From General to Clinical: Adapting Foundation Models for Medical Images

Kick-off Presentation

Feb 4th, 2026

Who are we?

Computational Imaging and AI in Medicine

Prof. Dr. Julia Schnabel

Who are we?

Sameer Ambekar

Research interest:
- Distribution shifts
- Test-time adaptation
- Foundation models

Dr. Laura Daza

Research interest:
- Multi-modal Learning
- Foundation models
- Segmentation

Structure of the seminar

Week	Session
1	Introduction to the seminar
2	How to read a paper and do poster presentations
3	Theory on self-supervised learning
4	Invited talk (SSL)
5	Student presentations
6	Student presentations
7	FM in the natural and medical domains
8	Invited talk (FM)
9	Student presentations
10	Student presentations
	Christmas break
11	How to adapt models to new tasks and domains
12	Invited talk (adaptation)
13	Student presentations
14	Student presentations
15	Poster presentations

Deliverables & grading

Oral presentation:
- 30 min presentations – 10 min questions
- Presentation date depends on the topic

Poster presentation:
- 3 – 5 min pitch
- End of the semester

Paper selection:
- We provide a list of papers
- Assignment of papers for presentations after the seminar (moodle)
- Papers on both CV methodologies or adaptations to medical images

Grading:
- We will take into account the following things:
  - Oral presentation
  - Poster presentation
  - Attendance
  - Participation

Poster session at the end of the semester

Goals of the Seminar

Theoretical knowledge:

(i) Understand the principles of self supervised learning:
Learn how to leverage large quantities of data without the need of annotations

(ii) Learn what are foundation models:
Understand what are these large models that are so popular right now, how to train them and how to use them

(iii) Understand how to adapt foundation models to the medical domain:
Understand how to translate existing foundation models created in other domains to medical applications

Research skills:

- How to read and present a scientific paper
- How to design and present a scientific poster

Guest Lecture from W2025: For Transformers & LLMs: When Softmax Attention stops being truly sharp, & why they attend the first token due to Attention sinks?

UNIVERSITY OF OXFORD

Google DeepMind

softmax is not enough (for sharp size generalisation)

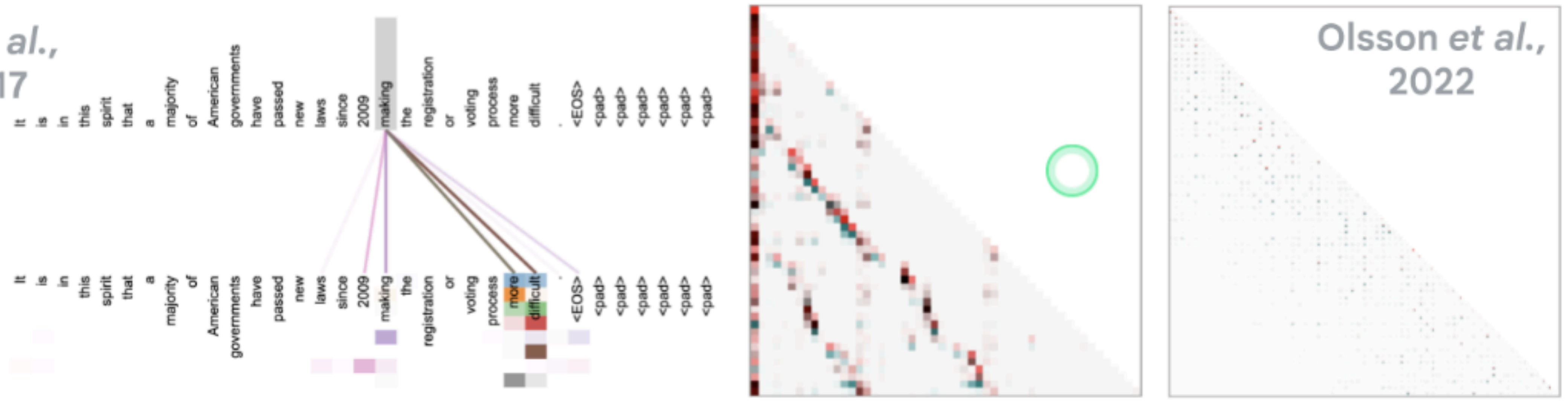
Petar Veličković · Christos Perivolaropoulos · Federico Barbero · Razvan Pascanu

How do Transformers choose what to focus on?

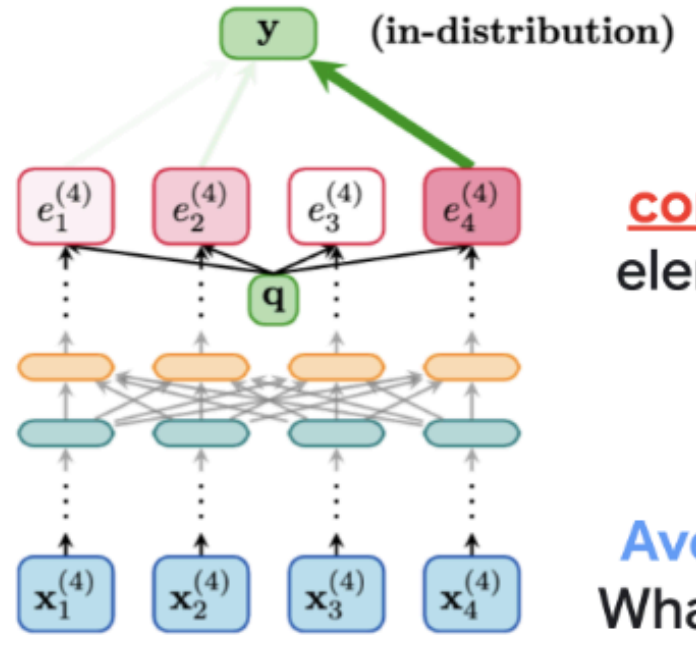
$$\texttt{softmax}_\theta(\mathbf{e}) = \left[\frac{\exp(e_1/\theta)}{\sum_k \exp(e_k/\theta)} \quad \dots \quad \frac{\exp(e_n/\theta)}{\sum_k \exp(e_k/\theta)}\right]$$

Vaswani *et al.*, NeurIPS'17

Olsson *et al.*, 2022



Key assumption: sharpness!



Focusing on a constant number of elements in the input

max is sharp
Average is not sharp
What about softmax?

[Register here](#)

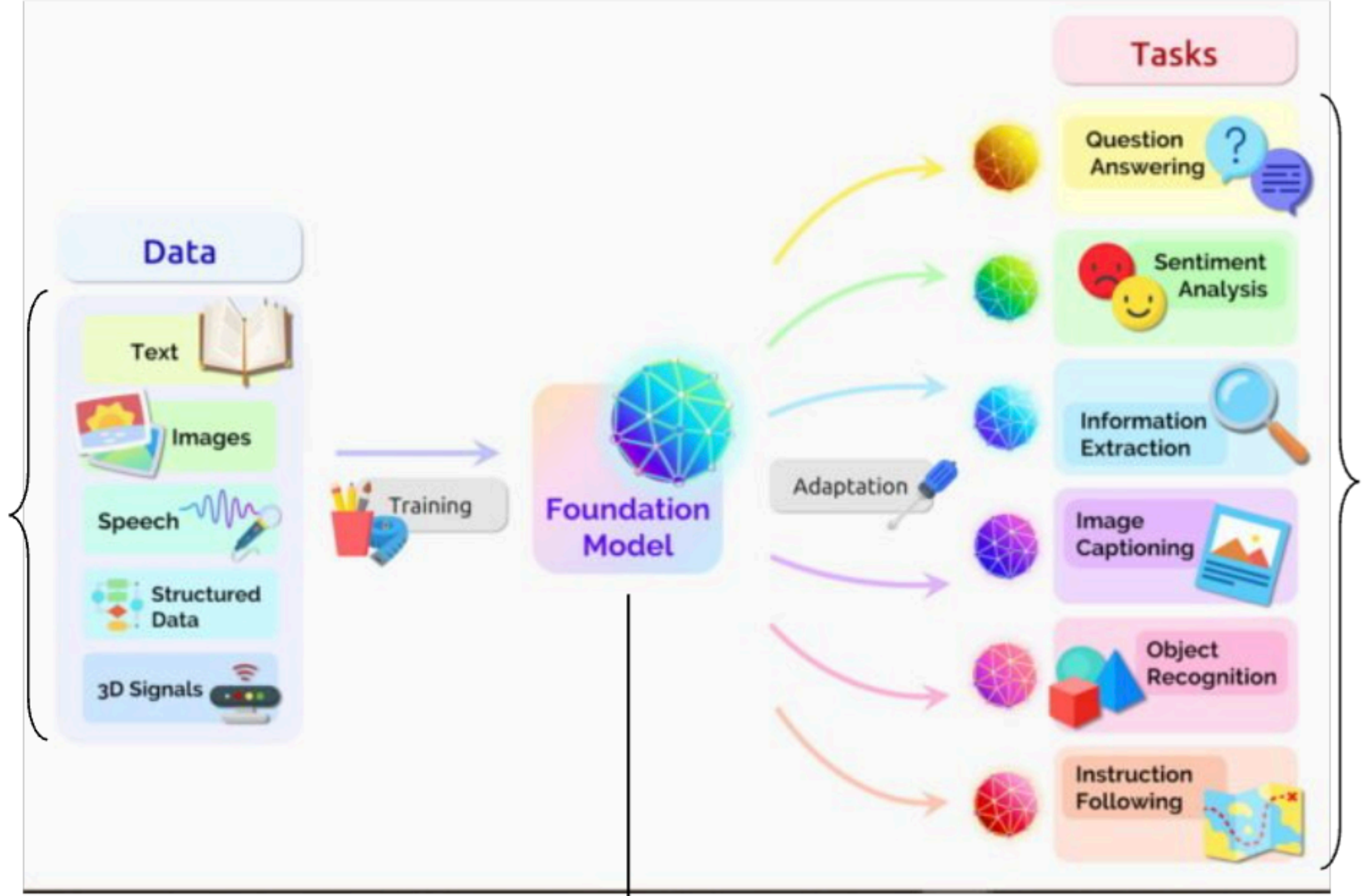
Guest Lecture by Christos Perivolaropoulos from Google Deepmind on January 22nd, 2pm-3pm via Zoom

Part of seminar: From General to Clinical: Adapting Foundation Models for Medical Images, TU Munich
By: Sameer Ambekar, Laura Daza under Prof. Julia Schnabel

Towards Foundation Models

Massive
amounts of data

Semi-, **self-**, unsupervised training



https://blogs.nvidia.com/wp-content/uploads/2023/03/Transformer-apps.jpg

Can be adapted to **many** downstream tasks

Huge models
(hundreds of billions of parameters)

Pretraining stage is extremely expensive

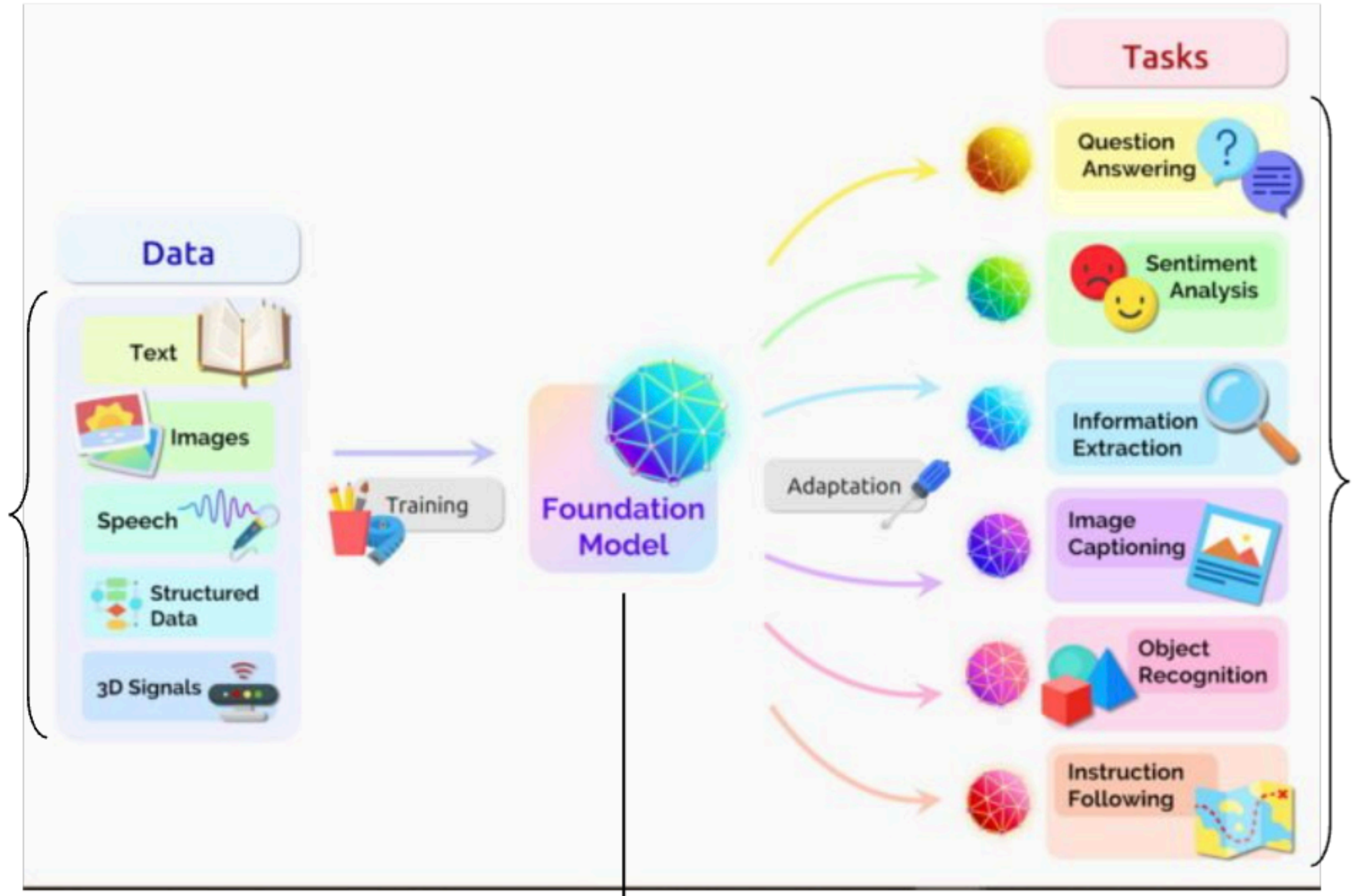
Fine tuning is way cheaper than training for every task

Towards Foundation Models

Millions or even billions of data samples
- Can be “easily” collected from the internet
- Many large companies create their own huge datasets

Massive amounts of data

Semi-, **self-**, unsupervised training



https://blogs.nvidia.com/wp-content/uploads/2023/03/Transformer-apps.jpg

Can be adapted to **many** downstream tasks

Huge models
(hundreds of billions of parameters)

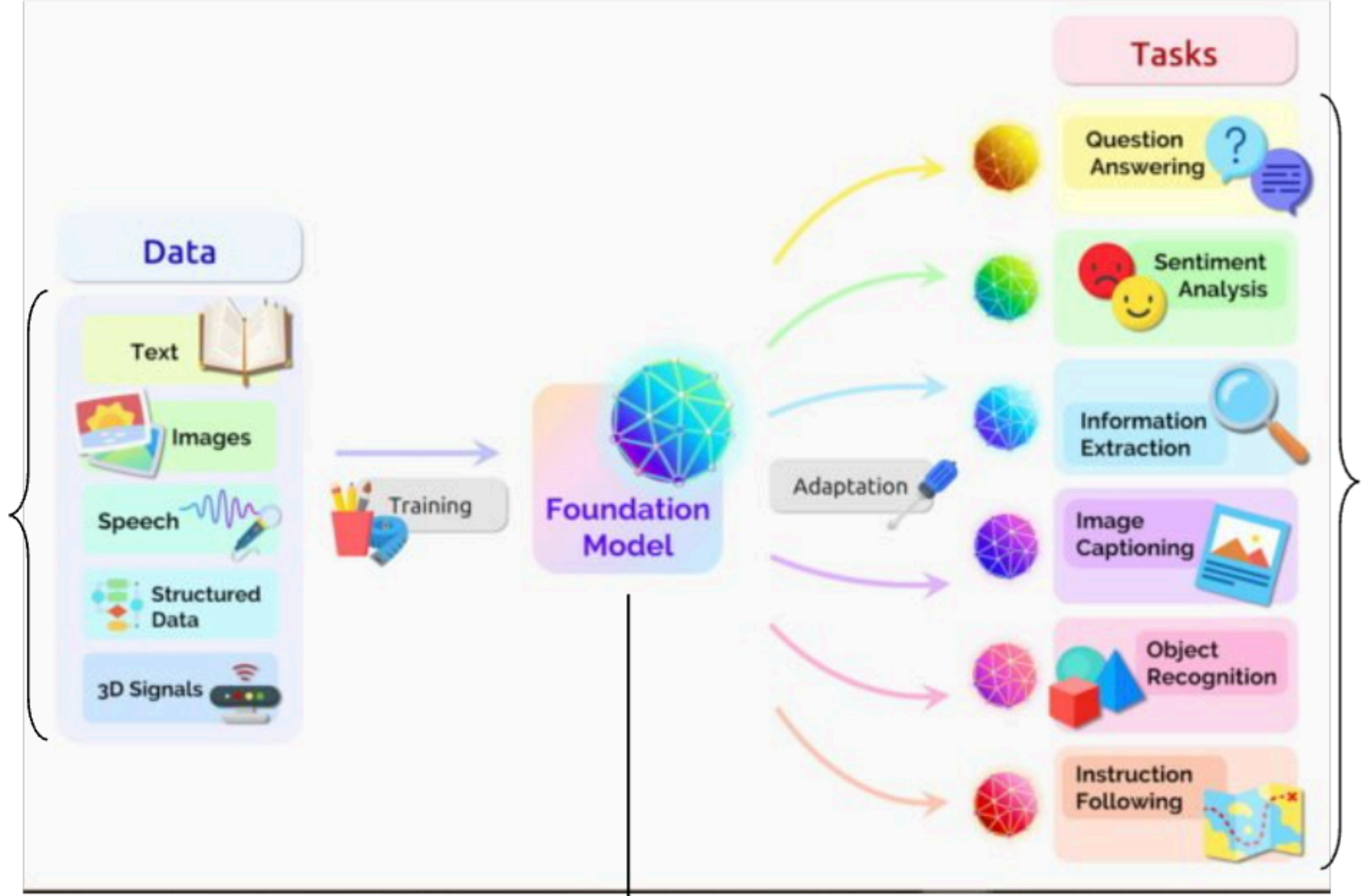
Pretraining stage is extremely expensive

Fine tuning is way cheaper than training for every task

Towards Foundation Models

Massive amounts of data

Semi-, **self-**, unsupervised training



https://blogs.nvidia.com/wp-content/uploads/2023/03/Transformer-apps.jpg

Can be adapted to **many** downstream tasks

Huge models
(hundreds of billions of parameters)

Can only be trained because of those massive datasets

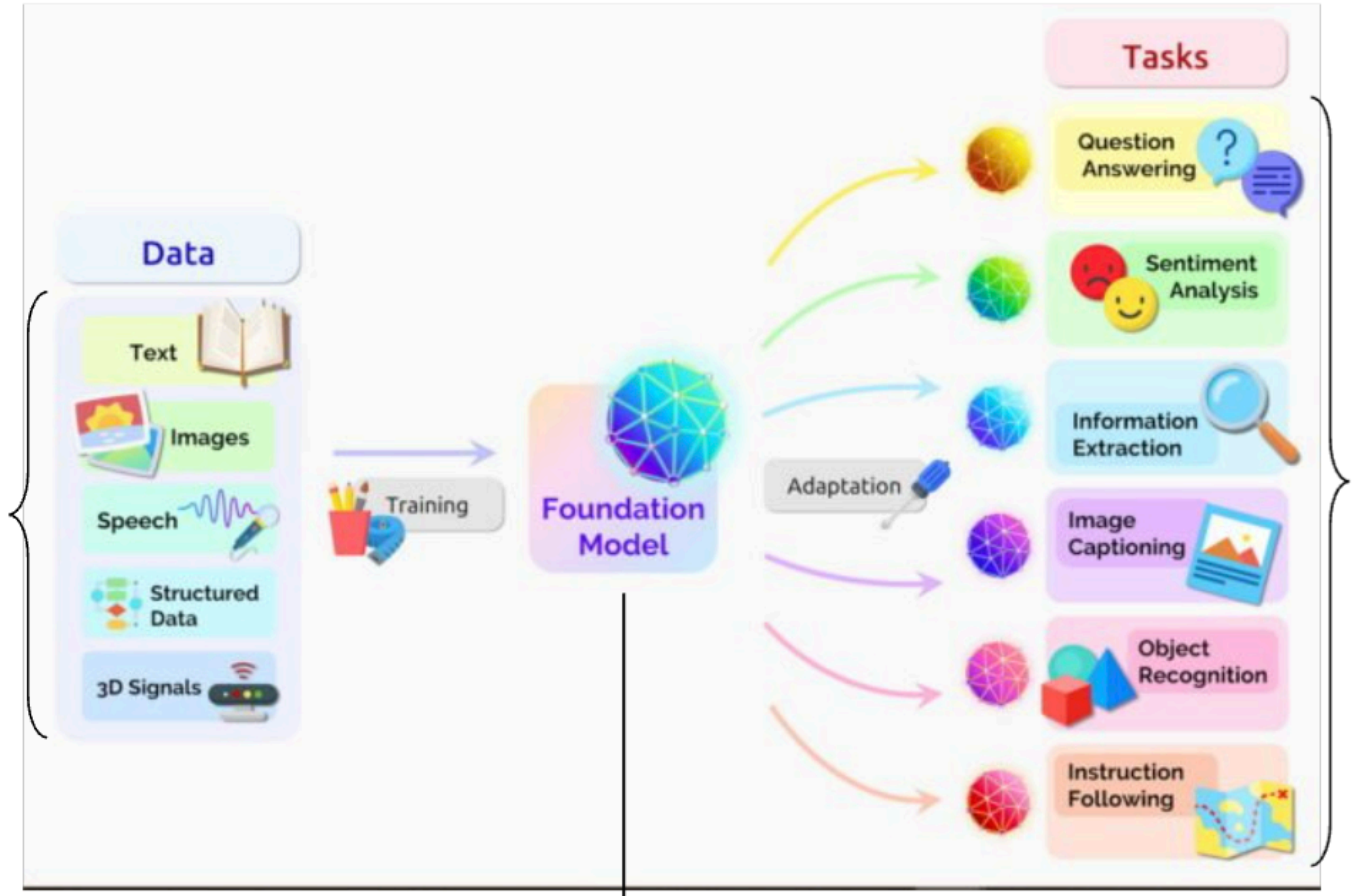
Pretraining stage is extremely expensive

Fine tuning is way cheaper than training for every task

Towards Foundation Models

Massive amounts of data

Semi-, **self-**, unsupervised training



https://blogs.nvidia.com/wp-content/uploads/2023/03/Transformer-apps.jpg

Can be adapted to **many** downstream tasks

Huge models
(hundreds of billions of parameters)

Pretraining stage is extremely expensive

Fine tuning is way cheaper than training for every task

Here we do need annotations, but not as much data anymore

- We can do standard supervised training starting from the FM
- We can leverage the models with the "N-shot strategies"*

*probably not an actual name

Leverage Foundation models with the "N-shot strategies"

Zero-Shot

```
Translate English to French:        task description
cheese =>                           prompt
```

Model predicts the answer with a natural language task description.

No Gradient updates

One-Shot

```
Translate English to French:        task description
sea otter => loutre de mer          example
cheese =>                           prompt
```

In addition to task description, the model sees a single example of the task.

No Gradient updates

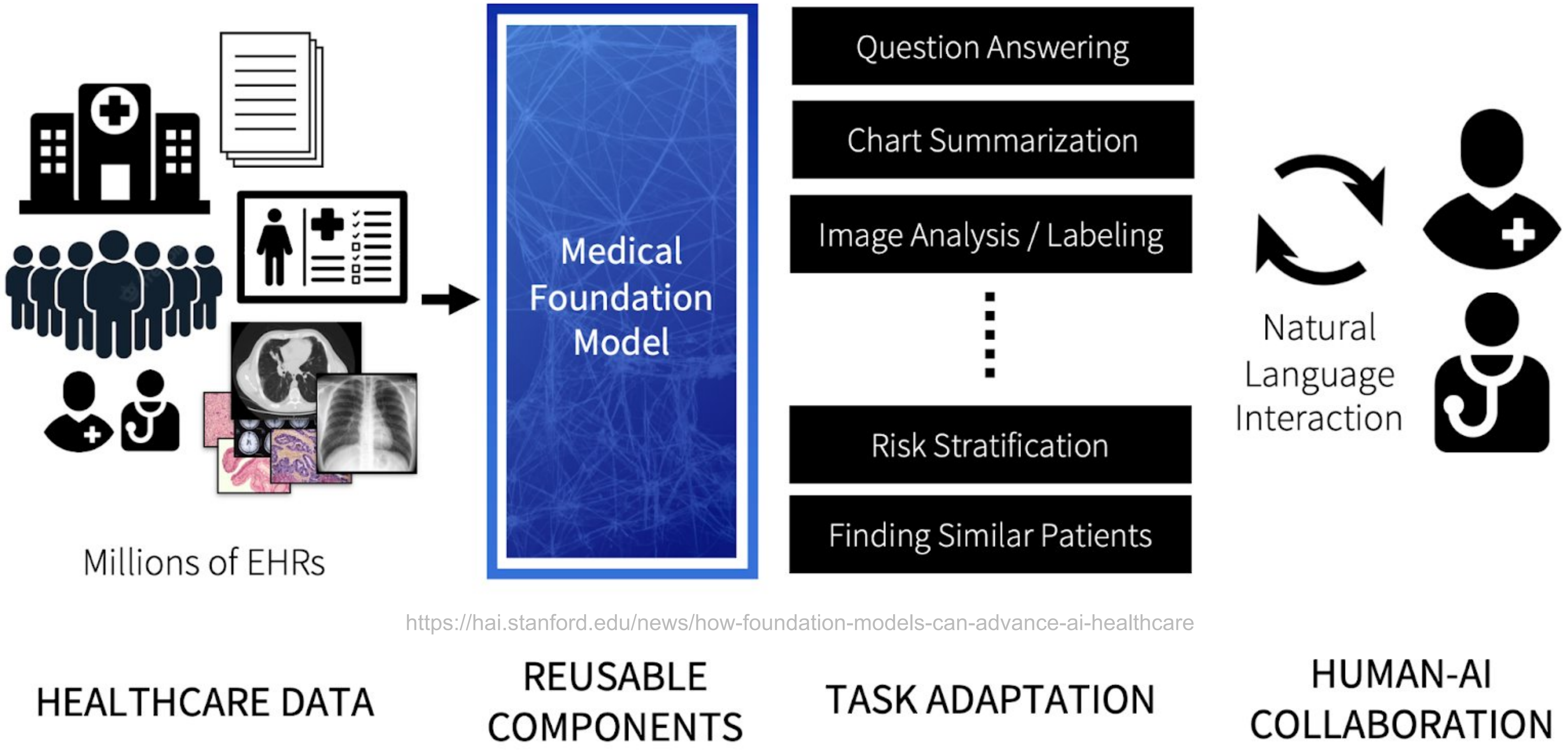
Few-Shot

```
Translate English to French:        task description
sea otter => loutre de mer          examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                           prompt
```

In addition to task description, the model sees a few examples of the task.

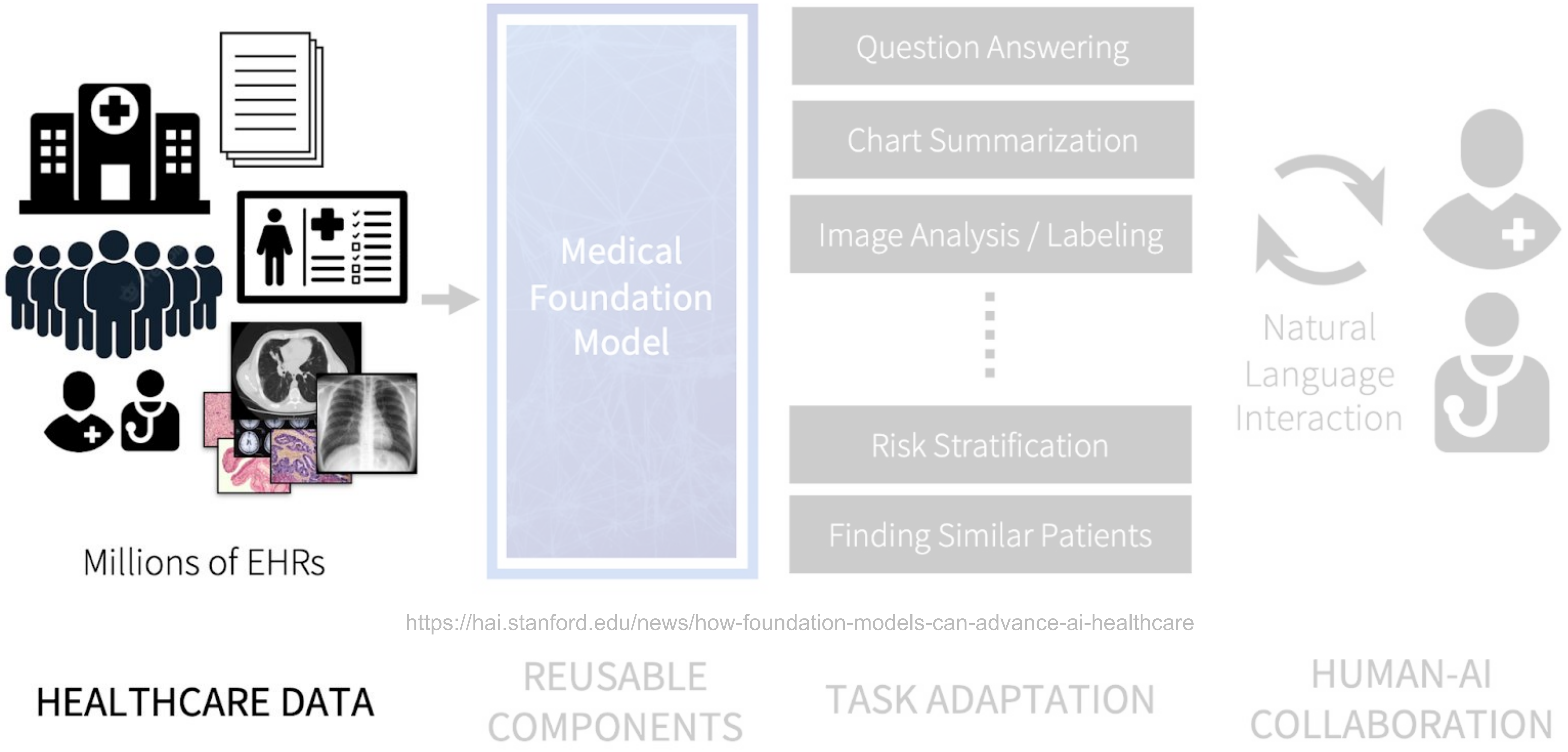
No Gradient updates

Towards Medical Foundation Models



https://hai.stanford.edu/news/how-foundation-models-can-advance-ai-healthcare

Towards Medical Foundation Models

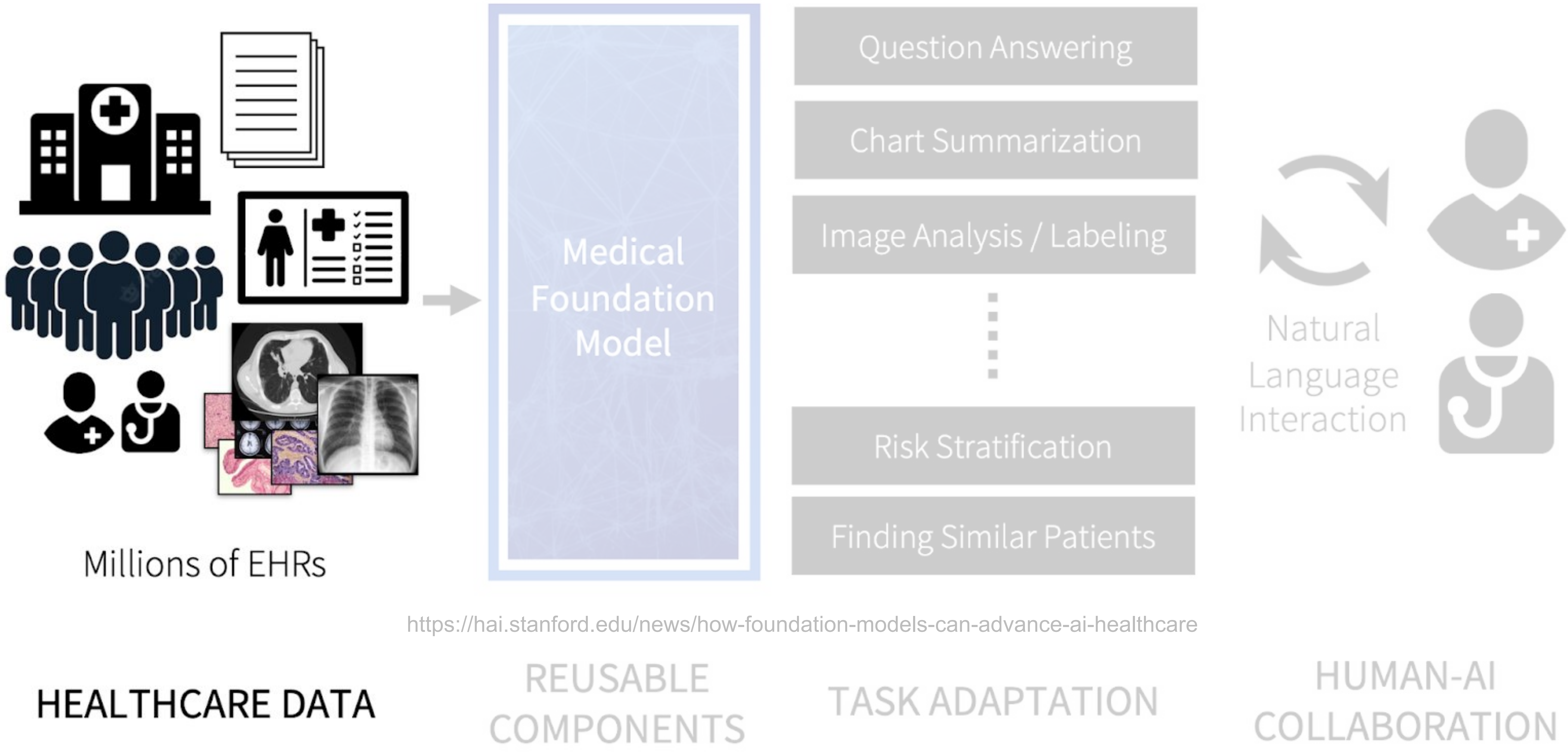


https://hai.stanford.edu/news/how-foundation-models-can-advance-ai-healthcare

Medical imaging can be very varied:

- 2D (x-rays, MR, histology, ultrasound) or 3D (CT, MR, ultrasound)
- Static images or videos (fMRI, cine MR, endoscopy, ultrasound)
- …

Towards Medical Foundation Models

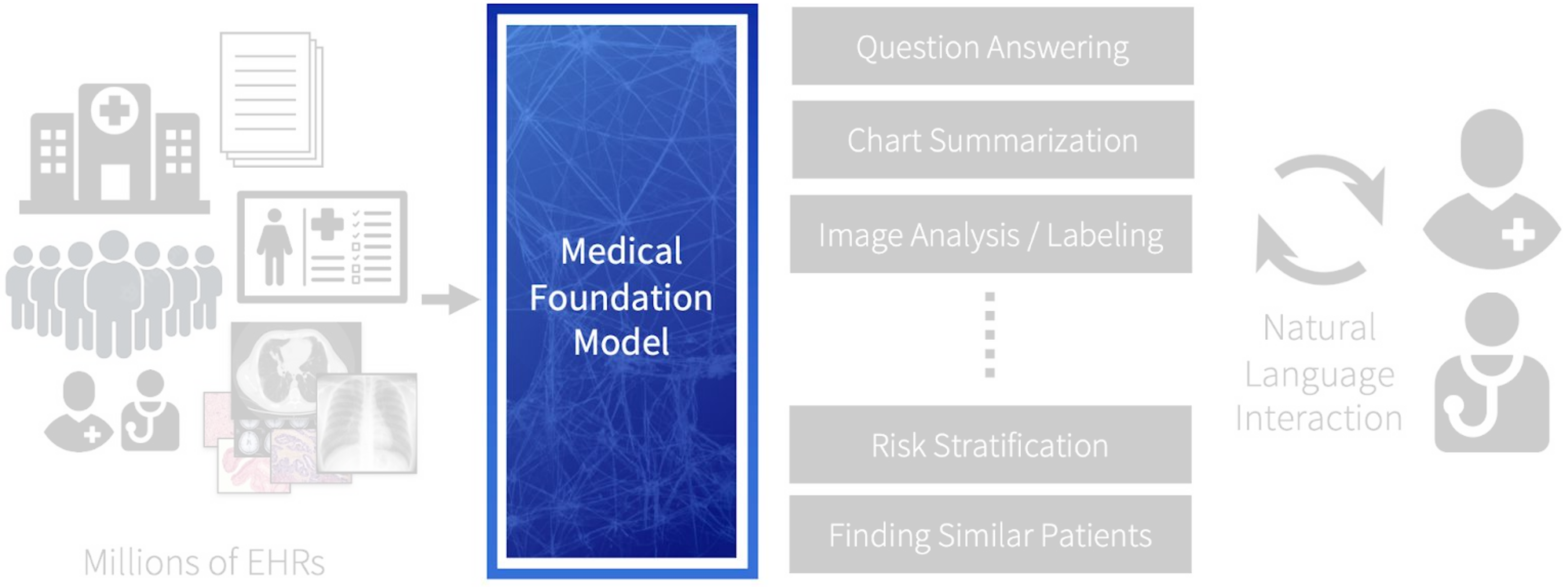


https://hai.stanford.edu/news/how-foundation-models-can-advance-ai-healthcare

Medical imaging can be very varied:
- 2D (x-rays, MR, histology, ultrasound) or 3D (CT, MR, ultrasound)
- Static images or videos (fMRI, cine MR, endoscopy, ultrasound)
- …

Medical data is much more difficult to collect and to annotate
- Smaller annotated datasets (hundreds o thousands of data)
- Highly unbalanced
- Small regions of interest

Towards Medical Foundation Models

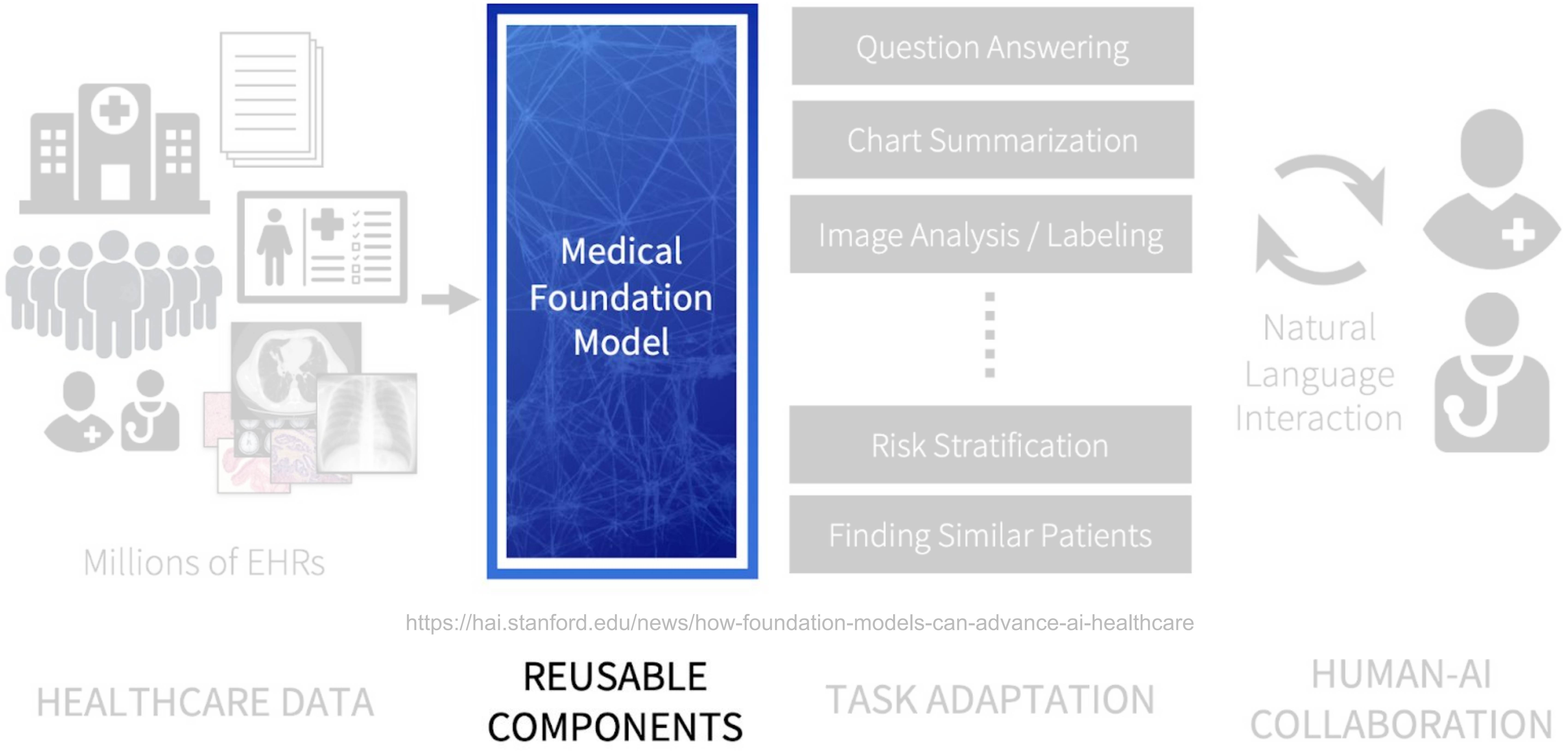


https://hai.stanford.edu/news/how-foundation-models-can-advance-ai-healthcare

Most existing big models are trained for 2D natural images. For the medical domain:

- Should we have a model per modality? Per spatial/temporal dimensionality? Number of channels?
- How do we get enough data to train them?

Towards Medical Foundation Models



Most existing big models are trained for 2D natural images. For the medical domain:

- Should we have a model per modality? Per spatial/temporal dimensionality? Number of channels?
- How do we get enough data to train them?

With SSL we don't need annotations

- Larger datasets are available. Even medical!

Maybe we don't need to start from scratch:

- How can we use the existing natural domain models as a starting point?

Why adaptation is feasible than training with all the images?

Foundation models serve as a promising backbone**:**

(i) Due to privacy reasons, it’s not easy to obtain billion-scale data

(ii) How can we leverage open source models that have already been exposed to web-scale natural images

(iii) Changing the model or its predictions is easier during inference (Adaptation):

Why adaptation is feasible than training with all the images?

Foundation models serve as a promising backbone:

(i) Data Privacy & Regulatory Constraints

Regulations prevent centralized billion-scale medical data aggregation

Foundation models bypass the need for massive proprietary datasets

(ii) Leverage Pre-trained Web-Scale Knowledge

Open-source models (SAM, DINOv2, CLIP) already encode rich visual representations from billions of natural images

Transfer universal features to medical domain with minimal task-specific data

(iii) Flexible Test/Inference-Time Adaptation

Update model behavior through lightweight adaptation (LoRA, adapters) without full retraining

Enable rapid customization for new imaging protocols or disease patterns

Papers for Teaching and Presentation (SSL) (Papers of your choice are fine too)

- Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PmLR, 2020.

- Caron and et al. Emerging properties in self-supervised vision transformers. ICCV, 2021

- Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *Proceedings of naacL-HLT*. Vol. 1. No. 2. 2019.

- K. He and et al. Masked autoencoders are scalable vision learners. CVPR, 2022.

- Huang, Ziyan, et al. "Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training." *arXiv preprint arXiv:2304.06716* (2023).

- Wald, Tassilo, et al. "An OpenMind for 3D medical vision self-supervised learning." *arXiv preprint arXiv:2412.17041* (2024).

Papers for Teaching and Presentation (FM) (Papers of your choice are fine too)

- R. Bommasani and et al. On the opportunities and risks of foundation models. arXiv, 2021.
- A. Radford and et al. Learning transferable visual models from natural language supervision. ICML, 2021
- Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
- Z. Liu and et al. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021
- Ma and et al. Medsam: Segment anything model for medical images. arXiv, 2023
- Zhao, Theodore, et al. "Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once." *arXiv preprint arXiv:2405.12971* (2024).

Papers for Teaching and Presentation (Adaptation) (Papers of your choice are fine too)

- Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models. arXiv 2021." ICLR 2022
- Veličković, Petar, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. "Softmax is not Enough (for Sharp Size Generalisation)." ICML 2025
- Ge and et al. Domain adaptation via prompt learning. IEEE TNNLS, 2023
- Y. Zhang and et al. Biomedclip: Open biomedical contrastive language-image pretraining. arXiv, 2023
- Hoopes, Andrew, et al. "Voxelprompt: A vision-language agent for grounded medical image analysis." *arXiv preprint arXiv:2410.08397* (2024).
- Yuan, Zheng, et al. "Improving biomedical pretrained language models with knowledge." *arXiv preprint arXiv:2104.10344* (2021).
- Remy, François, Kris Demuynck, and Thomas Demeester. "Biolord: Learning ontological representations from definitions (for biomedical concepts and their textual descriptions)." *arXiv preprint arXiv:2210.11892* (2022).