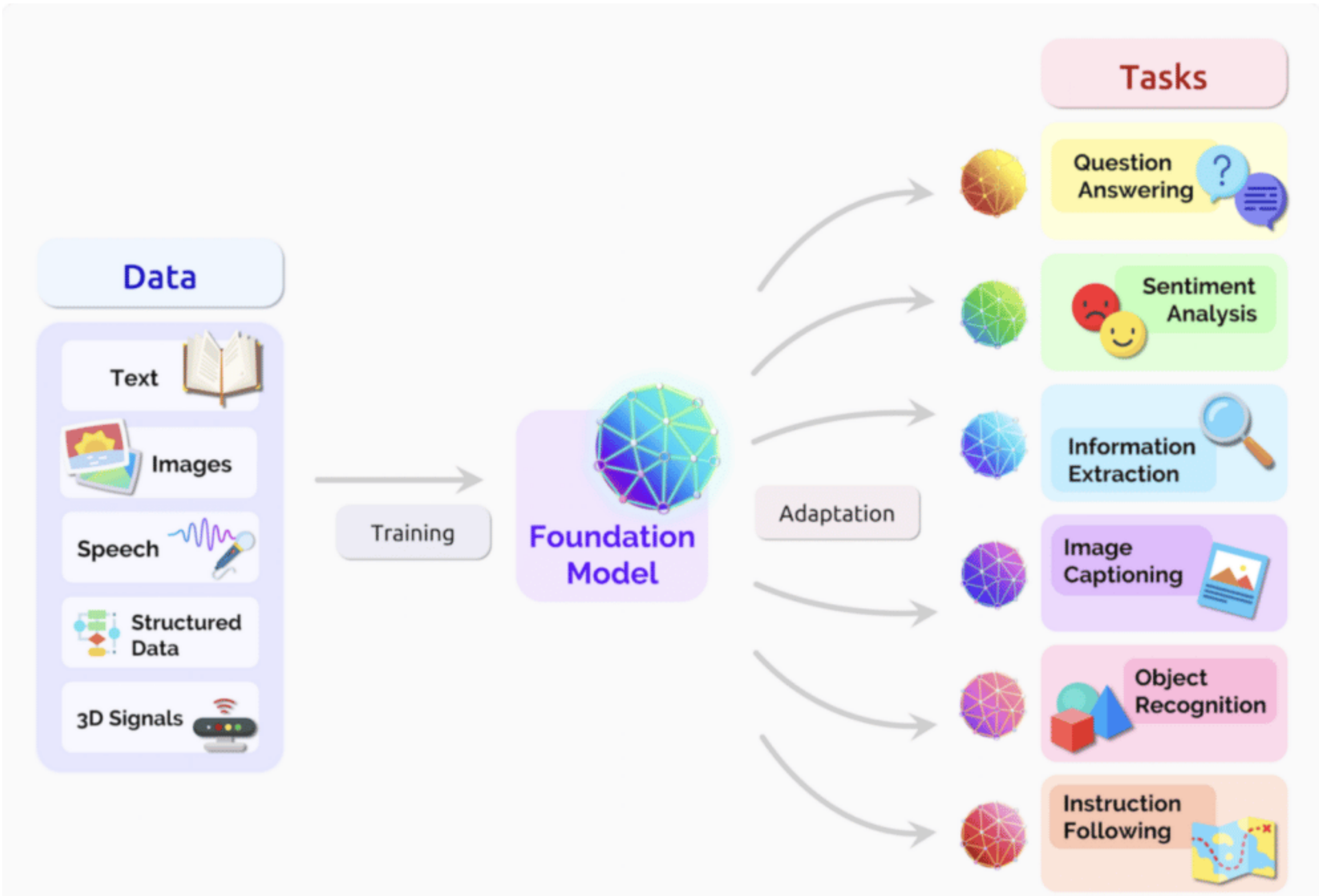


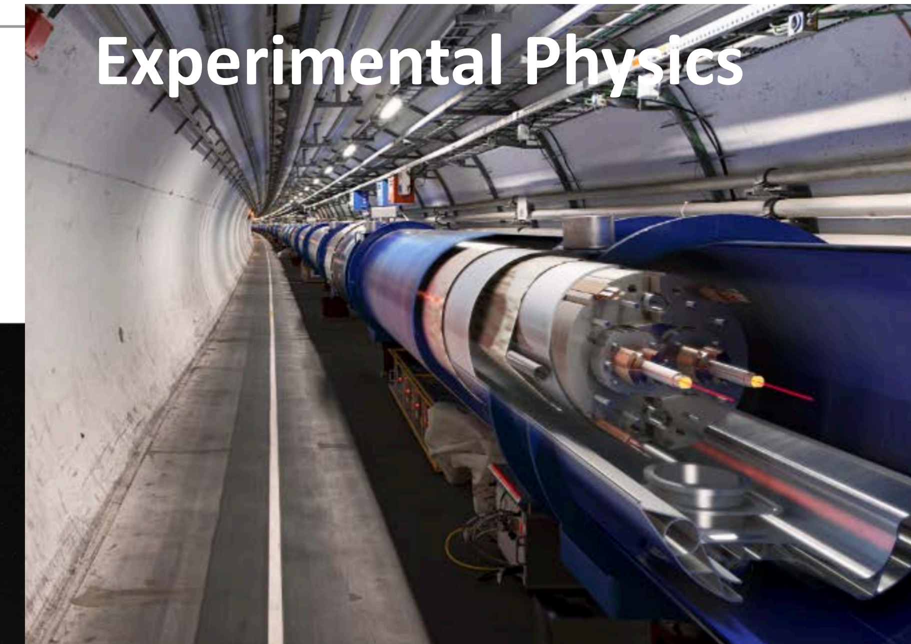
Training and fine-tuning foundation models: State of the art and future challenges

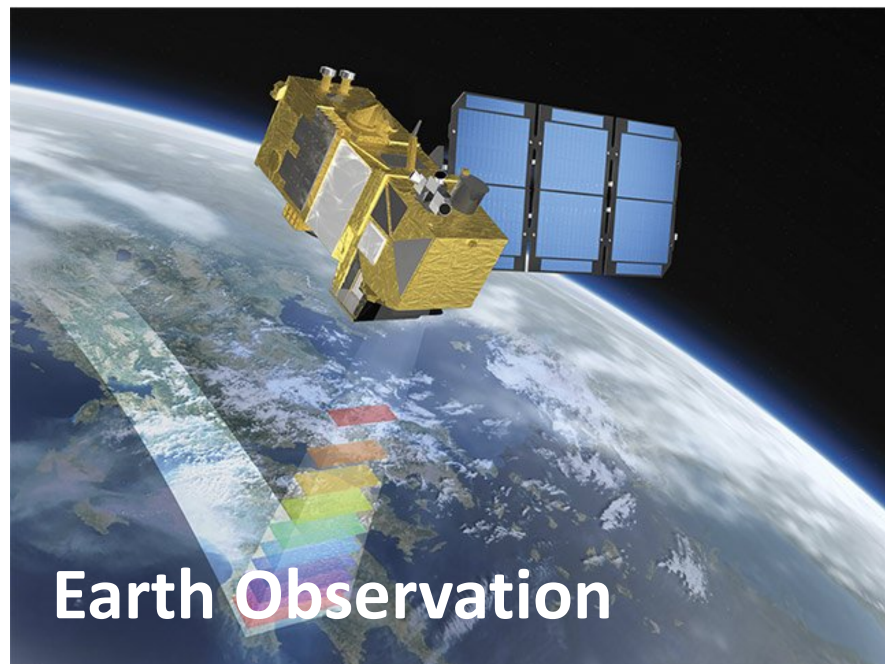
Ilaria Luise | Sofia Vallecorsa

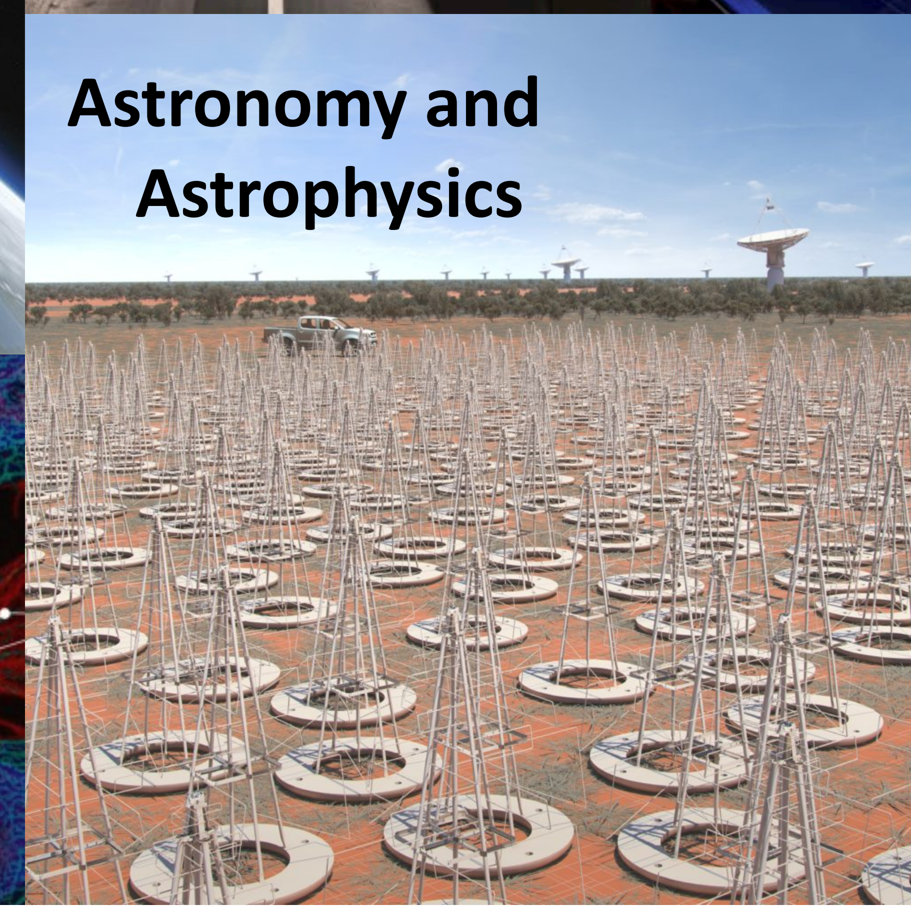
CERN OpenLab summer student lectures
26th July 2024

Big Data in Science

Science produces more data than ever before and at an unmatched pace in history





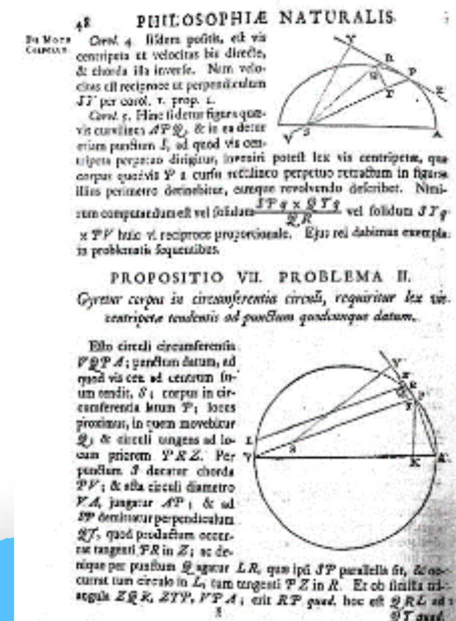


Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Four Paradigms of Scientific Research

4000 years

Empirical observations

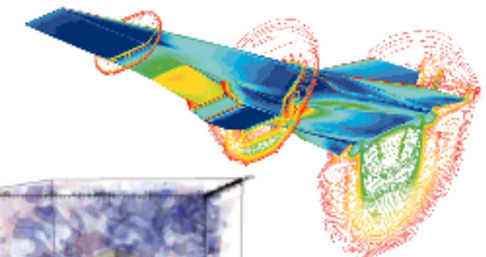


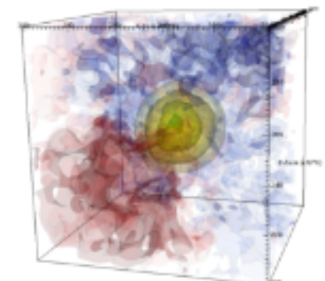
PHILOSOPHIÆ NATURALIS.

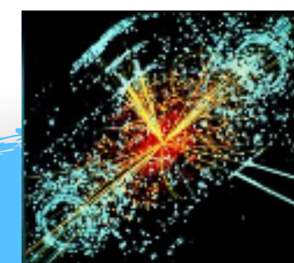
PROPOSITIO VII. PROBLEMA II.

500 years

Generalization
Theoretical models

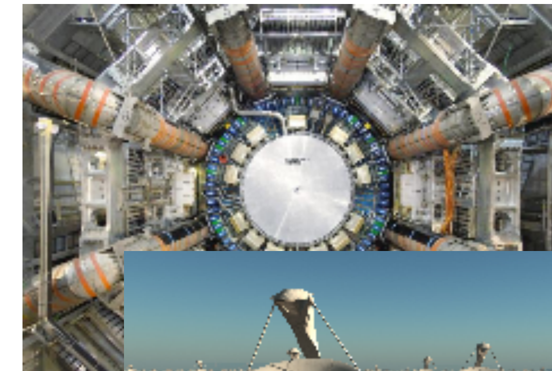


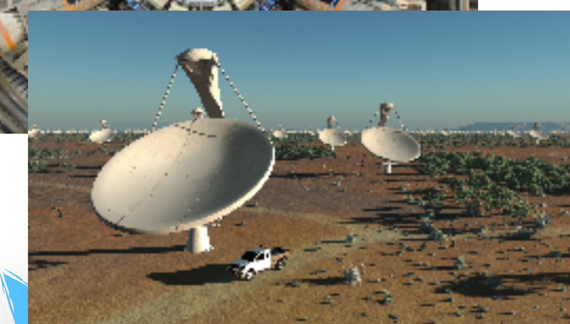




~50 years

Simulations
Computational sciences



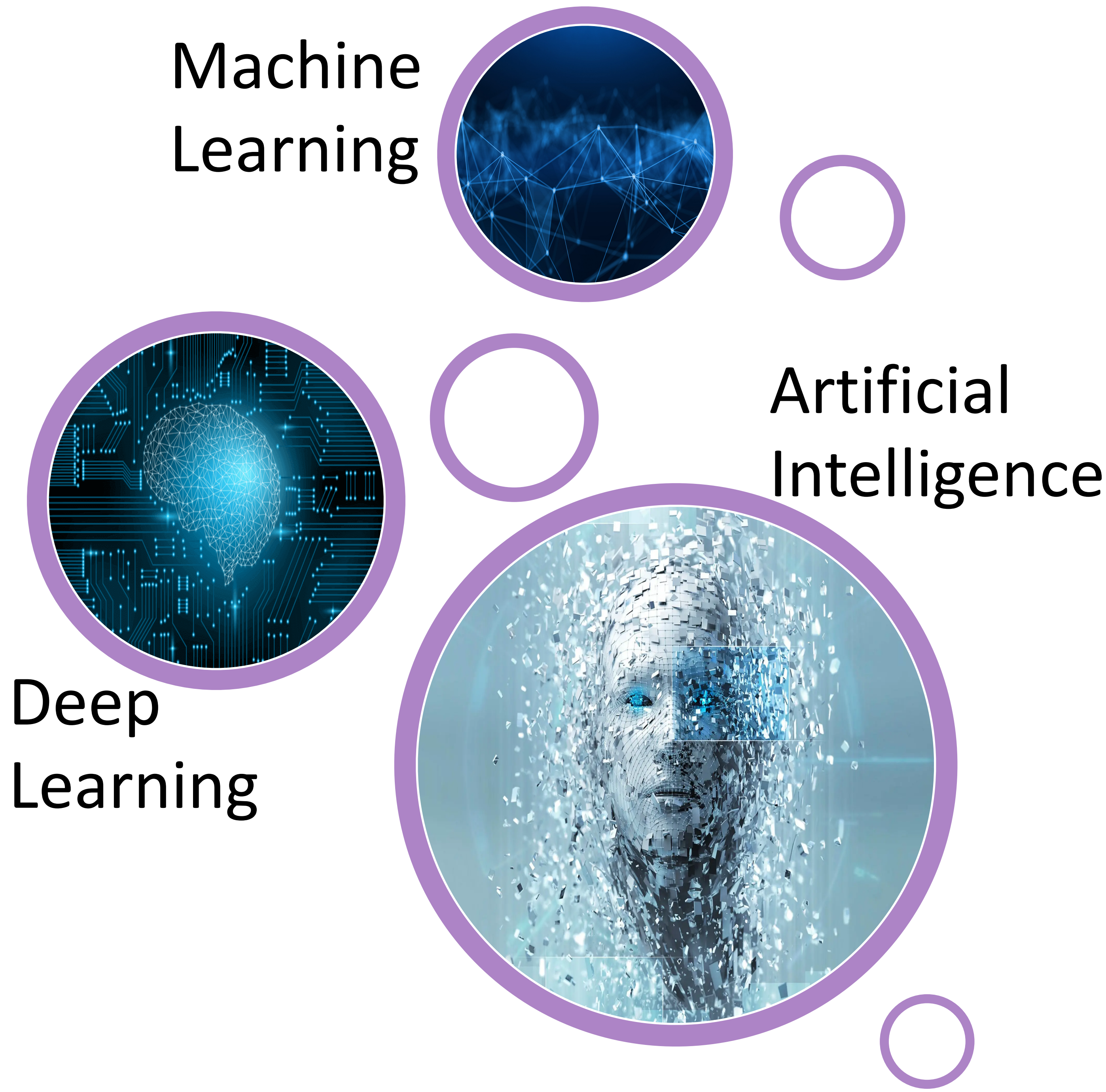


Today

Data-driven science

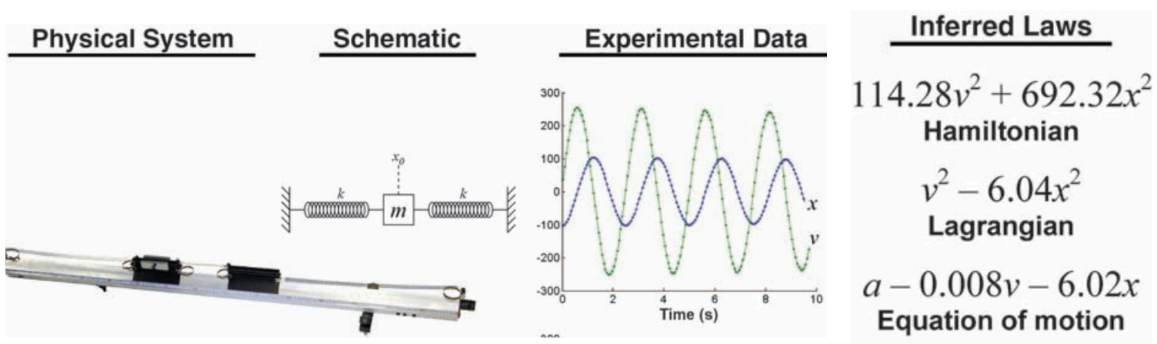
Data-driven science & AI

Is Artificial Intelligence just a **refined, faster** approach to computational science?

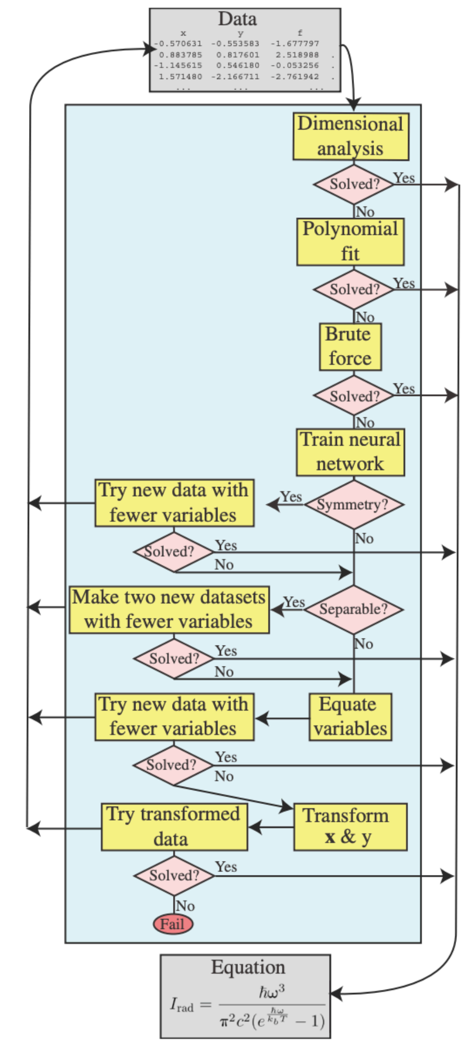


Rediscovering physics

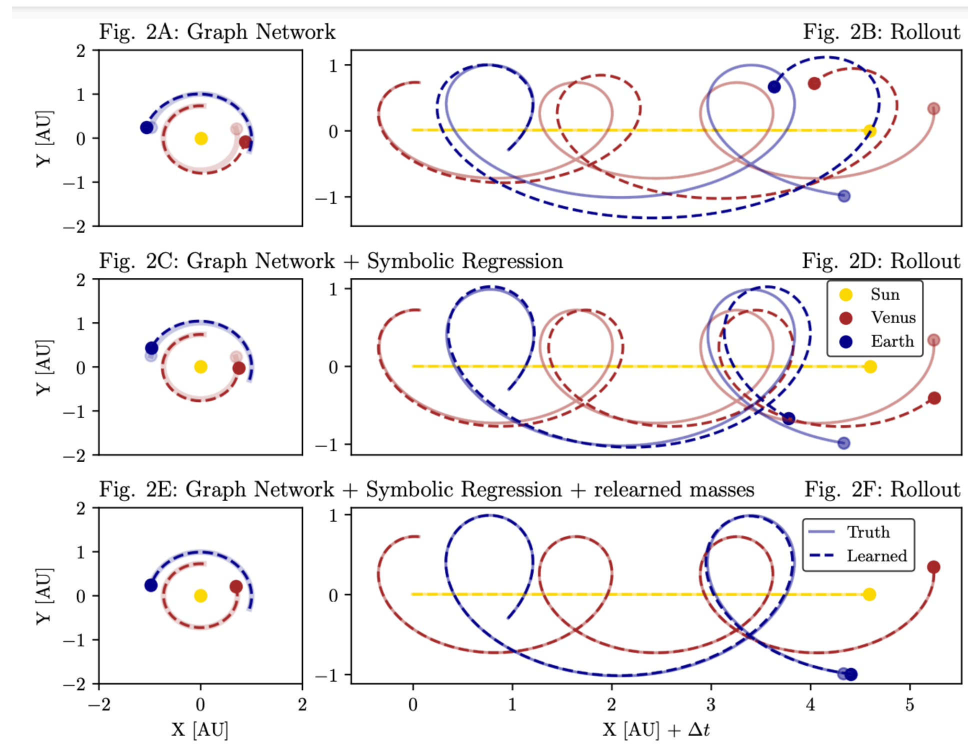
Schmidt, Michael, and Hod Lipson. "**Distilling free-form natural laws from experimental data**." *science* 324.5923 (2009): 81-85.



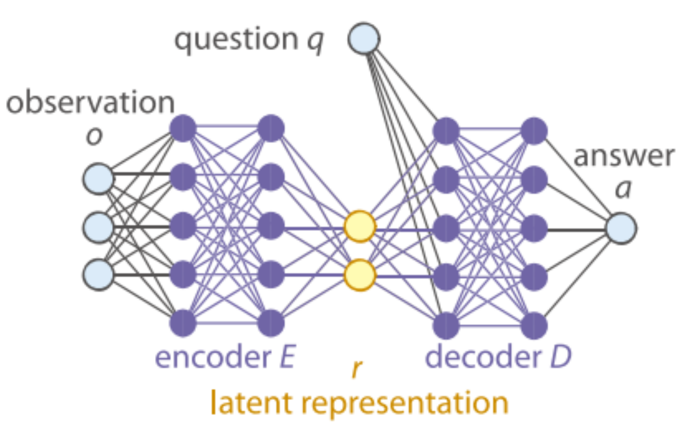
Udrescu, Silviu-Marian, and Max Tegmark. "**AI Feynman: A physics-inspired method for symbolic regression**." *Science Advances* 6.16 (2020): eaay2631.

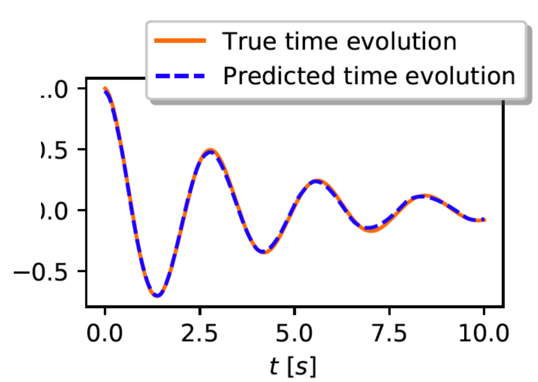


Lemos, Pablo, et al. "**Rediscovering orbital mechanics with machine learning**." *arXiv:2202.02306* (2022)



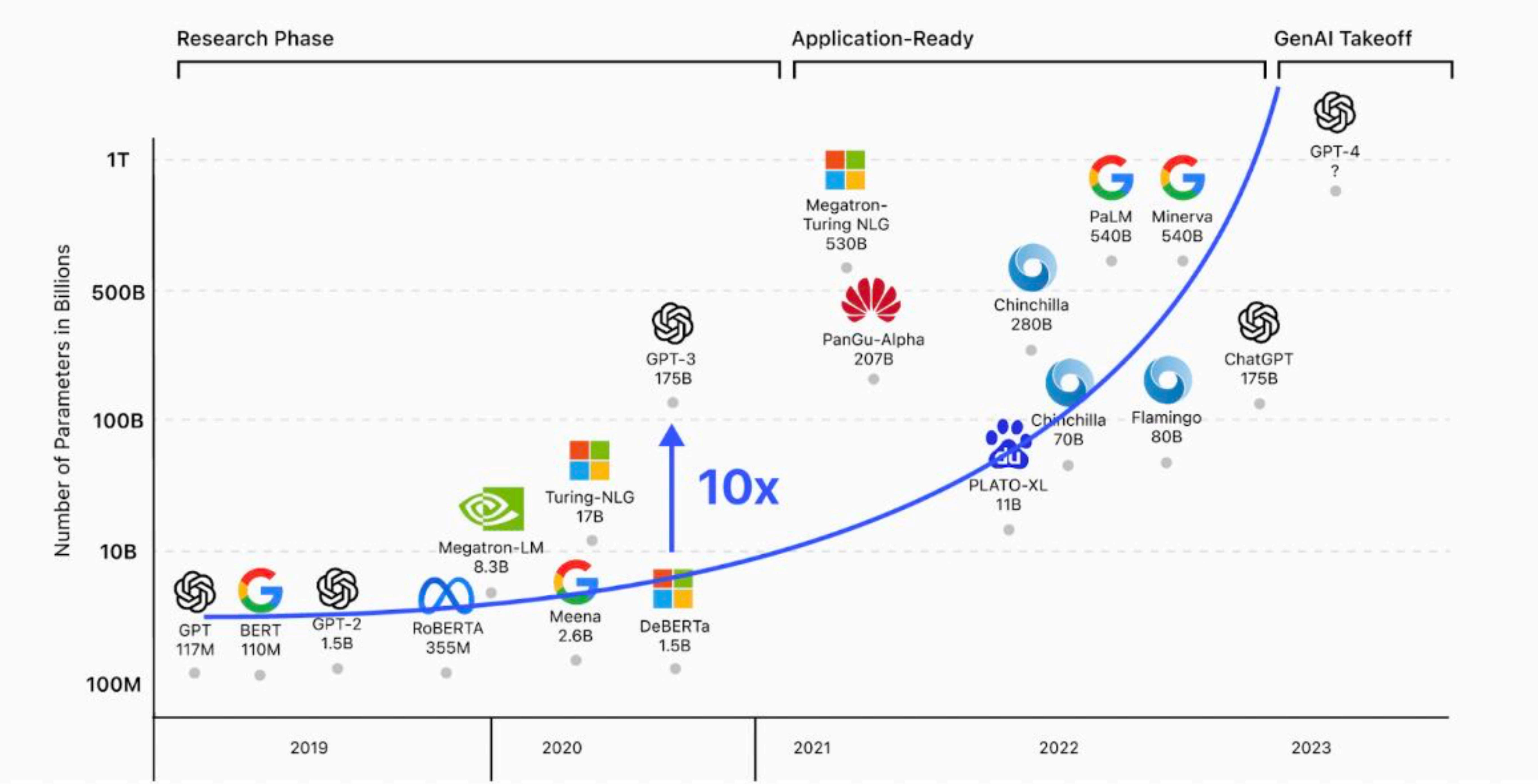
Iten, Raban, et al. "**Discovering physical concepts with neural networks**." *Physical review letters* 124.1 (2020): 010508.





Can we train AI to understand **physics itself** in order to achieve new discoveries ?

Existing models

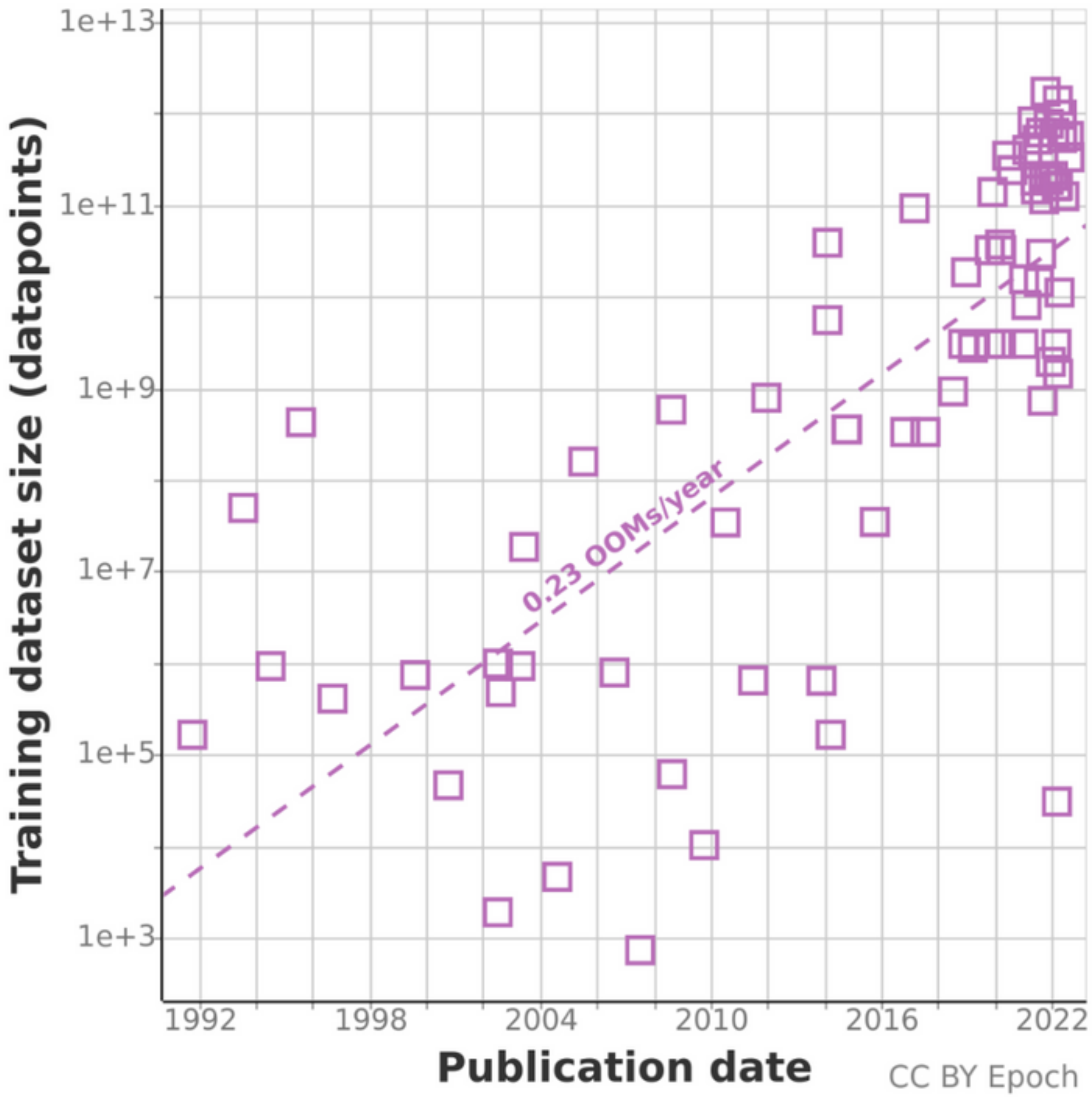


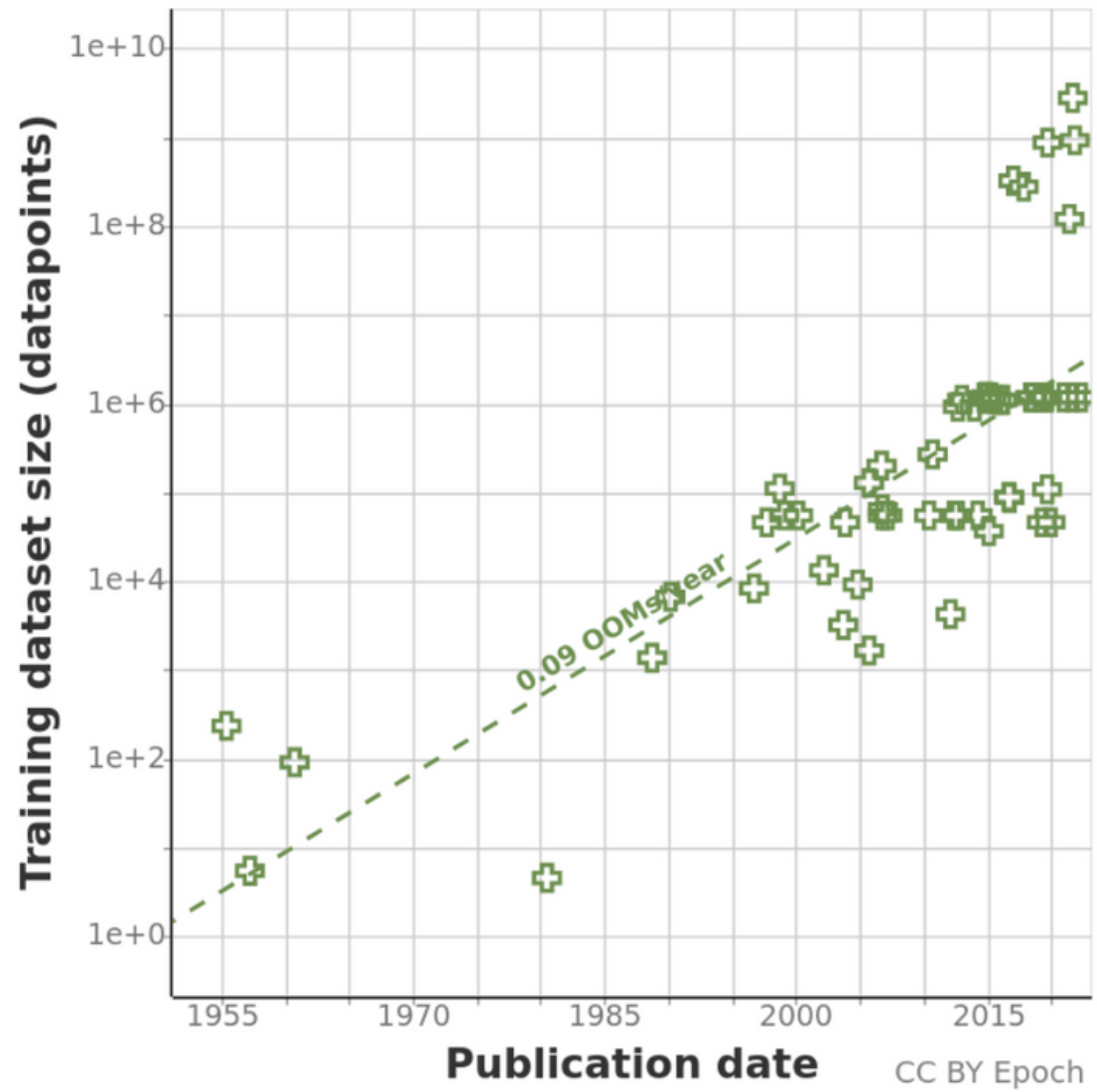
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Dataset sizes

Domain	Data points
Vision	#Images (eg: a model trained on 3B images has a dataset size of 3B)
Language	#Words (eg: a model trained on 1T English tokens has a dataset size of ~750B words, the exact quantity depends on the tokenization)

Training datasets for language (left) and vision (right)





https://epochai.org/blog/trends-in-training-dataset-sizes

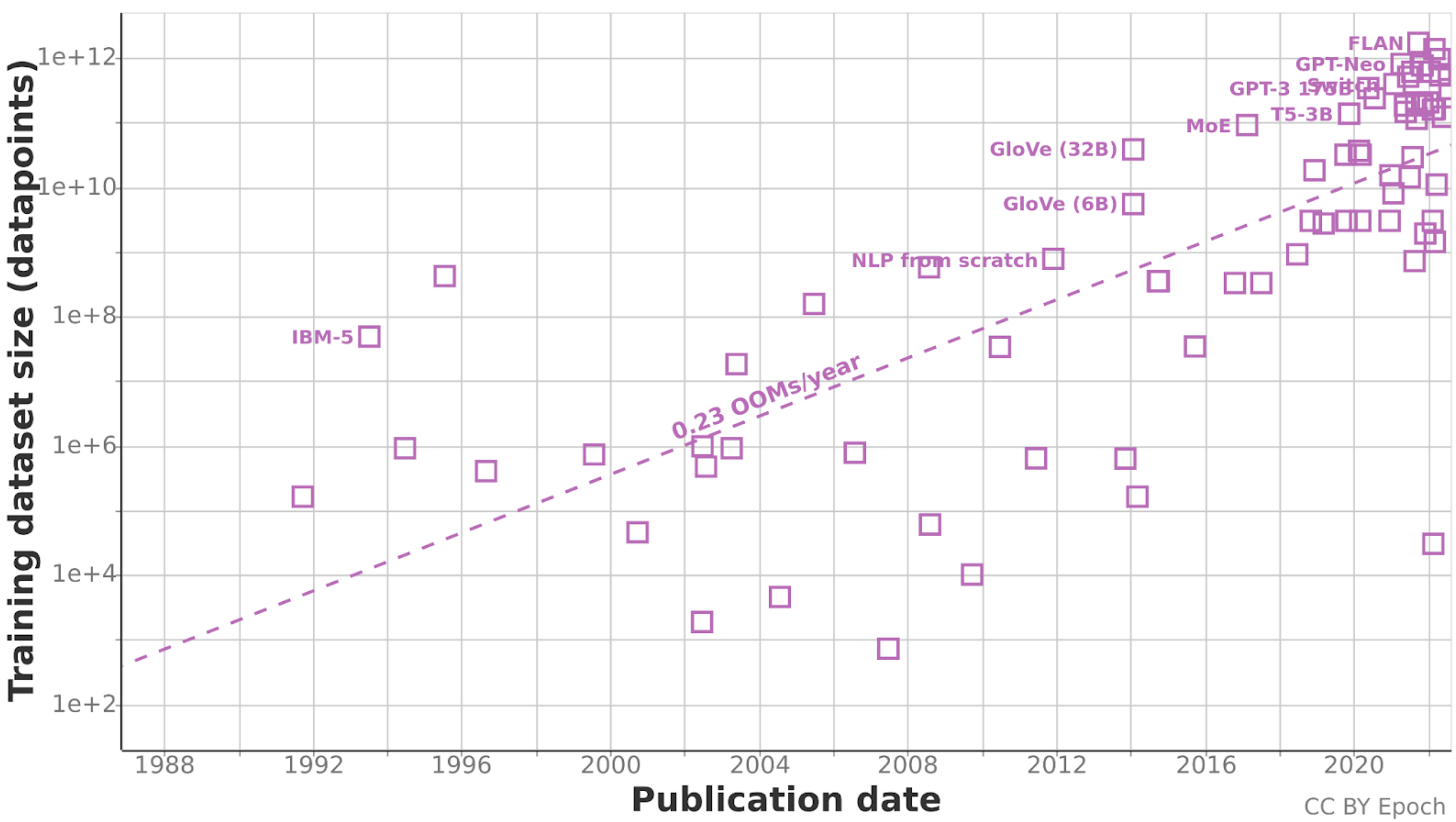
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Machine learning at scale, for science

Machine learning has been proven a very good tool to:

- Extract information from (very large) datasets
- Efficiently analyse very large amounts of data
- Easily handle data from different sources
- Scalability to HPC environments

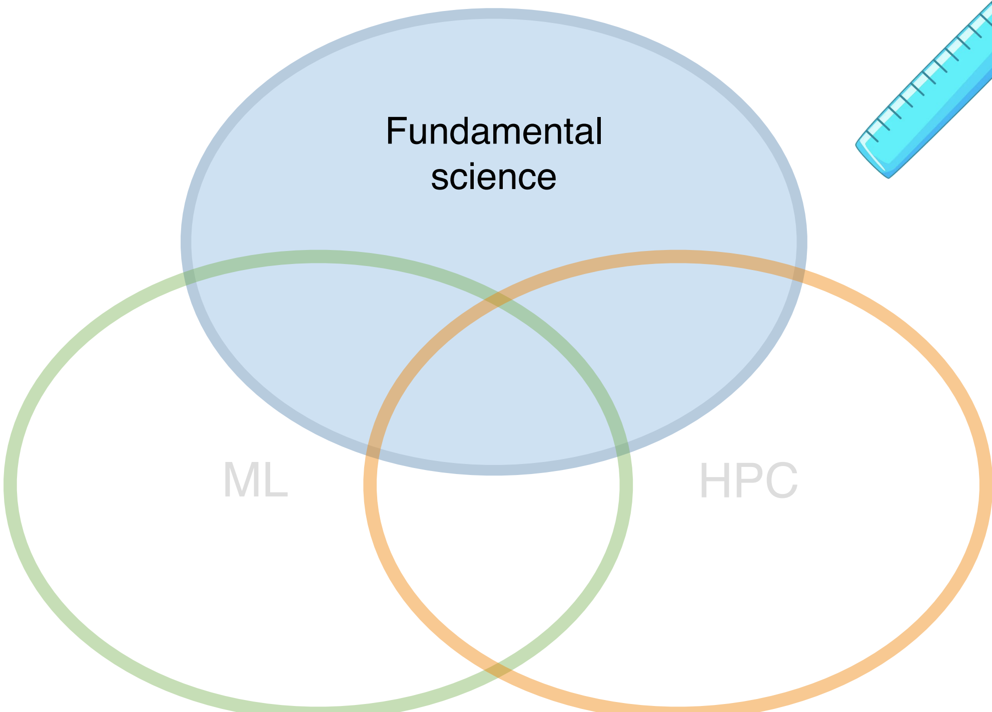
Observation based datasets in physics are comparable or larger than these!



Can we use these tools for fully data-driven science?

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Scientific opportunities



Multi-scale dependencies:
- **Model complex higher-order, statistical relationships between observations, fields, …**
- improve current simulations

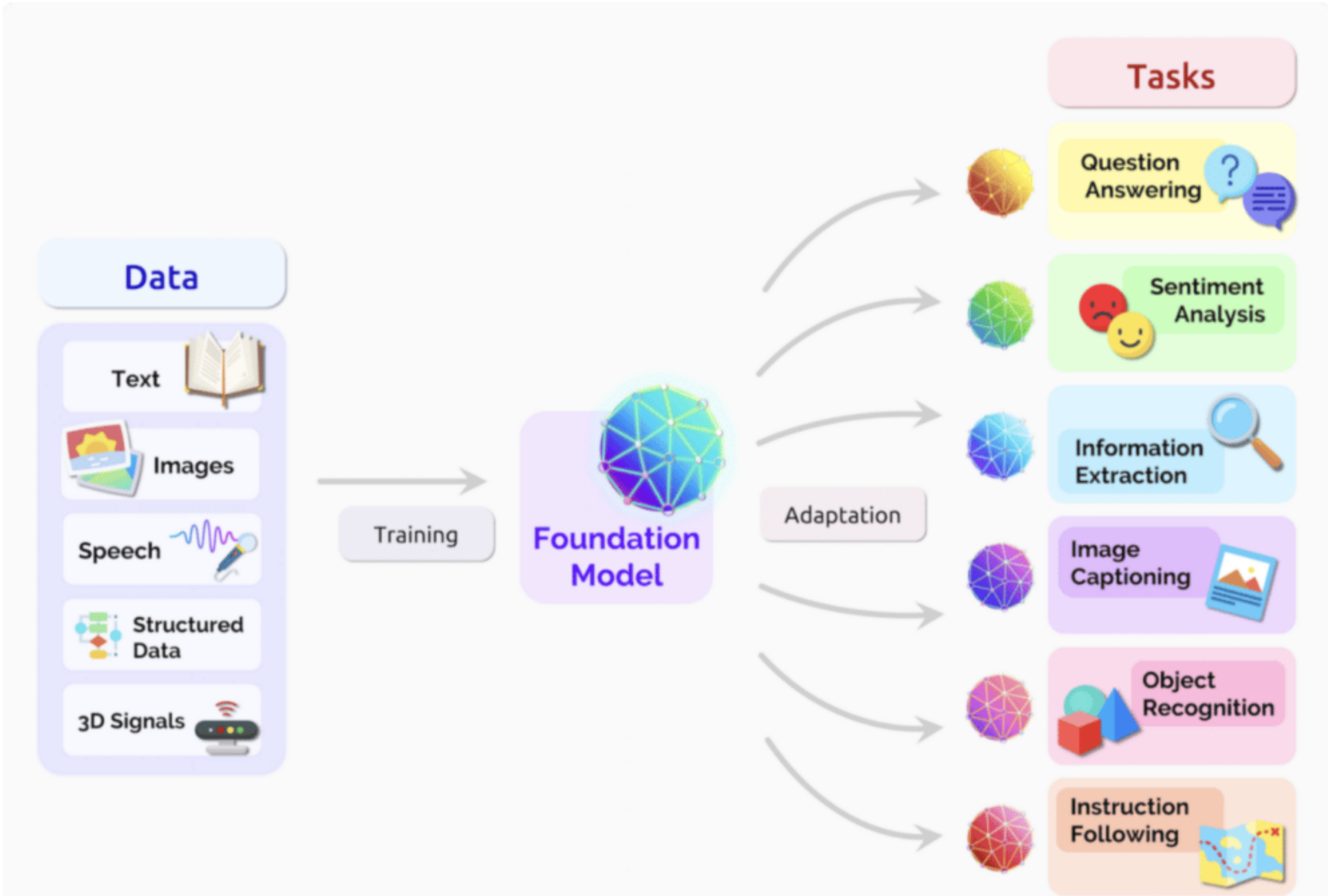
Compact representations:
- **Condense dataset information in a compact representation**
- eg. condense the info in a few GB rather than TB

Multi-source models:
- **Enable multimodal and multi-source learning**
- eg. build models based on scientific data, GDP, birth rate etc..

New discoveries:
- **Explore the potential of unsupervised learning to extract new information directly from data**
- Learn unknown correlation patterns

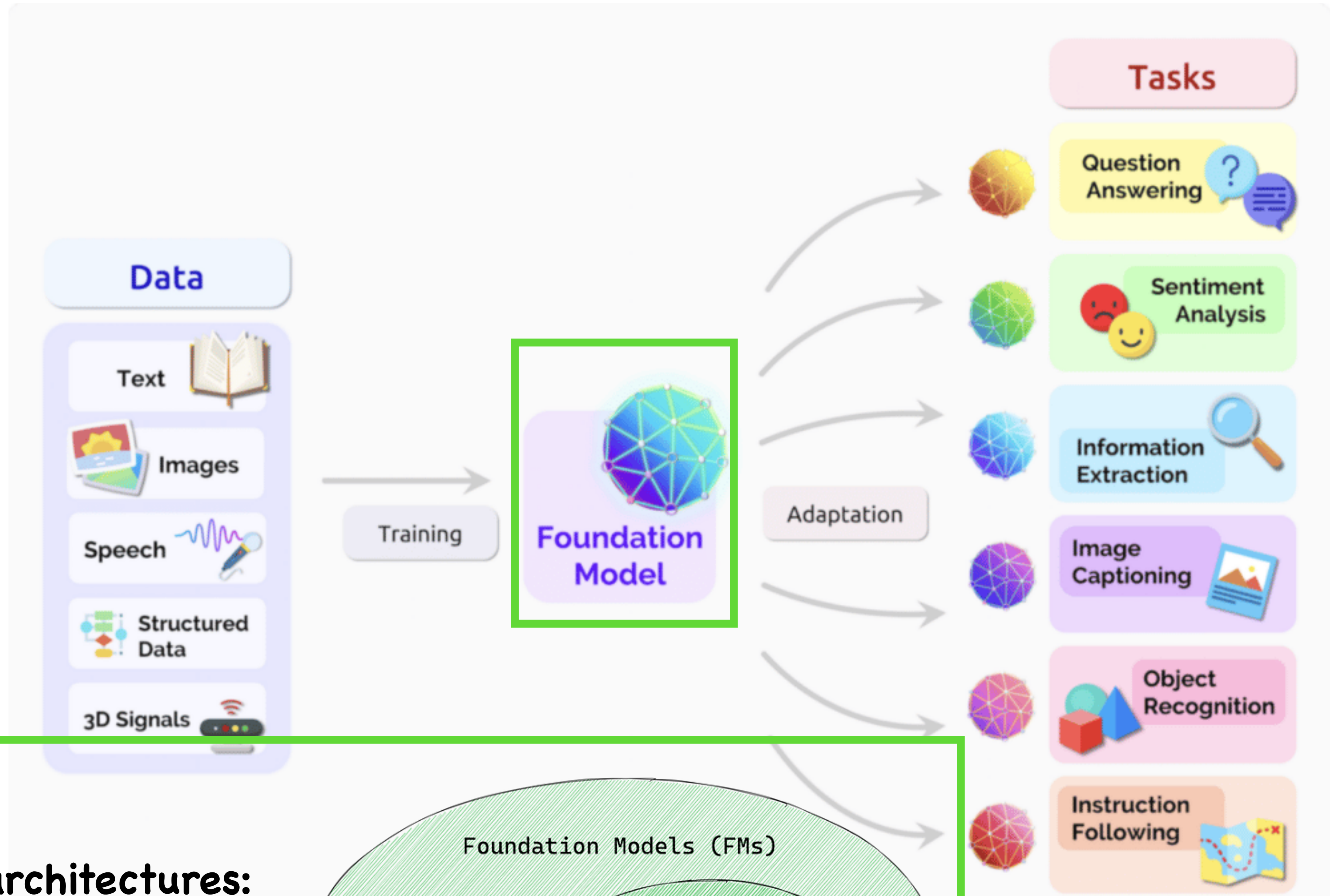
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Introduction

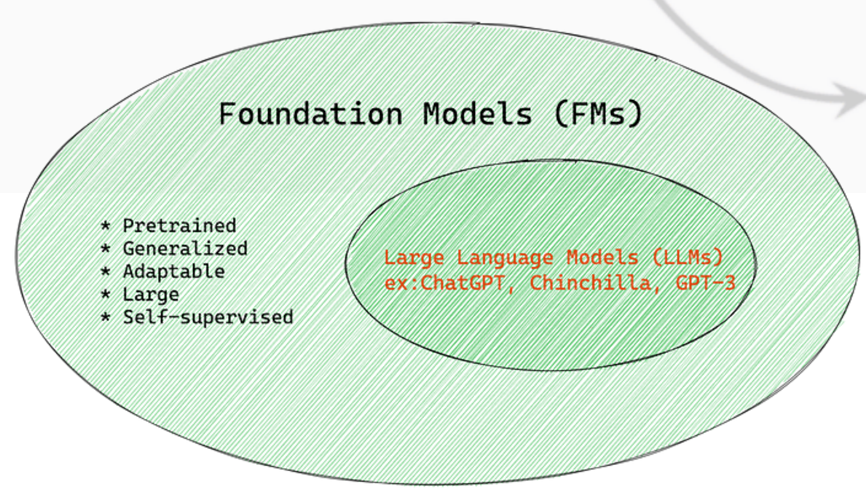


Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Introduction



Model architectures:
Generative Models
Attention & Self-Attention
Transformers



FMs are models trained on broad data (using self-supervision at scale) that can be adapted to a wide range of downstream tasks.
https://hai.stanford.edu/news/reflections-foundation-models

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Representational power

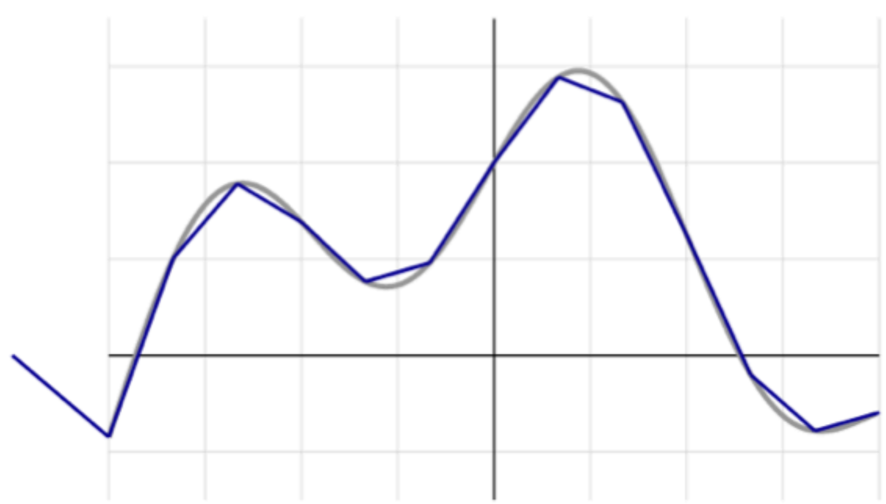
See M. Kagan lecture on July 5th :
https://indico.cern.ch/event/1392500/

Universal approximation theorem

77

- Feed-forward neural network with a single hidden layer containing a finite number of non-linear neurons (ReLU, Sigmoid, and others) can approximate continuous functions arbitrarily well on a compact space of $\mathbb{R}^n$

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



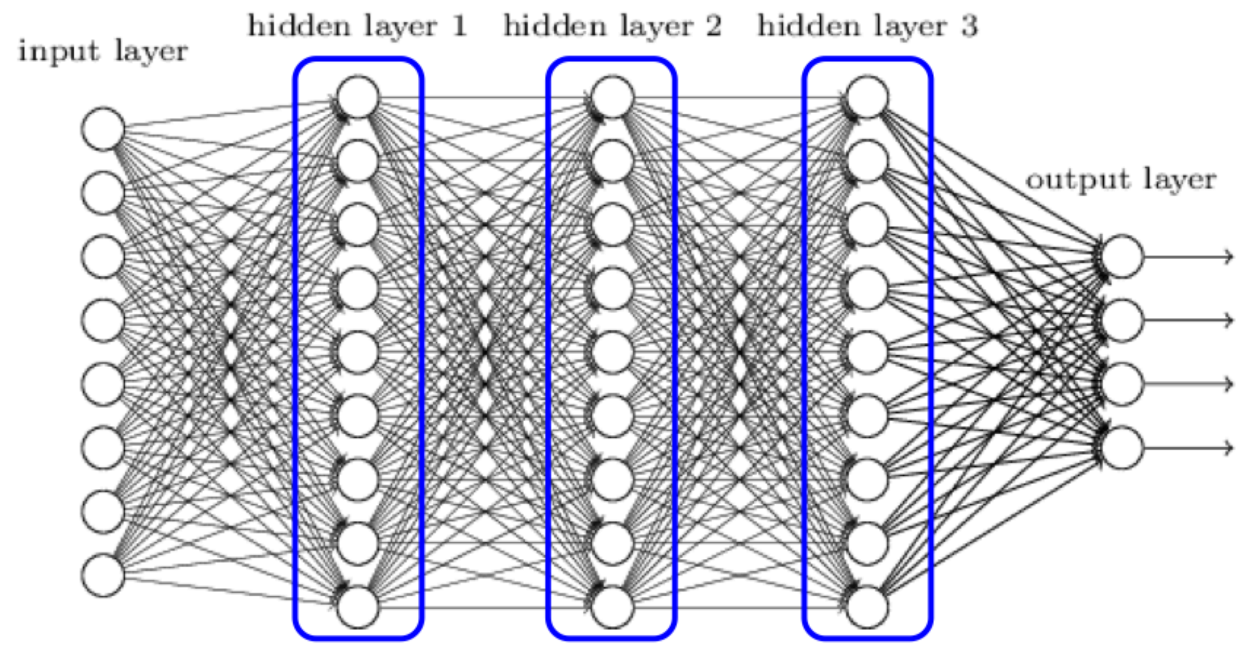
Fleuret, Deep Learning Course

Deep Neural Networks

See M. Kagan lecture on July 5th :
https://indico.cern.ch/event/1392500/

Deep Neural Networks

4



- As data complexity grows, need exponentially large number of neurons in a single-hidden-layer network to capture all structure in data
- Deep networks *factorize learning* of structure in data across layers
- Large datasets, fast computing (GPU / TPU) and new training procedures / network structures made training possible

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Generative Models and Representation Learning

Generative models

The problem:

Assume data sample follows p_{data} distribution

Can we draw samples from distribution p_{model} such that $p_{model} \approx p_{data}$?

31

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Generative models

The problem:

Assume data sample follows p_{data} distribution

Can we draw samples from distribution p_{model} such that $p_{model} \approx p_{data}$?

Maximum Likelihood Estimator:

- Assume some form for p_{model} (prior knowledge, parameterized by θ)
- draw samples from $p_{\theta *}$

$$\theta^* = \arg\max_{\theta} \sum_{\mathbf{x} \in \mathcal{D}} \log(p_{\text{model}}(\mathbf{x}; \theta))$$

Generative models don't look for mathematical expression of p_{model}

Train NN as a generator $g : \mathbb{R}^m \to \mathbb{R}^n$

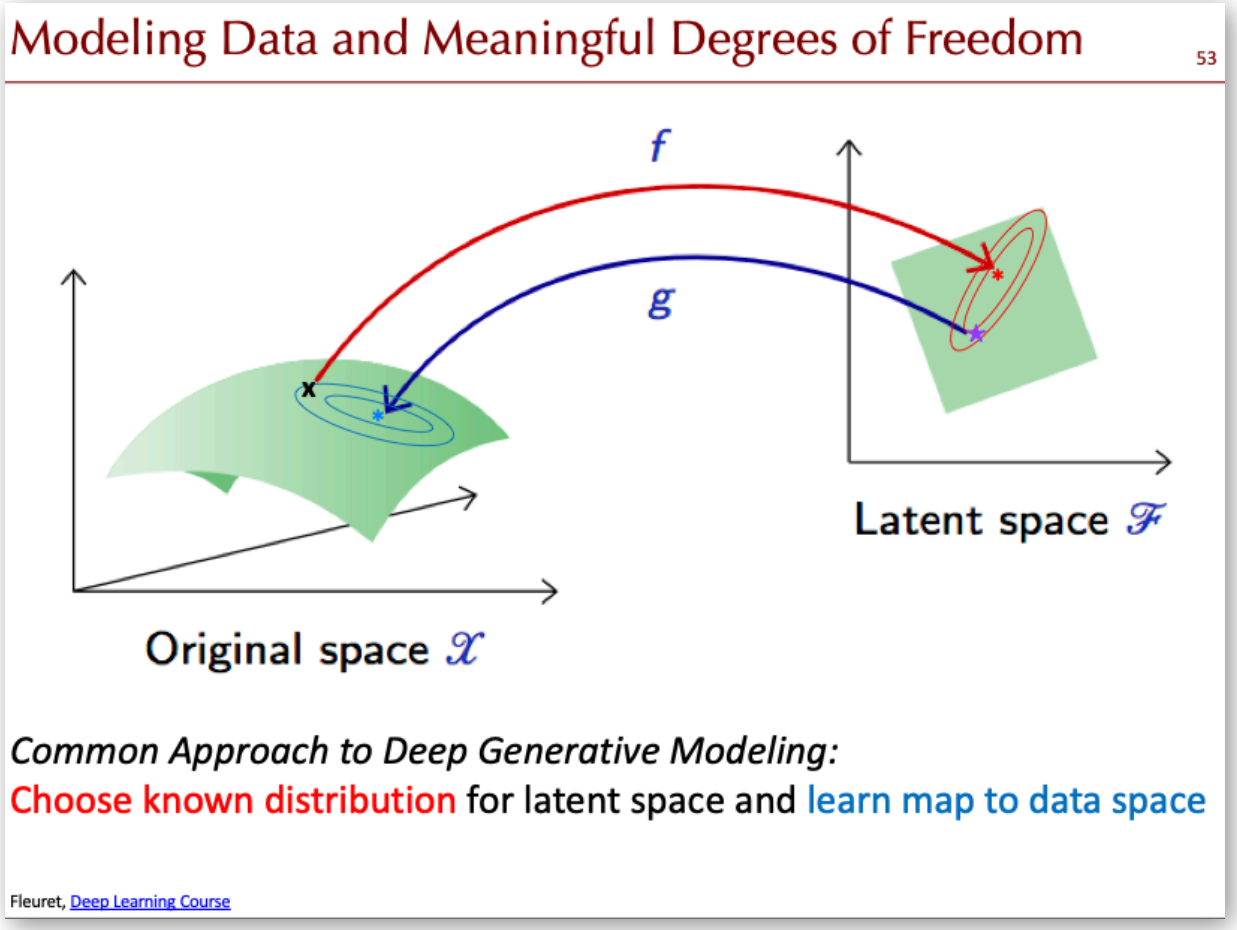
that maps samples from a tractable distribution supported in $\mathbb{R}^m$ to points in $\mathbb{R}^n$

31

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Latent Representation

See M. Kagan lecture on July 5th :
https://indico.cern.ch/event/1392500/

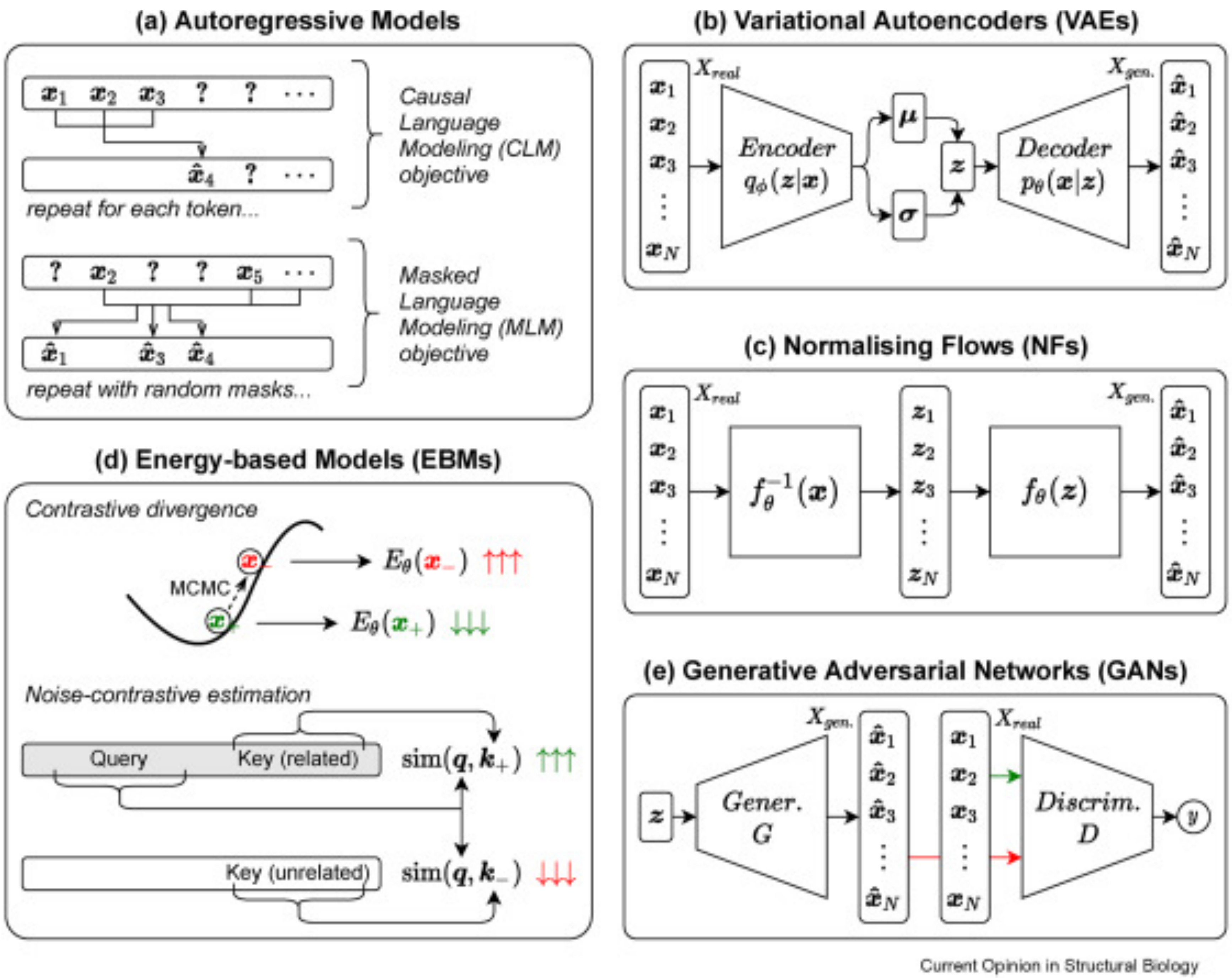


- Information content is preserved within a **hidden manifold with lower dimension**
- Can manipulate **latent space** (style specification, hypothesis testing directly in data, …)
- Can optimise latent representation according to a specific task (**guided compression**)
- Can help with **multi-modality**

NB: Problems exhibiting complex symmetries may benefit from latent space representations connected to the specific underlying symmetry group!

Deep Generative Models

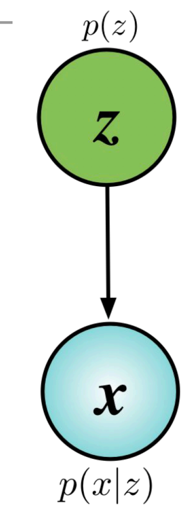
Deep models allow **higher levels of abstractions** and **improve generalization** wrt to shallow models



See Danilo Rezende tutorial on Deep Generative Models

Auto-Encoders

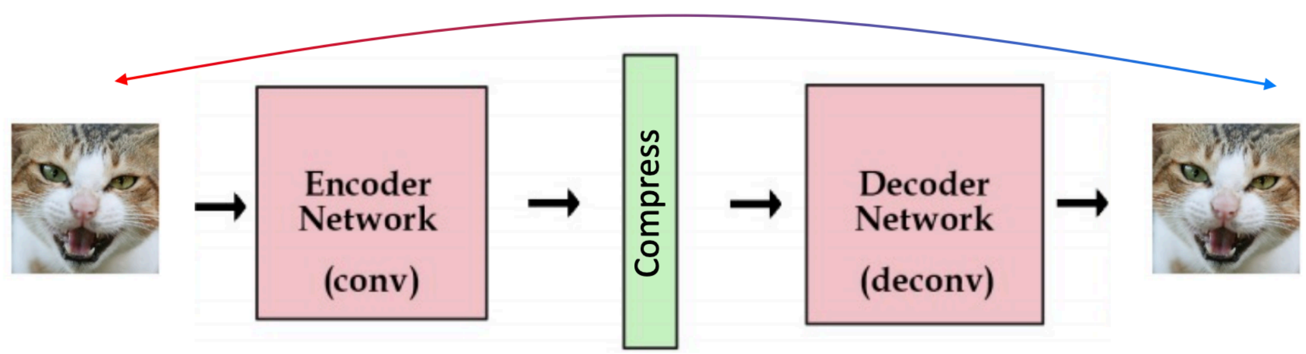
Examples of latent variables models (and implicit..)



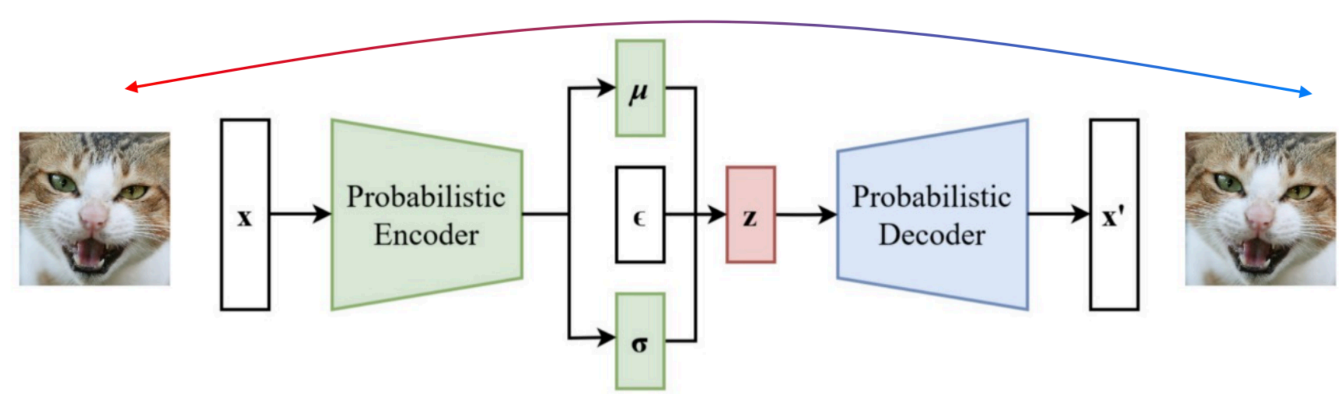
$x \in \mathbb{R}^{d_x} \quad z \in \mathbb{R}^{d_z} \quad \theta \in \mathbb{R}^{d_\theta}$

$\mathcal{D} = \{x_i\} \quad i \in \{1, ..., N\}$

Ex. Auto-Encoder



Ex. Variational Auto-Encoder

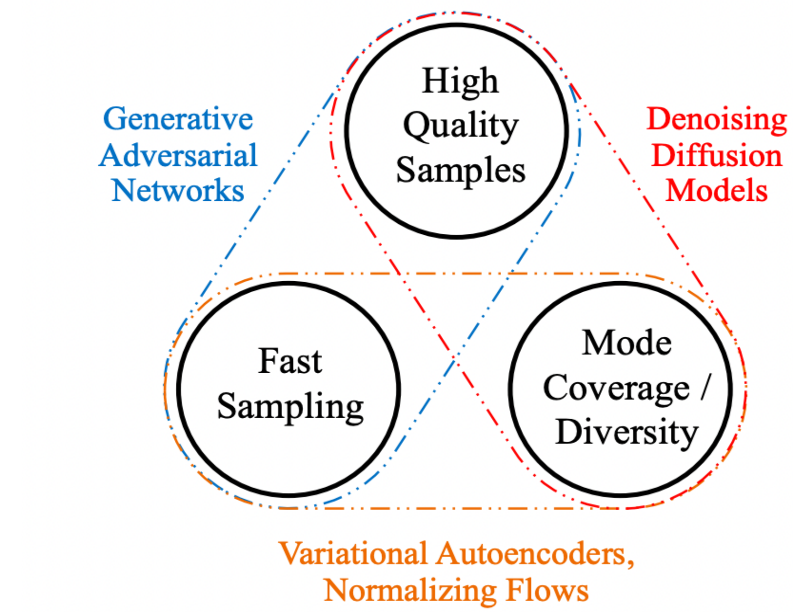


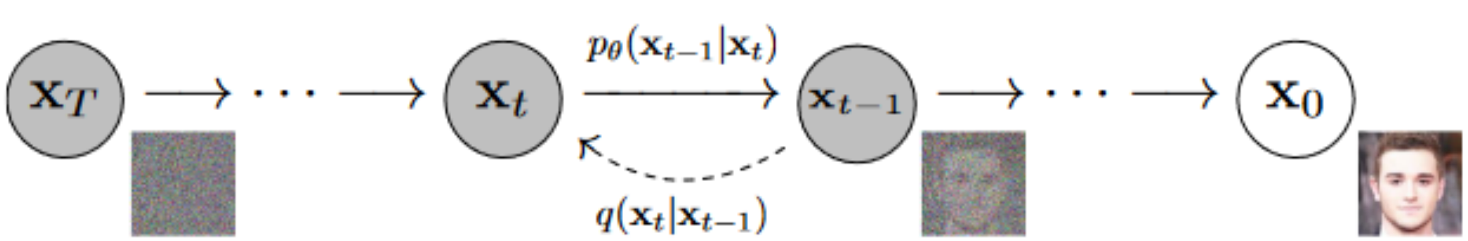
Explicit constraints on encoded representations (learn the **latent variable distribution**)

Two components in the loss function (**reconstruction loss and KL divergence** to constrain latent to prior)

Diffusion models

- **Parametrized Markov Chains** trained using variational inference to produce samples matching the data after finite time.
  - Chain transitions are **reverse diffusions** (gradually adding noise to the data)
- Ex. DDPM (Diffusion Denoising Probabilistic Models) based on U-Net architecture, https://arxiv.org/pdf/2006.11239.pdf:
  - Iteratively add Gaussian noise to input image, eventually reaching pure noise
  - Generation process **inverts the diffusion:** start from pure noise sample, then iteratively de-noise it.





Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Attention and Transformers

A step back

See M. Kagan lecture on July 5th :
https://indico.cern.ch/event/1392500/

Recurrent States

29

- Input sequence $x \in S(\mathbb{R}^m)$ of *variable* length $T(x)$
- Recurrent model maintain a **recurrent state** $\boldsymbol{h}_t \in \mathbb{R}^q$ updated at each time step t. For $t = 1, \dots, T(x)$:

$$\boldsymbol{h}_{t+1} = \phi(\boldsymbol{x}_t, \boldsymbol{h}_t; \theta)$$

  – Simplest model:

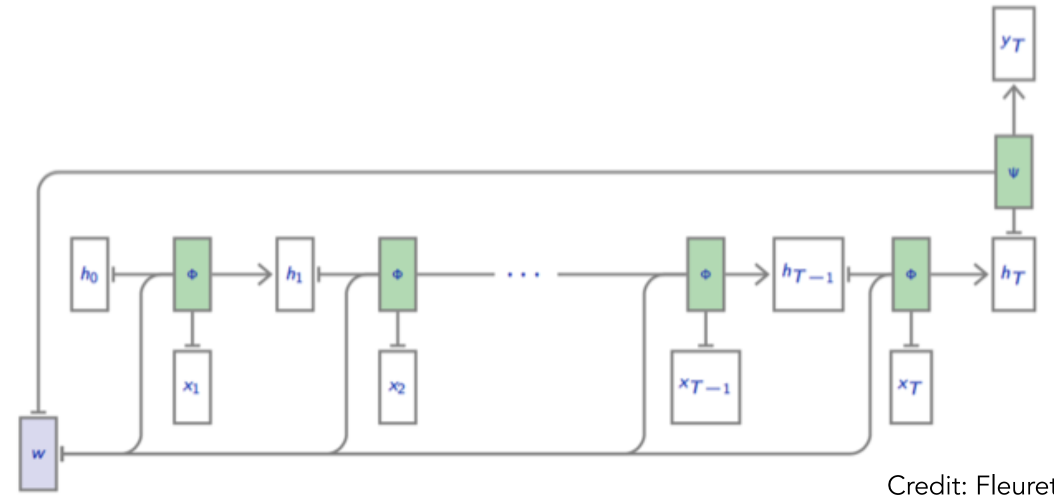
$$\phi(\boldsymbol{x}_t, \boldsymbol{h}_t; W, U) = \sigma(W\boldsymbol{x}_t + U\boldsymbol{h}_t)$$

- Predictions can be made at any time t from the recurrent state

$$\boldsymbol{y}_t = \psi(\boldsymbol{h}_t; \theta)$$

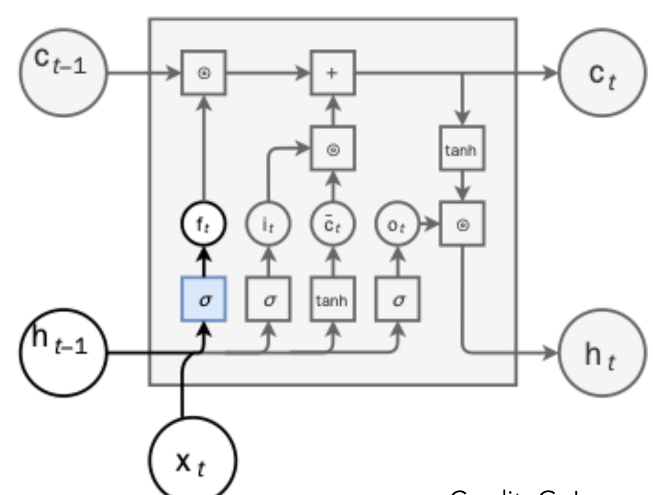
Credit: F. Fleuret

Recurrent Networks:



Credit: Fleuret

LSTMs:



Credit: G. Louppe

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

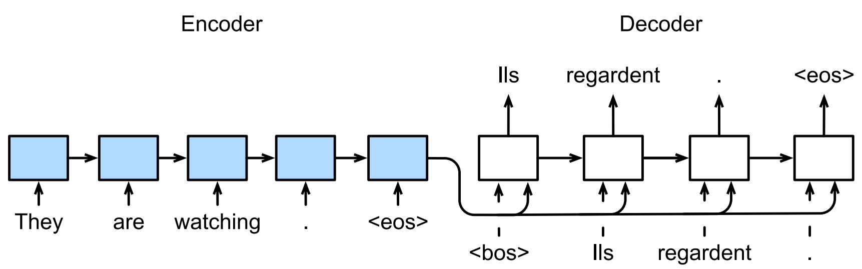
Bottleneck → Attention!

Seq2seq models analyse sequences

- Predict probability distributions of the next token given previous context
- Encoder compresses the sequence in a fixed size vector

Fixed size latent vector is a bottleneck

Decoder **next-step generation is suboptimal** since latent vector contains the same information



Credit: d2l.ai

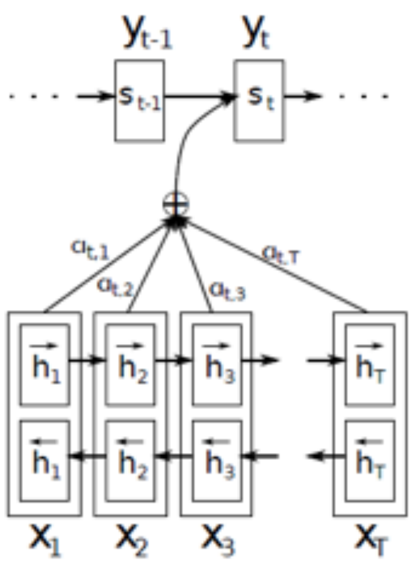
Can we avoid compression and pass the decoder entire input?

- Need a mechanism to **focus on most relevant** input tokens at each prediction step
- Introduce **softmax to calculate probability** (maintain differentiable architecture)

Output is **independent of the order** of input examples (set instead of sequences)

Use **relationships between input elements** (as graph representation).

Attention mechanism as originally formulated in a bi-directional LSTM Auto-Encoder https://arxiv.org/abs/1409.0473



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Attention mechanism

See tutorial G.. Weiss tutorial at IML workshop : https://indico.cern.ch/event/1297159/

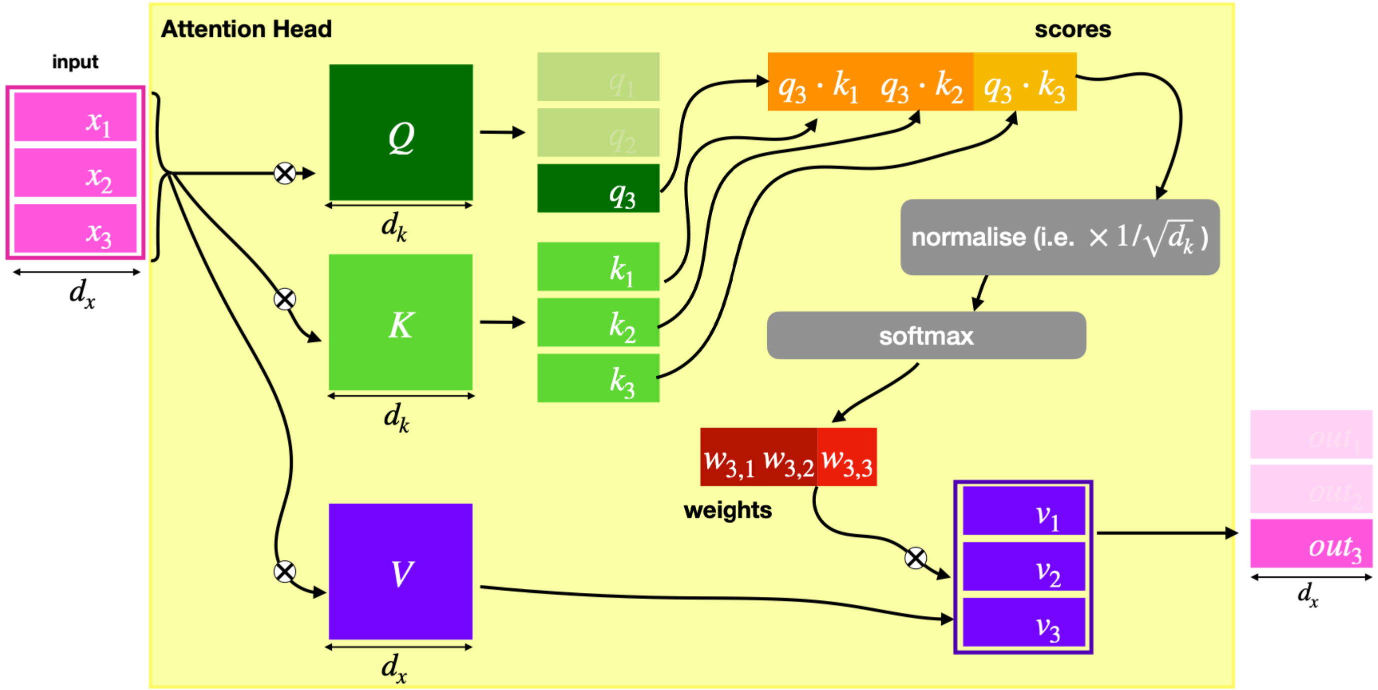
A key-value database (differentiable, entries are continues vectors):

$$Q = \{q_1, q_2, \ldots, q_m\} \text{ Queries}$$
$$K = \{k_1, k_2, \ldots, k_n\} \text{ Keys}$$
$$V = \{v_1, v_2, \ldots, v_n\} \text{ Values}$$

A normalised **similarity** function between query-key pairs:

$$S_{ij} = \text{Similarity}(q_i, k_j) \qquad A_{ij} = \text{Normalize}(S_{ij}) = \frac{e^{S_{ij}}}{\sum_{l=1}^{n} e^{S_{il}}}$$

A **weighted average** over values {V}, based on similarity: $O_i = A_{ij} V^j$



Credit: G. Weiss

NB. Weights are probabilities (use softmax)

Self-attention uses same input for values, keys and queries.
Focus on relationship between elements (adds context)

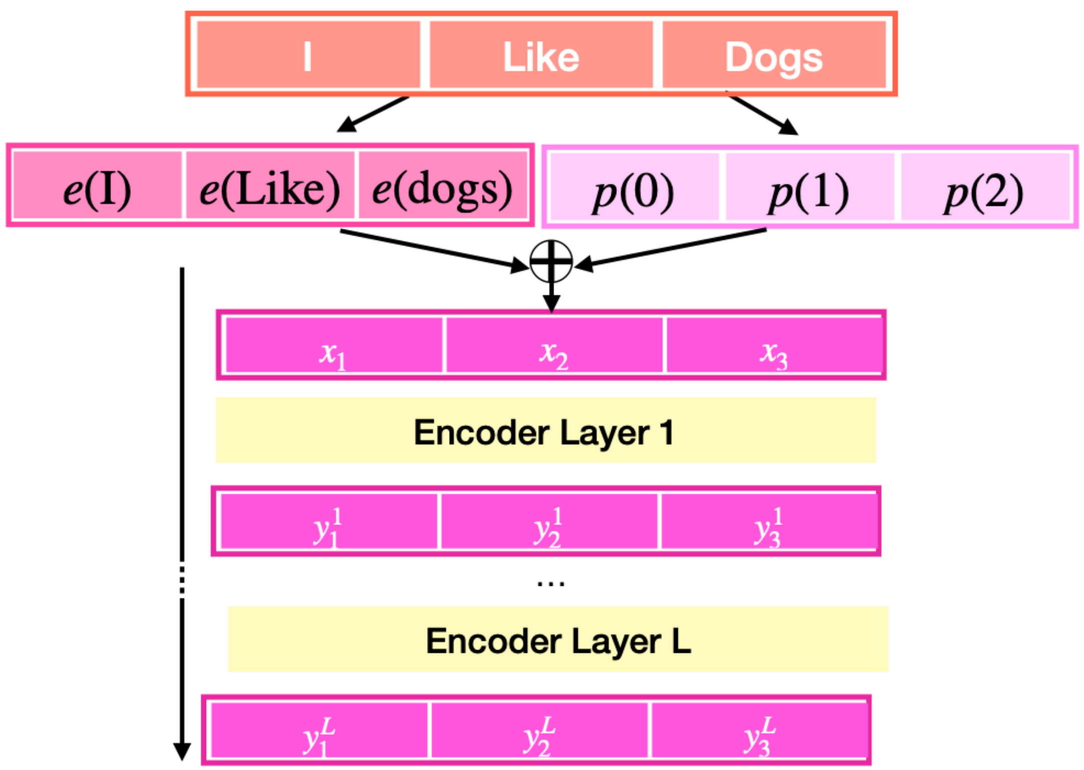
$$\text{Similarity}(q_i, k_j) = \frac{q_i \cdot k_j}{\sqrt{D}}$$

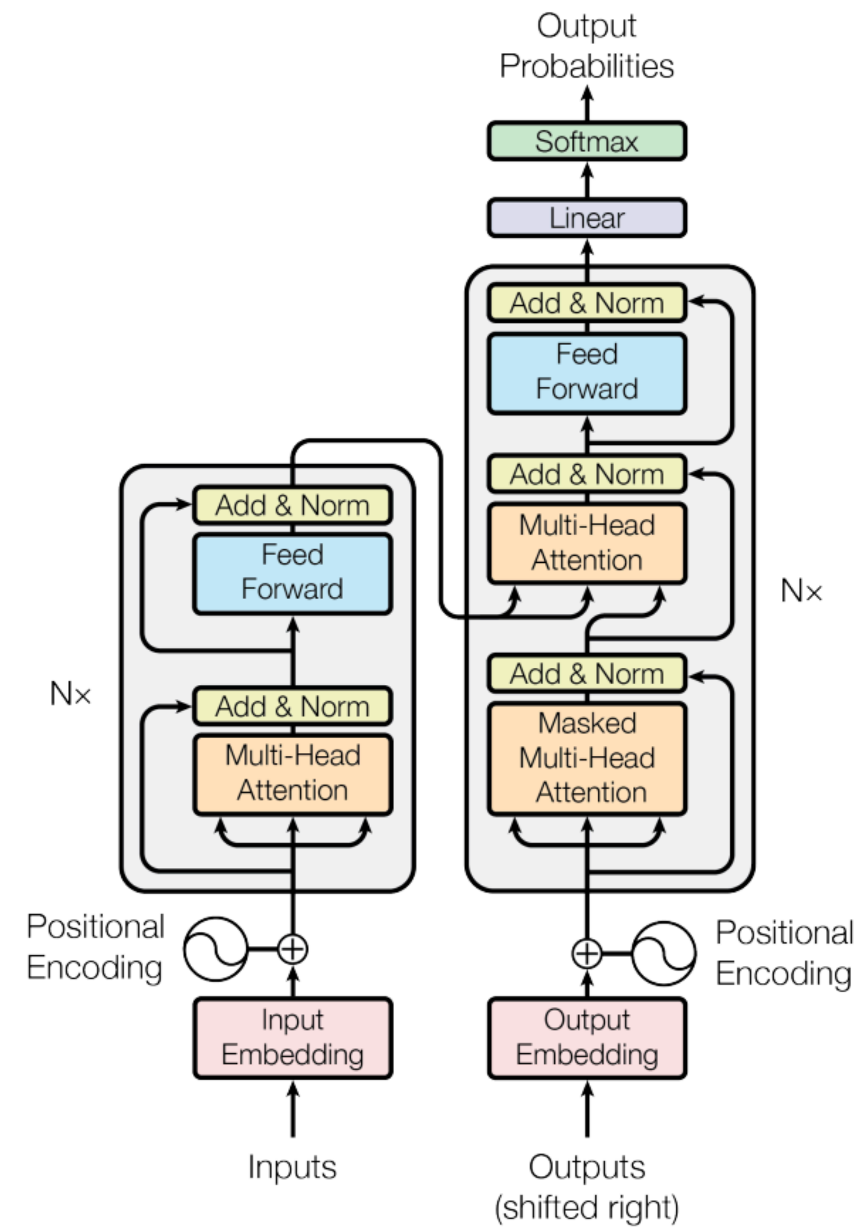
Multi-head attention splits input token in subgroups and processes them in parallel

NB: Scaled dot-product is permutation equivariant

Transformers

See tutorial G.. Weiss tutorial at IML workshop :
https://indico.cern.ch/event/1297159/





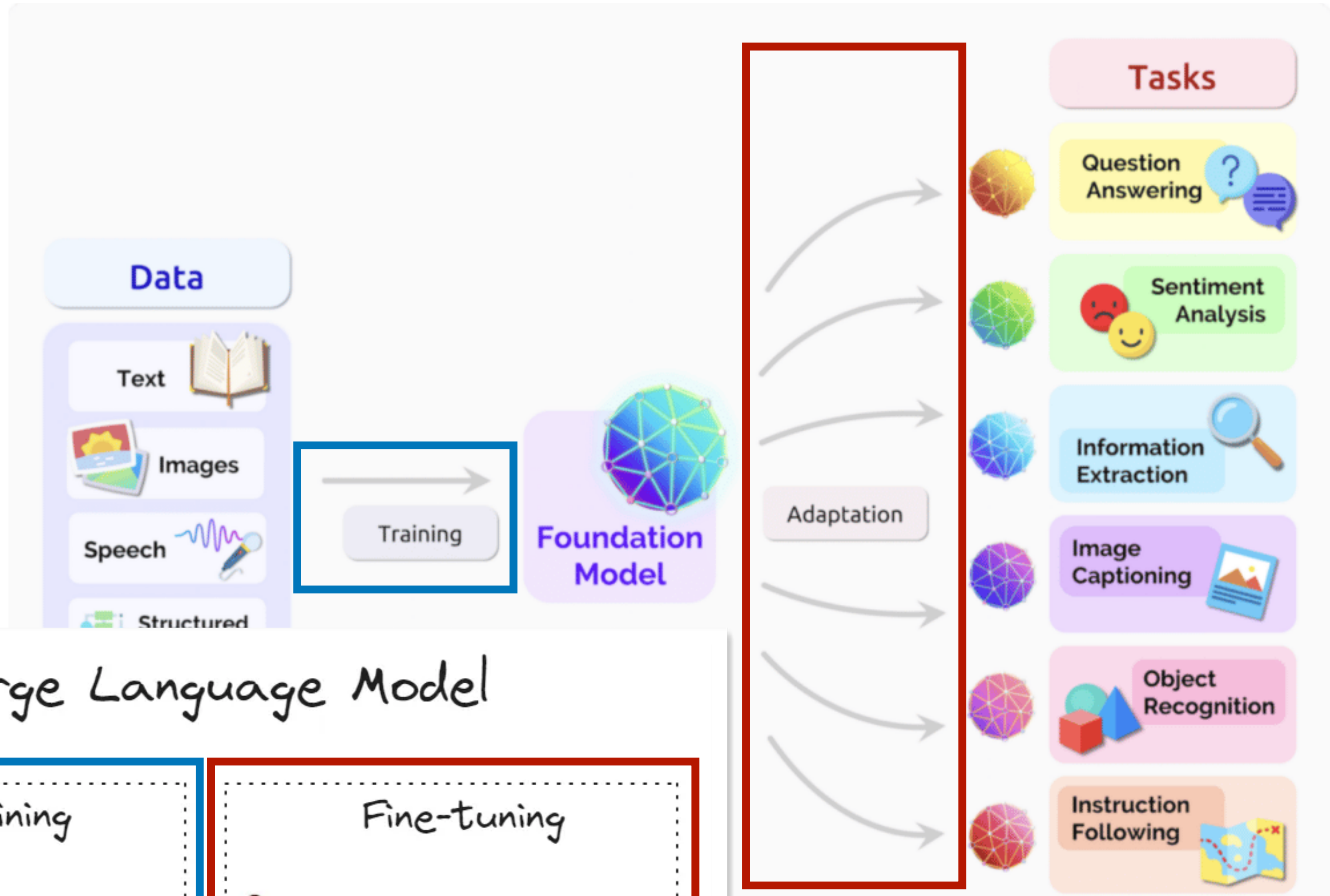
Transformer components include:

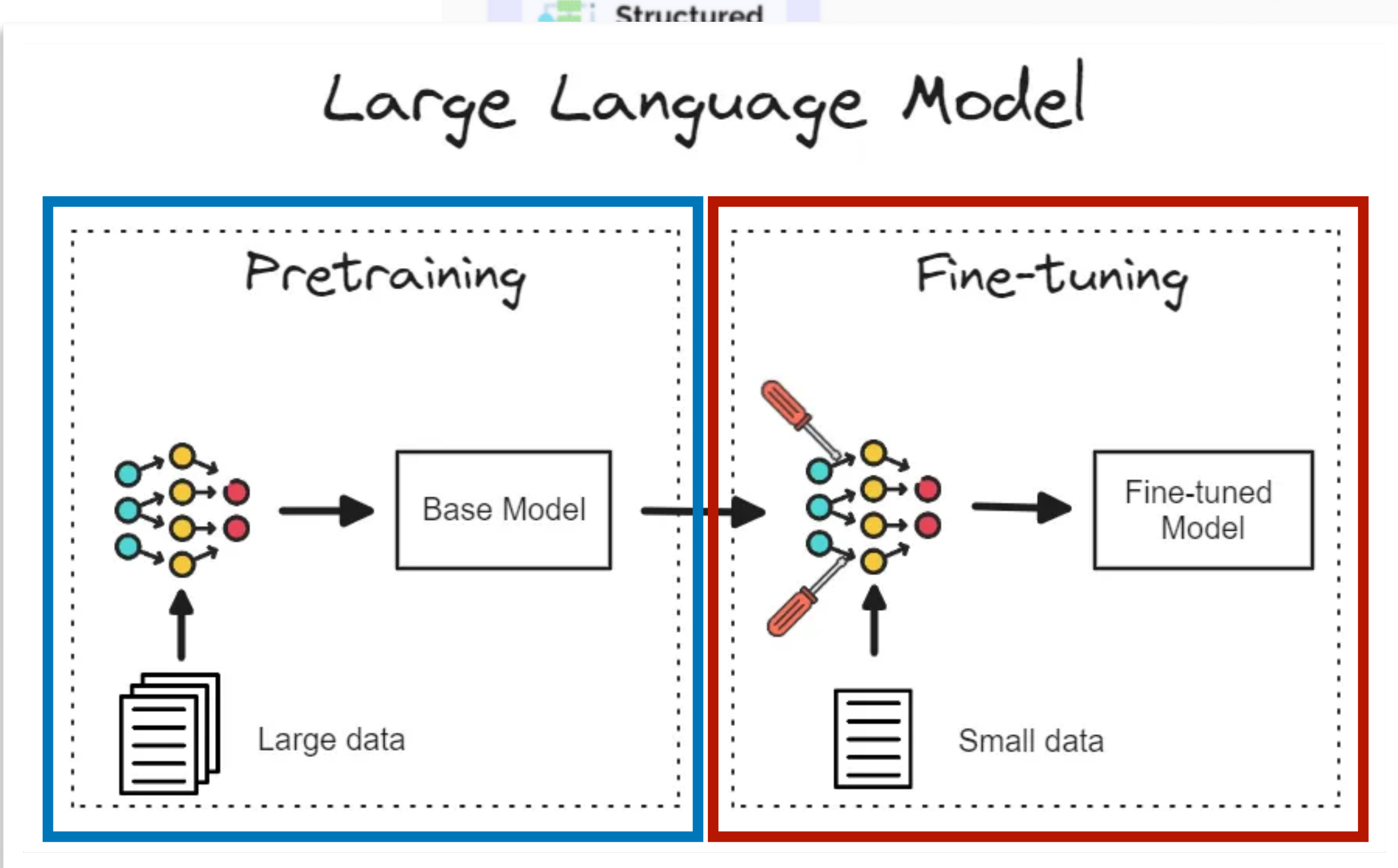
- Multi Head **Attention**
- **Normalisation** layers
- Position Independent **Feed Forward Layers**
- **Skip Connections**

NB. All tokens are processed in parallel

Vaswani et al., *Advances in Neural Information Processing Systems*, 2017, 5998–6008

Introduction

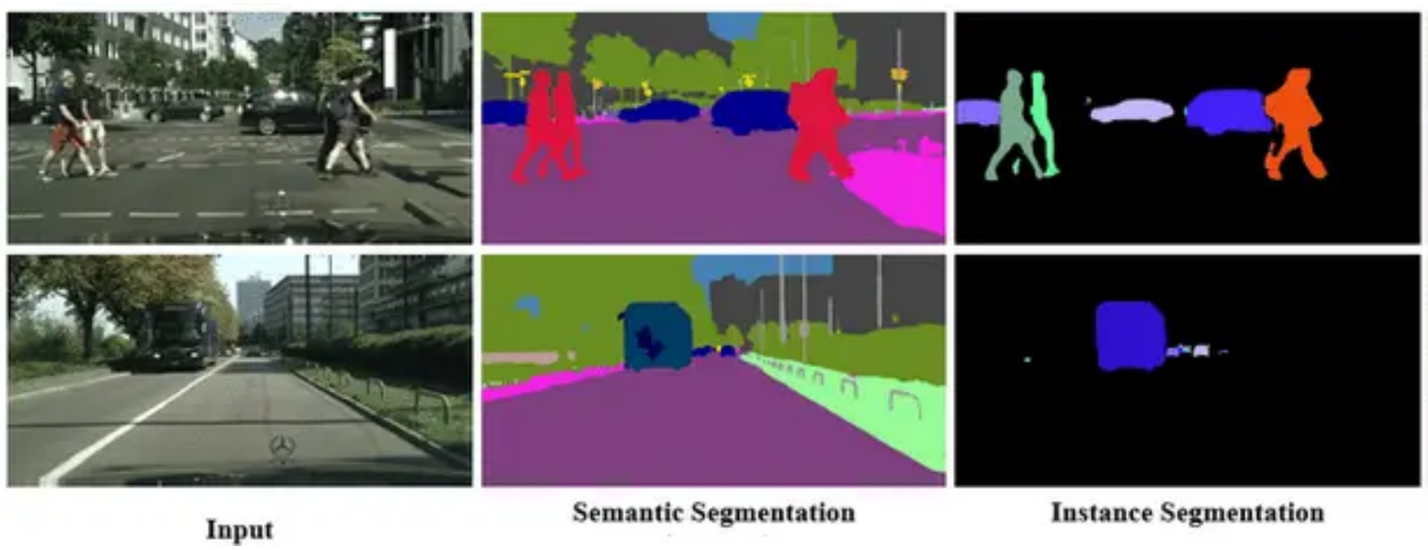




ise@cern.ch I sofia.vallecorsa@cern.ch

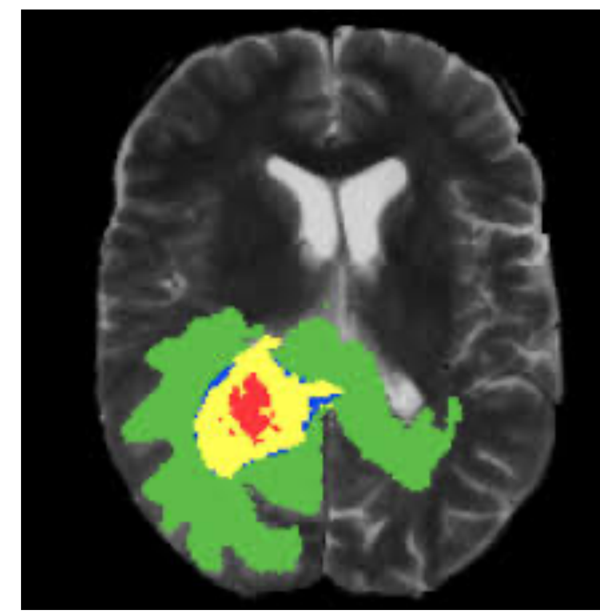
A concrete example

Downstream scientific application: detect brain cancer with machine learning



We can adapt a general model to brain images to improve accuracy

We would now need a much smaller dataset to "fine-tune" the model for the task



Pre-training: learn how to segment images (aka cluster pixels together into the different objects):

- Learn how to detect edges
- Learn how to cluster objects with the same e.g. colour …

These skills can be leant from a large general dataset that has nothing to do with brain images

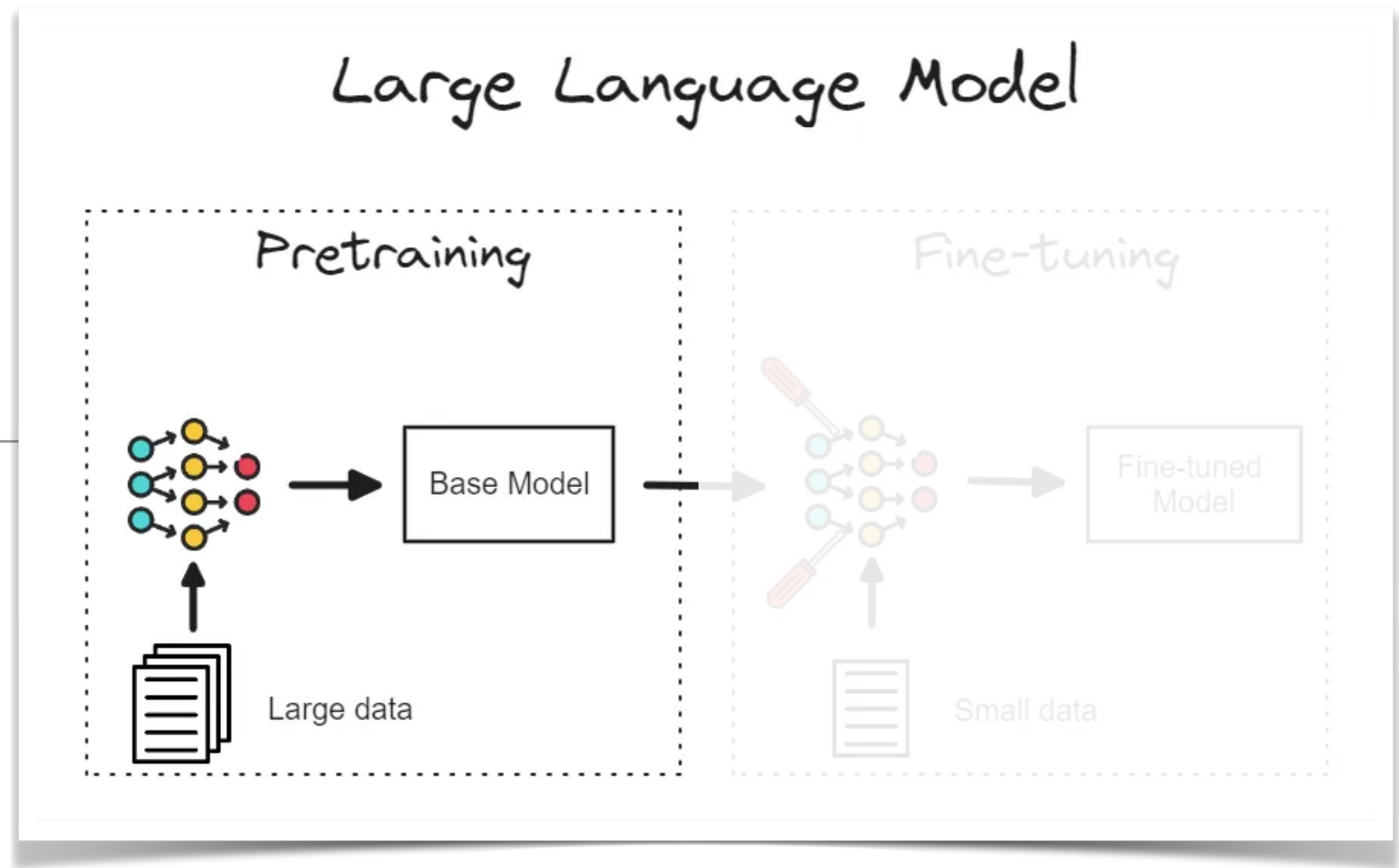
Brain images:

- costly
- Not many available
- Sensitive data: Privacy and access problems

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training

Basic concepts

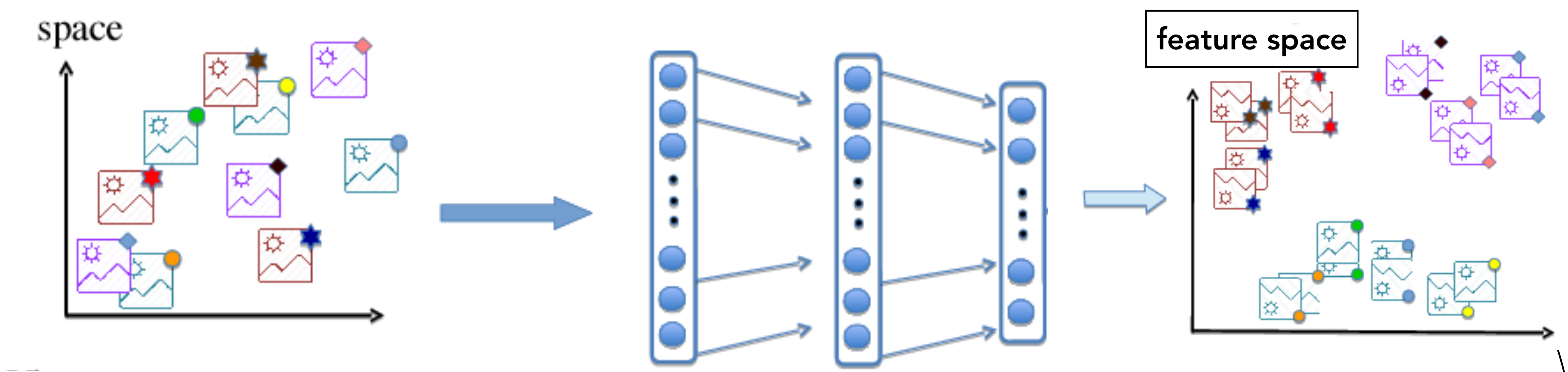


Main goal

Pre-training:
"train a model on a large dataset to learn general features and patterns before fine-tuning it for specific tasks or domains"

Representation learning:

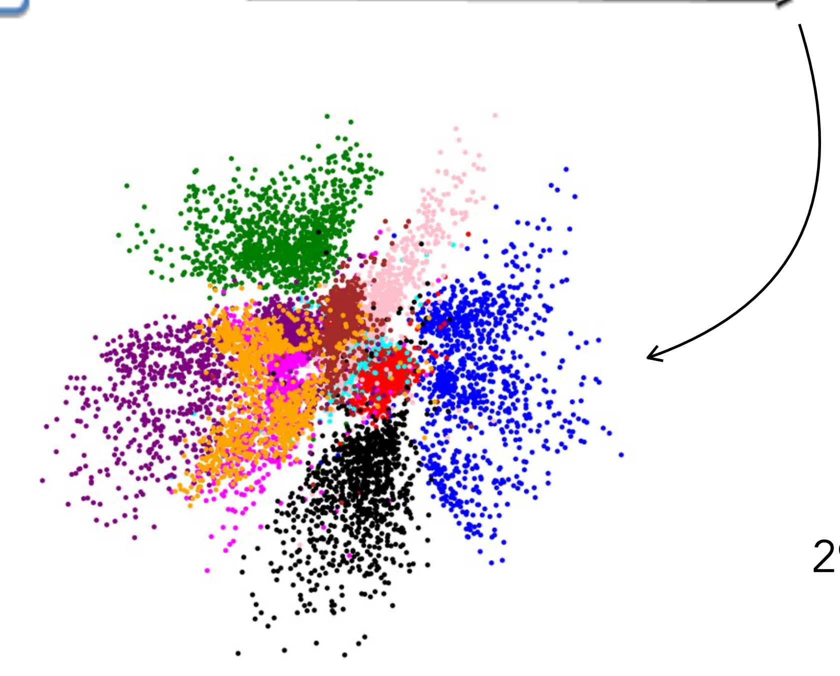
- Learn a **task-independent representation** of the data in the **feature space** of the neural network



Effective "encoding" of the data useful for many applications

$$L_j : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Data are clustered according to common properties in the "feature space"



https://arxiv.org/pdf/1206.5538v3.pdf

Advantages of the pre-training step

- **Improved Performance:**
  - **Better Generalization** to new tasks.
  - **Higher Accuracy** of the fine-tuning step compared to training from scratch.

- **Reduced Training Time:**
  - **Faster Convergence** during fine-tuning.
  - **Less Computational Resources,** since the model starts with a good initialization.

- **Data Efficiency:**
  - **Less Data Required** during fine-tuning. This is particularly beneficial for tasks where labeled data is scarce or expensive to obtain.
  - Applicability to **Multimodal and Multitask Learning**

- **Handling Overfitting:**
  - **Robustness:** Starting from a pre-trained model can help mitigate overfitting, especially when the target dataset is small, by leveraging the broad knowledge encoded during pre-training.

- **Feature Extraction:**
  - **Rich Feature Representations:** encapsulate complex correlations into an abstract representation
  - **Versatility:** Pre-trained models can be adapted to various downstream tasks.

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

… and some drawbacks

- **Data Dependency:**
  - Pre-training heavily relies on the availability and quality of large-scale datasets, posing **challenges in domains with limited data accessibility**.
- **Task Specificity:**
  - While pre-training initialises models with generalised knowledge, **fine-tuning for specific tasks may require additional data and computational resources**, impacting the overall training process.
- **Overfitting Risks:**
  - In certain scenarios, **pre-trained models may exhibit overfitting tendencies if not rigorously fine-tuned**, affecting their adaptability to new datasets.

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Workflow

Data-preprocessing
e.g. normalisation, augmentation

↓

Embedding
Project the data into the feature space

↓

Training
Learn the correlations in the projected feature space

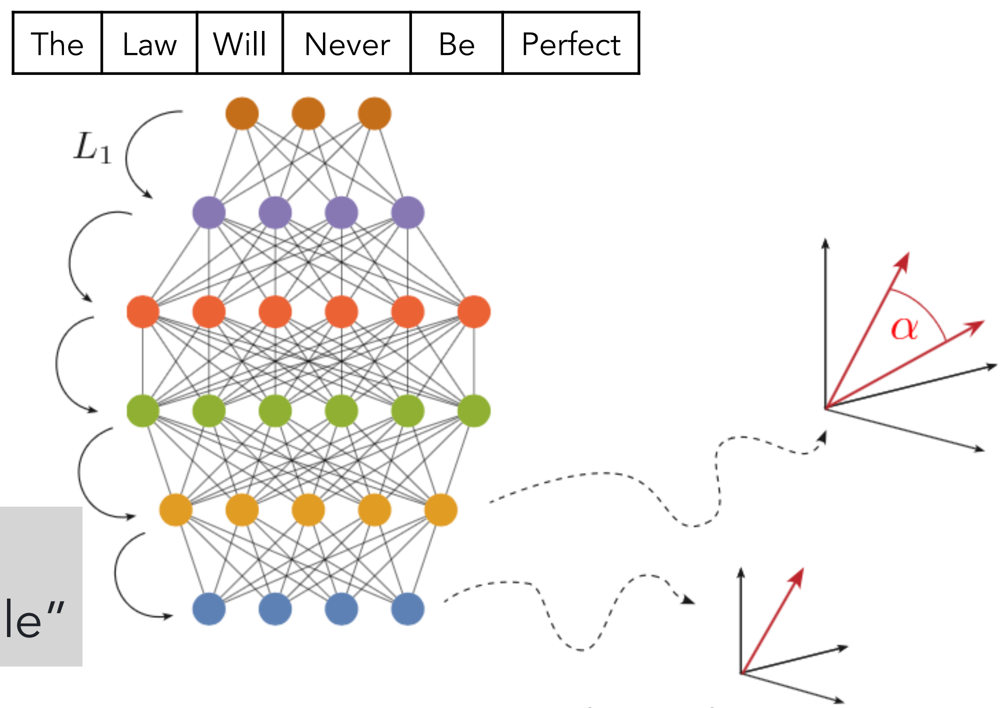
↓

Re-shuffle
project the data at a different “angle”

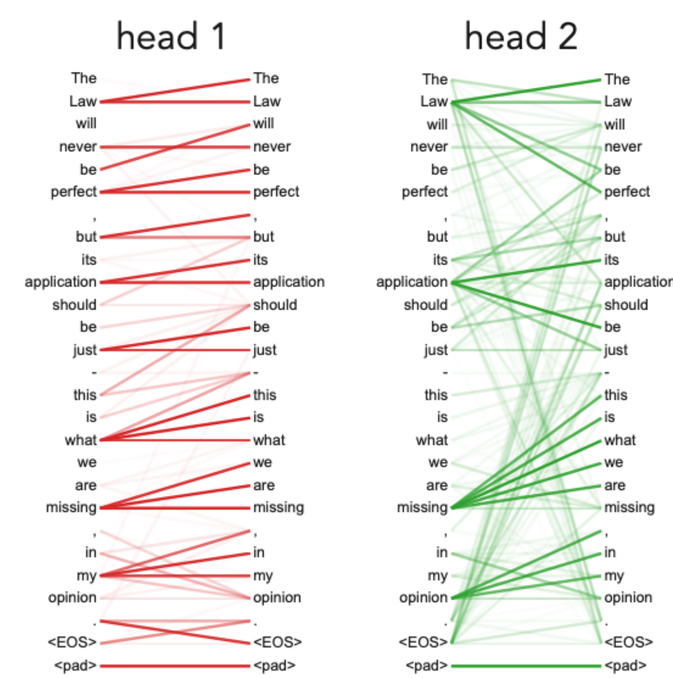
↓

Loss calculation

Important step: the embedding!
Project the data into a vector space
→ multimodality



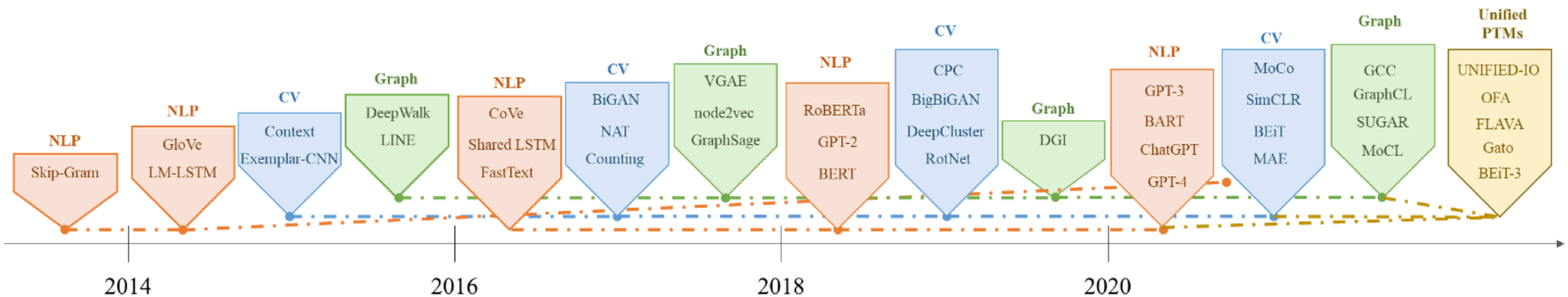
See each word as a vector in a complex space

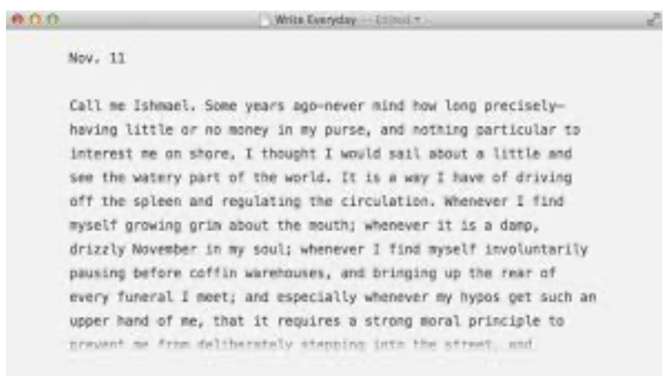


“Attention is all you Need” Vaswani 2017

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

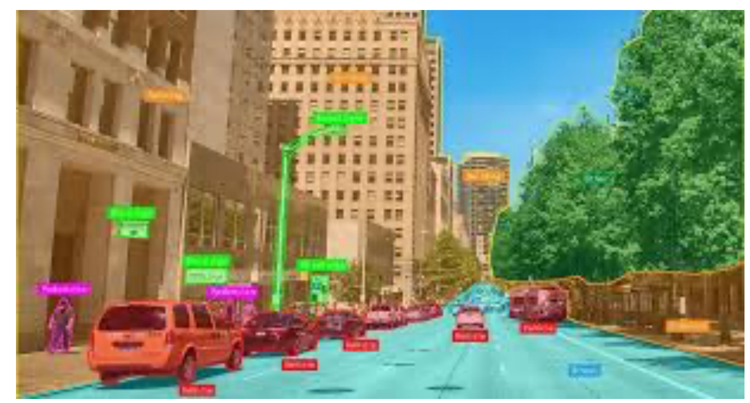
Types of pre-trained models




Nov. 11

Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and

NLP: Natural Language Processing



CV: Computer Vision

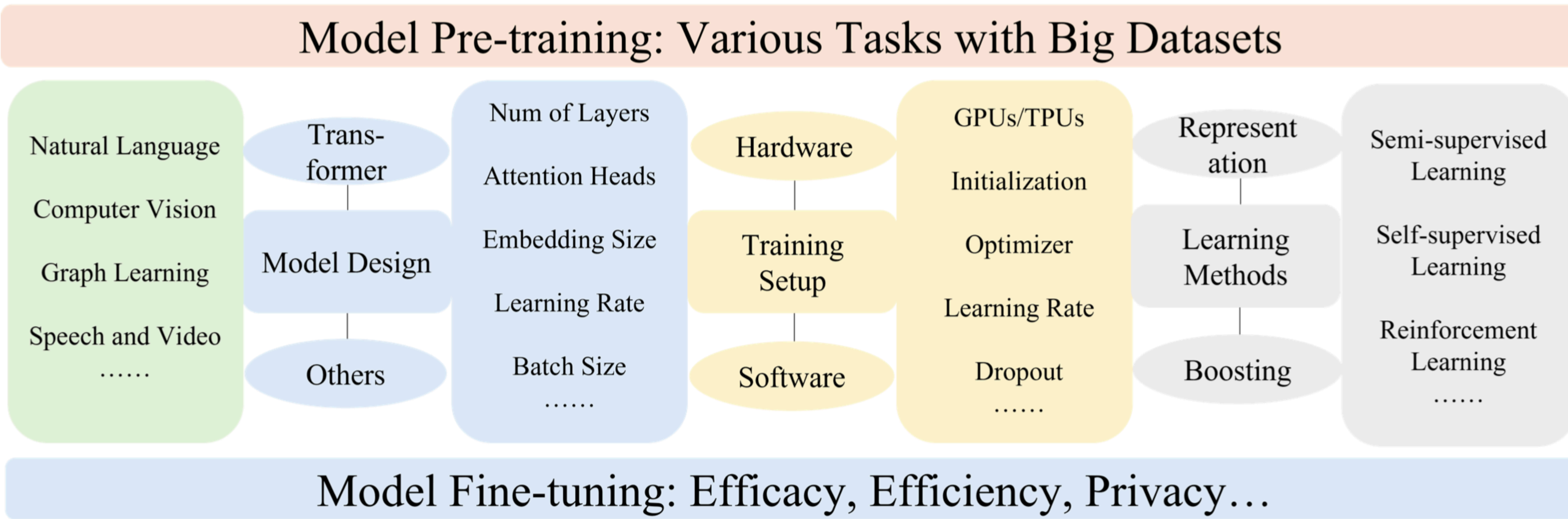
Graphs: Graph Learning (*not covered here*)

Unified Pre-trained Models

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Types of pre-trained models

Depending on the type of dataset (text, images, etc..) there are many choices to be done:



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

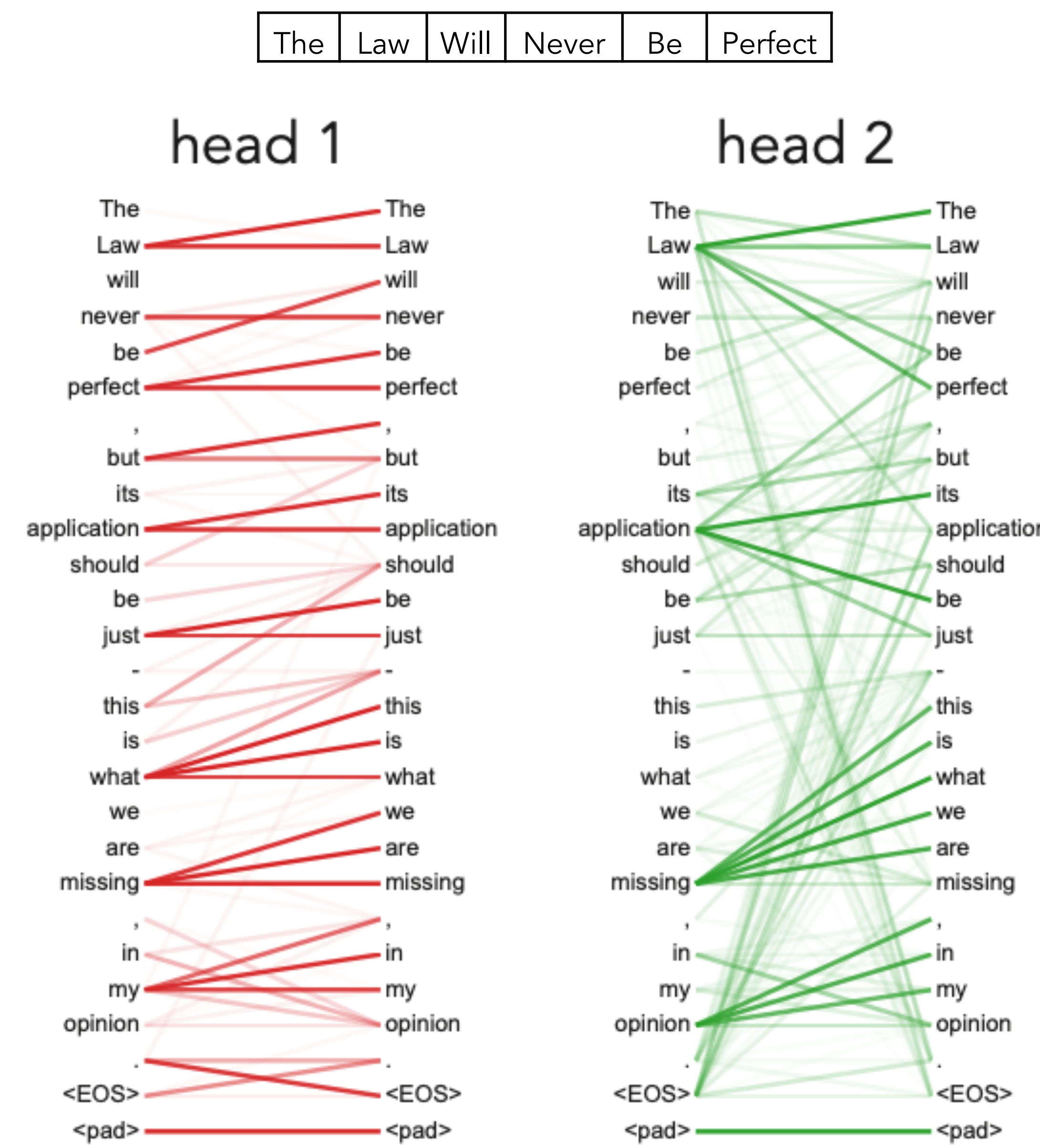
How do we pre-train?

Pre-training: Natural Language Processing

- **Mask Language Modelling (MLM):** mask some words randomly in the input sequence and predict them back.
- **Denoising AutoEncoder (DAE):** Add noise to the original text and reconstruct the original input.
- **Replaced Token Detection (RTP):** replace tokens with other random tokens and discriminate which tokens have been replaced.

Sentences (not covered here):

- ***Next Sentence Prediction (NSP):*** *binary classification task. Predict whether a given sentence is the direct continuation of a preceding sentence.*
- ***Sentence Order Prediction (SOP):*** *binary or multi-classification task. It learns to determine the correct order of a given set of sentences*



"Attention is all you Need" Vaswani 2017

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training NLP: Mask Language Modelling

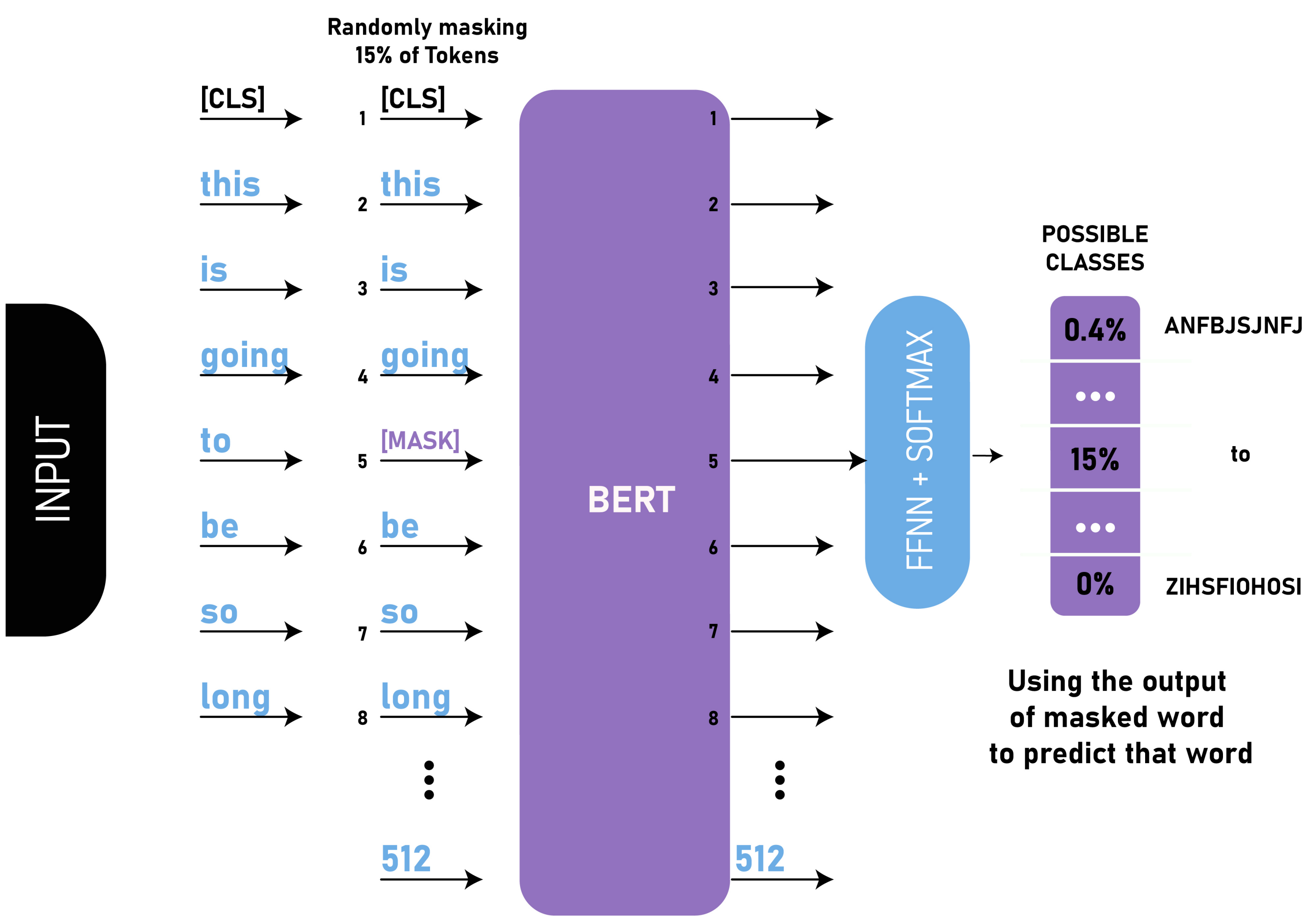
How Does It Work?

- **Input Text:** Take a large corpus of text.
- **Masking:** Randomly mask a portion of the tokens in the input text (typically 15%).
- **Model Training:** Train the model to predict the masked tokens based on the surrounding context.

Example model: BERT (Bidirectional Encoder Representations from Transformers)

Contextual Understanding: *Models learn bidirectional context, understanding the meaning of words in relation to their surrounding text.*

Bidirectional Context: *Unlike traditional language models that predict the next word, masked language models learn from both left and right contexts.*



https://www.geeksforgeeks.org/understanding-bert-nlp/

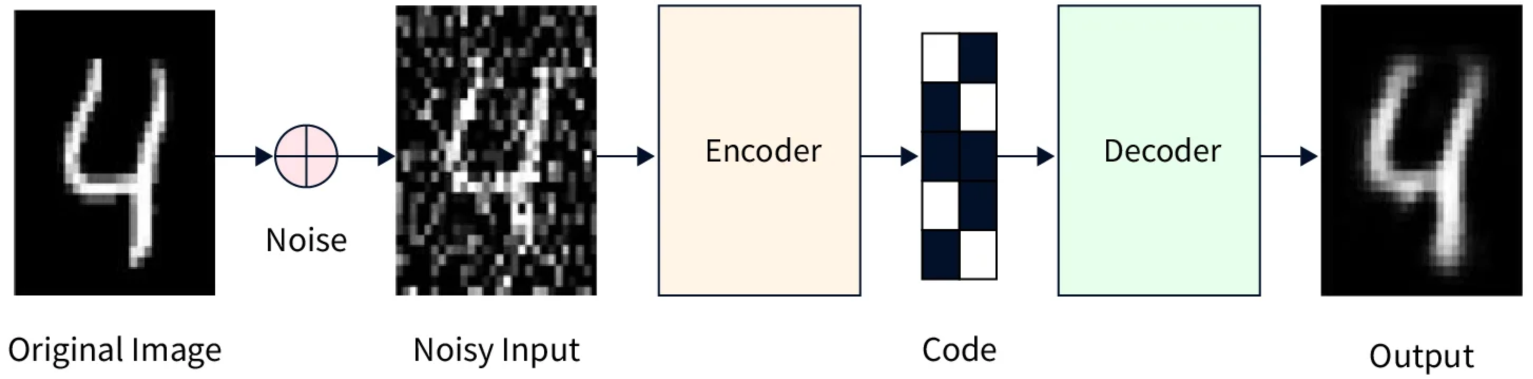
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training: Denoising AutoEncoder

How Does It Work?

- **Input Corruption: Introduce noise to the input data (e.g., Gaussian noise, masking).**
  - Example: Original input: [0.1, 0.2, 0.3, 0.4] -> Noisy input: [0.1, 0.0, 0.3, 0.0].
- **Encoding:** The encoder processes the noisy input to produce a compressed representation.
  - This step captures the essential features while ignoring the noise.
- **Decoding:** The decoder reconstructs the original input from the latent representation.
  - It aims to remove the noise and recover the clean data.
- **Loss Calculation:** Compute the loss by measuring the difference between the original input and the reconstructed output (e.g. Mean Squared Error)

Robust Feature Learning:
Learns to extract robust features that are resilient to noise.



Noise Handling:
Effective in learning representations that are less sensitive to noise and corruption in the input data.

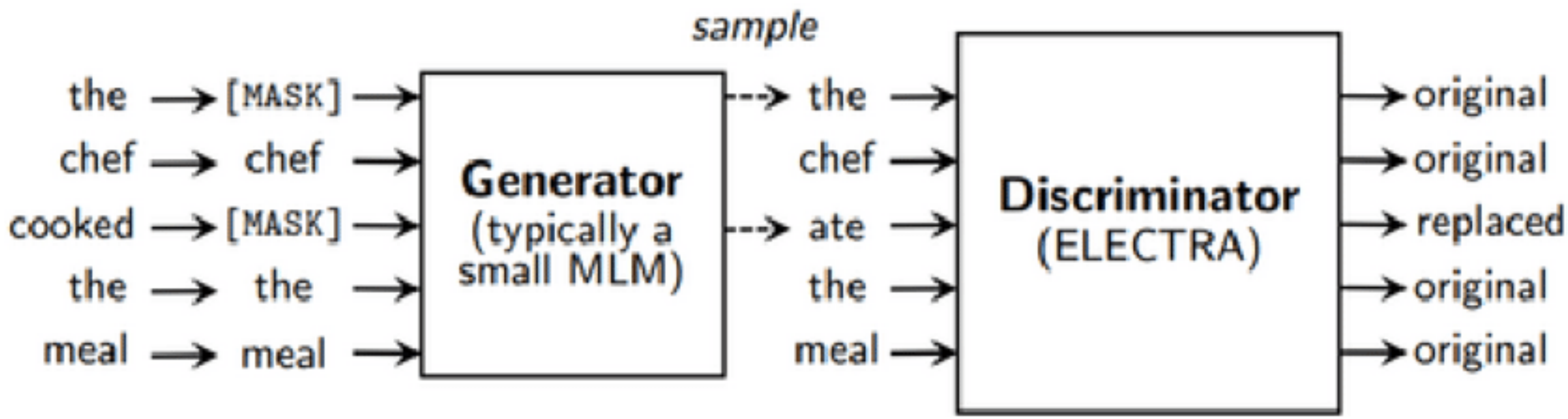
https://www.scaler.com/topics/deep-learning/denoising-autoencoder/

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training: Replaced Token Detection

How Does It Work?

- **Input Preparation:** Randomly select some tokens in the text to be replaced with incorrect tokens (e.g., tokens from a different context or completely random tokens).
- **Task Formulation:** The model is given the modified text and tasked with identifying which tokens have been replaced.
    - Example:
        - Original Sentence: "The cat sat on the mat."
        - Modified Sentence: "The cat sat on the dog."
- **Loss Function:** Typically involves a binary classification loss where the model predicts whether each token is correct or replaced.



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

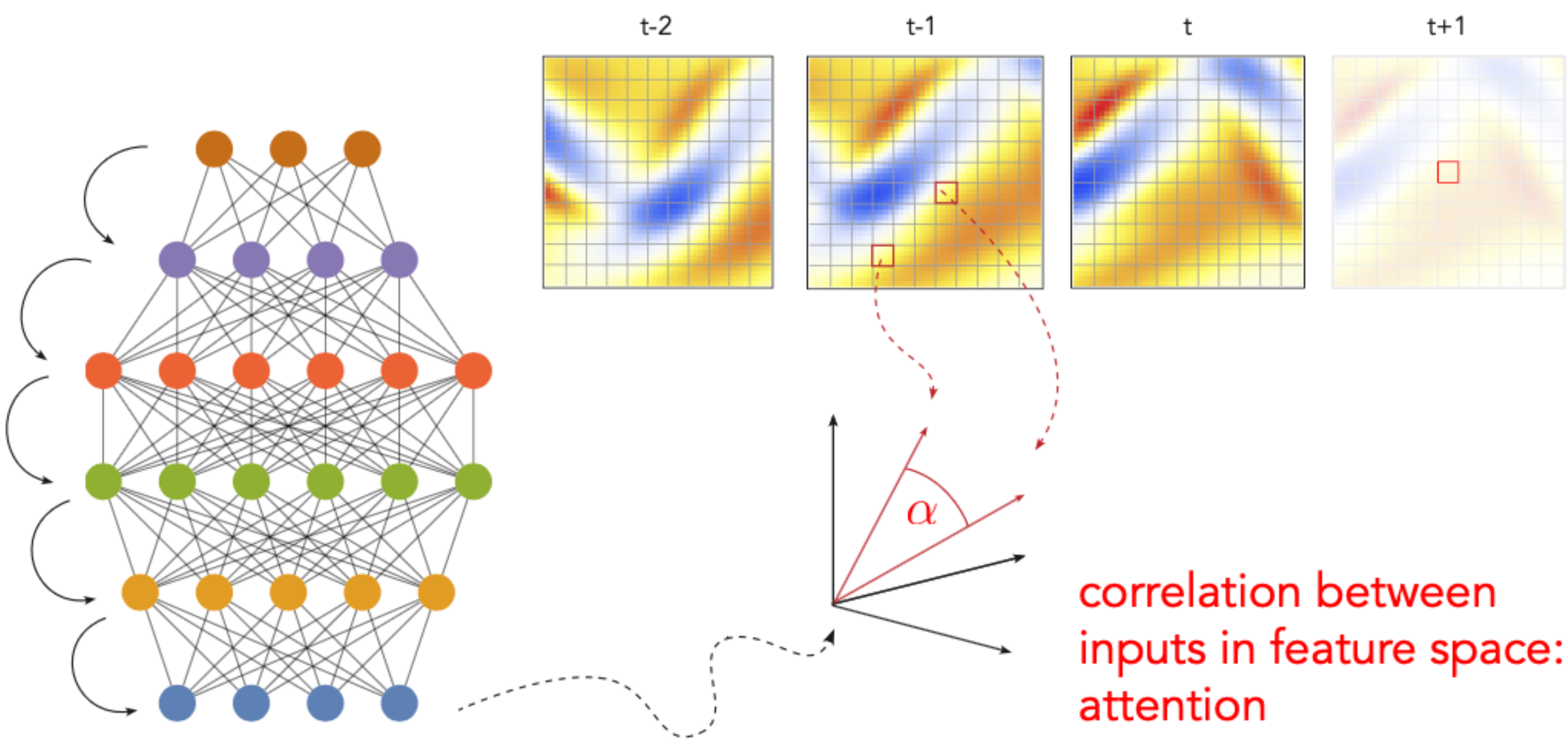
Pre-training: Computer Vision

- Data reconstruction tasks
- Specific pretext tasks
- Frame order tasks (not covered)
- Miscellaneous

Complication: what is a token?

Single pixels carry too little information.

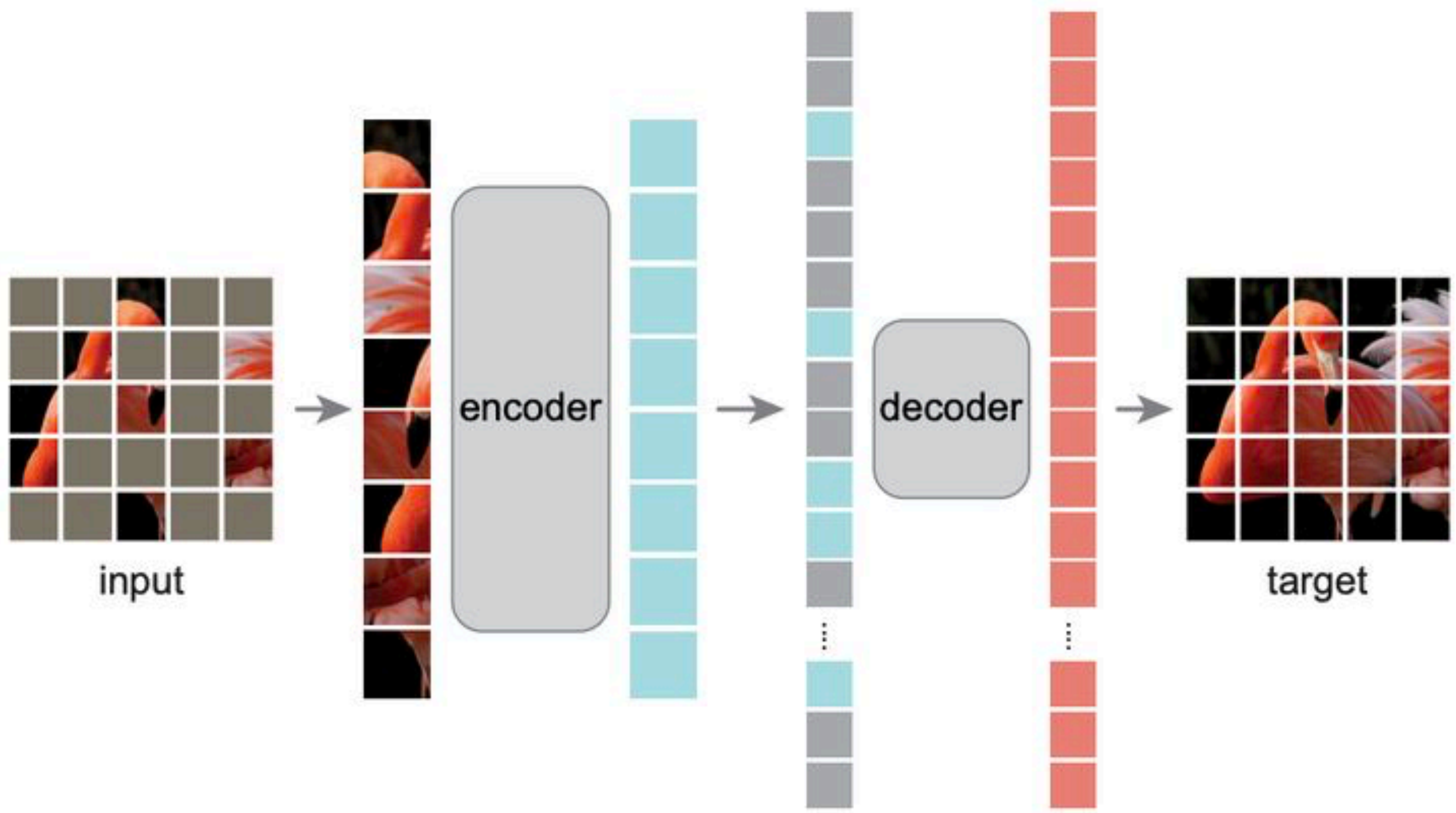
trade-off between token-size and information in each token



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training: Data reconstruction tasks

Image Inpainting: Learn to fill in missing parts of an image.
The model is trained to predict missing regions given the context of the surrounding pixels.

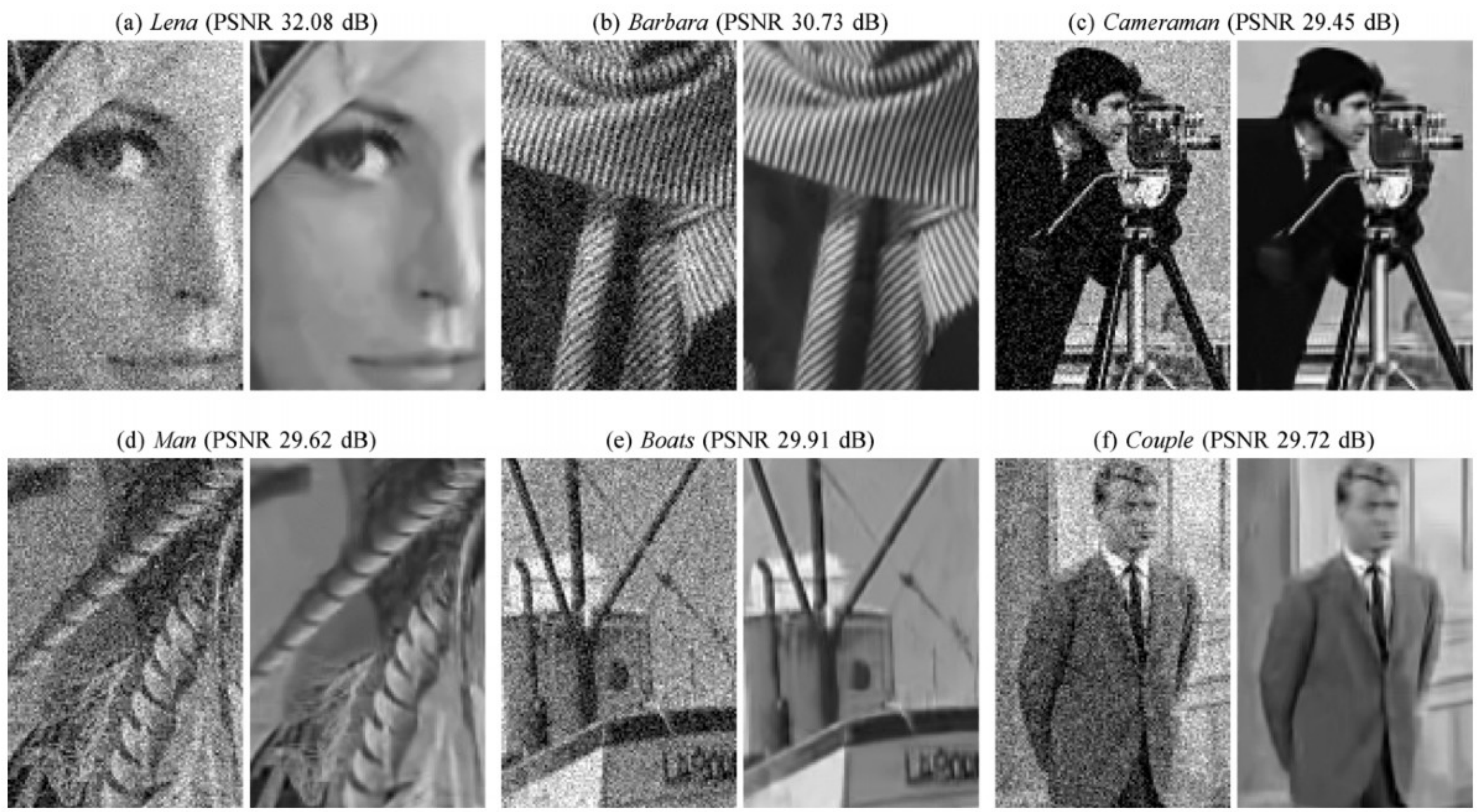


Example: Removing a portion of an image and training the model to reconstruct the removed region.

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training: specific pretext tasks

Image Denoising: Remove noise from an image, generating a clear version from a noisy input.

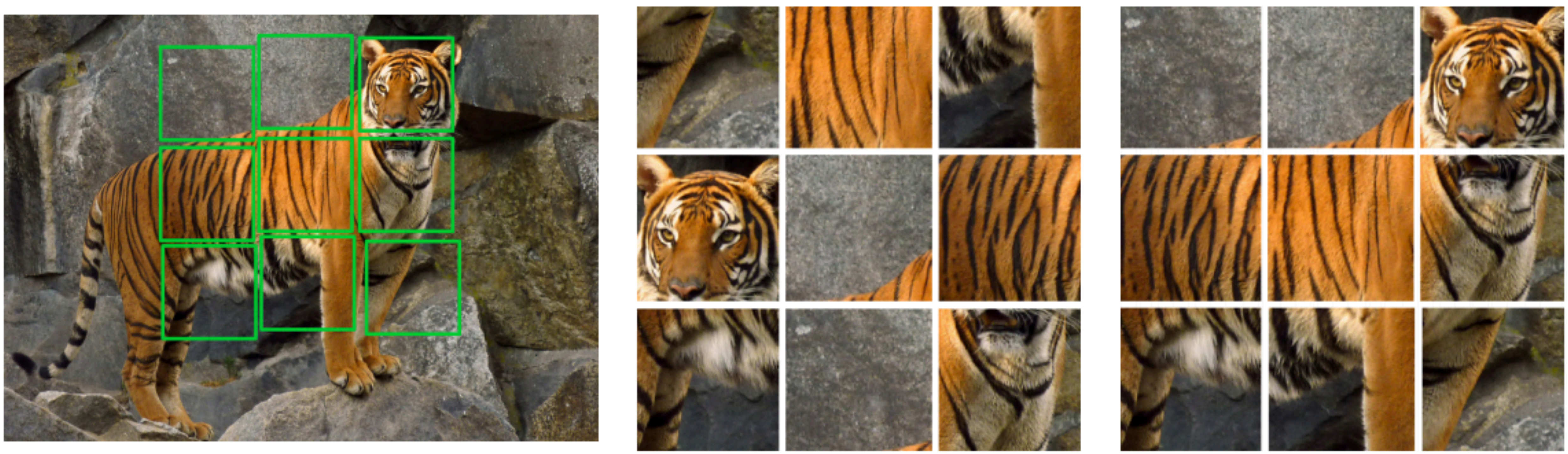


Example: A noisy image is input into the model, which predicts and outputs a clean, noise-free version. *(Different from diffusion models)*

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training: specific pretext tasks

Jigsaw Puzzle Solving: Divide images into patches, shuffle them, and train the model to predict the correct arrangement of the patches.



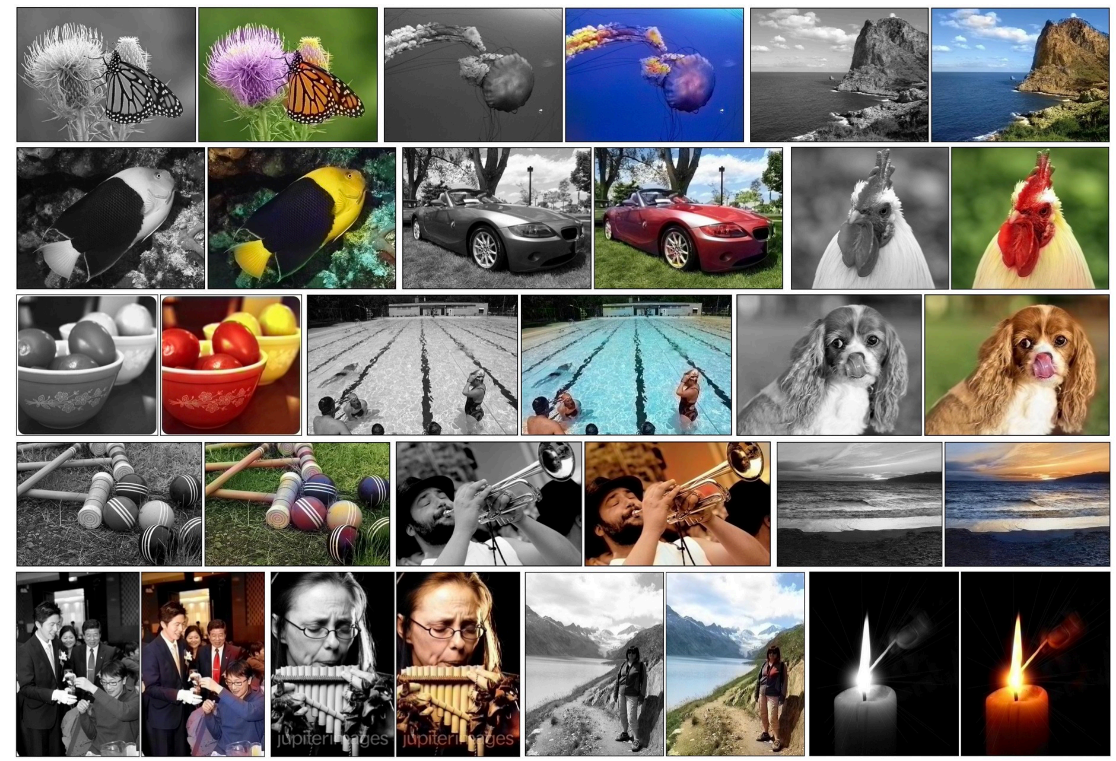
Example: Splitting an image into a 3x3 grid, shuffling the patches, and training the model to solve the puzzle.

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Pre-training: other specific pretext tasks

Colourisation: Convert grayscale images to color. The model learns to predict the colours from the grayscale input.

- **Example:** Training the model to colourise black and white images.



Style Transfer: Transfer artistic styles from one image to another while preserving the original content. The model learns to separate and apply style and content features.

- **Example:** Applying the style of a famous painting to a photograph.

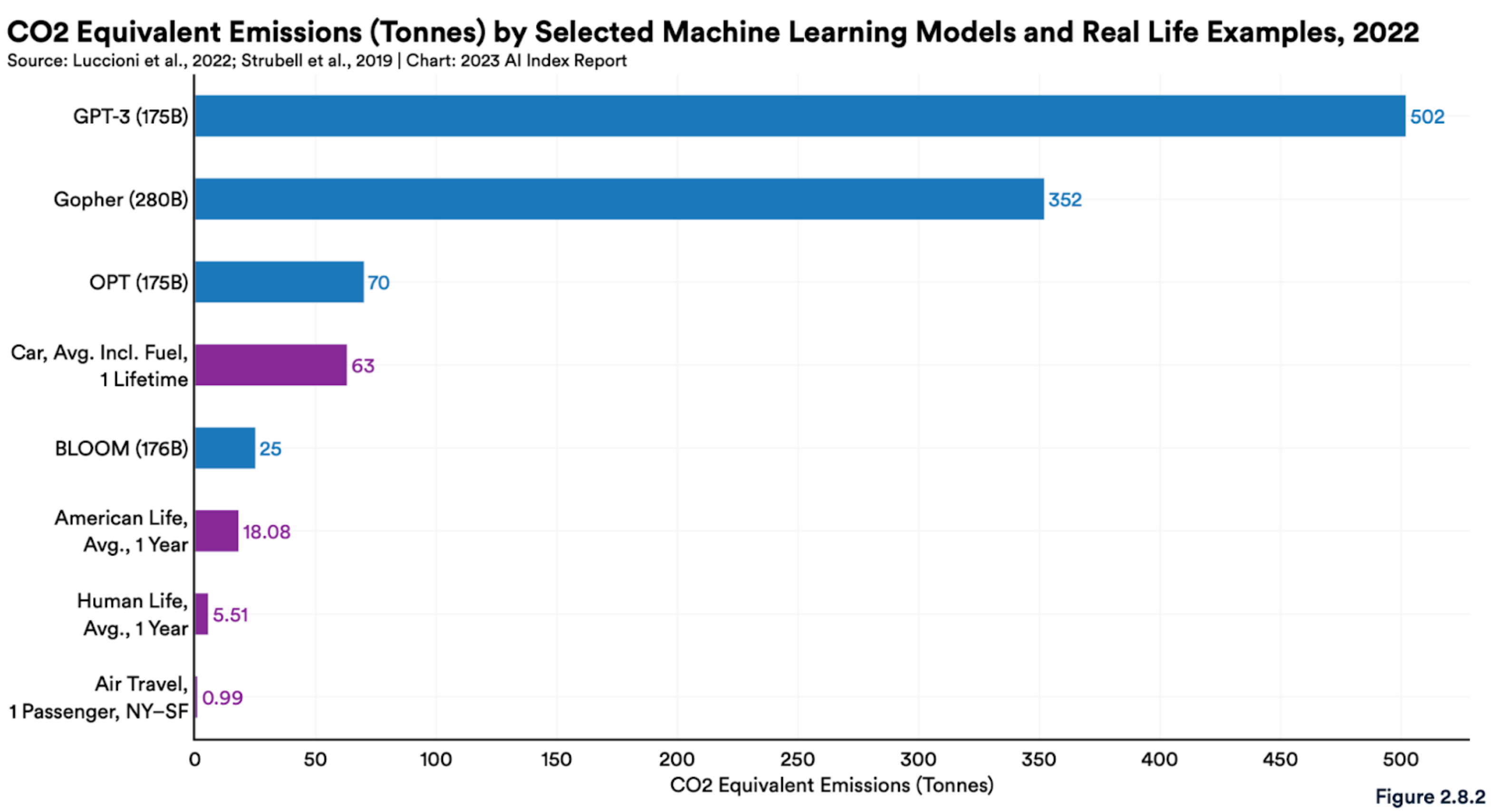
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Hardware and footprint

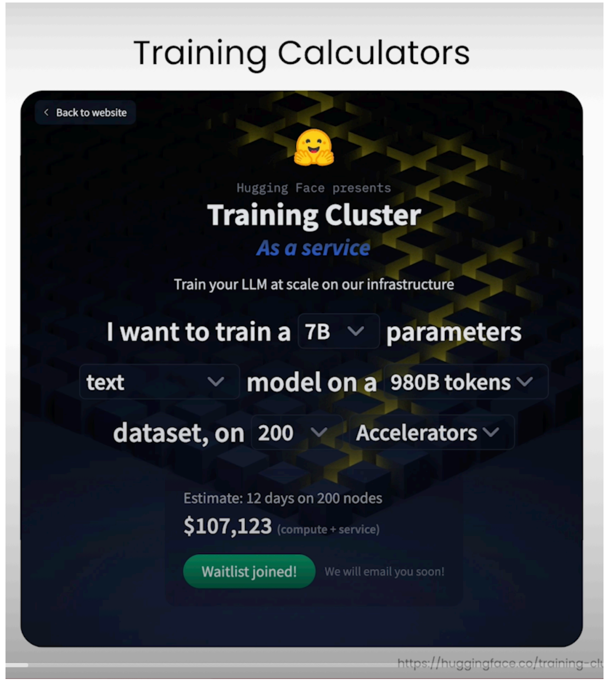
Computing Resources: Distributed computing

- **High-Performance GPUs:** Foundation models often require GPUs or TPUs.
  - **Example:** NVIDIA A100, Google TPU v4.

- **High RAM and storage capacities** are needed to manage large datasets and model checkpoints.
  - hundreds of terabytes of storage and several terabytes of RAM.



Training cost calculator

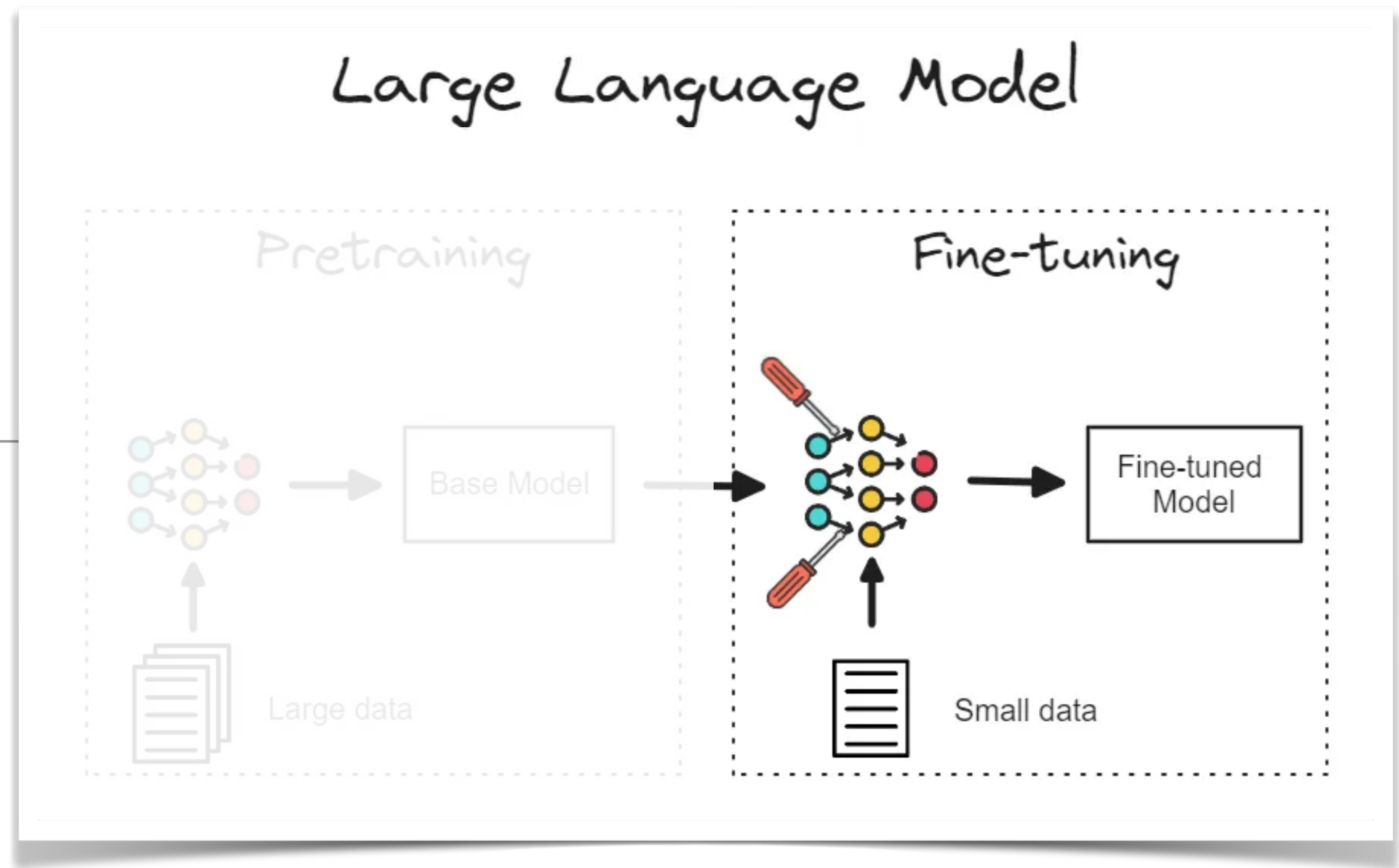


https://analyticsindiamag.com/ai-origins-evolution/the-environmental-impact-of-llms/

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

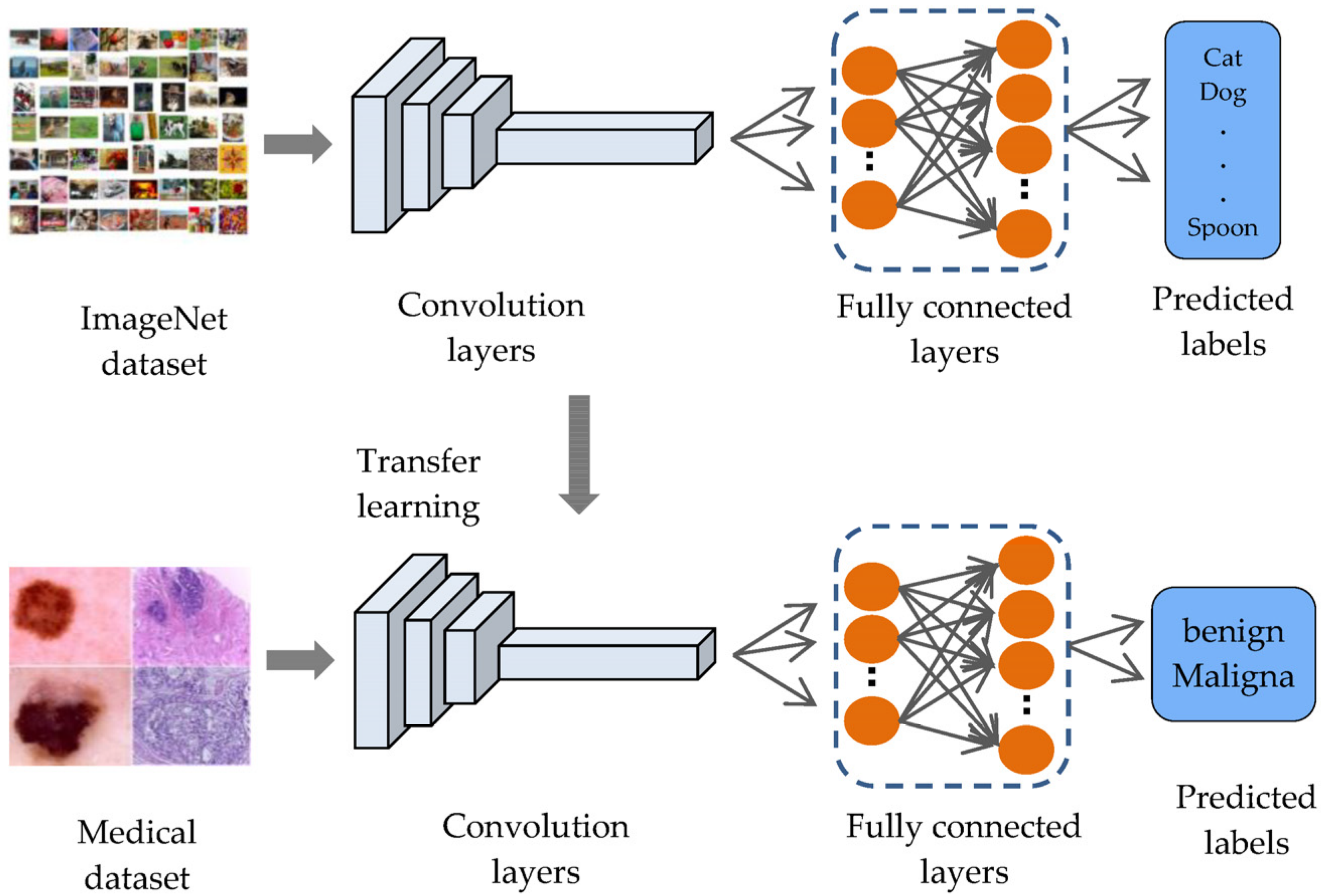
Fine-tuning

Basic concepts



Introduction

Fine-tuning:
"the process of adapting a pre-trained model to a specific task by training it on a smaller, task-specific dataset."



leverage the knowledge learned from a large, general dataset and refine the model's performance on a more specific or targeted problem.

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Fine tuning - overview

1. **Pre-Trained Model:**
   Use a model that has been pre-trained on a large dataset (e.g., ImageNet for images, large text corpora for NLP).

2. **Replace the Final Layers:**
   Replace or modify the final layers of the model to fit the specific output requirements of the target task.
   Example: Change the output layer from 1000 classes (ImageNet) to 10 classes (custom dataset).

3. **Continue the training on the Target Dataset:**
   Task: Fine-tune the model by training it on a smaller, task-specific dataset.
   Optimisation: Use a smaller learning rate to avoid overwriting the pre-learned features.

4. **Evaluate and Adjust:**
   Monitoring: Evaluate the model's performance on the validation set.
   Tuning: Adjust hyper-parameters and training duration as needed.

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

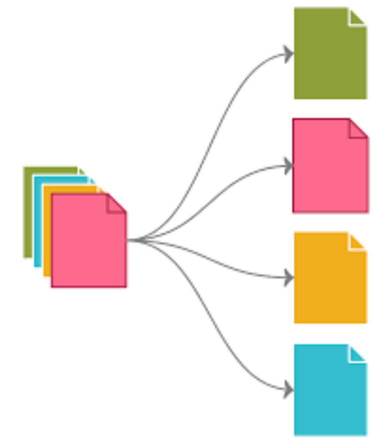
Fine-tuning in NLP - examples

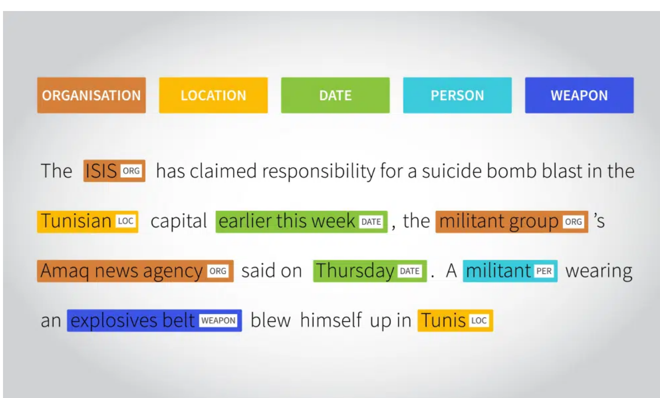
Text Classification:

- **Task:** Classify movie reviews as positive or negative.
- **Example:** Using a pre-trained BERT model, fine-tune it on a dataset of labeled movie reviews to classify sentiment.

Steps:

1. Load a pre-trained BERT model.
2. Replace the final classification layer with a binary classifier.
3. Train the model on the labeled sentiment dataset.





Named Entity Recognition (NER):

- **Task:** Identify entities like names, dates, and locations in text.
- **Example:** Fine-tuning a pre-trained RoBERTa model on a labeled NER dataset such as CoNLL-2003.

Steps:

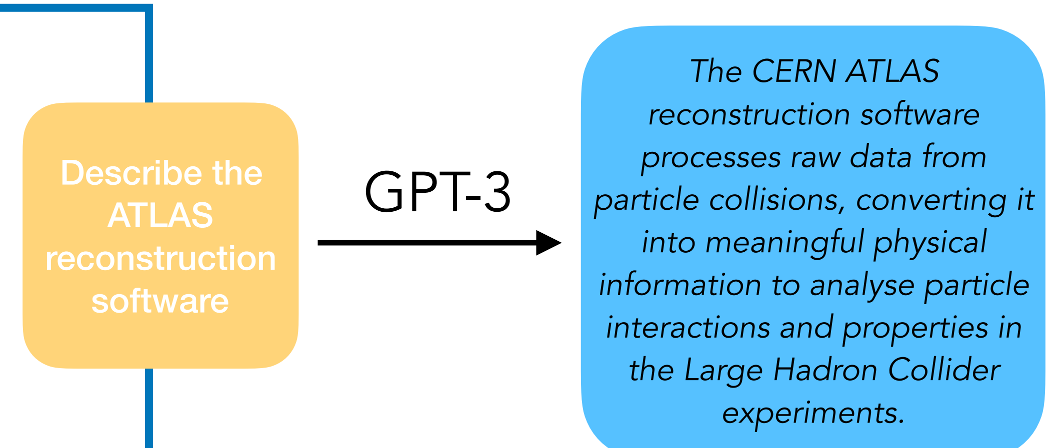
1. Load a pre-trained RoBERTa model.
2. Replace the output layer with a sequence tagging head.
3. Train the model on the NER dataset.

Text Generation (e.g. expert chat-bots):

- **Task:** Generate coherent text based on a prompt.
- **Example:** Fine-tuning GPT-3 or GPT-2 on a specific genre of text (e.g., technical manuals, creative writing).

Steps:

1. Load a pre-trained GPT model.
2. Fine-tune on a corpus of text specific to the desired genre.
3. Use the model to generate text in the target domain.



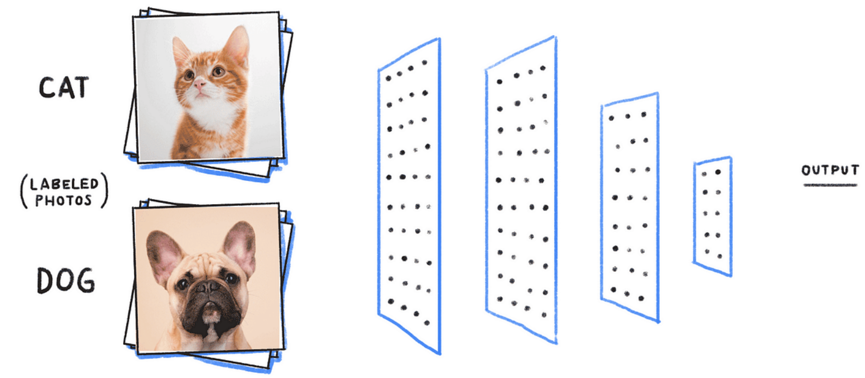
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

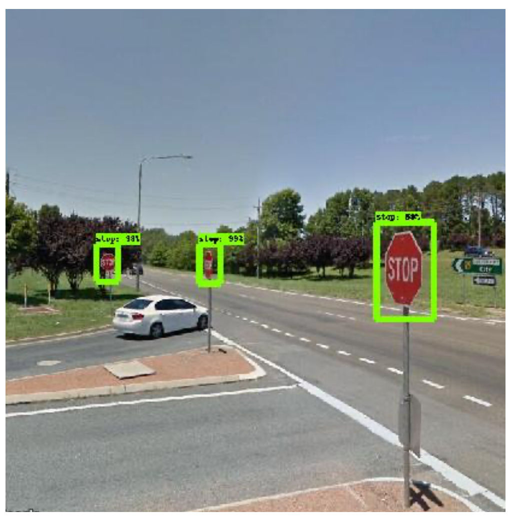
Fine-tuning in Computer Vision - examples

Image Classification:

- **Task:** Classify images into categories (e.g., cats vs. dogs).
- **Example:** Fine-tuning a pre-trained ResNet model on a dataset of pet images.

  Steps:
  1. Load a pre-trained ResNet model.
  2. Replace the final classification layer to match the number of target classes.
  3. Train the model on the pet image dataset.





Object Detection:

- **Task:** Detect and localise objects in images.
- **Example:** Fine-tuning a pre-trained YOLOv3 or Faster R-CNN model on a custom dataset of street signs.

  Steps:
  1. Load a pre-trained object detection model.
  2. Adjust the model for the specific number of object classes.
  3. Train on the labeled object detection dataset.

Image Segmentation:

- **Task:** Segment objects within an image.
- **Example:** Fine-tuning a pre-trained U-Net model on medical imaging data to segment tumours.

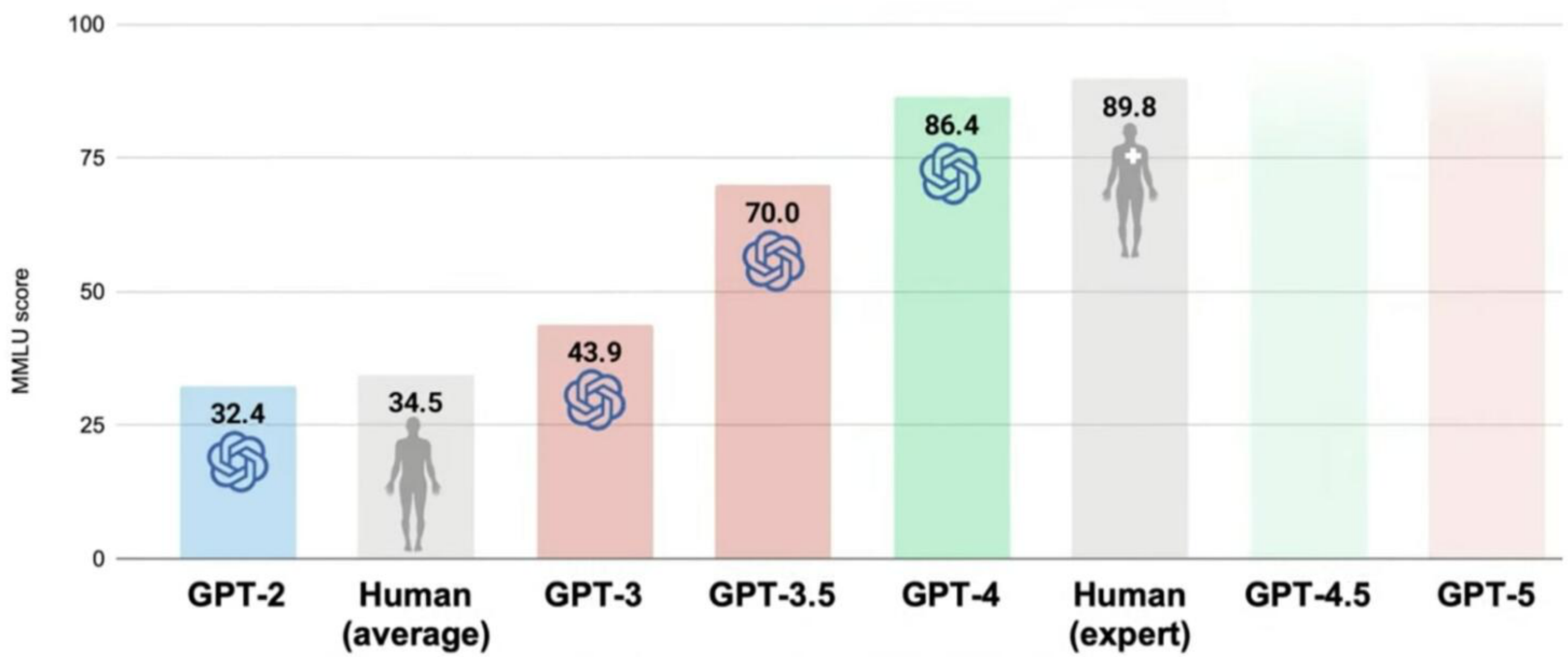
  Steps:
  1. Load a pre-trained U-Net model.
  2. Replace the output layer for segmentation tasks.
  3. Train the model on annotated medical images.

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Benchmarking & model performance

MMLU (Massive Multitask Language Understanding)

MMLU is a benchmark designed to quantify the model knowledge on a variety of language understanding tasks across different domains and topics (STEM, humanities, ..)



Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	Llama 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9

Benchmarking Metrics:

- *Accuracy*
- *F1 Score*

Subjects:

- *Language*
- *Math*
- *Social Science*
- *Humanities*
- *…*

Other evaluation metrics:

- Bilingual EvaLuation Understudy (BELU) BLEU
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation).
- METEOR: explicitly sorted translation evaluation metric.
- Perplexity Perplexity is also called the degree of confusion.

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{N}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

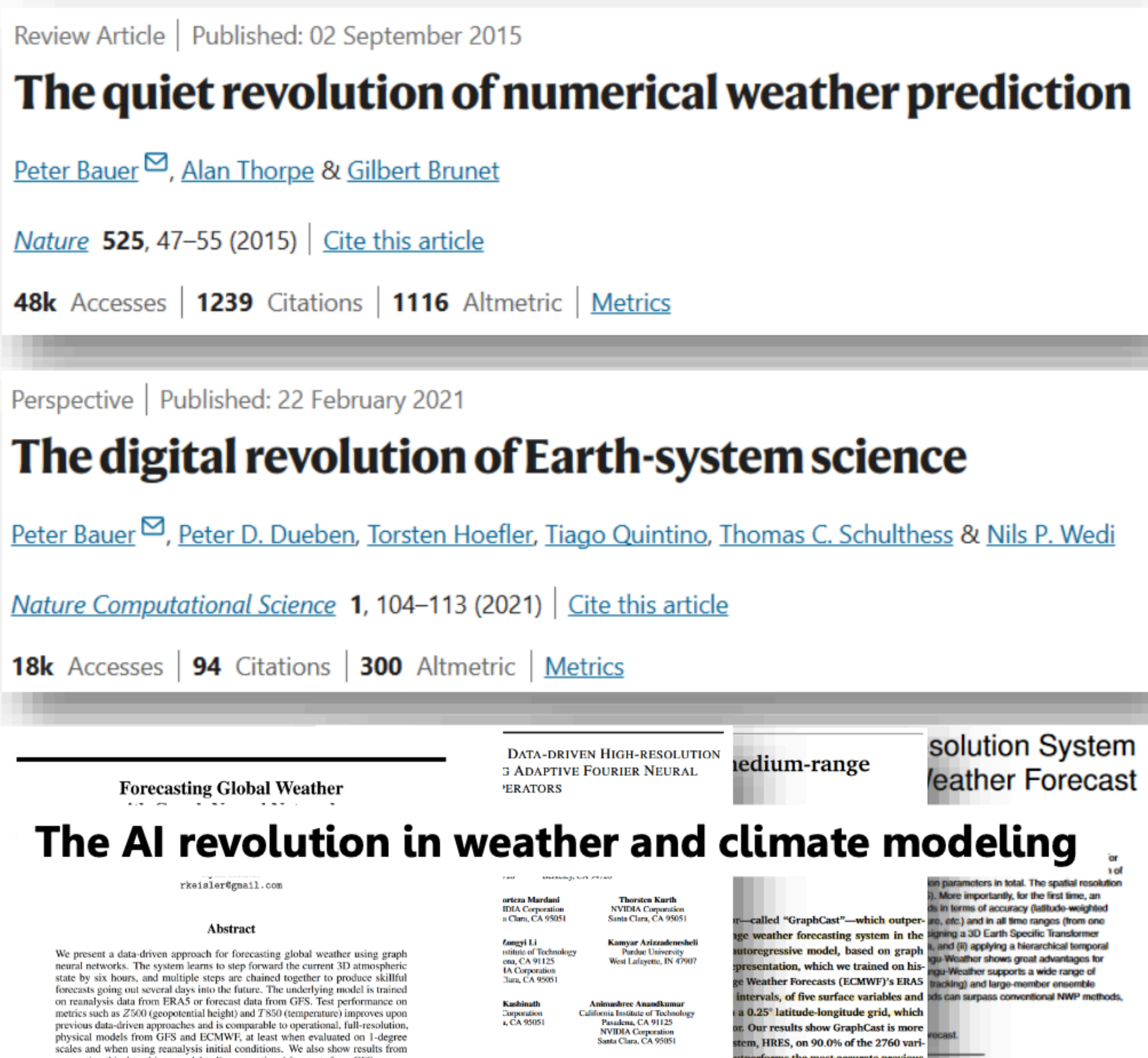
Image from LifeArchitect.ai

Building a foundation model for science

AtmoRep

The first breakthrough: weather & climate

Large datasets:
First time that an AI-model trained on TBs of pre-processed observations outperforms the numerical models for a 10 day forecasts


Review Article | Published: 02 September 2015

The quiet revolution of numerical weather prediction

Peter Bauer, Alan Thorpe & Gilbert Brunet

Nature 525, 47–55 (2015) | Cite this article

48k Accesses | 1239 Citations | 1116 Altmetric | Metrics

Perspective | Published: 22 February 2021

The digital revolution of Earth-system science

Peter Bauer, Peter D. Dueben, Torsten Hoefler, Tiago Quintino, Thomas C. Schulthess & Nils P. Wedi

Nature Computational Science 1, 104–113 (2021) | Cite this article

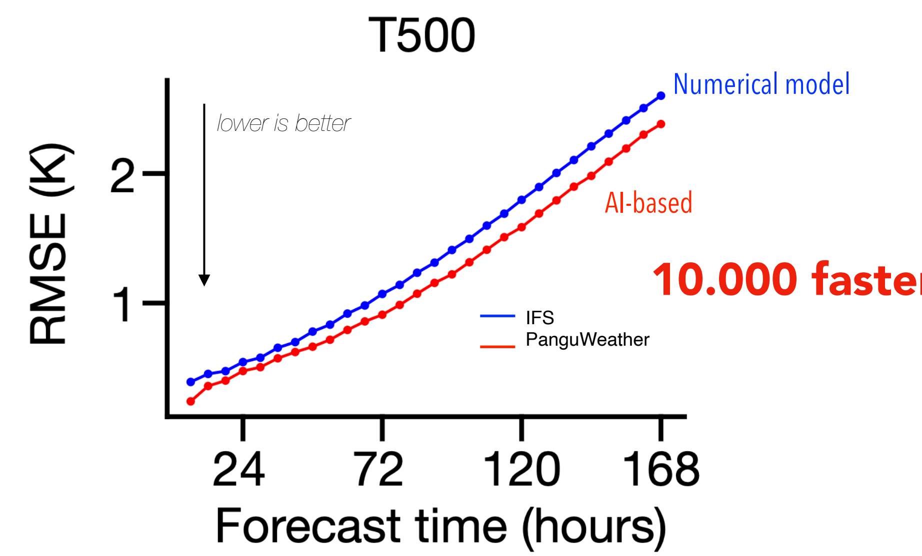
18k Accesses | 94 Citations | 300 Altmetric | Metrics

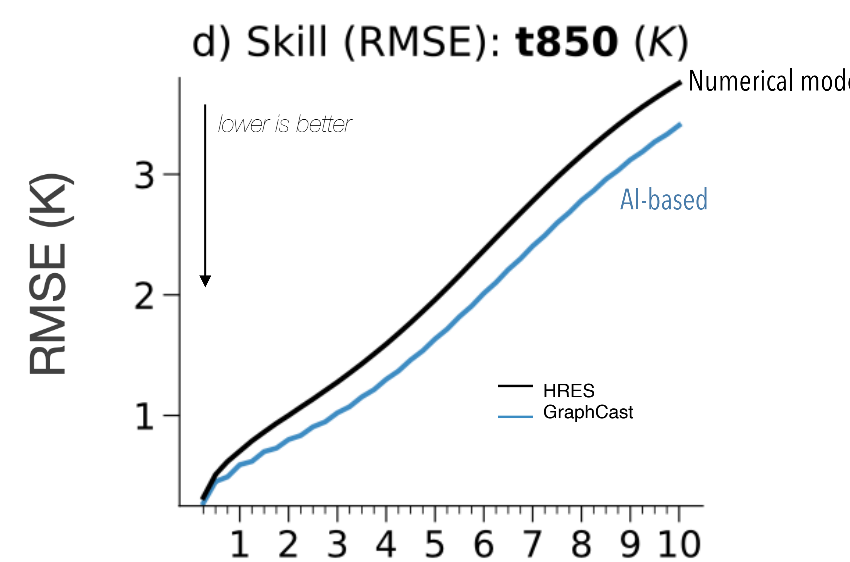
The AI revolution in weather and climate modeling

1960-2010

2005-2025

2022-





All these models have been trained on a *single* task: weather forecasting

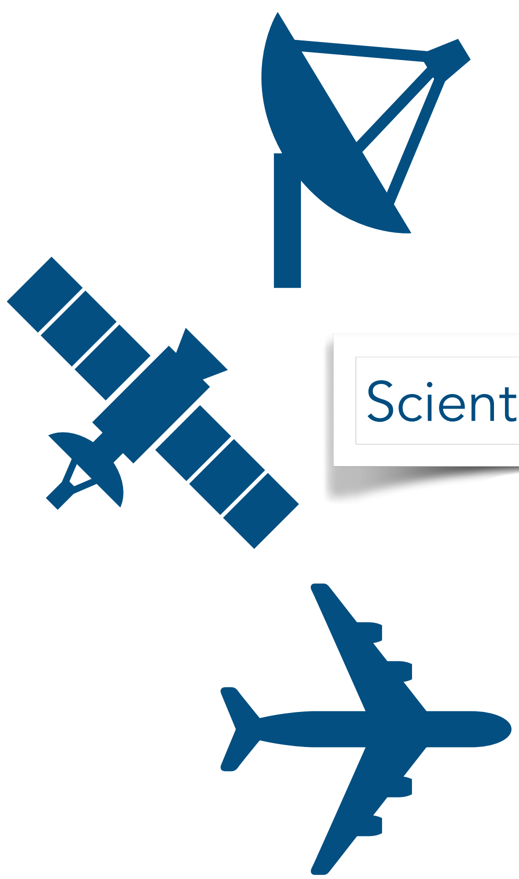
From Martin Schultz - ESA-ECMWF workshop 2024

Can we go beyond?

Building foundation models for science

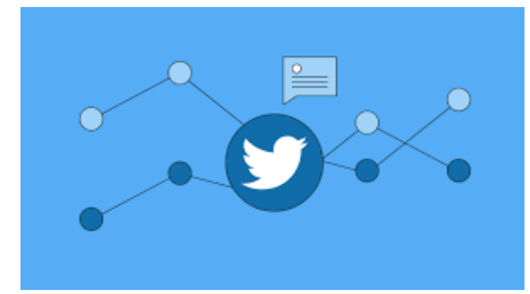
Multimodality

Data are getting **more and more multi-modal and the relationship between them is very complex to model**
(and requires all kinds of approximations)

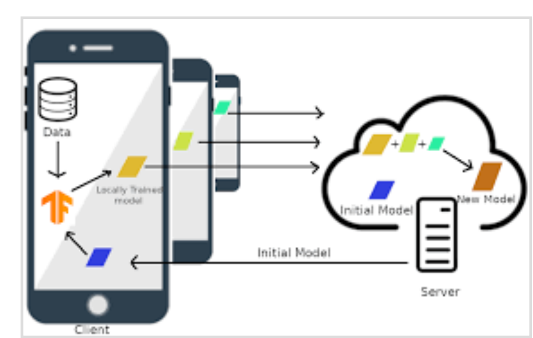


Scientific Data → **Policy-oriented scientific models** ← New data types

Social media



Economic growth GDPs, birth rates…



Data from distributed devices

Conventional approaches for analysing and processing the data come to their limits

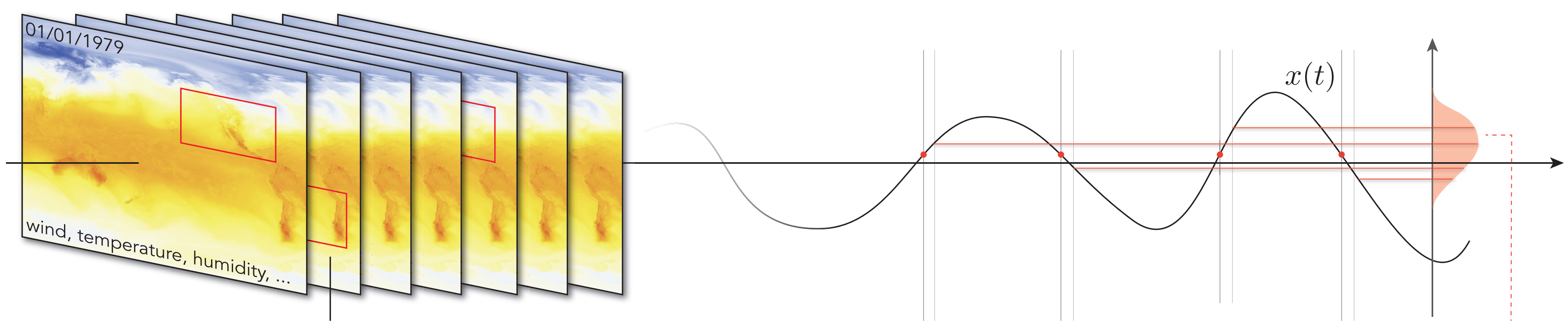
What is a task-independent model for us?

Encapsulate the spatio-temporal evolution of a dynamical system

Probability of getting the state y given the initial state x and the auxiliary info α

$$p(y \,|\, x, \alpha)$$

Auxiliary info: position, absolute time etc..



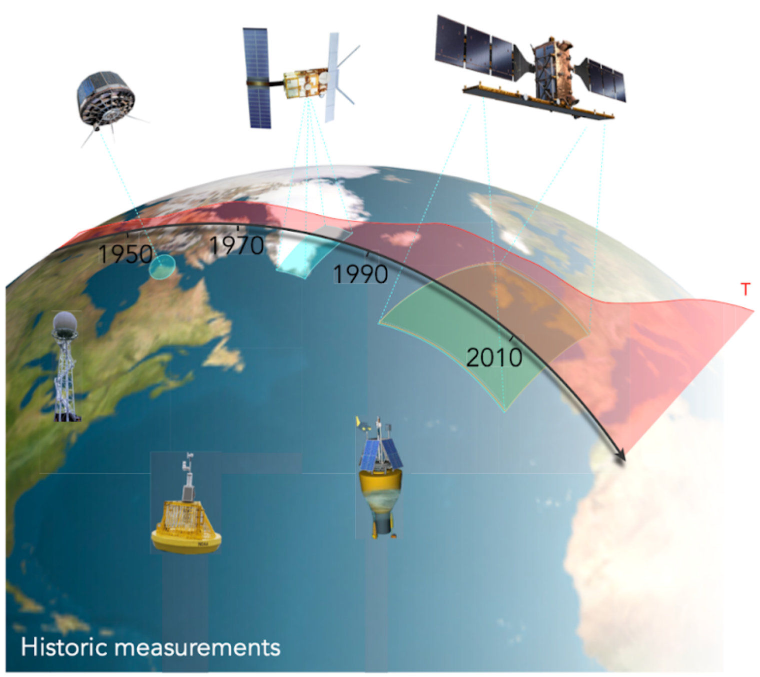
Training

$$p(y \,|\, x, \alpha) \approx p_\theta(y \,|\, x, \alpha)$$

foundation model:
neural network that models data distribution for a specific domain

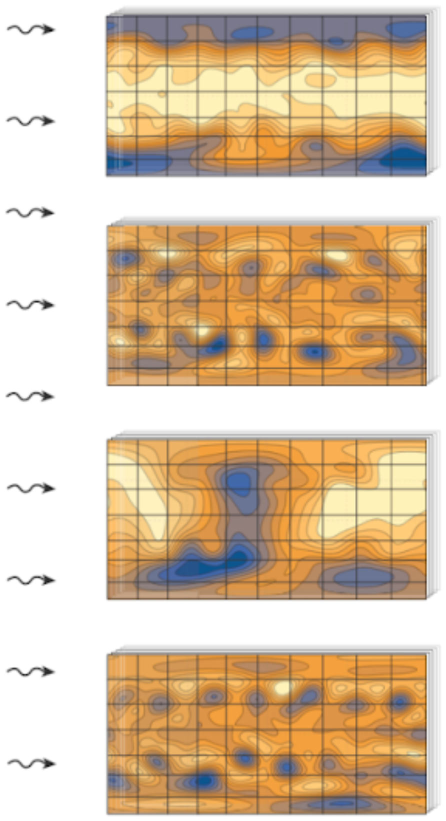
The distribution can be approximated by a large neural network

The project in a nutshell



Observations

ERA5 reanalysis

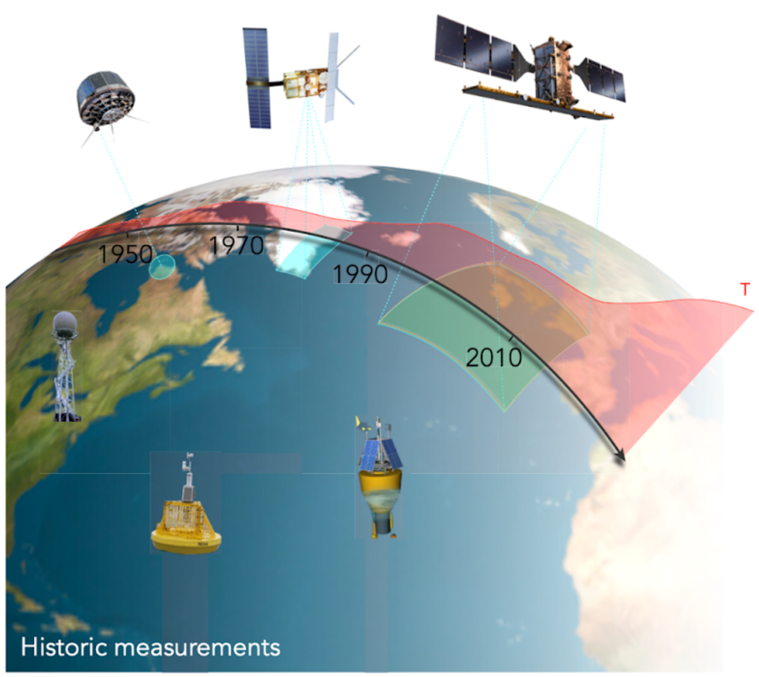


50 TB

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

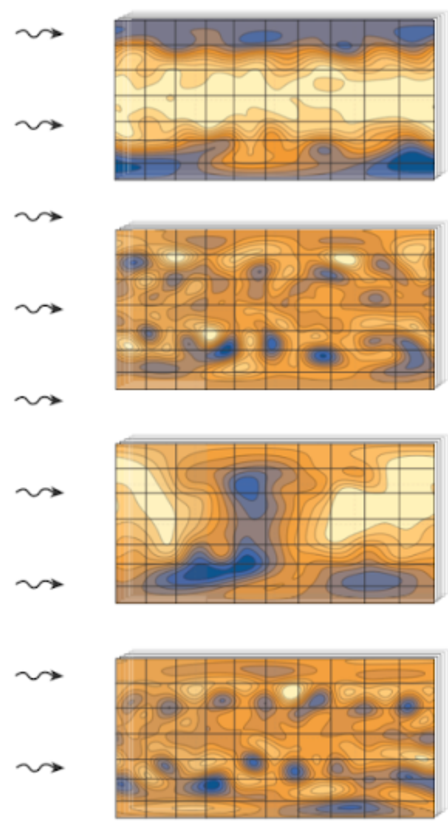
The project in a nutshell

Spatio-temporal representation of atmospheric dynamics

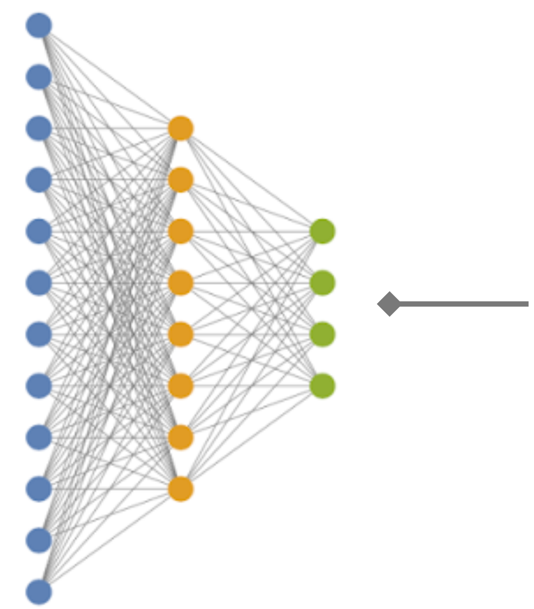


Observations

ERA5 reanalysis

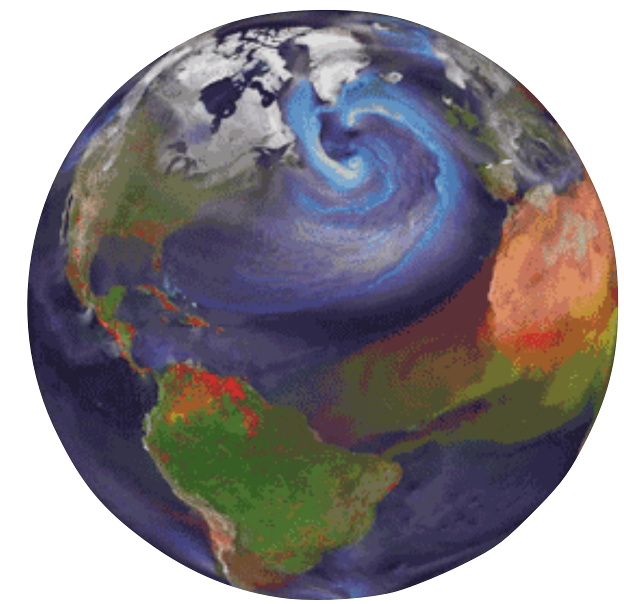


50 TB



large scale machine learning

Transformers architecture



Model
given by the trained neural network

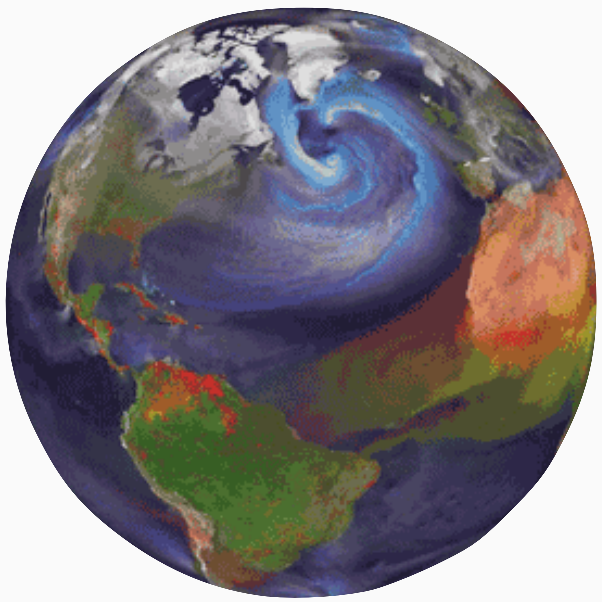
R&D at Juelich SSC:
$4x10^6$ GPU hours granted in 2023

JÜLICH Forschungszentrum

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Applications: one model for multiple purposes

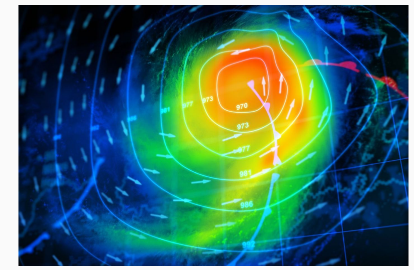
Spatio-temporal representation of atmospheric dynamics



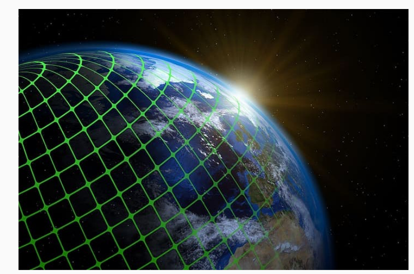
Model
given by the trained neural network

Task-dependent adaptable smaller networks

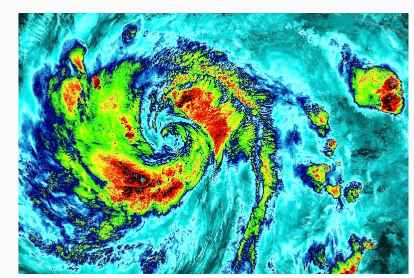
Adaptation



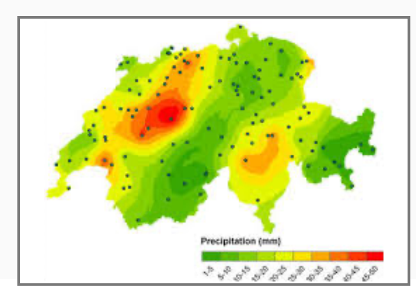
Weather predictions



Downscaling



Bias corrections



Spatio-temporal Interpolations (WIP)

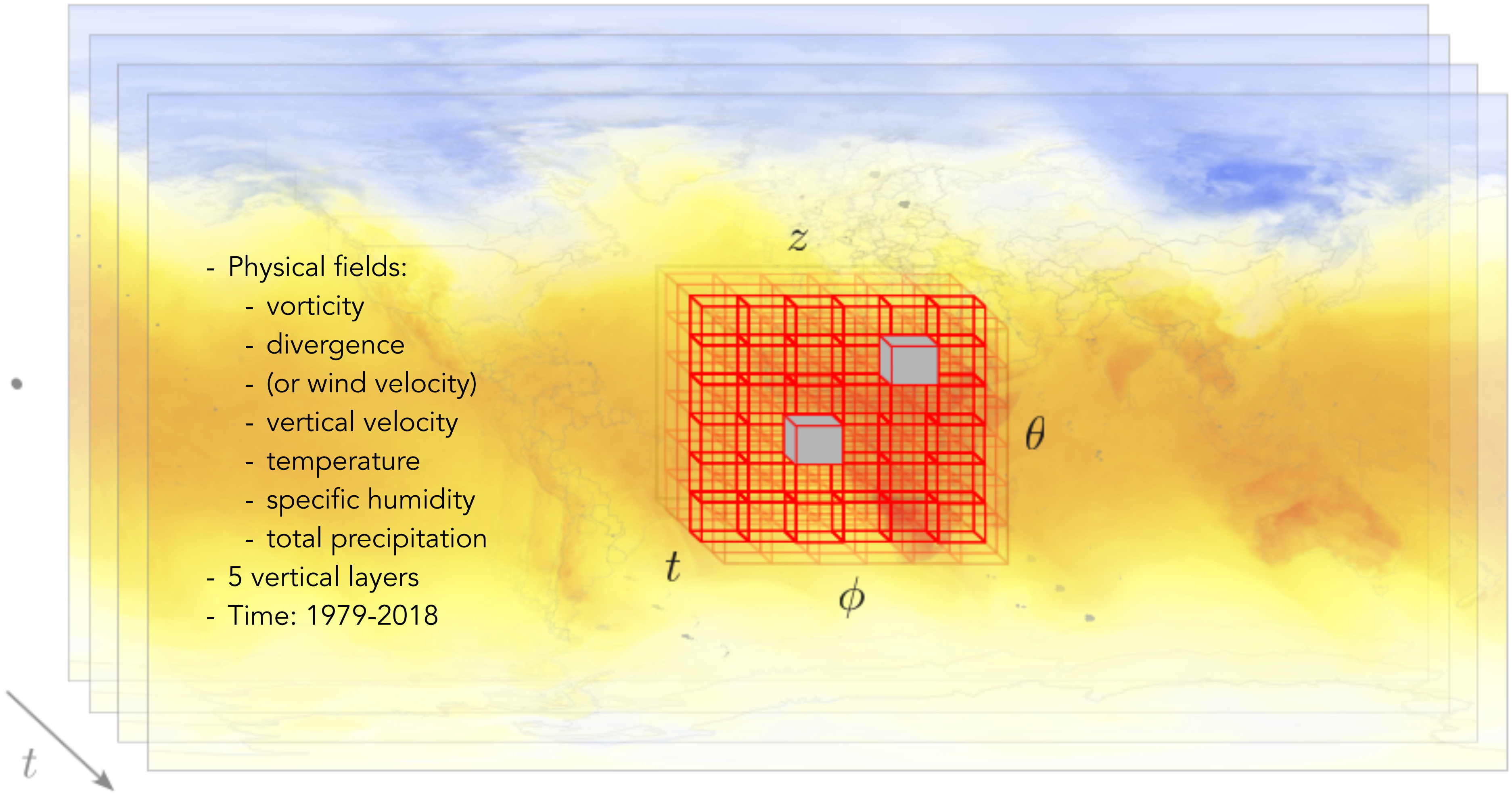
Physics-related applications = uncertainties
Need for a ***stochastic approach***

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

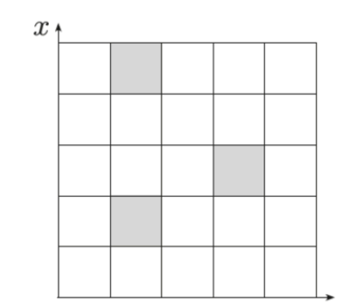
Key Ingredient: The training protocol

Use an extension of BERT masked language modelling from self-supervised trainings in NLP

Random sampling of neighbourhoods for training



BERT



Split cube in small space-time regions (3D cubes) → tokens

Mask random tokens within the hyper-cube and predict them

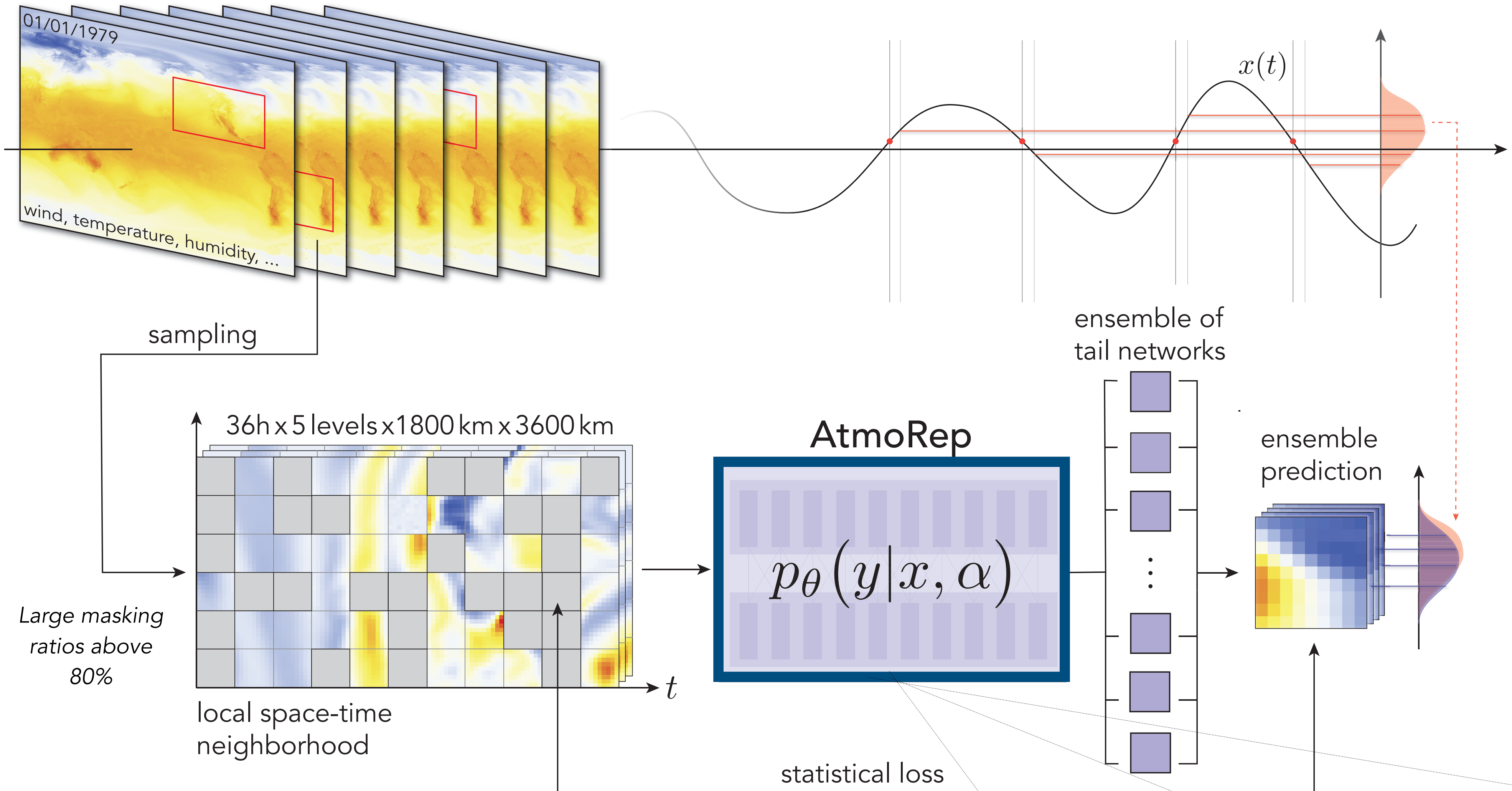
Large masking ratios above 80% using full masking, noise and climatology

Default: 12 x 6 x 12 tokens with 3 x 9 x 9 grid points

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

The AtmoRep workflow

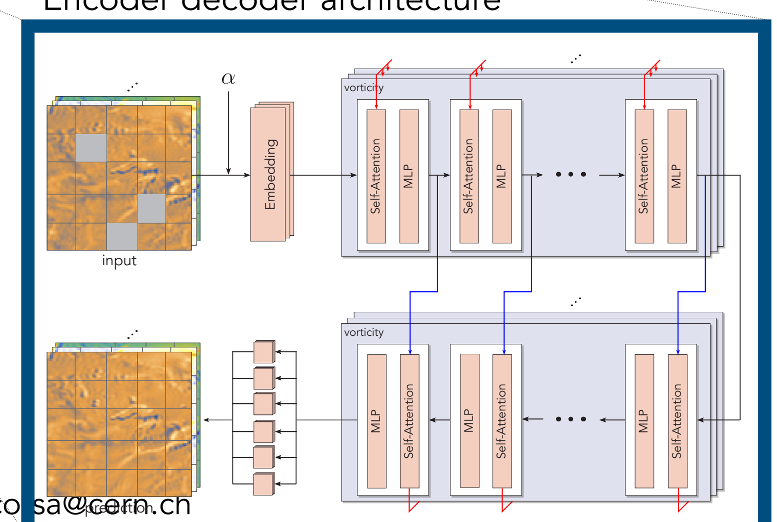
pre-processed historical observational record $x(t)$ (ERA5 reanalysis)



New stochastic approach

ensemble predictions with 16 members

Encoder decoder architecture



Approximate the 4-Dim PDF of the process using a Transformers-based network with 3.5 billion parameters

Task-specific fine-tuning

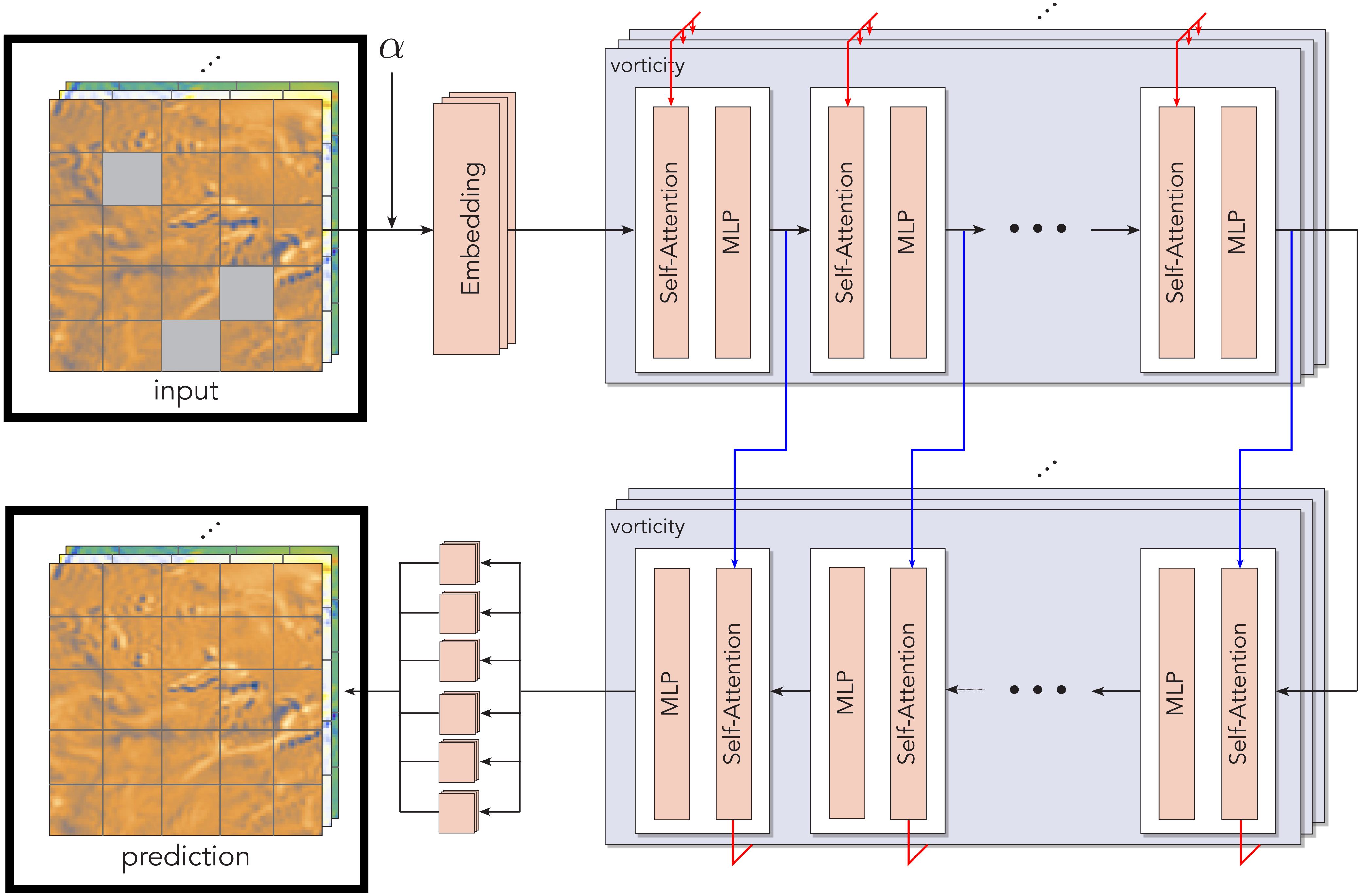
Goal: improve model performance for a specific task
e.g. forecasting, downscaling…

Examples:

e.g. fix masking scheme

OR

Change target dataset



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

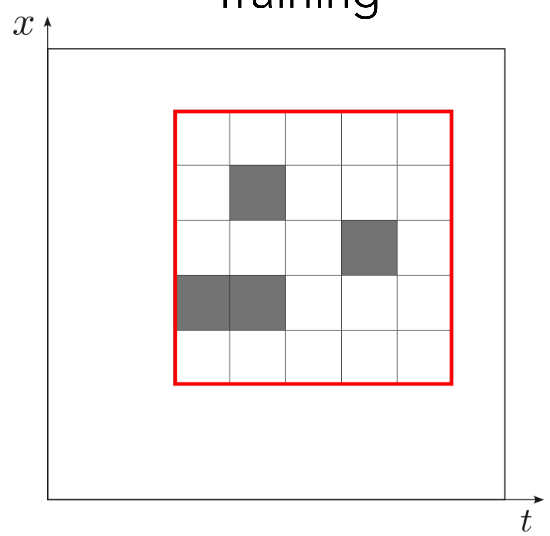
Task-specific fine-tuning

Goal: improve model performance for a specific task
e.g. forecasting, downscaling…

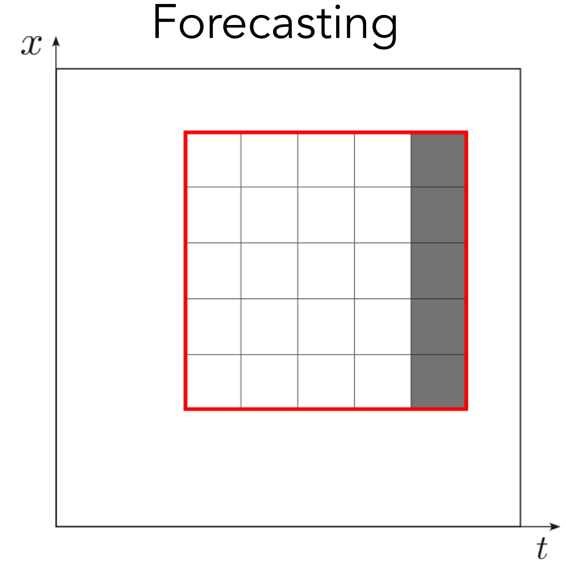
Examples:

e.g. fix masking scheme

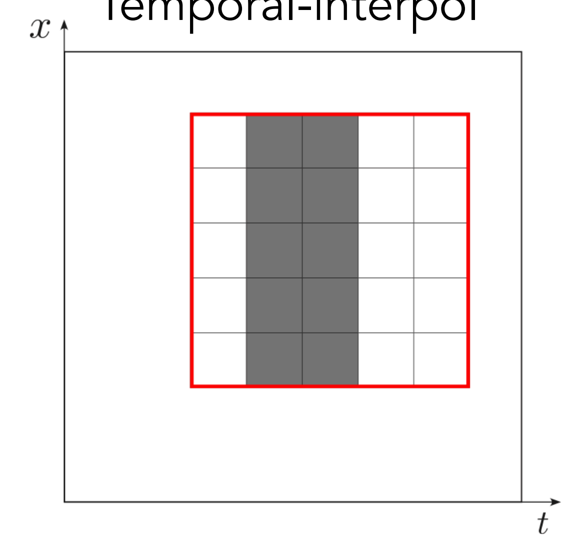
Training



Forecasting

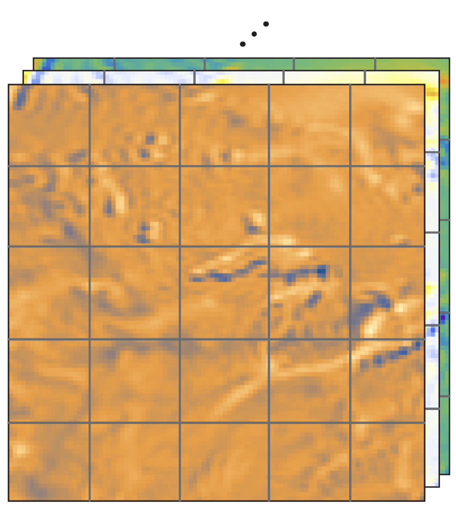


Temporal-interpol



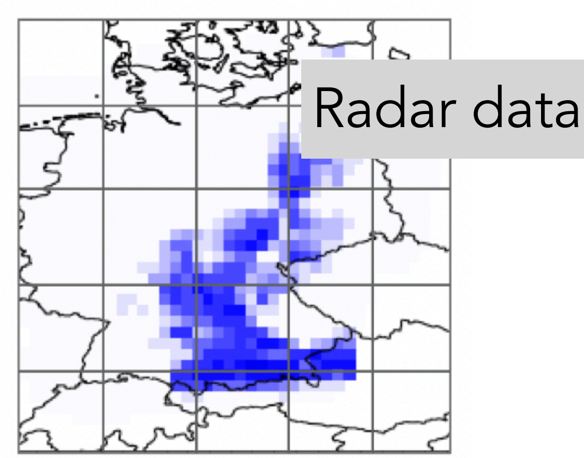
OR

Change target dataset

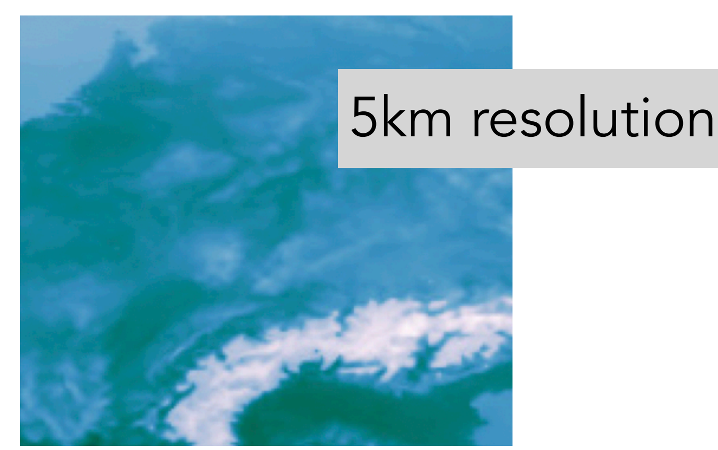


ERA5

Radklim data



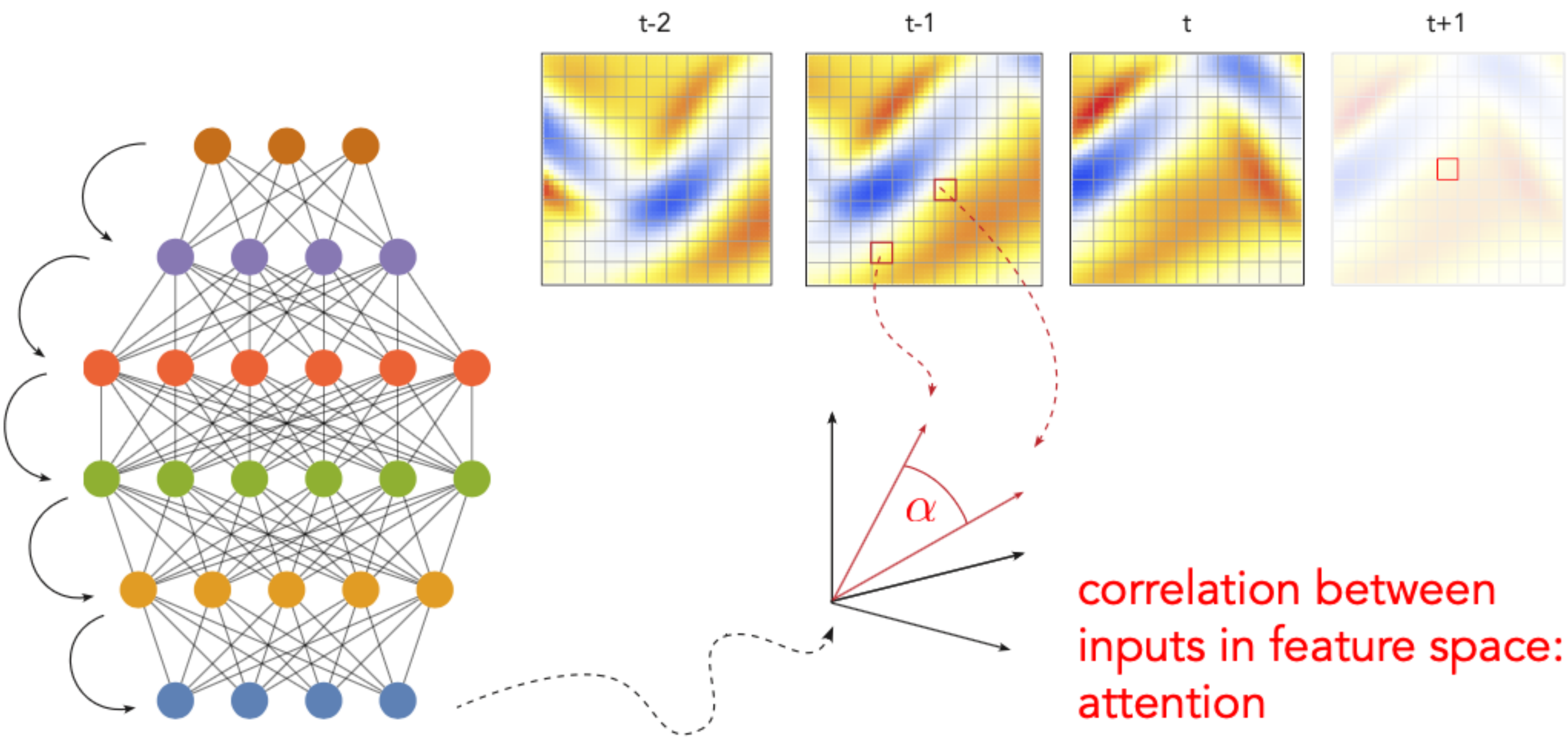
COSMO-REA6



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Attention maps and interpretability

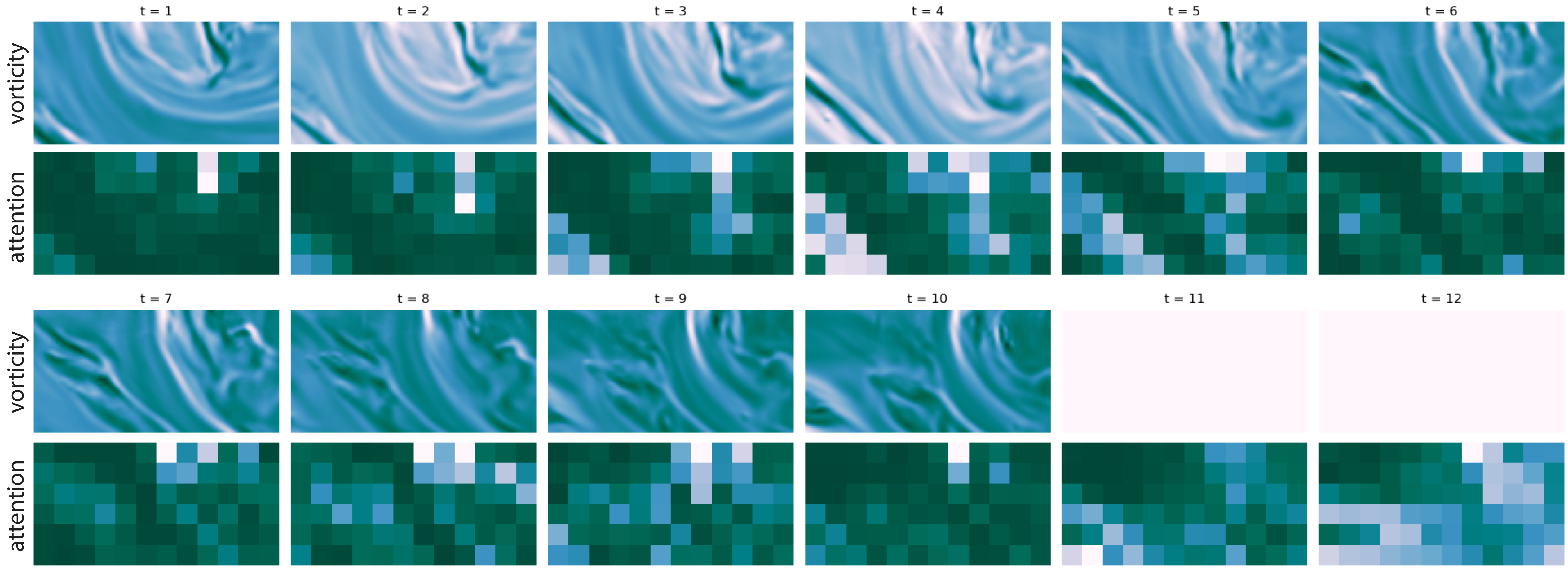
Inspect the self-attention mechanism:
can we identify physics phenomena (e.g. hurricane formation) before they are even created?



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Attention maps and interpretability

Inspect the self-attention mechanism:
can we identify physics phenomena (e.g. hurricane formation) before they are even created?



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

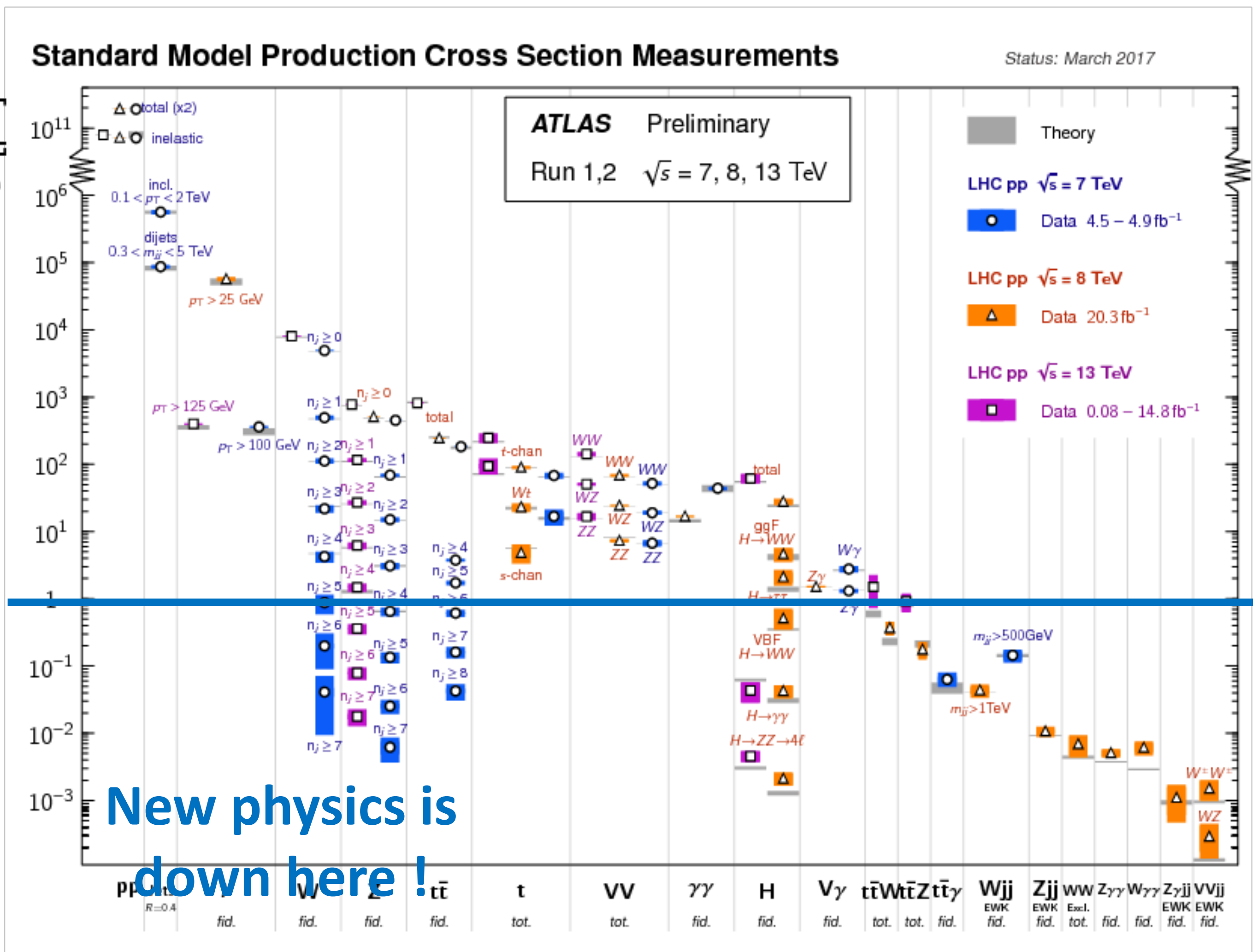
So WHERE and HOW can we use Foundation Models in HEP?

NB: LLMs are quickly entering our domain

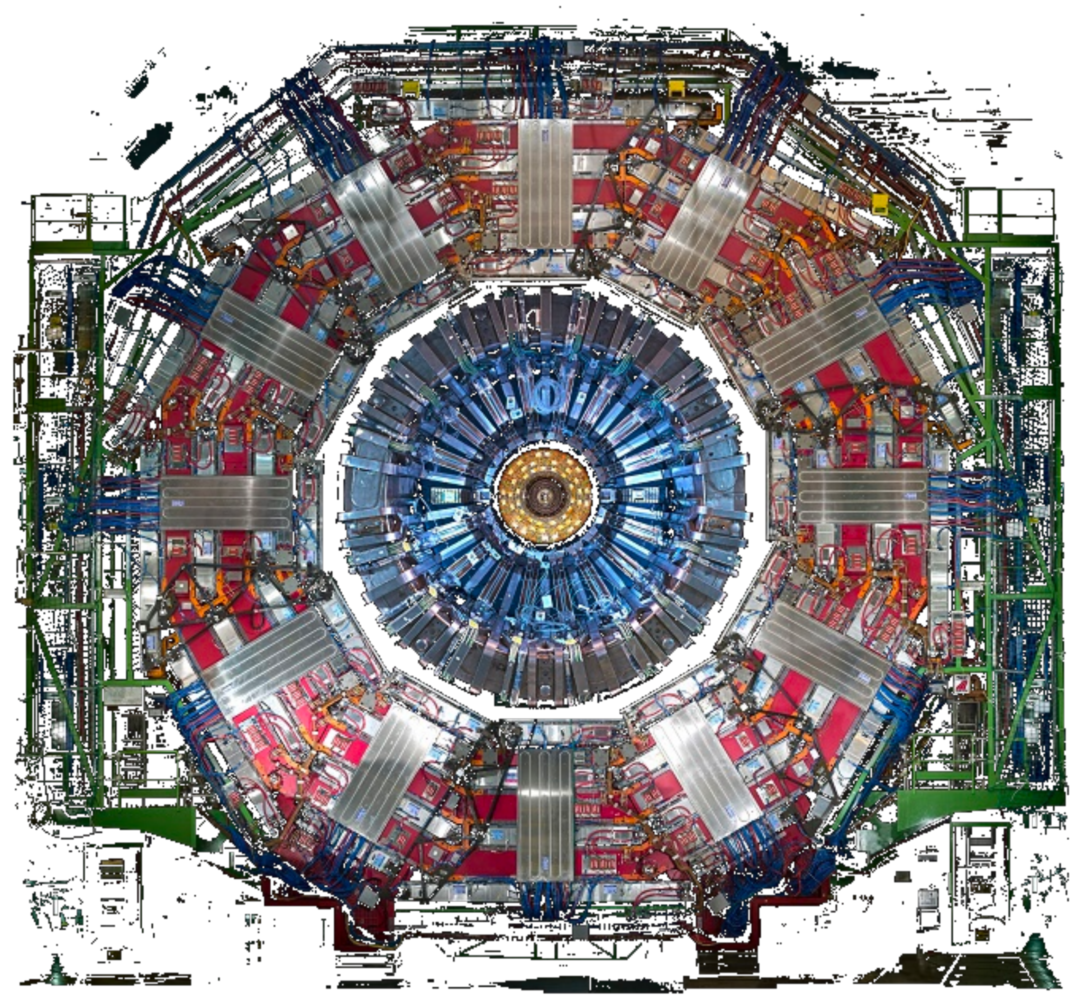
So WHERE and HOW can we use Foundation Models in HEP?

NB: LLMs are quickly entering our domain

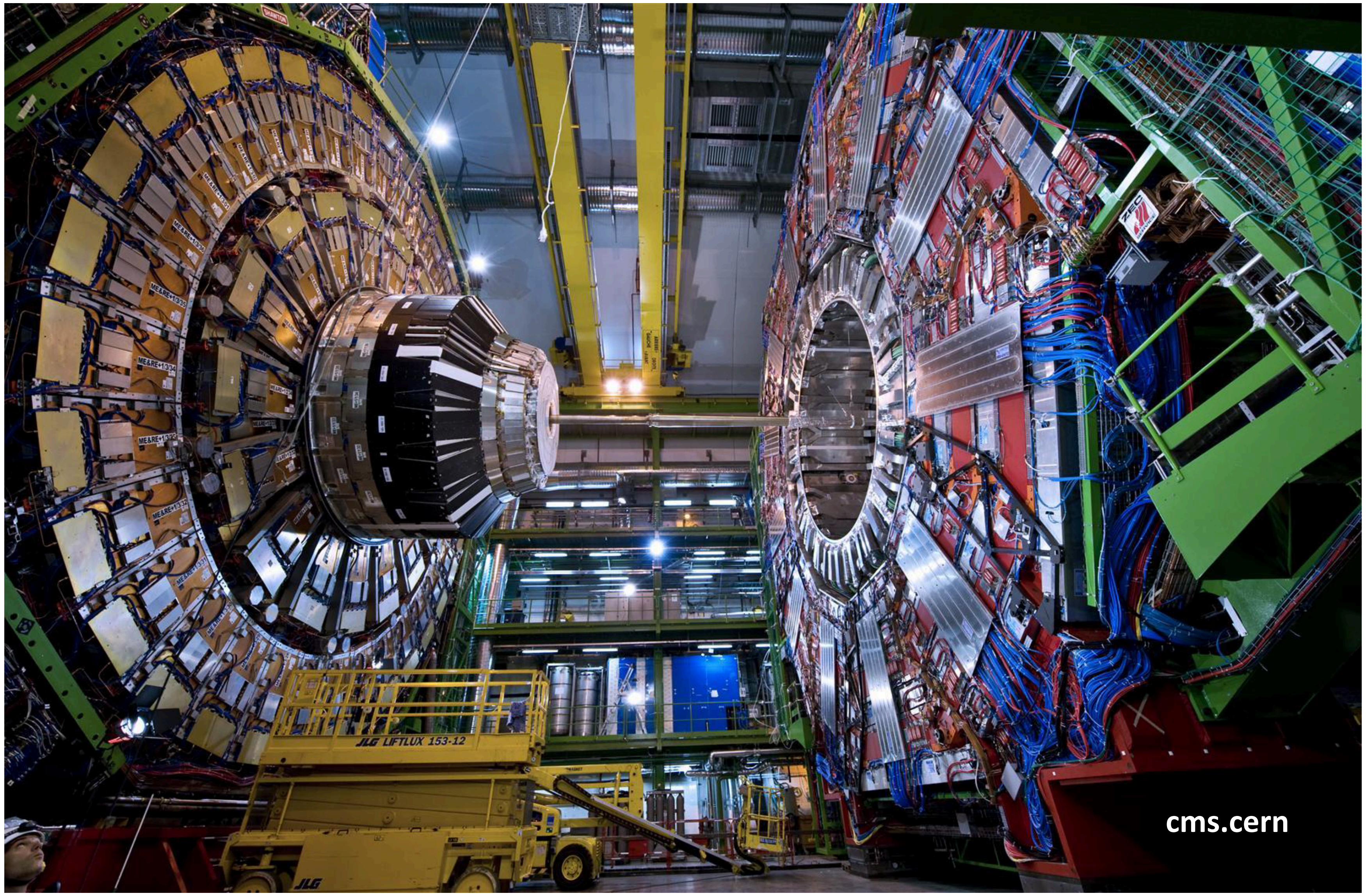
New Physics search as a Big Data problem

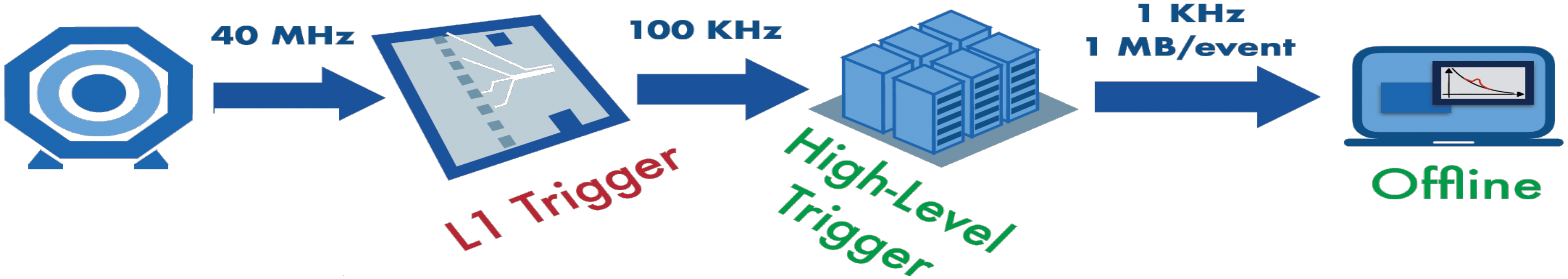


> 600 PB of collisions data



LHC data processing

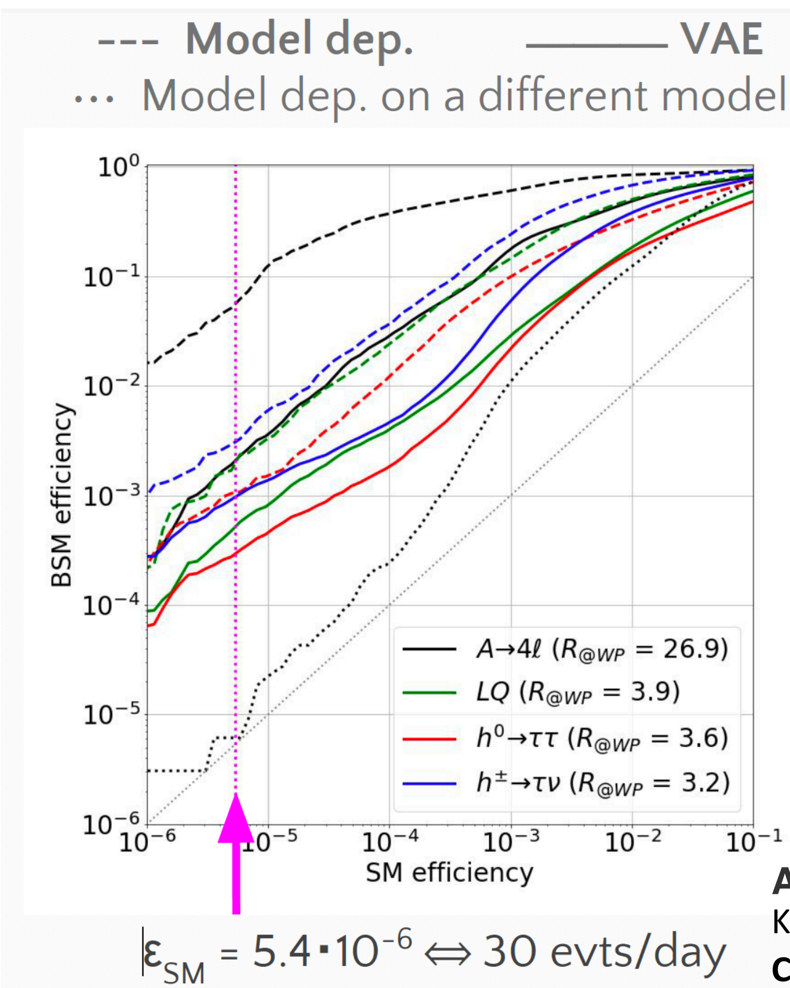




Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Selecting the unknown

Unsupervised and model independent tools for new physics searches

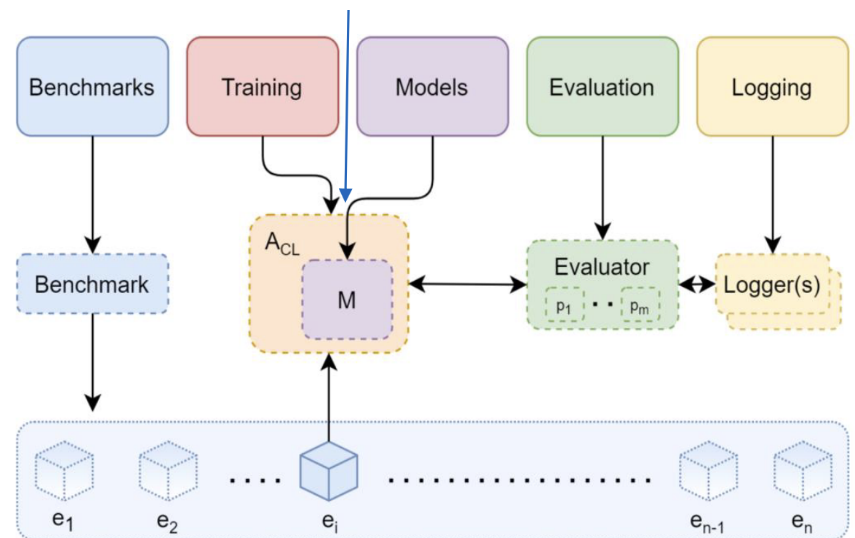


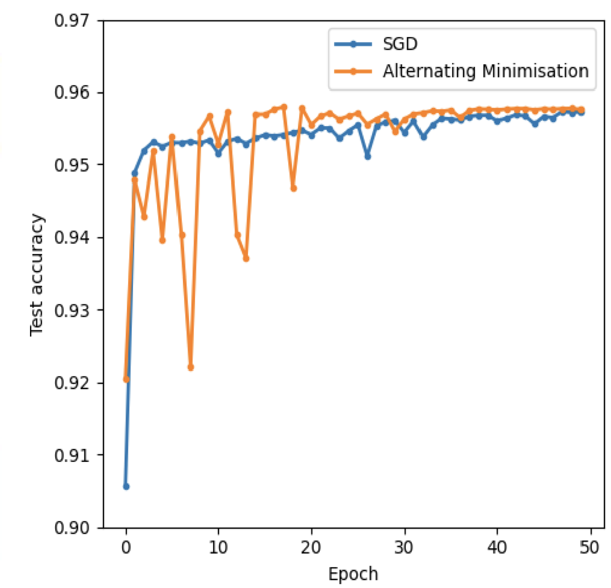
Arxiv:1811.10276. Evolved into:

Knapp, Oliver, et al. "**Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark**." *The European Physical Journal Plus* 136.2 (2021): 236.

Continual learning for online environment

Useful with **changing conditions**. Avoid retraining. Strong computational constraints. Proposed lightweight alternative to SGD.





Embedded Continual Learning for HEP, CHEP2023

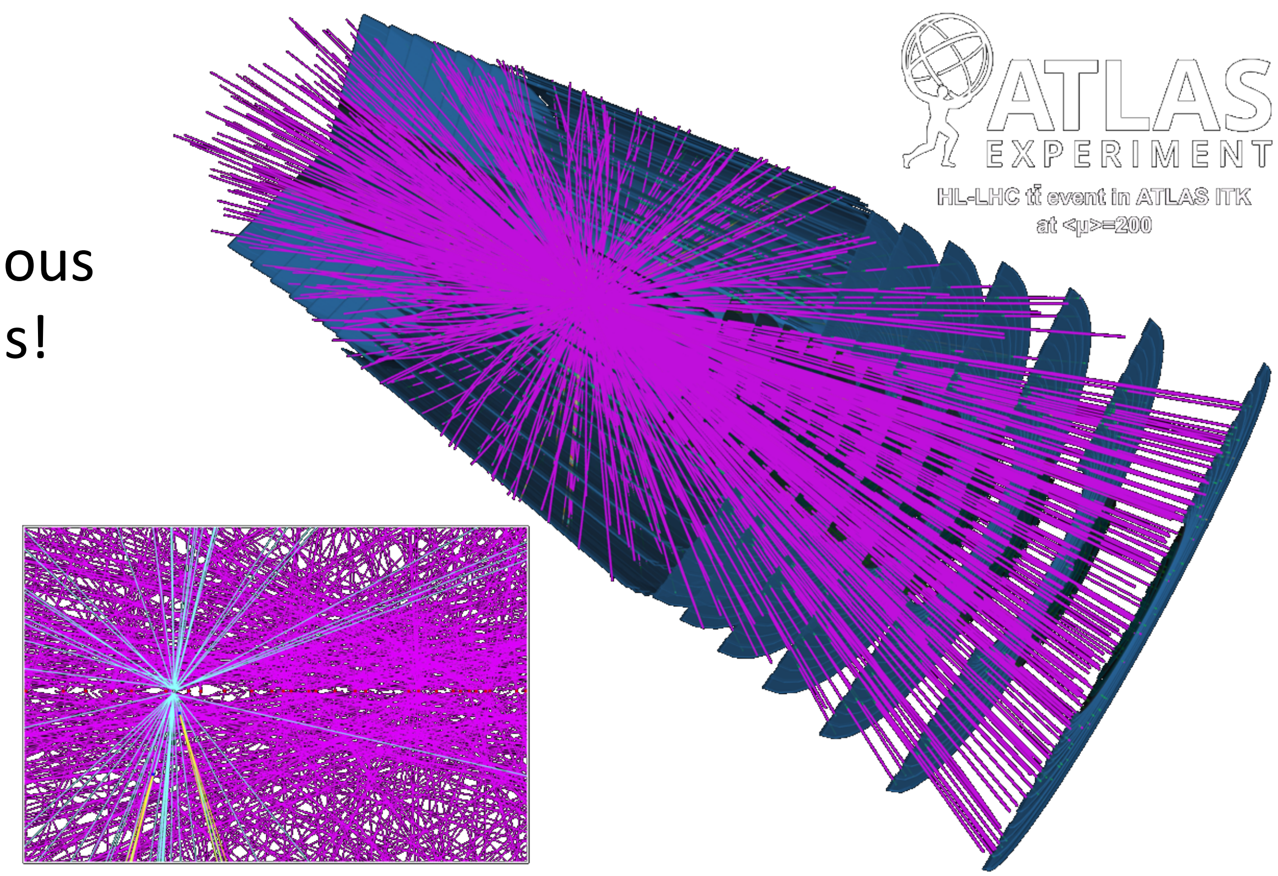
PATTERN RECOGNITION

BE QUIET DARLING. LET PATTERN RECOGNITION HAVE ITS WAY

William Gibson

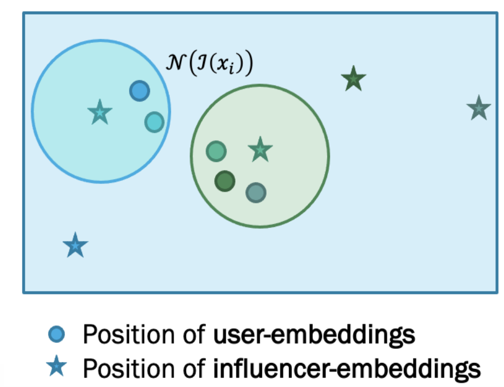
Multiple data processing tasks are formulated as pattern recognition and solved with AI:
Point clouds and transformers, geometric learning & GNN, RNN, CNN, etc…

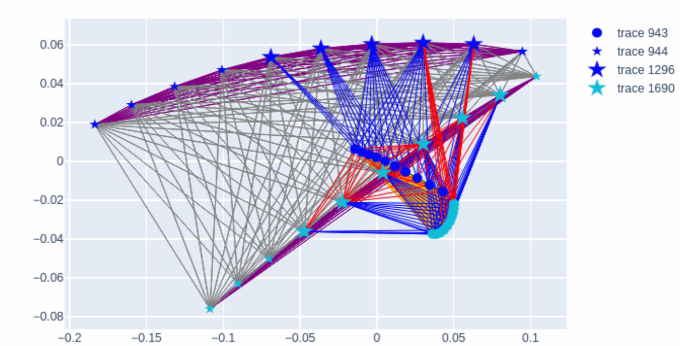
200
multaneous
collisions!



Reconstruct particle trajectories using **the influencer loss ! (social media inspired)**

An Object Condensation Pipeline for Charged Particle Tracking, CHEP2023





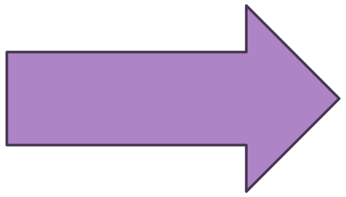
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

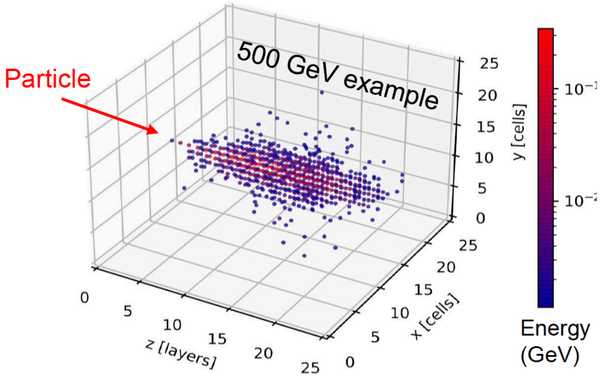
Synthetic data generation

CYBERSPACE. A CONSENSUAL HALLUCINATION EXPERIENCED DAILY BY BILLIONS OF LEGITIMATE OPERATORS

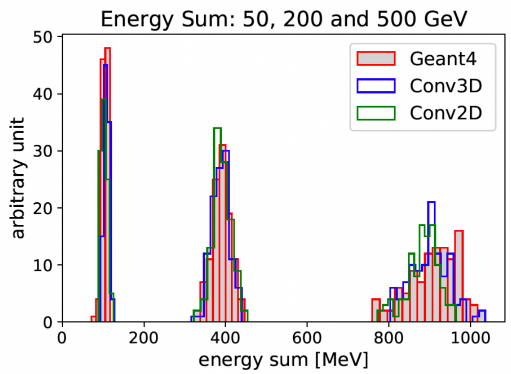
William Gibson

A major task, requiring high accuracy.
It is computationally expensive **(typically Monte Carlo based)**
Ideal task for state-of the-art generative AI

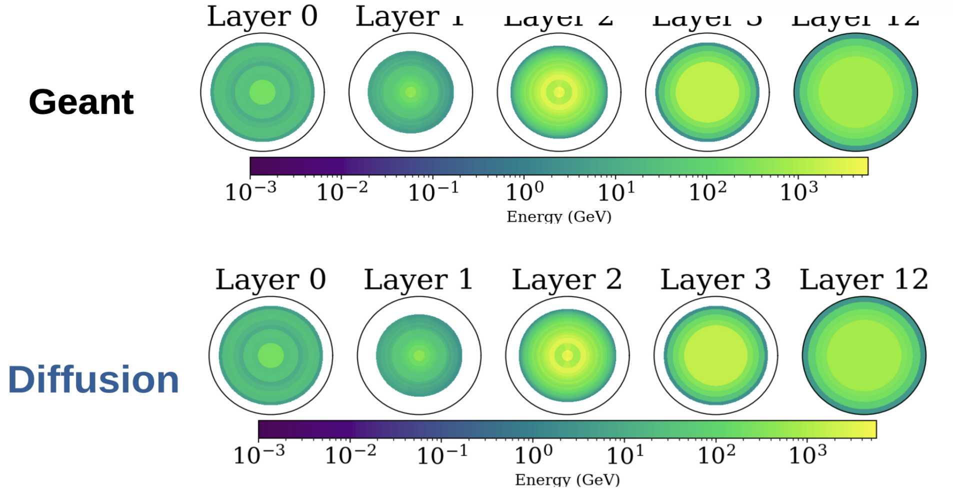




Rehm, Florian, et al. *arXiv:2105.08960* (2021).



Diffusion models for shower generation, CHEP2023



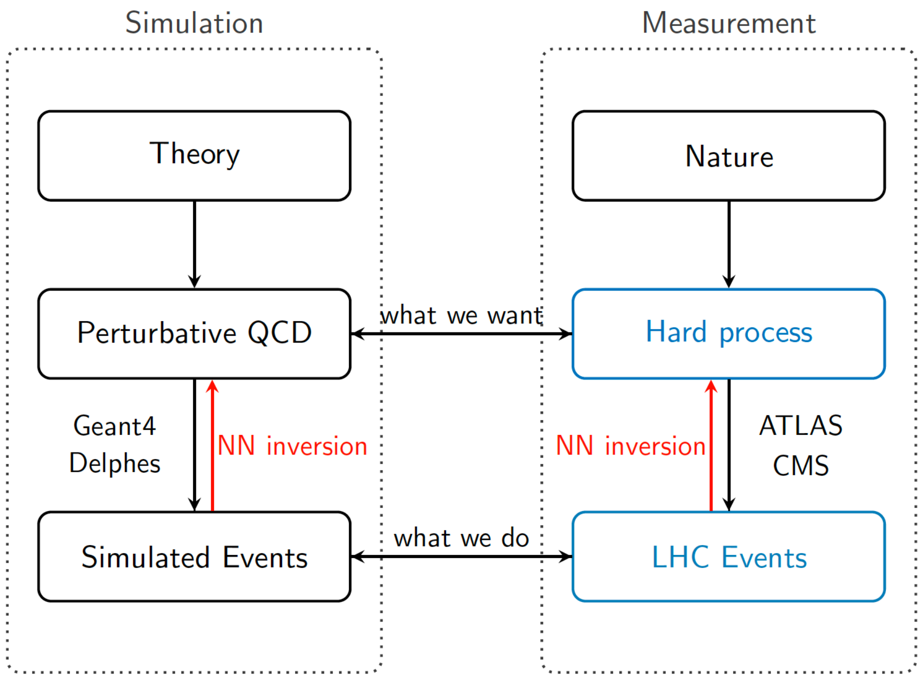
Inverting the experiments

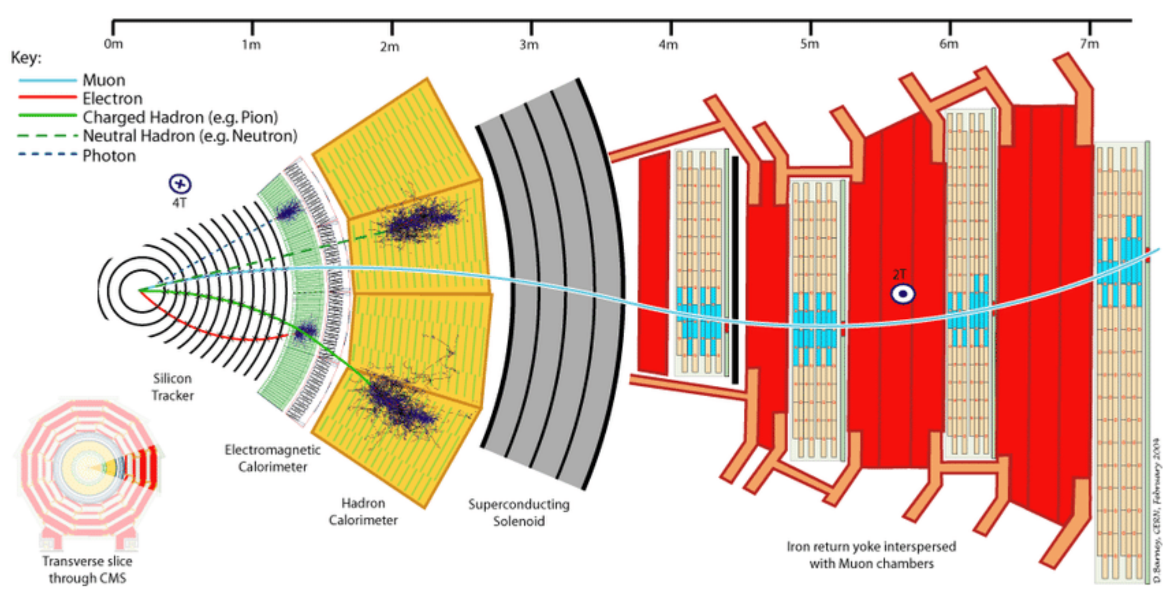
TIME MOVES IN ONE DIRECTION. MEMORY ANOTHER. WE ARE THAT STRANGE SPECIES THAT CONSTRUCTS ARTEFACTS INTENDED TO COUNTER THE NATURAL FLOW OF FORGETTING

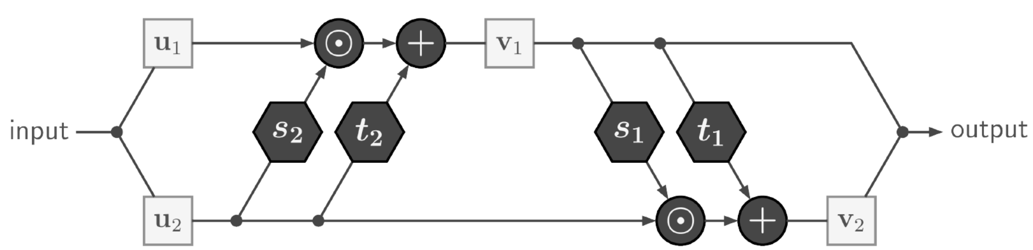
William Gibson

Detectors measure the results of **particle interactions** with matter but we need the **particle production processes**

- **Compare experimental data to theory** through invertible networks !







arxiv:1808.04730
arxiv:2006.06685

Automation

THE FUTURE IS ALREADY HERE – IT'S JUST NOT EVENLY DISTRIBUTED YET

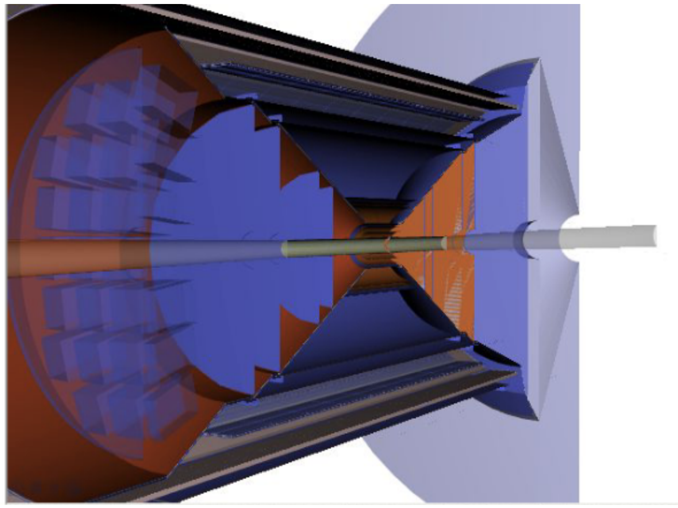
William Gibson

Autonomous inspection and environmental measurements
Autonomous control systems

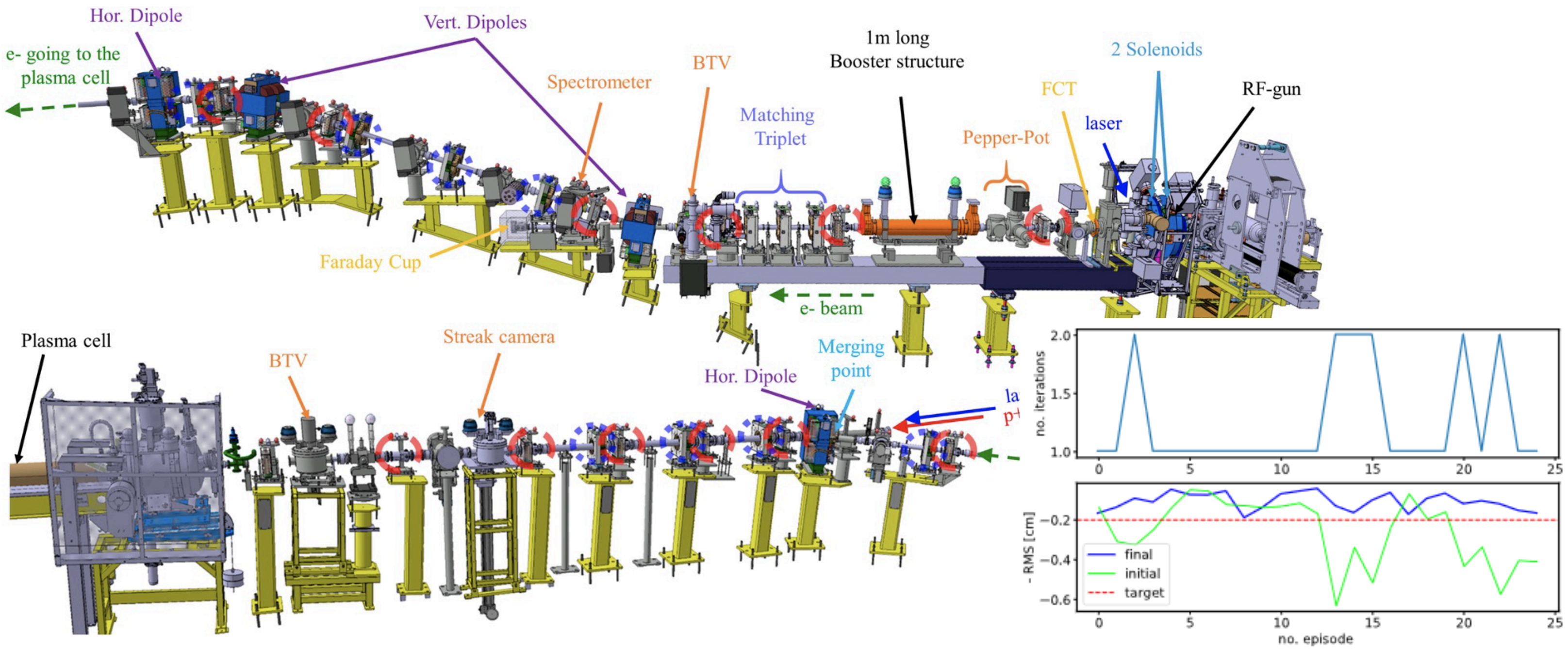
CERN Academic Lecture Series: Robotics activities at CERN - Robotic Solutions for remote maintenance, 2022, https://indico.cern.ch/event/1055745/

EPIC: First large scale experiment designed using AI/ML !

Artificial Intelligence and Machine Learning for EPIC: an Overview, CHEP2023



Reinforcement Learning agents for Beam Control at CERN



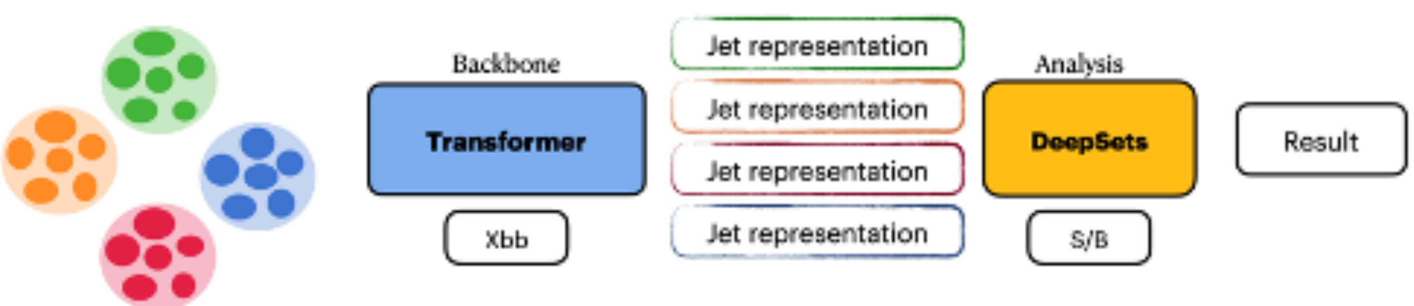
Kain, Verena, et al. "**Sample-efficient reinforcement learning for CERN accelerator control**." *Physical Review Accelerators and Beams* 23.12 (2020): 124801.

So WHERE and HOW can we use Foundation Models in HEP?

NB: LLMs are quickly entering our domain

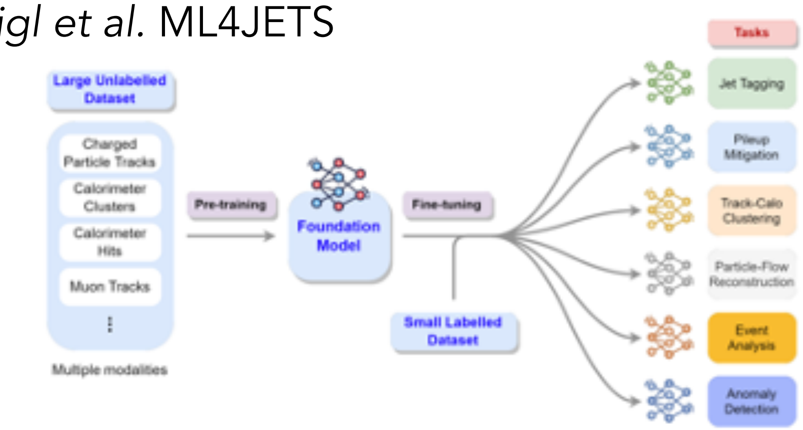
Foundation Models in HEP

Multiple studies in HEP (transformers, self supervision, fine tuning for HEP data, etc..)
A topic present in **many conferences and workshops**, (IML, ACAT, CHEP, ML4JET, …)
Direct **application of LLMs** to HEP (information mining, coding, etc..)

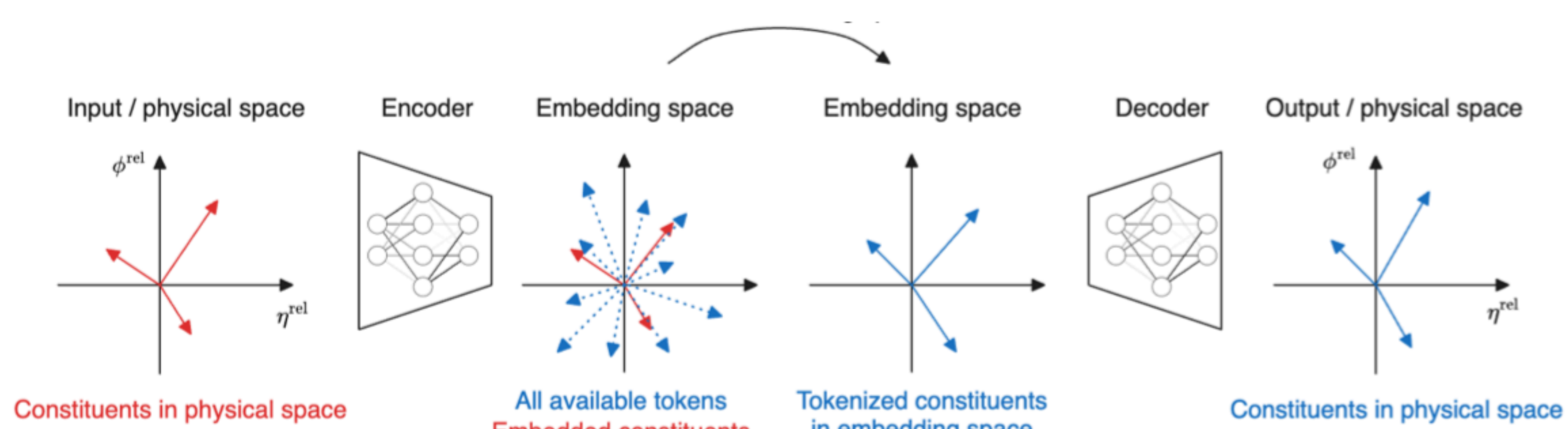


Masked particle modelling, *M. Leigh et al.* ML4JETS

Finetuning foundation models for analysis optimisation,
M. Vigl et al. ML4JETS



What is the best way to represent HEP data for input to a foundation model?



Credit A. Hallin

Simulating particle jets

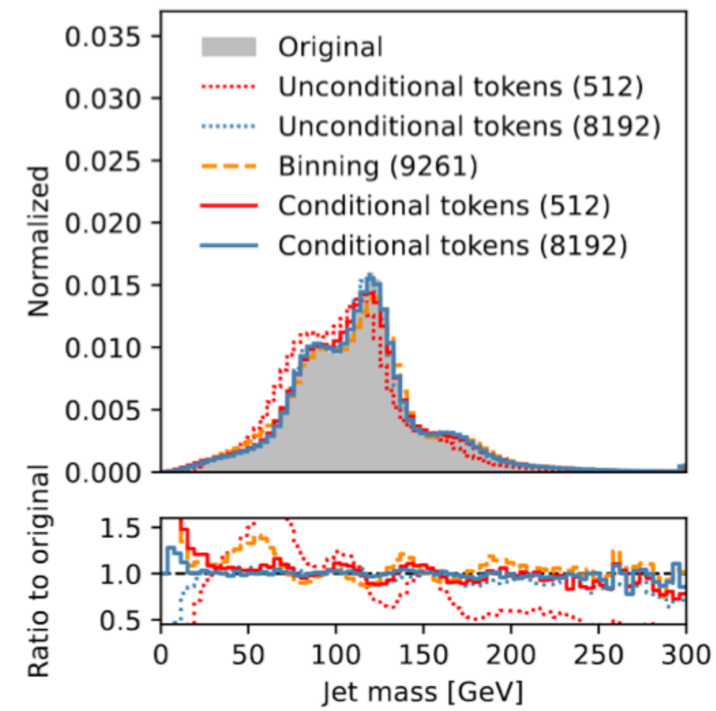
Anna Hallin et al. arxiv: 2403.05618

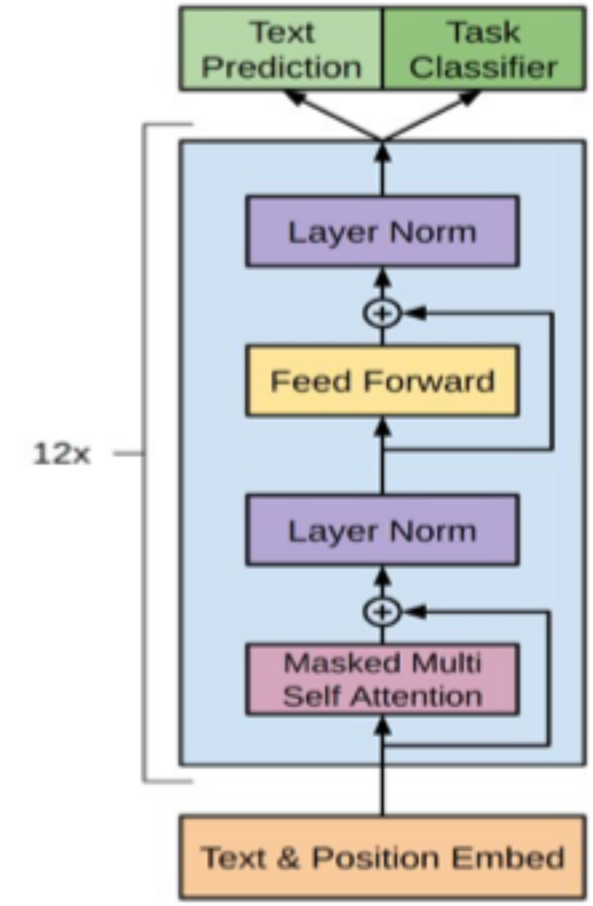
Particle and jets are interpreted as words and sentences.

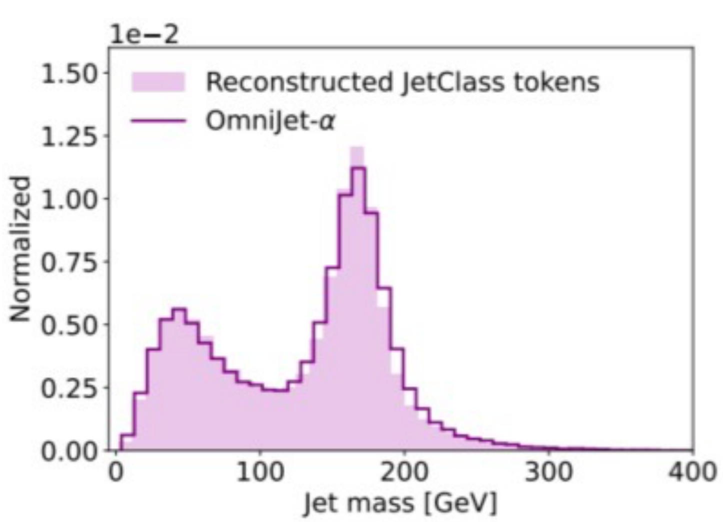
Use transformers as NLP to perform jet classification and generation

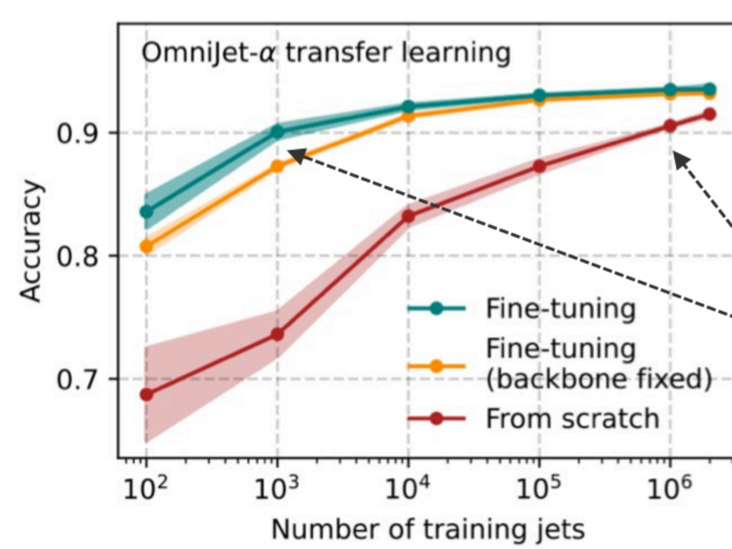
Transformers expect tokens

What happens to the continuous physics information ?









Pre-trained model requires only 1000 training events to reach the same accuracy level that the "from scratch" model reaches with 1M events

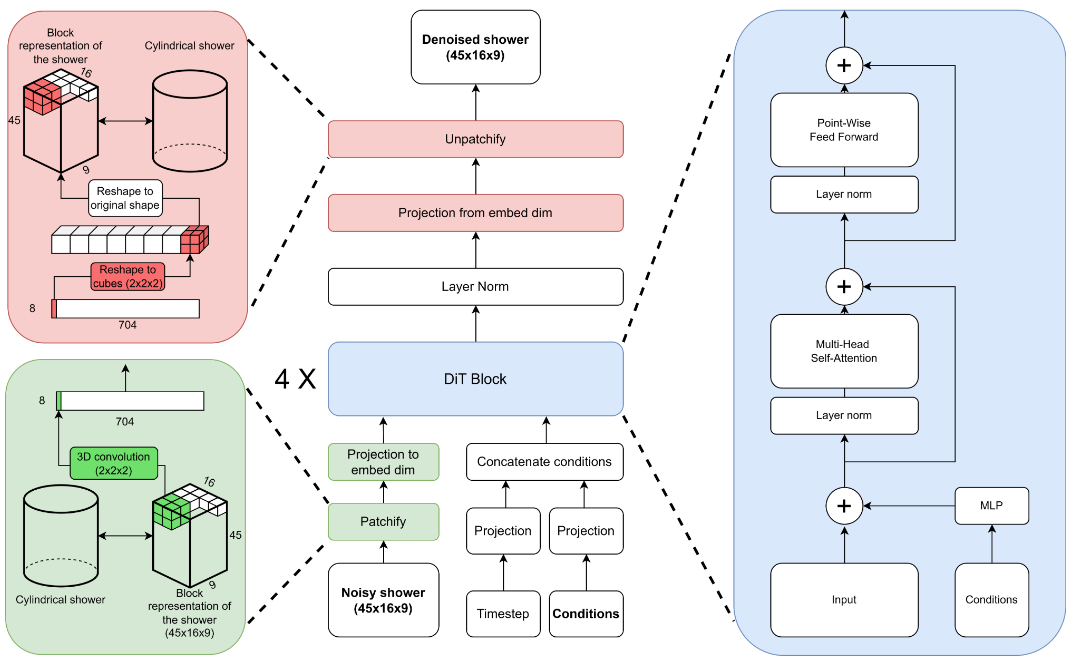
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

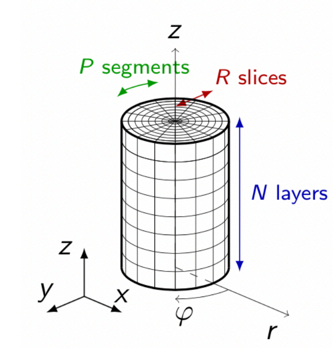
Diffusion Transformers for detector simulation

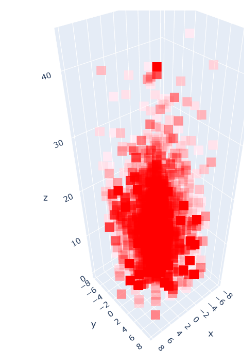
Renato Cardoso, et al.
CHEP 2023

A **generalized architecture** that works with any type of data

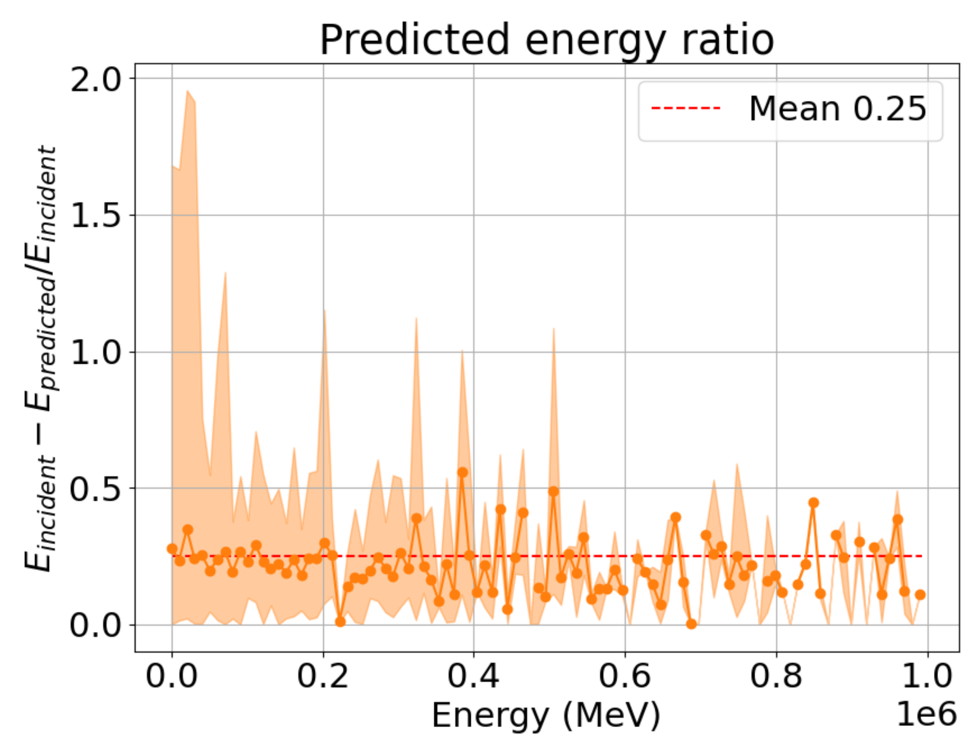
It models long-range dependencies via attention mechanism

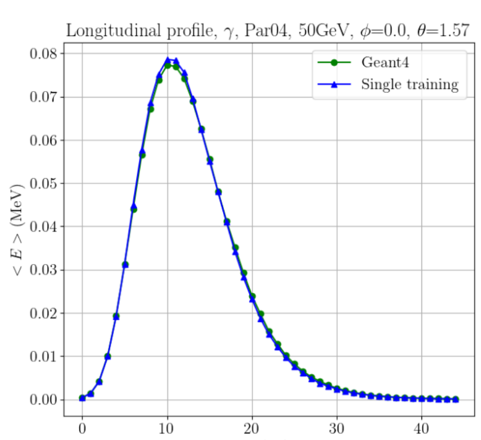


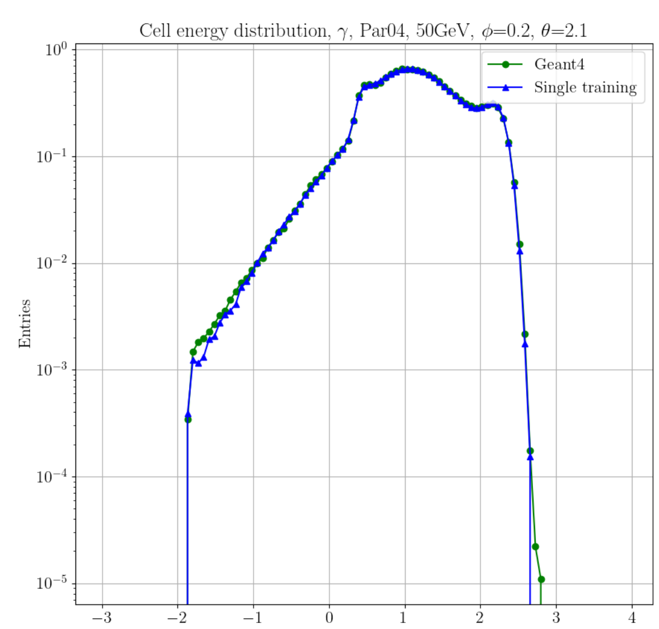




Adaptability to Multi-Tasking:
From image generation to regression

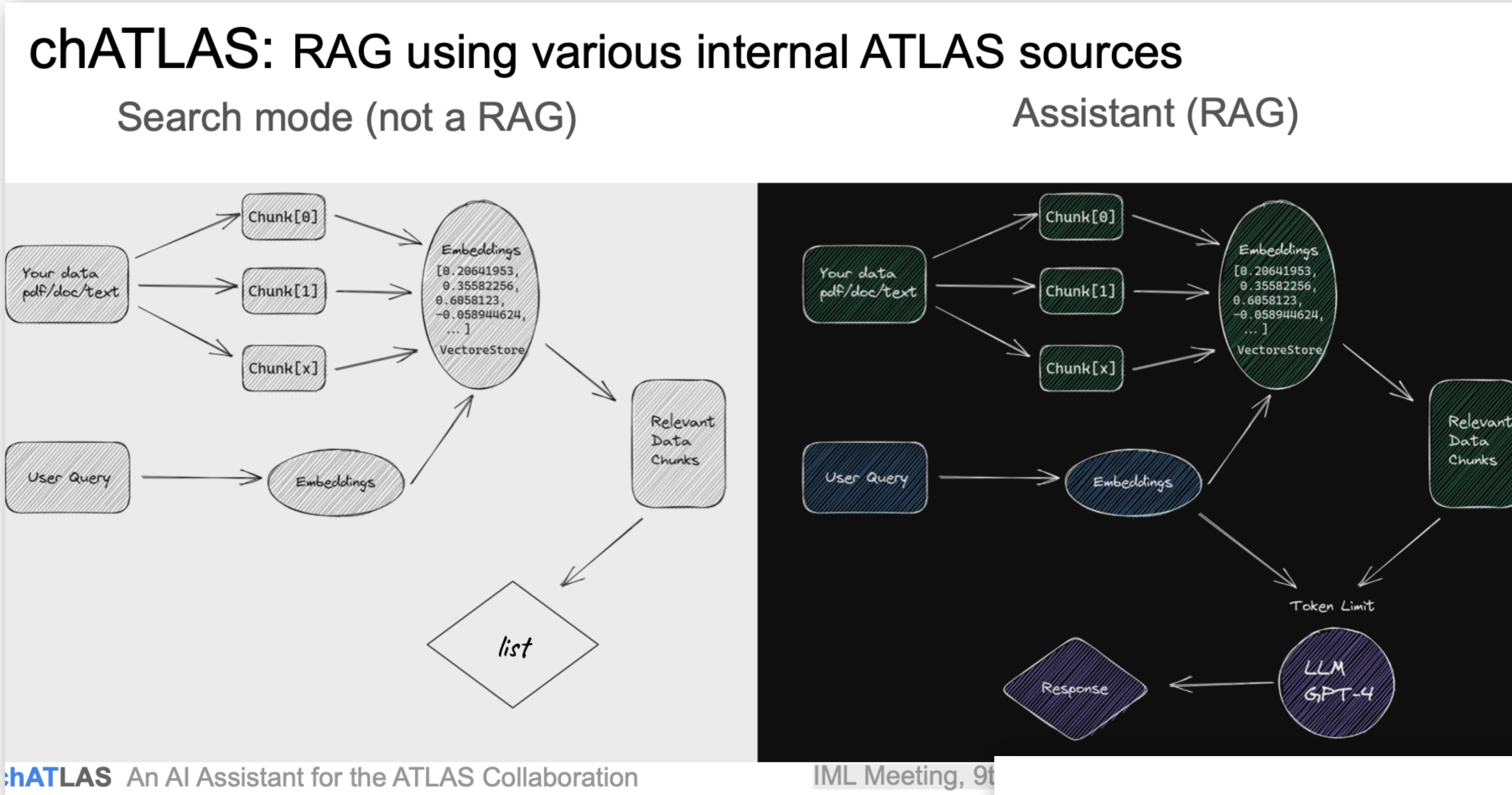






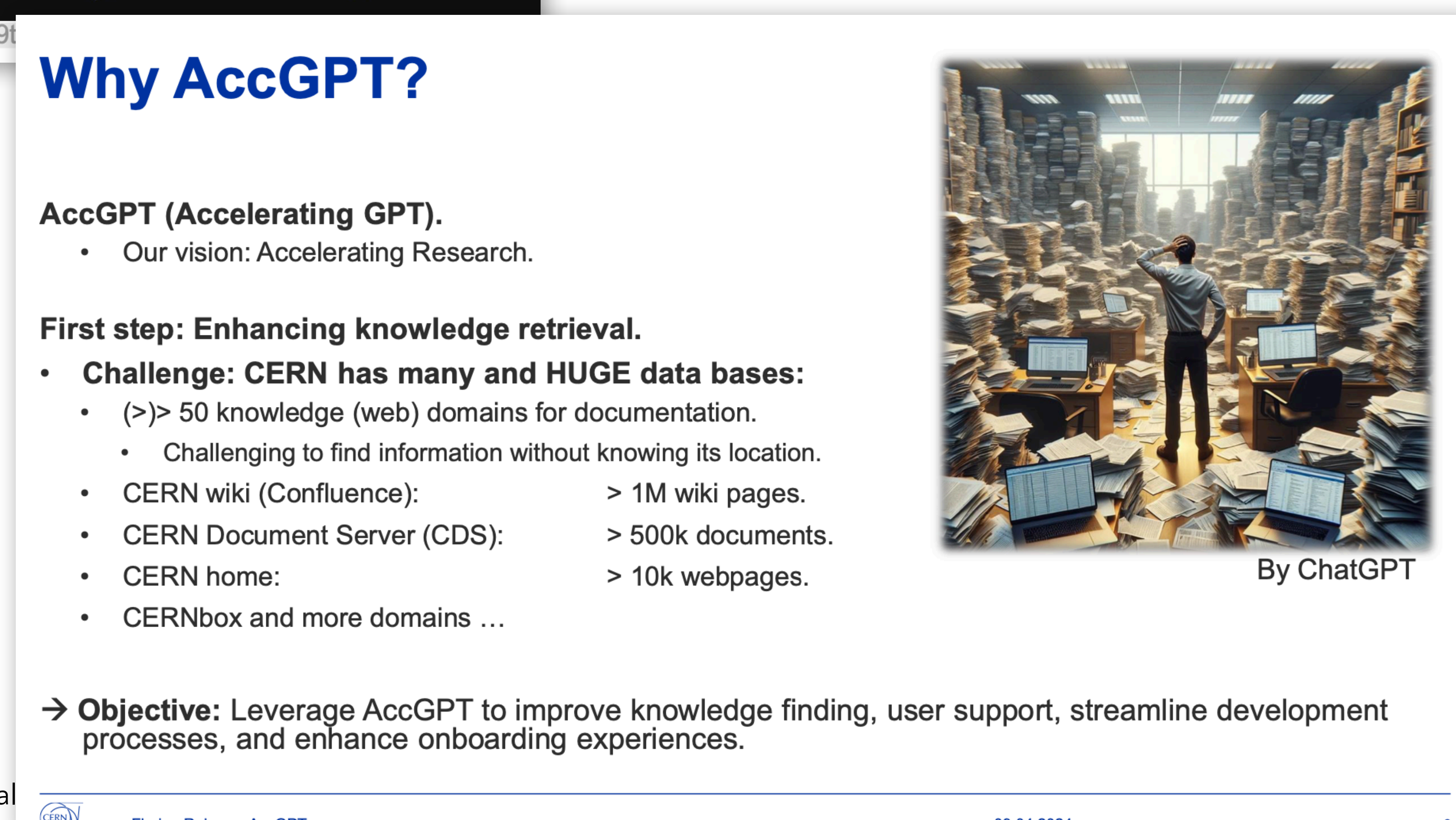
aria.luise@cern.ch |

LLMs as scientific Assistants



D. T. Murnane,
IML: https://indico.cern.ch/event/1395528/

F. Rehm,
IML: https://indico.cern.ch/event/1395528/



Resources

Comprehensive overview of PTMs (2023):

https://arxiv.org/pdf/2302.09419

How to stay up to date?

- https://alphasignal.ai/
- https://www.deeplearning.ai/the-batch/

Wanna learn more about foundation models?

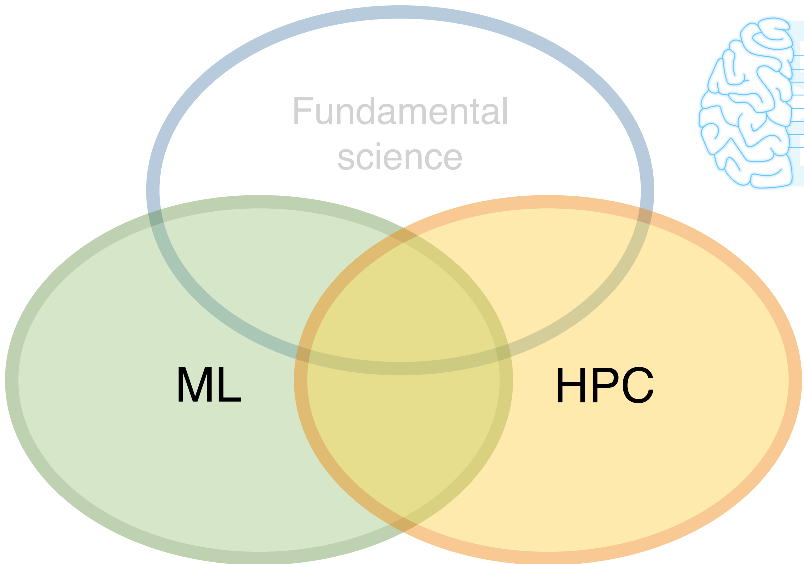
- Coursera - Introduction to foundation models
- https://crfm.stanford.edu/

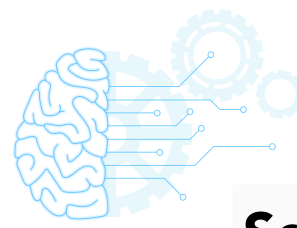
Center for Research on Foundation Models

Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Backup

Future challenges





Scaling:

- **Efficiently scaling distributed training to larger models**
- Develop the software infrastructure and model architecture suitable for such big models

Accessibility:

- **Deployment of the models on the cloud**
- we need an integration of the HPC centers to provide **seamless access** and data movement in the background (example: Google Cloud)

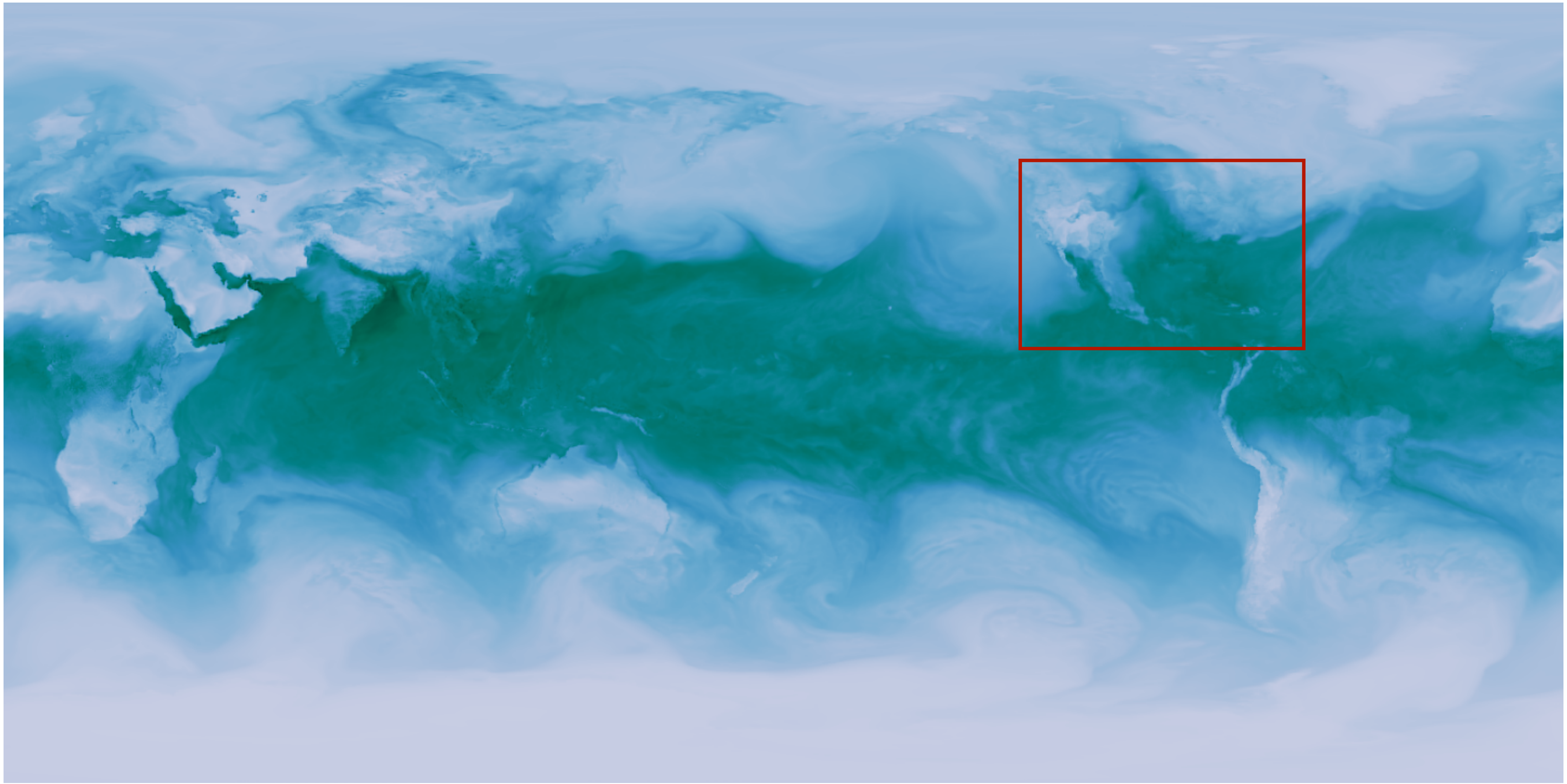
Maintenance:

- **How to integrate new incoming data**
- How to **expand** to new fields/variables without fully retraining the model each time?

From Martin Schultz - ESA-ECMWF workshop 2024

Results: Target - ERA5

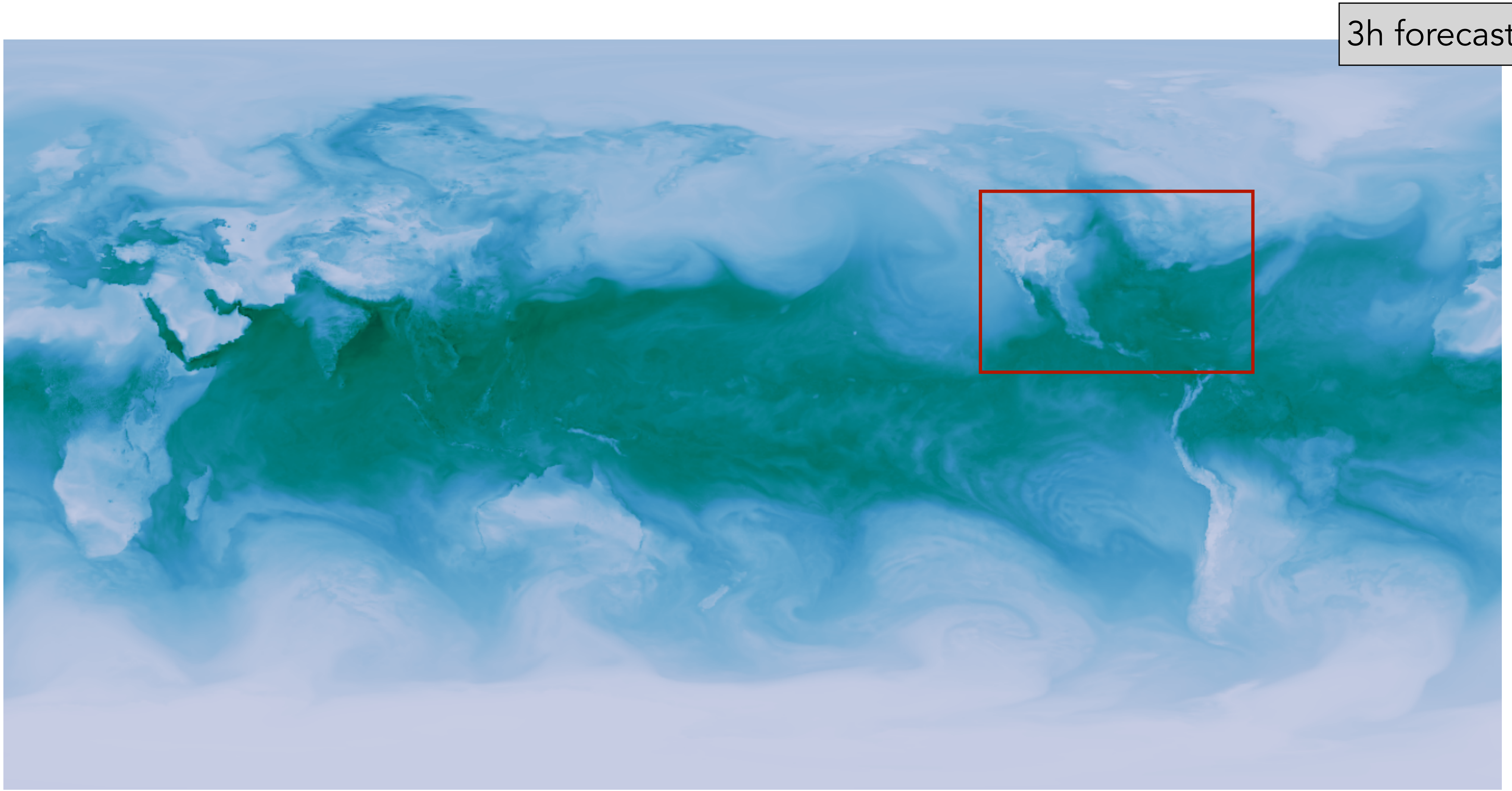
specific humidity, June 15th 2018 13:00 UTC



Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

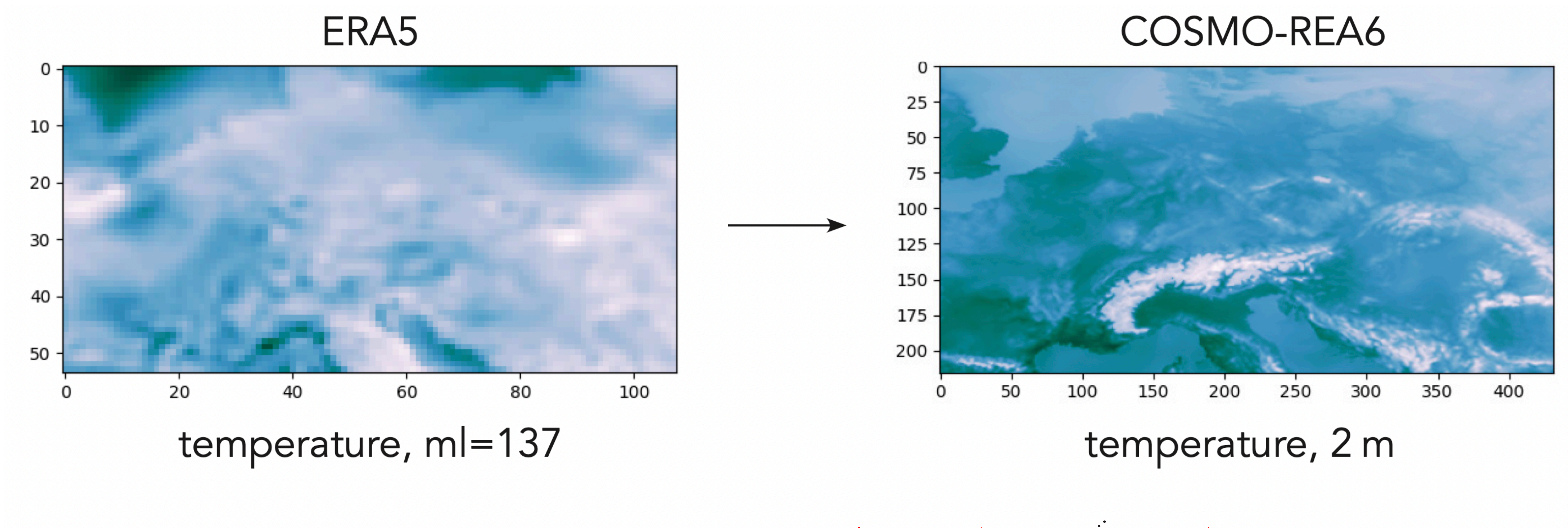
Results: Prediction - AtmoRep

specific humidity, June 15th 2018 13:00 UTC

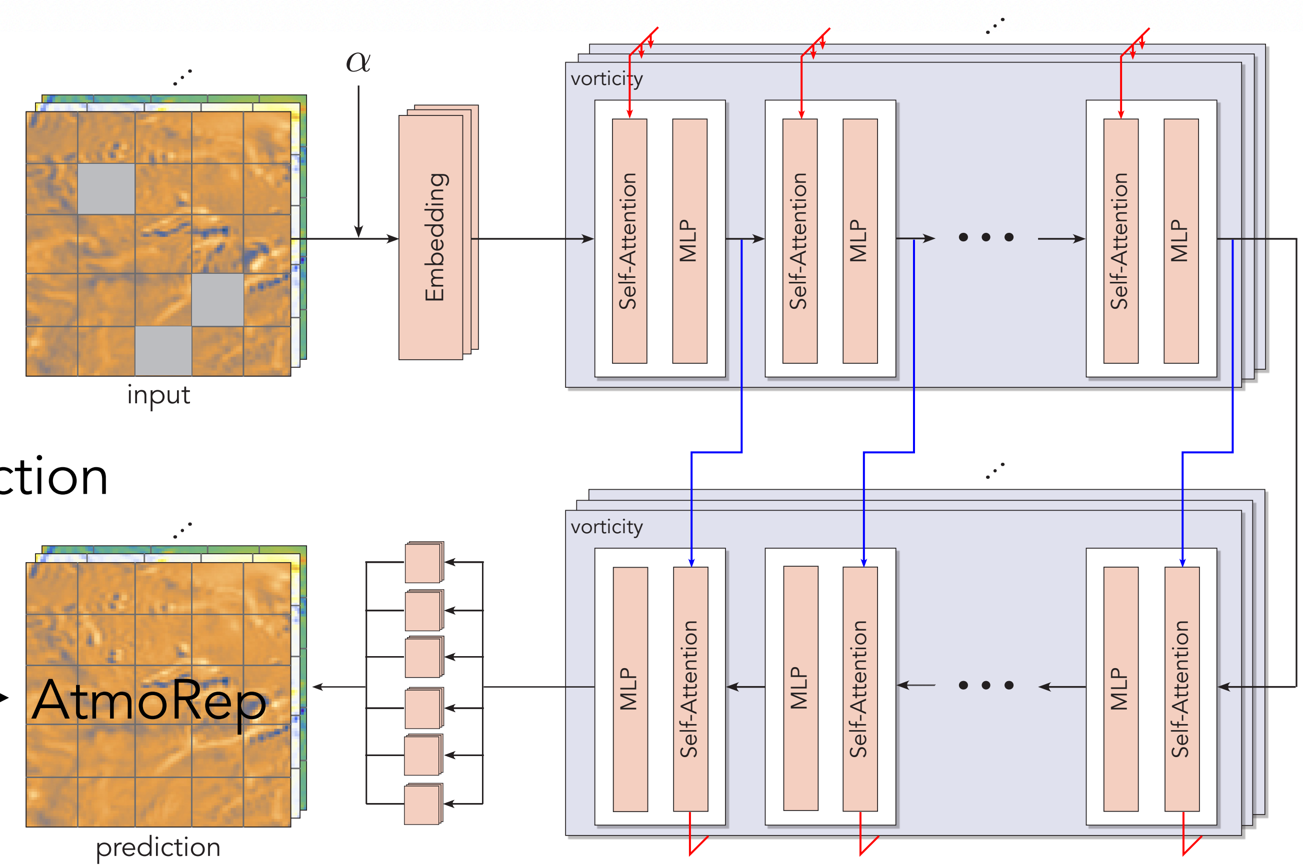


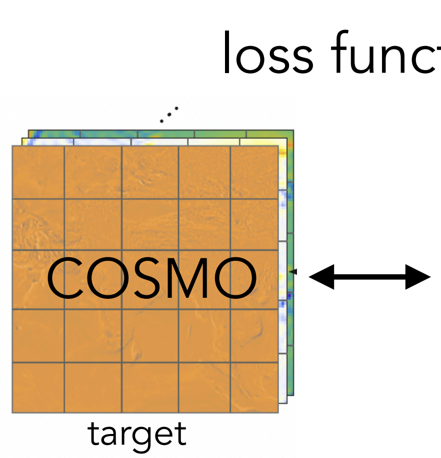
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Downscaling



Use COSMO REA6 data as ***target*** for the ***loss minimisation***



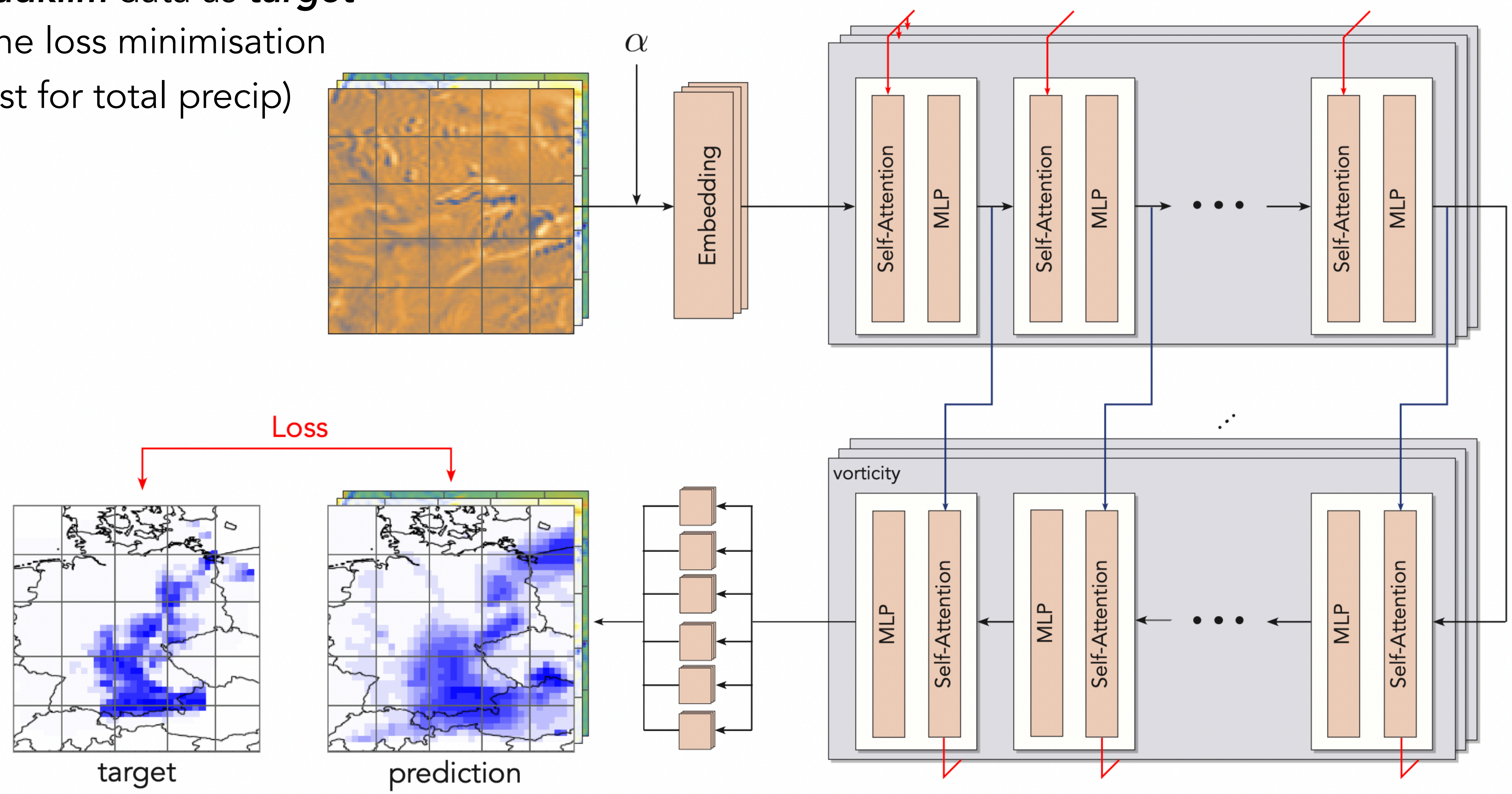


Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Bias corrections

Precipitation rates are known to be suboptimal in ERA5
Use RADKLIM radar data to fine-tune the precipitation rates in AtmoRep

Use ***Radklim*** data as ***target*** for the loss minimisation (just for total precip)



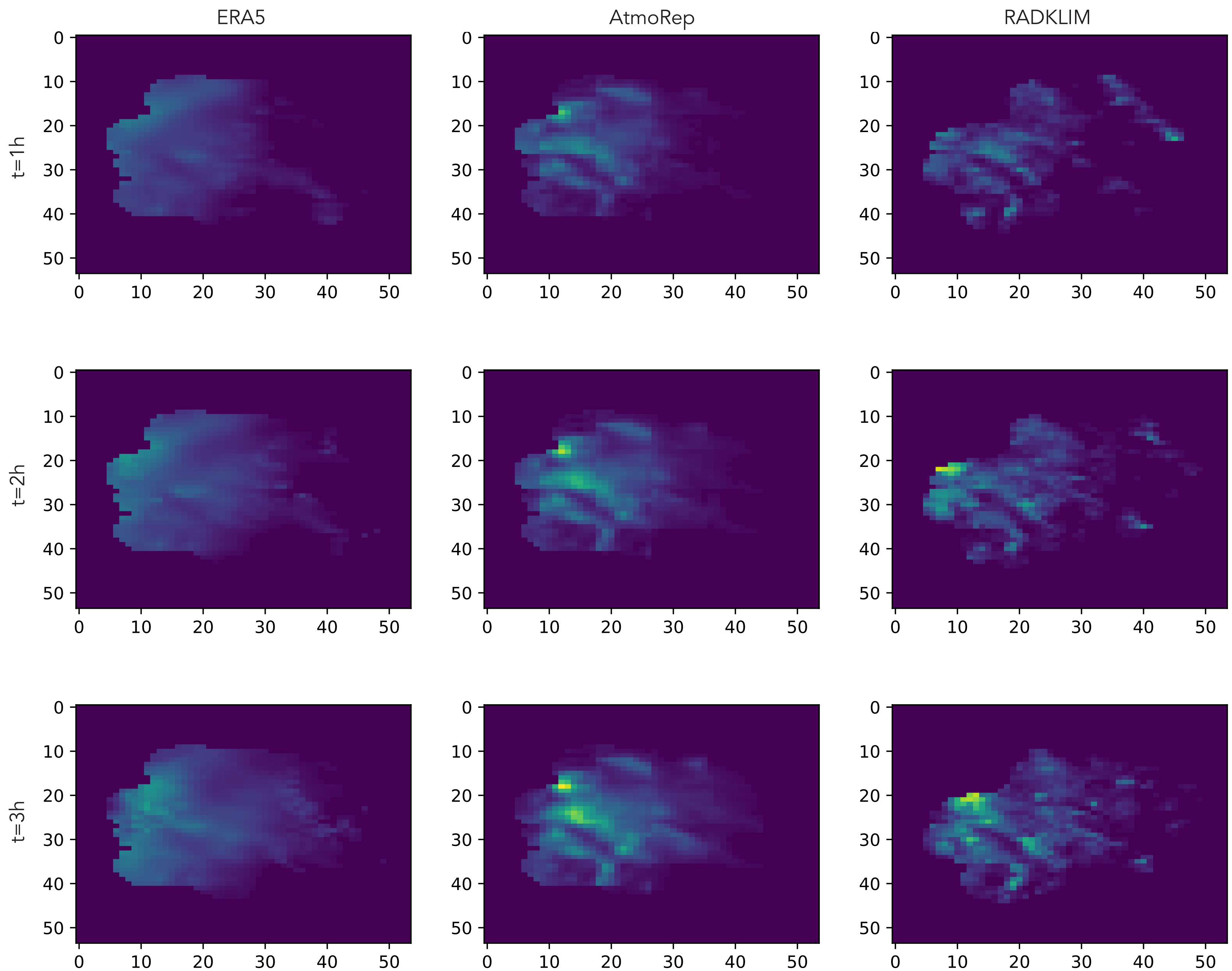
Radklim

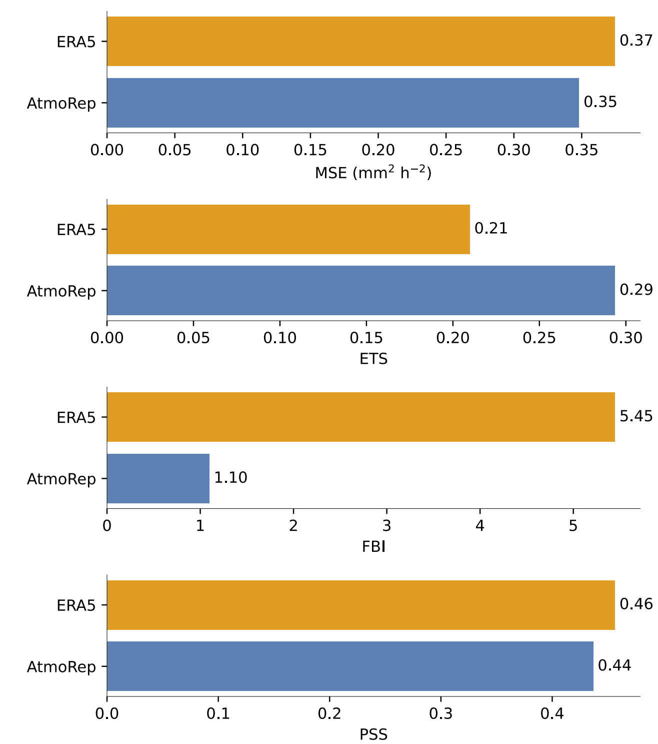
AtmoRep

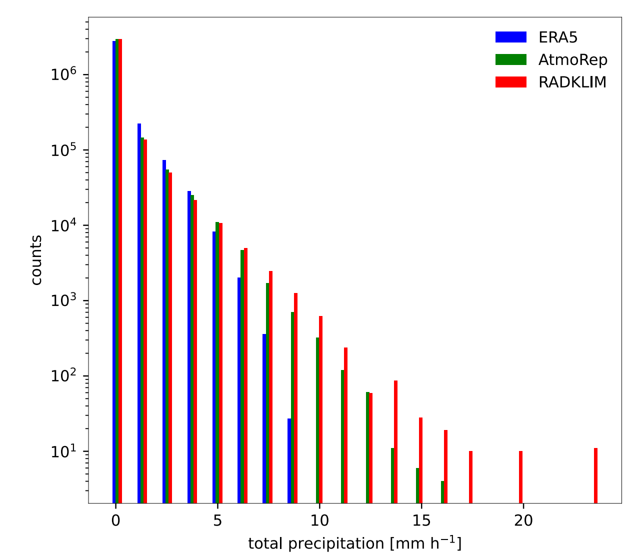
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Bias corrections: Results

Precipitation rates are known to be suboptimal in ERA5
Use RADKLIM radar data to fine-tune the precipitation rates in AtmoRep







Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch