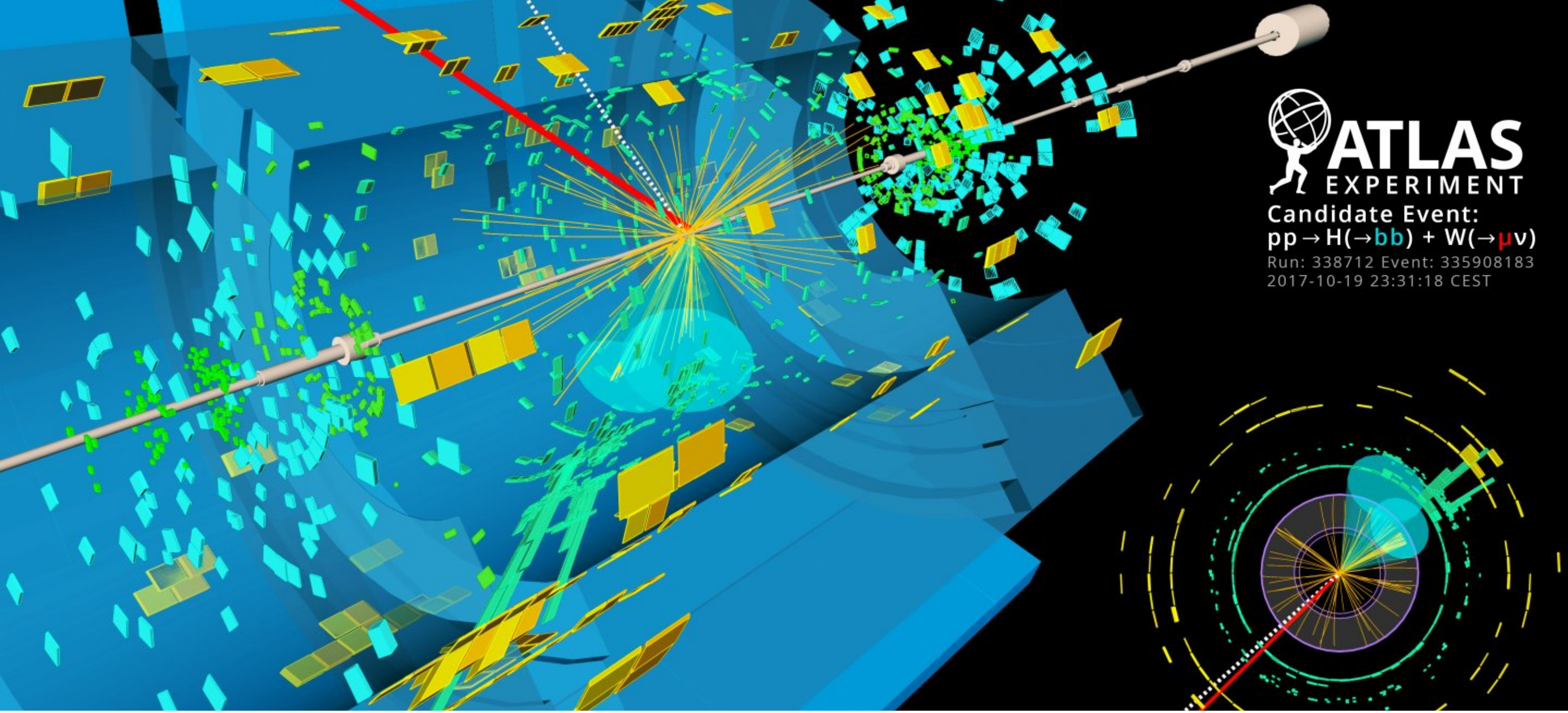


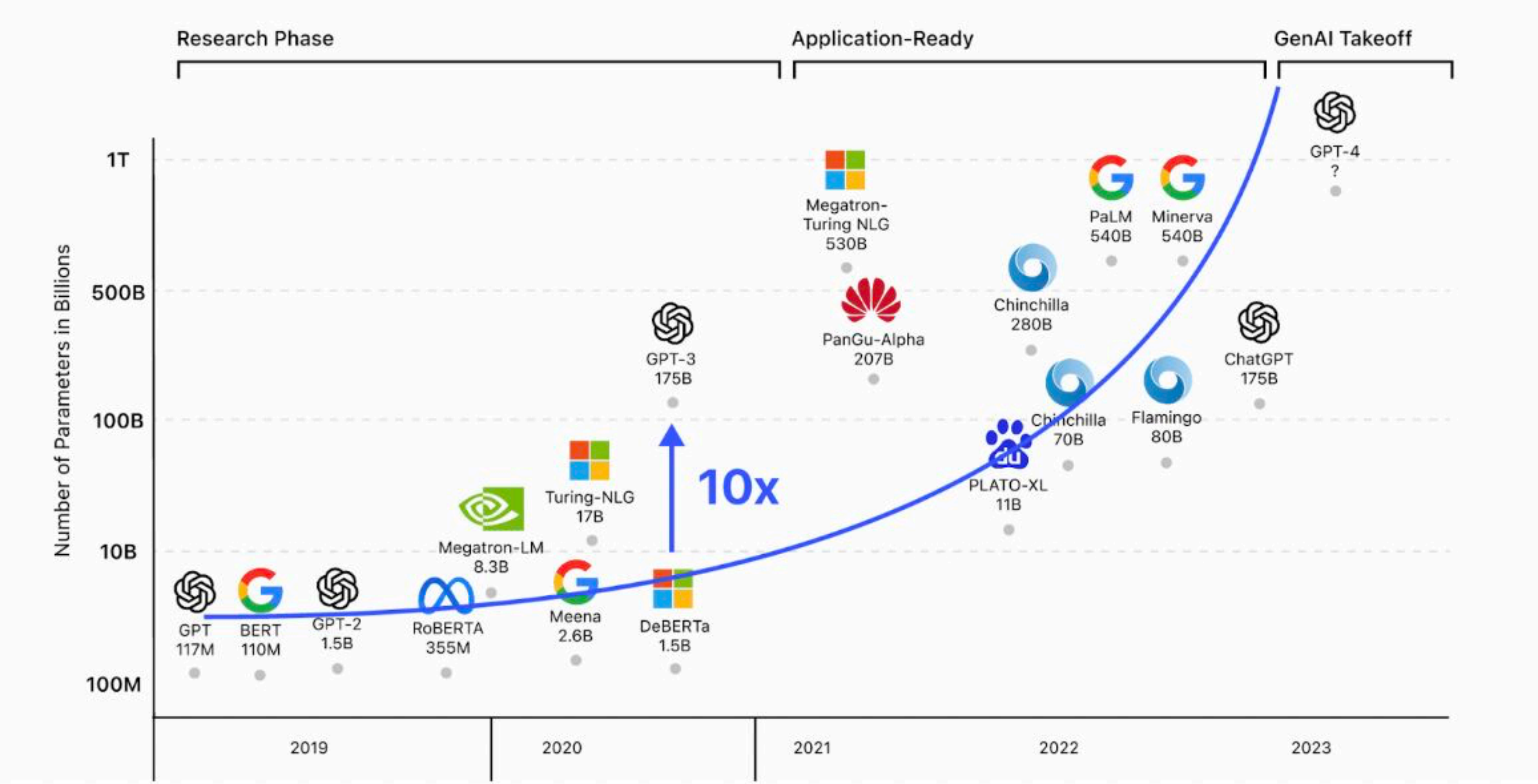
ML in Data Analysis: Foundation Models

Lecture 3

Sofia Vallecorsa | Ilaria Luise

Thematic CERN School of Computing on Machine Learning
17th October 2024

Then.. AI TakeOff….



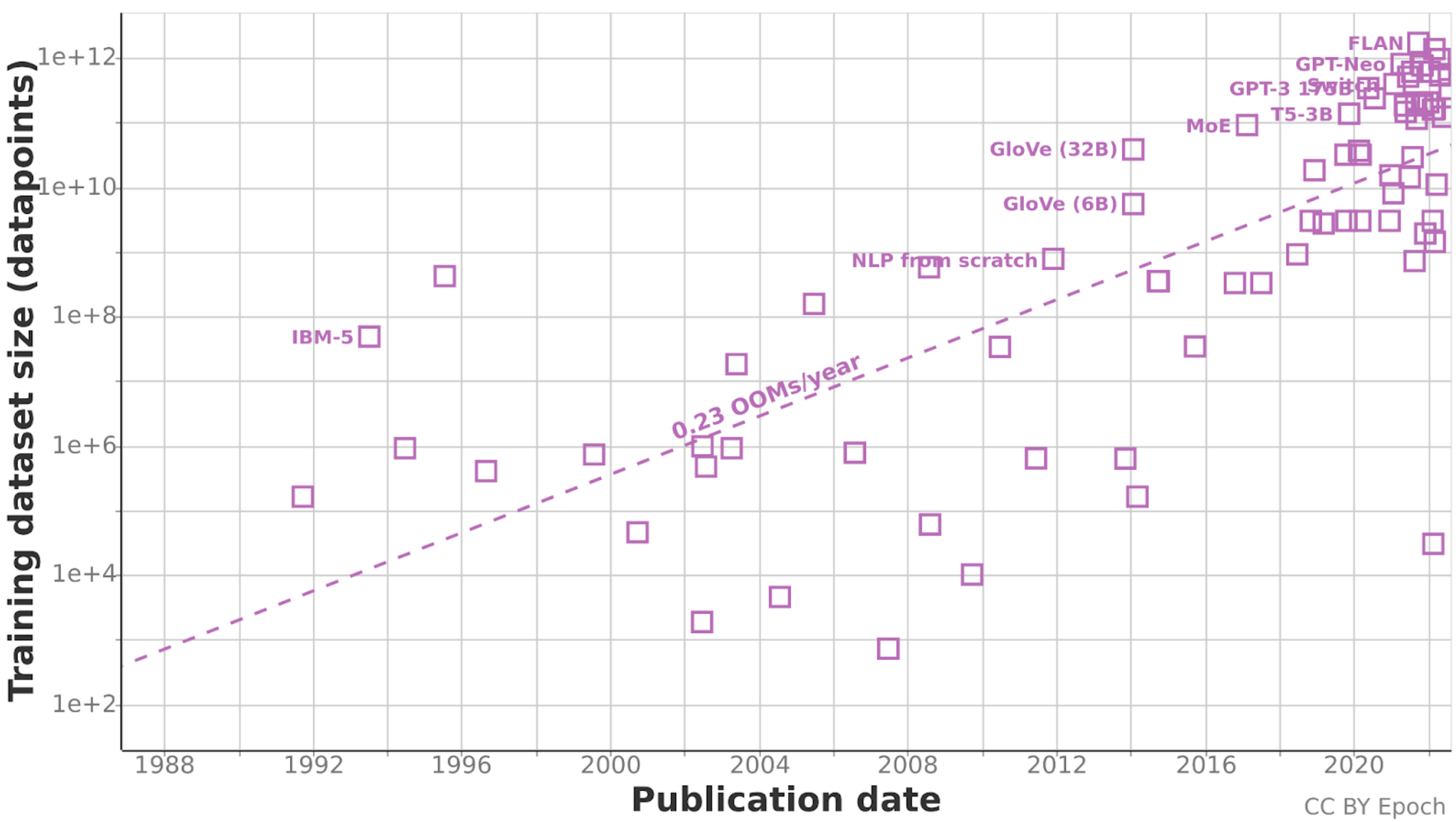
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

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!

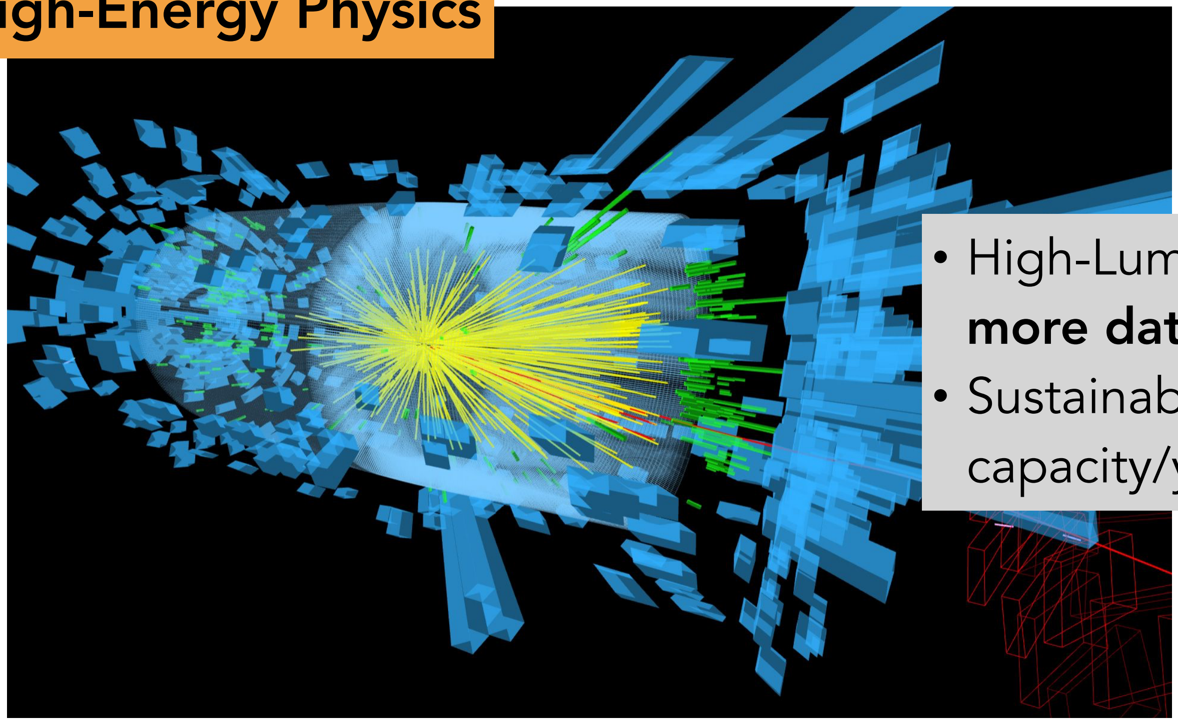


Can we use these tools for fully data-driven science?

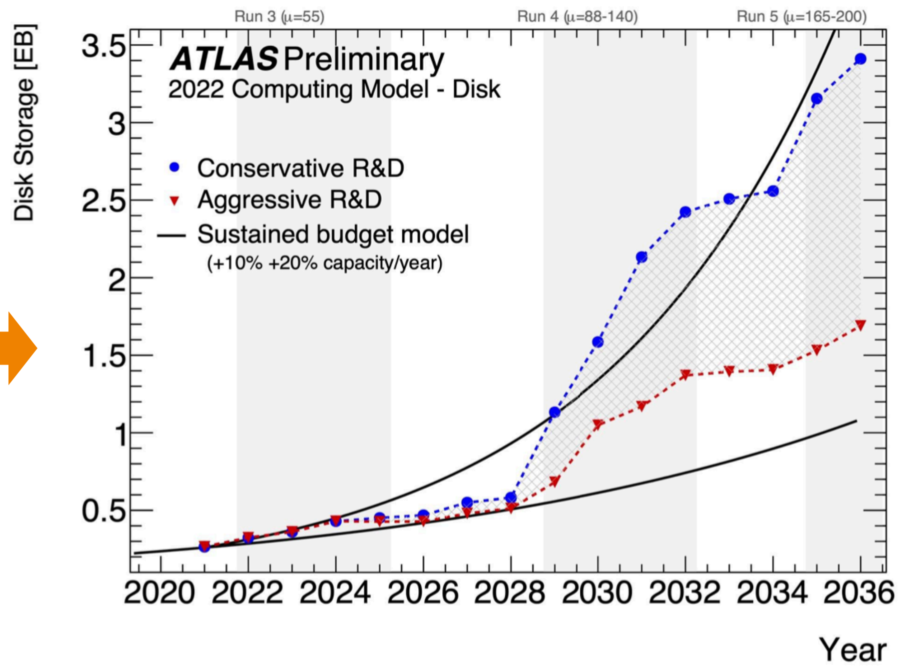
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

The future of observational data

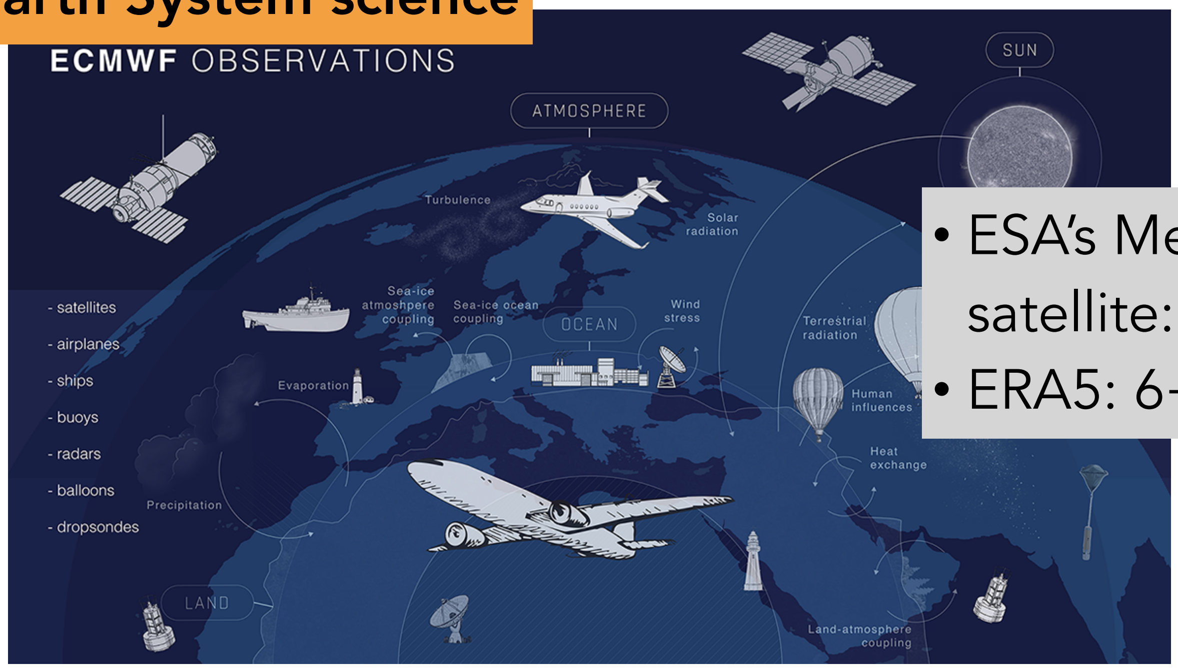
High-Energy Physics



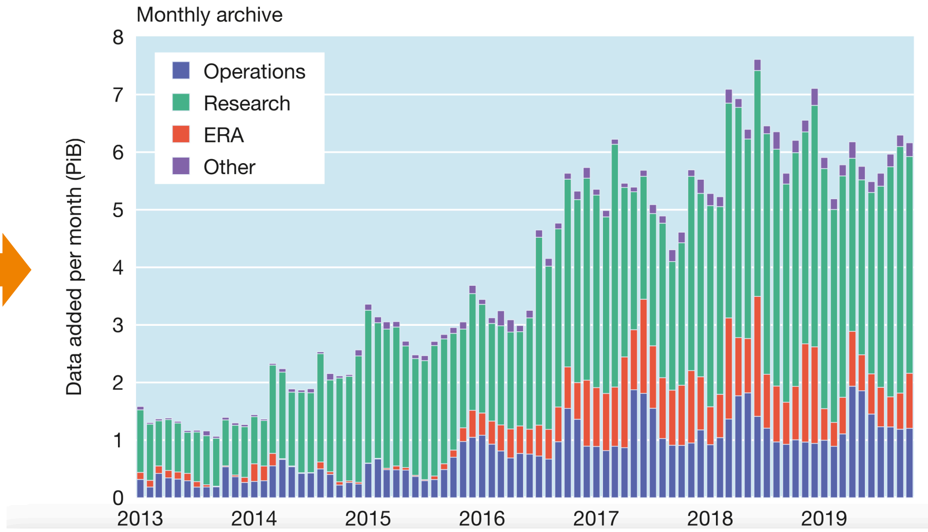
- High-Lumi LHC: **6 times more data by 2030**
- Sustainable model +20% capacity/year



Earth System science



- ESA's MetOp-SG satellite: **864 GB/day**
- ERA5: 6+ PB



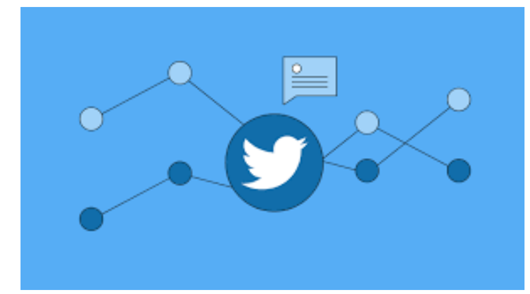
Need to find sustainable ways to store all these data

Ilaria Luise, CERN - ilaria.luise@cern.ch

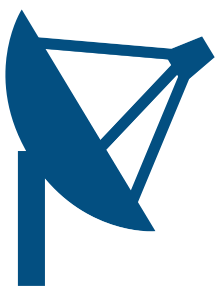
Multimodality

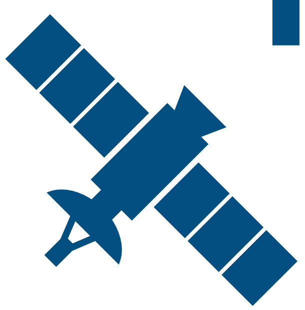
Data are getting **more and more multi-modal and the relationship between them is very complex to model** (and requires all kinds of approximations)

Social media

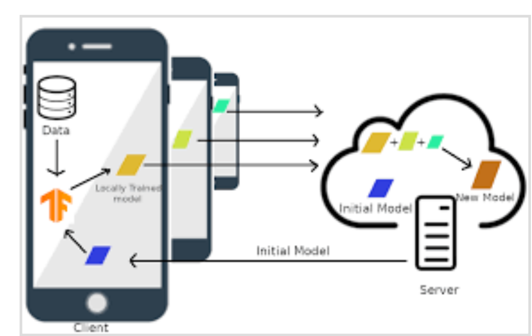


Economic growth
GDPs, birth rates…





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

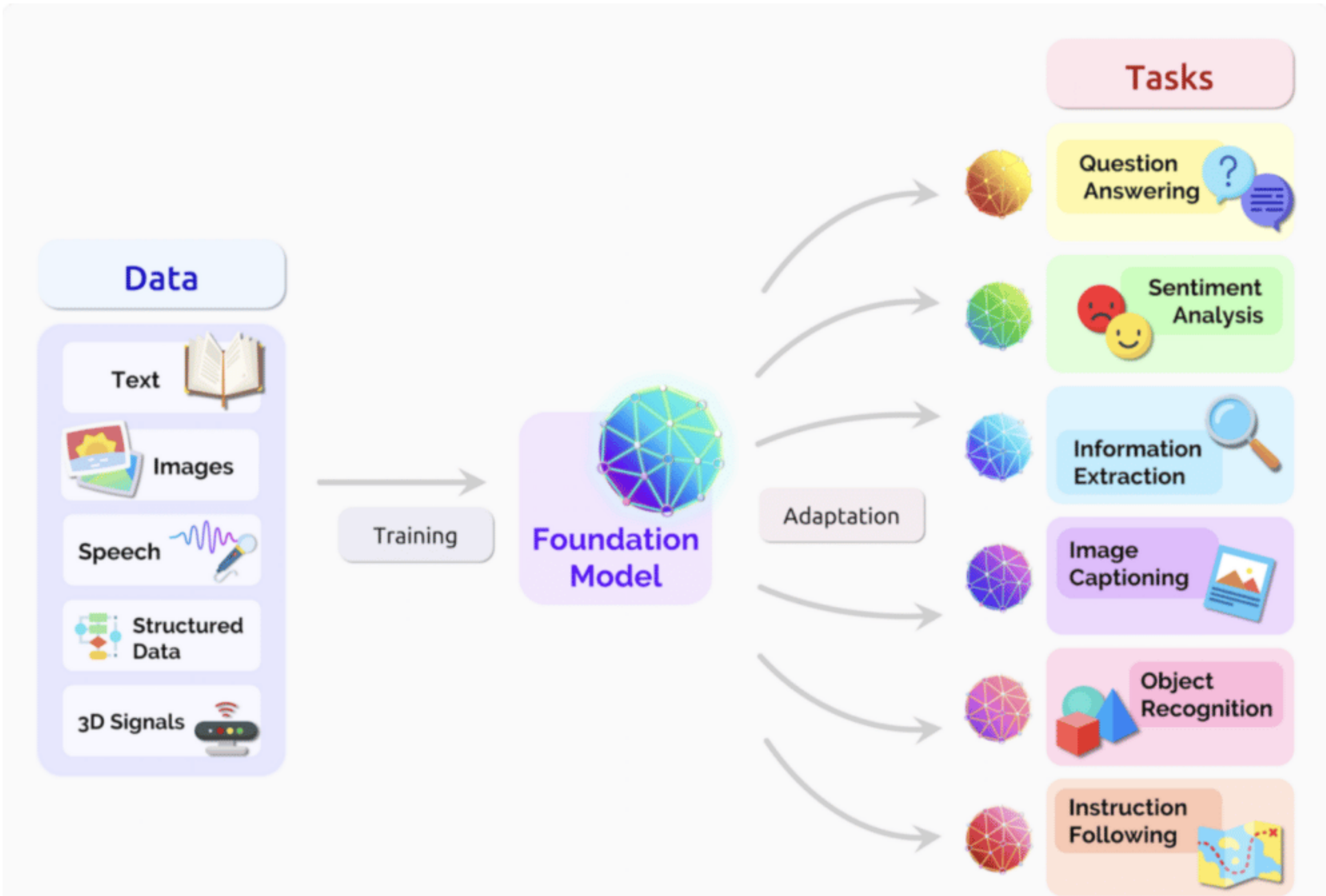


Data from distributed devices

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

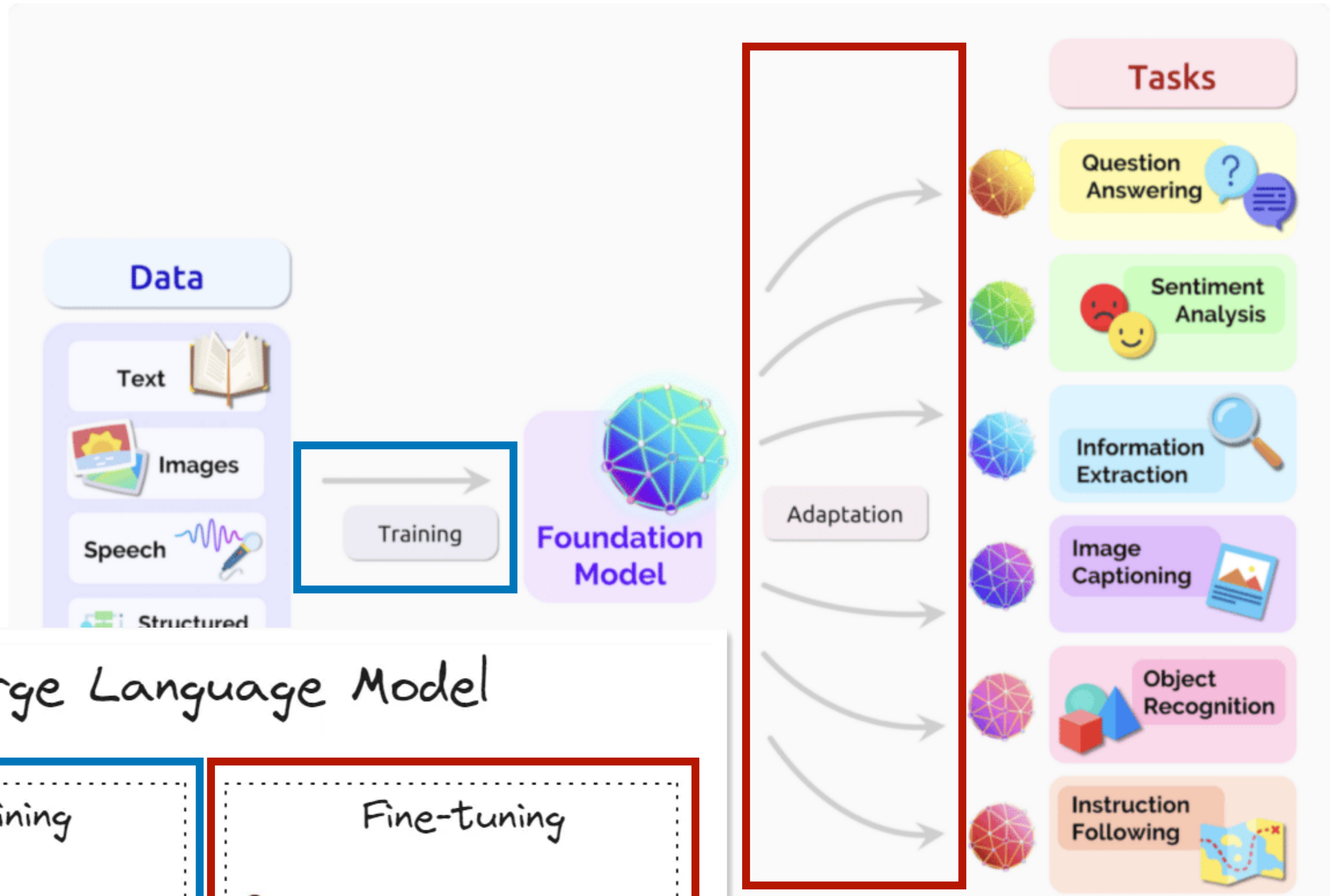
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

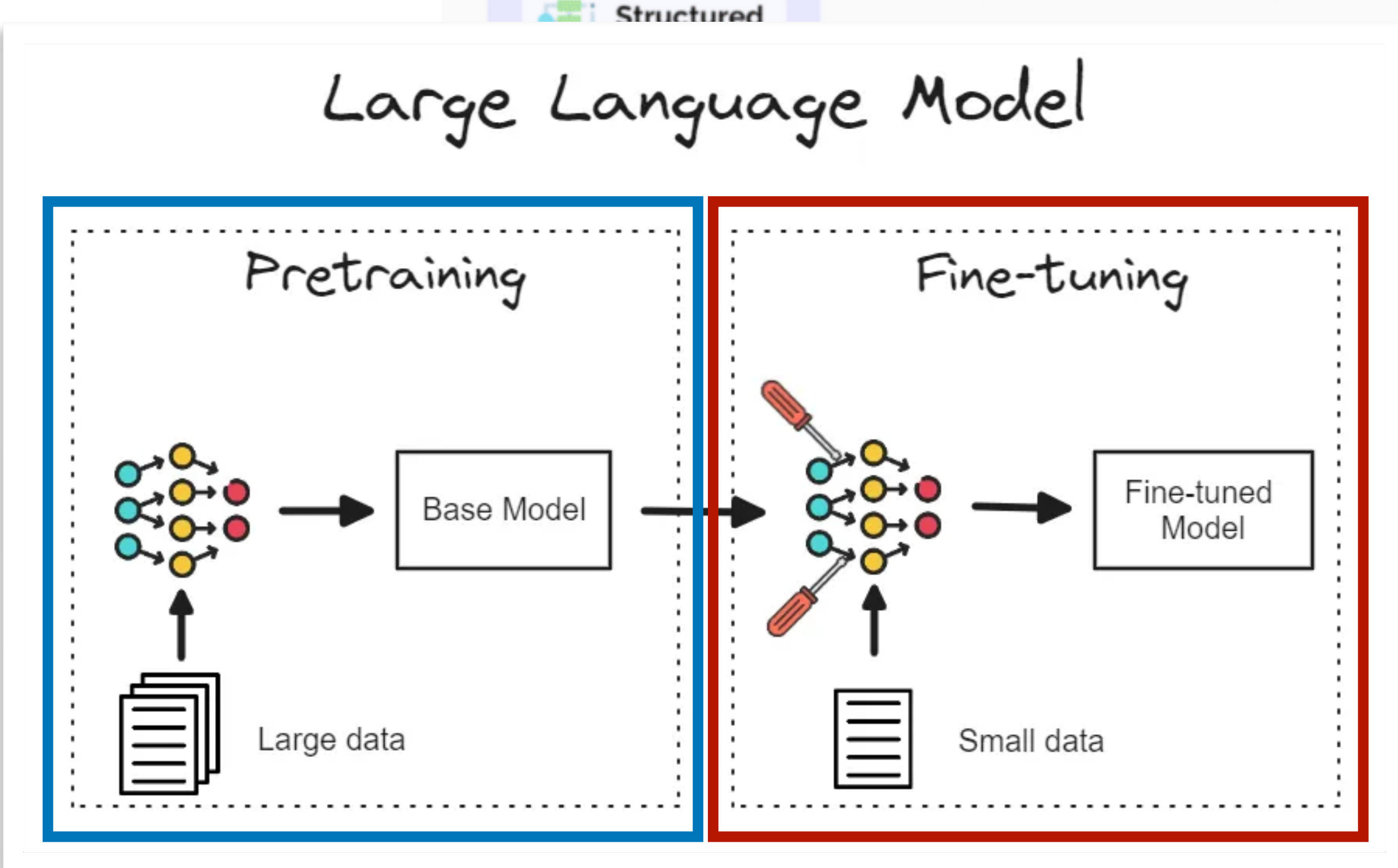
Introduction



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

Introduction

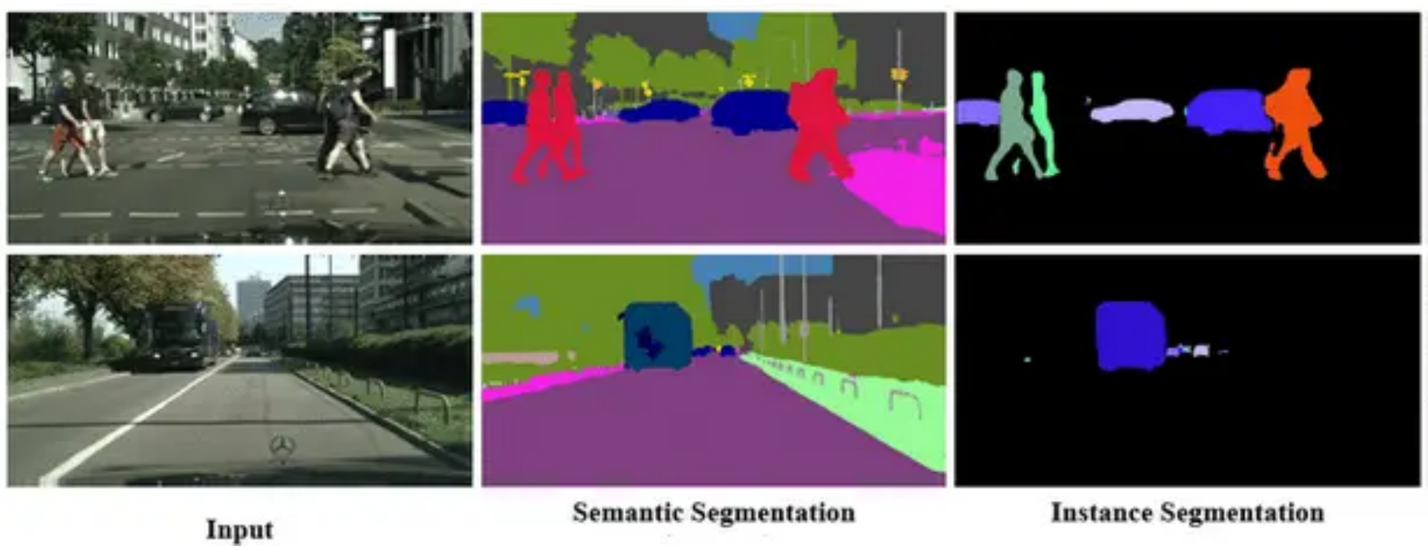




ise@cern.ch I sofia.vallecorsa@cern.ch

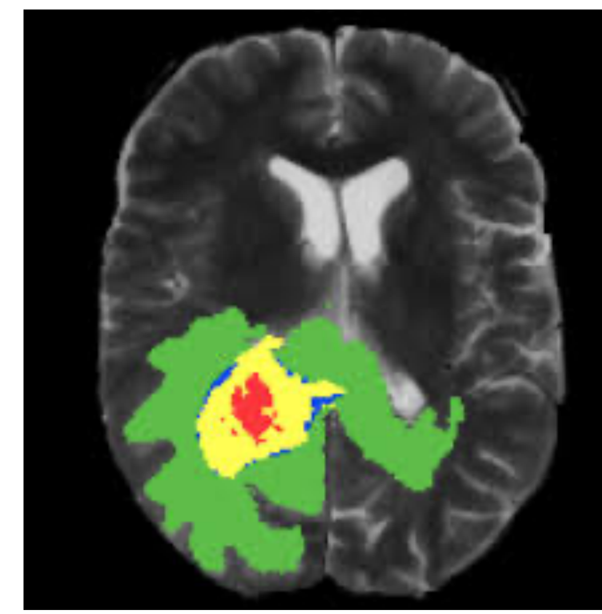
A concrete example

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



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

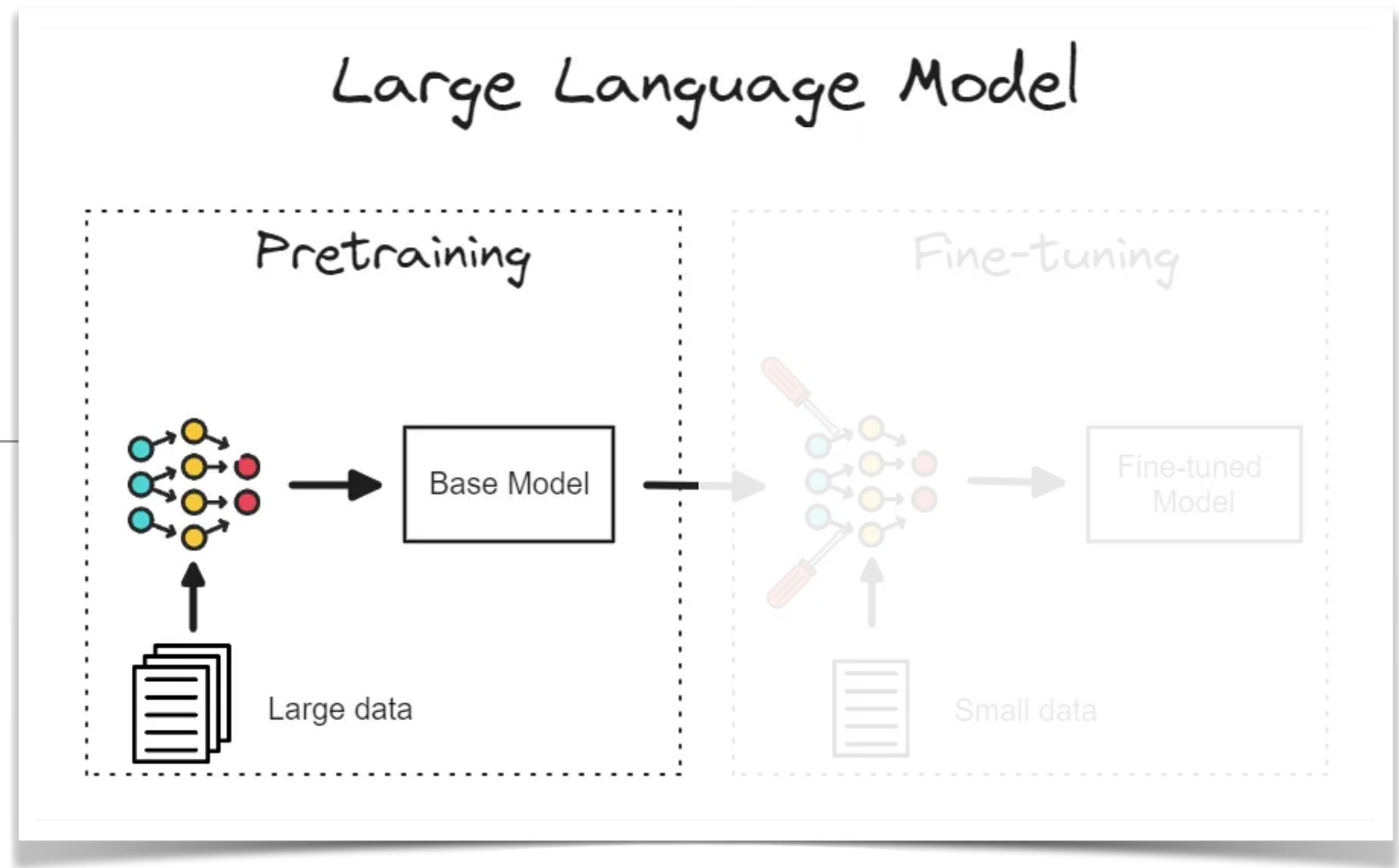
Pre-training: learn how to segment images (aka cluster pixels together into the different objects):
- Learn how to detect edges
- Learn how to cluster objects with the same e.g. colour …

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

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

Pre-training

Basic concepts



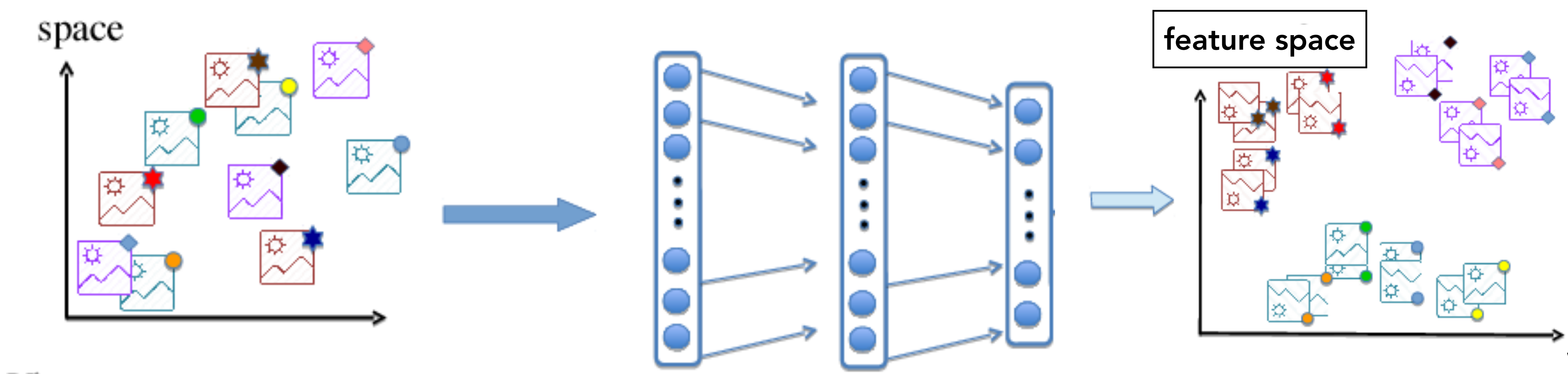
Main goal

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:

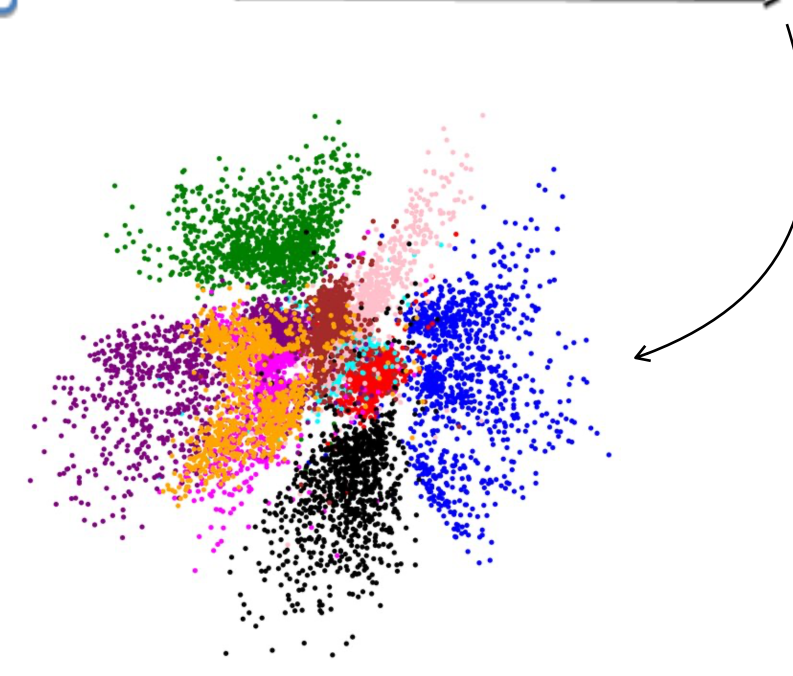
- Learn a **task-independent representation** of the data in the **feature space** of the neural network



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

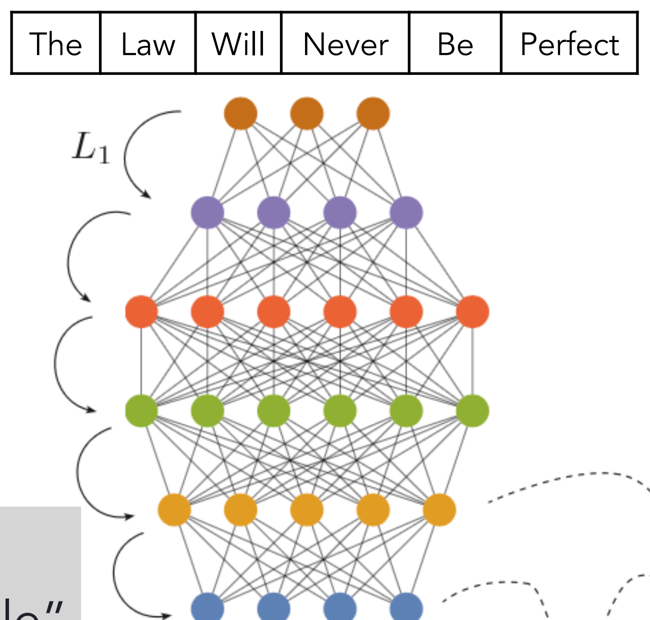
↓

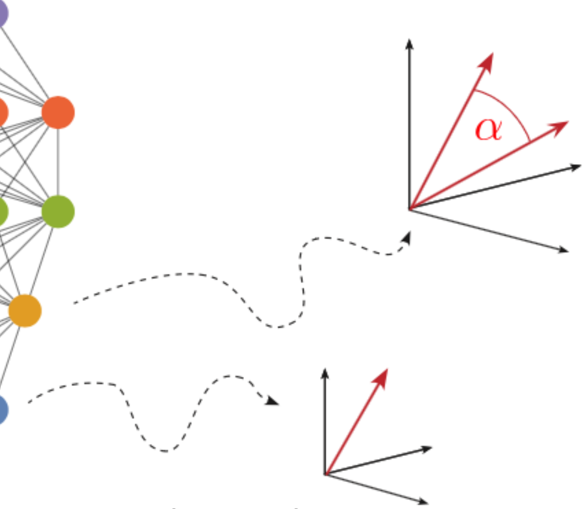
Re-shuffle
project the data at a different "angle"

↓

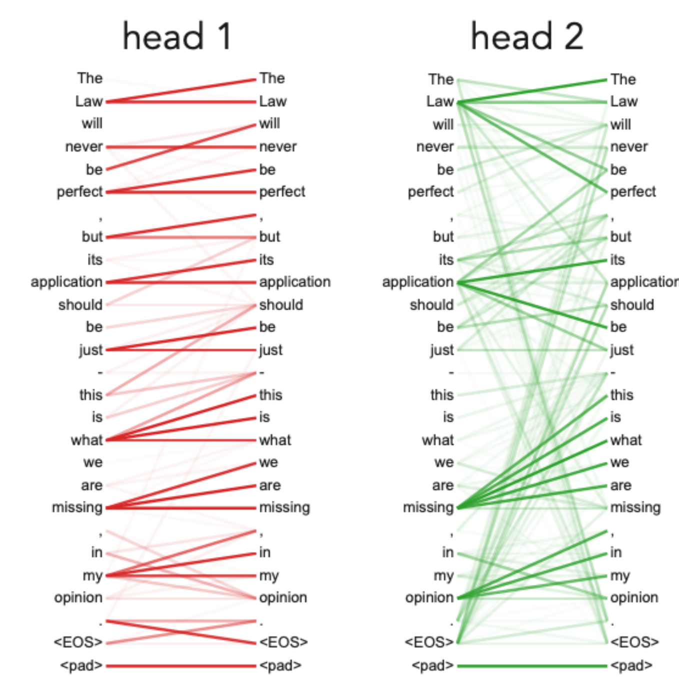
Loss calculation

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





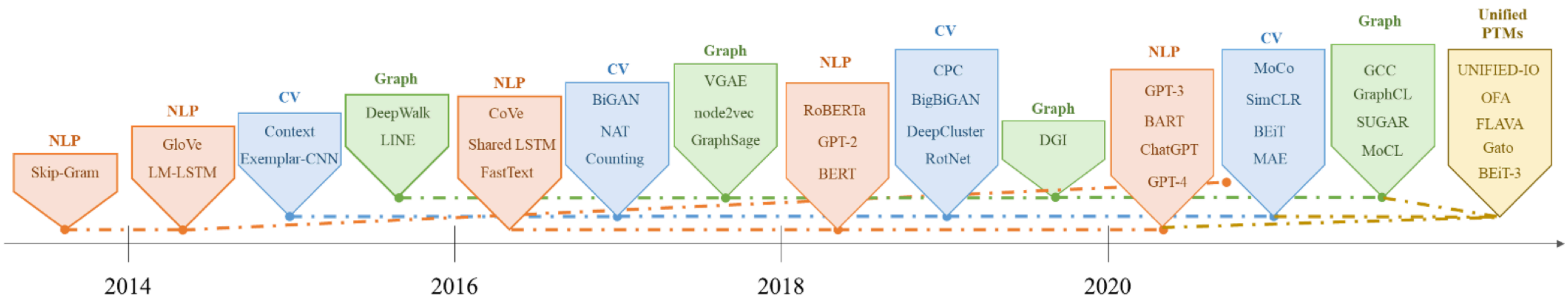
See each word as a vector in a complex space

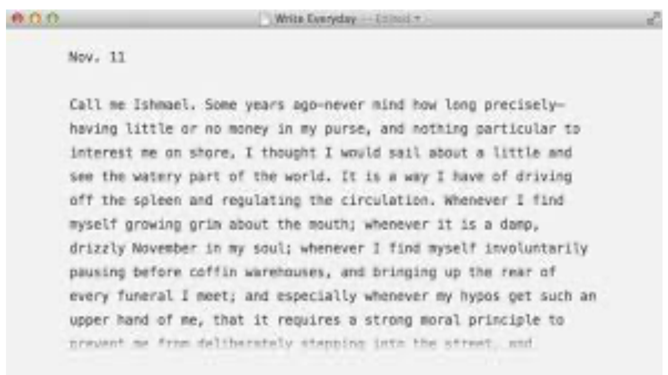


"Attention is all you Need" Vaswani 2017

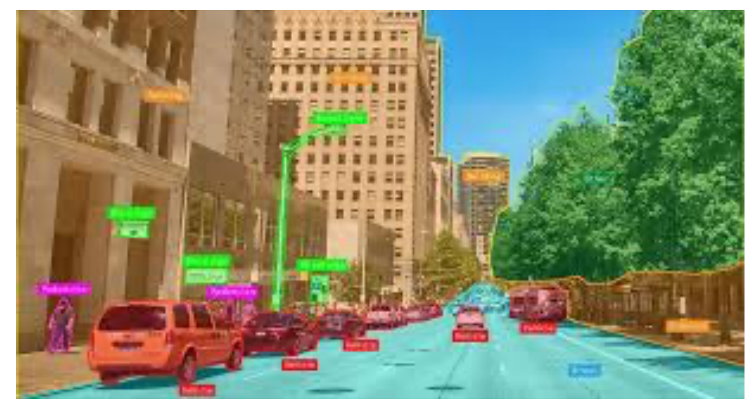
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Types of pre-trained models





NLP: Natural Language Processing



CV: Computer Vision

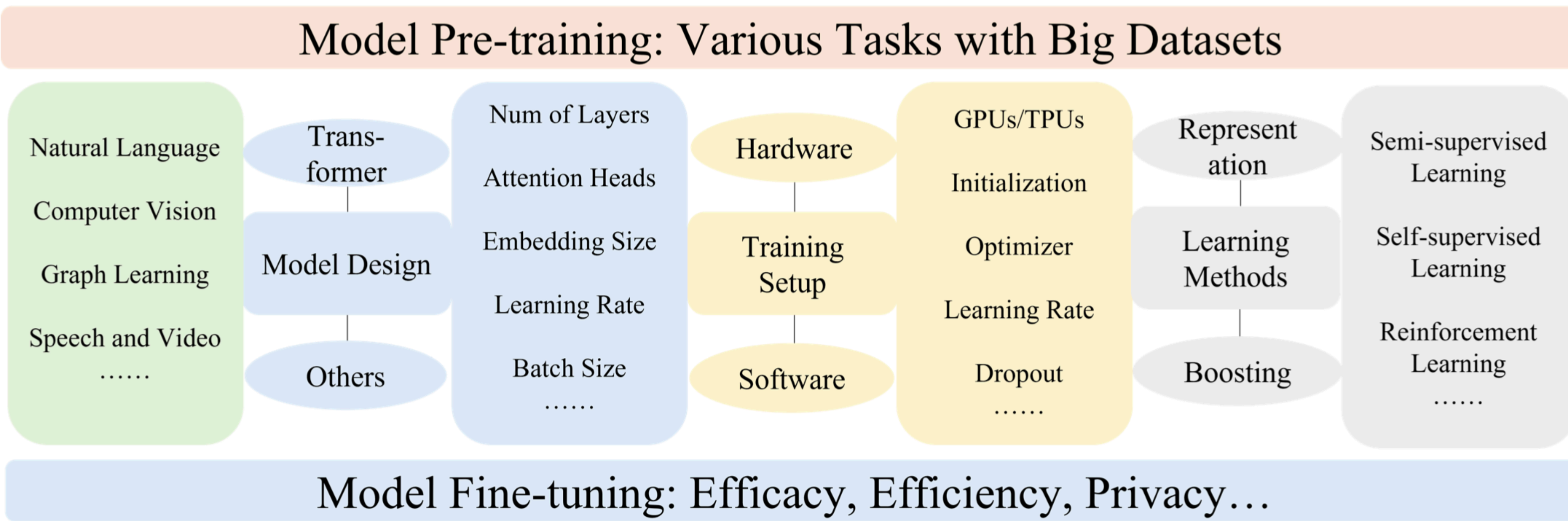
Graphs: Graph Learning (*not covered here*)

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

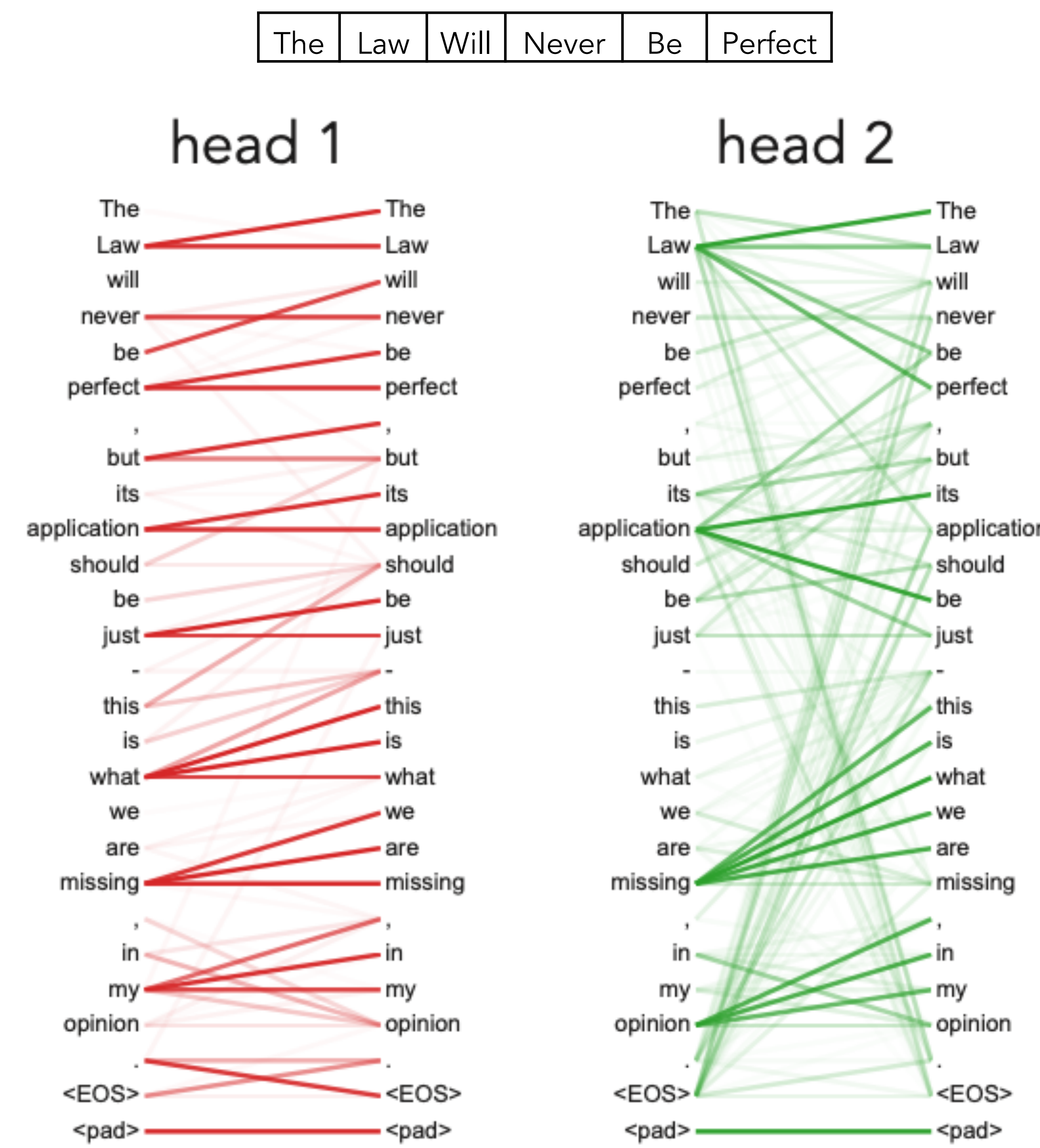
How do we pre-train?

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

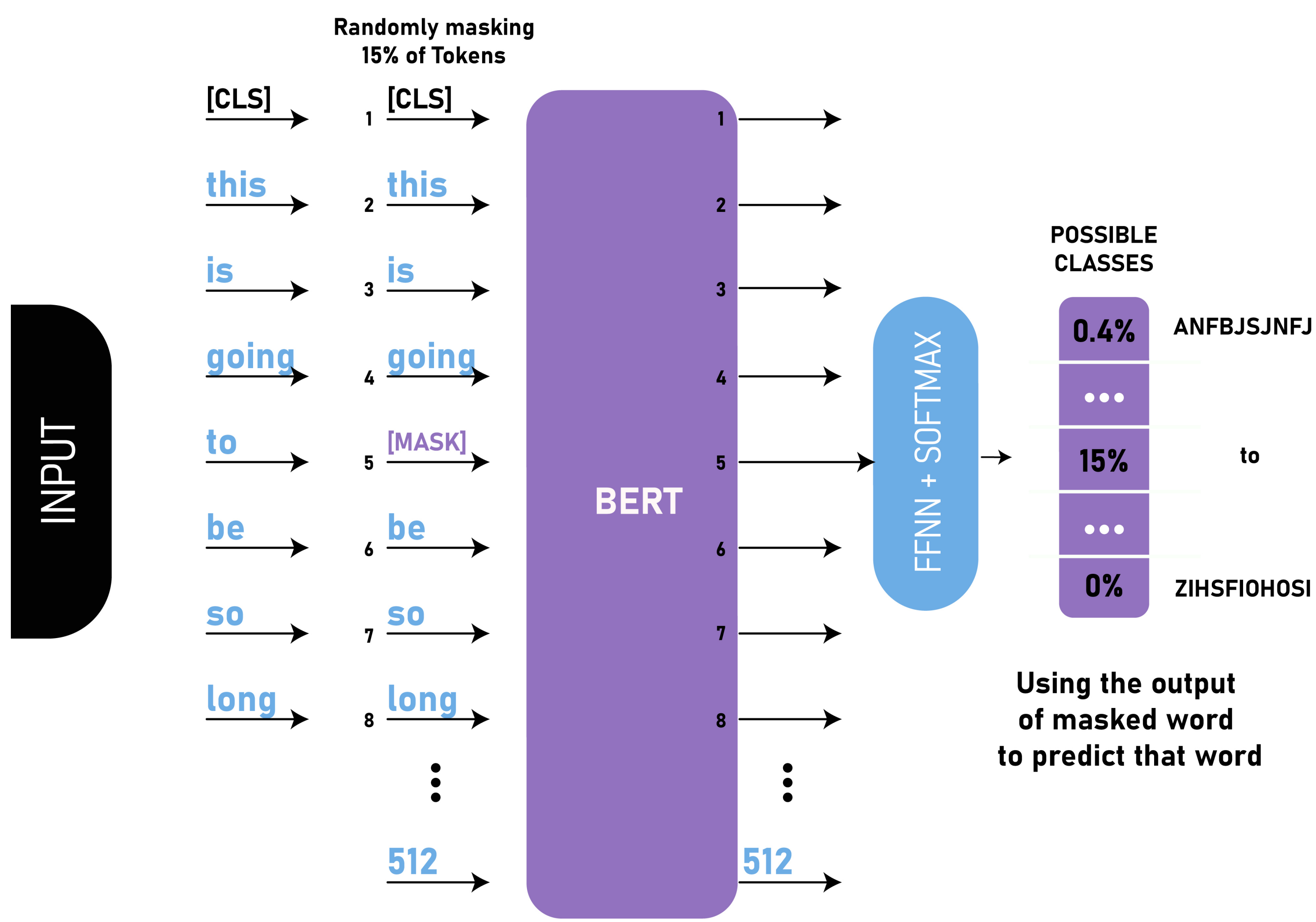
How Does It Work?

- **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

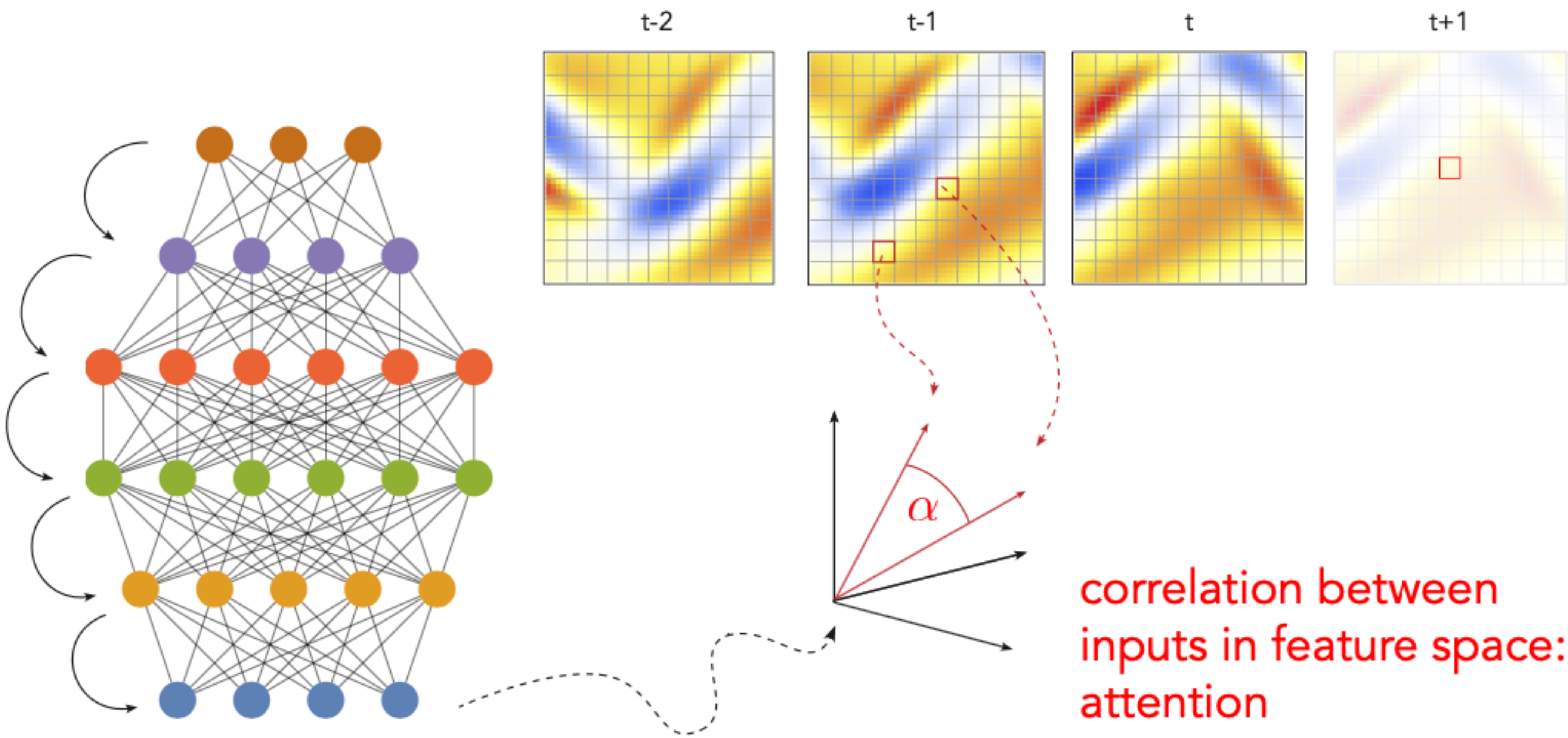
Pre-training: Computer Vision

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

Complication: what is a token?

Single pixels carry too little information.

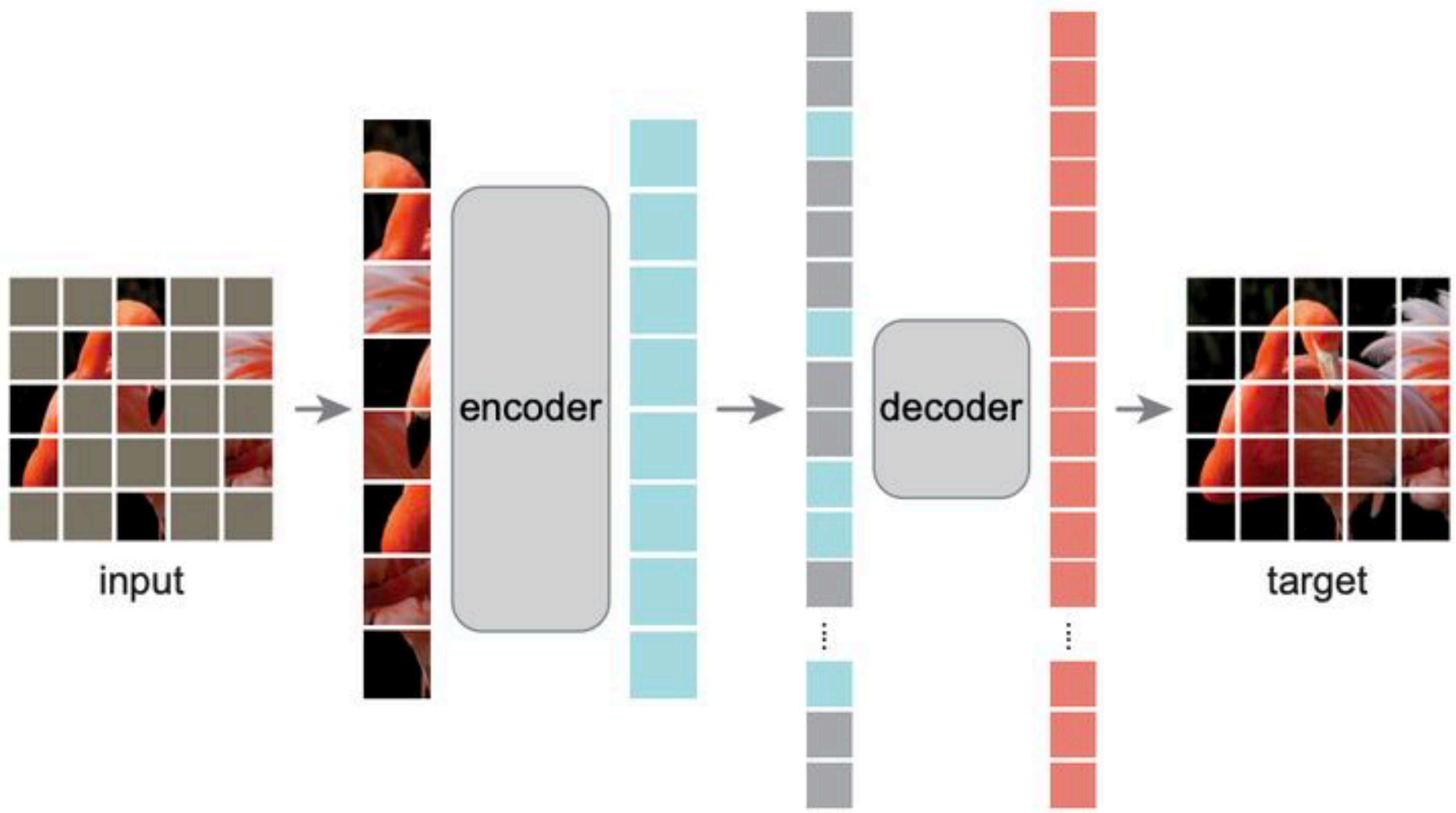
trade-off between token-size and information in each token



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.



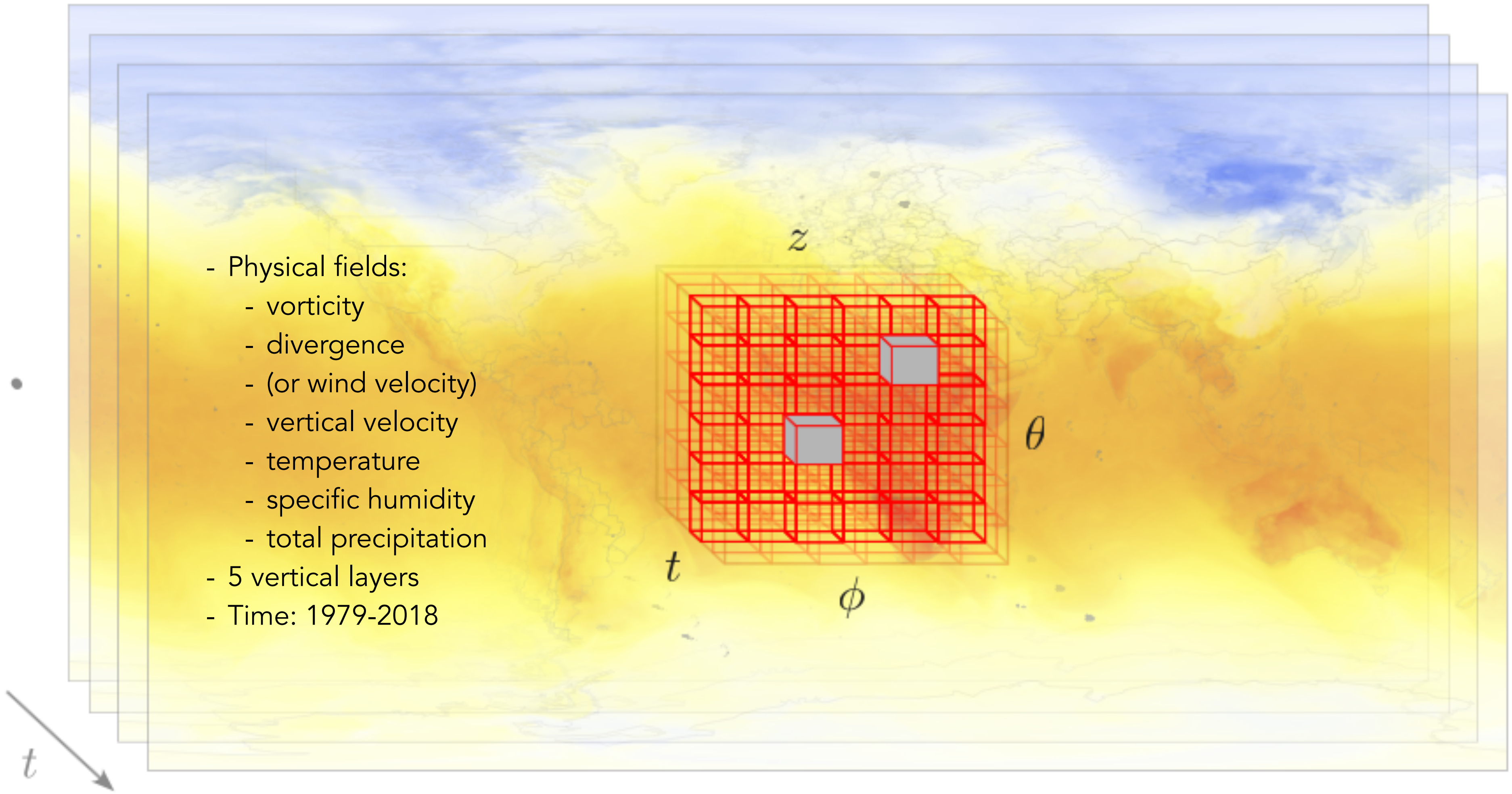
Example: Removing a portion of an image and training the model to reconstruct the removed region.

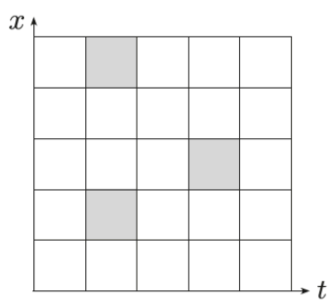
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

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





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

Mask random tokens within the hyper-cube and predict them

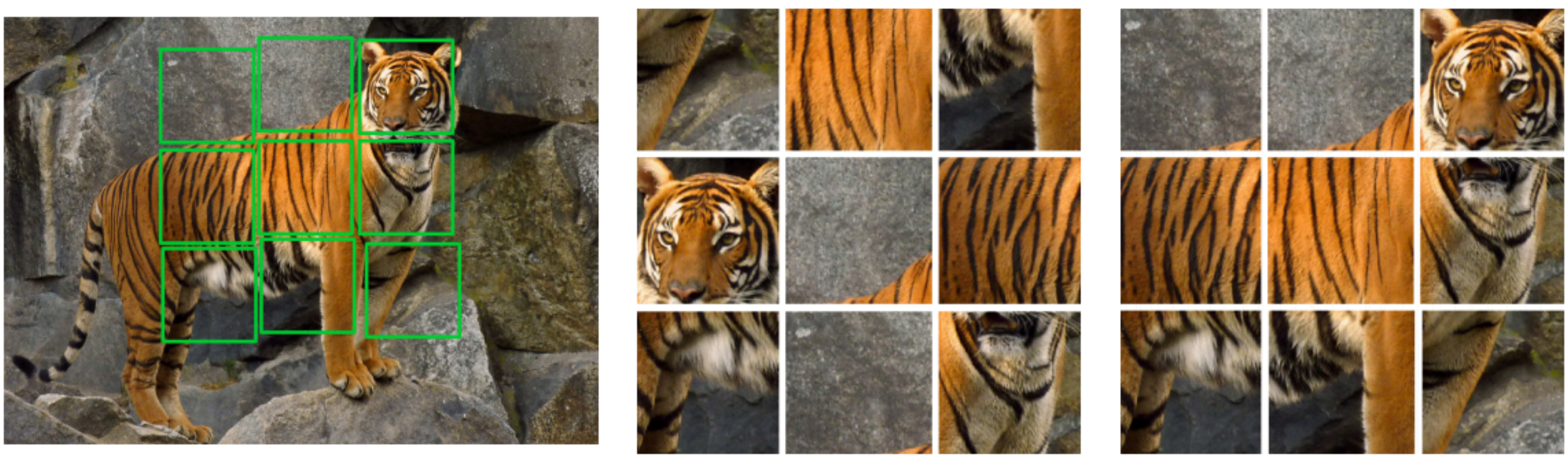
Large masking ratios above 80% using full masking, noise and climatology

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

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.



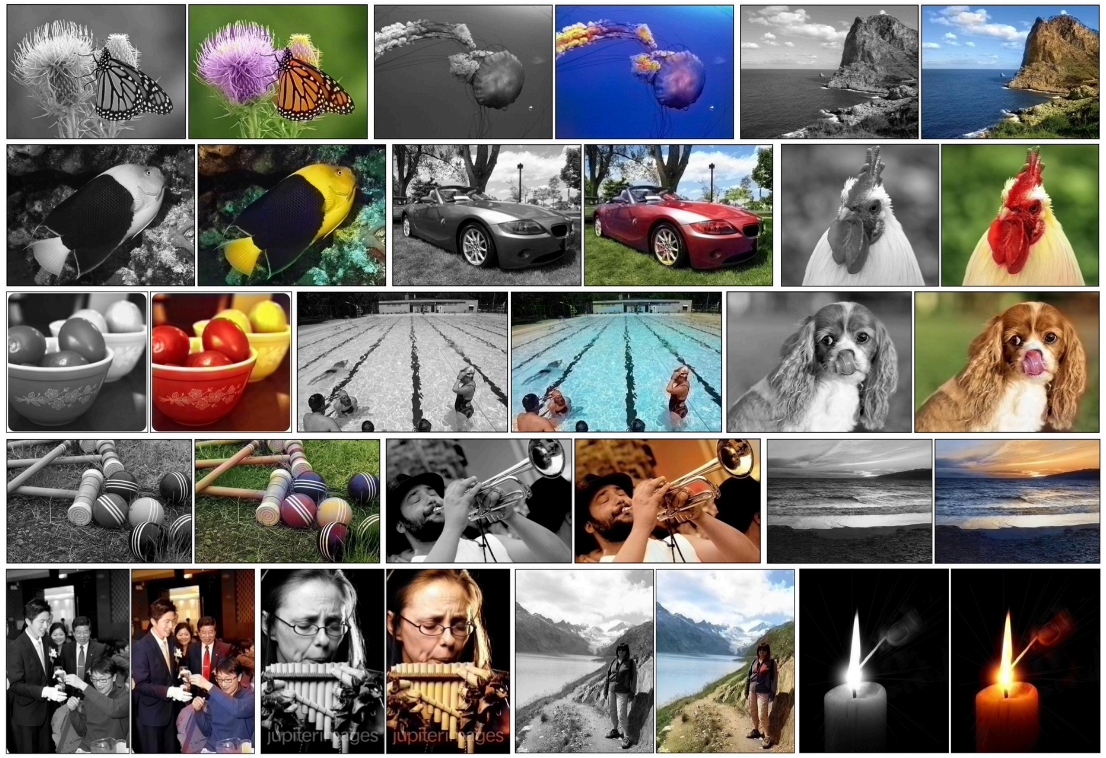
Example: Splitting an image into a 3x3 grid, shuffling the patches, and training the model to solve the puzzle.

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.

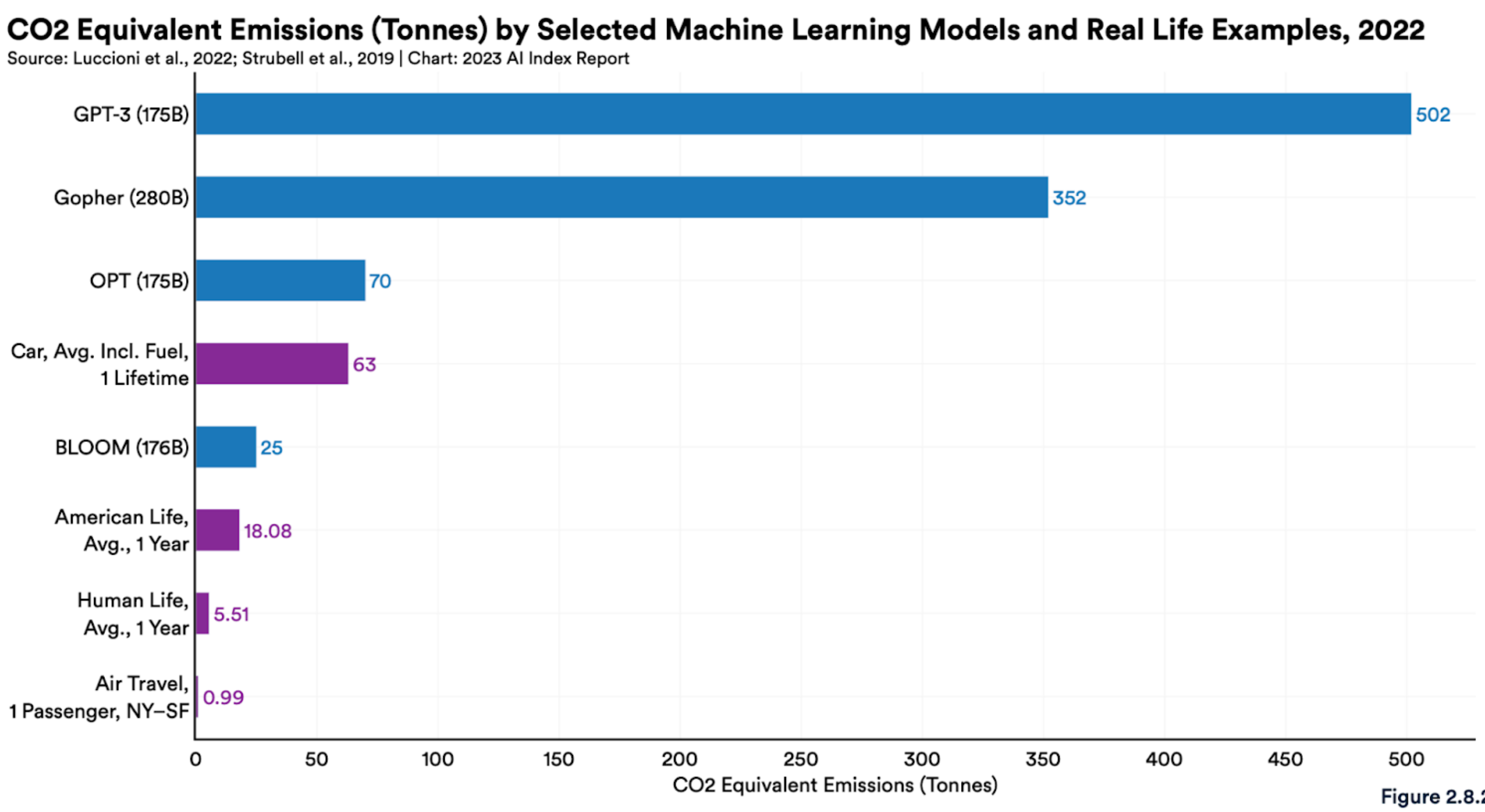
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

Hardware and footprint

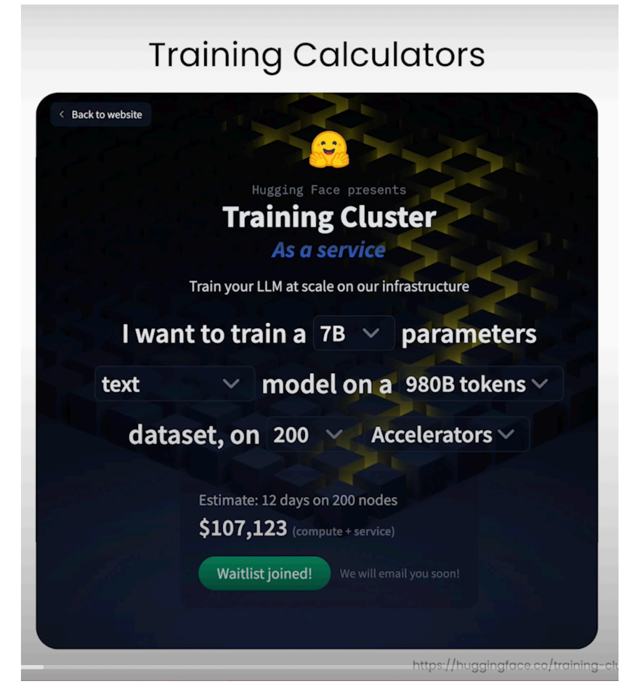
Computing Resources: Distributed computing

- **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

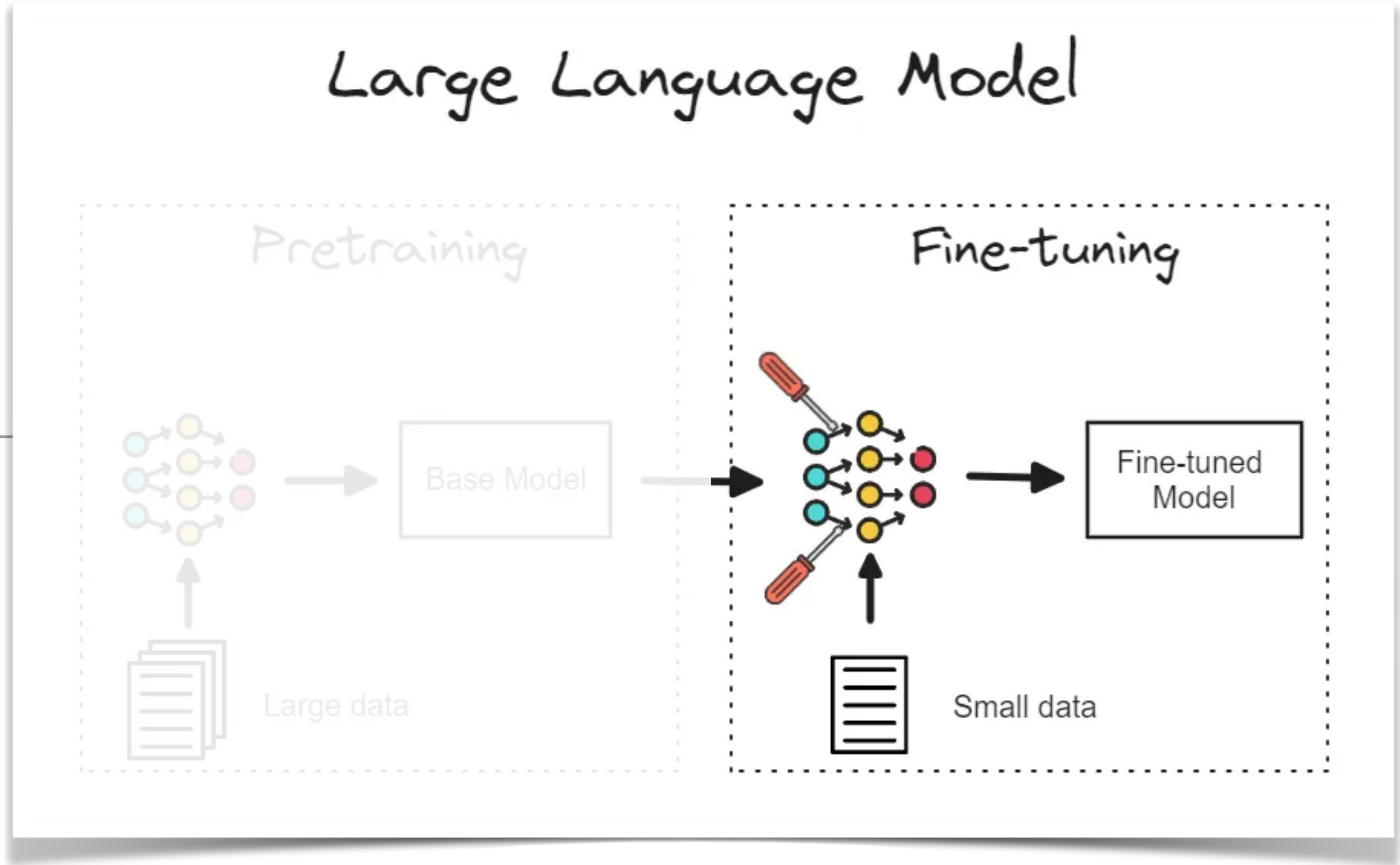


https://analyticsindiamag.com/ai-origins-evolution/the-environmental-impact-of-llms/

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

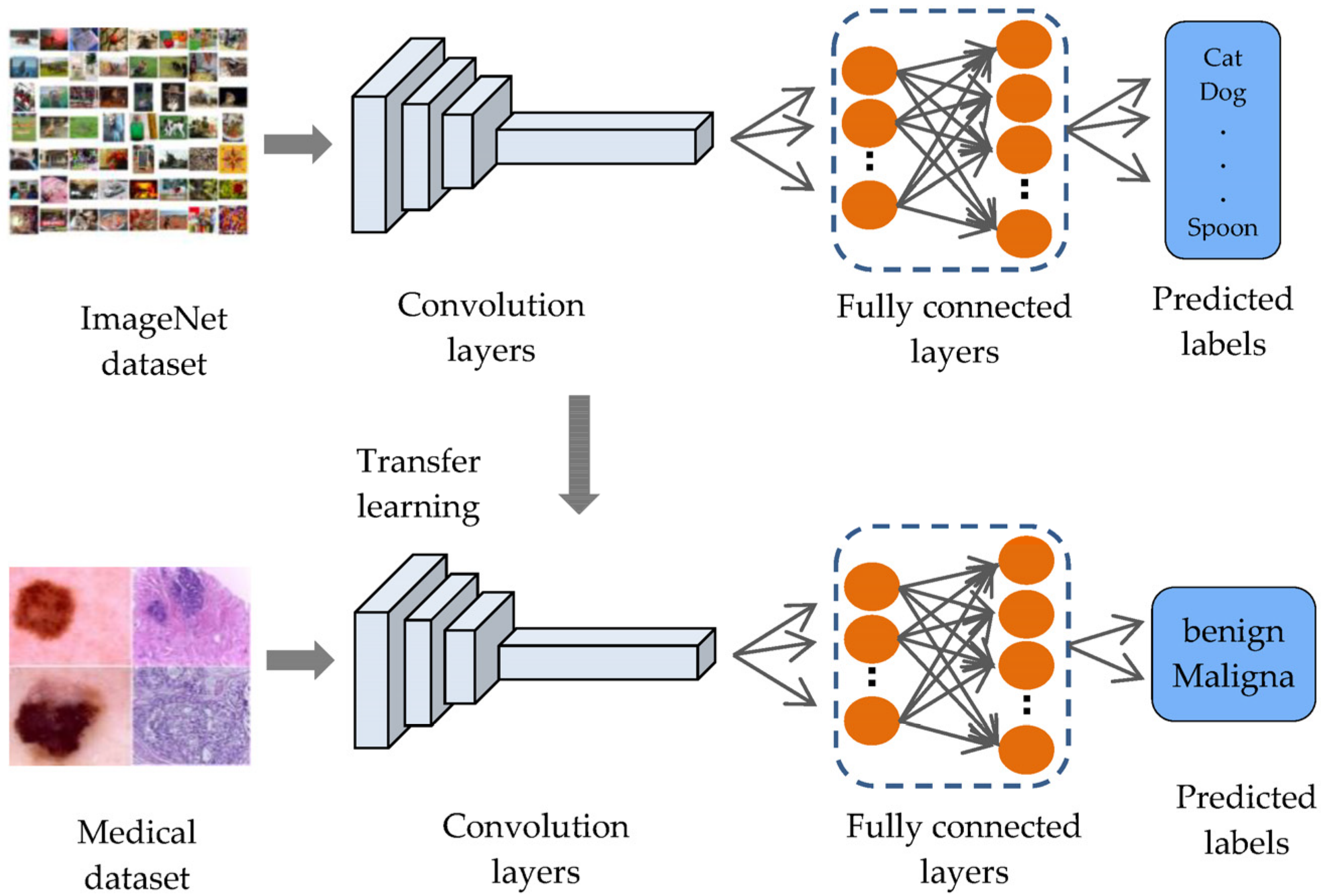
Fine-tuning

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.

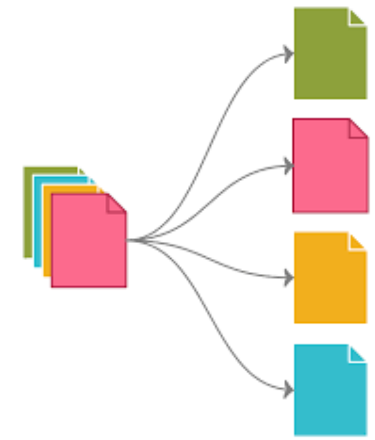
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

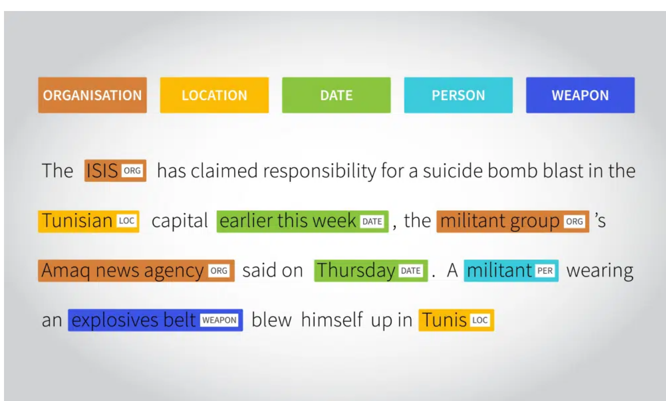
Fine-tuning in NLP - examples

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):

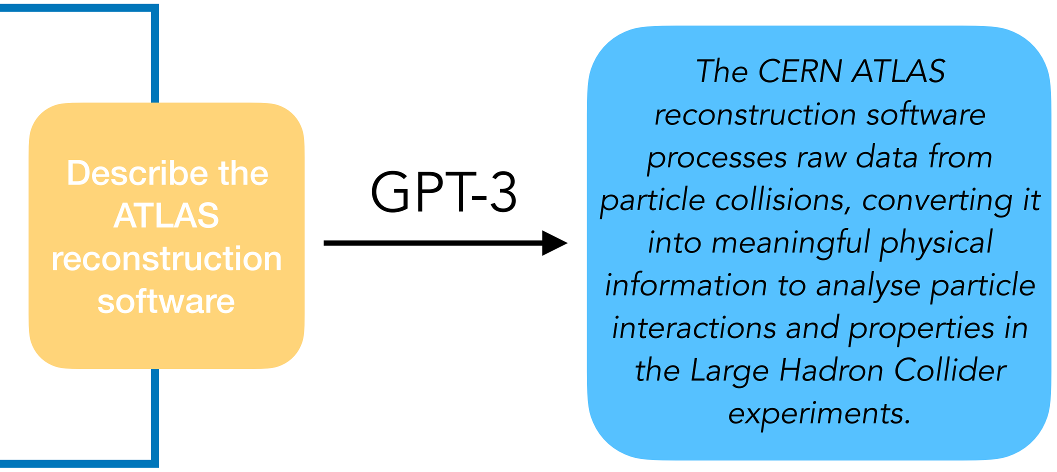
- **Task:** Identify entities like names, dates, and locations in text.
- **Example:** Fine-tuning a pre-trained RoBERTa model on a labeled NER dataset such as CoNLL-2003.

  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.



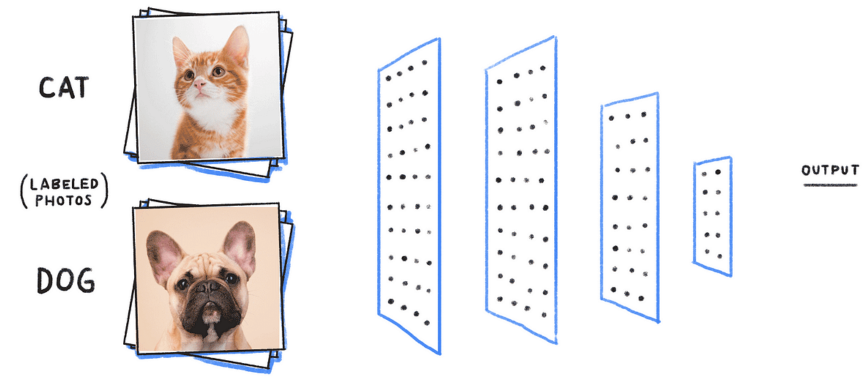
Ilaria Luise, Sofia Vallecorsa CERN - ilaria.luise@cern.ch I sofia.vallecorsa@cern.ch

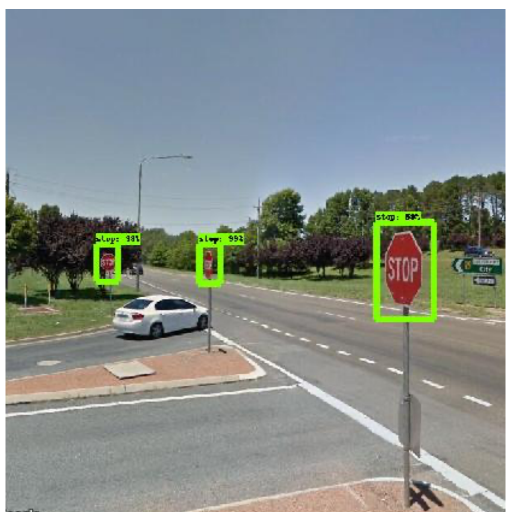
Fine-tuning in Computer Vision - examples

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.

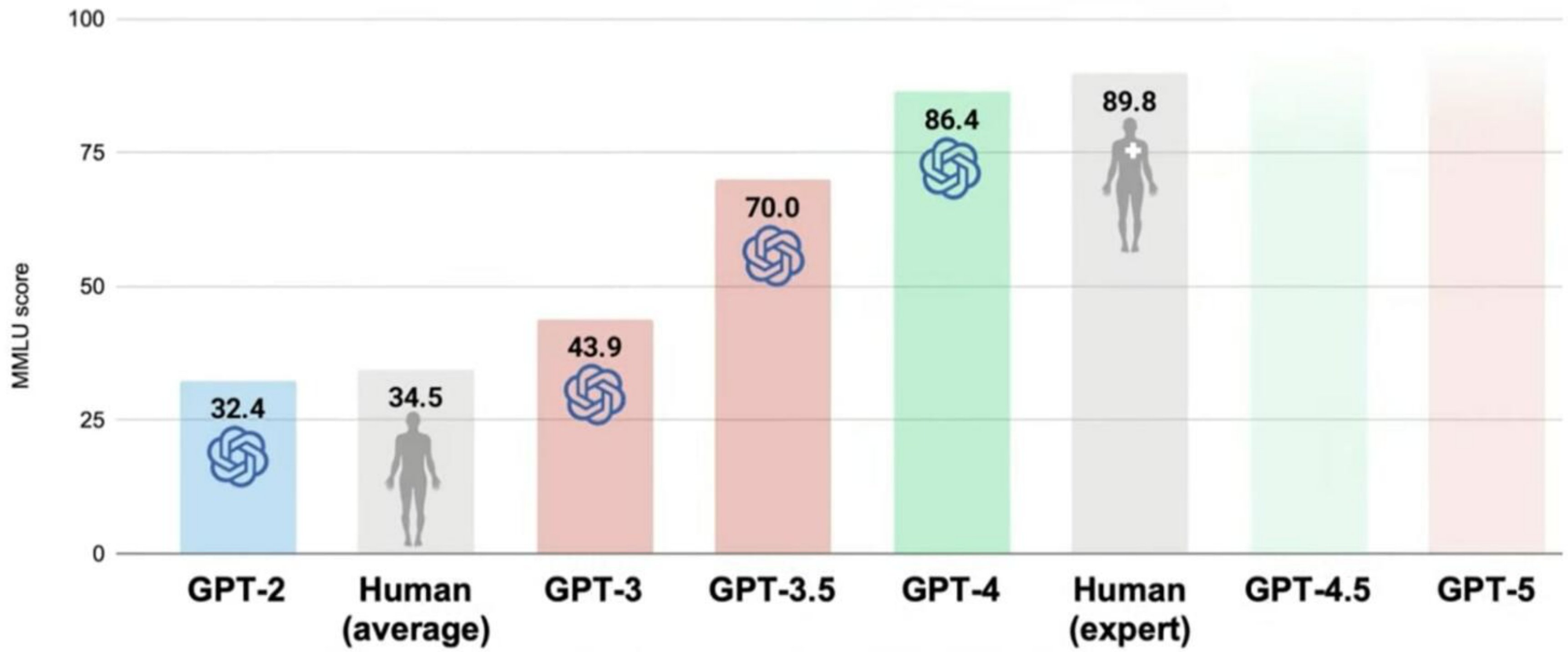
  Steps:
  1. Load a pre-trained U-Net model.
  2. Replace the output layer for segmentation tasks.
  3. Train the model on annotated medical images.

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*
- *…*

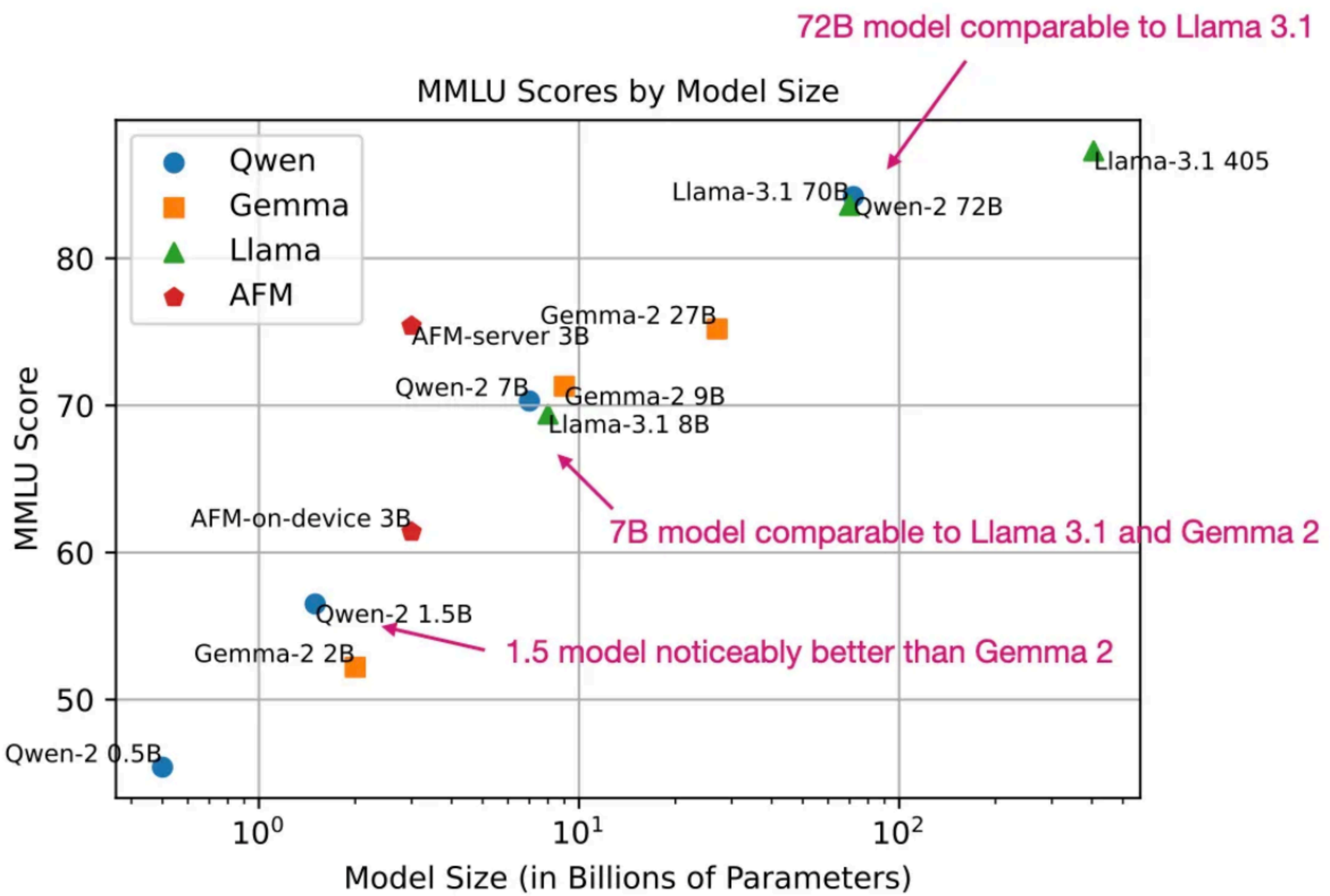
Other evaluation metrics:
- Bilingual EvaLuation Understudy (BELU) BLEU
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation).
- METEOR: explicitly sorted translation evaluation metric.
- Perplexity Perplexity is also called the degree of confusion.

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

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

Image from LifeArchitect.ai

LLMs benchmarks



MMLU benchmark scores for the latest open-weight models (higher values are better). I collected the scores for this plot from the official research papers of each model.

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

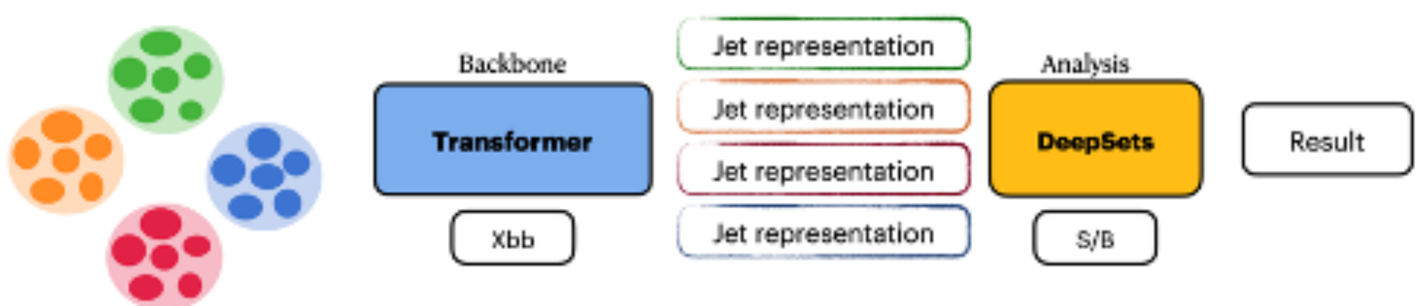
NB: LLMs are quickly entering our domain

Foundation Models in HEP

Multiple studies in HEP (transformers, self supervision, fine tuning for HEP data, etc..)

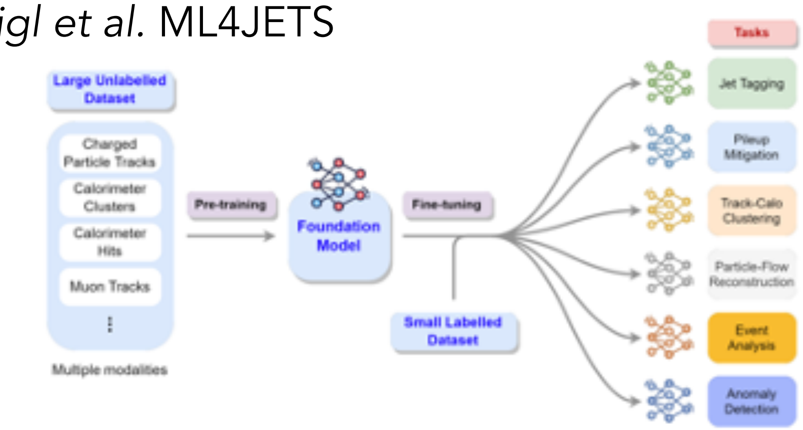
A topic present in **many conferences and workshops**, (IML, ACAT, CHEP, ML4JET, …)

Direct **application of LLMs** to HEP (information mining, coding, etc..)

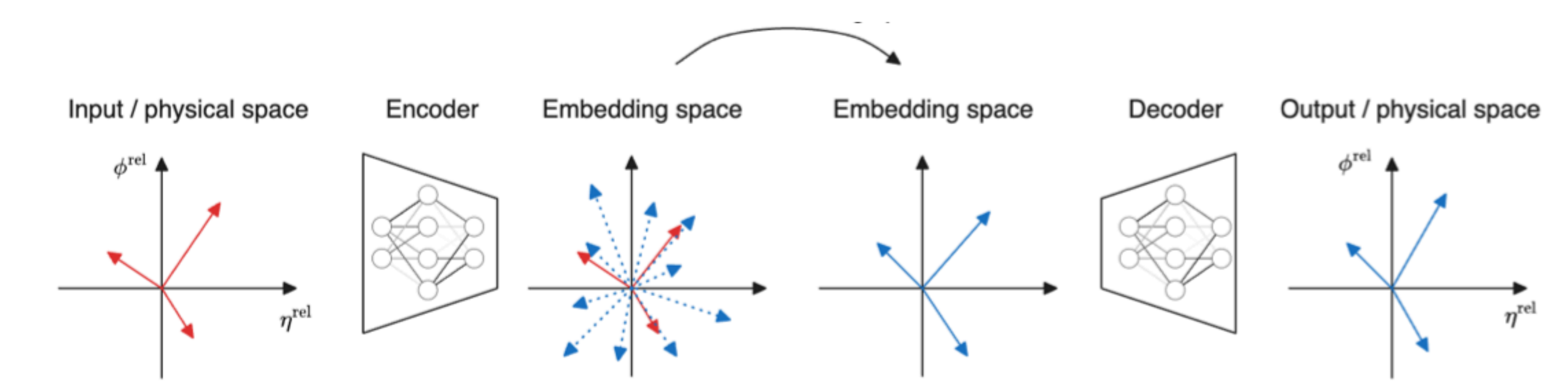


Masked particle modelling, *M. Leigh et al.* ML4JETS

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

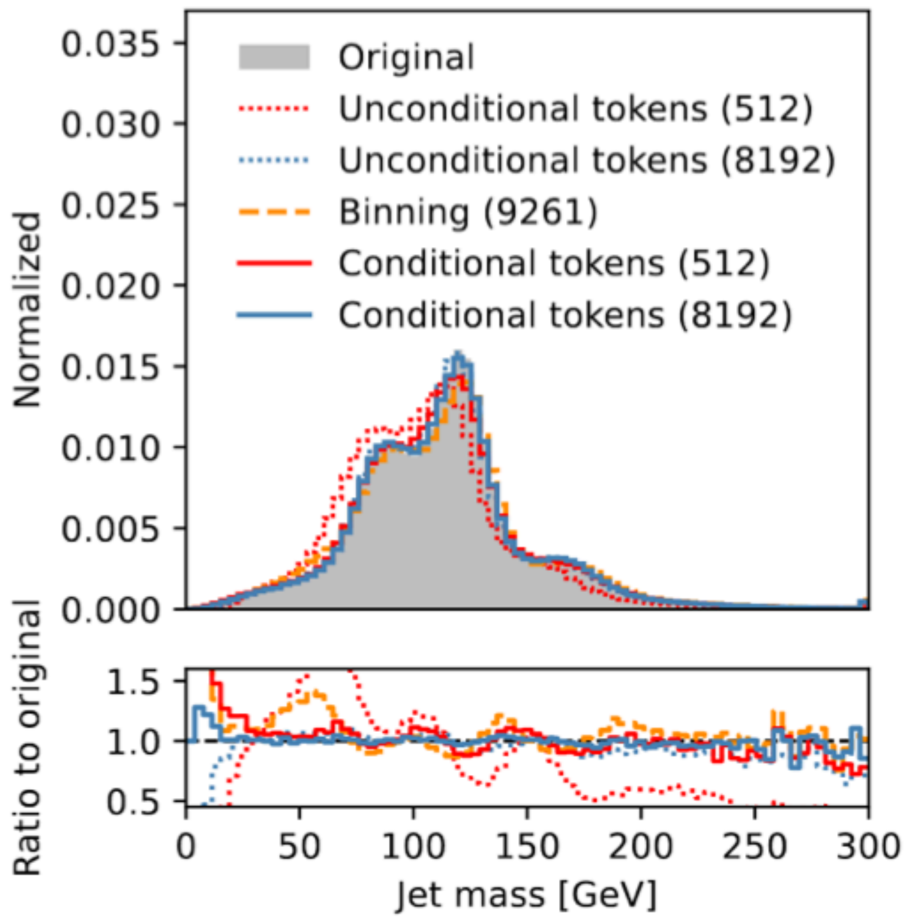
Simulating particle jets

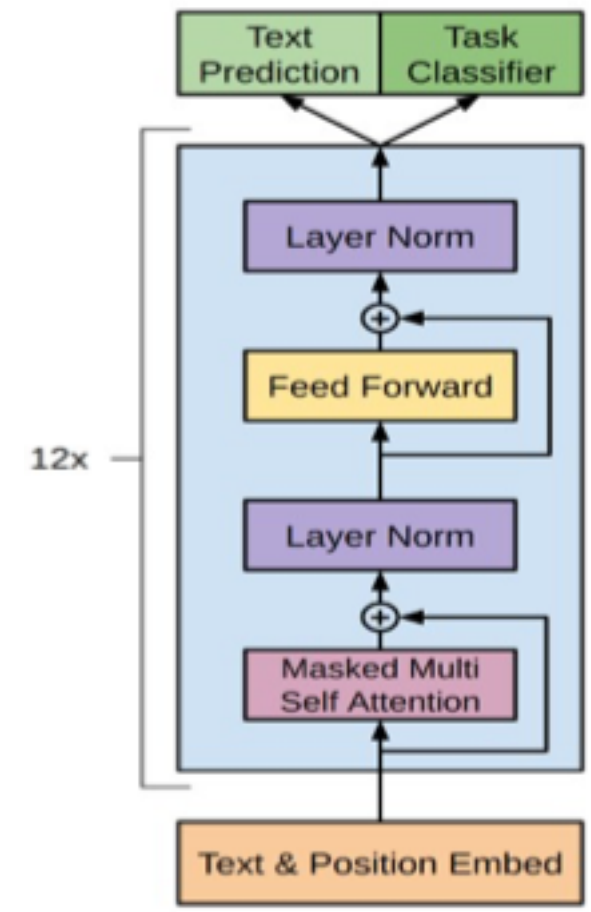
Anna Hallin et al. arxiv: 2403.05618

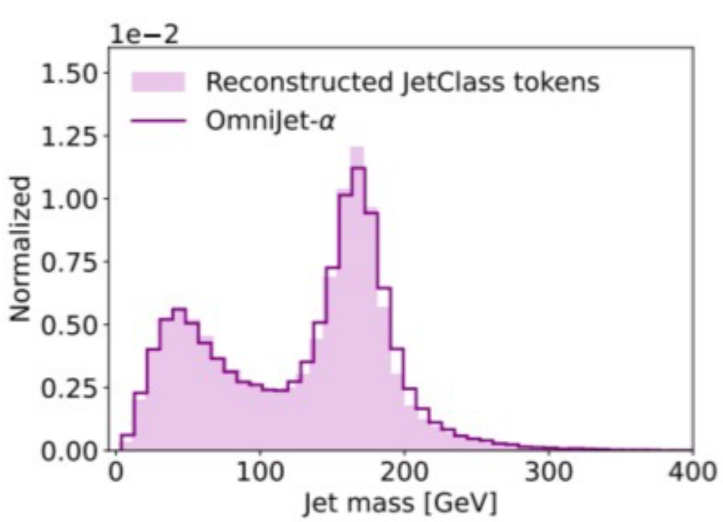
Particle and jets are interpreted as words and sentences.

Use transformers as NLP to perform jet classification and generation

- Transformers expect tokens
- What happens to the continuous physics information ?

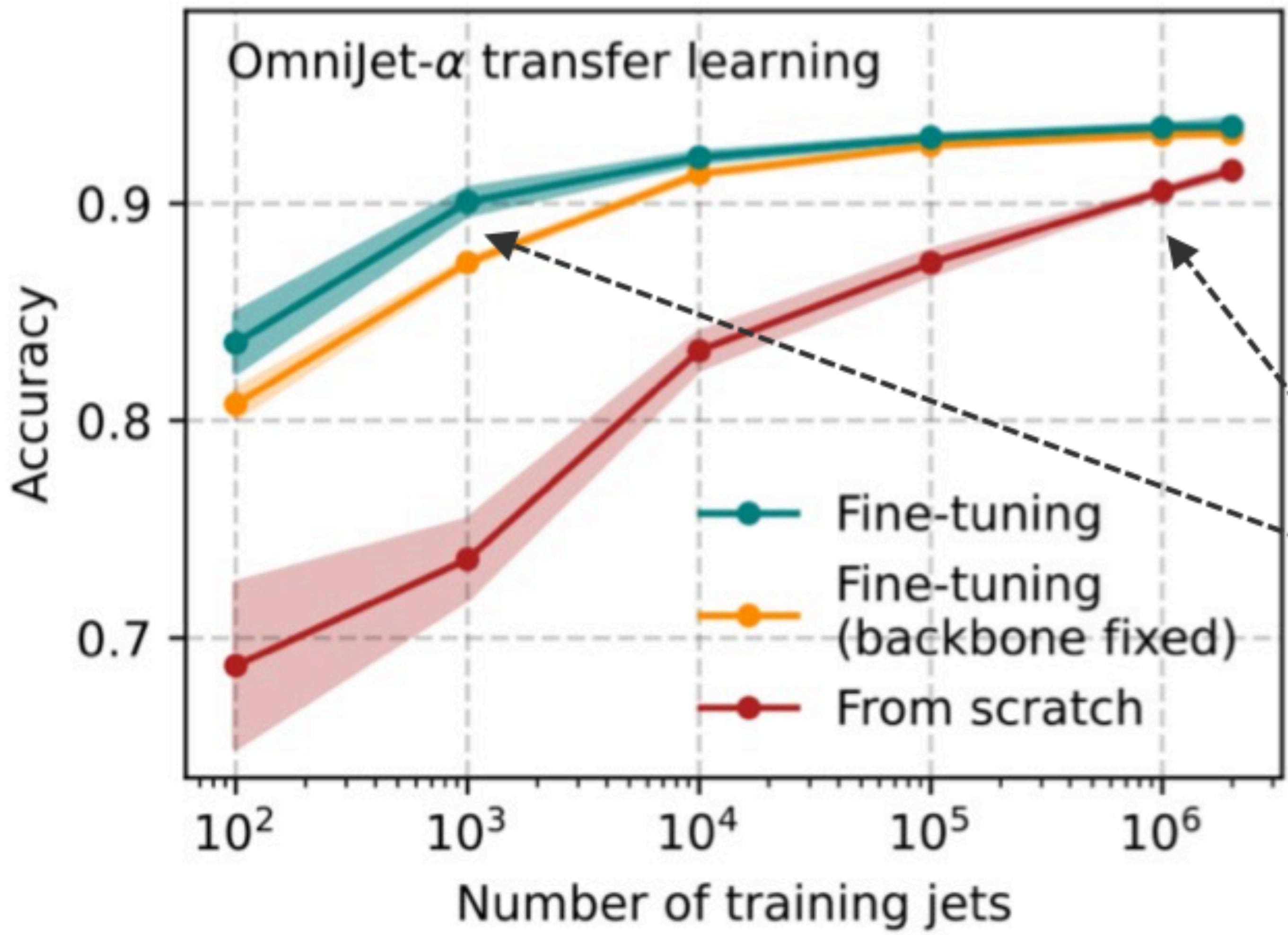






Simulating particle jets

Anna Hallin et al. arxiv: 2403.05618

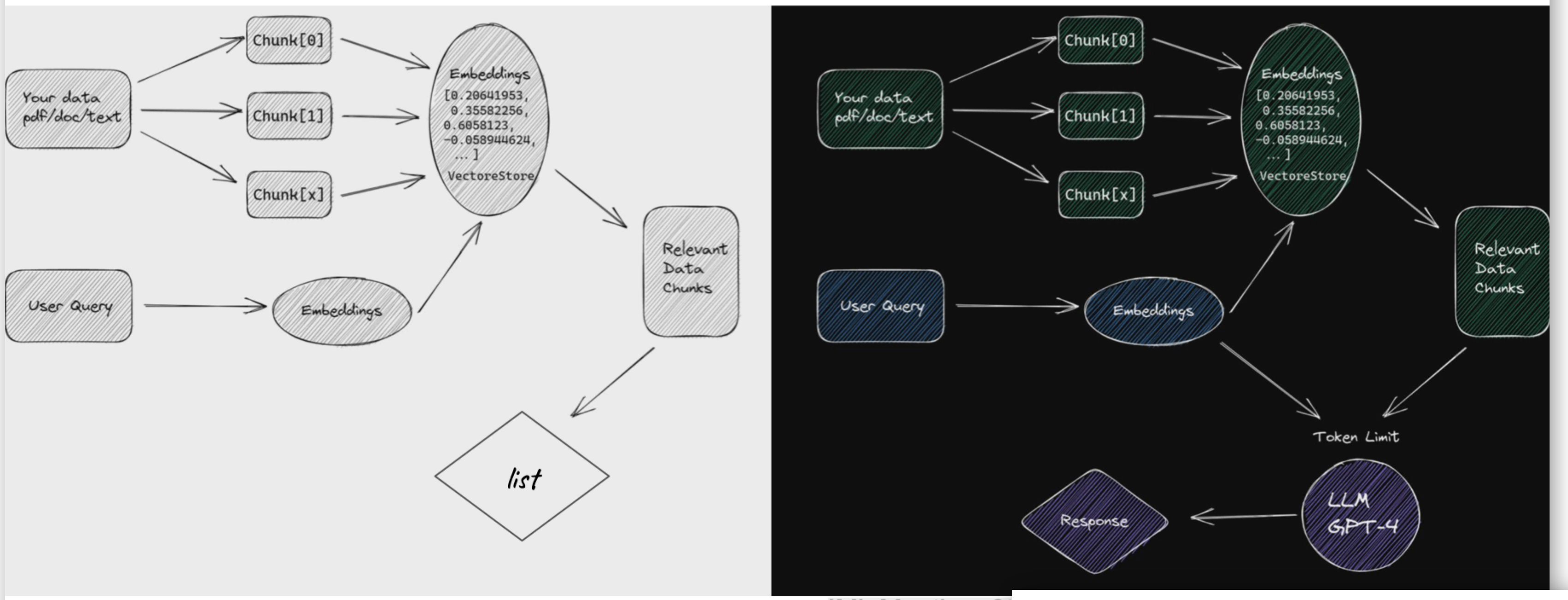


Pre-trained model requires only 1000 training events to reach the same accuracy level that the "from scratch" model reaches with 1M events

LLMs as scientific Assistants

chATLAS: RAG using various internal ATLAS sources

Search mode (not a RAG) | Assistant (RAG)



:hATLAS An AI Assistant for the ATLAS Collaboration IML Meeting, 9t

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

Why AccGPT?

AccGPT (Accelerating GPT).
- Our vision: Accelerating Research.

First step: Enhancing knowledge retrieval.
- **Challenge: CERN has many and HUGE data bases:**
  - (>)> 50 knowledge (web) domains for documentation.
    - Challenging to find information without knowing its location.
  - CERN wiki (Confluence): > 1M wiki pages.
  - CERN Document Server (CDS): > 500k documents.
  - CERN home: > 10k webpages.
  - CERNbox and more domains …

By ChatGPT

→ **Objective:** Leverage AccGPT to improve knowledge finding, user support, streamline development processes, and enhance onboarding experiences.

Florian Rehm - AccGPT 09.04.2024 3

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