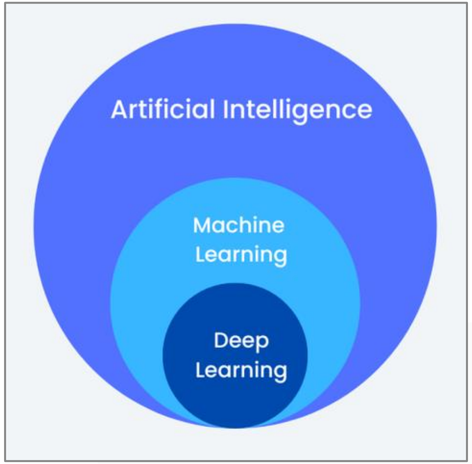
Symposium:
Qu’est-ce que l’IA ? Quels enjeux pour la Recherche ?

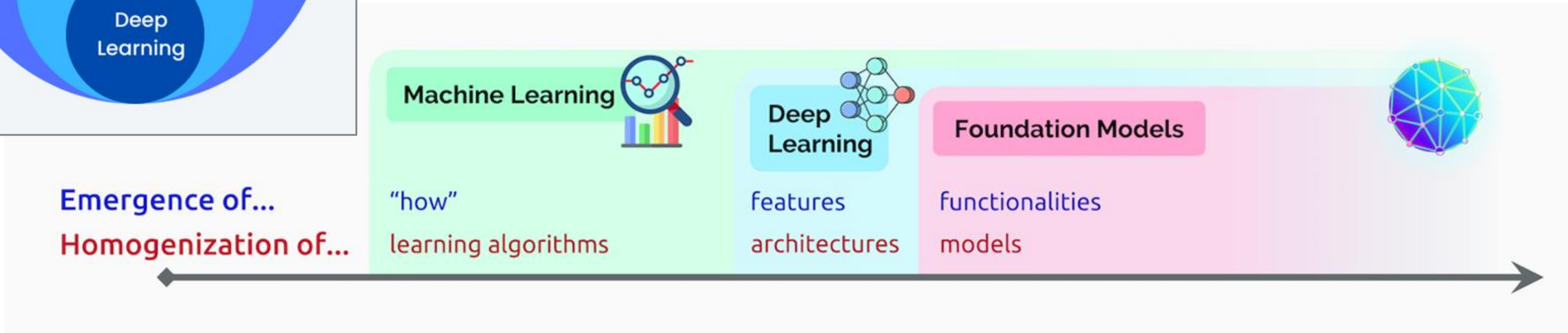
From ML to foundation models

Ana M. Barragán Montero

20/05/2025

From ML to foundation models





From Bommasani et al. 2022

What is a foundation model?

ChatGPT

LLaMA

deepseek

Also called **LLM (Large Language Models)**
when they deal with language only … But many of them **multimodal** now!

What is a foundation model?

arXiv
https://arxiv.org › pdf PDF

On the Opportunities and Risks of Foundation Models

by R Bommasani · 2021 · Cited by 5300 — This report provides a thorough account of the **opportunities and risks of foundation models**, ranging from their capabilities (e.g., language, ...

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

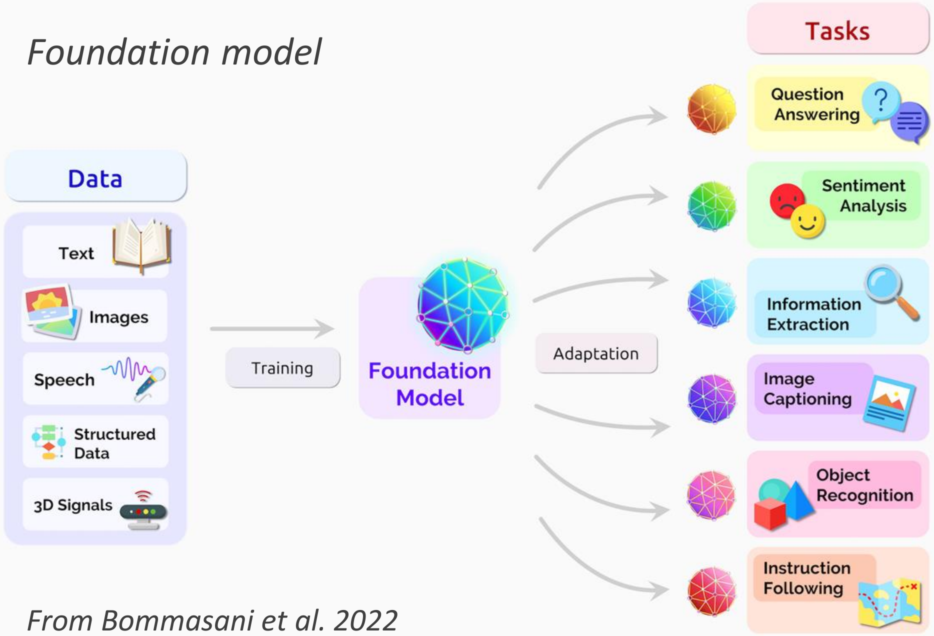
arXiv:2108.07258v3 [cs.LG] 12 Jul 2022

214 pages!

What is a foundation model?

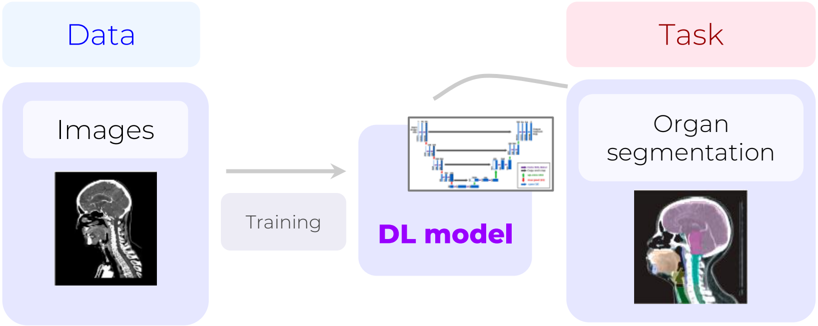
*“A foundation model is any model that is **trained on broad data** (generally using self-supervision at scale) that **can be adapted** (e.g., fine-tuned) **to a wide range of downstream tasks**”*

Foundation model



From Bommasani et al. 2022

Regular (single-task) model



What is a foundation model?

*"A foundation model is any model that is **trained on broad data** (generally using self-supervision at scale) that **can be adapted** (e.g., fine-tuned) **to a wide range of downstream tasks**"*

Foundation model

General AI

nature

Explore content | **About the journal** | **Publish with us**

nature > perspectives > article

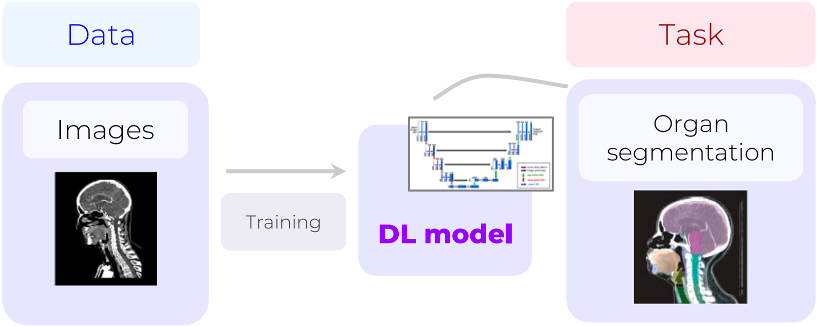
Perspective | Published: 12 April 2023

Foundation models for generalist medical artificial intelligence

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol & Pranav Rajpurkar

Regular (single-task) model

Narrow AI

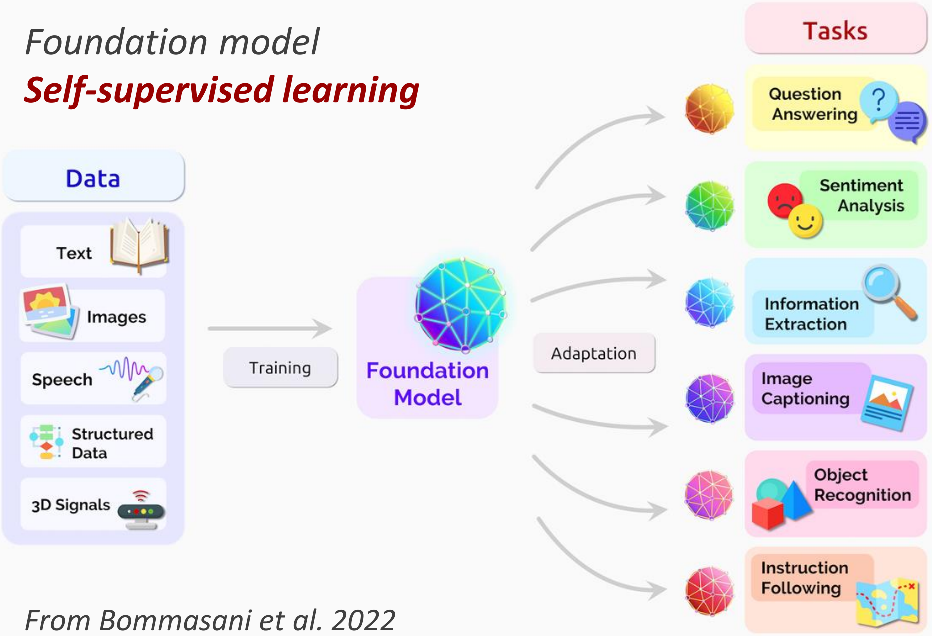


What is a foundation model?

*"A foundation model is any model that is **trained on broad data** (generally using self-supervision at scale) that **can be adapted** (e.g., fine-tuned) **to a wide range of downstream tasks**"*

Foundation model

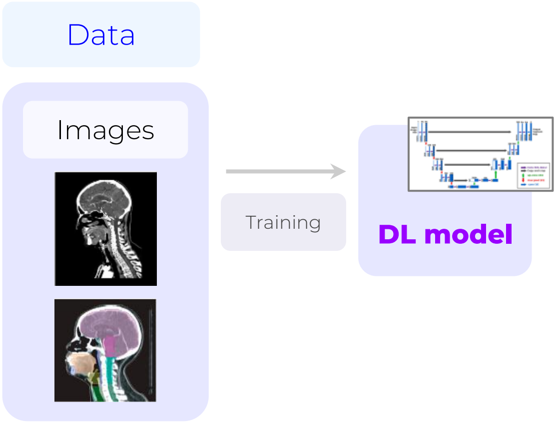
Self-supervised learning



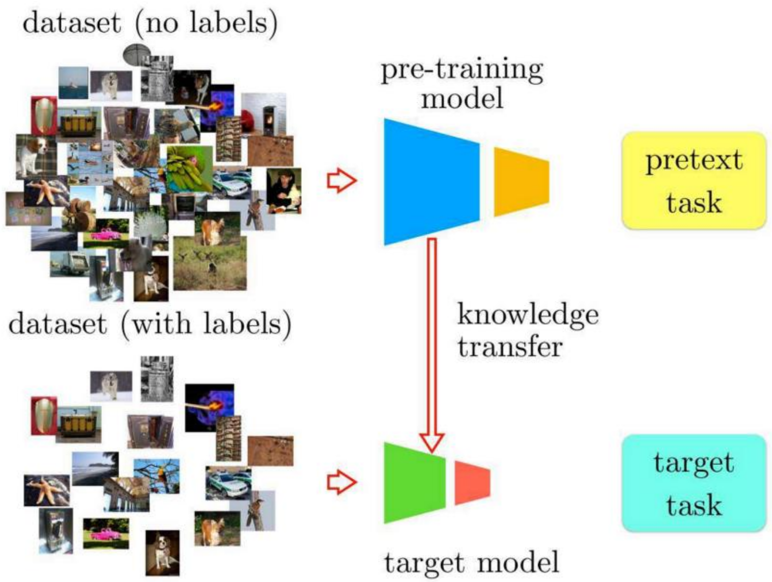
From Bommasani et al. 2022

Regular (single-task) model

Supervised learning

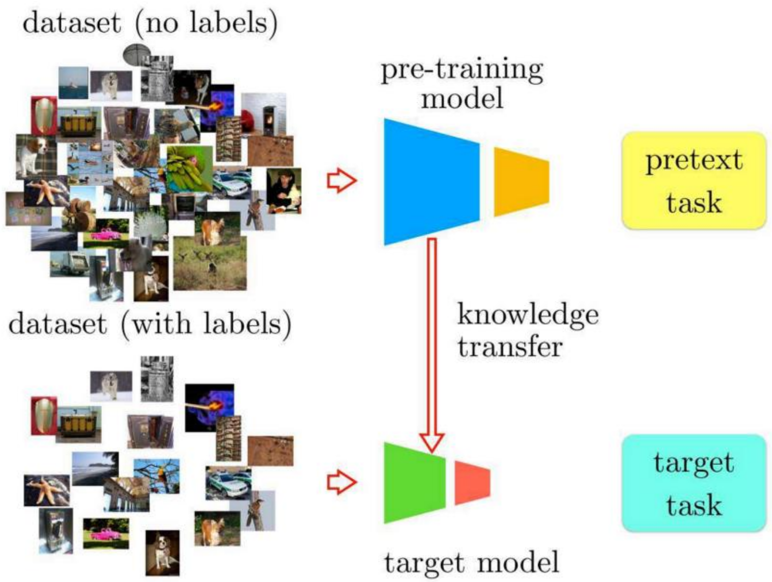


Self-supervision vanilla model

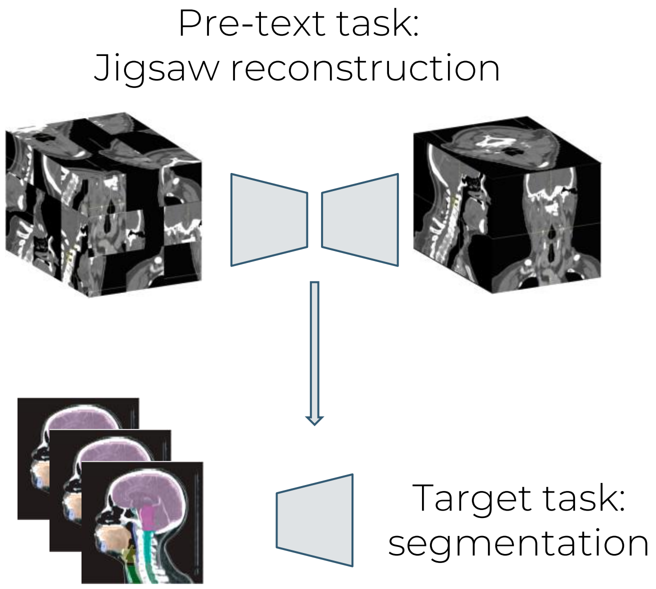


From Noroozi et al, 2018 **"Boosting Self-Supervised Learning via Knowledge Transfer"**

Self-supervision vanilla model



From Noroozi et al, 2018 **"Boosting Self-Supervised Learning via Knowledge Transfer"**



Taleb et al., 2020, **3D Self-Supervised Methods for Medical Imaging**

Self-supervision: transformers (ViT)

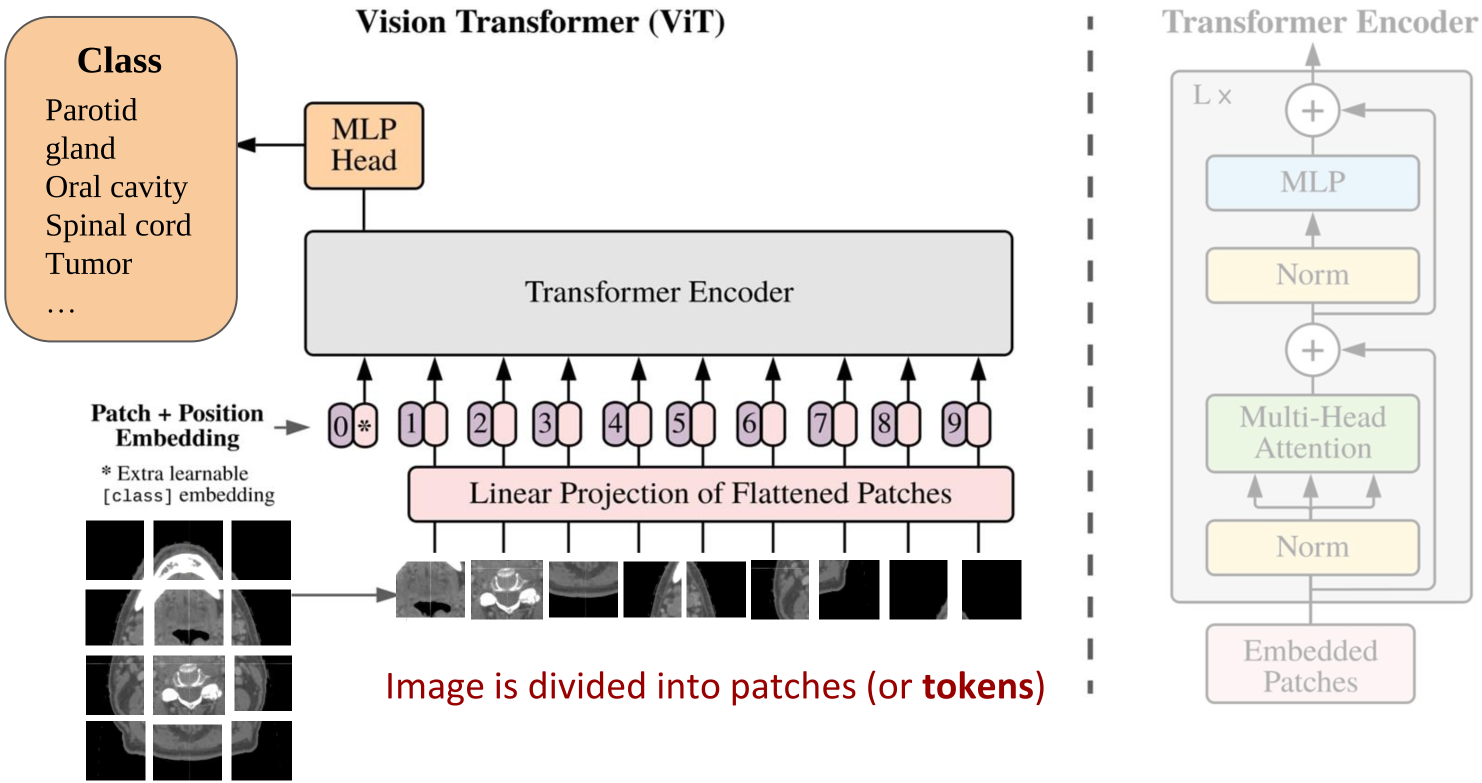
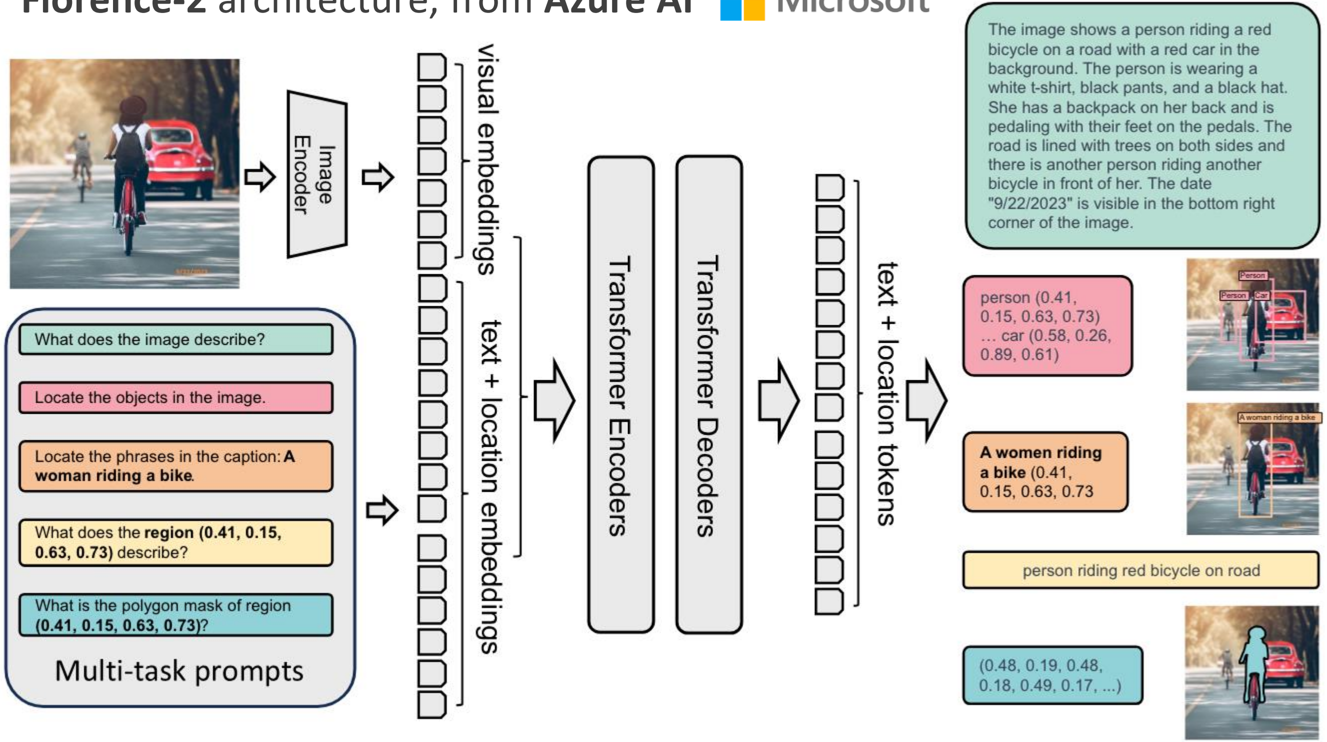


Image is divided into patches (or **tokens**)

Adapted from **Dosovitskiy et al. 2021**, “An image is worth 16x16 words”

Towards multimodal self-supervision

Florence-2 architecture, from **Azure AI** Microsoft



Potential of foundation models

- *Self-supervision* – less need of annotated data for the fine-tuned (target) task
- *Zero-shot learning (in-context learning)*
  Classify new, unseen categories without requiring any specific examples of those categories during training

Applications in research?
Examples from the medical field

SAM (Segment Anything Model)

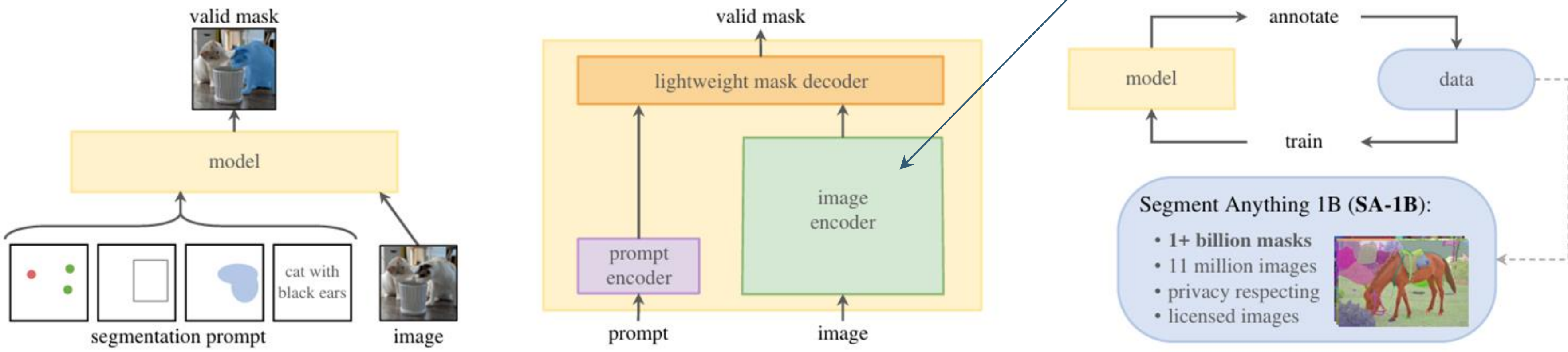
Meta AI

arXiv
https://arxiv.org › cs

[2304.02643] Segment Anything

by A Kirillov · 2023 · Cited by 10376 — We introduce the **Segment Anything (SA) project**: a new task, model, and dataset for image segmentation. Using our efficient model in a data ...

Transformer-based



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

MedSAM (fine-tune SAM on med. images)


nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

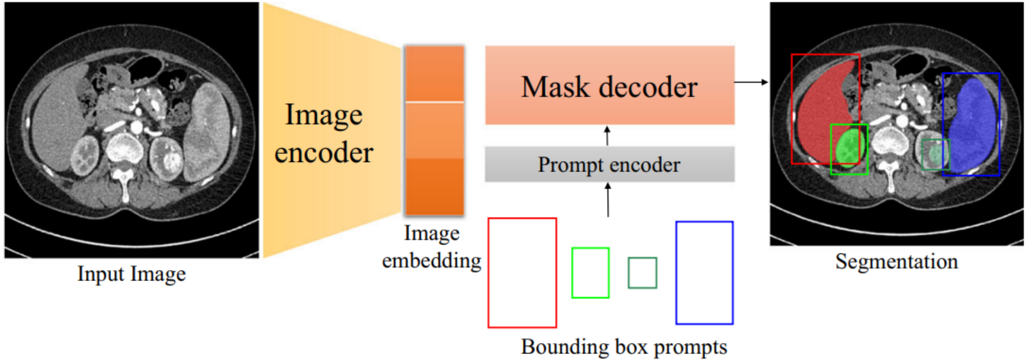
nature > nature communications > articles > article

Article | Open access | Published: 22 January 2024

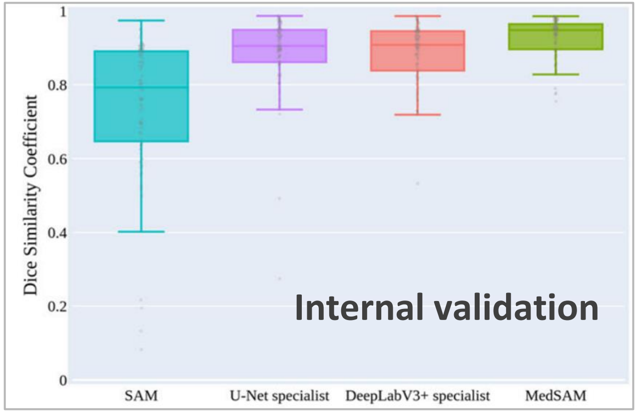
Segment anything in medical images

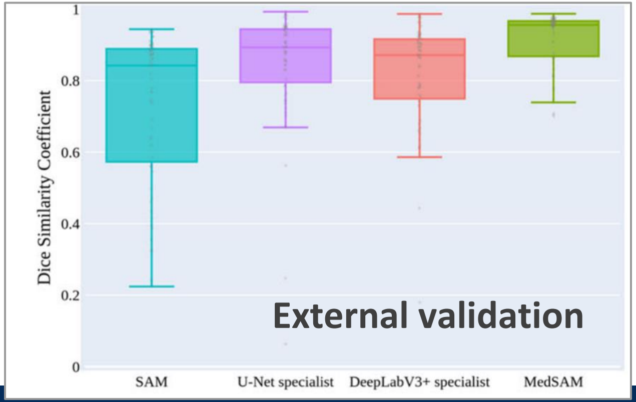
Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You & Bo Wang

- 1.5M image-mask pairs
- 10 imaging modalities
- 30 cancer types
- Internal and external task validation
- Open-source github.com/bowang-lab/MedSAM



MedSAM (fine-tune SAM on med. images)





- 1.5M image-mask pairs
- 10 imaging modalities
- 30 cancer types
- **Internal and external task validation**
- Open-source github.com/bowang-lab/MedSAM

UNet not really used as current practice in RT:
! bounding boxes in each 2D slice
! one model per image modality (e.g. CT)

UNet versus foundation models

20th International Conference on the use of Computers in Radiation therapy 8 - 11 July 2024, Lyon, France

Segment anything model for head and neck tumor segmentation with CT, PET and MRI multi-modality images

Jintao Ren[1,2], Mathis Rasmussen[1,2], Jasper Nijkamp[1,2], Jesper Grau Eriksen[1,3] and Stine Korreman[1,2]

[1]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[2]Danish Center for Particle Therapy, Aarhus University Hospital, Aar...
[3]Department of Experimental Clinical Oncology, Aarhus University Hospita...

MEDICAL PHYSICS
The International Journal of Medical Physics Research and Practice

TECHNICAL NOTE | Full Access

Technical note: Generalizable and promptable artificial intelligence model to augment clinical delineation in radiation oncology

Lian Zhang, Zhengliang Liu, Lu Zhang, Zihao Wu, Xiaowei Yu, Jason Holmes, Hongying Feng, Haixing Dai, Xiang Li, Quanzheng Li, William W. Wong, Sujay A. Vora, Dajiang Zhu, Tianming Liu, Wei Liu

First published: 06 February 2024 | https://doi.org/10.1002/mp.16965

Strahlentherapie und Onkologie (2025) 201:255–265
https://doi.org/10.1007/s00066-024-02313-8

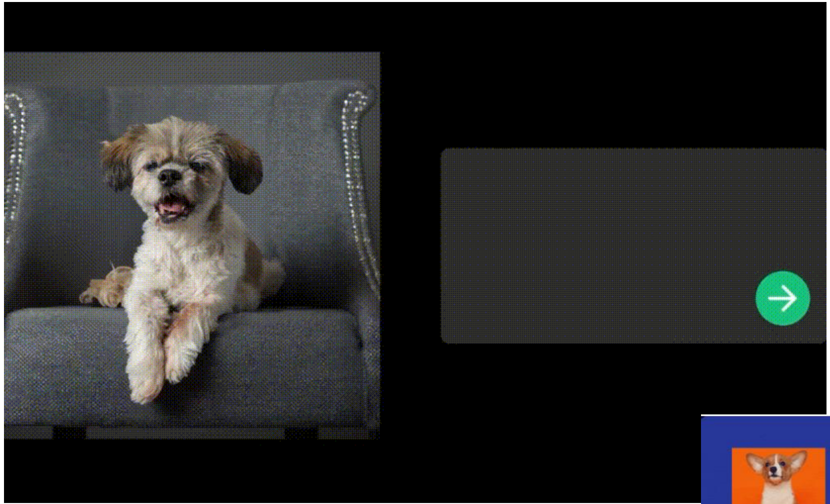
ORIGINAL ARTICLE

The Segment Anything foundation model achieves favorable brain tumor auto-segmentation accuracy in MRI to support radiotherapy treatment planning

Florian Putz[1,2,3] · Sogand Beirami[1,2] · Manuel Alexander Schmidt[2,3,4] · Matthias Stefan May[2,3,5] · Johanna Grigo[1,2] · Thomas Weissmann[1,2] · Philipp Schubert[1,2] · Daniel Höfler[1,2] · Ahmed Gomaa[1,2] · Ben Tkhayat Hassen[1,2] · Sebastian Lettmaier[1,2] · Benjamin Frey[1,2] · Udo S. Gaipl[1,2] · Luitpold V. Distel[1,2] · Sabine Semrau[1,2] · Christoph Bert[1,2] · Rainer Fietkau[1,2,3] · Yixing Huang[1,2]

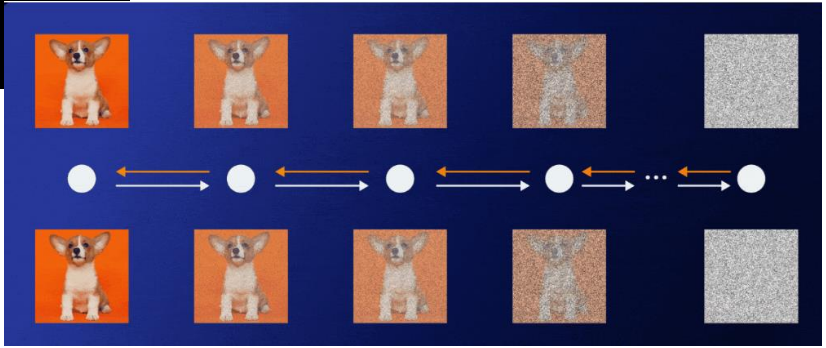
Received: 15 September 2024 / Accepted: 22 September 2024 / Published online: 6 November 2024

Synthetic image generation



DALL·E 2

Diffusion model



Text conditioned!

Synthetic image generation

nature medicine

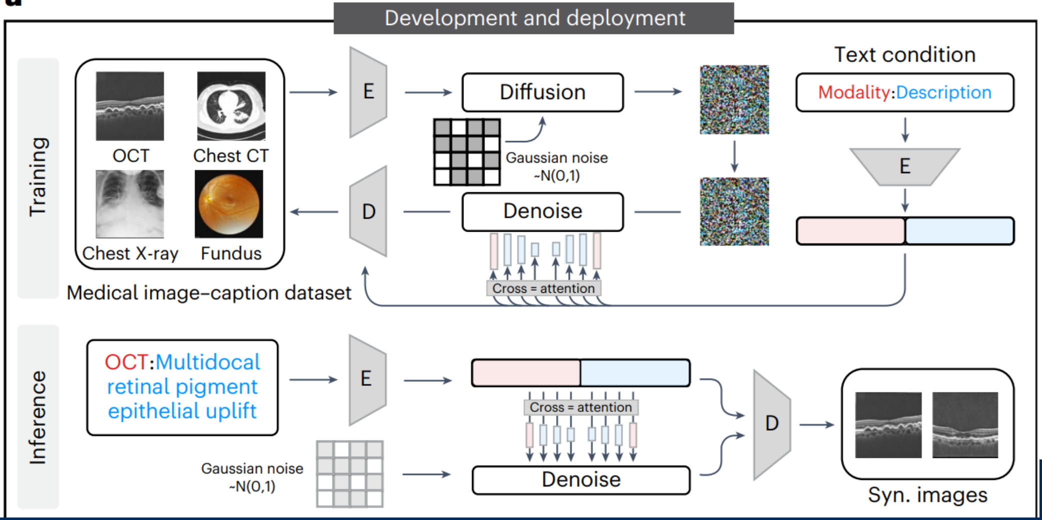
Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature medicine › articles › article

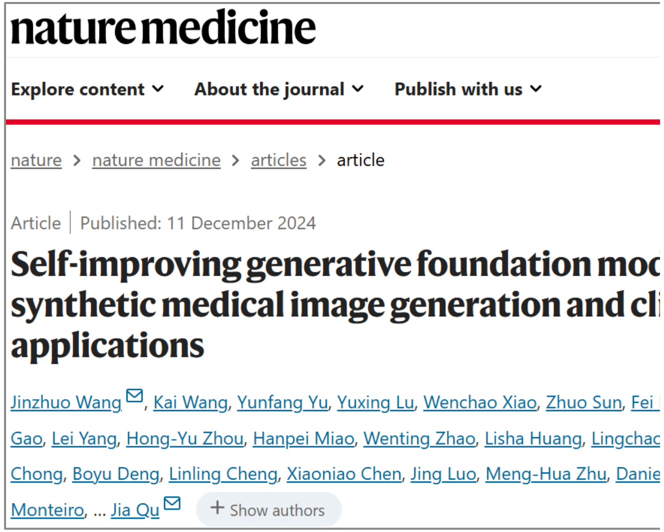
Article | Published: 11 December 2024

Self-improving generative foundation model for synthetic medical image generation and clinical applications

Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fe
Gao, Lei Yang, Hong-Yu Zhou, Hanpei Miao, Wenting Zhao, Lisha Huang, Lingcha
Chong, Boyu Deng, Linling Cheng, Xiaoniao Chen, Jing Luo, Meng-Hua Zhu, Dan
Monteiro, ... Jia Qu + Show authors



Synthetic image generation


nature medicine

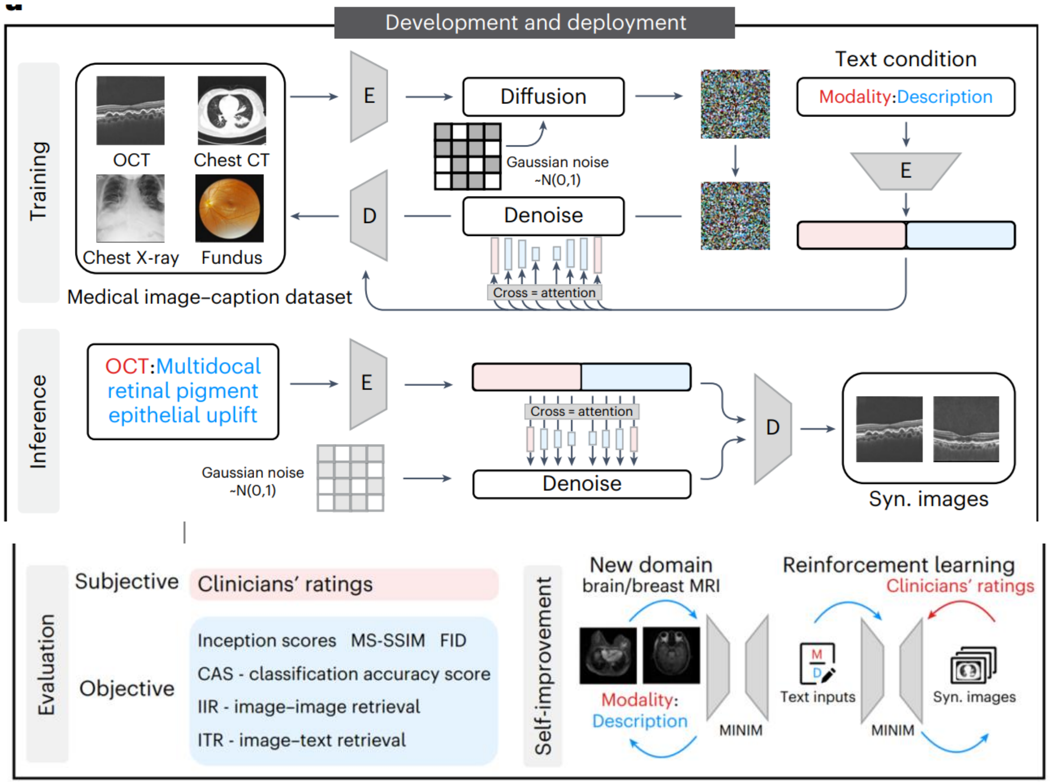
Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature medicine > articles > article

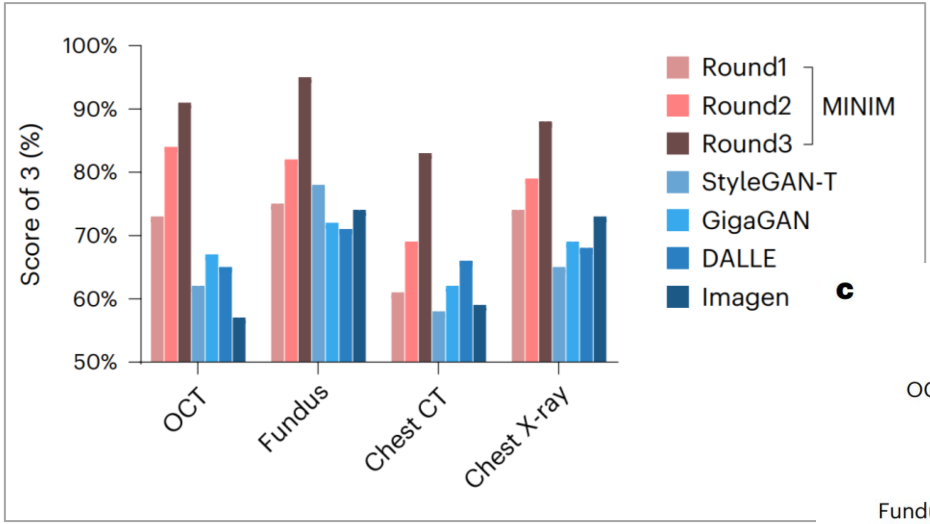
Article | Published: 11 December 2024

Self-improving generative foundation moc synthetic medical image generation and cli applications

Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Gao, Lei Yang, Hong-Yu Zhou, Hanpei Miao, Wenting Zhao, Lisha Huang, Lingchac Chong, Boyu Deng, Linling Cheng, Xiaoniao Chen, Jing Luo, Meng-Hua Zhu, Danie Monteiro, ... Jia Qu + Show authors



Synthetic image generation



nature medicine

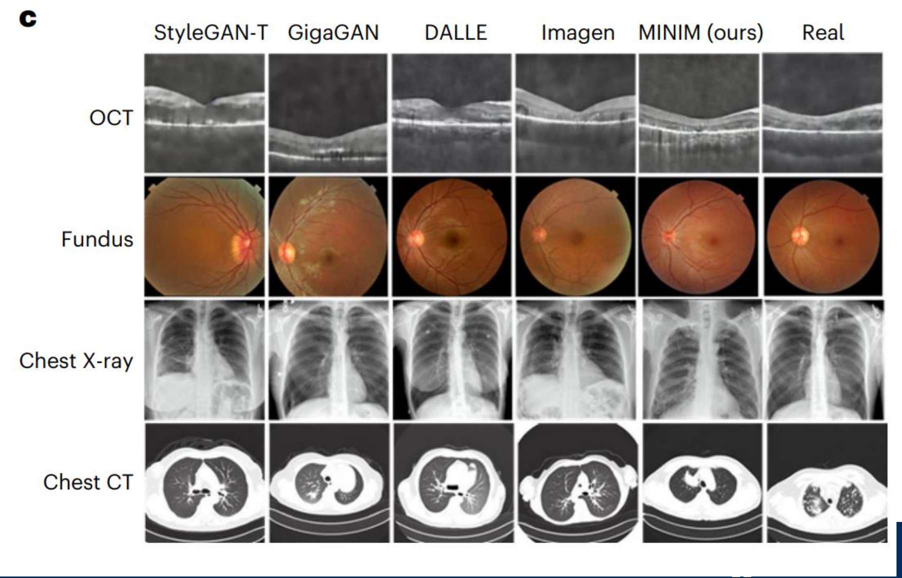
Explore content About the journal Publish with us

nature > nature medicine > articles > article

Article | Published: 11 December 2024

Self-improving generative foundation model for synthetic medical image generation and clinical applications

Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Liu, Zixing Zou, Yuanxu Gao, Lei Yang, Hong-Yu Zhou, Hanpei Miao, Wenting Zhao, Lisha Huang, Lingchao Zeng, Rui Guo, Ieng Chong, Boyu Deng, Linling Cheng, Xiaoniao Chen, Jing Luo, Meng-Hua Zhu, Daniel Baptista-Hon, Olivia Monteiro, ... Jia Qu + Show authors



Virtual imaging trials

MEDICAL PHYSICS
The International Journal of Medical Physics Research and Practice

REVIEW ARTICLE

Toward widespread use of virtual trials in medical imaging innovation and regulatory science

Ehsan Abadi, Bruno Barufaldi, Miguel Lago, Andreu Badal, Claudia Mello-Thoms, Nick Bottenus, Kristen A. Wangerin, Mitchell Goldburgh, Lawrence Tarbox ... See all authors

First published: 06 October 2024 | https://doi.org/10.1002/mp.17442 | Citations: 6

Collecting real images (for planning studies, clinical trials, etc.) involves:

- High cost
- Time-consuming
- Privacy concerns issues for data sharing

AAPM Task Group (TG387)
Requirements for VIT

- Diverse and realistic digital patient representations,
- Integration of physics and biology
- Development of **robust validation frameworks**

arXiv > cs > arXiv:2308.02463

Search...

Help | A

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 4 Aug 2023 (v1), last revised 16 Nov 2023 (this version, v5)]

Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, Weidi Xie

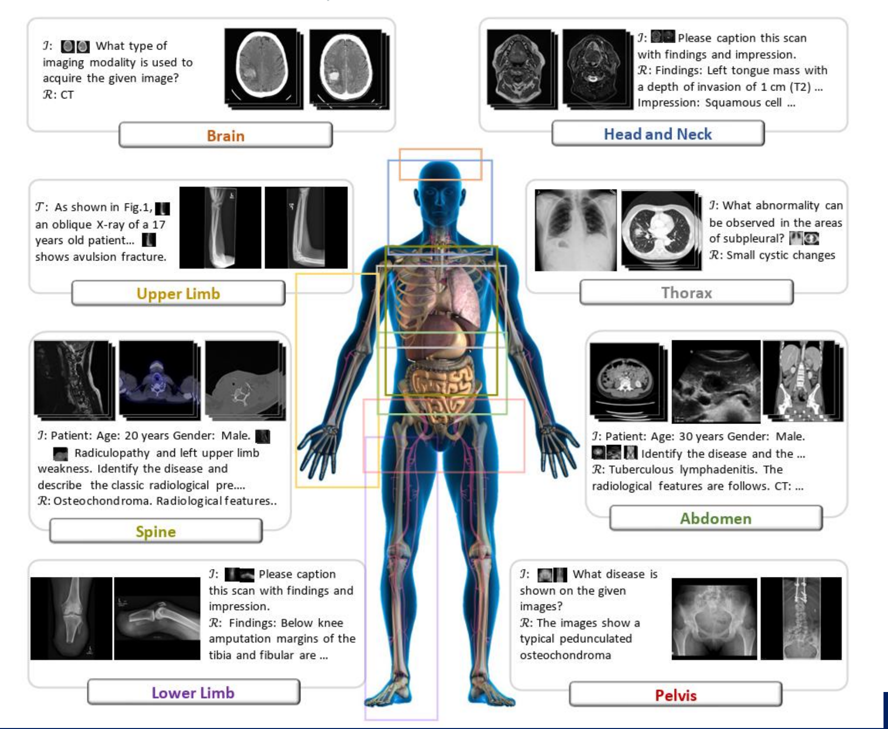
arXiv > cs > arXiv:2406.06512

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 10 Jun 2024]

Merlin: A Vision Language Foundation Model for 3D Computed Tomography

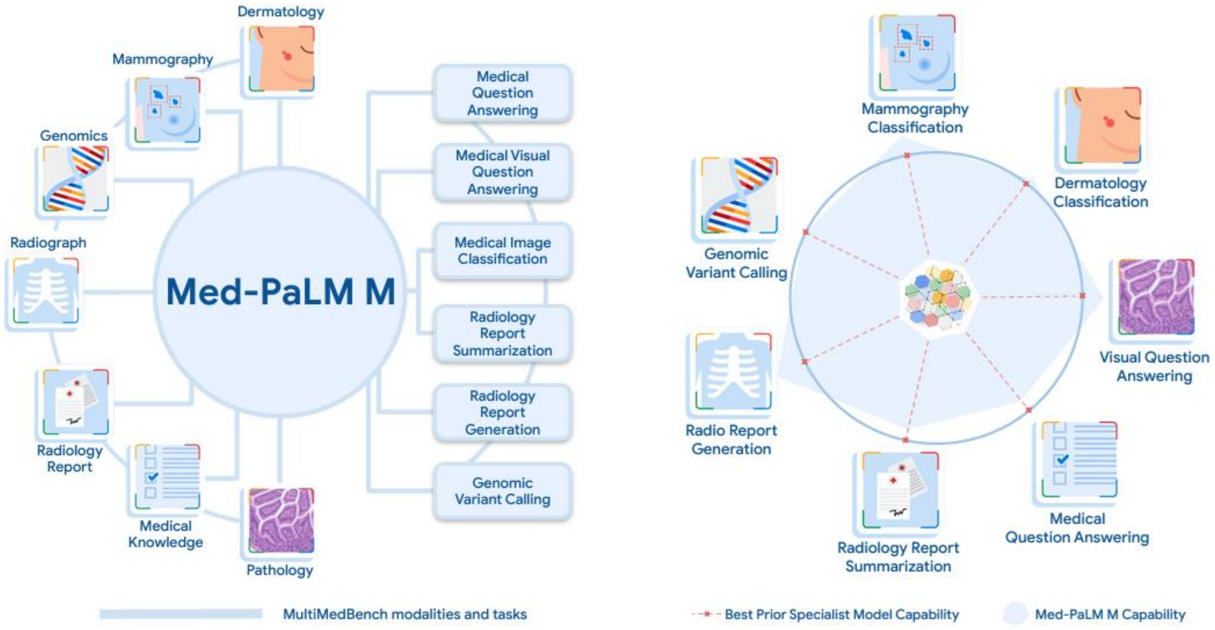
Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschal
Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, Christian Bluethgen, Malte Engmann Kjeldskov Jensen, Sophie Ost
Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson A
Neera Ahuja, Jason Fries, Nigam H. Shah, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jas
Gatidis, Akshay S. Chaudhari



Towards Generalist Biomedical AI

Tao Tu[*, ‡, 1], Shekoofeh Azizi[*, ‡, 2],
Danny Driess[2], Mike Schaekermann[1], Mohamed Amin[1], Pi-Chuan Chang[1], Andrew Carroll[1],
Chuck Lau[1], Ryutaro Tanno[2], Ira Ktena[2], Basil Mustafa[2], Aakanksha Chowdhery[2], Yun Liu[1],
Simon Kornblith[2], David Fleet[2], Philip Mansfield[1], Sushant Prakash[1], Renee Wong[1], Sunny Virmani[1],
Christopher Semturs[1], S Sara Mahdavi[2], Bradley Green[1], Ewa Dominowska[1], Blaise Aguera y Arcas[1],
Joelle Barral[2], Dale Webster[1], Greg S. Corrado[1], Yossi Matias[1], Karan Singhal[1], Pete Florence[2],
Alan Karthikesalingam[†, ‡, 1] and Vivek Natarajan[†, ‡, 1]

[1]Google Research, [2]Google DeepMind



Let’s step back and think

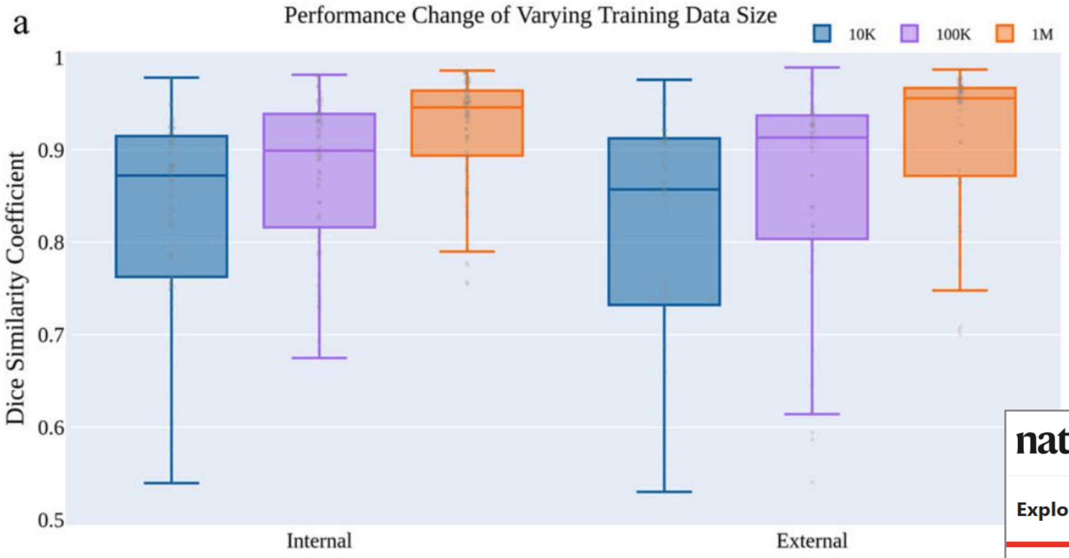
Let’s step back and think

- Model and dataset size, governance
- Computational & Environmental Costs
- Transparency and reproducibility
- Testing and reliability

Let’s step back and think

- **Model and dataset size, governance**
- Computational & Environmental Costs
- Transparency and reproducibility
- Testing and reliability

How much data we need for a FM?



- *Need less annotated data but much more non-annotated data to reach a good performance*
- *But probably OK because pre-train only once?*


nature communications

Explore content About the journal Publish with us

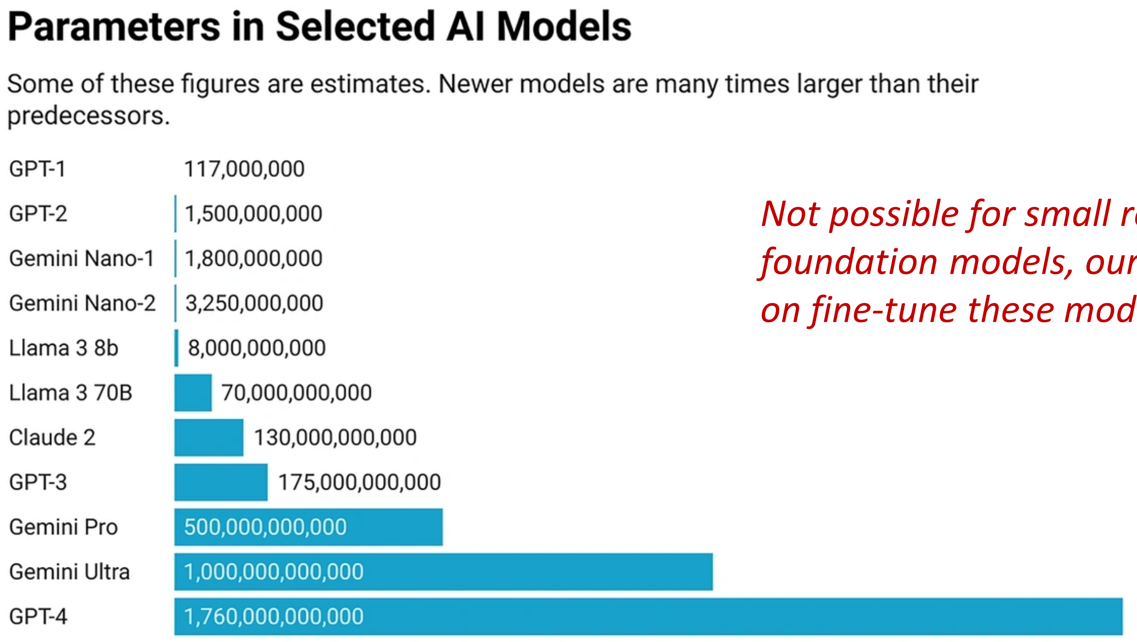
nature > nature communications > articles > article

Article | Open access | Published: 22 January 2024

Segment anything in medical images

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You & Bo Wang

Increasing number of parameters



Not possible for small research labs to create foundation models, our work will be restricted on fine-tune these models for our taks

https://explodingtopics.com/blog/gpt-parameters

Some examples 2024 / 2025

Model	Creator(s)	Type	Date	Open source	Params
LLaMA 4	Meta	Multimodal	2025	Yes	400B (MoE)
DeepSeek V3.1	DeepSeek AI	Multimodal	2025	Yes	560 B
Phi-4	Microsoft	Multimodal	2025	Yes	5.6B
Florence-2	Microsoft	Multimodal	2024	Yes	0.8B
Qwen2.5-VL	Alibaba	Multimodal	2025	Yes	72B
Pixtral Large	Mistral AI	Multimodal	2024	Yes	124B
Gemini 2.5	Google DeepMind	Multimodal	2025	No	128B x 16 (MoE)
GPT-4o	OpenAI	Multimodal	2024	No	1.8T (MoE)
Claude 3.5	Anthropic	Multimodal	2024	No	175B

*MoE = mixture of experts

*More at **github.com/uncbiag/Awesome-Foundation-Models***

FM versus classical UNet

Foundation model

General AI

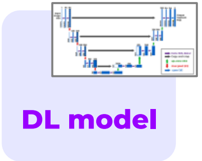
70B parameters

15T tokens for training set

LLaMA 2

Regular (single-task) model

Narrow AI



3.3M parameters

~ 6M tokens (**100 patients**) for training set

(100 x 200 slices x 300 tokens pr. slice)

PAPER

3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture

Dan Nguyen, Xun Jia, David Sher, Mu-Han Lin, Zohaib Iqbal, Hui Liu and Steve Jiang

Published 18 March 2019 • © 2019 Institute of Physics and Engineering in Medicine

Physics in Medicine & Biology, Volume 64, Number 6

Citation Dan Nguyen *et al* 2019 *Phys. Med. Biol.* **64** 065020

DOI 10.1088/1361-6560/ab039b

Let’s step back and think

- Model and dataset size, governance
- **Computational & Environmental Costs**
- Transparency and reproducibility
- Testing and reliability

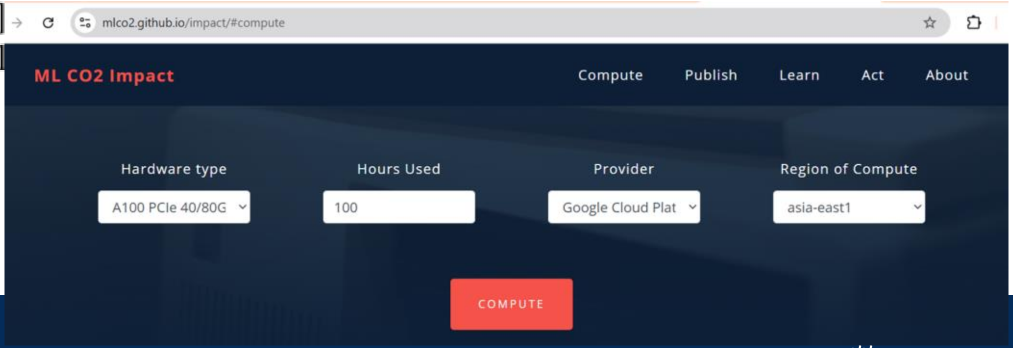
Computational & Environmental Costs

Trade-off between performance and sustainability

SAM
(Segment Anything) → **MedSAM** → **RT task**

SAM was trained on 256 A100 GPUS for 68 hours. We acknowledge the environmental impact and cost of training large scale models. The environmental impact of training the released SAM model is approximately 6963 kWh resulting in an estimated 2.8 metric tons of carbon dioxide given the specific data center used, using the calculation described in [77] and the ML CO_2 Impact calculator [61]. This is equivalent to ∼7k miles driven by the average gasoline-powered passenger vehicle in the US [101]

retraining and lower the barrier to entry for large scal



Computational & Environmental Costs

Trade-off between performance and sustainability

SAM (Segment Anything) → **MedSAM** → **RT task**

SAM was trained on 256 A100 GPUS for 68 hours. We acknowledge the environmental impact and cost of training large scale models. The environmental impact of training the released SAM model is approximately 6963 kWh resulting in an estimated 2.8 metric tons of carbon dioxide given the specific data center used, using the calculation described in [77] and the ML CO_2 Impact calculator [61]. This is equivalent to ~7k miles driven by the average gasoline-powered passenger vehicle in the US [101]. We released the SAM models to both reduce the need for retraining and lower the barrier to entry for large scale vision research.

BUT what if we end-up needing 200 UNets for all our-tasks?

Computational & Environmental Costs

It projects that **electricity demand from data centres worldwide is set to more than double by 2030** to around 945 terawatt-hours (TWh), …. <u>AI will be the most significant driver of this increase, with electricity demand from AI-optimised data centres projected to more than quadruple by 2030.</u>

Before, data centers (e.g. CECI) only used by computer science research, now they are <u>more and more used by researchers in many different fields</u>
Do we need more investment in data centers to keep up innovation?

Let’s step back and think

- Model and dataset size, governance
- Computational & Environmental Costs
- **Transparency and reproducibility**
- Testing and reliability

Generic tools for model & data reporting

Tool	Reference	Summary
ML Canvas	L. Dorard, 2015, https://www.ownml.co/machine-learning-canvas	Generic and simple template for a model card, available in different formats: PDF, Word, html, OpenDoc
Model Card Google	Mitchell et al., arXiv 2019	First model card template, by Google. Generic to any AI model. Includes model details, train / evaluation data, and performance
Factsheets IBM	M. Arnold et al., IBM Journal of Research and Development,2019	Collection of relevant information (facts) to promote transparency during the creation and deployment of an AI model, by IBM
Datasheets Microsoft	Gebru et al., arXiv 2021	First standard for datasets, by Microsoft. Generic to any dataset. Includes motivation, composition, collection process, cleaning, labelling, uses, maintenance, etc.
HuggingFace model creator HUGGING FACE	Ozoani 2022, hugginface.co/docs/hub/en/model-card-annotated	Tool to help operationalize model cards and create your own for specific applications

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.

Format and content choice let to the users - in order to be generic to any applications

Guidelines for reporting AI research*

**specific to healthcare*

Tool	Reference	Summary
CLAIM	Mongan et al., Radiology: AI 2020	Checklist for transparency in research papers involving AI in medical imaging
Model Fact Labels	Sendak 2020, npj Digital Medicine	1-page with relevant information to support clinicians for AI-based decision making
MINIMAR	Hernandez-Boussard, Jamia 2020	Checklist for MINimum Information for Medical AI Reporting
SPIRIT-AI **CONSORT-AI**	Cruz Rivera et al. Lancet Digit Health 2020	Guidelines for clinical trials protocols and reports for interventions involving AI
CLAMP	Naqa , Med Phys 2021	Methodology in sufficient detail to allow replication in publications
STARD-AI	Sounderajah et al., BMJ Open 2021	AI version of the Standards for Reporting of Diagnostic Accuracy Study checklist
DECIDE-AI	Vasey, BMJ 2022	Reporting Checklist for decision support systems (academic)
PRISMA-AI	Cacciaman et al., Nat Med 2023	Guidelines for systematic reviews and meta-analysis of AI interventions
CLEAR	Kocak et al., Insights Imaging 2023	CheckList for EvaluAtion of Radiomics research (CLEAR)
TRIPOD+AI	Collins et al., BMJ 2024	Reporting of studies that develop a prediction model or evaluate its performance
PROBAST+AI	Moons, BMJ 2025	Quality, risk of bias, and applicability assessment for prediction models using AI

Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update

Section/Topic	No.	Item
TITLE/ABSTRACT		
	1	Identification as a study of AI methodology, specifying the category of technology used (eg, deep learning)
ABSTRACT		
	2	Summary of study design, methods, results, and conclusions
INTRODUCTION		
	3	Scientific and/or clinical background, including the intended use and role of the AI approach
	4	Study aims, objectives, and hypotheses
METHODS		
Study Design	5	Prospective or retrospective study
	6	Study goal
Data	7	Data sources
	8	Inclusion and exclusion criteria
	9	Data preprocessing
	10	Selection of data subsets
	11	De-identification methods
	12	How missing data were handled
	13	Image acquisition protocol
Reference Standard	14	Definition of method(s) used to obtain reference standard
	15	Rationale for choosing the reference standard
	16	Source of reference standard annotations
	17	Annotation of test set
	18	Measures of inter- and intrarater variability of features described by the annotators
Data Partitions	19	How data were assigned to partitions
	20	Level at which partitions are disjoint

The risk of general statements & wording

ELSEVIER

Radiotherapy and Oncology
Volume 194, May 2024, 110196

Original Article

Assessment of bias in scoring of AI-based radiotherapy segmentation and planning studies using modified TRIPOD and PROBAST guidelines as an example

Coen Hurkmans [a,b], Jean-Emmanuel Bibault [c], Enrico Clementel [d], Jennifer Dhont [e,f], Wouter van Elmpt [g], Georgios Kantidakis [d], Nicolaus Andratschke [h]

- TRIPOD and PROBAST checklist items were adapted for AI reporting (Delphi process)
- 10 articles were scored by 6 co-authors
- For 41 items (out of 61) no statistically significant kappa was obtained indicating that the **level of agreement among multiple observers is due to chance alone**
- This raises concerns about the applicability of such checklists to objectively score articles for AI applications
- New checklists should not use subjective words nor composite questions.

The risk of general statements & wording

Radiotherapy and Oncology
Volume 194, May 2024, 110196

Original Article

Assessment of bias in scoring of AI-based radiotherapy segmentation and planning studies using modified TRIPOD and PROBAST guidelines as an example

Coen Hurkmans [a,b], Jean-Emmanuel Bibault [c], Enrico Clementel [d], Jennifer Dhont [e,f], Wouter van Elmpt [g], Georgios Kantidakis [d], Nicolaus Andratschke [h]

1. Checklist items composed of sub-questions/summations (e.g., "e.g., objectives, sample size, input parameters, statistical analysis, study design and conclusions") could be further clarified or clarified how this should be scored if this is partly answered.

2. Items with subjective words like "clearly", "appropriate" or "explain" can be problematic as leaving more space to subjective interpretation and may even drift over time if the field gets more mature. It was however decided not to replace them as otherwise these items would start to substantially deviate from the original items.

Let’s step back and think

- Model and dataset size, governance
- Computational & Environmental Costs
- Transparency and reproducibility
- **Testing and reliability** (knowing when to trust the AI?)

  - ➢ Some methods (e.g. ensembling for UQ) can be still applied, but not others (e.g. GradCAM for explainability)
  - ➢ Testing still needs to be done task-specific!
  - ➢ A lot to be explored! (*main research in my opinion in the AI field*)

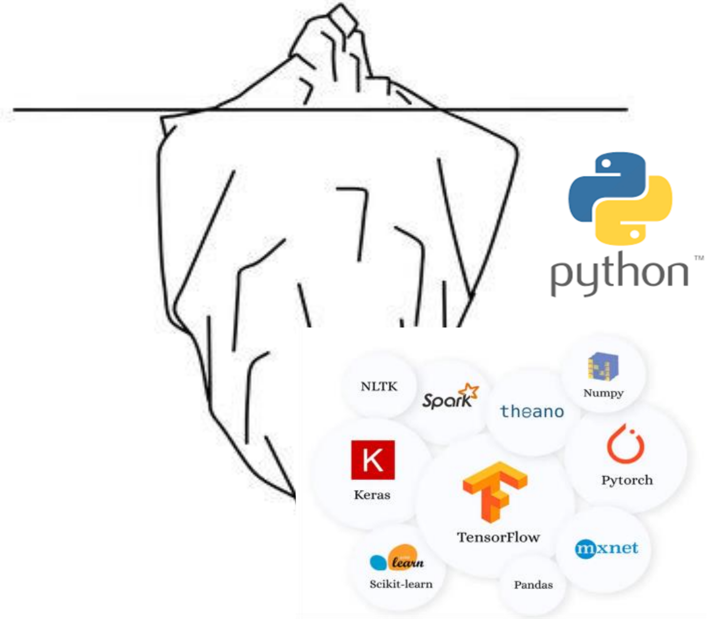
Ana Barragán is research associate with

fnrs
LA LIBERTÉ DE CHERCHER

MIRO TEAM MEMBERS
John Lee
Edmond Sterpin
Xavier Geets
Ana Barragan Montero
Geneviève Van Ooteghem
Dario Di Perri
Julien Pierrard
Danah Pross
Lies Verpoest
Marina Orts
Chuxin Zhang
Valentine Dormal
Eliot Peeters
Romain Schyns
Ellina Macaeva

Contact me at
ana.barragan@uclouvain.be !

miro
molecular imaging, radiotherapy & oncology


python
NLTK
Spark
theano
Numpy
Keras
Pytorch
TensorFlow
mxnet
learn
Scikit-learn
Pandas