



ALLEGATO A

Programma di training per infrastruttura di ricerca acquisita su fondi PNRR [SoBigData.it](https://www.sobigdata.it)

Periodo didattico: 3 mesi (14 settimane)

Impegno docenti: 3 corsi da 40 ore + 1 corso da 20 ore (140 ore totali). Lezioni: circa 2 pomeriggi a settimana (sessioni da 5 ore ciascuna), per 3 mesi e mezzo.

Obiettivo generale: fornire formazione operativa e riproducibile per ricercatori e tecnici sull'uso efficace delle macchine acquistate con il finanziamento, mettendo a disposizione competenze trasversali di Machine Learning, pratica su PyTorch e metodo su PyTorch Lightning per training distribuito e riproducibilità. Applicazione diretta della tecnologia a casi di studio reali con progetti imprenditoriali.

1) Struttura dei corsi (sintesi)

Istruttore 1 — Machine Learning (40 ore)

Abstract: corso teorico-pratico che copre i fondamenti delle machine learning moderno: formulazione dei problemi (classificazione, regressione, ranking), pipeline di preprocessing, valutazione (metriche e protocolli sperimentali), regolarizzazione, ottimizzazione, modelli lineari, alberi, ensemble, e introduzione ai modelli neurali per il deep learning. Si pone particolare attenzione a buone pratiche sperimentali, gestione dei dati e principi per rendere gli esperimenti riproducibili, includendo nozioni operative sull'uso delle macchine acquistate: copia e sincronizzazione del codice tramite GitHub sulle macchine acquistate, creazione e gestione di ambienti Python isolati (virtualenv/conda) per garantire replicabilità e coerenza delle installazioni software.

Istruttore 2 — PyTorch (40 ore)

Abstract: corso pratico mirato a insegnare l'ecosistema PyTorch: tensori e operazioni, costruzione di modelli custom, gestione dei dataset e dataloader, strategie di ottimizzazione, debugging, salvataggio e caricamento di checkpoint, tecniche di accelerazione (mixed precision), e integrazione con strumenti di monitoraggio (TensorBoard, Weights & Biases). Particolare attenzione sarà dedicata all'uso delle GPU delle macchine acquistate: gestione di device multipli in PyTorch, allocazione della memoria e best practice per il trasferimento dati CPU↔GPU, utilizzo di `torch.cuda` per profiling.

Istruttore 3 — PyTorch Lightning (40 ore)

Abstract: corso focalizzato su PyTorch Lightning come livello di astrazione per accelerare e rendere riproducibile il training su GPU: struttura di `LightningModule` e `Trainer`, checkpointing, callback, logging, strategie di schedulazione learning rate e orchestrazione di job (es. uso di script batch / scheduler). L'enfasi sarà posta sul corretto utilizzo delle GPU disponibili sulle macchine acquistate, in particolare sull'uso di più GPU all'interno di una singola macchina, con esempi pratici su configurazione, profiling e best practice per garantire stabilità ed efficienza degli esperimenti.



Istruttore 4 — Imprenditorialità digitale (20 ore)

Abstract: corso pratico per applicare le tecnologie apprese nei moduli precedenti a un caso di business, con il coinvolgimento di founder di startup tech. La prima settimana parte dall'introduzione al significato di fare startup, e si concentra sulla scelta ragionata di un ambito industriale al quale applicare la tecnologia, sull'analisi e del problema di un gruppo specifico di utenti e sull'identificazione della relativa soluzione offerta. La seconda settimana stimola i partecipanti a interagire realmente con i possibili clienti, iniziando a studiare i loro comportamenti e a proporre loro il concetto della soluzione.

2) Timeline didattica (proposta per 14 settimane — divisa per settimane e argomenti)

Nota logistica: 2 pomeriggi a settimana, sessioni da 5 ore (lezione + laboratorio). Le settimane si possono sovrapporre parzialmente tra i corsi.

Timeline 1 — Machine Learning (40 ore)

- **Settimana 1:** introduzione, setup ambienti, formulazione problemi ML.
- **Settimana 2:** metriche e pipeline dati, feature engineering.
- **Settimana 3:** modelli lineari, alberi, ensemble; introduzione reti neurali.
- **Settimana 4:** transfer learning, fine-tuning, data augmentation, consolidamento concetti e revisione.

Timeline 2 — PyTorch (40 ore)

- **Settimana 1:** setup PyTorch, tensori, autograd, modelli semplici.
- **Settimana 2:** dataset e DataLoader, preprocessing, checkpointing e logging.
- **Settimana 3:** mixed precision, gestione memoria GPU, ottimizzazione, debugging.
- **Settimana 4:** approfondimento modelli complessi (CNN, RNN, Transformer).

Timeline 3 — PyTorch Lightning (40 ore)

- **Settimana 1:** introduzione Lightning, `LightningModule/DataModule`, primo esempio.
- **Settimana 2:** Trainer, checkpoint, callback, logging.
- **Settimana 3:** multi-GPU su singola macchina, profiling, gestione batch.
- **Settimana 4:** consolidamento concetti, esercitazioni pratiche su training e logging.

Timeline 4 — Imprenditoria digitale (20 ore)

- **Settimana 1:** cosa significa fare startup, scelta del mercato a partire dalla tecnologia, identificazione del problema del cliente
- **Settimana 2:** customer discovery e validazione

3) Modalità pratiche e uso delle macchine acquistate

- **Accesso e ambiente:** i partecipanti useranno le macchine via SSH, con ambienti isolati e riproducibili (virtualenv/conda o Docker).
- **Gestione codice e materiali:** repository GitHub centrale con branch per ogni corso / esercitazione; notebook e script versionati.
- **Uso GPU e best practice:** esercitazioni pratiche sull'utilizzo di GPU, incluse



sessioni dedicate all'uso di più GPU su una singola macchina. Introduzione a gestione della memoria GPU e tecniche per evitare out-of-memory (profiling con `torch.cuda` e `nvidia-smi`), attivazione di mixed precision e suggerimenti operativi per avviare, monitorare e terminare job in modo sicuro.

- **Logging e monitoraggio esperimenti:** integrazione con TensorBoard o W&B per tracciare metriche.

4) Come la formazione si lega al progetto

1. **Capacity building e sostenibilità:** il training serve a far acquisire competenze agli utilizzatori delle macchine finanziate, massimizzandone l'utilizzo scientifico e la sostenibilità.
2. **Attività di formazione:** le risorse acquistate supportano l'attività di formazione stessa.
3. **Allineamento con obiettivi SoBigData++:** il programma enfatizza FAIR/FACT, riproducibilità e governance dei dati e facilita l'integrazione dei nodi locali nella rete europea.