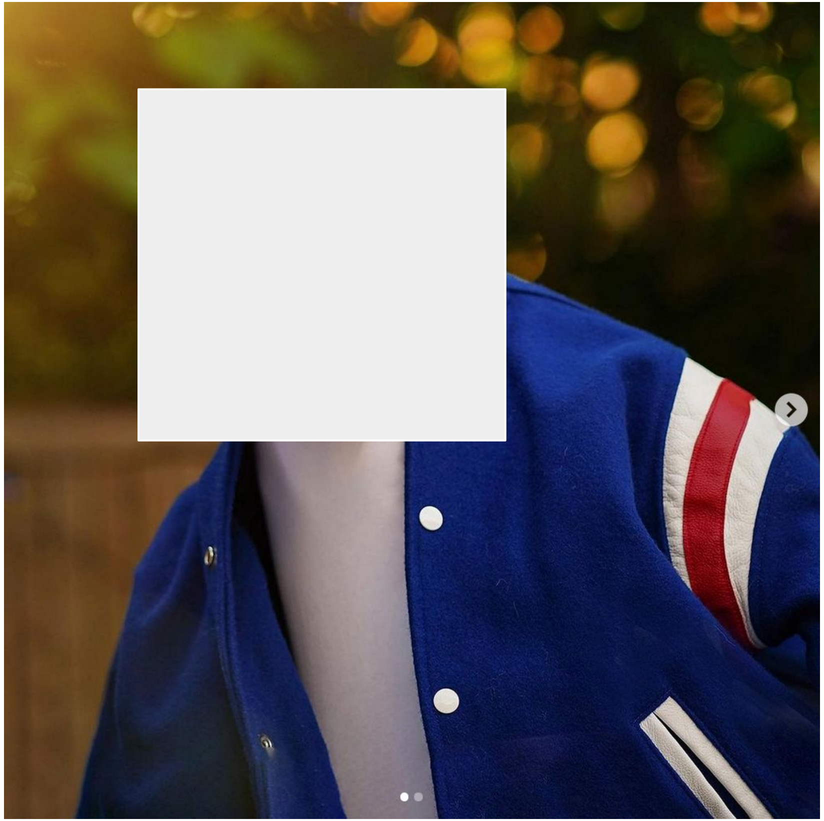
Machine Learning per la fisica delle alte energie: stato dell'arte e prospettive future

SCUOLA NORMALE SUPERIORE

INFN Istituto Nazionale di Fisica Nucleare

Francesco Vaselli
Scuola Normale Superiore & INFN Pisa

Un uomo o una donna?



Nessuno dei due!

Come ho potuto ingannarvi?

Sfruttando l’attenzione:
come i dettagli ci permettono di predire la parte mancante

Possiamo replicare lo stesso meccanismo nel ML?

Si, ed e’ la più grande innovazione degli ultimi anni!

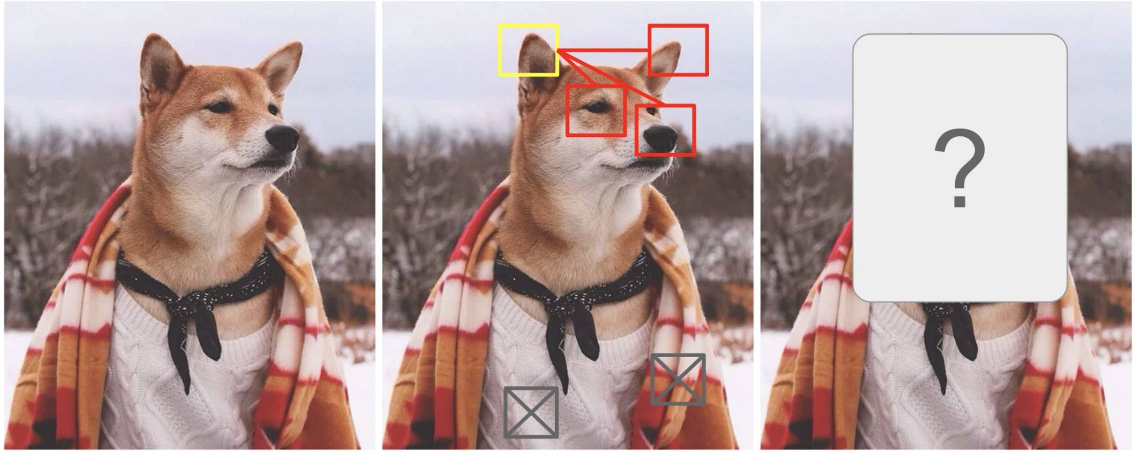
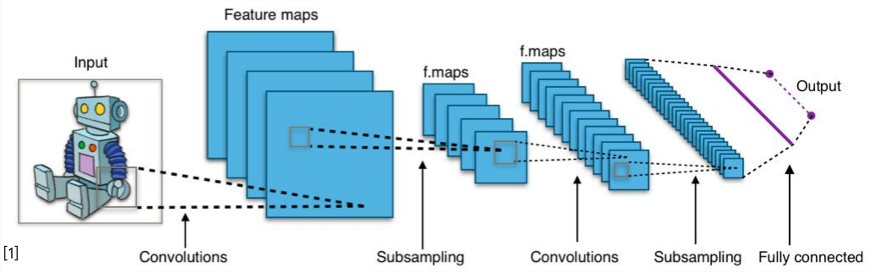


image source: Lilian Weng blog

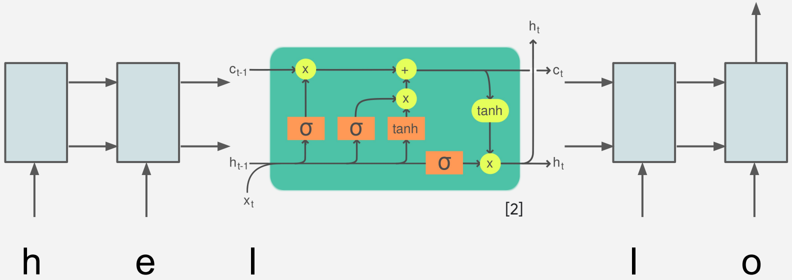
Computer Vision

Convolutional NNs (+ResNets)



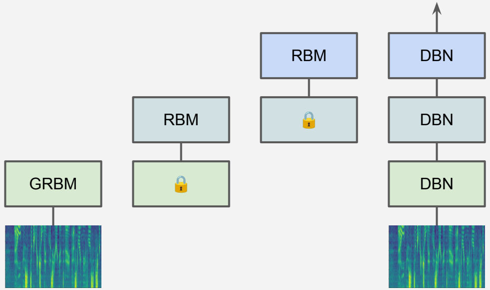
Natural Lang. Proc.

Recurrent NNs (+LSTMs)



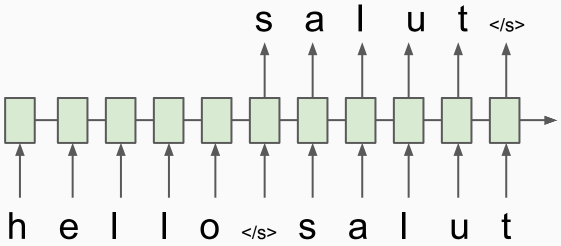
Speech

Deep Belief Nets (+non-DL)



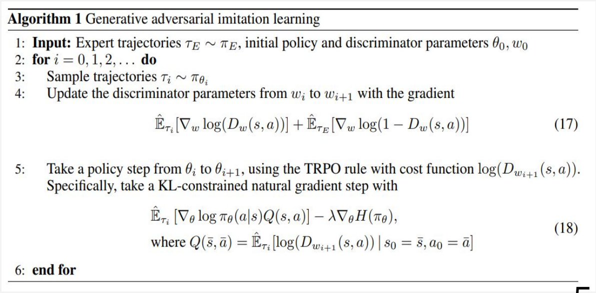
Translation

Seq2Seq



RL

BC/GAIL



Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

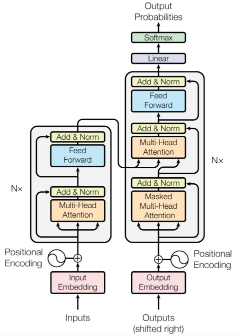
$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \quad (17)$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

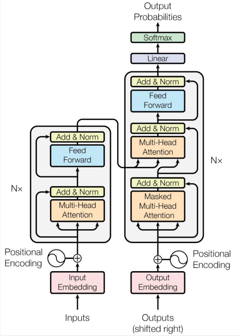
$$\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta),$$
$$\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}] \quad (18)$$

6: **end for**

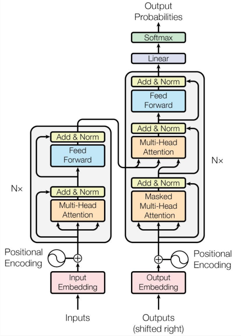
Computer Vision



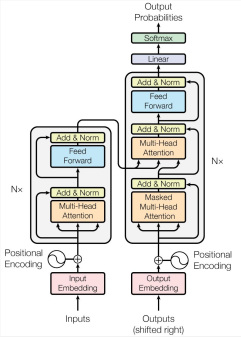
Natural Lang. Proc.



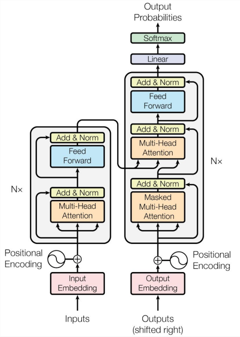
Reinf. Learning



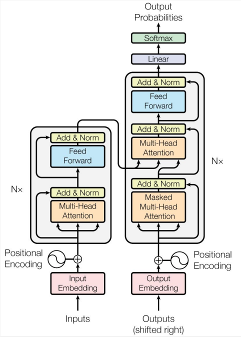
Speech



Translation



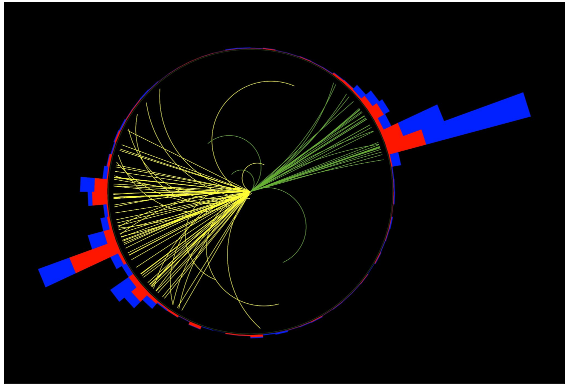
Graphs/Science

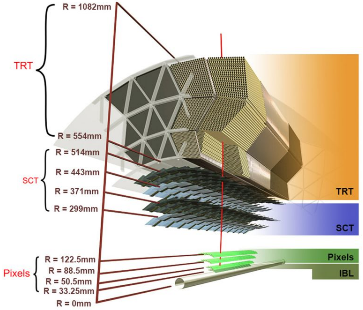


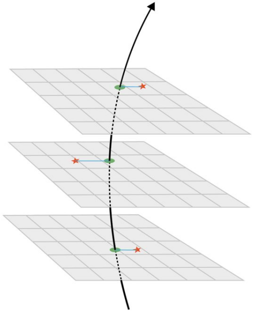
Transformer image source: "Attention Is All You Need" paper

L’attenzione come prodotto di matrici apprese!

X
see
that
girl
run

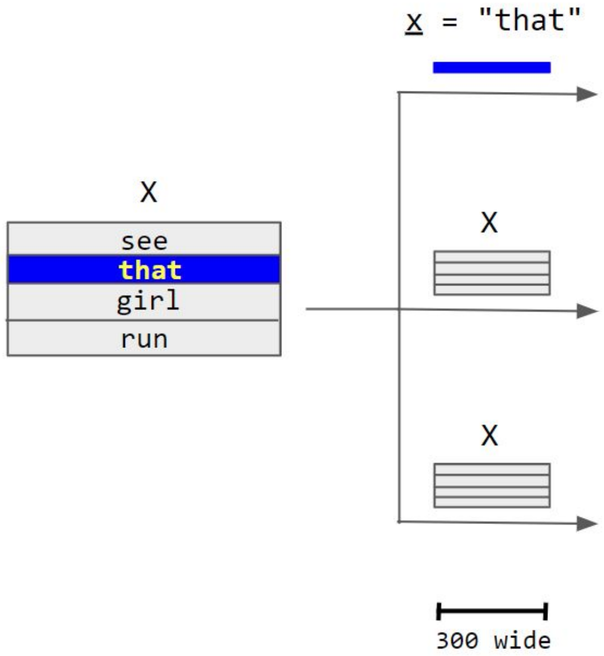




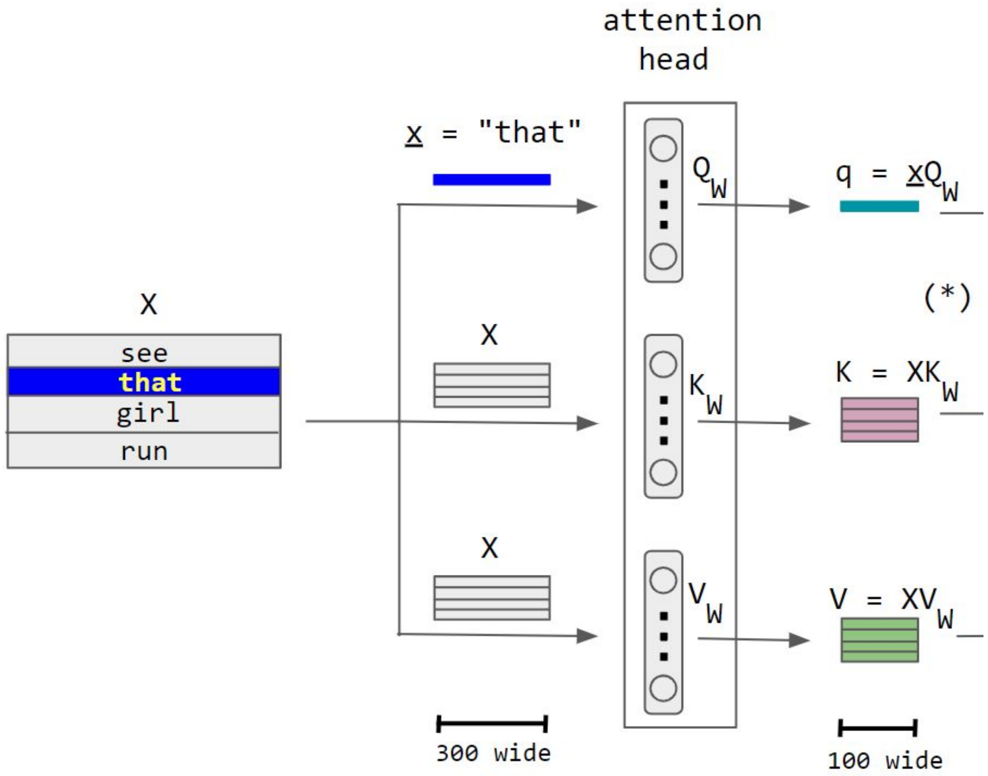


Immagini da CMS, ATLAS

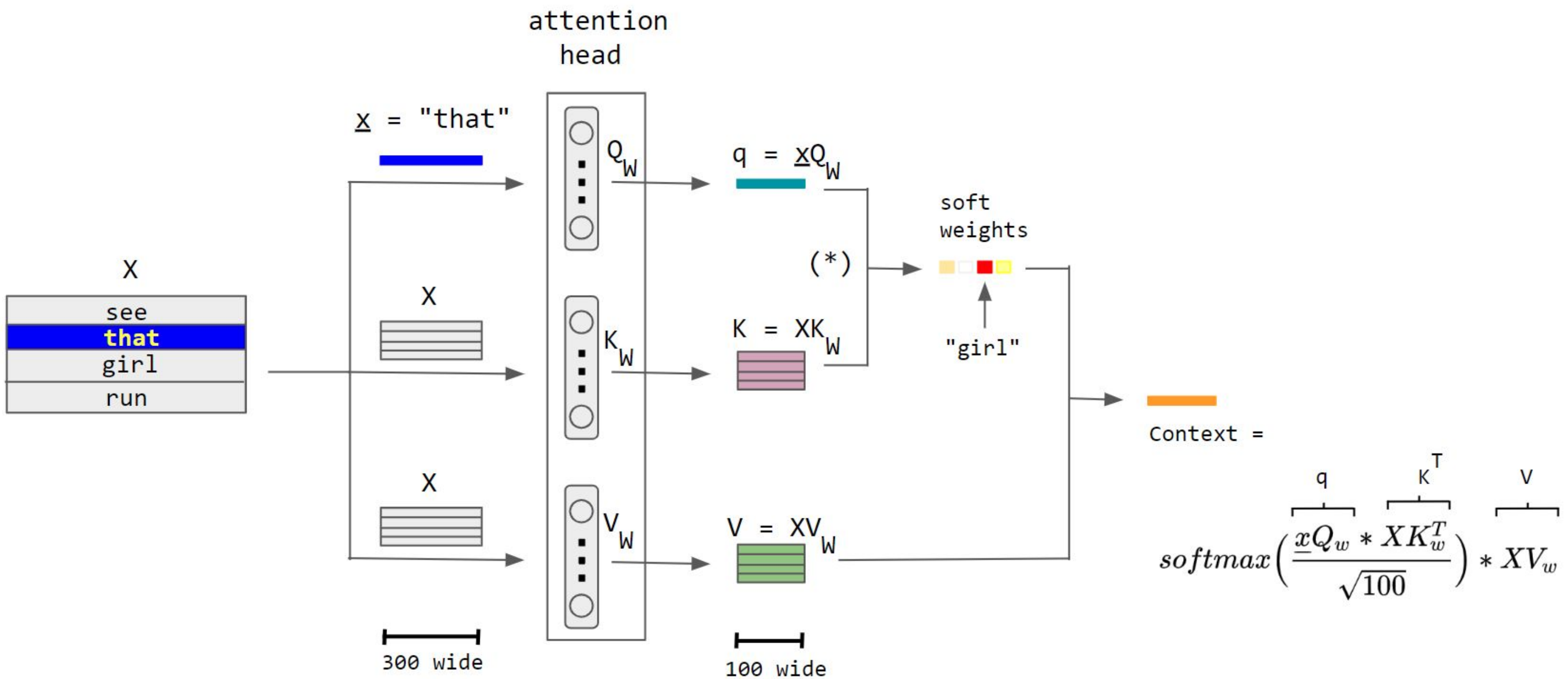
L’attenzione come prodotto di matrici apprese!



L’attenzione come prodotto di matrici apprese!



L'attenzione come prodotto di matrici apprese!

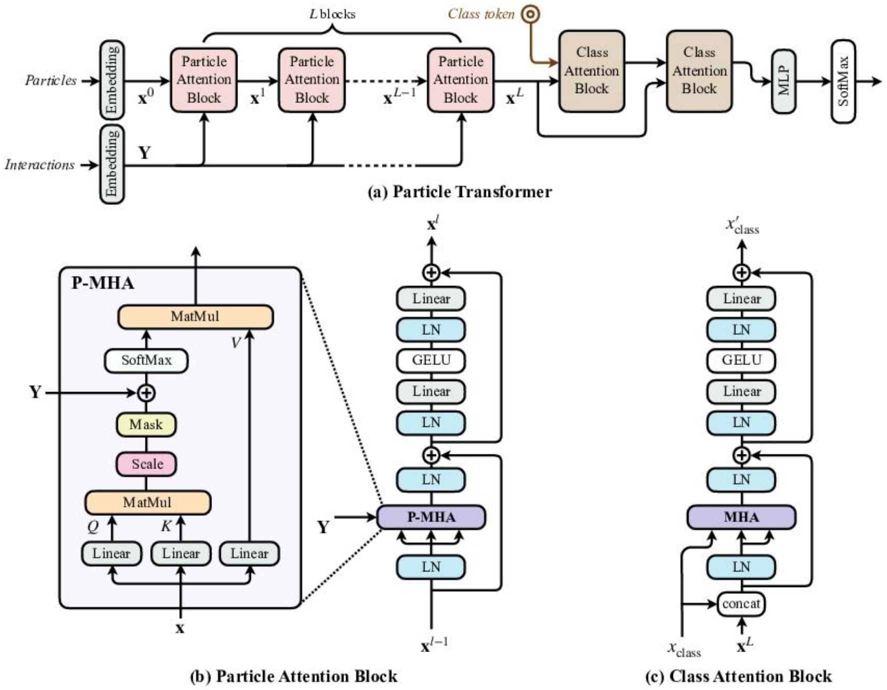


Un nuovo strumento (Transformer) dalle molte applicazioni

Qualsiasi cosa puo’ essere vista come una sequenza e passata a questi modelli!

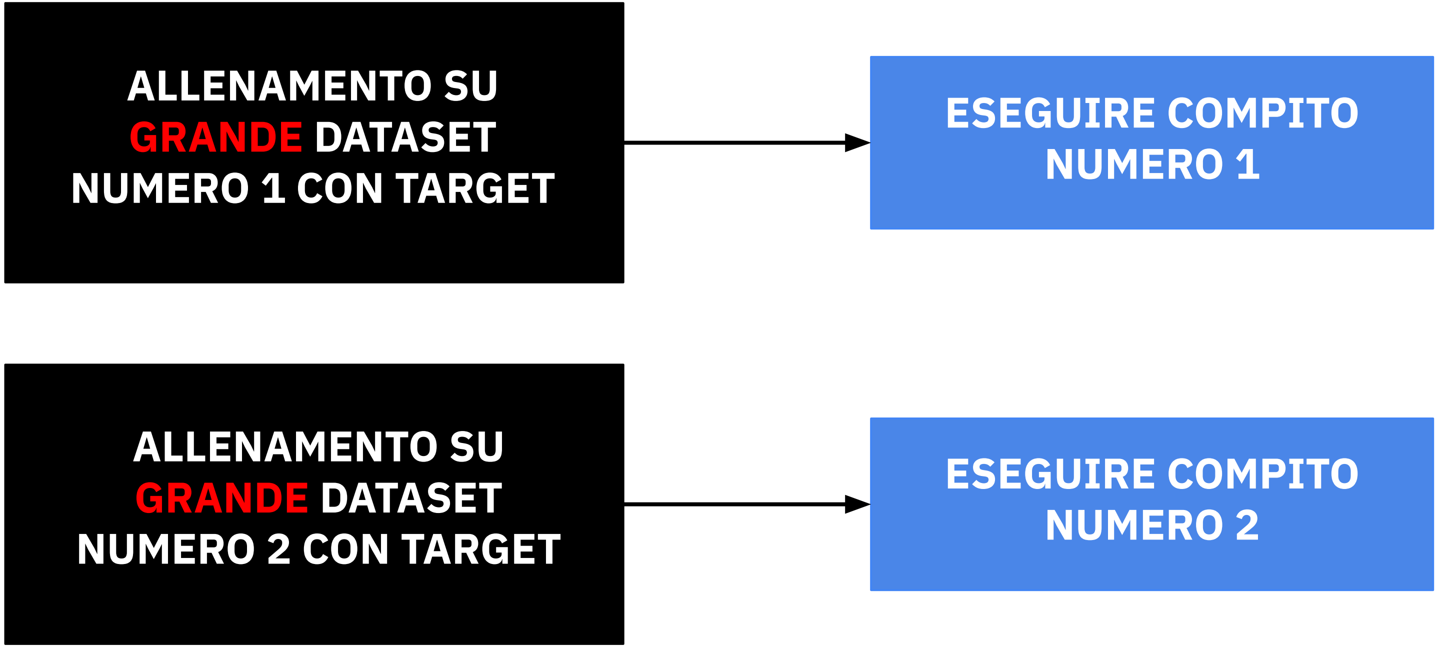
Tutta una serie di problemi diversi puo’ ora essere approcciata con la stessa architettura

Calorimetria, Ricostruzione, Mitigazione del Pileup, Tagging, Tracciamento…



Immagini da https://cds.cern.ch/record/2802264/plots

Dobbiamo ancora allenare su problemi diversi!



Riusciamo a trovare le 2 differenze?

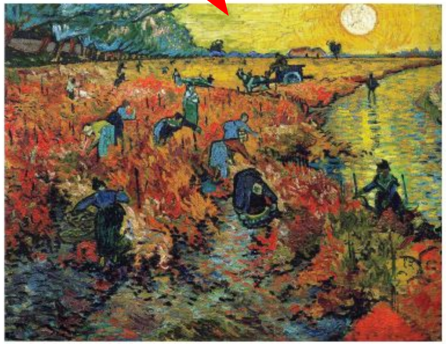
Riusciamo a trovare le 2 differenze?

Cosa abbiamo imparato?

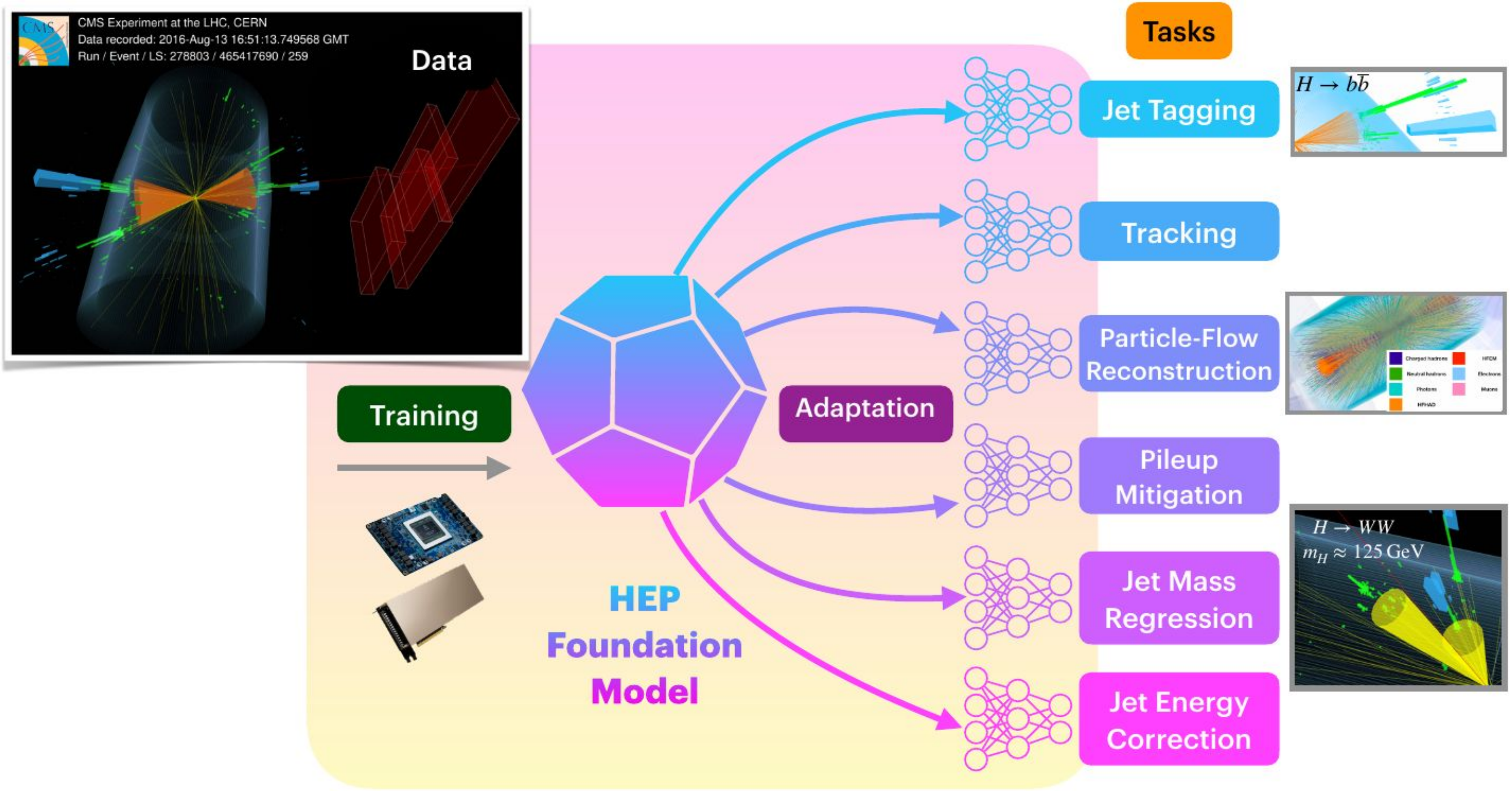
Ora sappiamo anche distinguere i van Gogh da tutti gli altri stili!

Nel confrontare due esempi simili, abbiamo cominciato a imparare qualcosa nello spazio di *tutti* i dipinti!

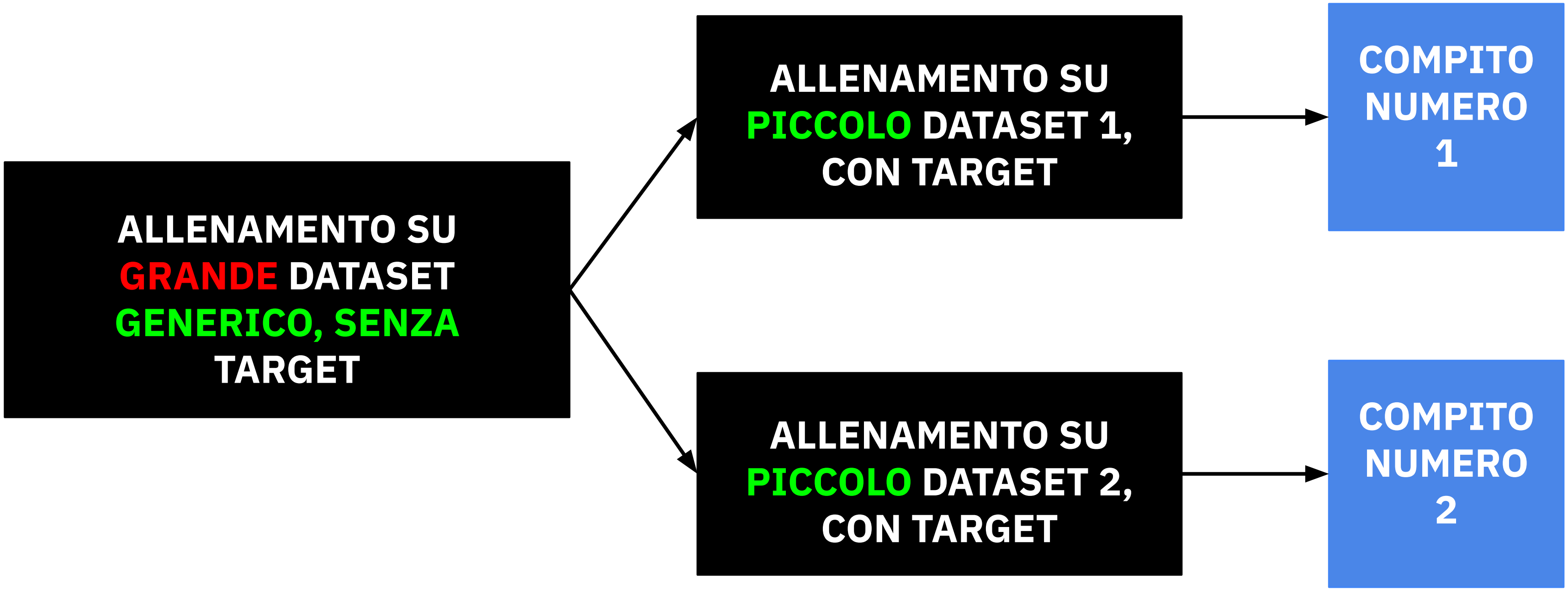
Come ripropongo la stessa cosa per la fisica?



Foundation models per le alte energie!



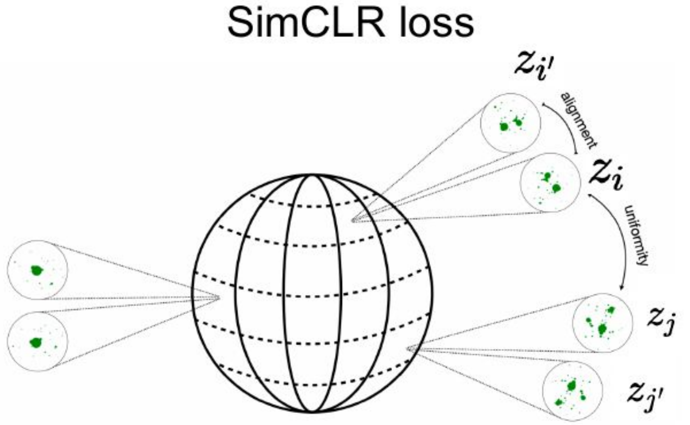
Un cambio di paradigma per ottimizzare!



Ci serve una nuova funzione di costo per imparare lo spazio dei dati!

Ci serve una funzione di costo che crei lo spazio confrontando due esempi alla volta

Vogliamo esempi simili vicini, differenti lontani



$$\mathcal{L}_i = -\log \frac{e^{s(z_i, z_i')/\tau}}{\sum_{j \neq i \in \text{batch}} \left[e^{s(z_i, z_j)/\tau} + e^{s(z_i, z_j')/\tau} \right]}$$

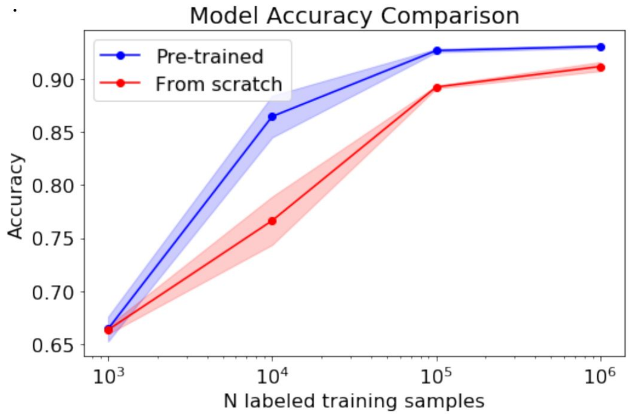
2108.04253

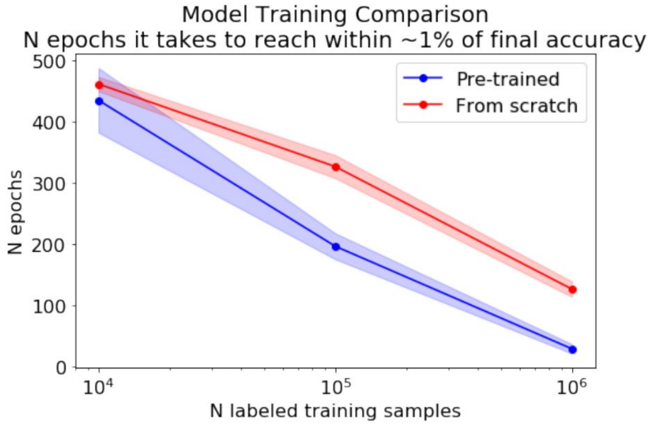
Un’applicazione al jet tagging

Si allena un Transformer su 100 Milioni di Jets senza label da 10 classi diverse

Si perfeziona su 1.2 Milioni di Jet da Top o da QCD per imparare a discriminare tra i due

Performance migliori e in meno tempo!





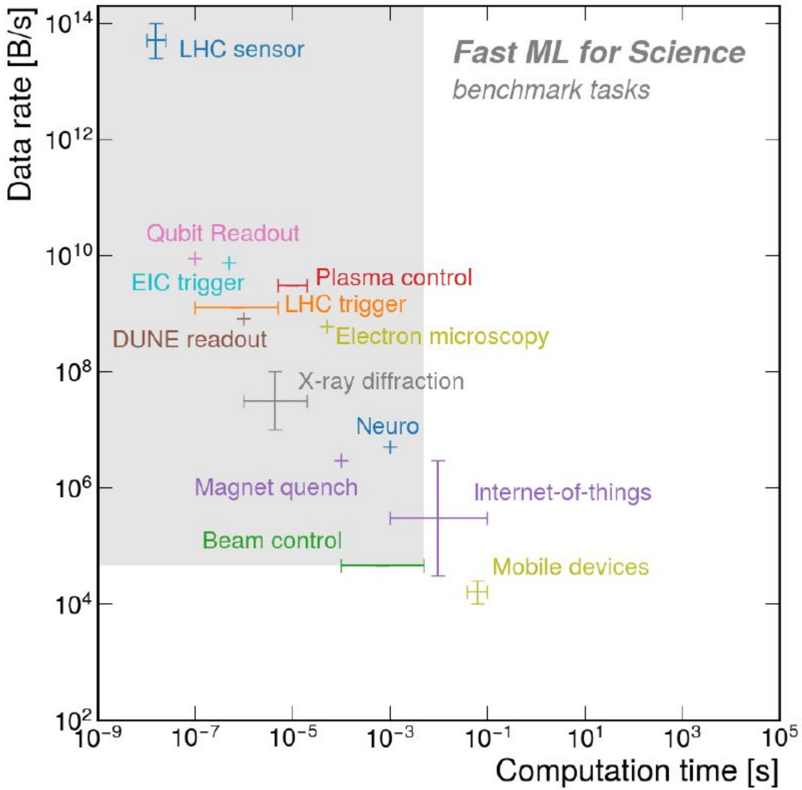
Da Zihan Zhao, Self-Supervised Learning (SSL) for Jet Tagging

La fisica non e' l'industria!

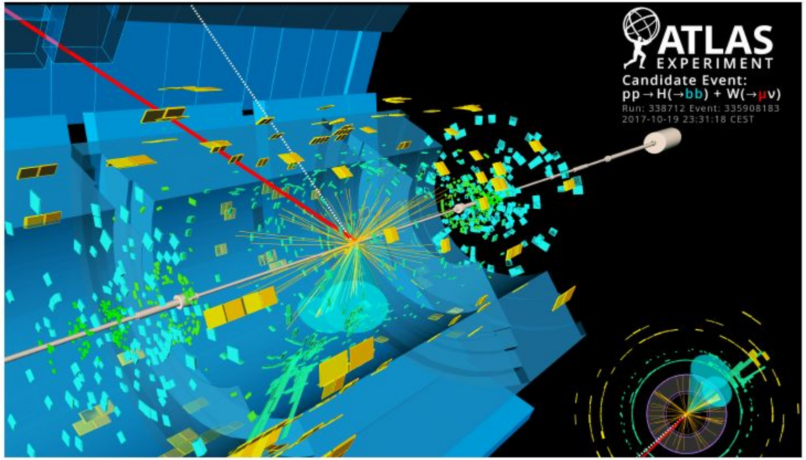
Necessita' e scale differenti

La velocità e' un fattore chiave che spesso non ha eguali negli altri domini!

Non tutte le architetture sono adatte a tutti gli scenari!



Vogliamo già delegare tutto a un unico modello?

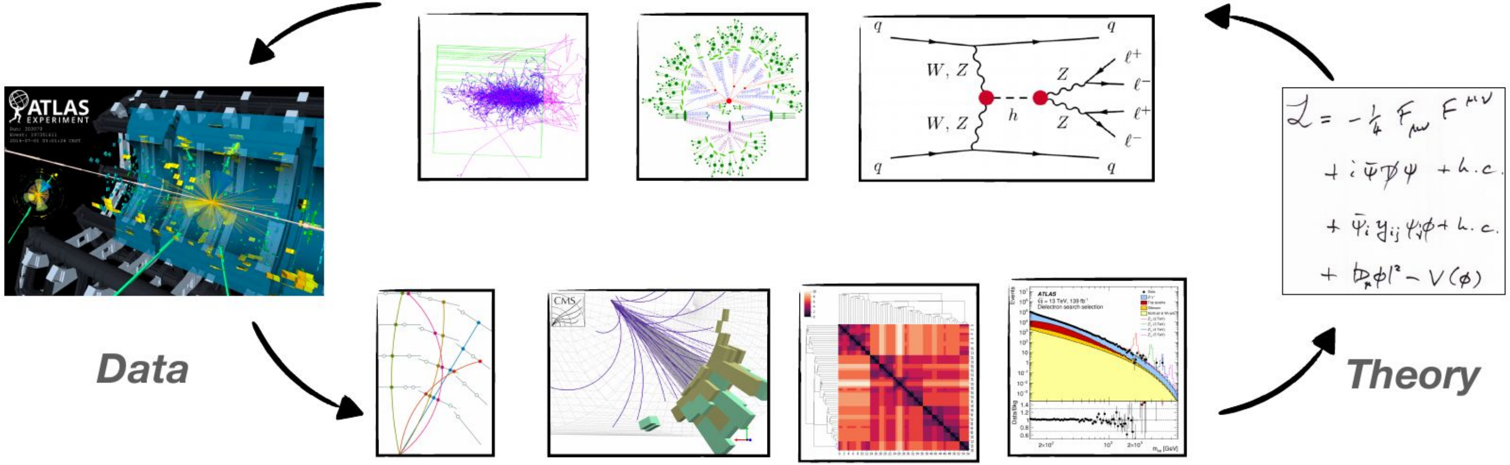


Transformer or whatever comes next.

AI: Trust me,
This is a Higgs Decay

Human: Ok.

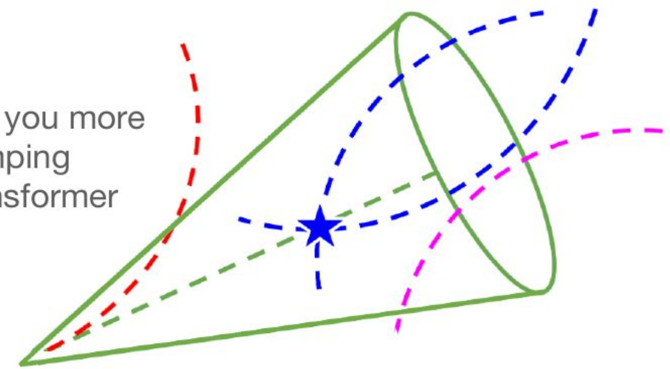
La catena gerarchica dei nostri dati e' alla base della comprensione fisica

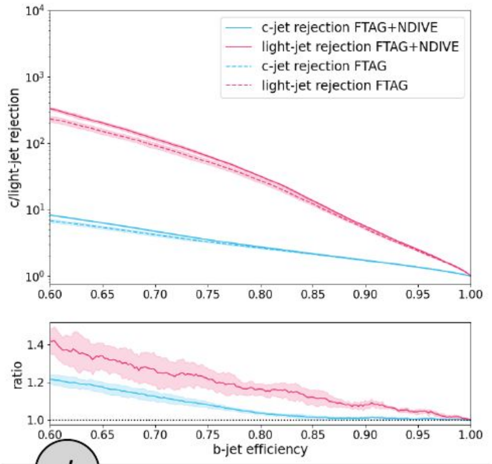


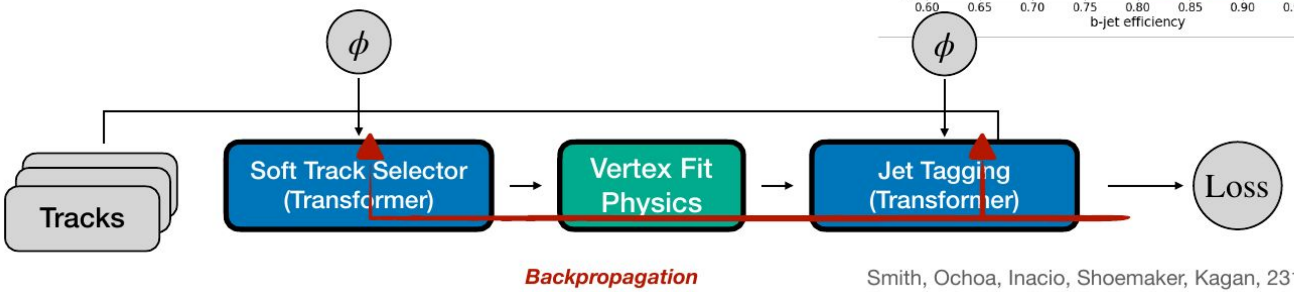
Forse non conviene delegare tutto a un unico modello...

Ci sono esempi dove l’informazione fisica fa ancora la differenza

Adding Physics **does give** you more performance than just dumping the raw data into a big transformer

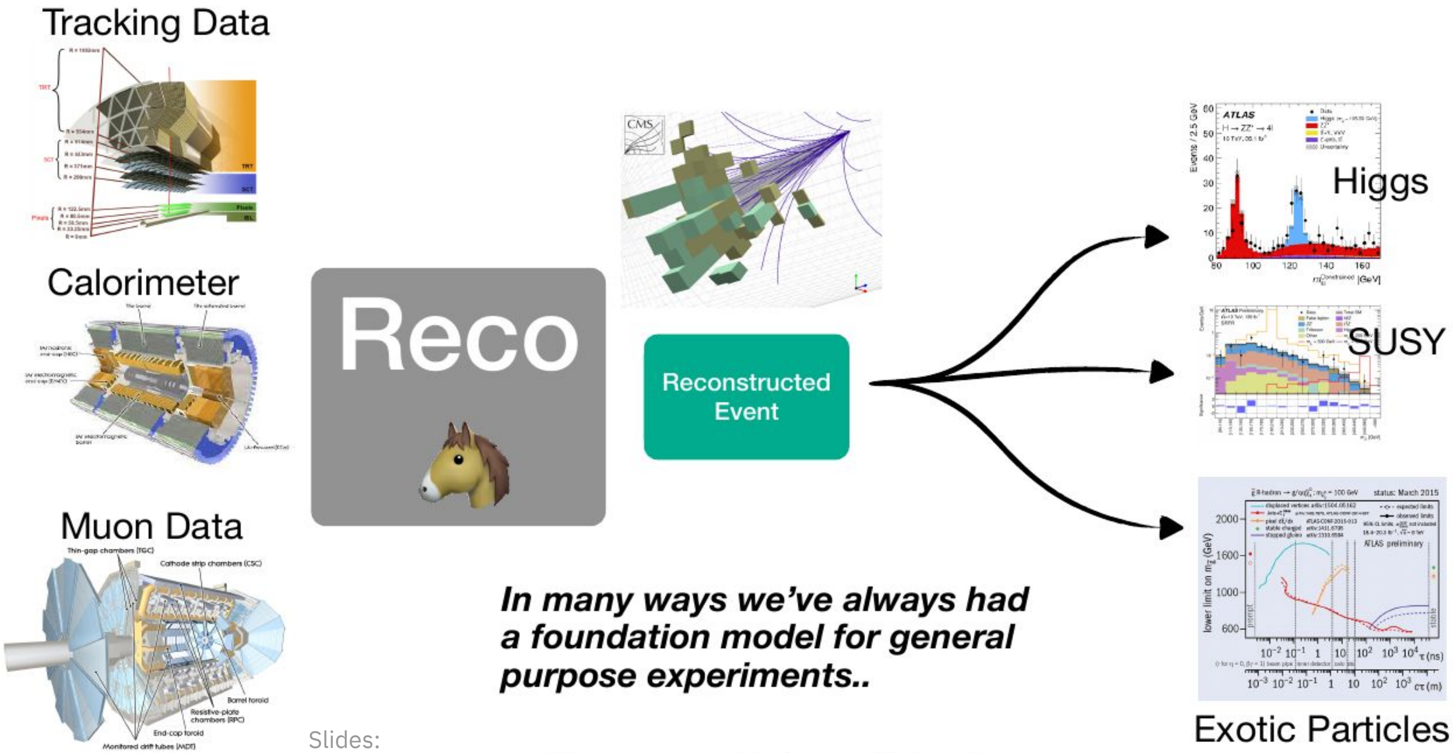






Smith, Ochoa, Inacio, Shoemaker, Kagan, 2310.12804

Per certi versi, abbiamo già dei foundation models!



Slides: https://indico.cern.ch/event/1202995/contributions/5241156/attachments/2744152/4774403/HN_2023_Talk.pdf

Conclusioni

Sono tempi stimolanti per il ML nelle alte energie

Possiamo sfruttare e innovare le soluzioni sviluppate negli altri domini

L’attenzione e i Transformers stanno soppiantando le altre soluzioni – ma sono abbastanza veloci?

I foundation models sono una possibilità – ma e’ la soluzione giusta per il nostro campo? Il modo in cui lavoriamo e’ già allineato con questi strumenti, ma serve cautela per andare avanti!

francesco.vaselli@cern.ch

Referenze

Immagini e idee raccolte da:

- Wikimedia,
- Lilian Weng, OpenAI,
- Lucas Beyer, lbeyer@google.com,
- Phil Harris(MIT), Towards the construction of Foundational Models at the LHC,
- Zihan Zhao, Self-Supervised Learning (SSL) for Jet Tagging
- Lukas Heinrich, TUM, Differentiable Programming in HEP