

ML-INFN

Tommaso Boccali (INFN Pisa / CERN)



Outline

- Perché?
- Cosa?
- Come?
- Chi?
- Quanto?

Perche'?

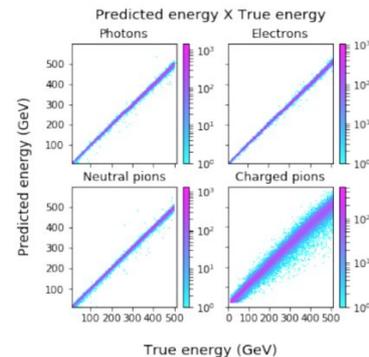
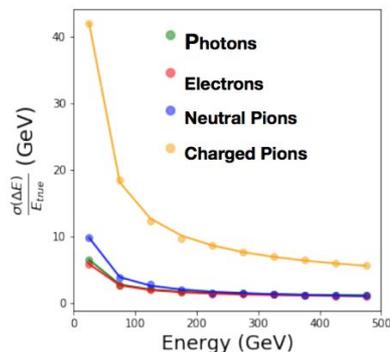
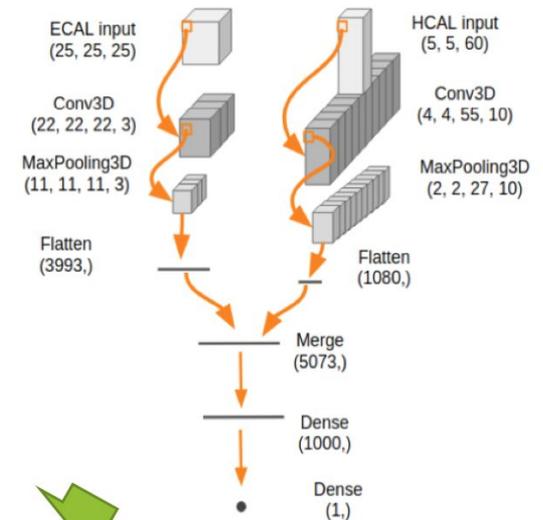
- Tecnologie ispirate alla Artificial Intelligence (AI) stanno diventando ubiquie nella ricerca di fisica
- Tools di AI possono variare da
 - “semplici tools” come quelli presenti in ROOT
 - Boost Decision Trees
 - Multivariate analysis
 - A sistemi Machine Learning → Deep Learning estremamente complessi
- In alcuni ambiti (p.e. Btagging a LHC) sistemi ML sono gia' adesso migliori (+ veloci, +performanti) di tools standard

Alcuni esempi

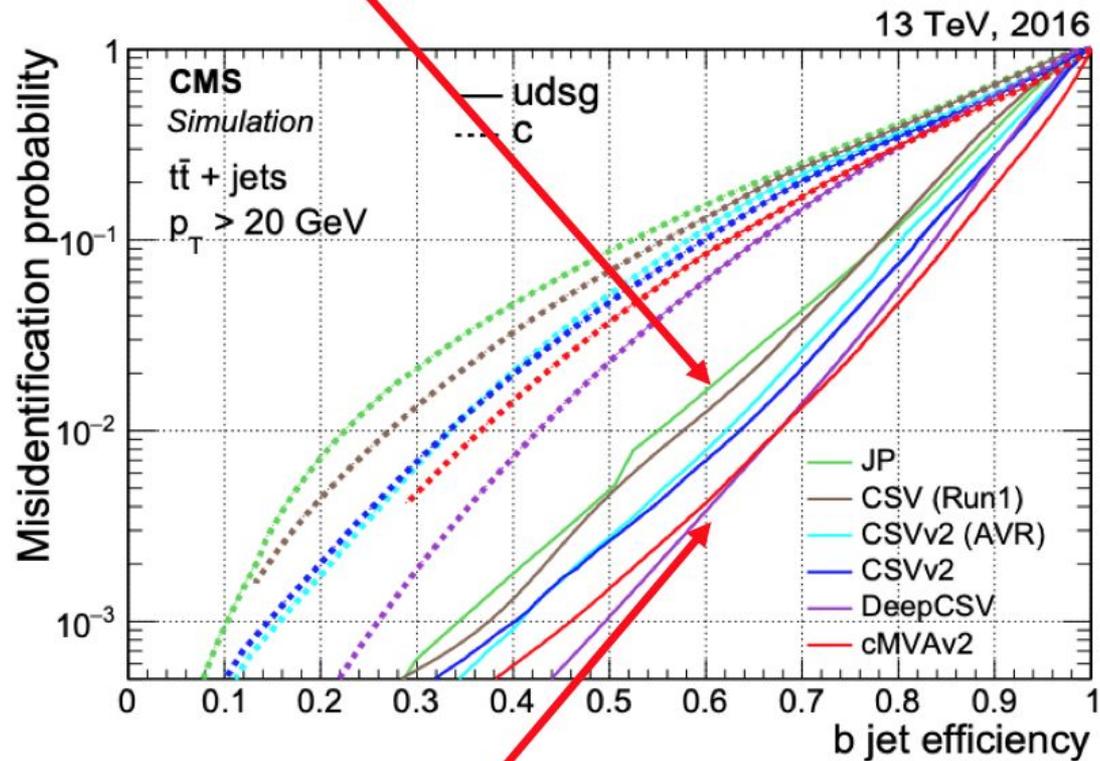
Jet energy Regression ...

CMS

- ▶ ... or how to measure the energy of the initial particle from the “image on calorimeters” (2 for CMS: ECAL and HCAL)
- ▶ Start with images (3rd dimension is time): so more a **movie** than an **image**
- ▶ Convolution layers do sampling and feature discovery
- ▶ MaxPooling reduces the parameter space
- ▶ The rest is mostly a **dense layer** with 1000 neurons
- ▶ Being a classical DNN (no recursion, no loops) its **timing is deterministic, and 1000x faster than standard approaches**



Typical classical algorithm:
60% efficiency for 50x rejection

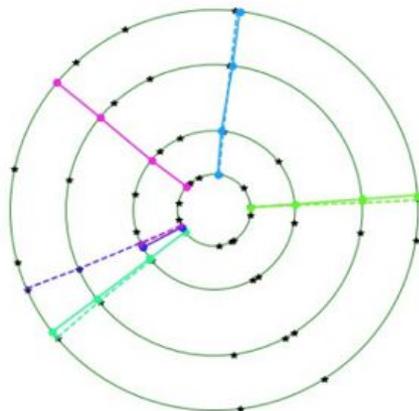


categorization

DeepCSV and other AI based
algorithms: 60% efficiency
for 300x rejection

Pattern Recognition with LSTM

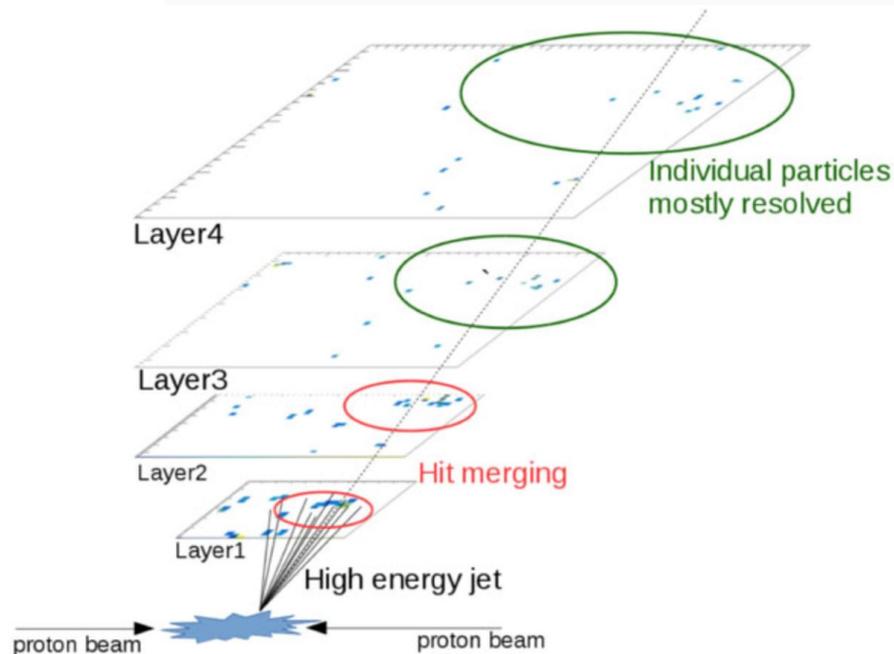
- Input sequence of hits per layers (one sequence per layer)
 - One LSTM cell per layer
- Output sequence of hits per candidates
 - Final LSTM runs for as many candidates the model can predict

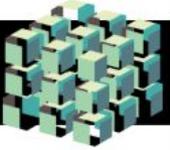


- Still work in progress
- Restricted to 4 layers (with seeding in mind)
- Work to some extent

Recurrent networks

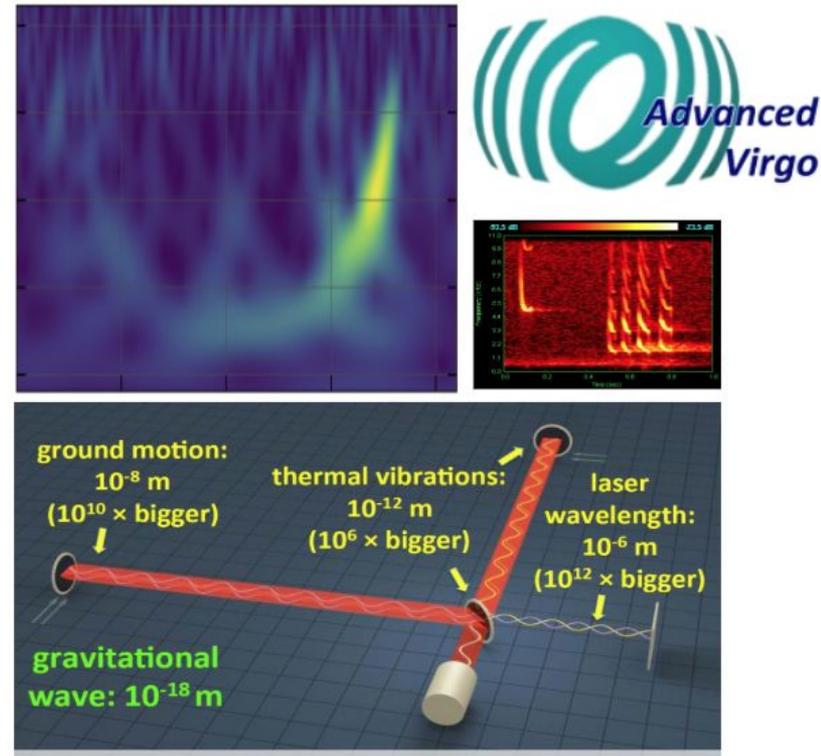
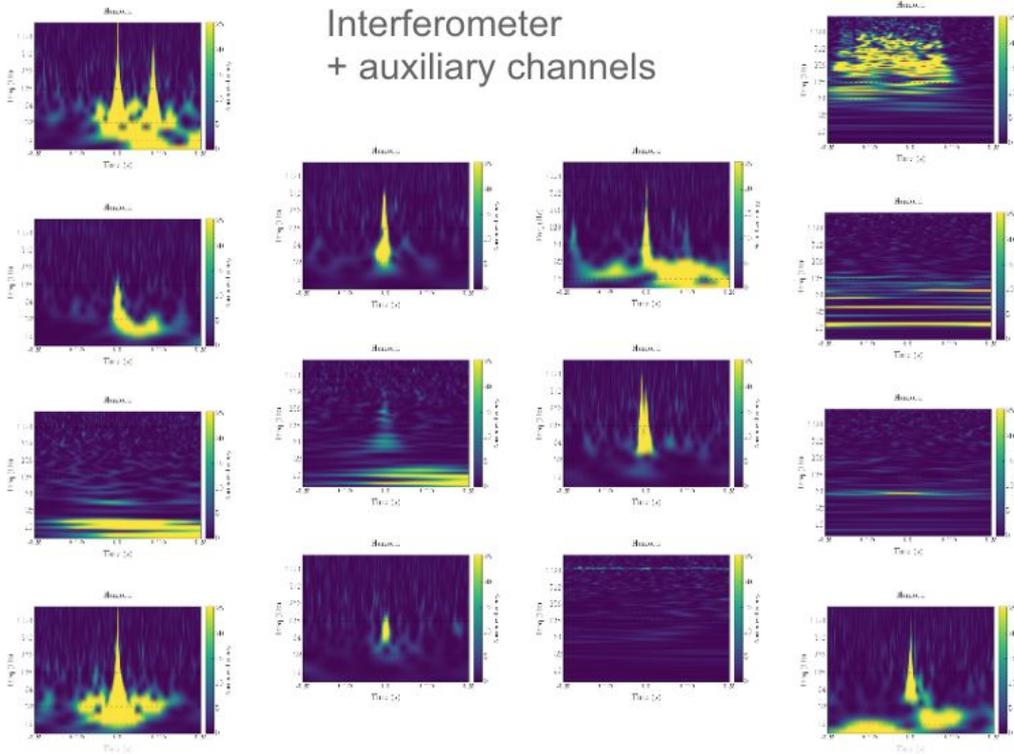
CMS Pisa





GW: transient noise detection

complex time-frequency patterns: multiple noise sources, loss of interesting signals

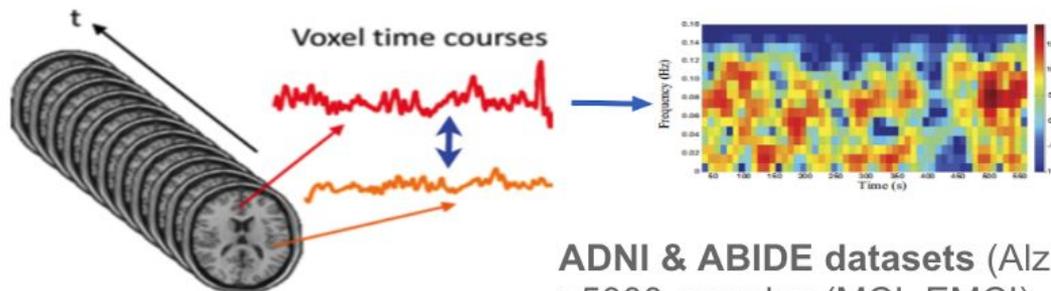
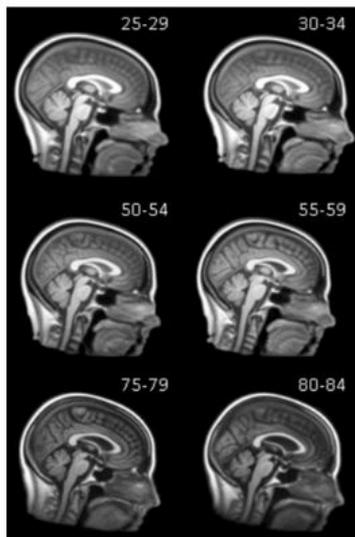


Current efforts focus on a mix of ML and distributed human vetting applied to time-frequency spectrograms (gravity SPY).

BIO: sample/brain region classification

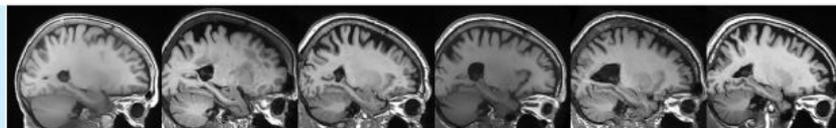


No clear background state > bad discrimination of mild and early



ADNI & ABIDE datasets (Alzheimer, Autism)
>5000 samples (MCI, EMCI)
MR (10^5 voxel), DTI, fMRI

Improved classification (confounders)
Early diagnosis
Disease progression tracking



IXI dataset (Apr 15)
>600 **healthy** samples
MR, DTI



Cosa?

- I ricercatori che utilizzano / vogliono utilizzare tecnologie AI devono (in ordine sparso)
 - Capire quale sia quella adeguata al loro problema
 - Sistemi ispirati a analisi immagini? A comprensione del linguaggio?
 - Categorizzazione / selezione o regressione?
 - Trovare / studiare i tools software adeguati
 - ROOT? Keras? Caffee? Theano? Tensorflow?
 - Trovare una piattaforma hardware adeguata
 - Quanta RAM?
 - Quanto disco veloce?
 - CPU / GPU?
 - Utilizzo proprietario / shared?

Come?

- Cercare una sistematizzazione delle possibilita' offerte ai ricercatori e una risposta guidata alle domande della pag precedente, nella forma di "Quale Soluzione?"
 - a. Creazione di Knowledge basis basate su soluzioni gia' in uso in INFN, elevate a "template" per gli utenti e messe a disposizione via Twiki etc; p.e.
 - "Problema di pattern recognition su immagini?"
 - → Feature detection in Fisica Medica
 - → cluster finding @ LHC
 - "Problema di analisi di serie temporali continue?"
 - → tools basati su analisi del linguaggio per anomaly detection a Virgo
 - → LSTM per Jet Reconstruction a LHC
 - "Problema di classificazione"?"
 - → GW discovery a Virgo
 - → bTagging a LHC
 - → diagnosi basata su features in fisica medica
 - "Ricerca di anomalie?"
 - → anomaly detection a LHC
 - → analisi dei logs a LHC
 - → noise @ Virgo

Come? #2

- b. Utilizzo delle esperienze presenti in INFN per Hackatons, Tutorial etc
- c. Technologies SW → nei template appena descritti
- d. Tecnologie HW? Utilizzo delle soluzioni già presenti in INFN (EOSC-Hub, Indigo-DataCloud, IDDL, ESCAPE, XDC, MSCA, ...) per la realizzazione di un'infrastruttura Cloud che sia
 - a. Basata su Servizi @CNAF su Cloud
 - b. Multi regione: permetta di inserire risorse GPU anche fuori dal CNAF
 - c. Unica: unico portale / accesso per richiedere e utilizzare risorse
 - d. Veloce: grande cache SSD davanti alle macchine, connesse allo storage di almeno Tier-1/Tier-2 per affettuare training con alto IO
 - e. Granulare: l'utente deve poter "liberamente" specifica quante risorse usare, per quanto tempo e di quale tipo

Chi?

- Hardware
 - Un sistema fascia “alta” (80kE) + disco SSD (40kE)
 - Un gettone di aggiornamento tecnologico per il secondo e il terzo anno
 - Devono essere un’infrastruttura INFN → finanziamento dalla Giunta
- 2 tipi distinti di persone:
 - Tecnologi che gestiscano l’infrastruttura sia CNAF sia Regionale
 - Proposta 1 FTE per servizi centrali + 0.5 FTE per gestire l’infrastruttura in ogni sito che entri nella Cloud
 - 1 FTE che pensi alla realizzazione, mantenimento e sistematizzazione delle KB e dei tutorial
 - Devono essere un’infrastruttura INFN → finanziamento dalla Giunta
 - Ricercatori che aiutino a elevare i loro algoritmi a “template”, e che aiutino nel training
 - Pagate su Missioni CSN5; preposte 2 settimane/anno per ognuno dei templates e per la partecipazione a training

Struttura (temporanea)

- WP1: “Infrastruttura”
 - Sviluppo e gestione delle risorse su Cloud, loro provisioning etc
- WP2: “Training e KB”
 - Gestione e mantenimento delle risorse condivise; organizzazione e gestione del training
- WP3: “Scientific use cases”
 - Raccogliere studi esistenti nella comunita', elevarli a “template” e partecipare al training

Lista temporanea dei “template”

1. Convolutional neural networks:
 - a. Immagini mediche
 - b. Applicazioni a HEP (pixel clustering, calorimeter clustering)
 - c. Mappe tempo frequenza dei segnali gravitazionali
2. Time series:
 - a. Long short-term memory (LSTM) su dati di Virgo
 - b. Analisi di immagini funzionali in Risonanza Magnetica
3. Algoritmi: esempi di GraphNet
 - a. Sostituzione di algoritmi classici con graphnet per tracking, particle flow, jet clustering HEP
 - b. Analisi di causalità in sistemi dinamici complessi (ambito detector characterizatio VIRGO)
4. Modelli supervisionati: selezione/classificazione; esempio di feed forward NN
 - a. Selezione delle posizioni ottimali di sensori sismici per la mitigazione del rumo Newtoniano (VIRGO)
 - b. Categorizzazione di jets per flavour in HEP
 - c. Categorizzazione di immagini mediche sane/con patologia
 - d. Classificazione dei segnali transienti
5. Modelli non supervisionati: categorizzazione e compressione; esempi di auto encoders:
 - a. Data compression in HEP
 - b. Data quality monitoring di detectors
 - c. Segnali anomali per nuova fisica
 - d. Virgo (per anomaly detections)

Quanto?

- Siamo ancora in alto mare, i numeri attuali dicono

	I anno	II anno	III anno	TOTALE
Hardware	120	20	20	160
Personale a contratto	91	91	91	273
Missioni	52	52	52	156
	263	163	163	589

Giunta

CSN5

Lista delle persone attualmente interessate (notare “attualmente”)

Attualmente il focus e' + sulla parte infrastrutturale che su quella scientifica ...

Il numero di persone interessate a quest'ultima e' enorme, e la lista qui a dx e' destinata a aumentare
(la richiesta di fare sigla csn5 e' della settimana scorsa ...)

	Nome	Sede
1	Alessandra Retico	PI
2	Andrea Chincarini	GE
3	Tommaso Boccali	PI
4	Elena Cuoco	EGO
5	Ian Harms	GSSI
6	Francesco Giacomini	CNAF
7	Stefano Dal Pra	CNAF
8	Elisabetta Ronchieri	CNAF
9	Luca Rei	GE
10	Enrico Peira	GE
11	Diego Bersanetti	GE
12	Stefano Bagnasco	TO
13	Federica Legger	TO
14	Sara Vallero	TO
15	Daniele Spiga	PG
16	Mirco Tracoli	PG
17	Daniele Bonacorsi	BO
18	Stefano Giagu	RM1
19	Francesco Conventi	NA