ML_INFN

Un approccio end-to-end all'utilizzo del Machine Learning per le linee di ricerca INFN

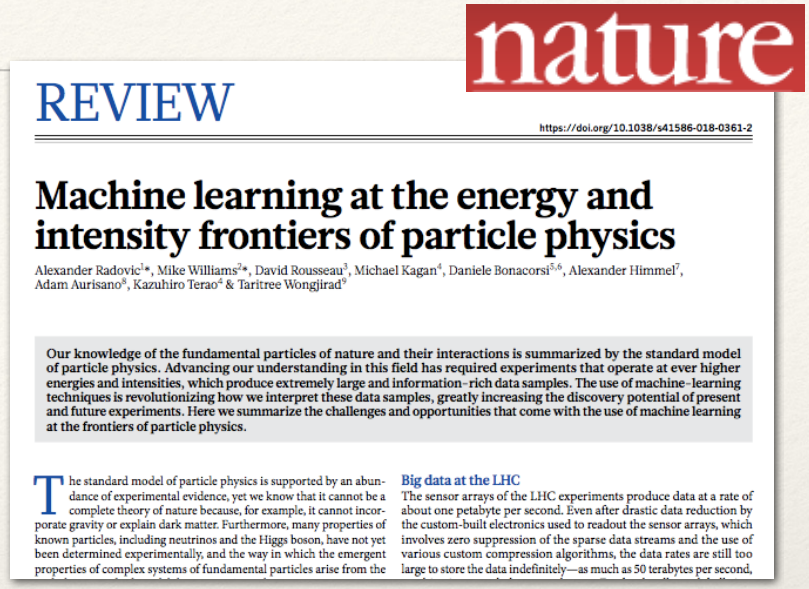
Daniele Bonacorsi
(a nome del gruppo ML_INFN)
crediti per il materiale: T. Boccali (INFN-Pisa)

Assemblea di Sezione INFN-Bologna - 4 Febbraio 2021

ML in HEP

ML ormai pervasivo in HEP.

Dall'adozione di modelli relativamente semplici (BDT) al design di reti neurali con architetture anche relativamente sofisticate: **whatever works!**


nature

REVIEW

https://doi.org/10.1038/s41586-018-0361-2

Machine learning at the energy and intensity frontiers of particle physics

Alexander Radovic[1]*, Mike Williams[2]*, David Rousseau[3], Michael Kagan[4], Daniele Bonacorsi[5,6], Alexander Himmel[7], Adam Aurisano[8], Kazuhiro Terao[4] & Taritree Wongjirad[9]

Our knowledge of the fundamental particles of nature and their interactions is summarized by the standard model of particle physics. Advancing our understanding in this field has required experiments that operate at ever higher energies and intensities, which produce extremely large and information-rich data samples. The use of machine-learning techniques is revolutionizing how we interpret these data samples, greatly increasing the discovery potential of present and future experiments. Here we summarize the challenges and opportunities that come with the use of machine learning at the frontiers of particle physics.

The standard model of particle physics is supported by an abundance of experimental evidence, yet we know that it cannot be a complete theory of nature because, for example, it cannot incorporate gravity or explain dark matter. Furthermore, many properties of known particles, including neutrinos and the Higgs boson, have not yet been determined experimentally, and the way in which the emergent properties of complex systems of fundamental particles arise from the

Big data at the LHC

The sensor arrays of the LHC experiments produce data at a rate of about one petabyte per second. Even after drastic data reduction by the custom-built electronics used to readout the sensor arrays, which involves zero suppression of the sparse data streams and the use of various custom compression algorithms, the data rates are still too large to store the data indefinitely—as much as 50 terabytes per second,

https://www.nature.com/articles/s41586-018-0361-2

Non sono idee "nuove".
Perchè un boost recente? (non solo HEP!)

- ingenti moli di dati, agilità "tecnica" nel loro accesso
- iterative gradient descent, etc.. → GPGPU, TPU
- frameworks di alto livello, open source, support community, interfacce ai principali linguaggi di programmazione

Stack tipico di un ML practitioner di oggi:

- PyTorch + fast.ai, oppure Keras + Tensorflow-GPU, o analoghi

"From zero to hero in ML" come buzzword: ma è davvero così facile?

Come divento "operativo" con il ML? *[1/2]*

Qualsiasi ricercatore che voglia (genericamente) "usare tecniche di ML" deve:

Valutare quale sia "**un approccio ML adeguato**" al problema

- a livello di task: classificazione? regressione? clustering? altro?
- a livello di approccio: simile alla computer vision? simile al processamento del linguaggio? altro?

Identificare il sw e dotarsi dei tools adeguati nell'ecosistema

- a livello dei dati da studiare: e.g. ROOT? DBs?
- a livello dei ML framework: Tensorflow? PyTorch? Keras? Caffee? fast.ai? MXNET? h20.ai? altro?

Reperire l'hw adeguato e correttamente dimensionato

- Quanta RAM?
- Quanto disco? e quanto veloce?
- CPU? GPU? altro?
- Utilizzo singolo? risorse proprietarie? shared?
- Configurazione dell'accesso ai dataset per il training? Complicazioni legate ai dati (e.g. privacy)?

Come divento "operativo" con il ML? *[2/2]*

Valutare l'approccio ML adeguato

Training "generale" su ML, e best practices in **Applied ML**, disponibilità online di una grande mole di "esempi" e di una comunità open-source molto vasta. Al limite, una "guida" per evitare rischio di disorientamento (e/o perdita di tempo)

Identificare il sw e dotarsi dei tools

Training "specifico" su ML con hands-on. Esistono esempi documentati di set-up e implementazioni a cui ispirarsi, ma l'efficacia dipende dall'esperienza: la disponibilità di esperti INFN per discussioni di persona sarebbe un asset

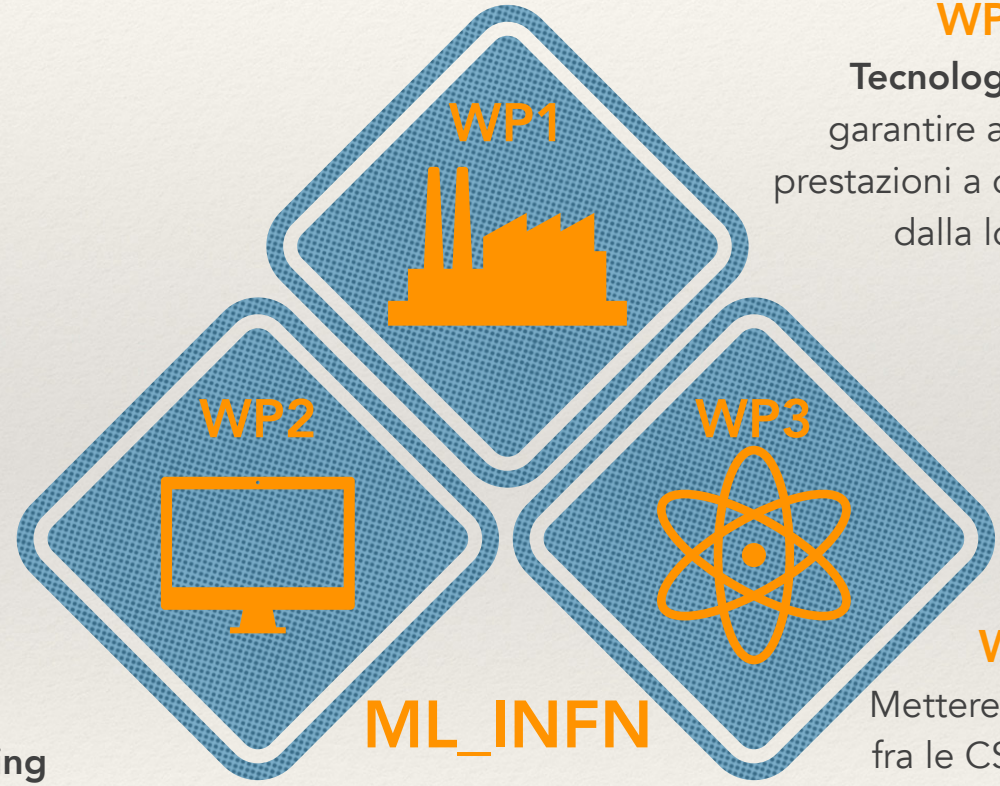
Reperire l'hw

Acquisti solo per casi o molto semplici, o molto complessi e dunque specifici e ben motivati. Per il grosso bulk degli use-cases, l'accesso "democratico" alle risorse INFN, slegato dalla contingenza geografica e dalle eventuali disponibilità locali, sarebbe un asset

La risposta di ML_INFN

Obiettivo di **ML_INFN** è dare una risposta efficace a ciascuno di questi punti, investendo su due punti di forza: 1) ottimizzazione delle risorse condivise su infrastrutture solide, e 2) razionalizzazione delle competenze presenti nell'ente.

Un progetto articolato in 3 WPs:



WP1 - Infrastruttura

Tecnologie informatiche in grado di garantire agli utenti un accesso ad alte prestazioni a dati remoti, indipendentemente dalla loro locazione geografica

WP2 - Formazione (stewardship)

Accesso a opportunità di **training** a tutti i livelli, inclusi hand-on intensivi con realizzazione di sistemi completi e funzionanti

WP3 - Casi scientifici

Mettere a sistema le conoscenze sparse fra le CSN in una **knowledge base** (KB) categorizzata per tecnologia, per guidare utenti in studi di ML con codice esistente e possibilità di contattare gli autori

Organizzazione

Periodo: 2020/21/22

Finanziamento:

- 2020: 40 k€ HW + missioni (COVID → tutto restituito)
- 2021: 20 k€ HW + 27 k€ missioni (?)

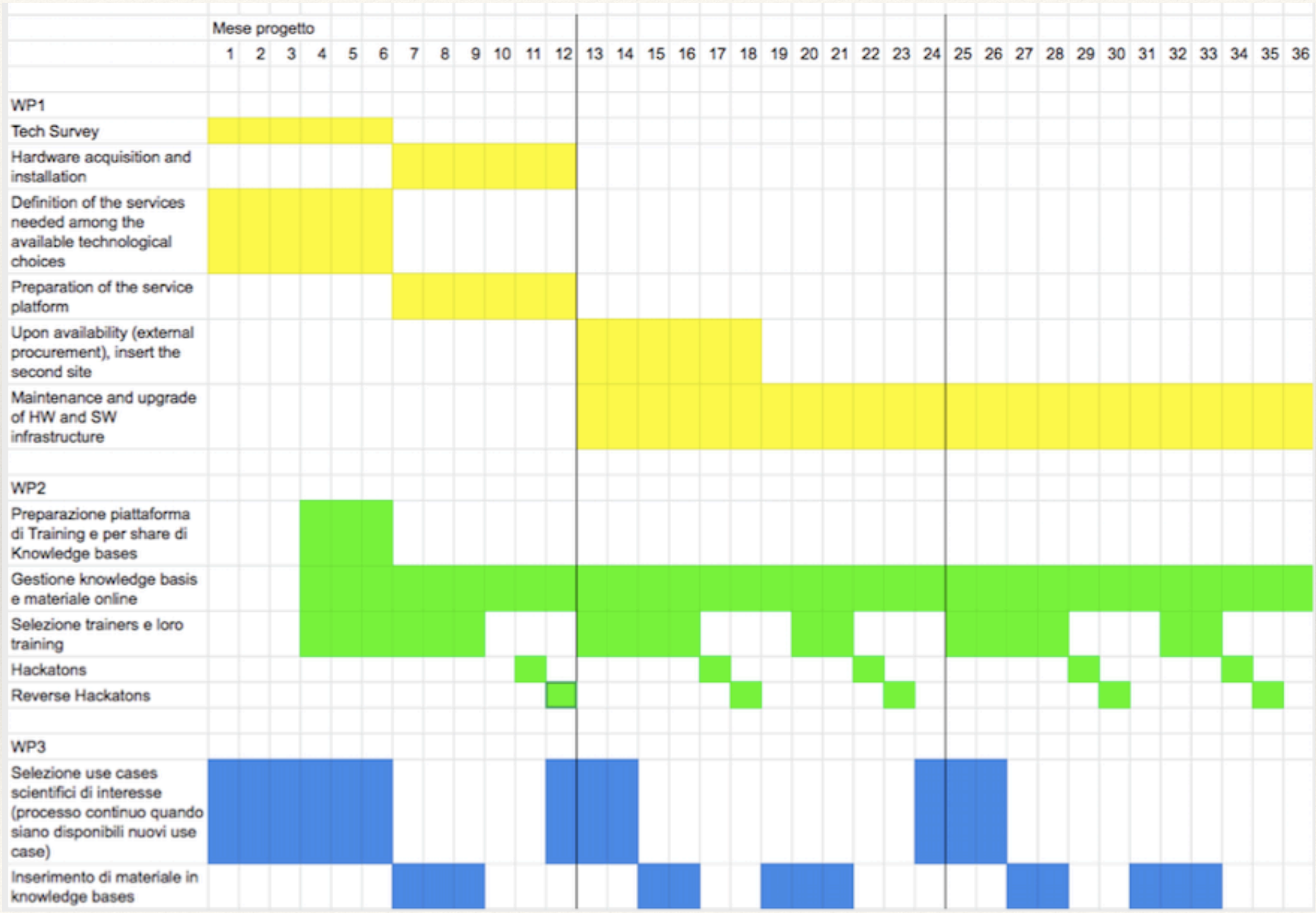
PI: T. Boccali (INFN Pisa)

Sedi partecipanti: **BA**, **BO**, **CNAF**, **FI**, **GE**, **NA**, **PD**, **PG**, **PI**, **RM1**, **TO** (+ **Univ di PI, FI, GE, PD, PG, Roma1, BO, NA, EGO, SNS**)

Persone: 76, per circa 12 FTE

- come “da design", molti contributori con ricerca non assorbita in ML_INFN contribuiscono alle attività con frazioni anche basse (e.g. 10%) di FTE
- singole sezioni più grandi sono FI, PD, PI; BO da sola tra le più piccole; se si conta BO+CNAF, diventa la più grande
  - Bonacorsi, Castellani, Clissa, Diotalevi, Giommi, Grandi, Remondini, Rinaldi (+ 10 persone CNAF)

Gantt


Mese progetto
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
WP1
Tech Survey
Hardware acquisition and installation
Definition of the services needed among the available technological choices
Preparation of the service platform
Upon availability (external procurement), insert the second site
Maintenance and upgrade of HW and SW infrastructure
WP2
Preparazione piattaforma di Training e per share di Knowledge bases
Gestione knowledge basis e materiale online
Selezione trainers e loro training
Hackatons
Reverse Hackatons
WP3
Selezione use cases scientifici di interesse (processo continuo quando siano disponibili nuovi use case)
Inserimento di materiale in knowledge bases

WP1 - Infrastruttura

INFN protagonista negli ultimi anni di progetti su transizione Grid→Cloud

- progetti EU: INDIGO-Datacloud, EOSC-*, ESCAPE, XDC, ..
- progetti nazionali: PON Sud, CloudPD, IPCEI, …

Architettura originale proposta era quella di una Cloud geograficamente distribuita, con accesso democratico per ogni utente INFN

- Inizialmente basata su risorse CNAF, ma poi allargabile a risorse ReCaS, IBiSCo, ..
- NB: questo era precedente alla nascita di INFNCloud. Da allora, convergenza totale.

Grande attenzione alla modalità di accesso ai training data: garanzia di accesso a dati remoti "come fosse locale" via protocolli AAI-safe

- tecnologie drop-in a-la Xcache (XDC, ESCAPE, IDDLS-CNS5)

"Voglio N macchine con 32 GB di RAM, 48 cores e 2 GPU V100 per i prossimi 5 mesi, con accesso a bassa latenza ai dati VIRGO residenti a Torino (100 TB). La macchina deve essere accessibile a questo gruppo di utenti"

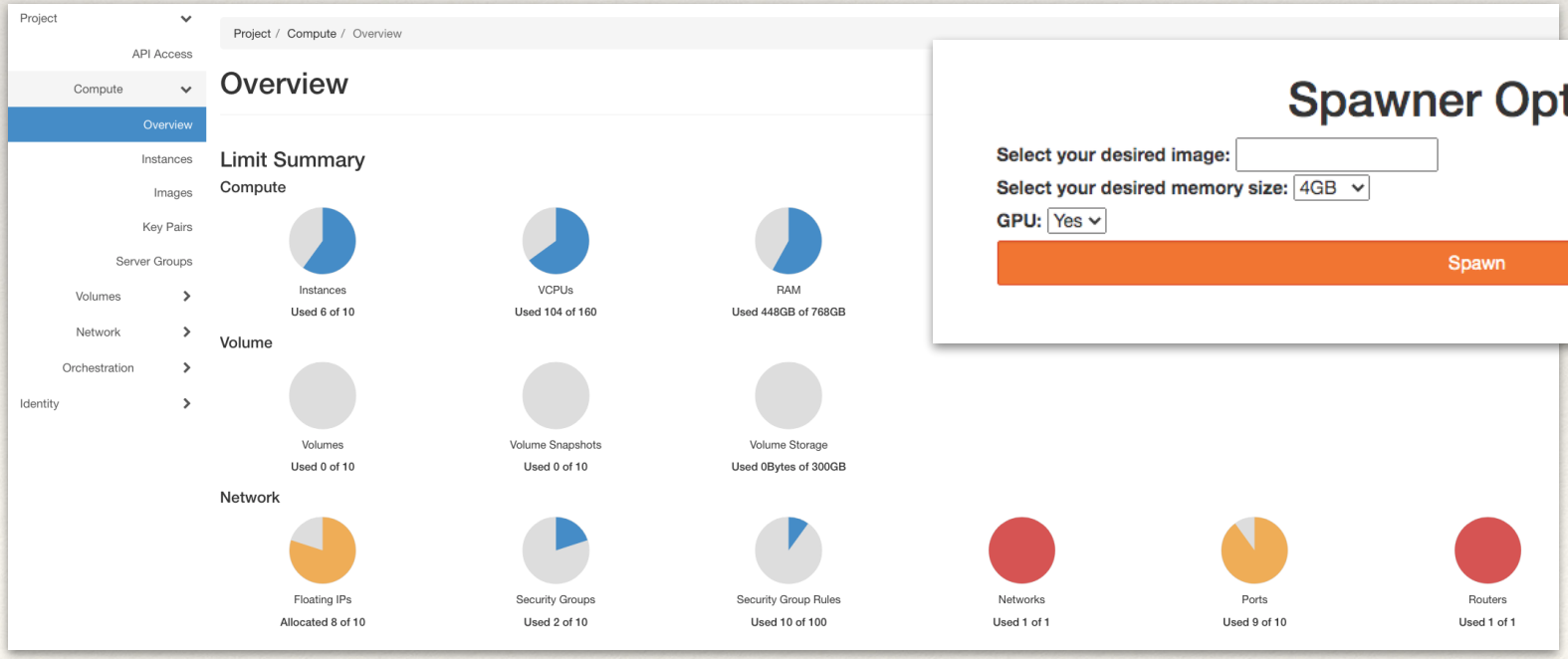
WP1 - Infrastruttura

Stato attuale

Hardware scelto, acquisito e online

- 6 Nvidia T4 + 36 TB NVme al CNAF
- attualmente operativo sotto Cloud@CNAF, a breve integrazione con INFNCloud
- entro i primi 6 mesi del 2021 volontà di inserire altre risorse "non finanziate" (PD, BA e TO sembrano in prima fila)

Test di macchine virtuali in corso da ~8 mesi



WP1 - Infrastruttura

Stato attuale

Full monitoring già disponibile per le macchine, e.g.:

Monitoring - **node** view

Monitoring - **GPU** view

Lavoro di preparazione per "gruppi di lavoro" via Jupyter hub quasi finito

- possibilità per un gruppo di lavoro di fare share di risorse ("macchina di gruppo"), integrata con AAI, includendo la possibilità di accesso a storage “multiplo”: Pandora, Dropbox, GoogleDrive, OneDrive, cernbox

WP2 - Training

Tagliare i tempi di ingresso proficuo del ML nelle attività di ricerca INFN, fornendo supporto alle singole ricerche

Iniziative di training (anche molto mirato) a vari livelli:

- **Hackathons**: 3-4 giorni full immersion, 100% hands-on, ricercatori e studenti vengono guidati alla realizzazione di use case end-to-end, dalla concettualizzazione all'implementazione
- **Reverse Hackathons**: i partecipanti a un primo hackathon (grosso) diventano tutor a un secondo hackathon (più piccolo) successivo: spinta ad approfondire le loro conoscenze, e successivamente viene stimolato il training reciproco
  - NB: emersa di recente come soluzione molto efficace per training pervasivo in ambito HEP
- **Webinars** (incluse partecipazioni remote / registrazione degli hackathon). Gli stessi esperti, sempre più coadiuvati dal personale formato, gestiscono anche una community di online presence, su cui discutere problemi, approcci, soluzioni
- Armonizzazione di iniziative spontanee e non coordinate di training sul ML all'INFN: workshop, seminari, tutorial, lezioni, .. WP2 intende creare un catalogo, fare revisione del materiale, dare indicazioni a utenti INFN interessati (quale training scegliere, cosa sapere prima, ..)

WP2 - Training

Stato attuale

Non abbiamo effettuato l'Hackathon 2020 (previsto nel Gantt) per ovvi motivi

- spostato al primo semestre 2021, in modalità online only

Stimolati dalla situazione contingente, avviata ricerca di tools adeguati a fare (anche) queste attività solo online

- e.g. Zammad al momento per il tutoraggio su lavoro di gruppo

Procede il lavoro di censimento delle esistenti iniziative di training e la loro integrazione organica nel progetto

WP3 - Casi Scientifici

Le ricerche ML-based all'INFN restano negli ambiti di origine, e il progetto ML_INFN non le "assorbe" né intende sostituirsi agli esperimenti che le realizzano, ma intende **facilitarne l'implementazione**.

Aspetto caratterizzante del WP3 è la **raccolta e messa a sistema delle esperienze esistenti nell'INFN**, mediante un apporto dei singoli ricercatori alle attività di training e alla realizzazione di **knowledge base (KB)** documentate e corredate da esempi completamente funzionanti.

Dalle iniziali manifestazioni di interesse, le macro aree emerse erano relativamente vaste e diversificate

- sia modelli supervisionati che modelli non supervisionati, architetture come Convolutional Neural Networks e Graph Neural Networks, nonché applicazioni per serie temporali (TS), trattamento di dati testuali (NLP), ..
- diversi livelli di maturazione delle attività: alcune convertibili in occasioni di training (entries nella KB) già nel primo anno, altre entreranno successivamente

WP3 - Casi Scientifici

Stato attuale

Conv NN: Applicazioni a HEP (pixel clustering, calorimeter clustering) → **INFN-PI (CMS)**
Clustering e analisi di log → **INFN-NA (CMS)**
Failure prediction and predictive maintenance → **INFN-BO/CNAF (Tier1)**
Data quality monitoring di detectors → **INFN-BA (CMS)**
Categorizzazione di jets per flavour in HEP →**INFN-PI (CMS)**
Applicazioni a HEP (pixel clustering, calorimeter clustering) → **INFN-NA (ATLAS)**
Predicting Dataset popularity with in data management → **INFN-BO (CMS)**
Monitoring data transfers → **INFN-BO (CMS)**
Applicazioni a HEP (pixel clustering, calorimeter clustering) → **INFN-ROMA1 (ATLAS)**
Categorizzazione di jets per flavour in HEP → **INFN-ROMA1 (ATLAS)**
Applicazioni a HEP (pixel clustering, calorimeter clustering) → **INFN-LE (ATLAS)**
Conv NN: modelli di regressione e classificazione Immagini mediche → **INFN-PI INFN-BA (AIM)**
Estrazione di features radiomiche da immagini mediche e classificazione → **INFN-FI (AIM)**
Long short-term memory (LSTM) su dati di Virgo → **INFN-GE (VIRGO)**
Utilizzo di NN per la mitigazione del rumore Newtoniano → **INFN-GE (VIRGO)**
Reti evolutive per l'ottimizzazione posizioni sensori sismici attorno a masse di test → **INFN-GE (VIRGO)**
Autoencoder su dati EEG → **INFN-GE (AIM)**
ML-pipelines con Spark → **INFN-PD**
Uso di Big Data tools per il processamento in real-time di data stream da DT di CMS → **INFN-PD (CMS)**
Algoritmi di ML nell'area di likelihood free inference, semi-/un-supervised searches → **INFN-PD**
Reti neurali profonde quantizzate e ultra-veloci su FPGA per applicazioni real-time → **INFN-RM1**
AE, VAE e InfoVAE per de-noising supervisionato e auto-supervisionato in imaging medico ad alto rumore → **INFN-RM1**

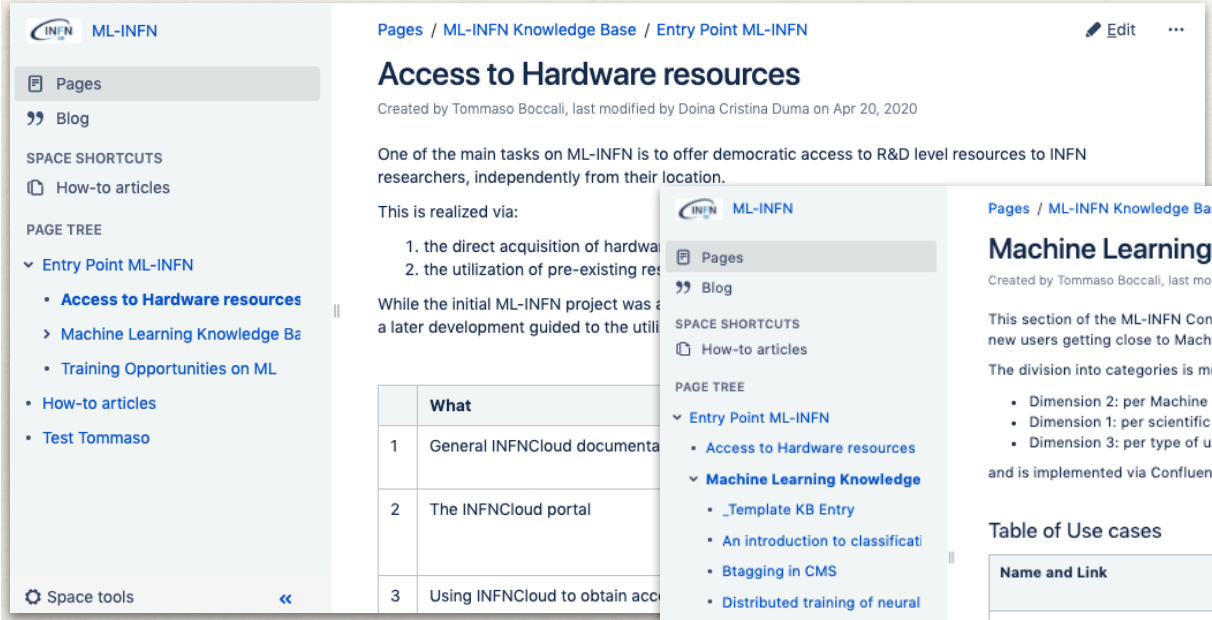
(… and more ….)

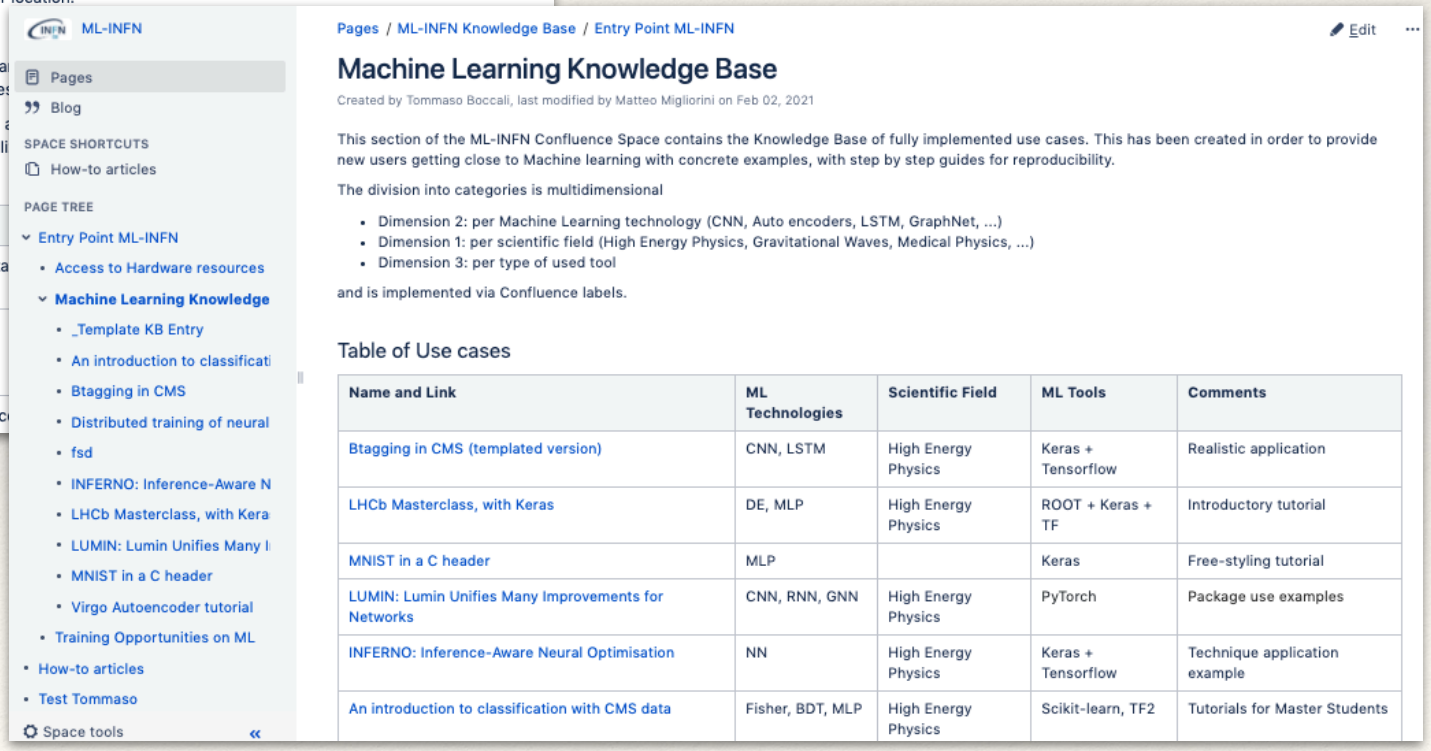
WP3 - Casi Scientifici

Stato attuale

Stato della Knowledge Base (KB) → visibile su Confluence INFN

- https://confluence.infn.it/display/MLINFN/Entry+Point+ML-INFN





Gli use-case già presenti saranno revisionati e usati per l'hackathon 2021

Conclusioni

ML_INFN sta lavorando per **facilitare e velocizzare l'accesso a tecniche di ML da parte di ricercatori coinvolti nelle linee di ricerc INFN**, mettendo a disposizione dell'ente competenze sul ML e risorse di calcolo esistenti, in modo efficace:

- condividendo le **risorse** di qualità (altrimenti difficili/impossibili da reperire) mediante tecnologia Cloud, democraticamente disponibili a tutti gli utenti
- creando opportunità di **training** a vari livelli e su varie tecnologie ML
- facendo confluire conoscenze e skill presenti nell'ente sul ML in attività specifiche di interesse dell'INFN in una **struttura documentale viva e consultabile**, in cui convivono codice, documentazione, e link attivi a persone di riferimento

crediti per il materiale: T. Boccali (INFN-Pisa)