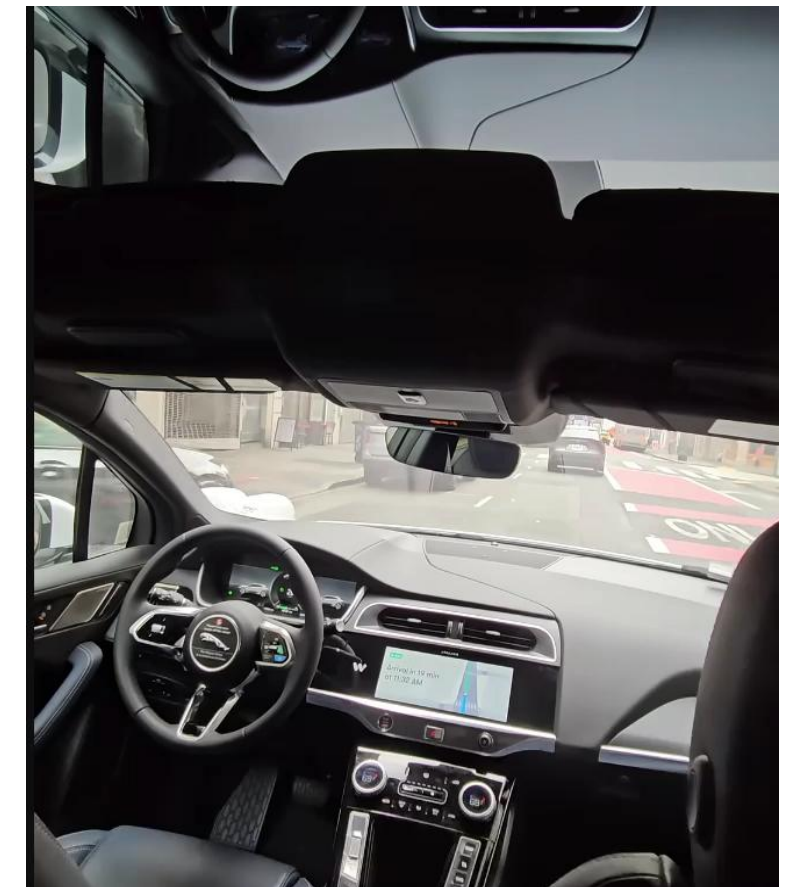
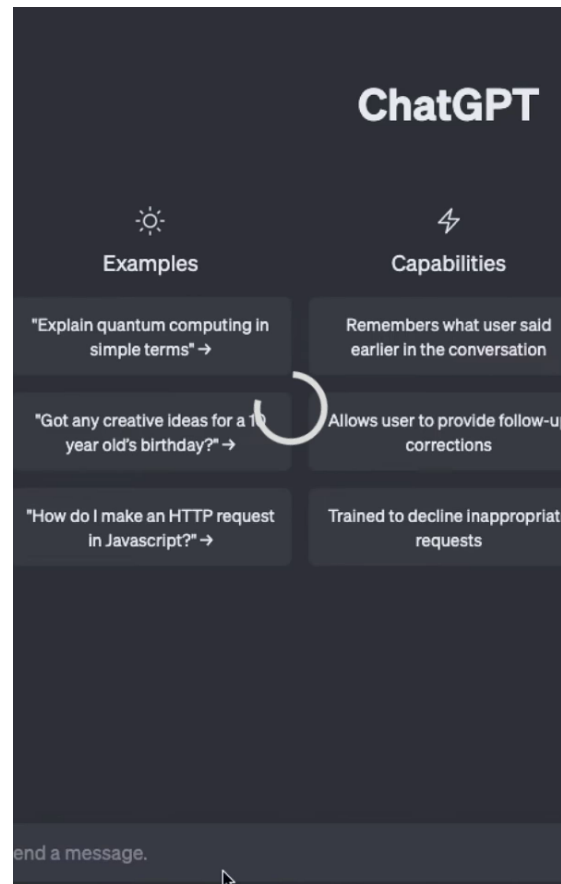


# L'Intelligenza Artificiale in Medicina come Supporto alla Decisione Clinica

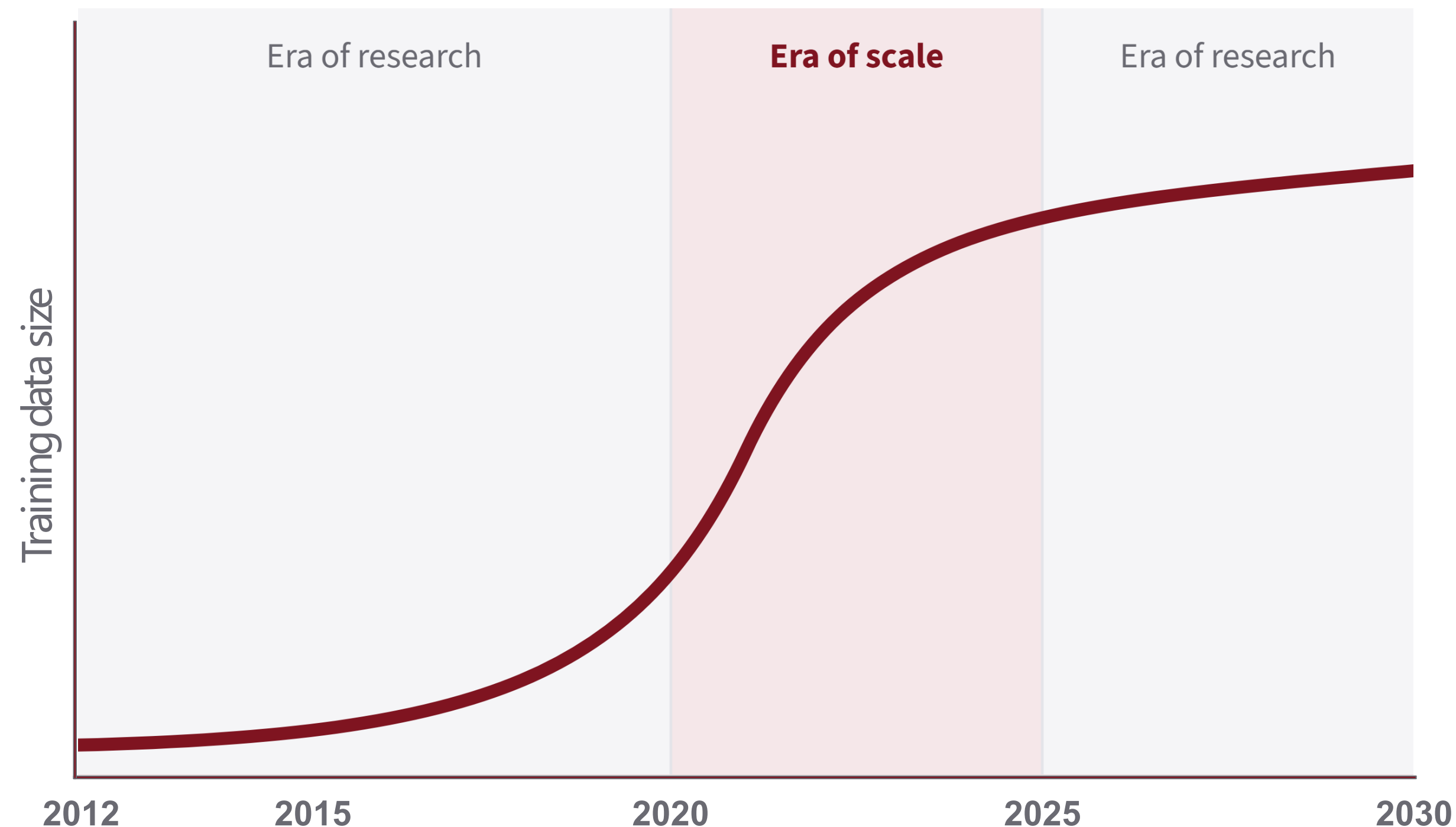
Shared Data · Reliable Supervision · Robust Models  
Actionable Outputs · Trusted Evaluation

**Federico Bolelli**

AlmageLab — University of Modena and Reggio Emilia · federico.bolelli@unimore.it



General models crossed into the **era of scale**.



VISION-LANGUAGE **650×**

SigLIP 2 10B

BiomedCLIP 15.3M

SEGMENTATION **700×**

SAM 1.1B

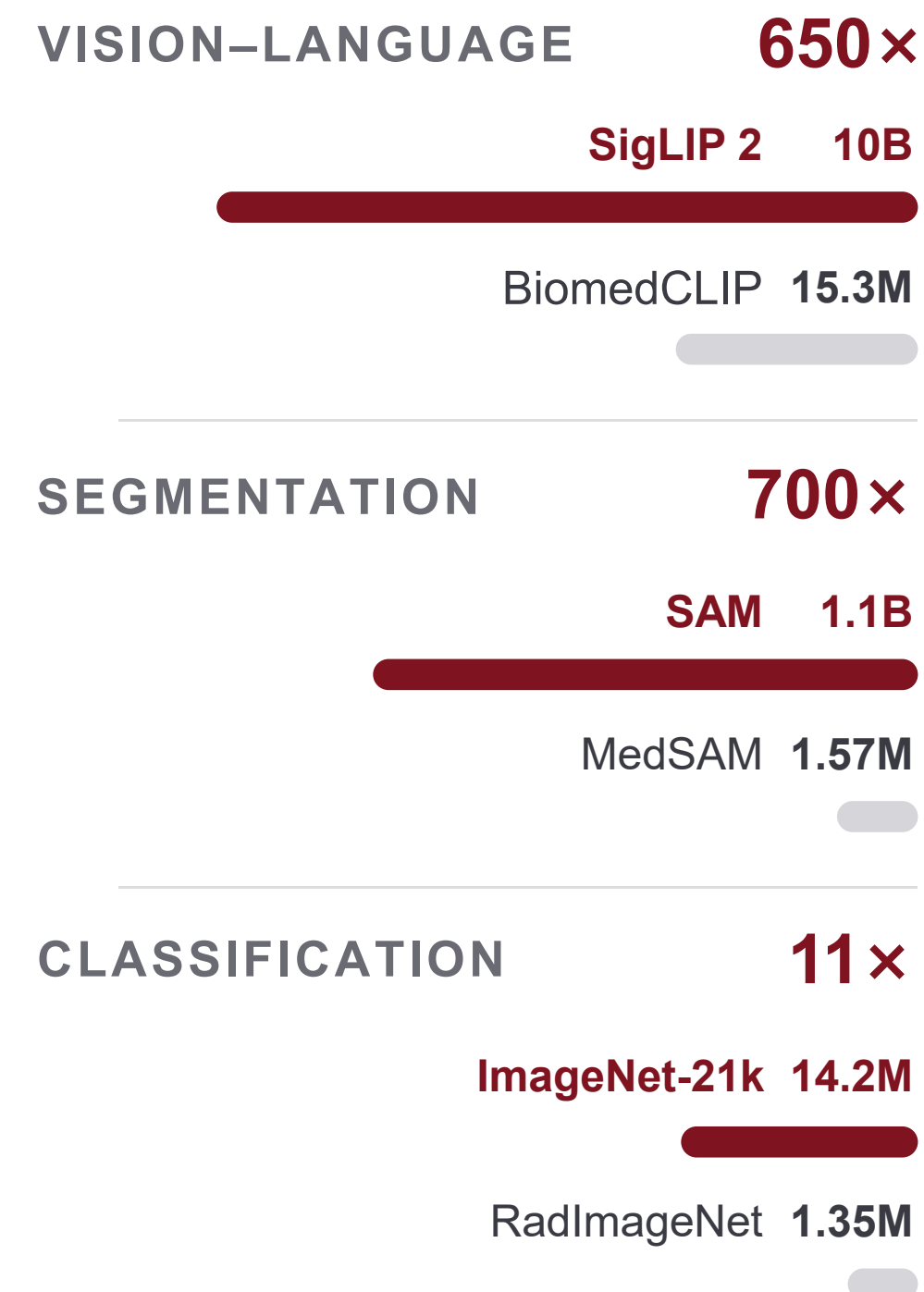
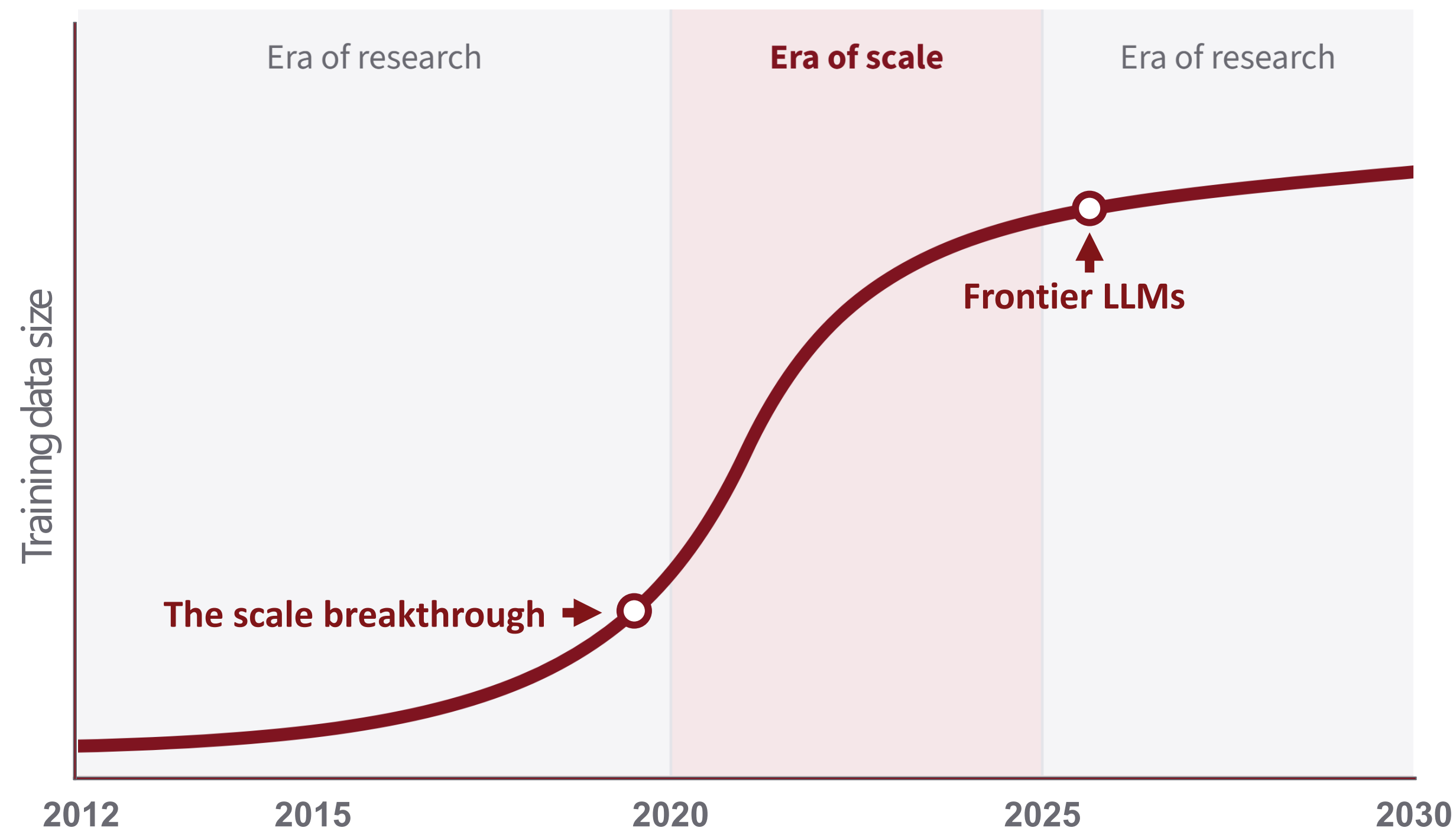
MedSAM 1.57M

CLASSIFICATION **11×**

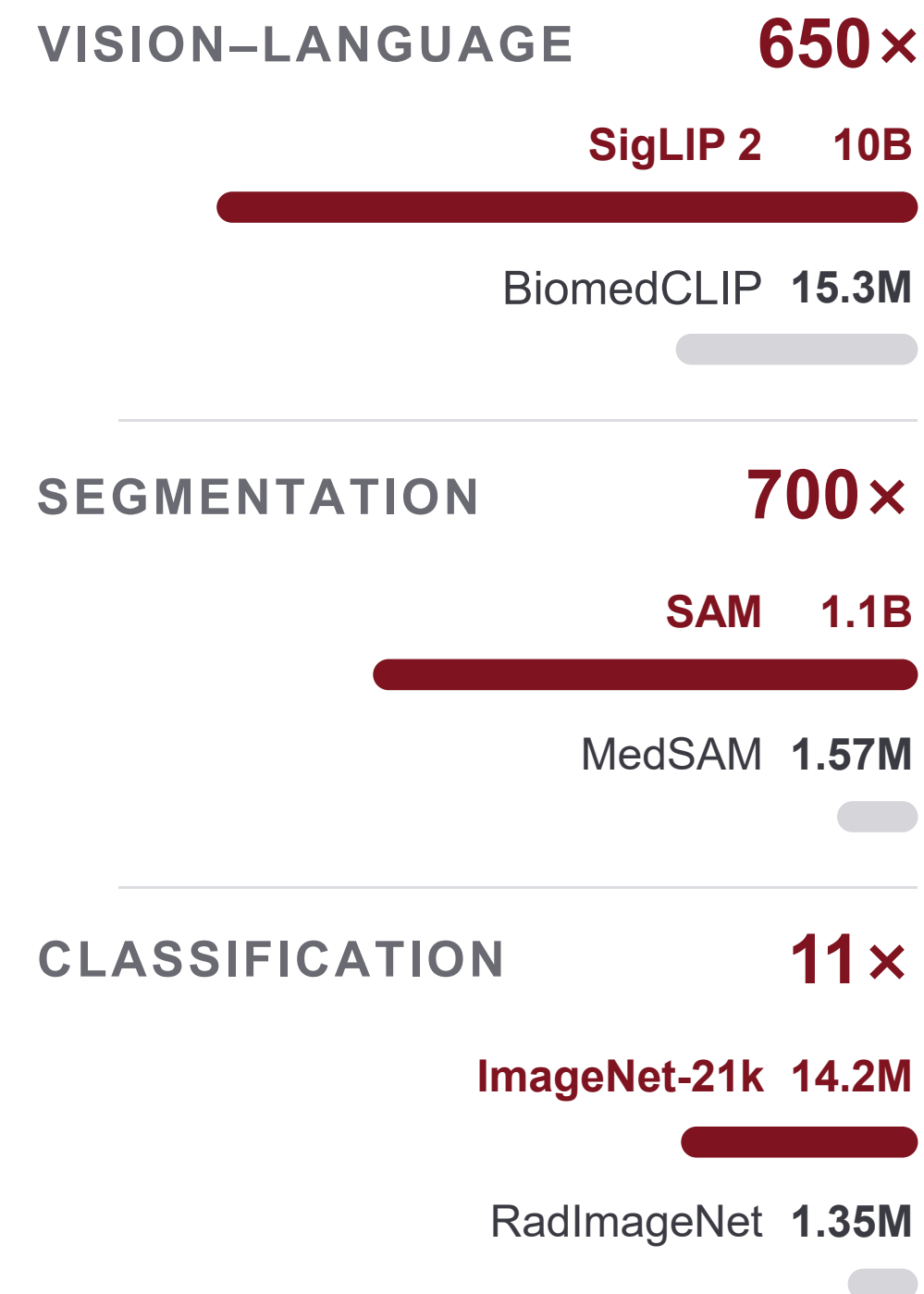
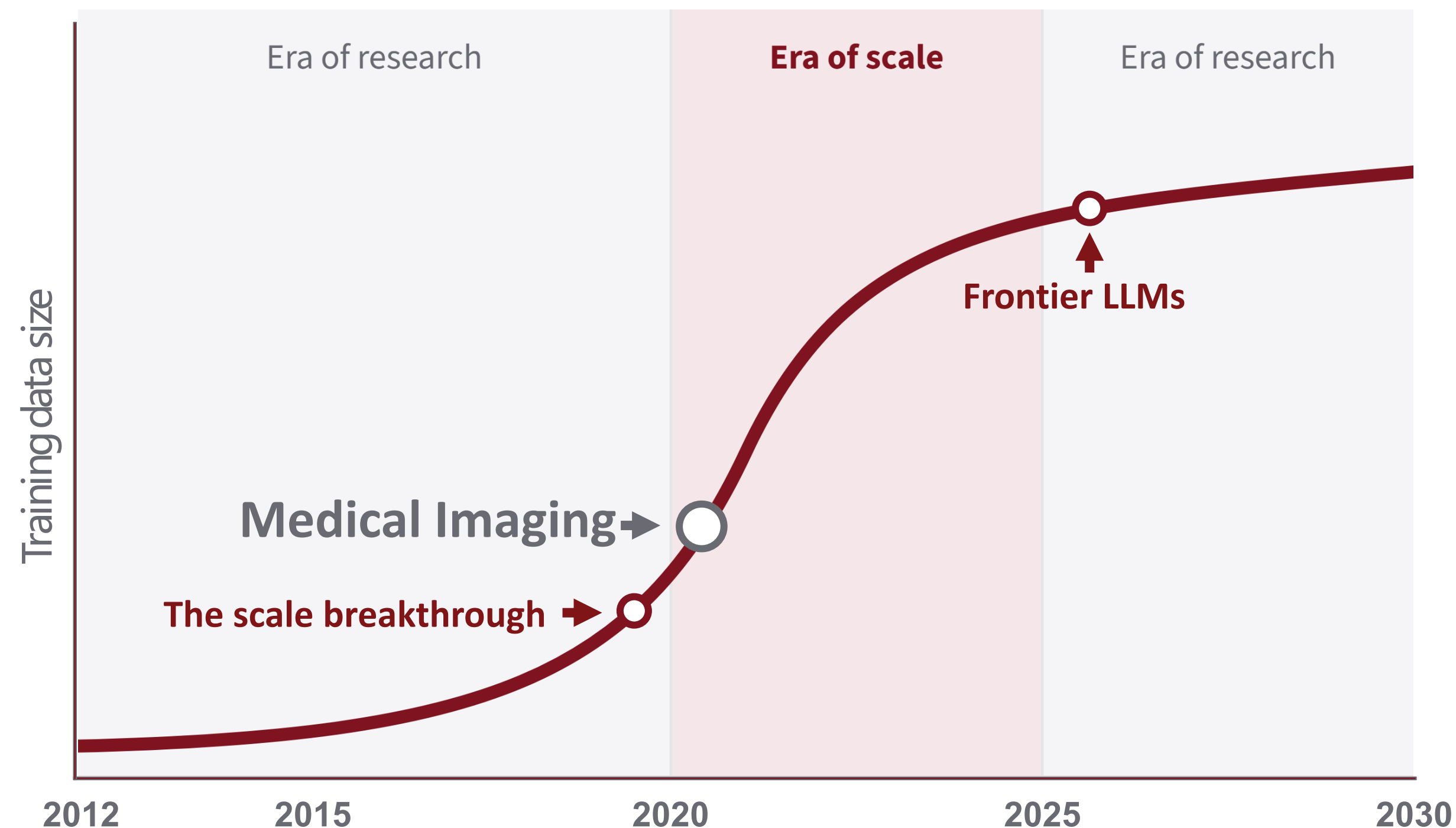
ImageNet-21k 14.2M

RadImageNet 1.35M

General models crossed into the **era of scale**.



General models crossed into the **era of scale**.



## WHERE IT ALREADY WORKS

### PATHOLOGY — BREAST, WSI

**0.994**  
algorithm

**0.884**  
pathologist

AUC of the best algorithm against the best pathologist in a time-constrained diagnostic setting. Seven of 32 algorithms exceeded a panel of 11 pathologists.

### DERMOSCOPY — MELANOMA

**0.86**  
CNN

**0.79**  
dermatologists

ROC AUC of a CNN against the mean of 58 dermatologists in a reader study.

### BREAST HISTOLOGY

**99%**  
panel of four

**85%**  
single observer

Accuracy on benign versus malignant discrimination on previously unseen slides; pooling four independent observers raises it to 99%.

[1] Ehteshami Bejnordi, Babak, et al. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer" *Jama* 2017.

[2] Haenssle, Holger A., et al. "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists." *Annals of Oncology* 2018.

[3] Levenson et al., "Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images" *PLoS ONE* 2015.

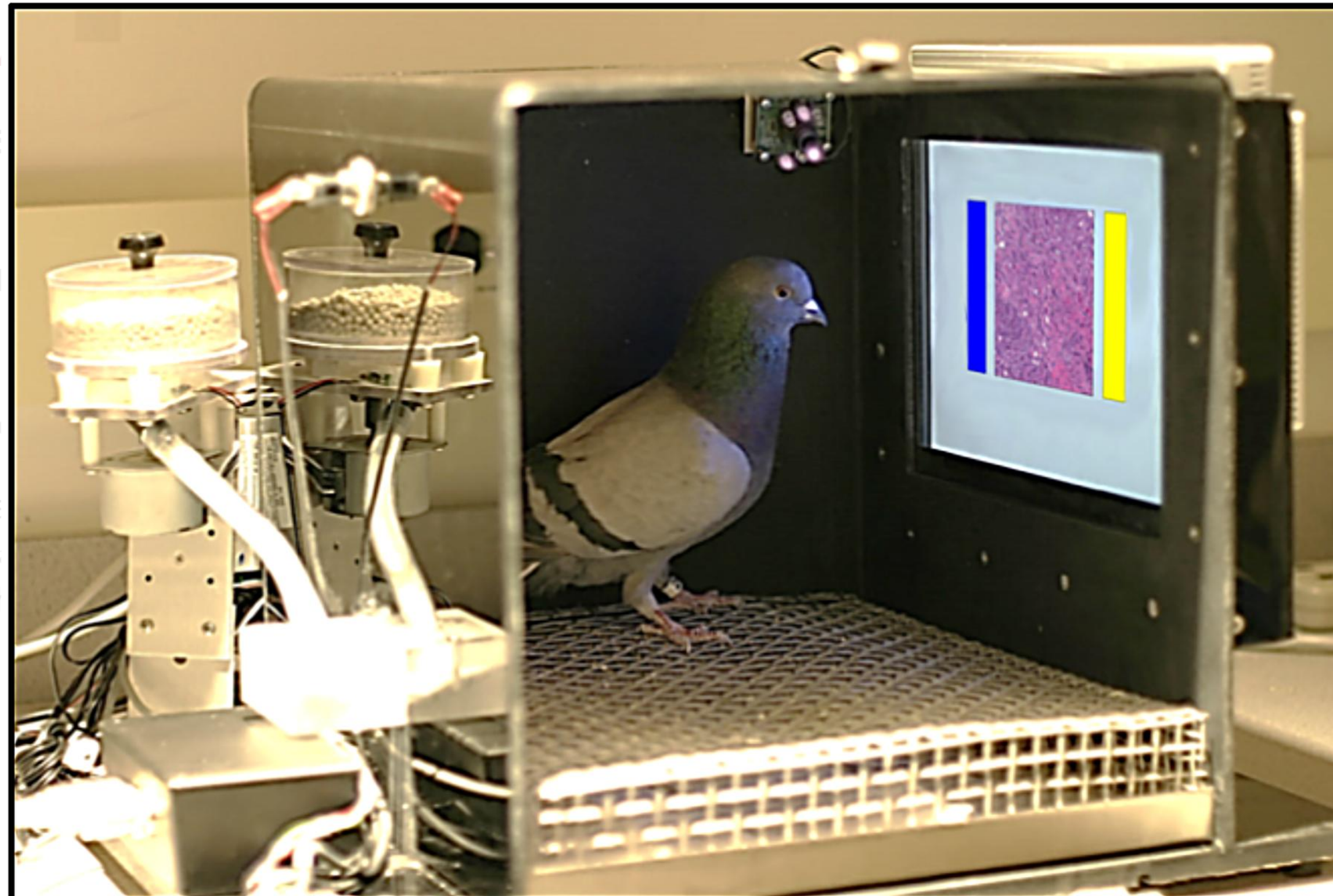
WHERE IT ALREADY WAS

PATHOLOGY — BREAST

0.994

0.884

AUC of the best algorithm a pathologist in a time-constrained setting. Seven of 32 algorithms panel of 11 pathologists.



PATHOLOGY

panel of four

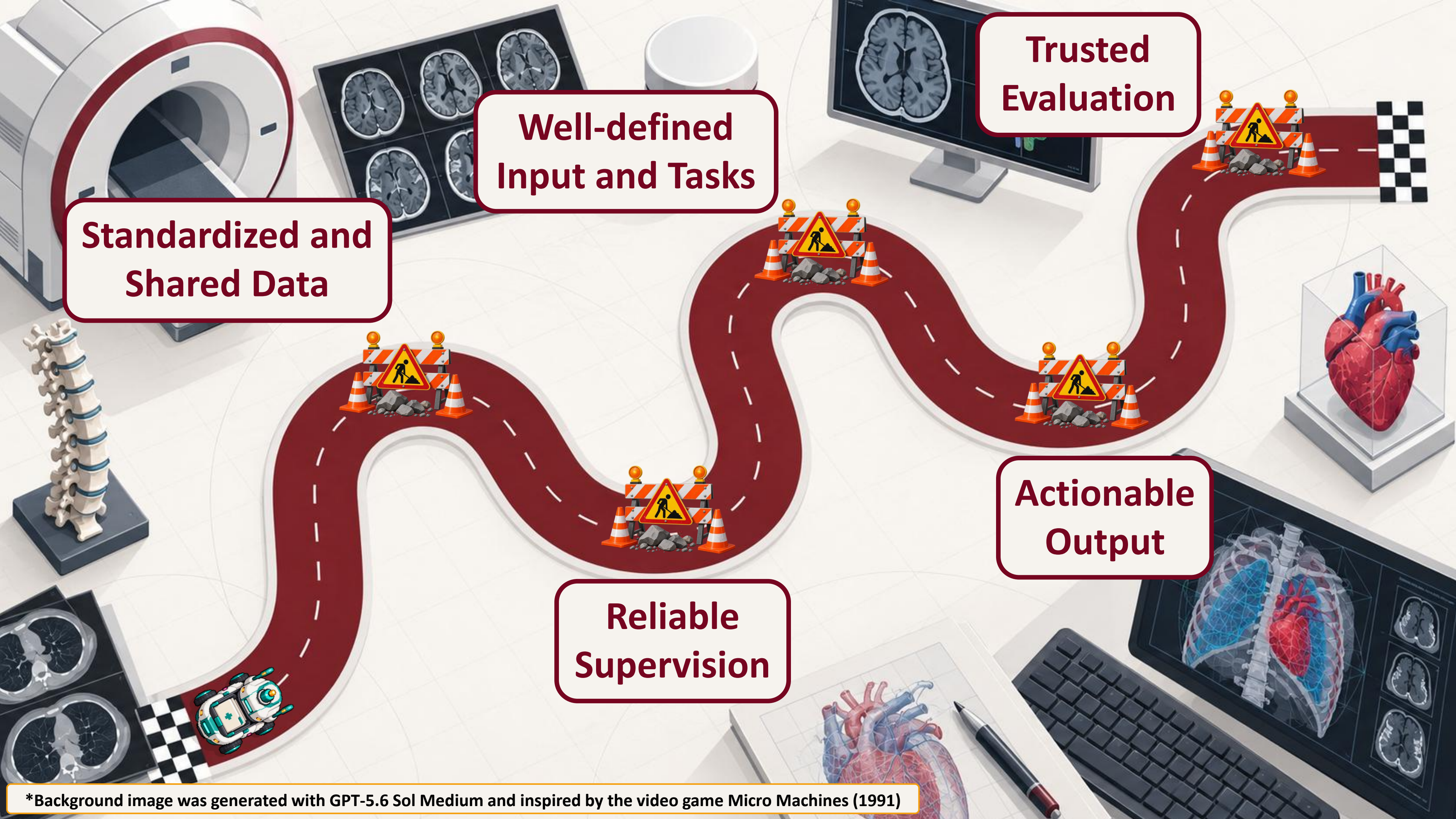
single observer

benign versus malignant  
previously unseen slides;  
independent observers raises it

[1] Ehteshami Bejnordi, Babak, et al. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer" *Jama* 2017.

[2] Haenssle, Holger A., et al. "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists." *Annals of Oncology* 2018.

[3] Levenson et al., "Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images" *PLoS ONE* 2015.



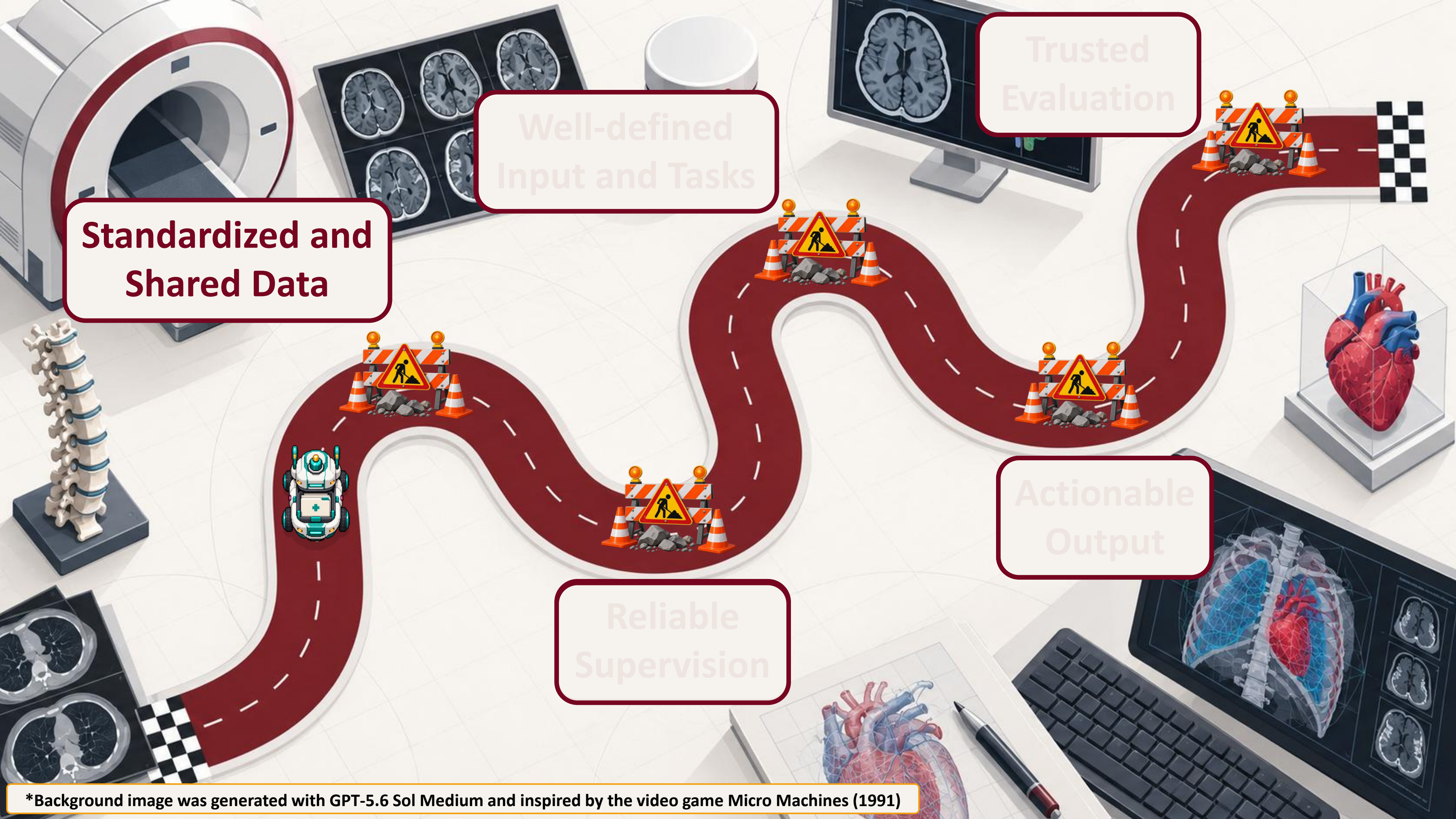
**Standardized and  
Shared Data**

**Well-defined  
Input and Tasks**

**Trusted  
Evaluation**

**Actionable  
Output**

**Reliable  
Supervision**



**Standardized and  
Shared Data**

**Well-defined  
Input and Tasks**

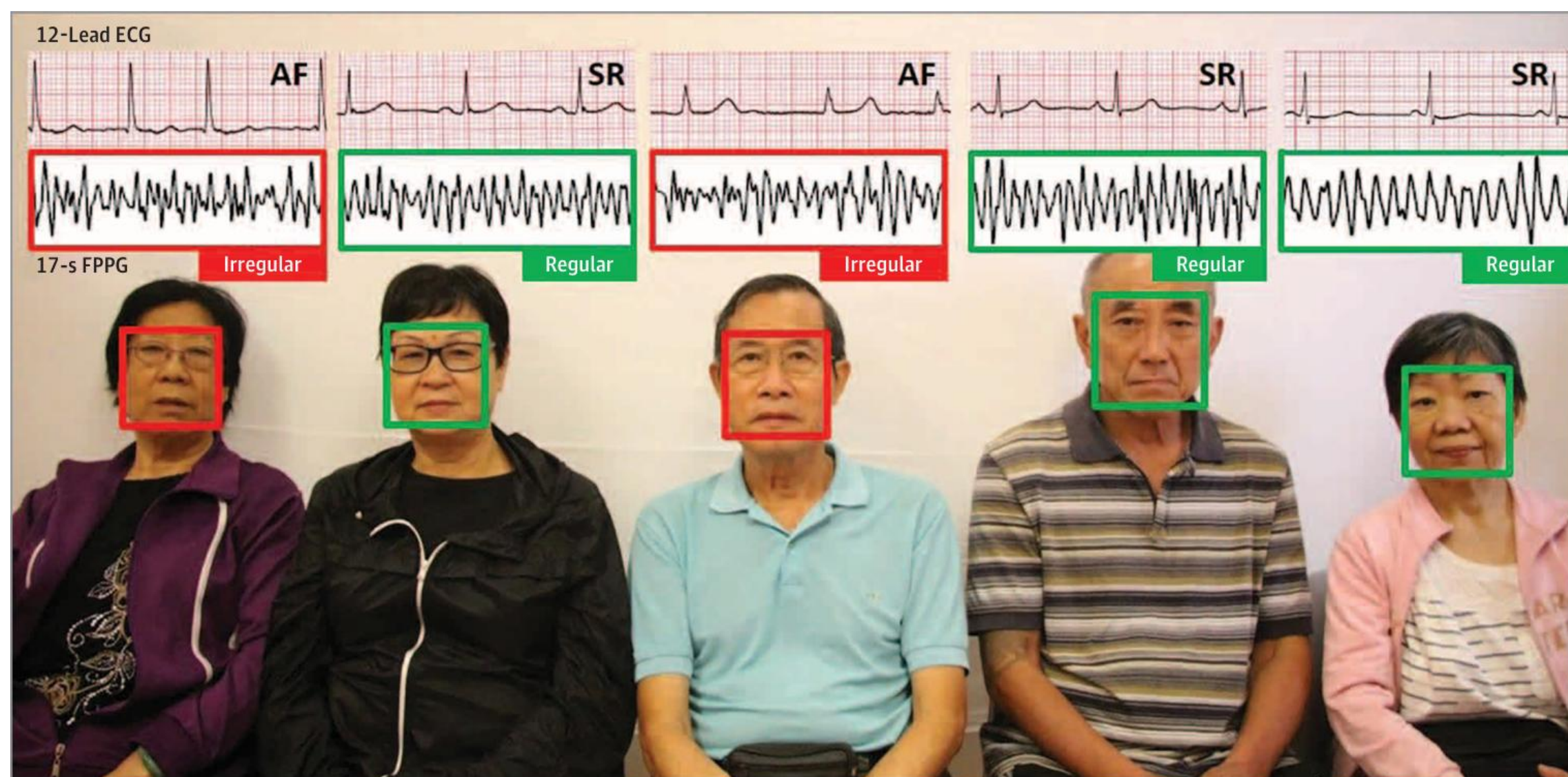
**Trusted  
Evaluation**

**Actionable  
Output**

**Reliable  
Supervision**

## CASE STUDY · STANDARDIZED AND SHARED DATA

Striking accuracy on private data cannot be verified, reproduced or certified.<sup>1</sup>



RESULT · 44 PATIENTS · 64 VIDEOS

**93.8%**

sensitivity

**98.1%**

specificity

### WHAT CANNOT BE CHECKED

Private data and a proprietary algorithm: **no one else can reproduce it.**

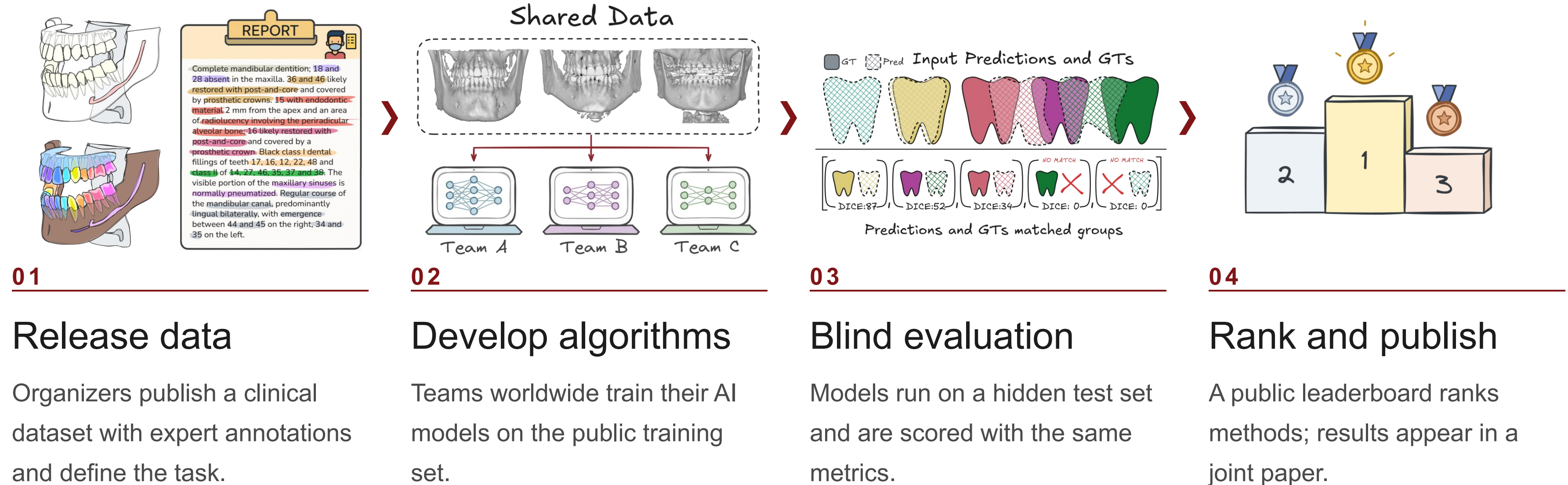
### WHAT A MEDICAL DEVICE NEEDS

- **Independent validation** on a public, standardized benchmark.
- A **reproducible model** that can be approved for decision support.

[1] Yan B.P., et al. "High-Throughput, Contact-Free Detection of Atrial Fibrillation From Video With Deep Learning." JAMA Cardiology 2020;5(1):105–107. Figure adapted from the paper. The algorithm (Cardio) is proprietary.

## A FAIR WAY TO BENCHMARK AI ALGORITHMS

An open competition: many teams solve **the same clinical task** on the same data.<sup>1</sup>



01

### Release data

Organizers publish a clinical dataset with expert annotations and define the task.

02

### Develop algorithms

Teams worldwide train their AI models on the public training set.

03

### Blind evaluation

Models run on a hidden test set and are scored with the same metrics.

04

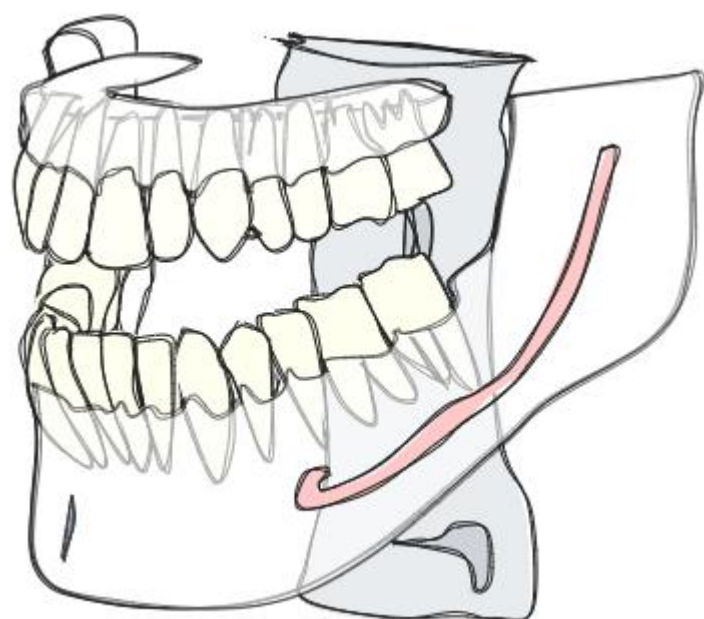
### Rank and publish

A public leaderboard ranks methods; results appear in a joint paper.

[1] L. Maier-Hein, et al. "Why rankings of biomedical image analysis competitions should be interpreted with care." Nature Communications 2018.



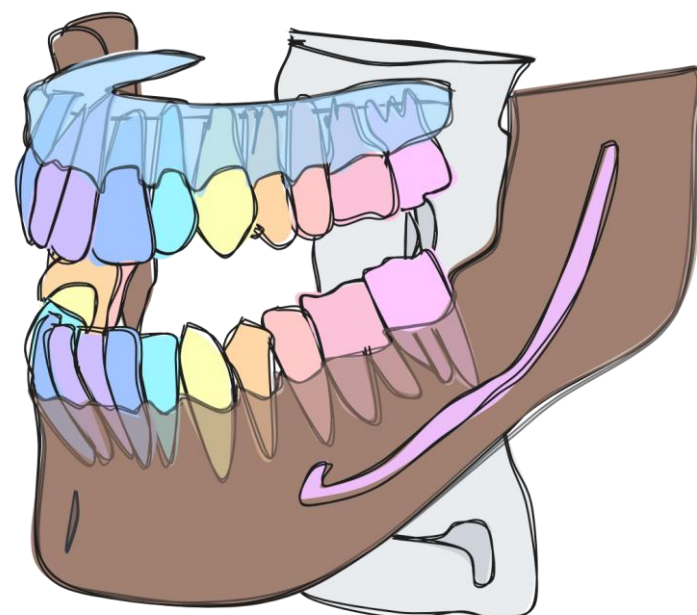
1 class



2023



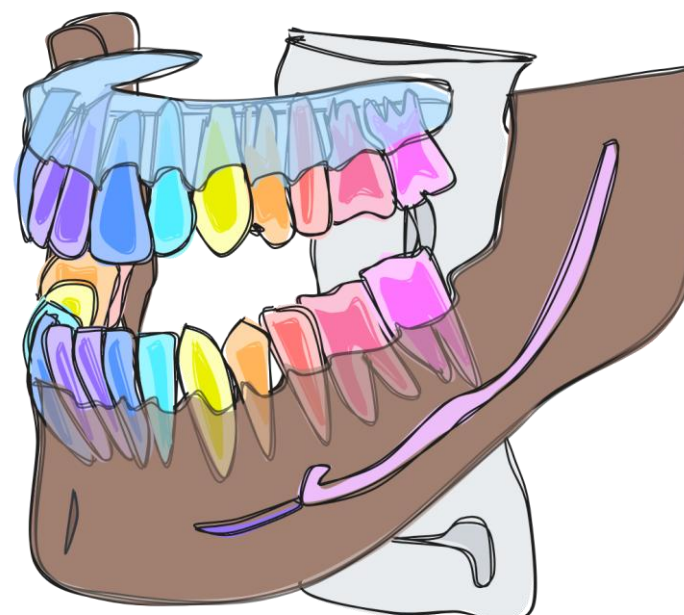
42 classes



2024



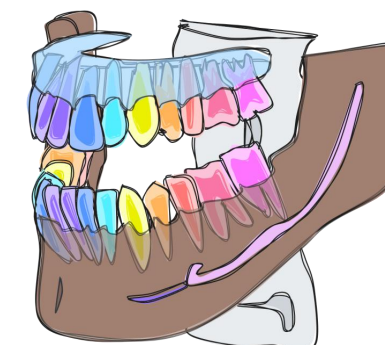
77 classes



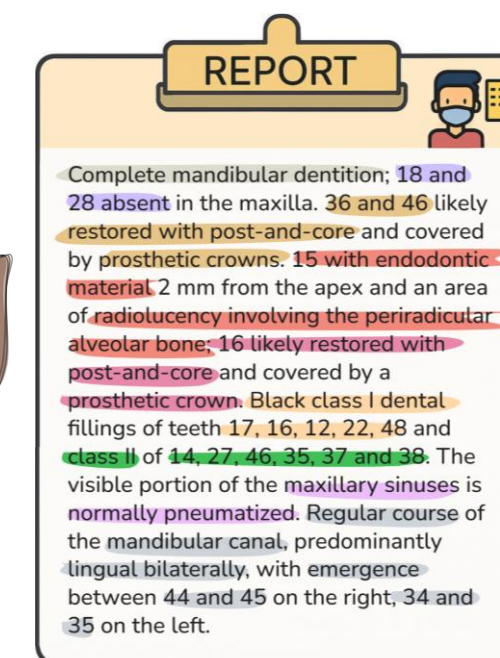
2025



77 classes + Reports



2026



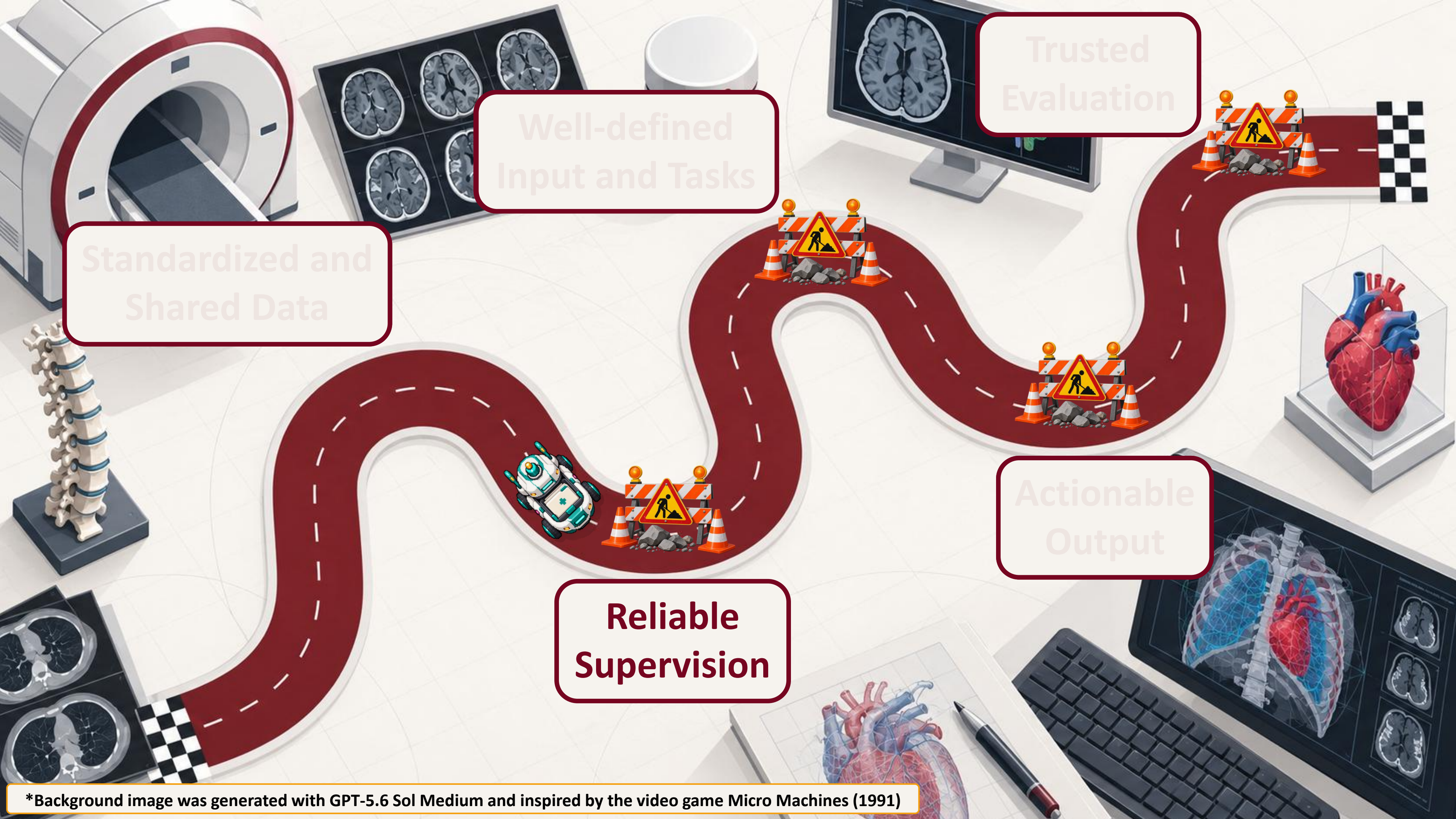
[1] Bolelli F., et al. "Segmenting the Inferior Alveolar Canal in CBCT Volumes: The ToothFairy Challenge" IEEE Transactions on Medical Imaging 2024.

[2] Bolelli F., et al. "Multi-structure segmentation in CBCT volumes: The ToothFairy2 challenge." Medical Image Analysis 2026.

[3] Lumetti L., et al. "ToothFairy3: Scaling CBCT Maxillofacial Segmentation to 77 Classes with U-Mamba2." MICCAI 2026.

[4] Lumetti L., et al. "Ontology-Grounded Structured Prediction for Dental CBCT Reporting." MICCAI 2026.





Standardized and  
Shared Data

Well-defined  
Input and Tasks

Trusted  
Evaluation

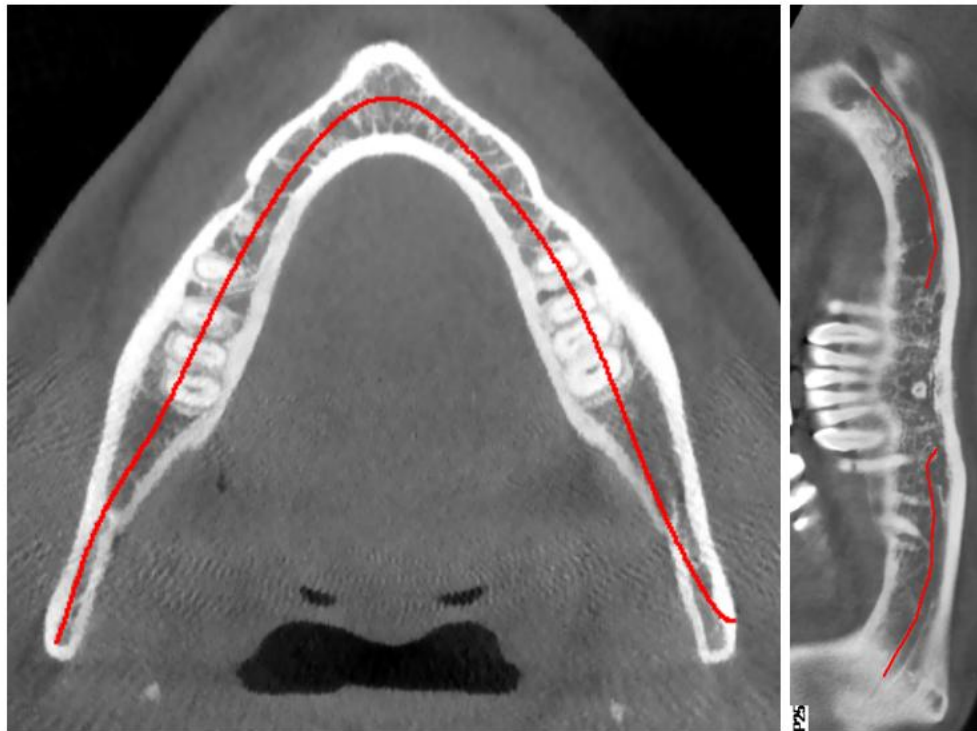
Actionable  
Output

Reliable  
Supervision

**Dense 3D masks are the exception. Sparse labels are what clinics already produce.**

01

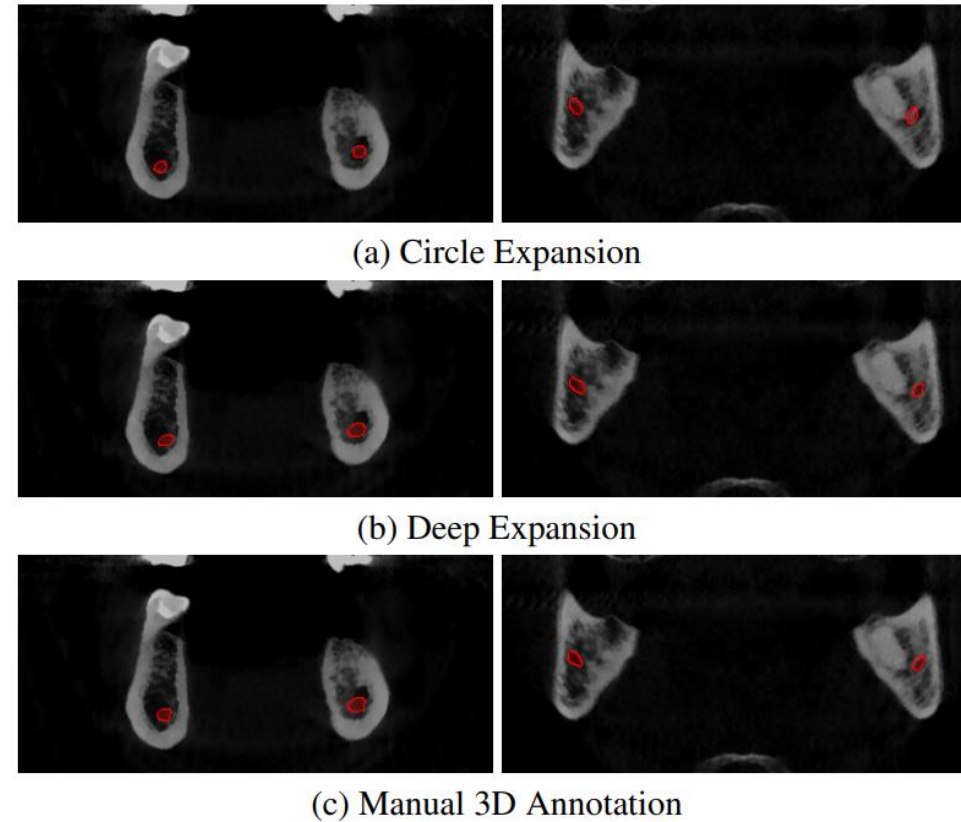
What the clinic gives you



SPARSE BY DEFAULT

02

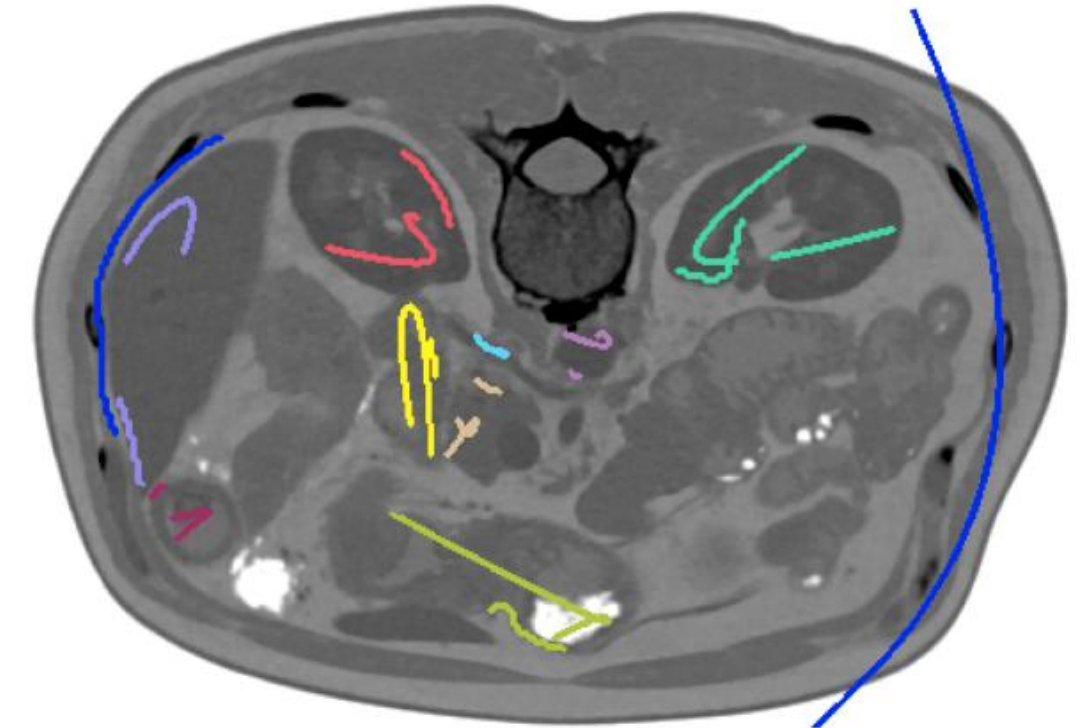
Propagate it



OUR ROUTE — CVPR 2022

03

Or ask for less



THE INTERACTIVE/SCRIBBLE ROUTE

[1] Cipriano M., et al. "Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation", CVPR 2022.

[2] Luo X., et al. "Scribble-Supervised Medical Image Segmentation via Dual-Branch Network and Dynamically Mixed Pseudo Labels Supervision", MICCAI 2022.

[3] Gotkowski, K. , et al. "Revisiting 3D Medical Scribble Supervision: Benchmarking Beyond Cardiac Segmentation", MICCAI 2025.




**Yggdrasil**
v3.3.0
ML mlugli



# Yggdrasil


ACTIVE

**MAXILLOFACIAL**

## Maxillo

Dental & maxillofacial imaging — bite classification, IOS, CBCT and panoramic extraction.

CBCT
IOS
Panoramic

4,100 patients
Enter →


ACTIVE

**NEUROIMAGING**

## Brain

Brain tumor MRI — multi-sequence review with AI-assisted captioning.

MRI T1/T2
FLAIR

AI Voice

1,037 patients
Enter →


ACTIVE

**SURGICAL VISION**

## Laparoscopy

Surgical video annotation — frame-accurate segmentation and tagging.

Video
Keyframes

Segmentation

4 patients
Enter →


ACTIVE

**UROLOGICAL ONCOLOGY**

## Urology

Urological oncology — MRI, WSI digital pathology and Confocale microscopy.

MRI
WSI
Confocale

0 patients
Enter →


ACTIVE

**CARDIOLOGY**

## Cardiology

ECG review — clinical-grid waveform plotting and arrhythmia annotation.

12-lead ECG
Rhythm

Captions

597 patients
Enter →

CONTINUE WHERE YOU LEFT OFF


**Patient Y**  
Maxillo


**027**  
Maxillo


**118**  
Maxillo

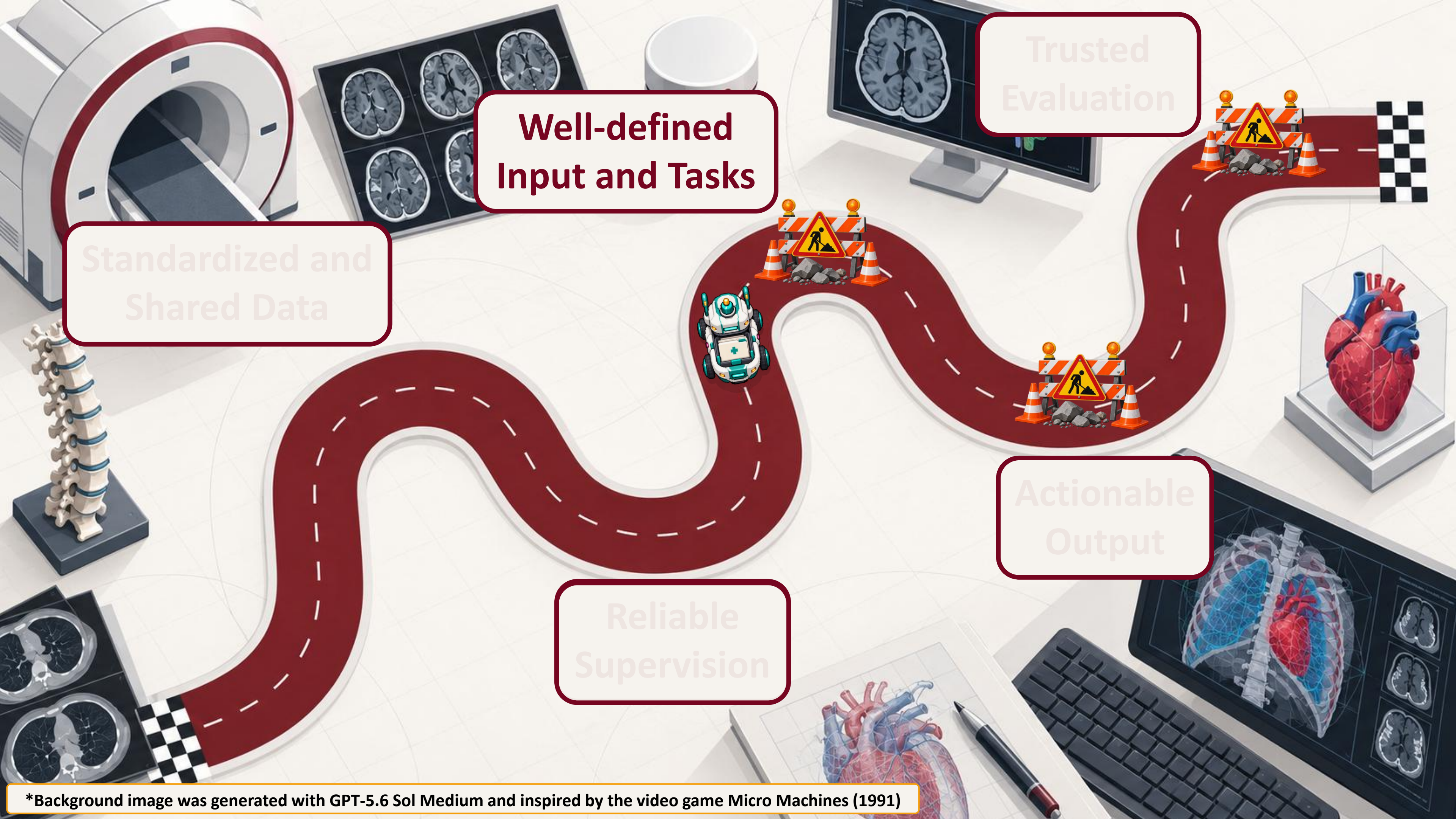

**Patient 7102**  
Maxillo


**A1000**  
Maxillo


**1091\_before**  
Cardiology

Powered by  AIimageLab

© 2026 Yggdrasil  v3.3.0



**Well-defined  
Input and Tasks**

**Standardized and  
Shared Data**

**Trusted  
Evaluation**

**Actionable  
Output**

**Reliable  
Supervision**

**A model that requires every input to be present is a model that cannot be deployed.**

## WHY AN INPUT GOES MISSING

-  Time in the acquisition slot
-  Acquisition costs
-  Patient condition
-  Equipment available at that center

**Accuracy should fall in proportion to the information lost — and the model should stay usable at every subset of inputs.**

## THE MRI EXAMPLE — Brain Tumor Segmentation

Medical AI often assumes that all imaging modalities are available, but real-world data is often incomplete. Robust models should account for missing modalities and maintain reliable performance even when some inputs are unavailable.

**Mamba-based  
Fusion Model**  
IM-Fuse

**A deployed model is not a finished artifact.**

## WHAT CHANGES AFTER DEPLOYMENT

New anatomical structures to cover

New scanners or centres appear

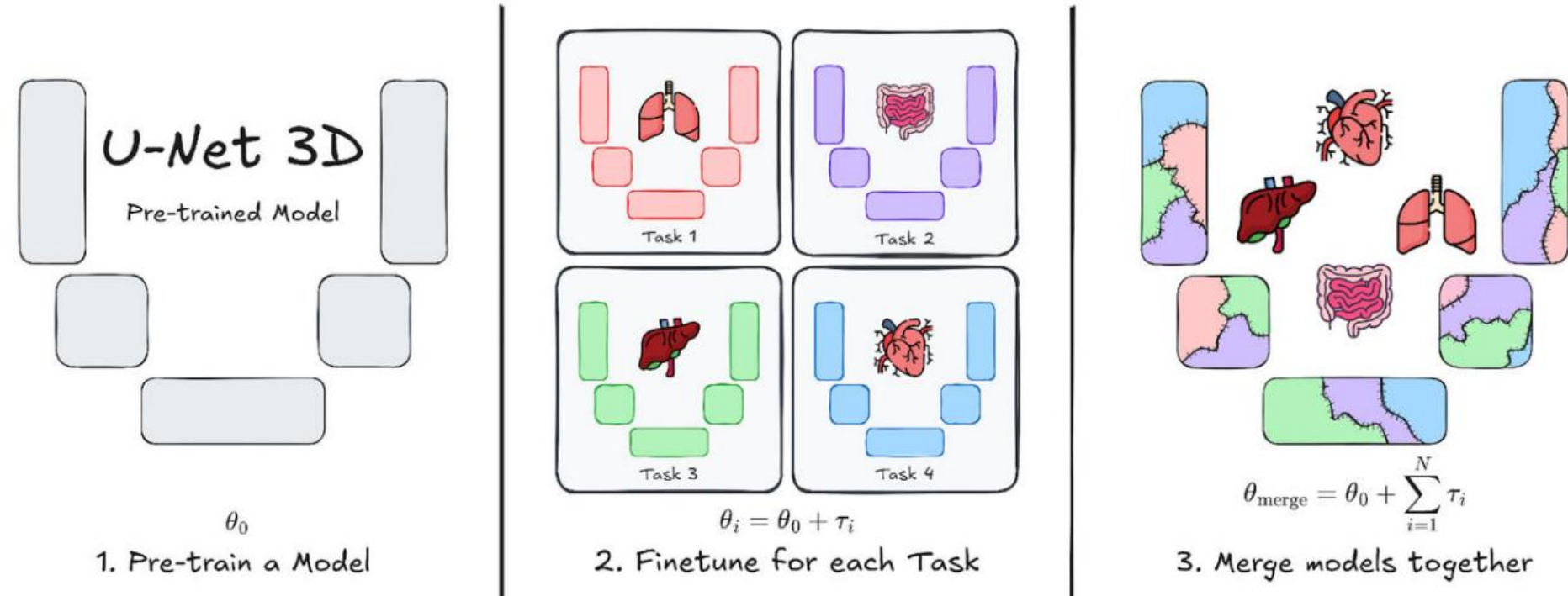
New clinical requirements emerge

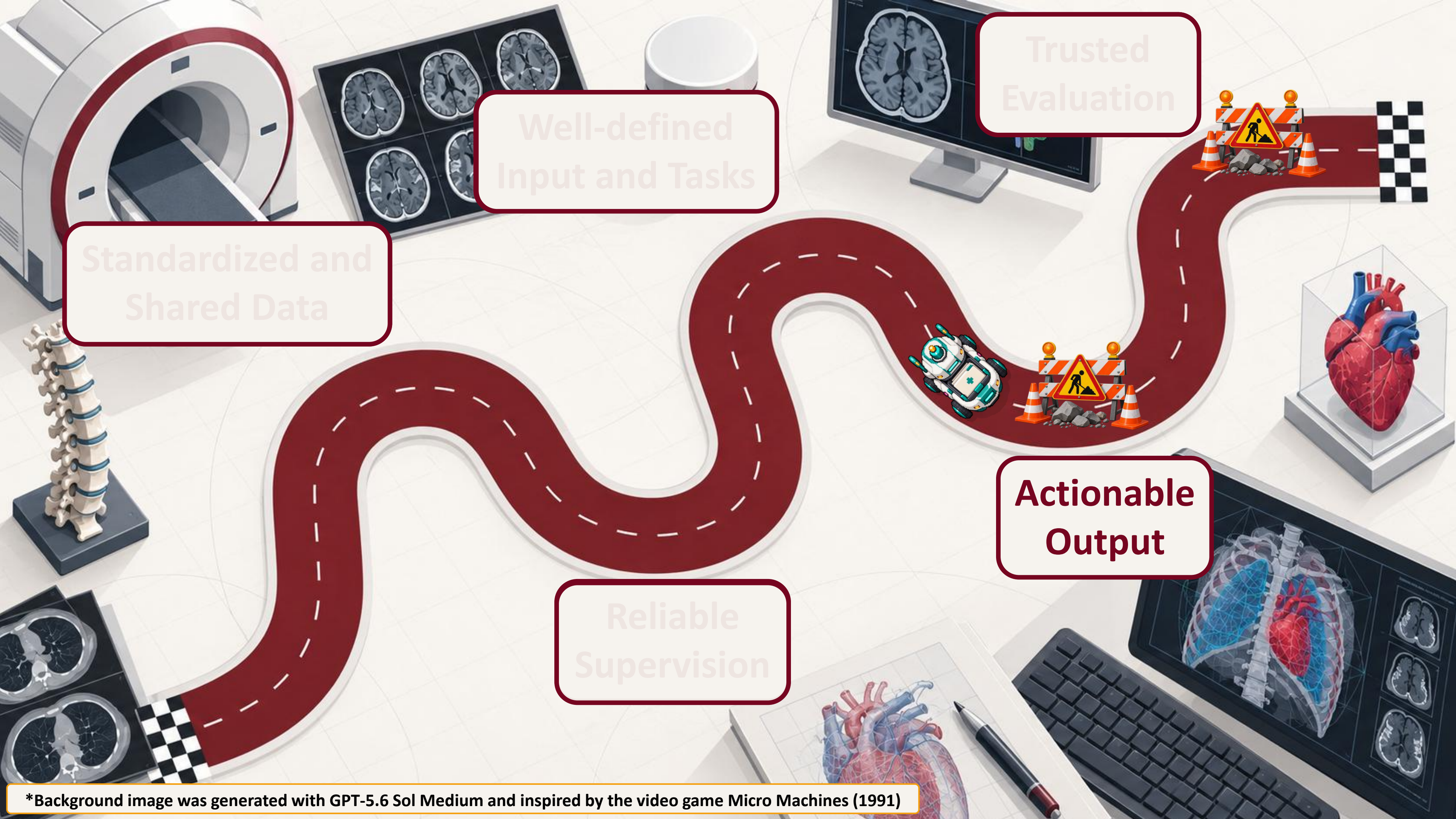
The original training data are gone

**Adding a structure should not mean retraining everything from scratch.**

## THE CT/CBCT EXAMPLE — U-NET TRANSPLANT

Instead of one monolithic network, train specialists separately and merge them. Pre-training that encourages wide minima makes those task vectors combine cleanly — evaluated on ToothFairy2 and BTCV Abdomen.





Standardized and  
Shared Data

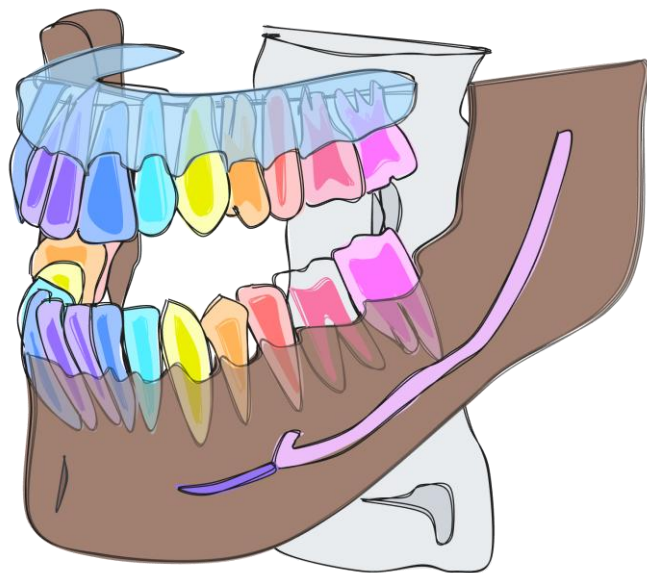
Well-defined  
Input and Tasks

Trusted  
Evaluation

Actionable  
Output

Reliable  
Supervision

## A patient is not a volume



SEGMENTED VOLUME

### PATH A — RAW GEOMETRY

tooth 38

1.8 cm<sup>3</sup>

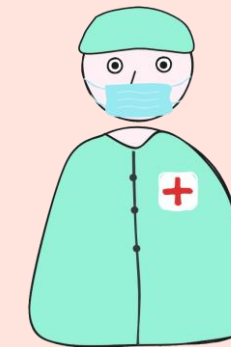
Accurate, but not yet a clinical statement



### PATH B — REPORT GENERATION

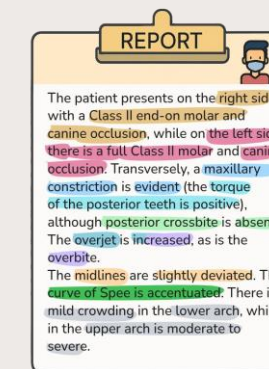
**Findings stated in clinical language**

Relations, severity, laterality; what the volume implies



### Clinician's intervention

A human has to translate the numbers into findings



### Directly actionable

The output now carries the clinical meaning

Segmentation describes anatomy. Reports communicate clinical meaning.

**ToothFairy4 moves the benchmark from masks to structured findings.**

## WHAT WE RELEASED

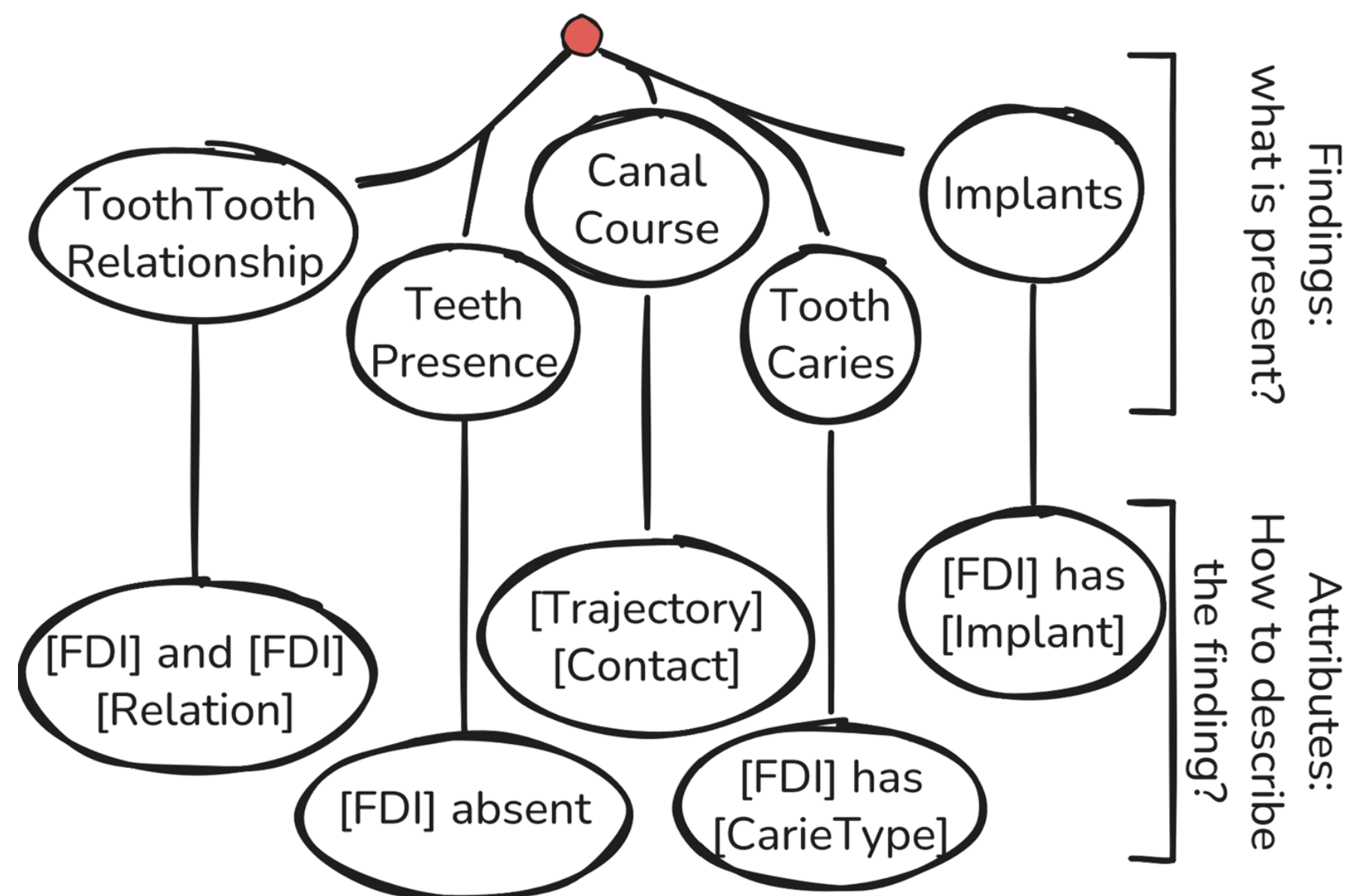
893 clinical reports for 529 public CBCT volumes

A clinician-designed OWL ontology, 13 finding types

Every report validated as an RDF/Turtle instance

Six metrics that decouple the three stages

**Reporting posed as structured prediction — not free text.**



**Trusted  
Evaluation**

**Well-defined  
Input and Tasks**

**Standardized and  
Shared Data**

**Actionable  
Output**

**Reliable  
Supervision**

## COMMON LEXICAL METRICS

**BLEU** n-gram precision

**METEOR** unigram alignment

**ROUGE** n-gram recall

All three score how a report is *phrased*.  
None of them check whether the finding is semantically **true**.

REF

Teeth 18, 28, 38, 48 are **absent**

GENERATED

Teeth 18, 28, 38, 48 are **endodontically treated**

6 of 7 tokens identical → high n-gram overlap

**the one token that carries the diagnosis**

**0.68** BLEU-4

reads as a strong match

**0.00** FINDING AGREEMENT

wrong diagnosis!

**So maybe a more structured metric that checks actual findings is needed....**

LMarena Dental



Competitions

Manage

Live users

Logout

## Competitions

Choose an active competition to vote or review its ranking.

ACTIVE EVALUATION

### B2T testset LMarena

Competition to decide the winner of the Bite2Text Challenge

Intraoral photos

IOS scans

Start voting →

[View leaderboard](#)

ACTIVE EVALUATION

### TF4 testset LMarena

Competition to decide the winner of the Toothfairy4 Challenge

CBCT

Panoramic

Start voting →

[View leaderboard](#)

## MAIN TAKEAWAYS

Every direction helps — **none** of them is a panacea

01

### Challenges solved comparison, not clinical evaluation

Adoption as a medical device also needs testing in the wild: prospective, multicentre, real patients.

02

### GDPR blocks most data collection and sharing

Few public datasets slow down benchmarking and the development of robust models.

03

### Expert annotation is still a bottleneck

Labelling 3D scans takes time, and experts disagree without a shared protocol.

04

### Interactive and agentic AI workflows are still early

Systems that combine tools, modalities and clinician feedback are not yet clinic-ready.

05

### The clinician stays in the loop

Verifying outputs, planning treatment, handling rare cases, informing the patient, owning the decision.

## AIMAGELAB — UNIVERSITY OF MODENA AND REGGIO EMILIA

### GROUP



F. Bolelli



K. Marchesini



E. Candeloro



O. Carpentiero



N. Morelli



E. Vezzali



L. Borghi



G. Casari



G. Rosati



Y. Assefa



M. Lugli



L. Lumetti



A. Veronese



C. Grana

### CLINICAL PARTNERS

A. Anesi · M. Di Bartolomeo · A. Pellacani · P. Minafra · G. Boriani · J.F. Imberti · G. Pellacani · L. Bertoni · G. Bianchi · S. Puliatti · G. Besutti · R. Magistroni · F. Benuzzi · F. Lui · L. Lombardo · F. Cremonini

### COLLABORATIONS

# L'Intelligenza Artificiale in Medicina come Supporto alla Decisione Clinica

## Thank You!

Shared Data · Reliable Supervision · Robust Models  
Actionable Outputs · Trusted Evaluation

**Federico Bolelli**

