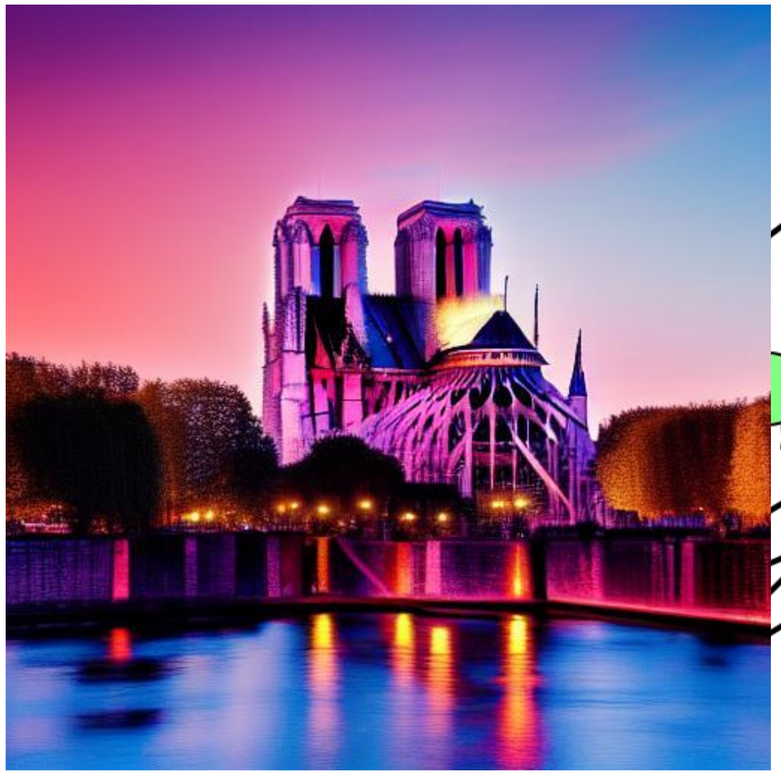
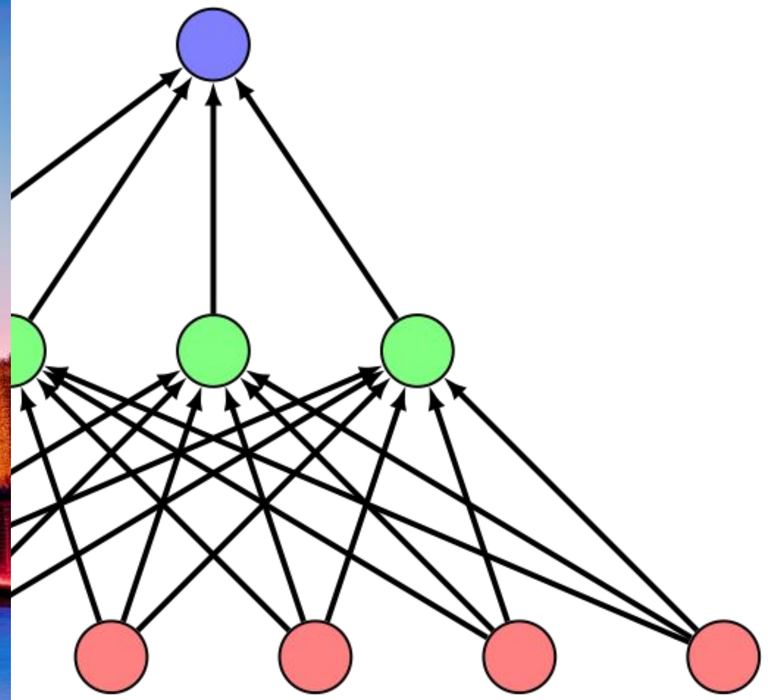


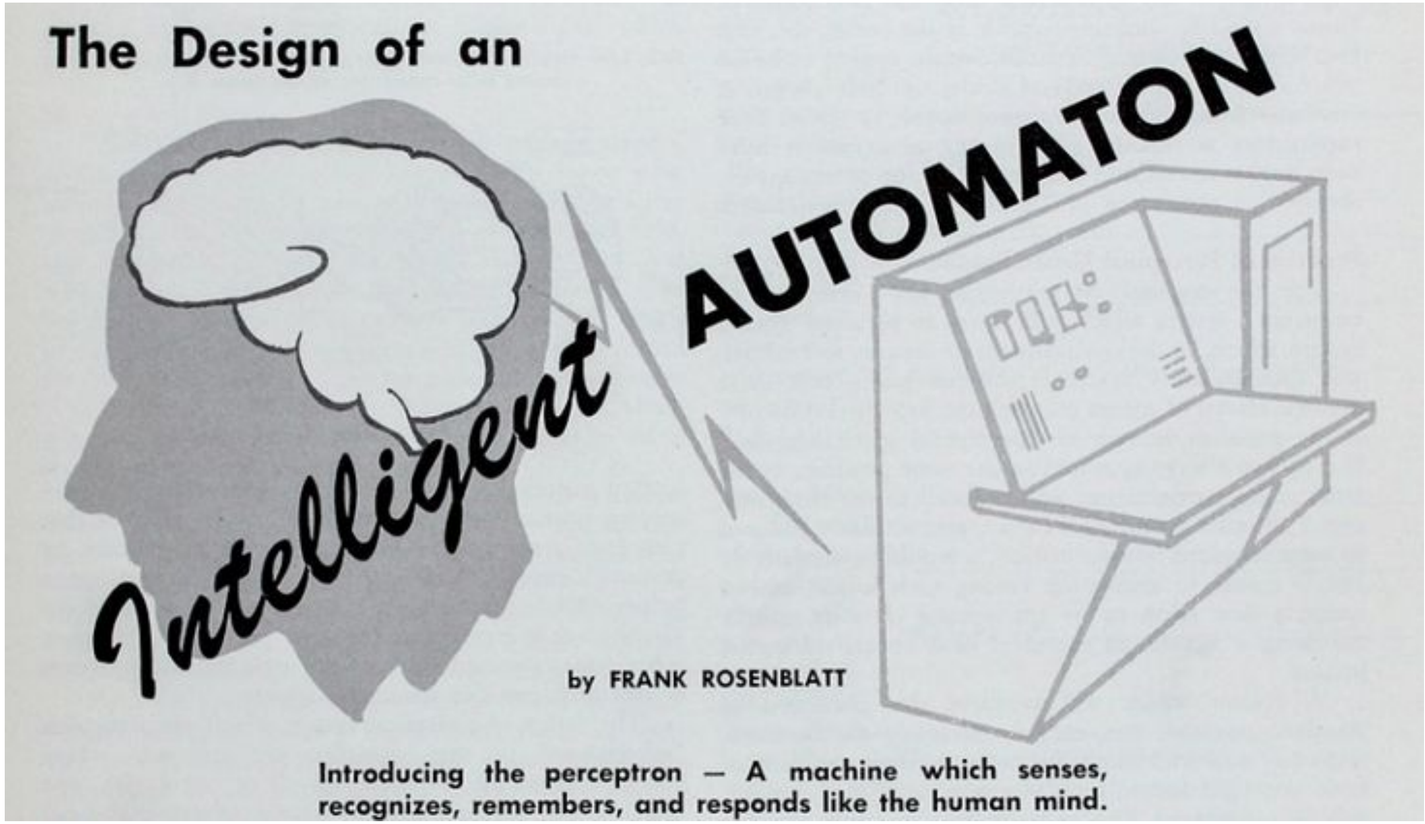


Modèles de fondation pour l'annotation, la segmentation et la reconnaissance de formes

Kévin Réby, UPR 2002 CNRS MAP Marseille
ERC n-Dame_Heritage

Une cathédrale de données numériques et connaissances pluridisciplinaires
Journées d'études du groupe de travail "données numériques" du chantier scientifique Notre-Dame de Paris
19 – 21 juin 2024, Campus CNRS Joseph Aiguier, Marseille

Deep Learning

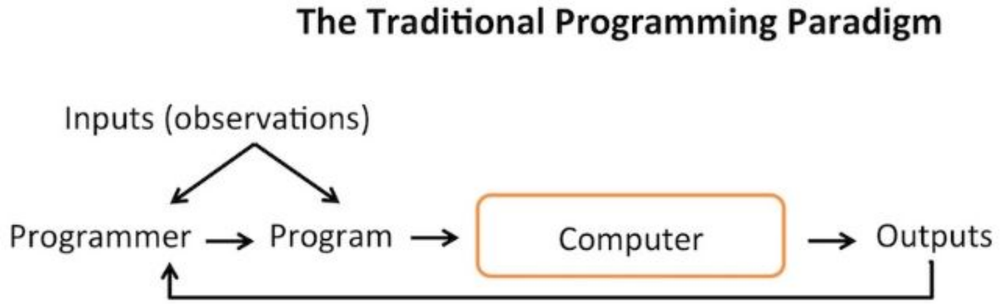


https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

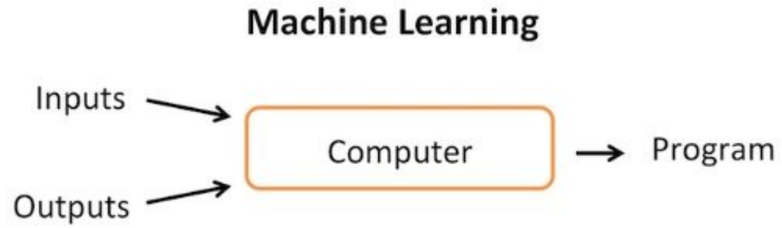
Big Data & Machine/Deep Learning

- Lots of unstructured/semi-structured data, such as text, images, and audio from experts
  - **quality & traceability**
- Natural Language Processing (**NLP**) tasks:
  - Named entity recognition, text classification
- **Image recognition** :
  - Object detection, Segmentation, Classification
- **Multimodal analysis**

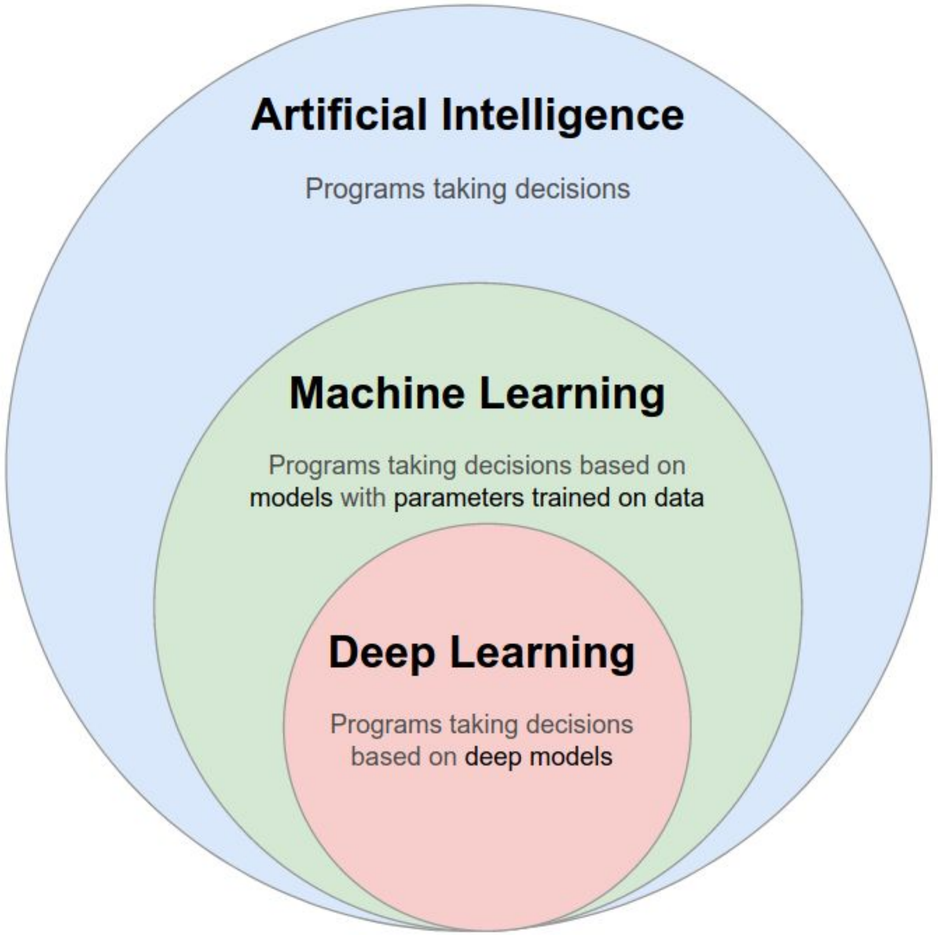
Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities?." *Journal of the Digital Humanities* 2.3 (2013).



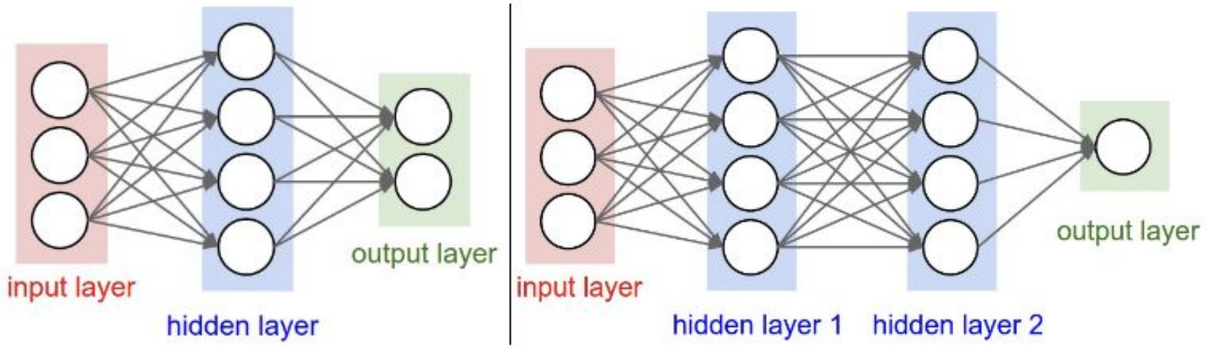
Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)



AI, Machine Learning & Deep Learning



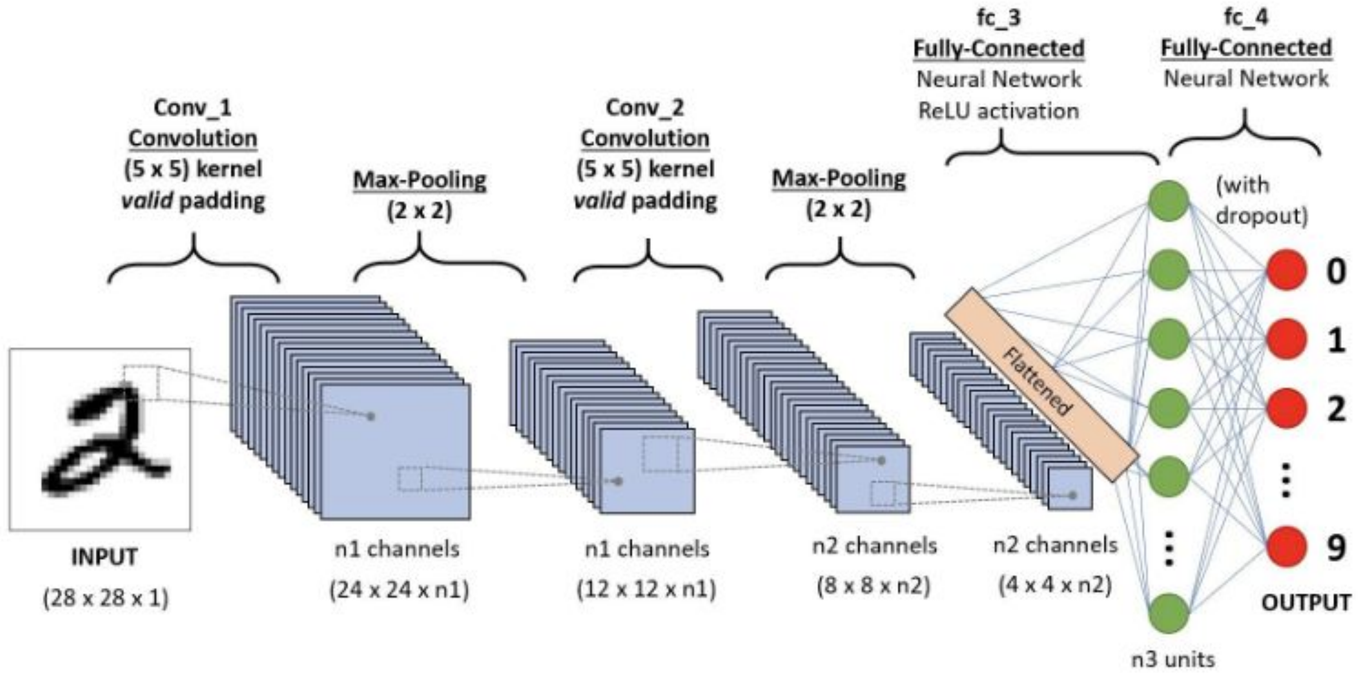
An artificial neural network is an organized **set of interconnected neurons** used to solve complex problems such as computer vision or natural language processing.



https://blog.sinatechnologie.com/presentation-des-reseaux-de-neurones-artificiels

Convolutional Neural Network (CNN)

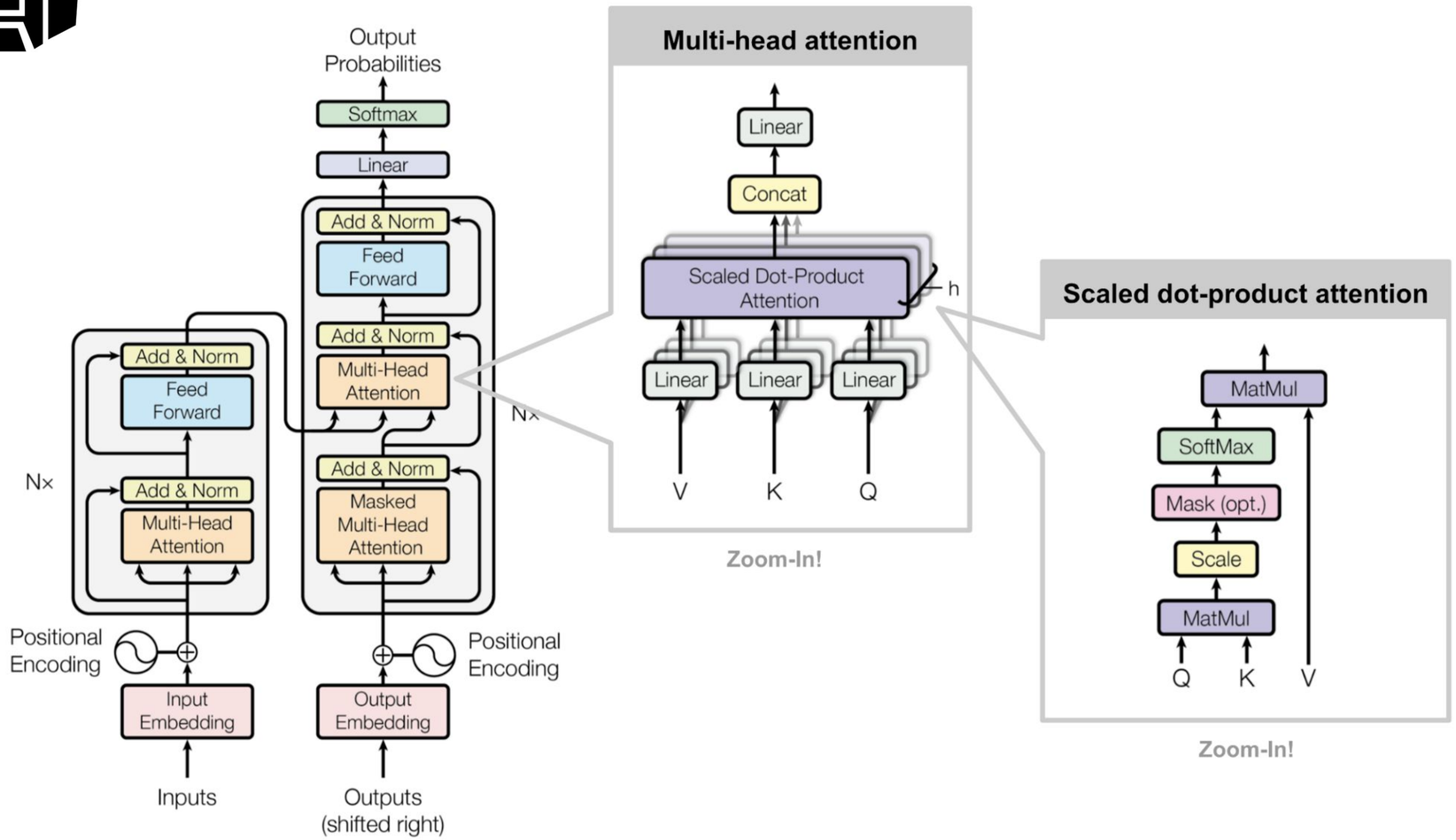
- CNNs were one of the key innovations that led to the deep neural network renaissance in **computer vision**
- Most well-known image recognition and classification algorithm.
- A typical CNN consists of a combination of
  - convolutional,
  - pooling,
  - and dense layers



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Transformer

- Seq2seq tasks without recurrent units
- **Encoder decoder** architecture
- **Multihead self attention**
- Large language models (**LLMs**) like as ChatGPT and BERT have achieved cutting-edge outcomes in a variety of NLP applications
- and have since expanded into other fields such as Computer Vision (VIT, Stable Diffusion, LayoutLM) and Audio (Whisper, XLS-R).



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

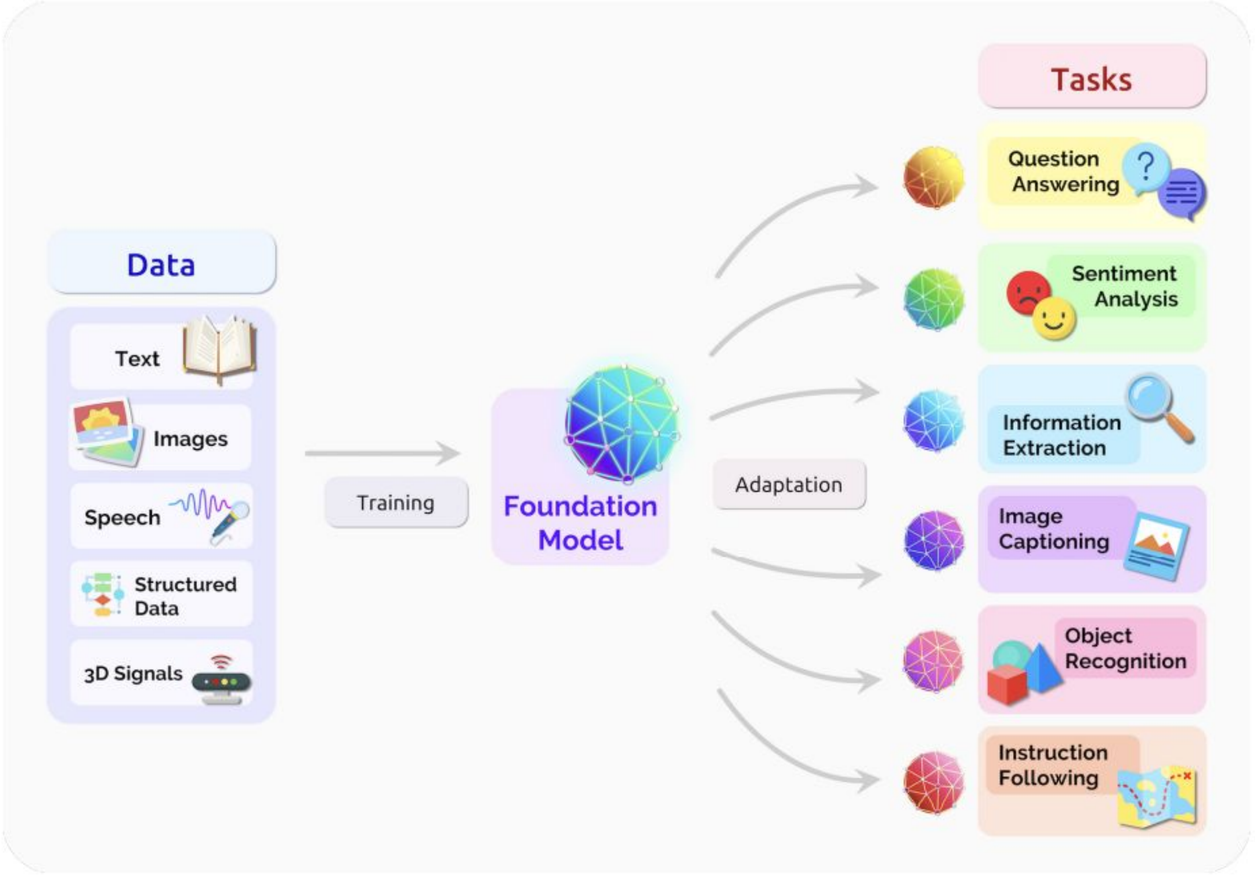
Foundation models

https://xkcd.com/1838/

Foundation model

- **Scaling** : large deep learning model (with several hundred billion parameters) that has been pre-trained on a massive amount of data (tens of millions of images), using several hundred of GPUs
- Without being expressly trained on them, foundation models can accomplish a variety of functions.
  ⇒ **Zero shot Learning**
- Have demonstrated great effectiveness in a variety of applications, including cross-modal retrieval, **zero-shot** categorization, and text-to-image/video/3D production.
- Multimodal foundation models are simultaneously trained using **many modalities**
- For example, given a short **natural language prompt** , DALLE or Midjourney, may execute tasks like answering questions, producing essays, or generating images.

- **Transfer learning** : process of applying "knowledge" from one task (for example, object detection in photographs) to another (for example, activity recognition in films).

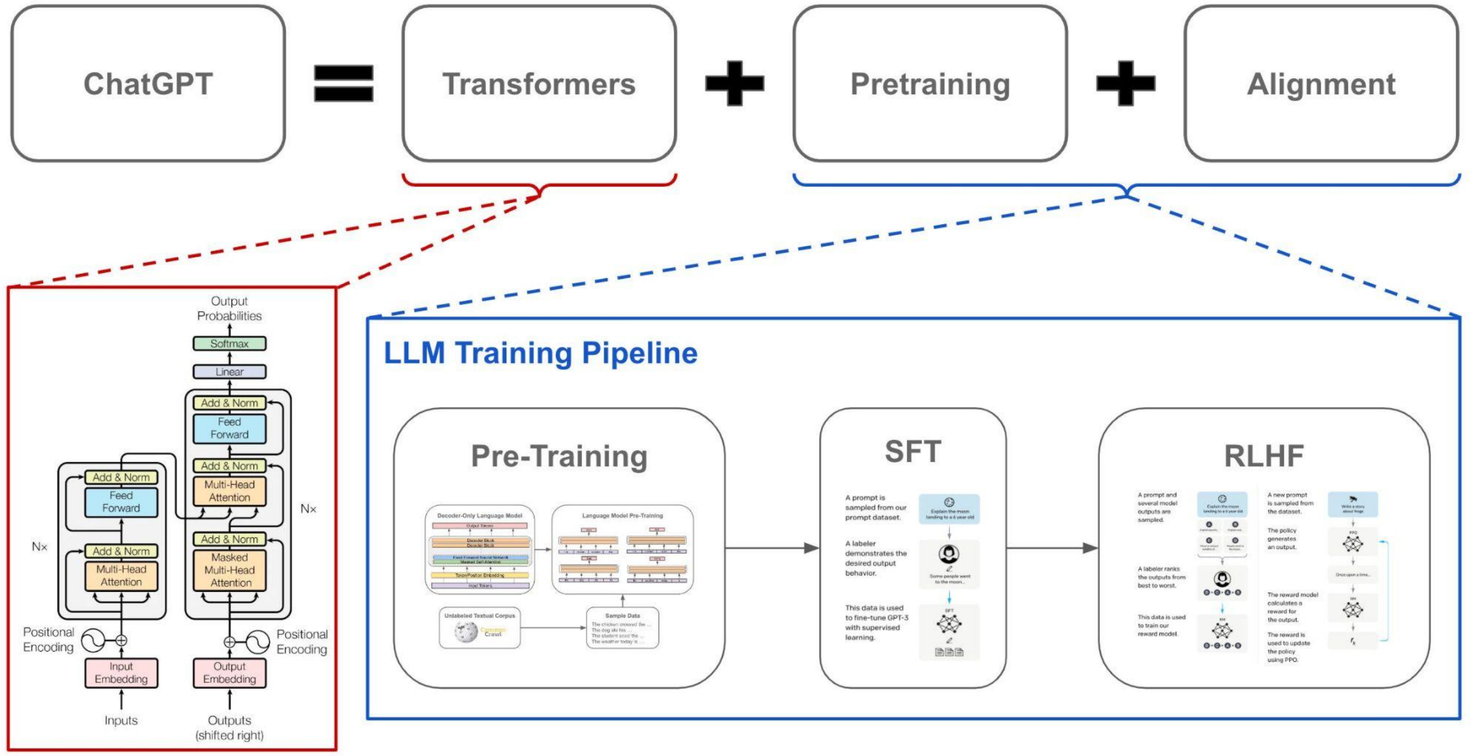


Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Large Language Models

Three main components :

1. ***Transformer*** *architecture*
2. *Language model* ***pretraining***
3. *The* ***alignment*** *process*



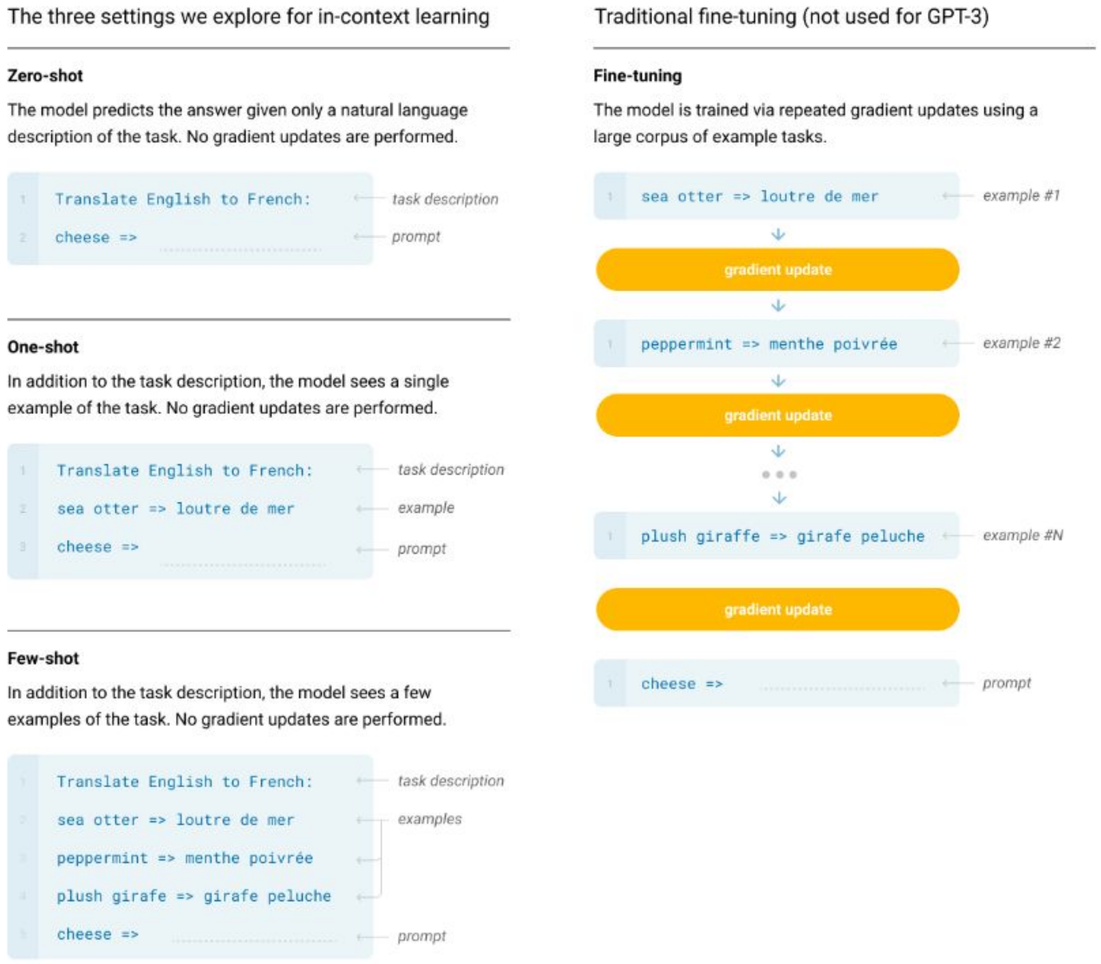
Cameron R. Wolfe, Deep (Learning) Focus

Generative Pre-trained Transformer 3

- System **prompt** (“useful assistant”)
- User initial question to start dialogue
- Start prediction
- Interaction with AI bots

⇒ With the proposal of GPT-3, we saw for the first time that **In Context Learning** (ICL) is an emergent ability of LLMs.

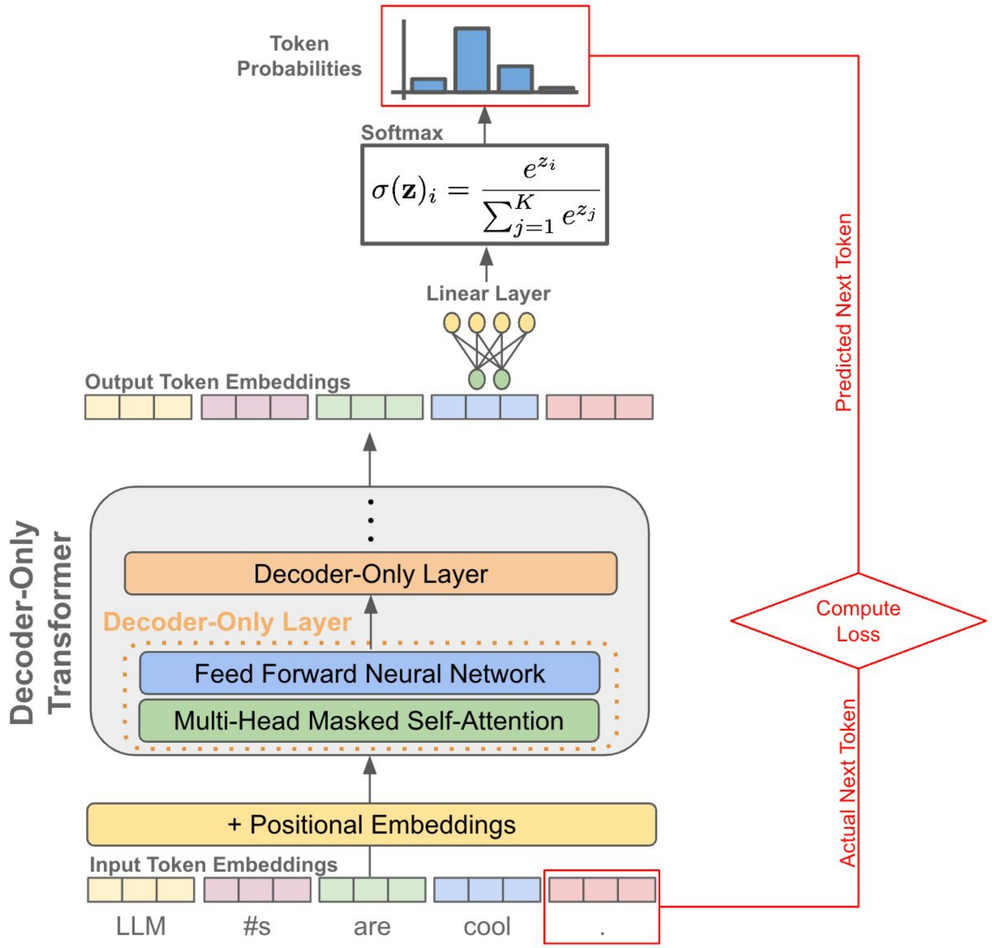
ICL = **use information in the context window to generate better output**



Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

LLM Architecture

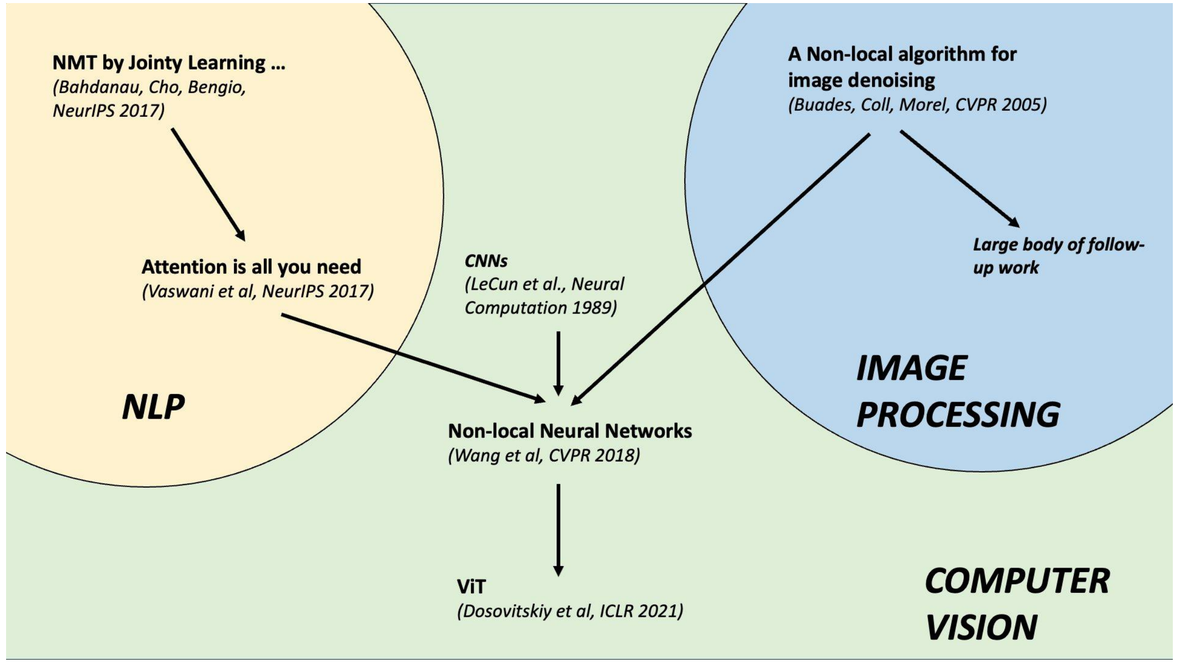
- **Generative** LLM: decoder only architecture → generate sequences of texts
- **Pretrained** :
  - takes a model at the start that's completely random (weights) = no knowledge
  - the learning objective is **next token prediction**
  - use giant corpus data from internet → unlabeled ⇒ self supervised learning
  - **general purpose** model
- Input → **Context Window**
- Embedding layer
- Processing the input: **Transformer**
  - Attention, Feed Forward
- Softmax → Probability distribution



Cameron R. Wolfe, Deep (Learning) Focus

Vision Language Model

- A VLM is a type of model that is designed to process and understand **both** visual and language information. VLM typically consist of two main components: a visual encoder and a language encoder.
- The **visual encoder** is responsible for processing the input image and extracting high-level visual features, such as objects, attributes, and relationships. This is usually done using a **vision transformer** .
- The **language encoder** is responsible for processing the input text and extracting high-level linguistic features, such as words, phrases, and syntax. This is usually done using a **transformer** .
- The two encoders are then connected by an **attention mechanism** , which allows the model to selectively focus on the relevant visual and language features for the task at hand.
- One example of a vision-language model is the **Segment Anything Model (SAM)** , which is a model that can segment any object in an image given a **prompt.**



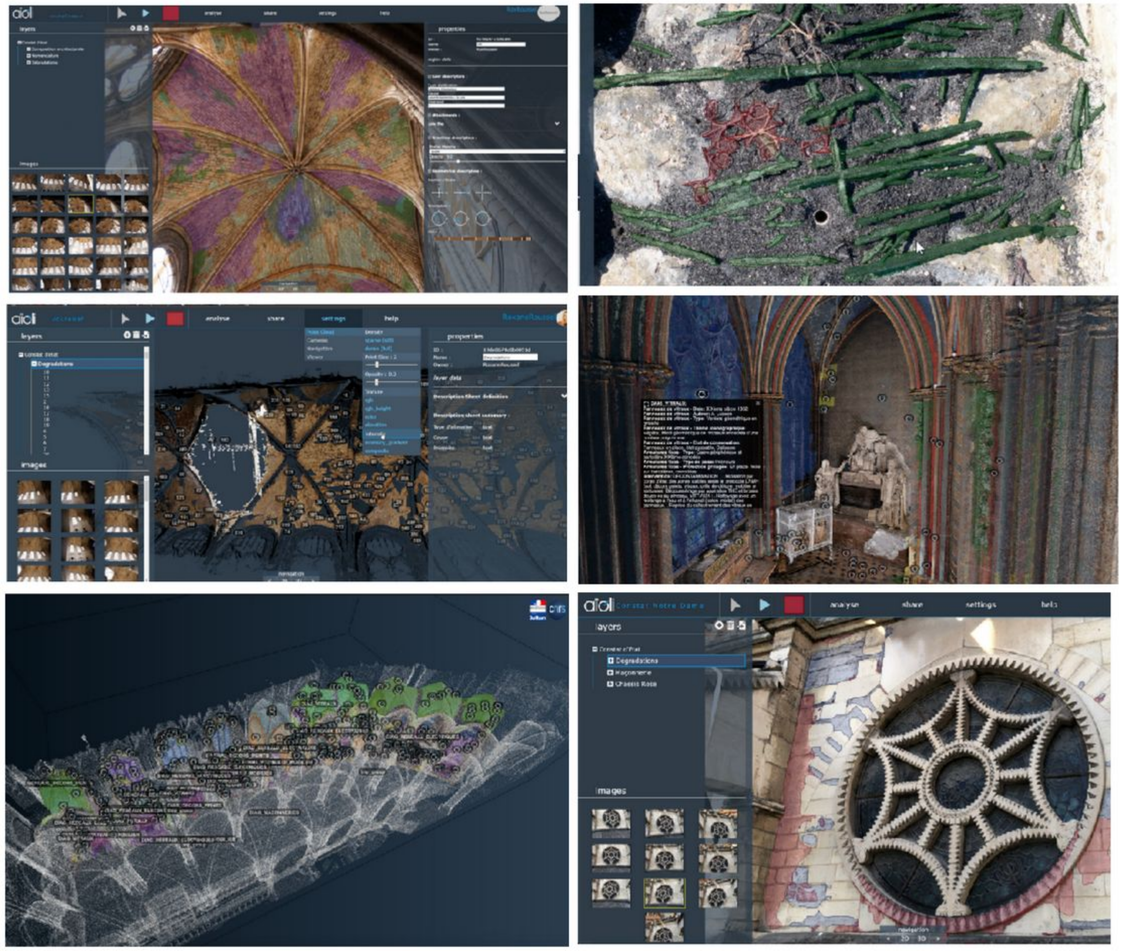
Christian Wolf, 2023

Vision Language Models

Detection & Segmentation

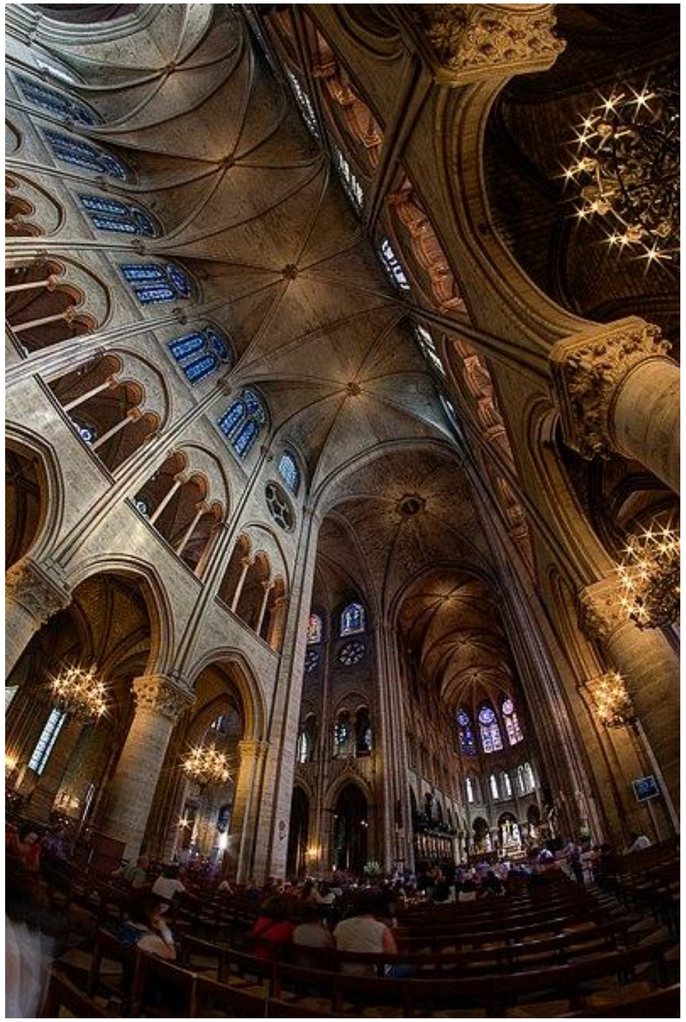
- **Object detection** = techniques like bounding boxes are used to localize objects within the image.
- **Semantic segmentation** = the model can assign a class label to every pixel, resulting in a detailed segmentation mask that precisely outlines different objects and regions in the image.

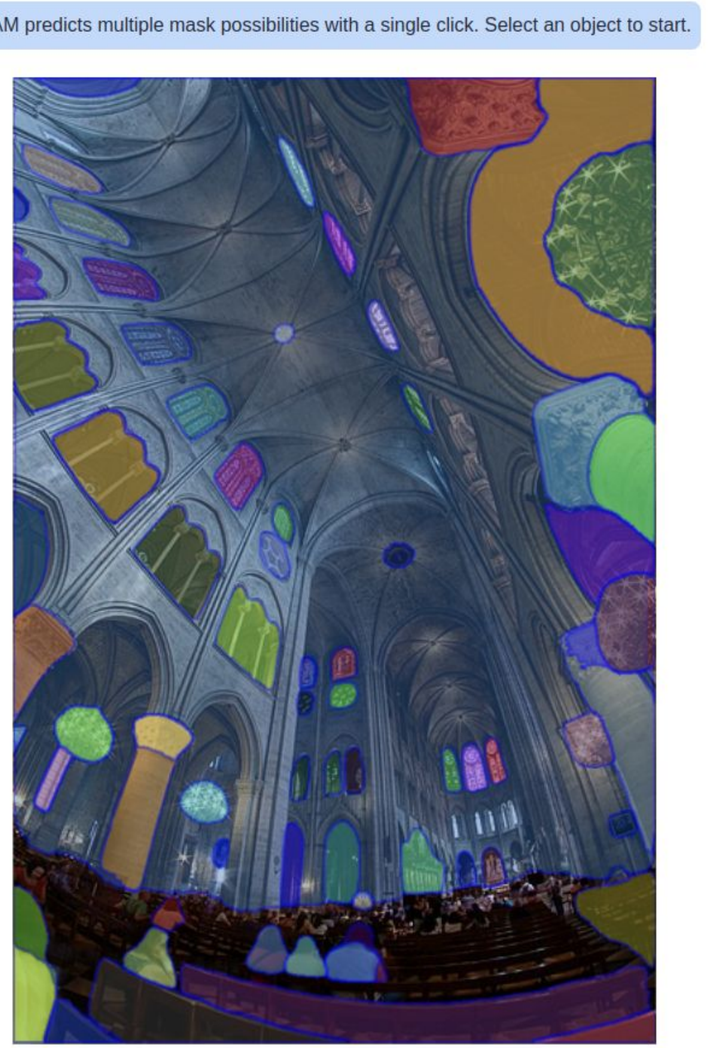
- Deep learning can act as a powerful tool for experts to automates tasks and opens doors to new applications.
- Learn from vast amounts of data and identify increasingly intricate features.
- Additionally, deep learning architectures can be adapted and fine-tuned for specific tasks and domains by leveraging pre-trained models on massive datasets.
  ⇒ **automatic annotation & segmentation**



Segment Anything Model (SAM)

- Foundation model for segmentation
- The model predicts binary masks that state the presence or not of the object of interest given an image.
- You can **prompt** multiple points for the same image, and predict a single mask.
- Generalization ⇒ **zero shot learning**





SA-1B Dataset

- SA-1B dataset:
  11 millions images
  1 billion annotations
- 256 GPUs
- Generalization
- Segmentation only (no labels)

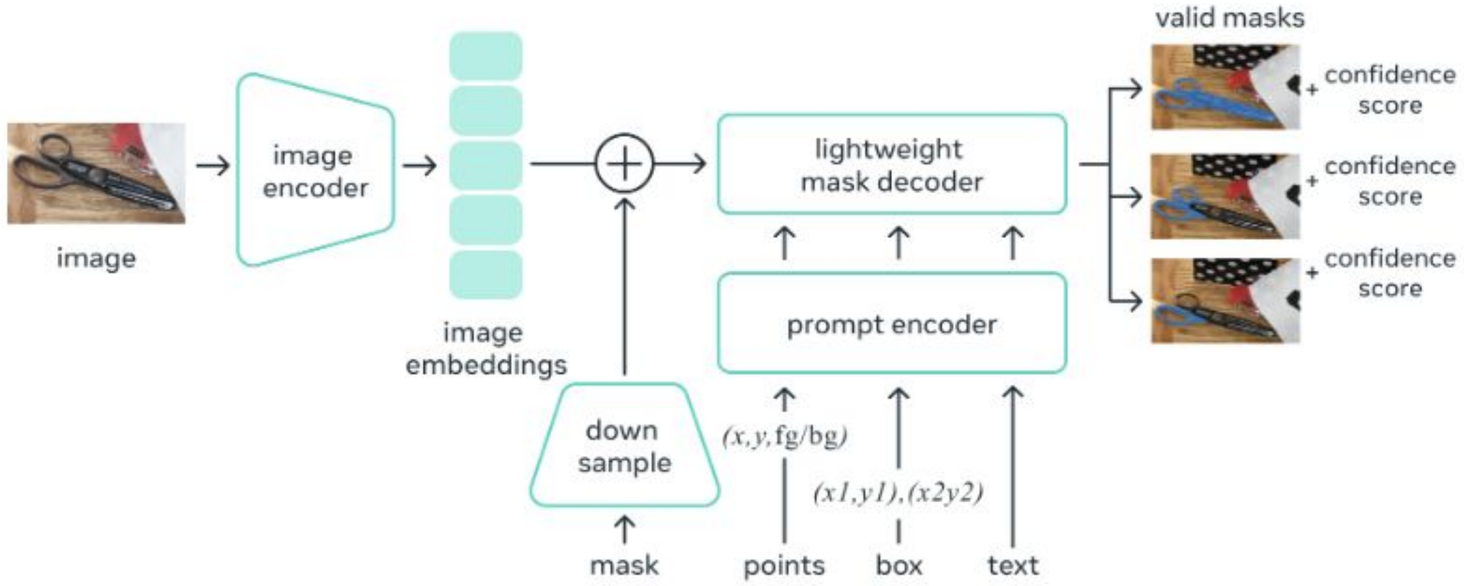
Kirillov, Alexander, et al. "Segment anything." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

SAM Architecture

Encoder :
Masked AutoEncoder (MAE)
Pre-trained Vision Transformer (ViT)

Decoder:
Transformer+MLP ⇒ classifier

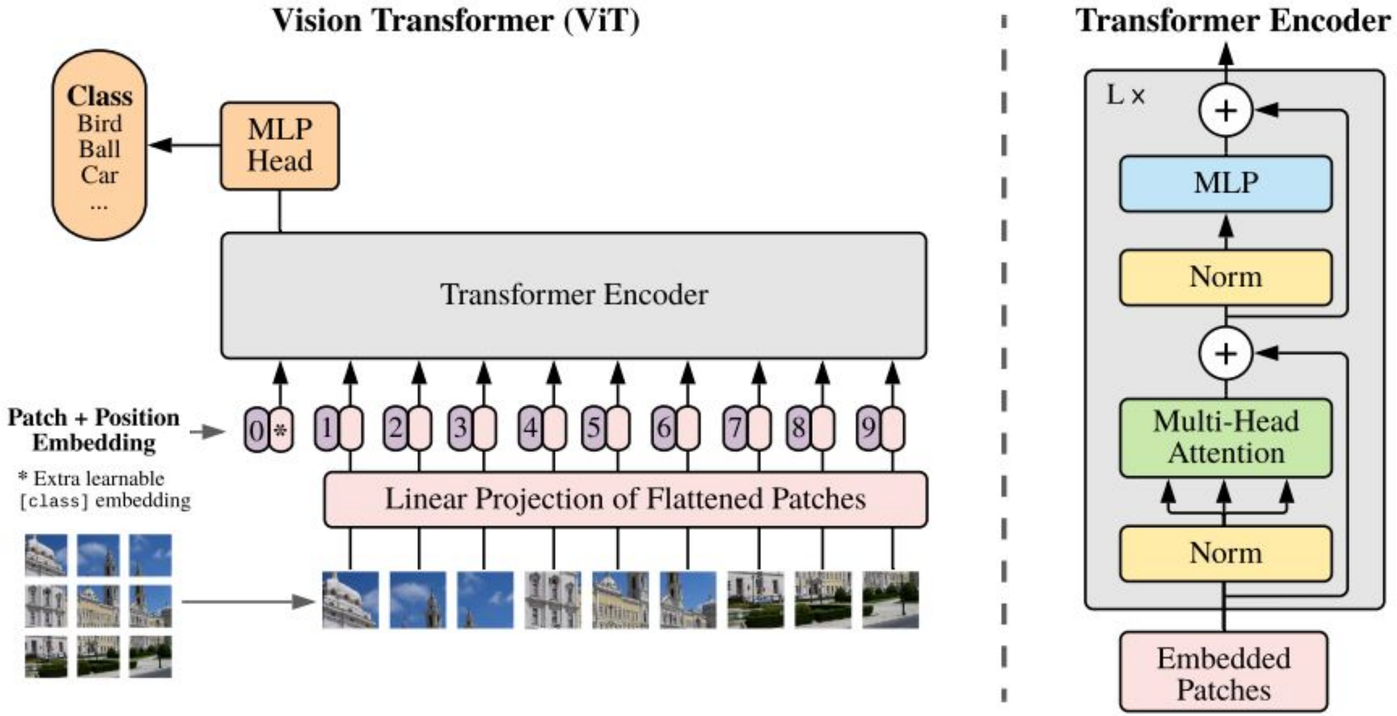
- The model predicts binary **masks** that state the presence or not of the object of interest given an image.
- You can **prompt** multiple points for the same image, and predict a single mask.



Kirillov, Alexander, et al. "Segment anything." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Vision Transformer

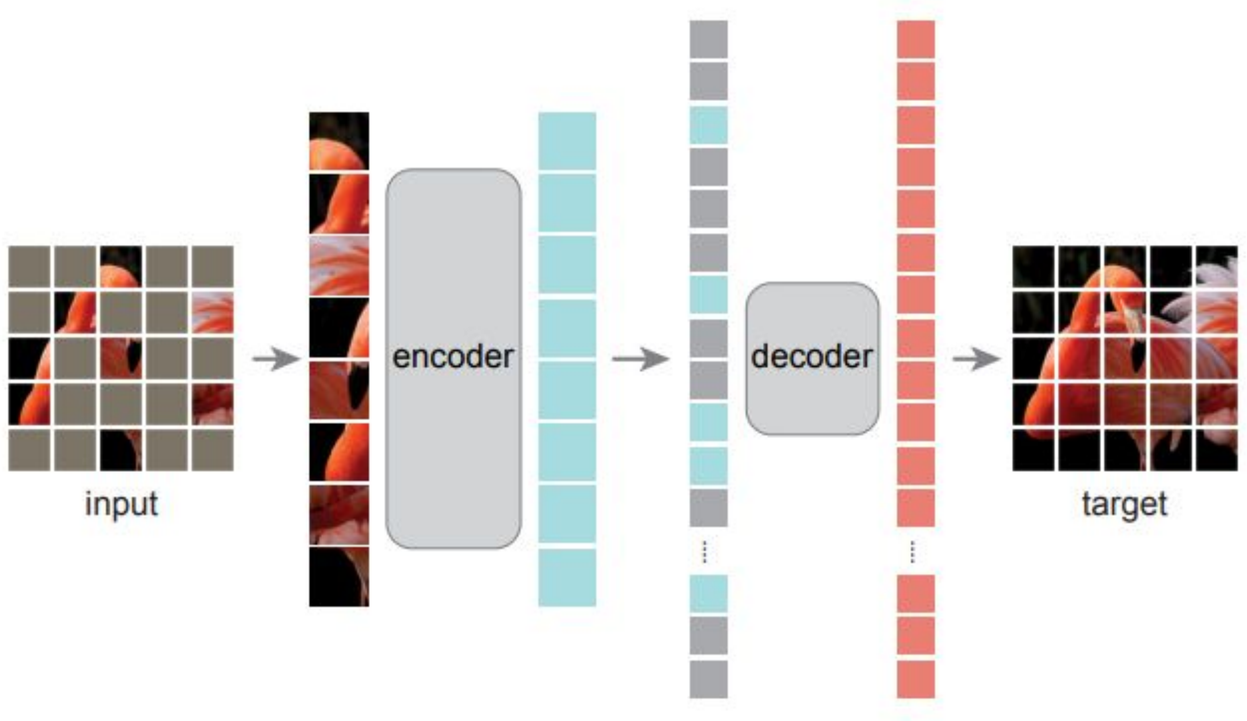
- Vision Transformers is an architecture that uses **self-attention mechanisms** to process images.
- The Vision Transformer Architecture consists of a series of transformer blocks.
- Each transformer block consists of two sub-layers:
  - a **multi-head self-attention layer**
  - and a **feed-forward layer** .



Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

Masked Auto Encoder

- Strategy that works well in computer vision: **masking a very high portion of random patches** .
- This strategy largely **reduces redundancy** and creates a challenging self supervisory task that requires holistic understanding beyond low-level image statistics.
- The autoencoder's decoder, which maps the latent representation back to the input, plays a different role between reconstructing text and images.



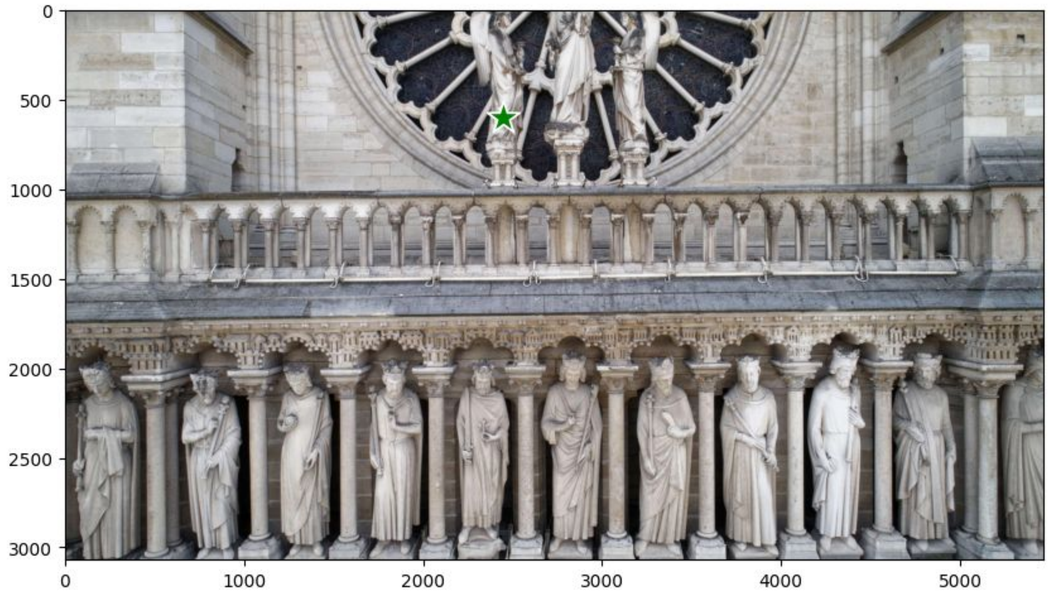
He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Automatic mask generation

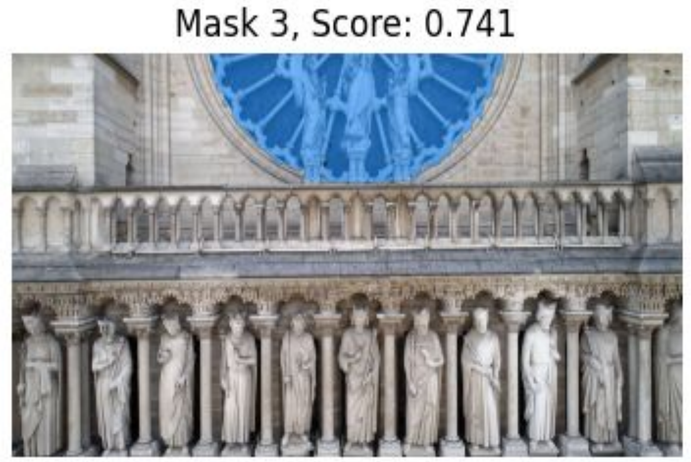
input = just the image, without prompt

Prompts

inputs = image + point

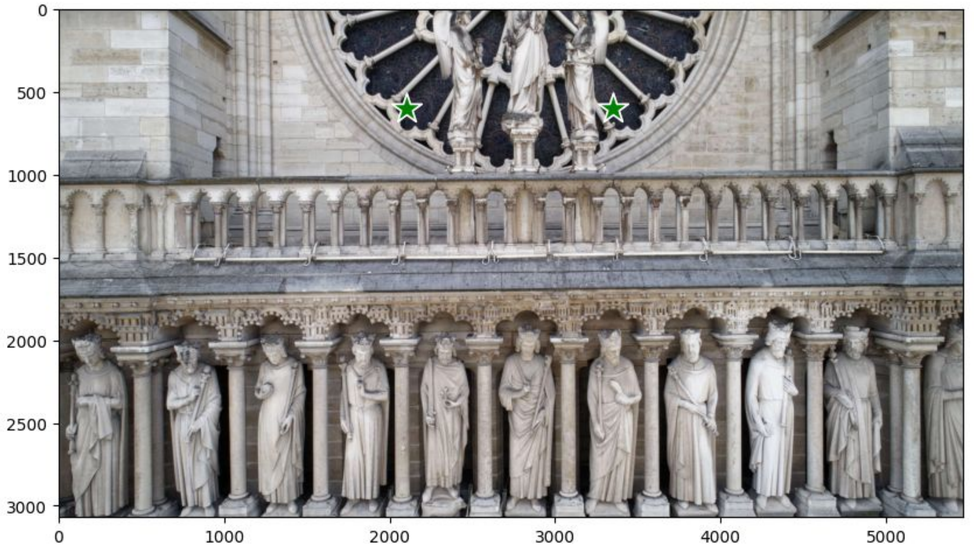


outputs

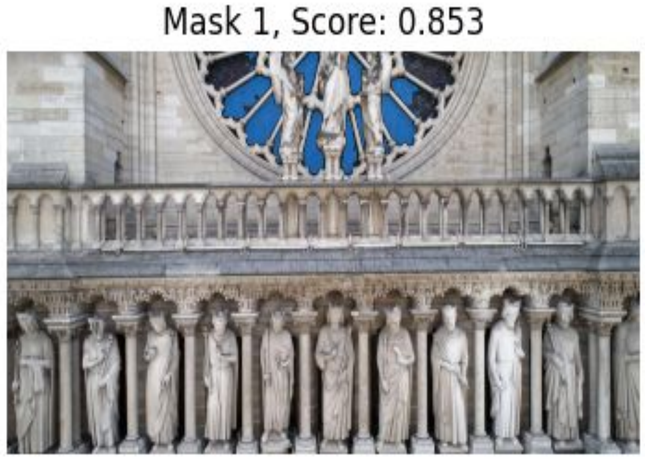


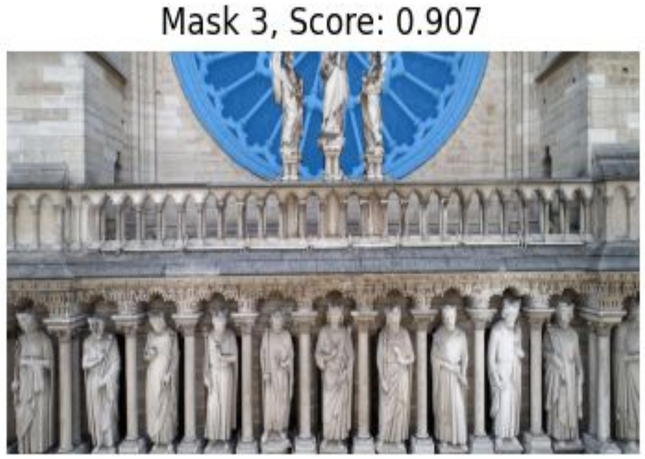
Example

inputs = image + multiples points



outputs

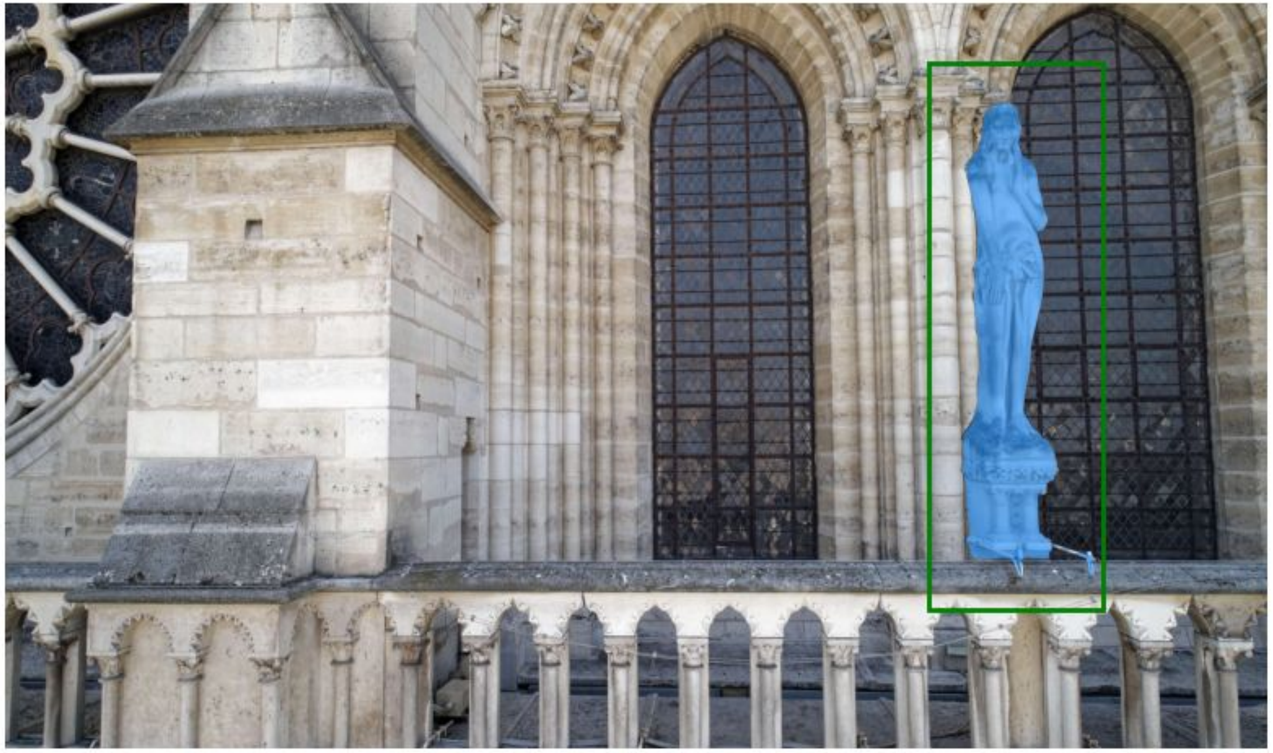




Input:

- **image**
- **prompt = points and/or bounding boxes**

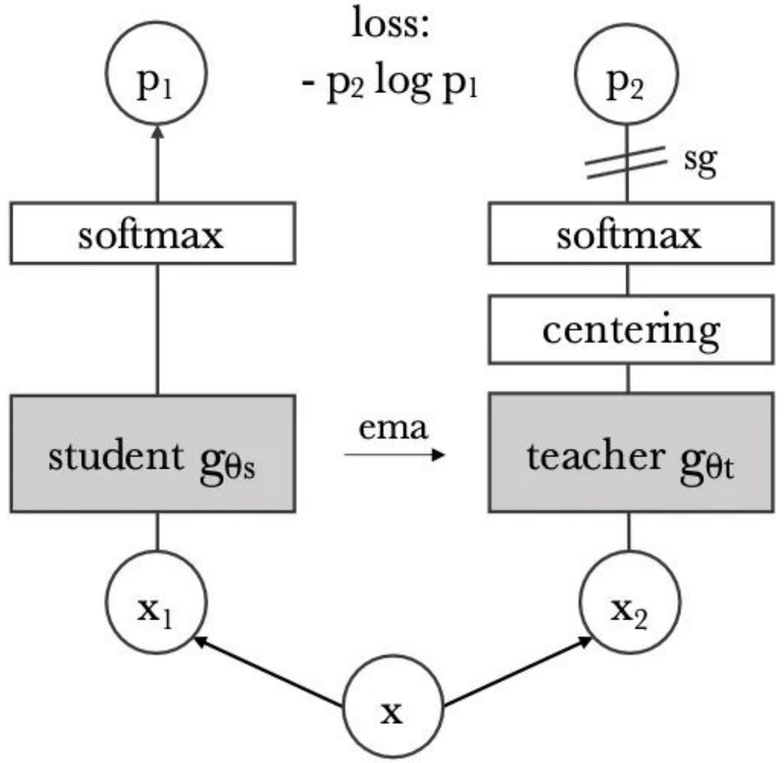
⇒ The model predicts much better results if input 2D points and/or input **bounding boxes** are provided



How to have bounding boxes ?

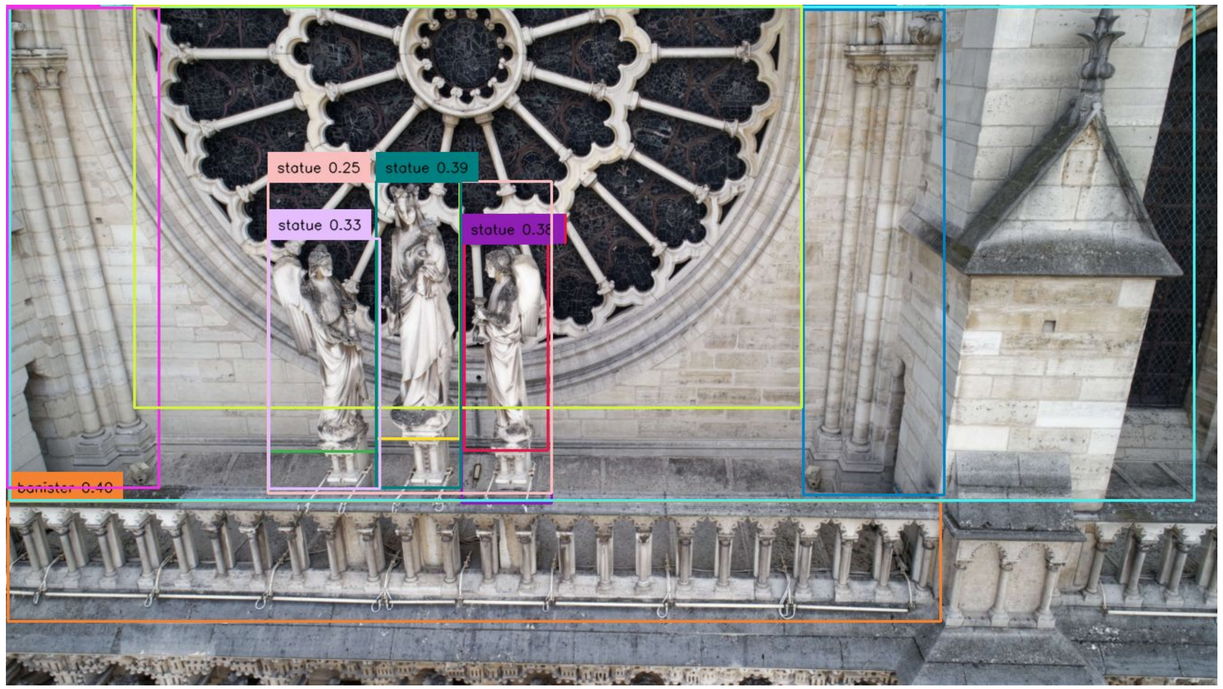
DiNO (self-distillation with no labels)

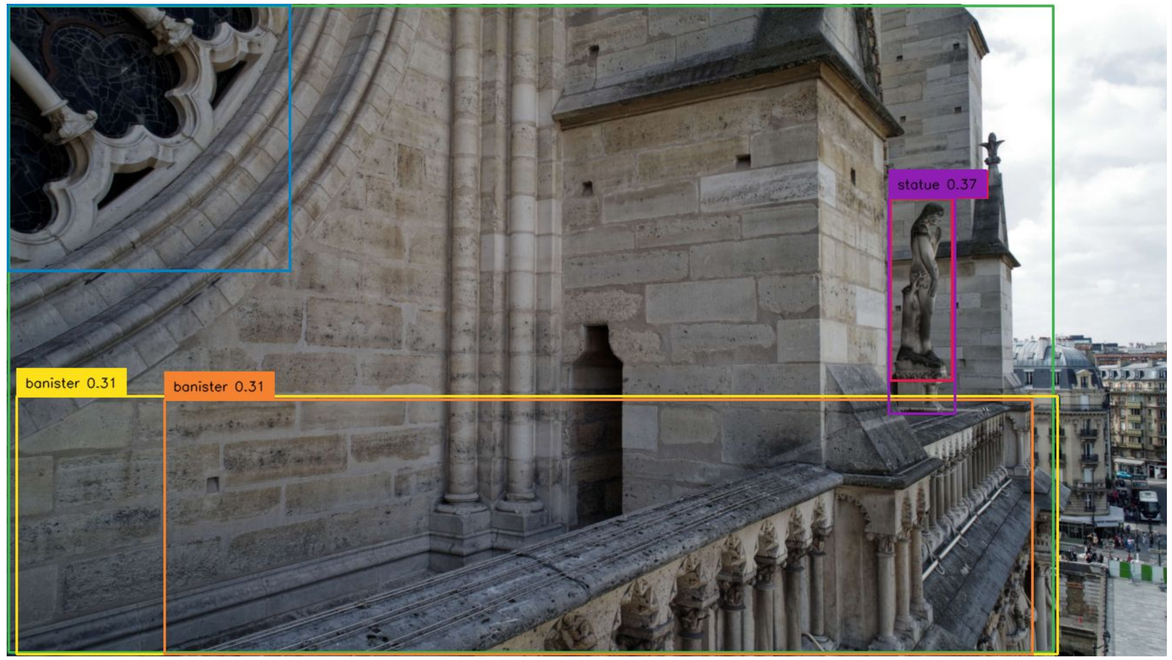
- A **Student** ViT learns to predict global features in an image from local patches supervised by the cross entropy loss from a momentum **Teacher** ViT's embeddings while doing centering and sharpening to prevent mode collapse
- There is a teacher and student network both having the same architecture, a **Vision Transformer(ViT)** .
- The teacher is a **momentum teacher** which means that it's weights are an exponentially weighted average of the student's.
- The network learns through a process called '**self-distillation** '.
- The teacher and student each predict a 1-dimensional embedding.
- A softmax along with cross entropy loss is applied to make student's distribution match the teacher's: The cross-entropy loss tries to make the two distributions the same just as in **knowledge distillation** .



Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

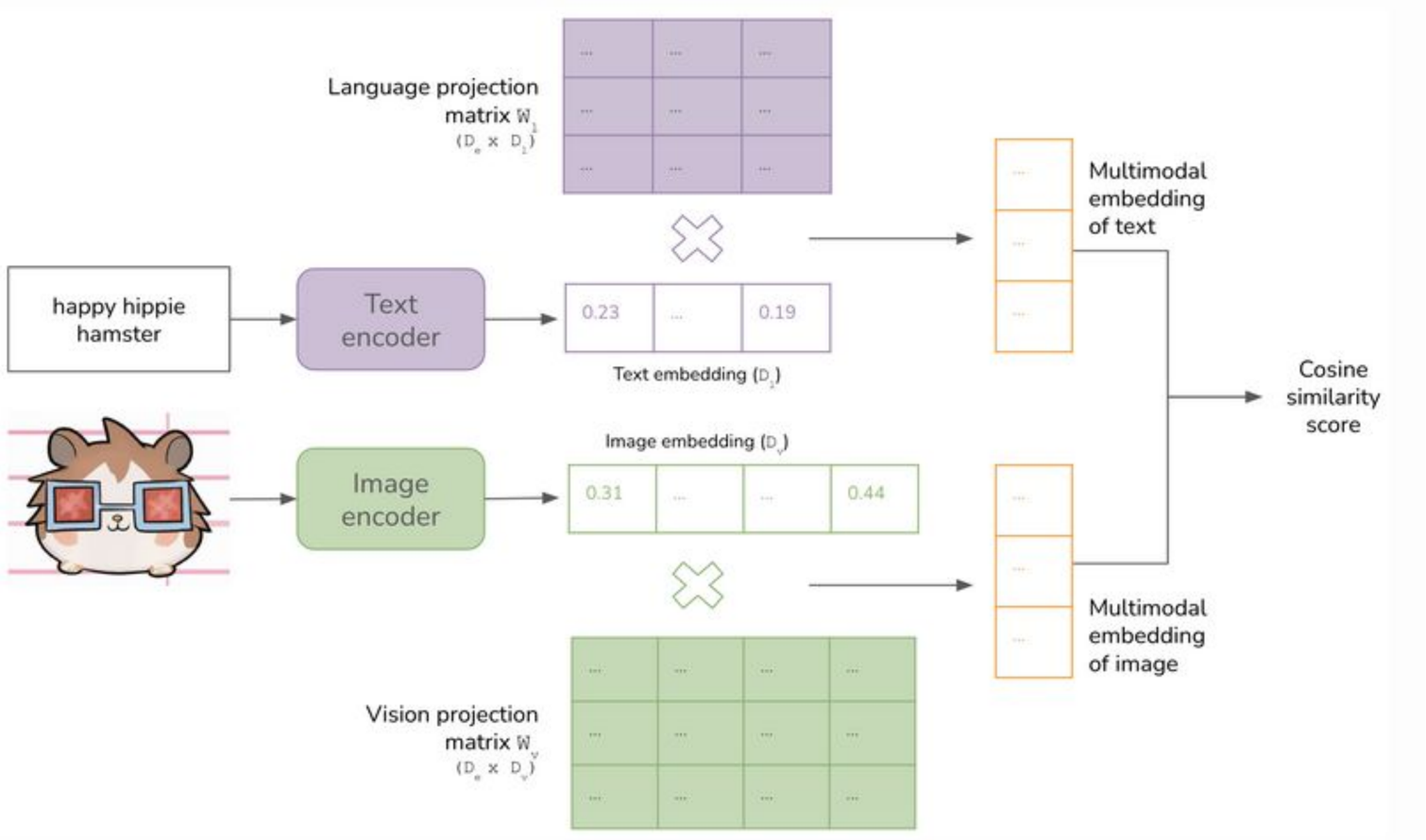
DiNO





But what about labels ?

CLIP (Contrastive Language-Image Pretraining)



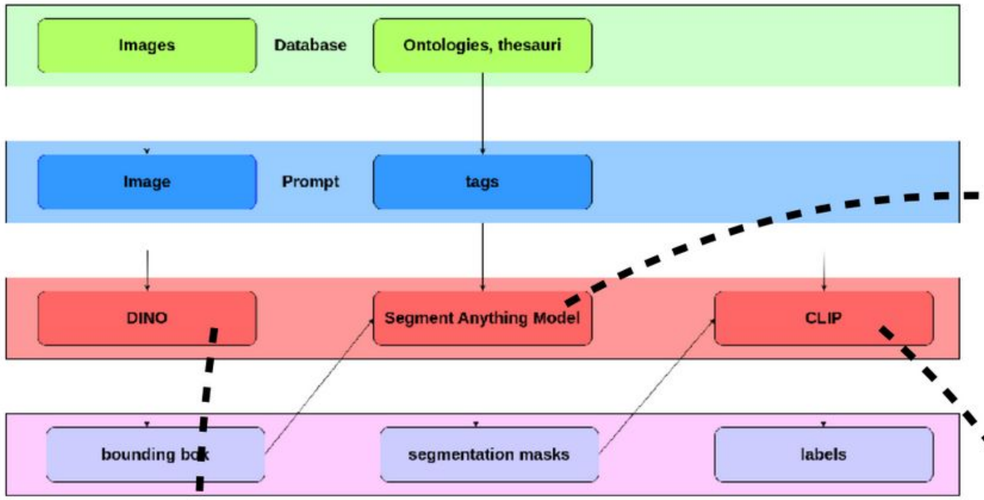
Both encoders and projection matrices are jointly trained together from scratch.

The training goal is to **maximize** the **similarity scores** of the right (image, text) pairings while **minimizing** the similarity scores of the wrong pairings (**contrastive learning**).

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP

labels are wrong !
⇒ need for controlled vocabularies !



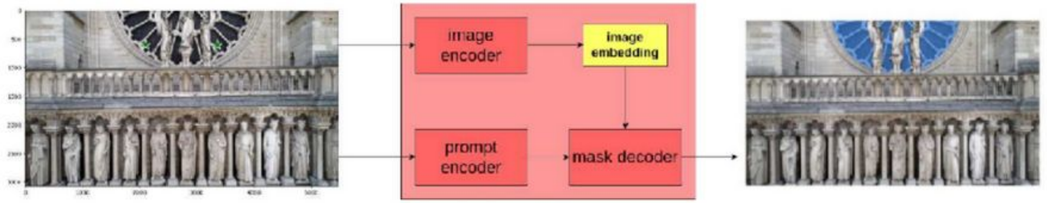


Figure 2. Overview of the Segment Anything Model.

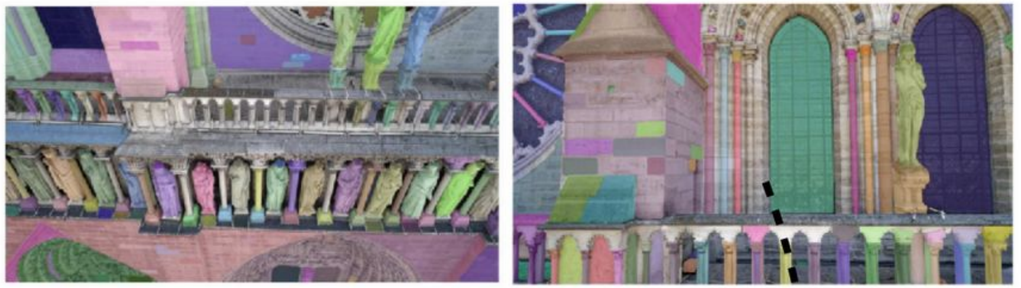
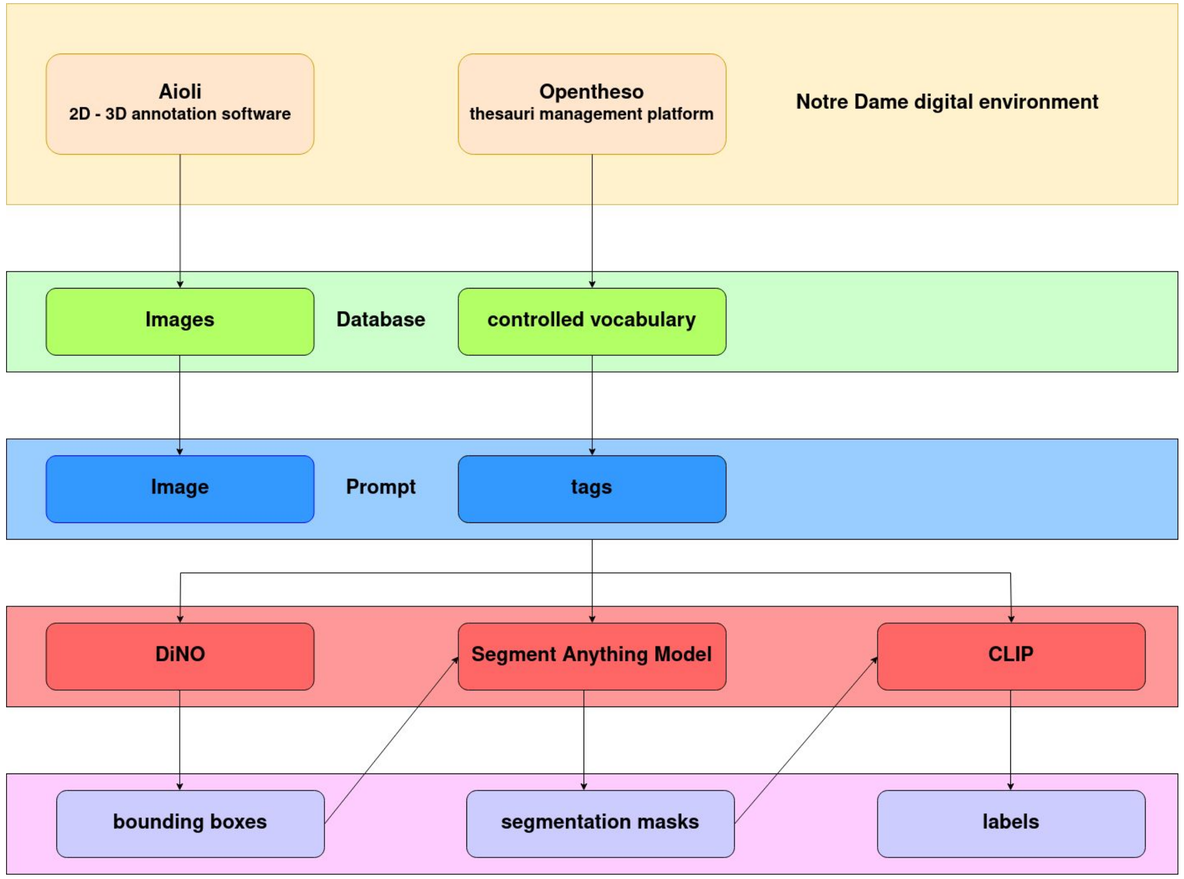
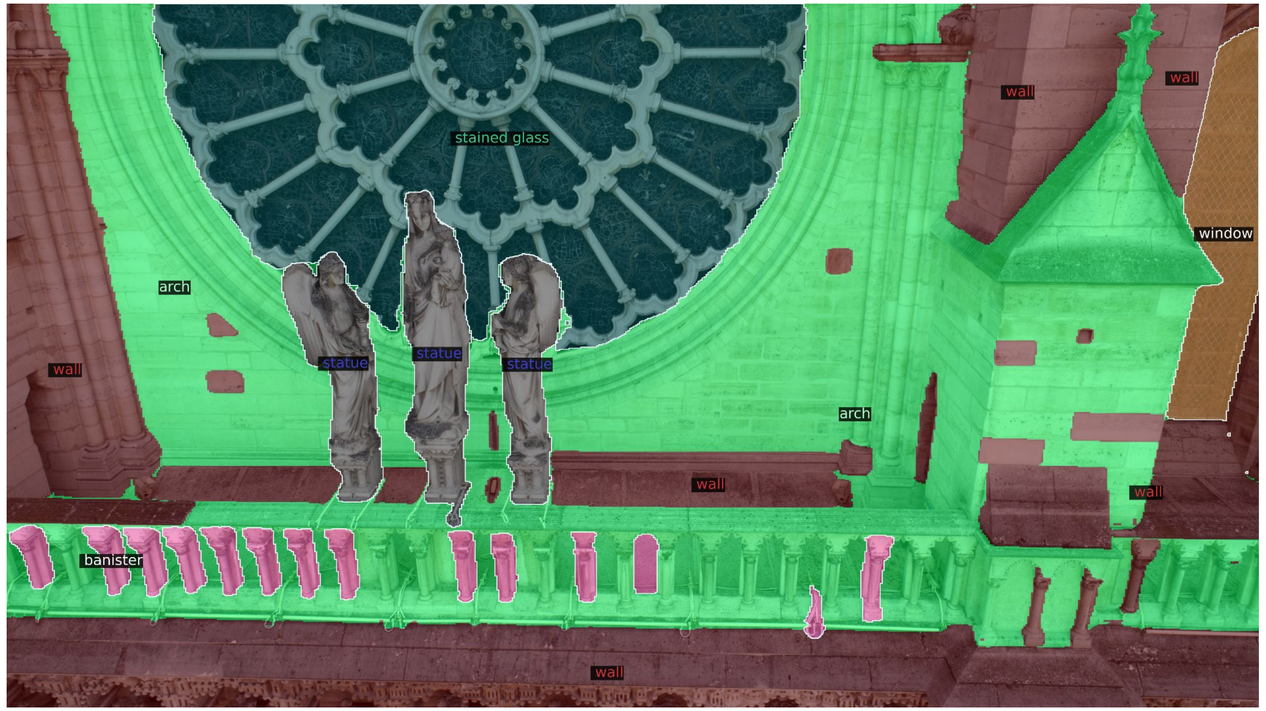


Figure 3. Examples of automatic segmentation using SAM.

Réby, Kévin, Anaïs Guillem, and L. de Luca. "Semantic Segmentation using Foundation Models for Cultural Heritage: an Experimental Study on Notre-Dame de Paris." *4th ICCV Workshop on Electronic Cultural Heritage*. 2023.

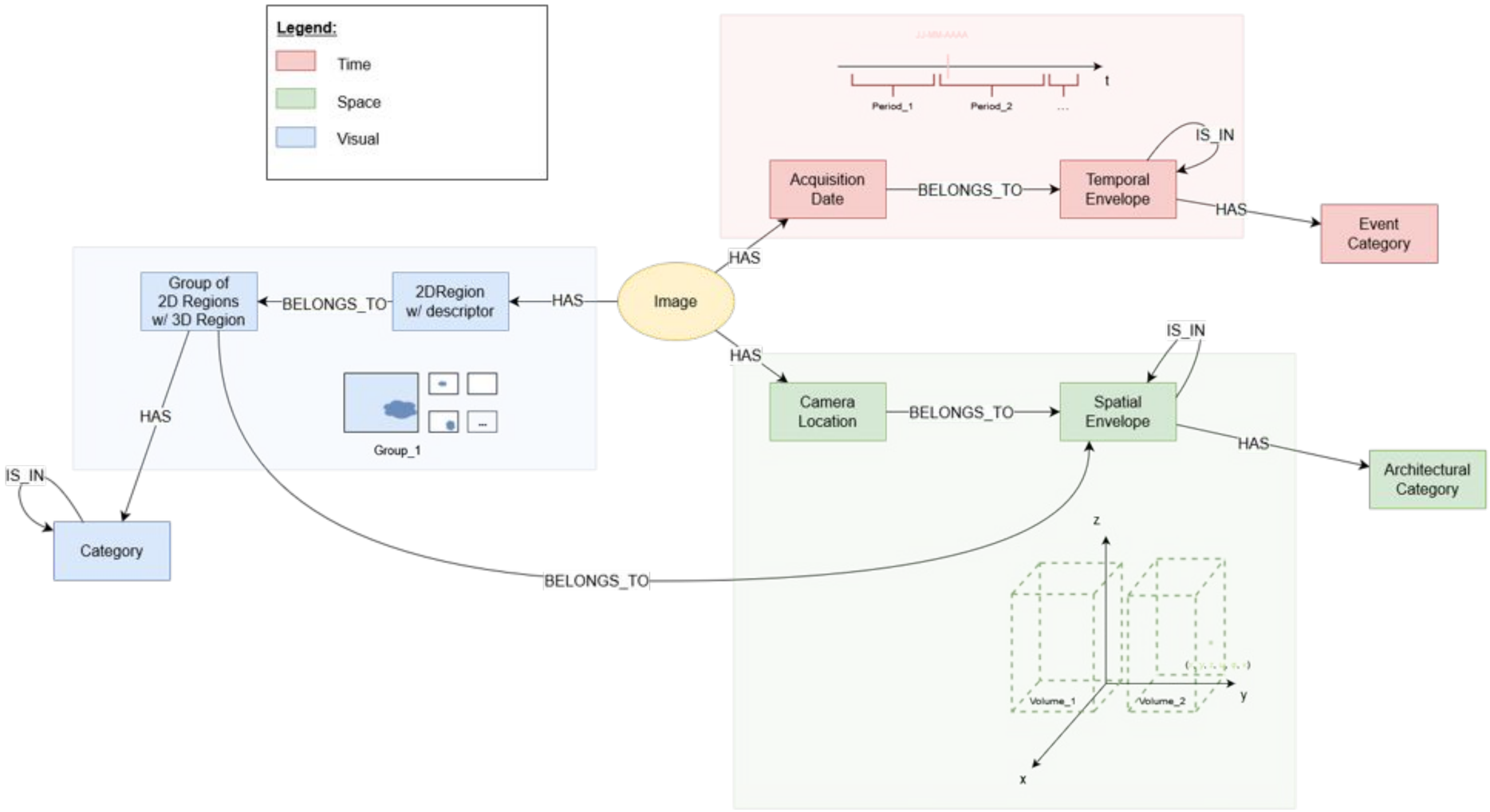
Foundations models for Detection & Semantic Segmentation





Réby, Kévin, Anaïs Guillem, and L. de Luca.,"Hybrid construction of Knowledge Graph and Deep Learning experiments for Notre-Dame de Paris' data", REAACH Symposium 2023

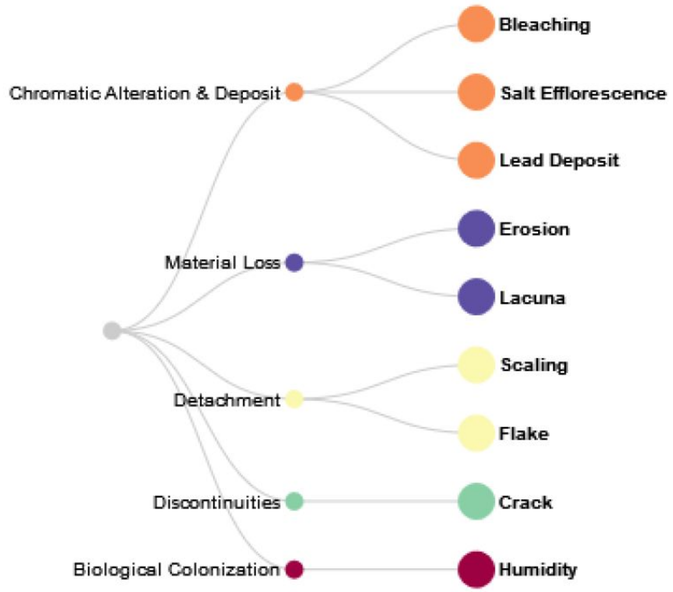
Knowledge Graph

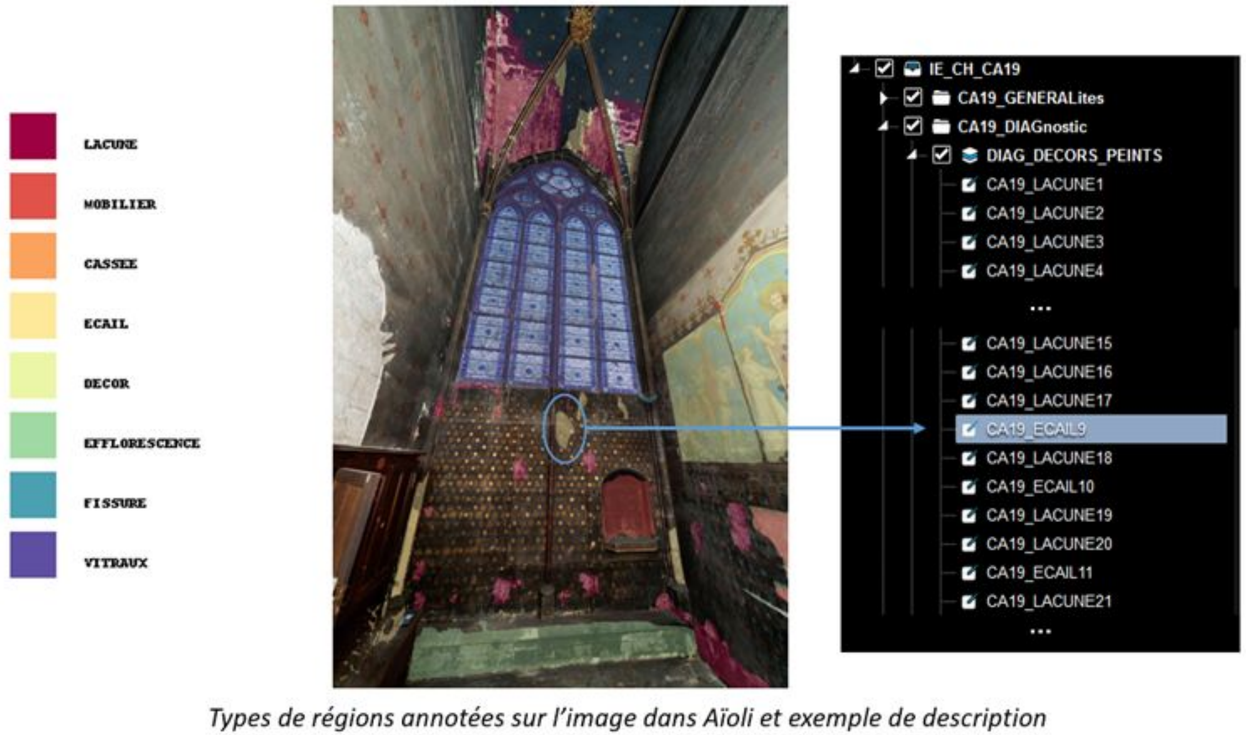


Laura Willot 2024,
ETIS, MAP, LASTIG, FSP

Degradation Dataset

Stone alteration patterns hierarchy :





Types de régions annotées sur l'image dans Aïoli et exemple de description

Laura Willot 2024,
ETIS, MAP, LASTIG, FSP

COCO format

1. **Informations**
2. **Images** : The dataset contains a list of images, each represented as an object with the following attributes:
   - id (unique integer identifier)
   - width and height (in pixels)
   - file_name (the file name of the image)
   - aioli_url (url to the image used in Aioli)
3. **Categories** : The dataset includes a list of object categories, each represented as a degradation phenomenon with the following attributes:
   - id (unique integer identifier)
   - name (the name of the category, e.g., "Bleaching", "Erosion", "Flake")
   - supercategory (the name of a higher-level category, e.g., "ChromaticAlteration&Deposit" for "Bleaching")
4. **Annotations** : The dataset contains a list of annotations, each representing an object instance in an image. The annotations have the following attributes:
   - id (unique integer identifier)
   - image_id (the id of the image this annotation refers to)
   - category_id (the id of the category this object belongs to)
   - **bbox** (a list of four floating-point numbers representing the bounding box of the object, in the format [x, y, width, height], where x and y are the coordinates of the top-left corner of the bounding box)
   - **segmentation** (list of pixel-wise segmentation masks for the object, represented as a list of polygons, where each polygon is a list of [x, y] coordinate pairs)


```
{
    "info": {
        "year": 2024,
        "date_created": "2024/05/29",
        "description": "IE_CH_CA19",
        "url": "https://public.aioli.map.cnrs.fr/spritz2_ldl/index.html?projectId=624c06c2bc7ad",
        "contributor": "MAP (CNRS, UPR 2002)",
        "version": "1.0"
    }
}
```

```
{
    "images": [
        {
            "id": 394351001015752577895758218986284477319272497 27,
            "file_name": "ITER_ca_19_04571.jpg",
            "date_captured": "2020-06-11 13:00:13",
            "aioli_url": "https://absinthe.aioli.map.cnrs.fr/workspace/usr/NDP/projects/624c06c2bc7ad/chantier/Images/ITER_ca_19_04571.JI
            "width": "6336",
            "height": "9504"
        },
        {
            "id": 394351001015752577895758218986284477319272497 28,
            "file_name": "ITER_ca_19_04572.jpg",
            "date_captured": "2020-06-11 13:00:50",
            "aioli_url": "https://absinthe.aioli.map.cnrs.fr/workspace/usr/NDP/projects/624c06c2bc7ad/chantier/Images/ITER_ca_19_04572.JI
            "width": "6336",
            "height": "9504"
        },
```

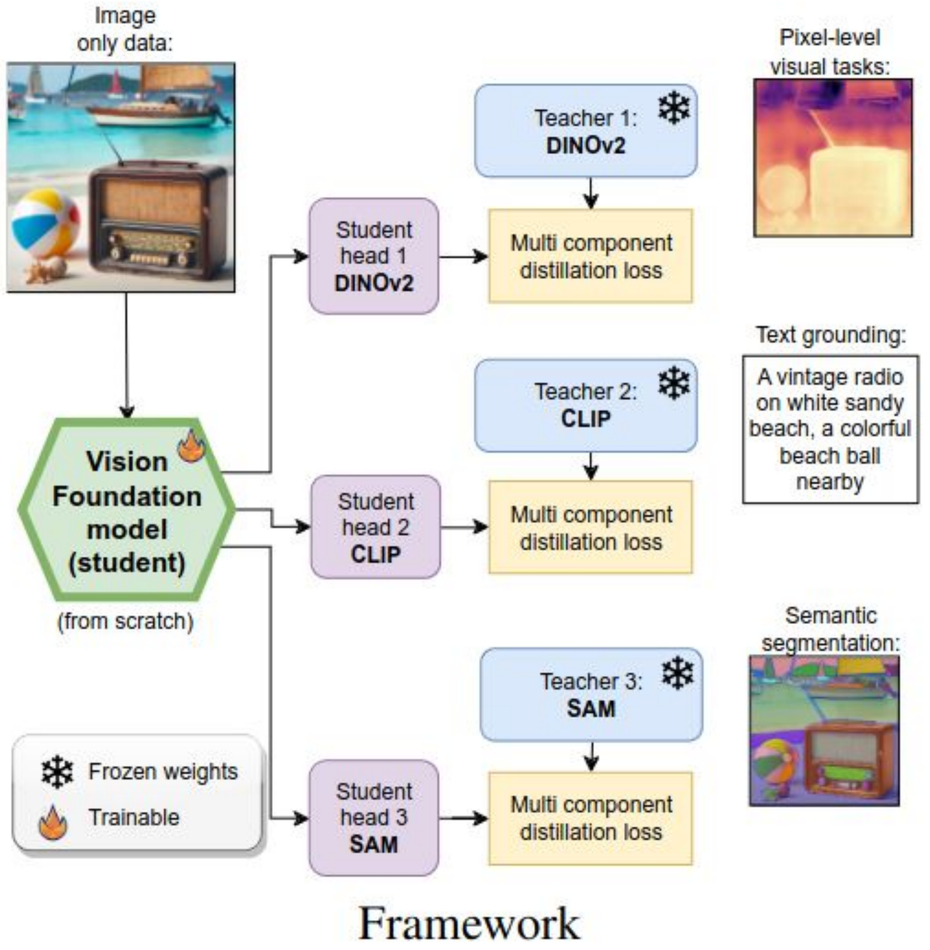
```
{
    "categories": [
        {
            "id": 2,
            "name": "Bleaching",
            "supercategory": "ChromaticAlteration&Deposit"
        },
        {
            "id": 5,
            "name": "SaltEfflorescence",
            "supercategory": "ChromaticAlteration&Deposit"
        },
        {
            "id": 6,
            "name": "LeadDeposit",
            "supercategory": "ChromaticAlteration&Deposit"
        },
        {
            "id": 1,
            "name": "Erosion",
            "supercategory": "MaterialLoss"
        },
        {
            "id": 8,
            "name": "Lacuna",
            "supercategory": "MaterialLoss"
        },
        {
            "id": 7,
            "name": "Scaling",
            "supercategory": "Detachment"
        },
        {
            "id": 9,
            "name": "Flake",
            "supercategory": "Detachment"
        },
        {
            "id": 3,
            "name": "Crack",
            "supercategory": "Discontinuities"
        },
        {
            "id": 4,
            "name": "Humidity",
            "supercategory": "BiologicalColonization"
        }
    ]
}
```

```
{
    "annotations": [
        {
            "id": 67083725421942087656052622490712771043516211607751 36,
            "image_id": 394351001015752577895758218986284477319272509 84,
            "category_id": 8,
            "bbox": [
                4369.460506135343,
                5783.439074012593,
                512.2218119739473,
                290.68233162931756
            ],
            "segmentation": [
                [
                    4726.463587205419,
                    6074.12140564191,
                    4696.830920395237,
                    6027.555789410735,
                    4719.408190339847,
                    6017.678234452701,
                    4720.819269712962,
                    5980.99017318271,
                    4734.930063431494,
                    5976.756935341076,
                    4729.285745951649,
                    5961.235063263491,
                    4737.752222177724,
```


Laura Willot 2024,
ETIS, MAP, LASTIG, FSP

Ongoing research

- Domain adaptation
  - Degradation phenomena
  - instance segmentation
- Fine-tuning models
- Segmentation dataset
  - COCO format
- Knowledge Distillation
- GPUs: Jean Zay



Ranzinger, Mike, et al. "AM-RADIO: Agglomerative Model--Reduce All Domains Into One." *arXiv preprint arXiv:2312.06709* (2023).

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation A0161015066 made by GENCI.

This work was funded since 2019 by the CNRS and the French Ministry of Culture within the framework of the national scientific action Notre-Dame de Paris, and since 2022 by the European Research Council (ERC Advanced Grant nDame_Heritage : n-Dimensional analysis and memorization ecosystem for building cathedrals of knowledge in Heritage Science). I also wish to thank Roxane Roussel, Florent Comte and AGP who shared their acquisition campaigns, and all the scientific partners and collaborators working on Notre-Dame de Paris.

Thank you !