Foundation models
&
scientific discovery

Tobias Golling, University of Geneva
Lukas Heinrich, Technical University of Munich

Setting the stage

- Supposed to be interactive
- Think out loud, brainstorm,…
- Comment & vote on SpeakUp

This is just the **first step!**

- Follow up, please fill out: https://bit.ly/eucaifcon24-wg1

https://web.speakup.info/room/join/36585

Let’s get right into it


Will we ever be able to talk to whales?

Part 1

The promise of foundation models

What **is** a foundation model?

You heard the word already dozens of times at this conference !

Definition?

Examples?

Characteristic features of a FM

Pre-train using SSL – no labels needed

“**Meaningful data representation**,” “Implicit model”

Transferrable & finetunable: *Easy* to adopt to multiple downstream tasks

Multimodality character: *one model to do it all* – common embedding / no pairing needed

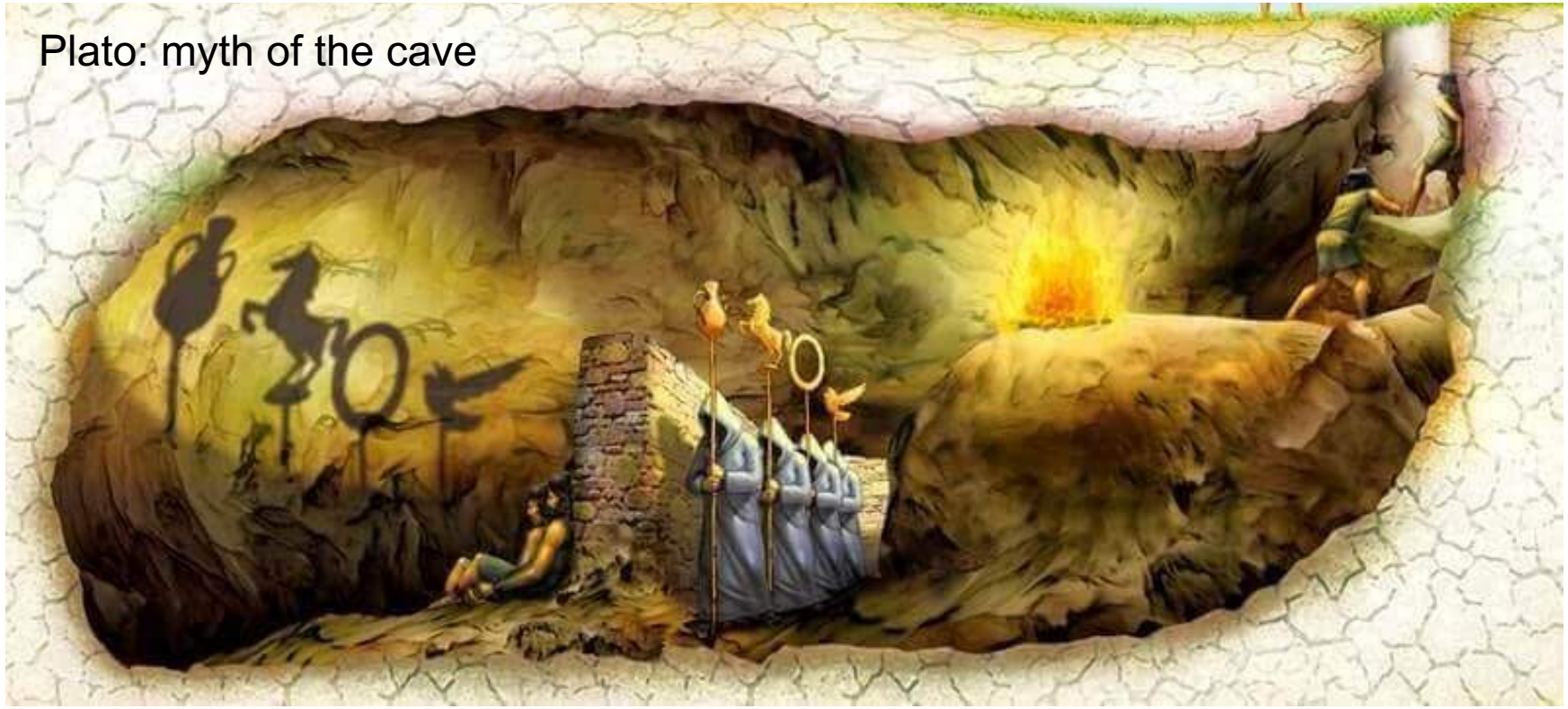
FMs = stochastic generative models with high expressiveness and outstanding interpolation and generalization power in ultra-sparse training data spaces of high dimensionality.

Pre-training

- Masking
- Augmentation
- … novel training schemes / physics-inspired?
- Need for auxiliary tasks?
- Encode physics as inductive bias? Flexible prior [Miles]

- **Evaluation** – go beyond downstream task?

Foundation models: learned *meaningful* **latent** representations



Learn true underlying objects (latent variables) from observed data (shadows)

Compare with **our** embedding spaces

- Our **reconstruction** is a common embedding space of our data
- Our **theory space** is a multi-modal common embedding space (e.g. combined fits, combination plots)

- **What does FM add to this?**
  - **End-to-end**
  - **Differentiable**
  - **Democratize** AI – commonly trained [Anna Scaife]
  - **Common** model across subdetectors, experiments,…
  - Independent of **theory**? Add theory as a modality?
  - …?
  - **AI oracle ⇔ interpretability**
    - Matt: machine understands & explains it to 5-year old
    - Miles: symbols → selection effect [+, x], *simplicity generalizes well*

Shared model / embedding space across

- ... multiple subdetectors?
- ... multiple LHC experiments?
- ... multiple HEP experiments?
- ... between Astro / HEP experiments?
- … beyond?
- … the utility of language & LLMs?

What can this EuCAIF WG do?

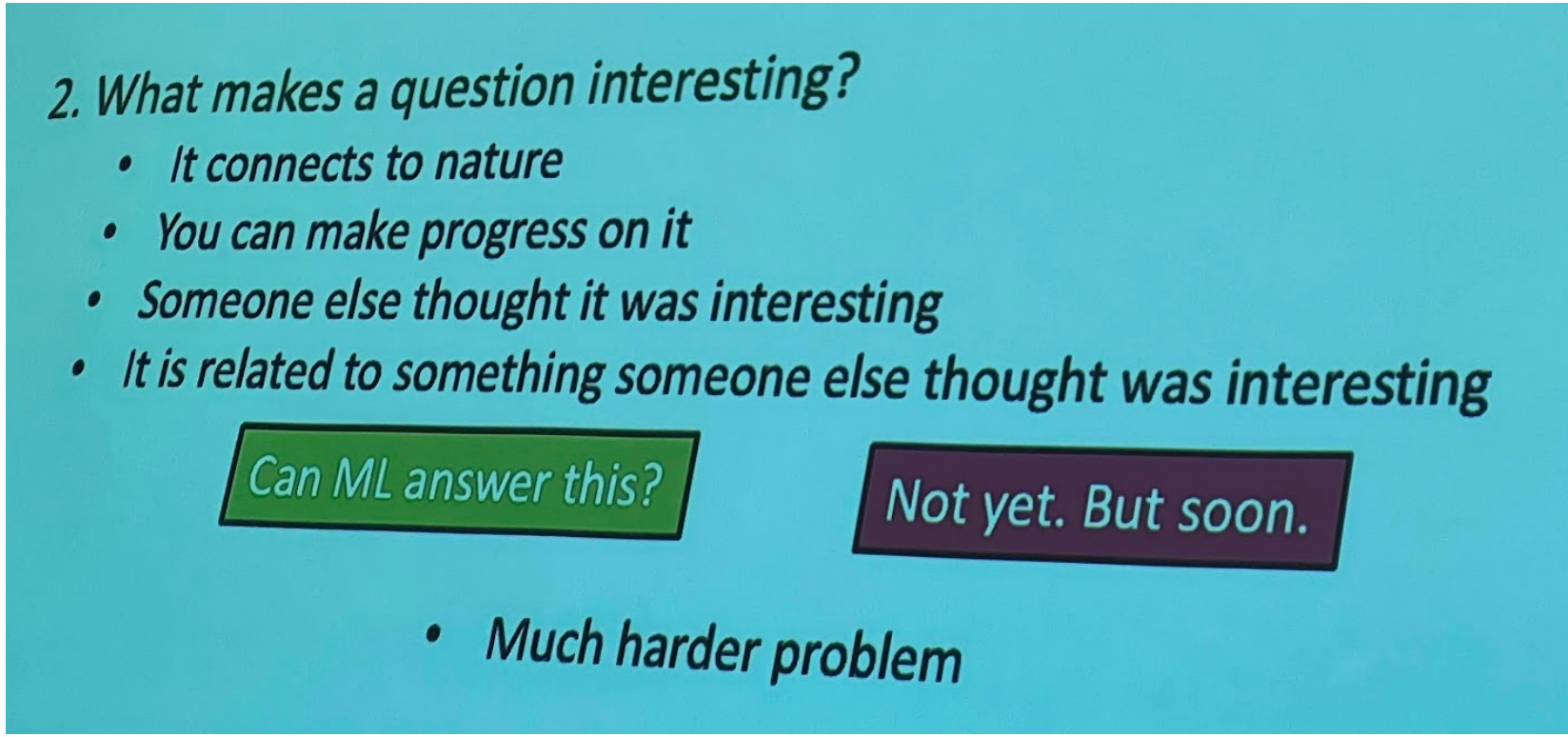
Community consensus [see ML4Jets]

Sample over all of you

Identify trends

Snapshot: white paper (?)

Continued effort: workshops, seminars, exchange,…



Matt's talk from this morning

[**Not** trying to direct research... let a thousand flowers bloom]

Facilitate collaboration

- What is needed?

- Harmonizing / publicizing **datasets**
  - **MNIST of PP** [challenges e.g. HiggsML, TrackML,…]
  - Scaling up: 100M jets → 10B jets → …
  - How to meet scaled up compute (GPU) needs?
- Community **benchmarks** & metrics / evaluation
- **Common software framework**?
- Interface with **experiments?**
- **...**
- (Big) European funding tools?

Questions, comments, suggestions,…

Reminder

Stay involved

Add your thoughts, comments, suggestions, preferences here: https://bit.ly/eucaifcon24-wg1

https://web.speakup.info/room/join/36585

Who is using foundation models?

Who **plans to use** foundation models?

More thoughts?

Part 2: discovery

Optimal search for the unknown

- Trade-off between **generality** and **specificity**
- Knob to tune pareto optimality between the endpoints: **supervised** & **unsupervised**
- What **metric** to assess performance – should **not** be known models
- What's the **follow-up strategy** after an "anomalous" signal ?
  - Balance cost of follow-up against frequency alerts ?

Automation Automation Automation

- Which tools are needed to automate analyses / interpretations efficiently
  - Fast approximate detector simulation
  - Fast approximate analysis implementations
  - Full recast capability
- Which Data Products are needed (and be public ?)
  - Public likelihoods (neural approximations?)
  - Public models

Interface: foundation models **&** discovery

- **Common / portable** model [efficient]
- Allows to **accelerate and automate** cycle of scientific method

Backup: one-slide primer on search strategy

- NP: best test statistics is full high-dim LH ratio = p_0/p_1

- Add realism:
  - Finite statistics + syst. ⇒ data selection, factorise LH ⇒ ∑ LH
  - Time sink: MC simulation/calibration/syst. & analysis optimization
  - Finite person power ⇒ **Automate**
  - Finite compute ⇒ **Fast solutions**
  - Finite belief in BSM ⇒ **Hedging / diverse approach**
    - Definition of **coverage (metrics)** / Bayesian: **vary prior (model / learn p_1)***
    - Tune *continuously* **power of test** vs. **assumptions**
  - Look elsewhere ⇒ **Minimum #tests**
  - Interpretation ⇒ Benchmarking, reinterpretation

***Flexible** enough to approximate true hypothesis & **regularized** enough to be insensitive to statistical fluctuations

Thank you !