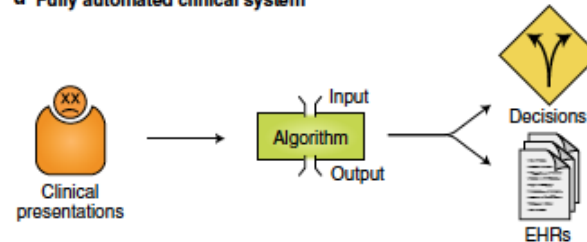
b Conventional decision support systems



c Integrative decision support systems



d Fully automated clinical system



Responsability

- Principes d'une approche responsable à l'appui d'une IA digne de confiance :
 - i) croissance inclusive, développement durable et bien être
 - ii) valeurs centrées sur l'humain et équité
 - iii) transparence et explicabilité
 - iv) robustesse, sûreté et sécurité
 - v) responsabilité.

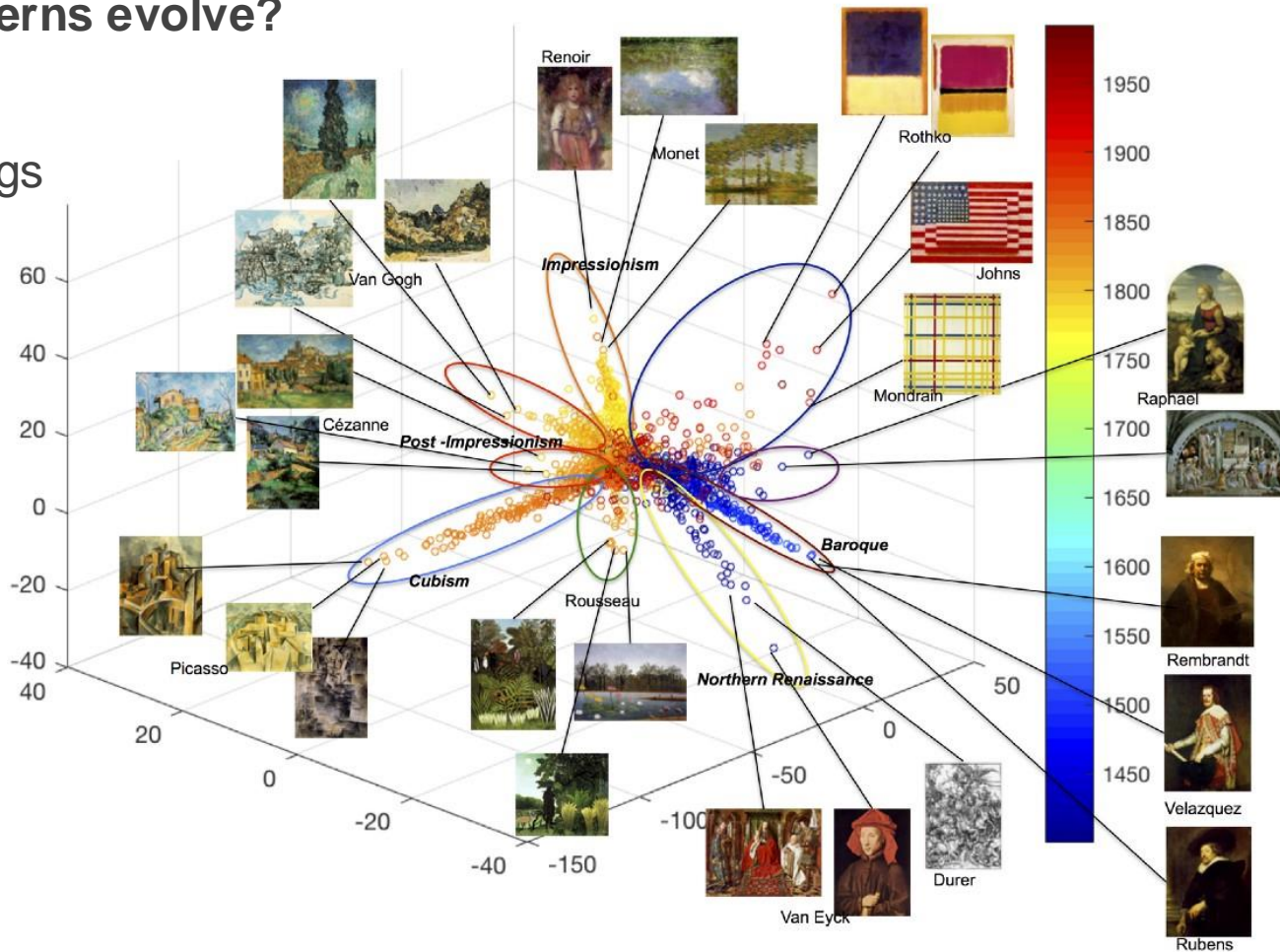
Elle appelle en outre les acteurs de l'IA à promouvoir et mettre en oeuvre ces cinq principes, selon leurs rôles respectifs.

OCDE 2019

Art History

How characteristics of style are identified?
How the patterns evolve?

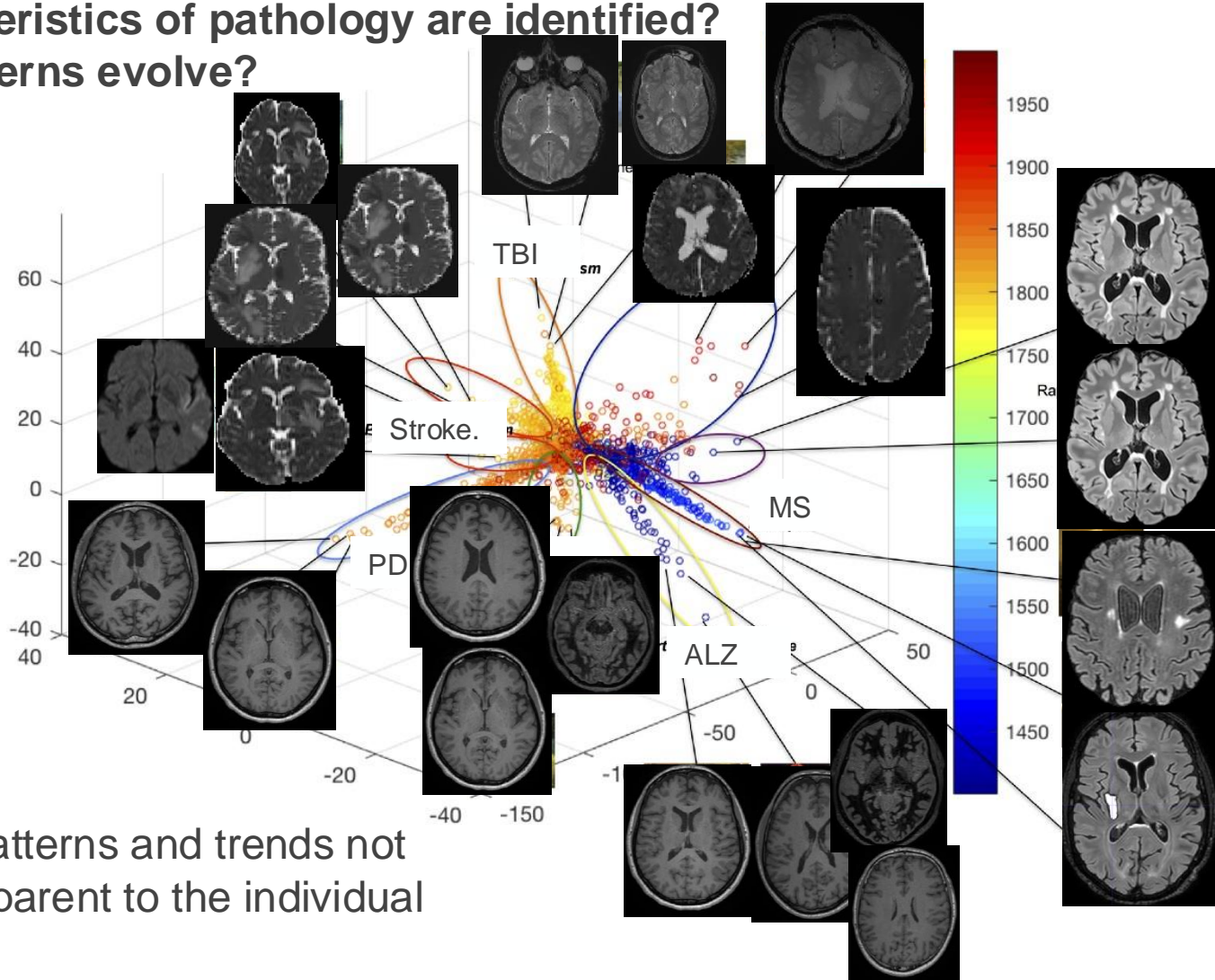
76921 paintings
Train(85%)
Val (9.5%)
Test (5.5%)



Elgammal et al. 2018 arxiv 1801.07729

e-Nosology

How characteristics of pathology are identified?
How the patterns evolve?



« to discover
fundamental patterns and trends not
necessarily apparent to the individual
human eye »

From Internist-I/QMR to Med-PaLM-I

Miller RA, Pople HE Jr, Myers JD.
INTERNIST-1, An Experimental Computer-
based Diagnostic Consultant for General
Internal Medicine. N Engl J Med
1982;307:468-76.

Knowledge-based
written in LISP

```
IMMUNOELECTROPHORESIS SERUM IGA INCREASED, IMMUNOELECTROPHORESIS SERUM IGG IN-
CREASED, IMMUNOELECTROPHORESIS SERUM IGM INCREASED, LDH BLOOD INCREASED, LE TEST
POSITIVE, LIVER RADIOISOTOPE SCAN IRREGULAR UPTAKE, PROTHROMBIN TIME INCREASED, RHEU-
MATOID FACTOR POSITIVE
DISCRIMINATE: HEPATITIS CHRONIC ACTIVE, BILIARY CIRRHOSIS PRIMARY
[Comment: At this point, INTERNIST-1 has constructed a master differential- diagnosis list, ordered its members
with a scoring algorithm, and then focused attention on the most promising problem area, which appropriately con-
tains two liver diseases. The DISREGARDING list consists of all findings that are inconsistent with the topmost (first-
listed) diagnosis; the CONSIDERING list includes all findings explained by the topmost diagnosis (i.e., chronic
active hepatitis). The number of plausible contenders in the problem area determines the strategy for questioning.
Questions are asked to discriminate between the two diagnoses.]
Please Enter Findings of LIPID(S) BLOOD
GO
CHOLESTEROL BLOOD DECREASED?
N/A
[Comment: INTERNIST-1 asks questions in two forms. A question of the style "Please Enter Findings of ..." asks
about a generic class of findings, such as serum lipids or chest x-ray films. The user enters specific (positive or
negative) findings and when finished types "GO." If there are no positive or negative find- ings to enter, typing "GO"
will cause the programs to ask the original single item that prompted the generic question. For example, an elevated
cholesterol level would favor primary biliary cirrhosis over chronic active hepatitis, whereas the converse situation
would apply if the cholesterol were decreased. Thus, INTERN- IST-1 asked a generic question about serum lipids. The
second form of question- ing is direct. INTERNIST-1 simply prints a manifestation name followed by a question
mark. The user can then respond by typing "YES" (present), "NO" (absent), or "N/A" (not available-either unknown
or not performed). The published CPC did not mention the serum cholesterol value, so the user entered N/A.]
Please Enter Findings of INSPECTION AND PALPATION SKIN
GO
SKIN XANTHOMATA?
NO
Please Enter Findings of BLOOD COUNT AND SMEAR
GO
RBC TARGET(S)
NO
[Comment: INTERNIST-1 uses information about the type of each manifestation-history, symptoms, signs, and labo-
ratory tests, both routine and two gra- dations of more expensive or invasive laboratory procedures. The types of
manifestations determine which questions are asked. Although the greatest dis- criminating power for resolving a
differential diagnosis often resides in an inva- sive test, such as a liver biopsy, INTERNIST-1 has been programmed to
exhaust the useful history and physical-examination findings before escalating its questioning to more expensive or
dangerous laboratory tests.]
Please Enter Findings of KIDNEY FUNCTION TEST(S)
GO
UREA NITROGEN BLOOD 60 TO 100?
NO
UREA NITROGEN BLOOD GTR THAN 100?
NO
Please Enter Findings of LIVER FUNCTION TEST(S)
GO
UROBILINOGEN URINE ABSENT?
N/A
[Comment: INTERNIST -1 did not seek to investigate renal function itself in the passage above. It was programmed
to recognize that the urinary urobilinogen value may be unreliable in the presence of marked azotemia, so it first asked
about the blood urea nitrogen.]
CONCLUDE: HEPATITIS CHRONIC ACTIVE
[Comment: It is likely that INTERNIST-1 had nearly made the diagnosis of chronic active hepatitis before it asked
```

From Miller IMIA Yearbook of Medical Informatics 2010

From Internist-I/QMR to Med-PaLM-II

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to common consumer search queries about health.

Question: How do you treat skin redness?

Complete Answer: It depends on the cause of the skin redness. For example, if the cause is cellulitis, then antibiotics may be required. However, this might be inappropriate for other causes of redness such as eczema. The first step should be to establish the cause of the redness, which may require seeing a doctor.

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to consumer questions about medications.

Question: Why take simvastatin?

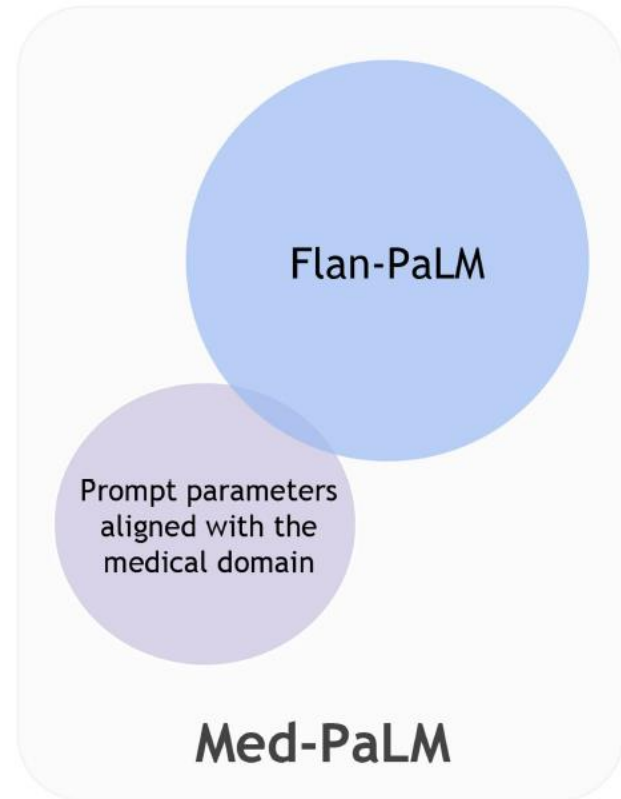
Complete Answer: Simvastatin is a medication that is used to lower cholesterol levels in the blood. Simvastatin is typically used to treat high cholesterol and to prevent heart disease, and to reduce the risk of heart attacks and strokes in those with the greatest risk of these happening (or re-occurring). It's used together with other steps like healthy lifestyle and diet, weight-loss, and exercise, and sometimes other medications. Simvastatin may also be used to decrease the amount of...

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to consumer questions about medications.

Question: How can i split a 3mg or 4mg waefin pill so i can get a 2.5mg pill?

Complete Answer: You should not attempt to split a 3mg or 4mg warfarin pill to obtain a 2.5mg dose. If you are prescribed warfarin in a 2.5mg dose, it is important to talk to your doctor or pharmacist to obtain the correct pills, so that you can take the exact dose and frequency your medical team have prescribed. Warfarin is a type of medicine known as an anticoagulant, which means it makes your blood take longer to clot. The exact amount required to achieve the desired effect can vary.

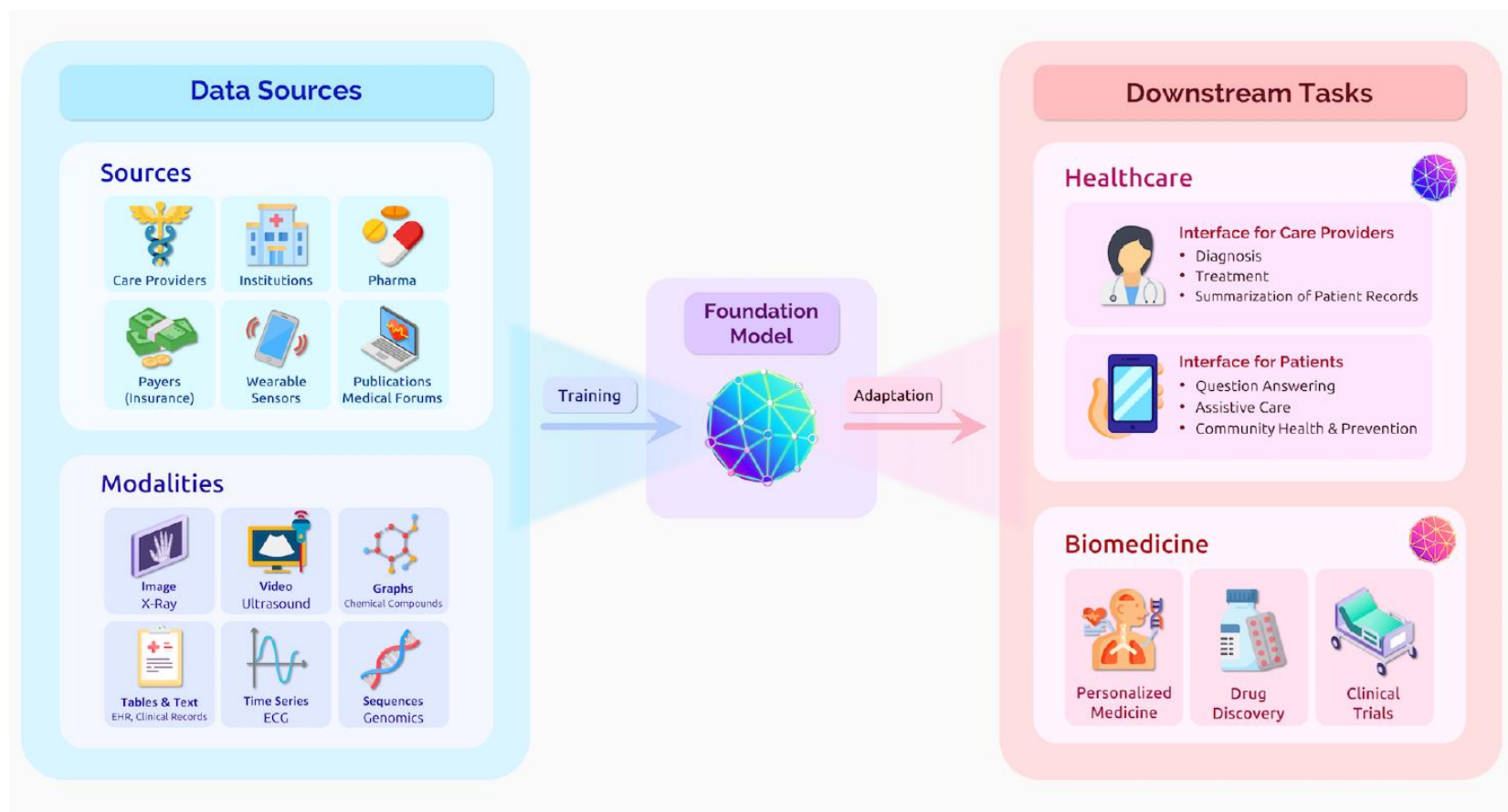
→
Instruction
Prompt Tuning



Large Language Model based

Singhal et al Nat Aug 2023

Foundation Models in Healthcare



Bommasani et al 2022

<https://arxiv.org/abs/2108.07258>

Some references

- Context

- <http://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/>

- Languages

- <https://www.tensorflow.org/>
- <http://torch.ch/>
- <http://scikit-learn.org/>
- <http://caffe.berkeleyvision.org/>
-

- Courses

- Karpathy <http://cs231n.github.io/convolutional-networks/>
- Collège de France : Y Le Cun (2015) et S. Maillard (2019 ...)
- Ng A <https://www.coursera.org/learn/machine-learning>
- Nielson M. NN & ML <http://neuralnetworksanddeeplearning.com/>
-