Scavare nei dati

Il *data mining* nella fisica delle particelle

di Cristiano Bozza

a.
Se l'informazione è il petrolio del XXI secolo, i *big data* degli esperimenti sono le miniere della nuova fisica.

Il modello standard, nonostante i suoi successi, è una rete nella quale siamo intrappolati. Cerchiamo una smagliatura per venirne fuori e avere una visuale più ampia, ma per ora non si è trovata. Con il progredire delle tecniche sperimentali abbiamo accesso a misure di precisione che in passato non erano nemmeno proponibili, ma trovare l'inconsistenza, l'anomalia, il fenomeno imprevisto è tutt'altro che banale. Così come accade nello studio dell'infinitamente piccolo, andando lontano nell'infinitamente grande stiamo assistendo a una esplosione di dati. L'astronomia multimessaggera, che combina telescopi ottici, radiotelescopi, onde gravitazionali e neutrini, dà la caccia a correlazioni tra segnali flebili e nascosti nel rumore dei fenomeni banali. Nessuna strada può essere lasciata non battuta. Non è soltanto un imperativo metodologico, ma anche economico: gli esperimenti costano sia in termini di risorse finanziarie che umane. Sono vere e proprie fabbriche di dati, di una qualità e abbondanza mai viste prima, eppure cerchiamo l'ago nel pagliaio a volte senza sapere nemmeno che aspetto abbia.

Non siamo soli: per quanto la fisica delle particelle elementari abbia avuto un ruolo trainante in passato nell'evoluzione dell'informatica, oggi in molti campi i ricercatori fronteggiano montagne di dati. Si pensi ad esempio alle redditizie attività di profilazione degli utenti, per ottimizzare le campagne pubblicitarie, messe in campo dalle maggiori aziende informatiche, i cui introiti sono per oltre il 90% dovuti alla capacità di analizzare i dati e intraprendere azioni conseguenti. Se è vero che l'informazione e la conoscenza sono il petrolio del XXI secolo, il *data mining* (che potremmo tradurre con "scavare nei dati") è una delle attività da intraprendere tanto per le applicazioni quanto per la ricerca pura.

Come si "scava nei dati"? Evidentemente non si possono seguire solo logiche preordinate. Esistono correlazioni ovvie di cui tenere conto nella costruzione delle basi di dati, come le condizioni operative della strumentazione. Tuttavia, l'analisi dei dati può richiedere decenni dopo la fine della presa dati di un esperimento. I database devono rispecchiare questa flessibilità. Nell'approccio relazionale, le connessioni logiche sono precostituite, anche se possono evolvere con il tempo, ma sono di tipo predicativo-binario. Un dato "è in relazione" oppure "non è in relazione" con un altro. Invece gli algoritmi di analisi, così come le esigenze, cambiano nel tempo.

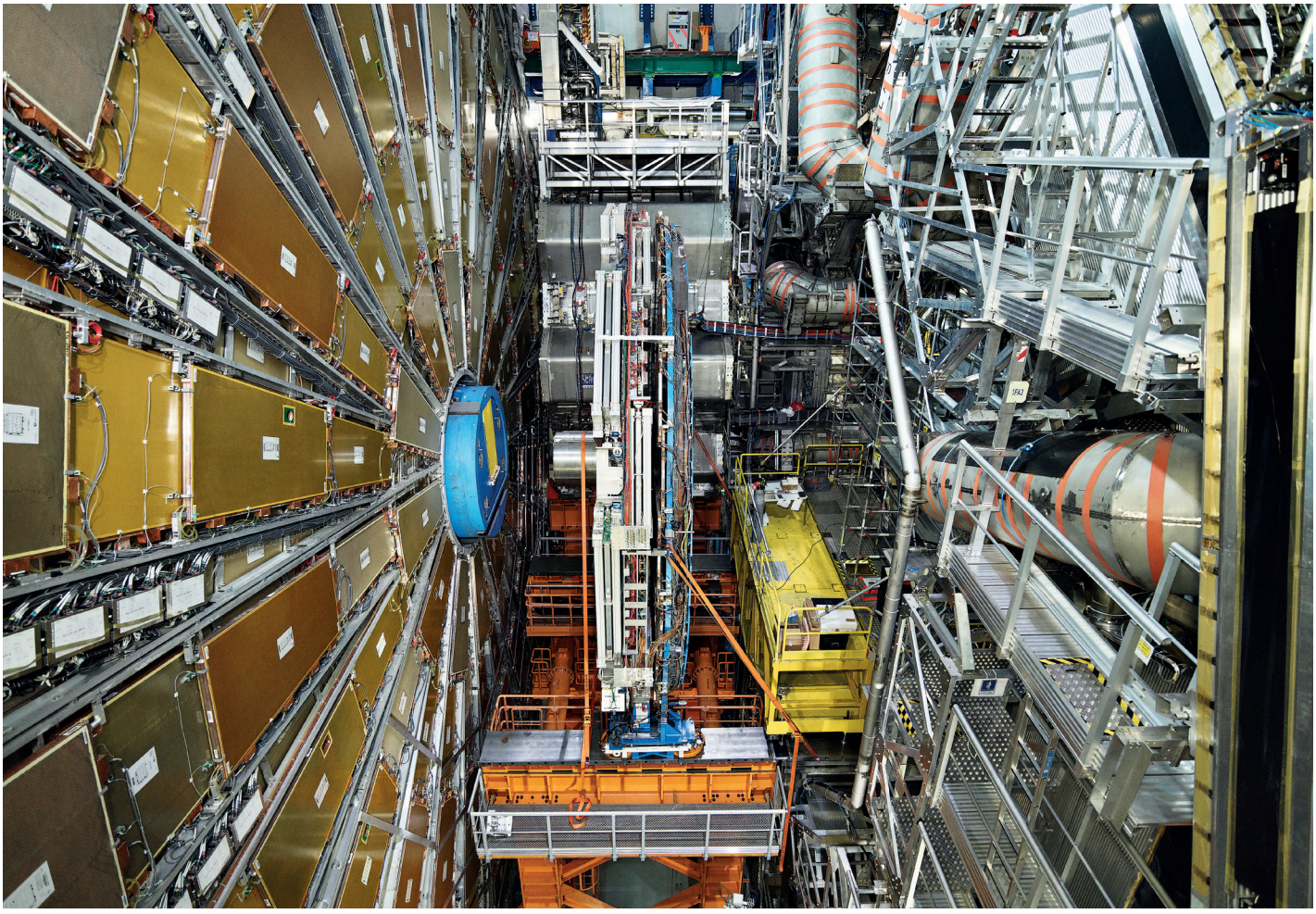
E l'approccio statistico allo studio non consente quasi mai di utilizzare logiche di tipo binario, per cui ogni dato ha un valore se è parte di una popolazione, ma non è individualmente distinguibile dal fondo (si pensi ad esempio agli eventi in cui si è osservato il bosone di Higgs). Per quanto si possa

b.
Il centro di calcolo del Cern: centinaia di migliaia di processori negli armadi refrigerati ricercano instancabilmente preziosi indizi che possano consentire alla scienza di aprire nuove prospettive.

tentare di restringere e rendere efficiente la ricerca, il grosso del lavoro è un "attacco a forza bruta" agli archivi di dati. Pur non potendosi prescindere dalla catalogazione dei *dataset*, almeno un sottoinsieme deve passare al vaglio di algoritmi di tipo statistico. La disponibilità di elaboratori più veloci e più economici, che consentono di costruire centri di supercalcolo sempre più potenti, non è l'unica strada. Rileggere i dati costa molto sia in termini di tempo che di energia. Ogni nuova sessione di analisi, con nuove selezioni, deve essere preparata con cura per ottimizzare i risultati. Non è affatto detto che i ricercatori riescano a "sistemare tutte le trappole" per catturare fenomeni elusivi, attraverso quello che tradizionalmente viene chiamato "*analytical processing*". Le tecniche di intelligenza artificiale, anzi, offrono numerosi vantaggi da questo punto di vista, poiché lo sperimentatore, almeno in teoria, deve "solo" preparare una simulazione, definire la tipologia di algoritmo e confrontare i risultati dell'addestramento su dati simulati con i dati reali. Tuttavia, un "albero decisionale" (Bdt) non prenderà in considerazione quantità che non siano state identificate in precedenza da chi dirige l'analisi. Similmente accade per reti neurali a pochi livelli (in uso da tempo nelle applicazioni più semplici). Significativi vantaggi possono venire dall'uso di tecniche più aggressive come le reti neurali convoluzionali (Cnn) che trattano i dati grezzi, opportunamente addestrate e rese robuste con "reti avversarie" (Gan) che le abituano a non farsi ingannare da segnali irrilevanti, rumore, disturbi, eventi occasionali (vd. p. 15, ndr). Rispetto alle tecniche di analisi tradizionali, l'intelligenza artificiale in genere fornisce maggiore velocità di elaborazione, perché in definitiva funzioni complicate dei dati sono approssimate da "mattoni" semplici. Si sposta l'onere della scoperta dei criteri ottimali di classificazione dal ricercatore alla macchina.

Tuttavia, si introducono due problemi nuovi, ossia quello della preparazione dei dati di addestramento ("*training set*"), che devono essere rappresentativi del fenomeno cercato, e quello dell'interpretazione statistica dei risultati. In altri termini si deve riguadagnare il controllo sulla significatività delle decisioni prese dalle macchine. Quanto possiamo fidarci di un classificatore automatico?

Quanto spesso si sbaglia, e di quanto? Lo stiamo usando in "*overfitting*", cioè ricorda perfettamente l'addestramento ma non è riuscito ad astrarre quantità rilevanti? La maggior parte di queste tecniche nasce al di fuori della fisica delle alte energie, in ambiti originali di applicazione, ad esempio nel commercio, nelle scienze sociali o nella sicurezza informatica, nei quali è possibile che vi sia una reale volontà di ingannare l'algoritmo. Si affrontano e si risolvono problematiche che nella fisica non si pongono: noi possiamo credere all' "onestà" dei nostri dati, altri no. D'altra parte possono esservi cause ignote o sottovalutate di malfunzionamento di un apparato o semplici correlazioni inesplorate che finiscono per mimare segnali statisticamente rilevanti. Questo porta un cambiamento rilevante nel profilo culturale del fisico: è relativamente meno importante la capacità di costruire programmi molto ottimizzati dal punto di vista delle prestazioni, perché strumenti aperti come TensorFlow, Keras o altri, sviluppati e supportati da grandi multinazionali, consentono di avere accesso allo stato dell'arte dal punto di vista tecnologico. Invece, è progressivamente più importante che il fisico si addentri in aspetti matematici, statistici e più astratti della teoria dell'informazione. Immagazzinare dati, rileggerli e rielaborarli può avere un impatto notevole in termini di tempo (sia umano che macchina) e di costo della memoria di massa. Una delle più forti limitazioni negli algoritmi di selezione degli eventi sperimentali (il cosiddetto *trigger* online, vd. in Asimmetrie n. 17 p. 33, ndr) viene dalla necessità di "tagliare" il più possibile, sacrificando a volte intere linee di ricerca. Tuttavia, la potenza di calcolo oggi disponibile è tale che è possibile far girare durante l'acquisizione parti della ricostruzione degli eventi che normalmente si eseguono in fase di analisi, o versioni accelerate di esse. Si sfuma così la distinzione tra online e offline, tra presa dati e analisi.
Nell'esperimento Cms del Cern questo è stato chiamato "*data scouting*", ossia



c.
La spettacolare macchina di Cms, in preparazione per una nuova presa dati. La complessità della strumentazione è seconda solo a quella delle informazioni prodotte.

d.
Quattro Detection Unit di Km3net pronte per essere deposte sul fondo del Mediterraneo. Negli abissi marini, dove migliaia di metri d'acqua schermano i raggi cosmici, cercheranno neutrini provenienti dal cosmo o dall'alta atmosfera.

andarsi a cercare già durante la presa dati ciò che può essere interessante prima che sia perso per sempre. Anche altri due esperimenti del Cern, Atlas e Lhcb, hanno implementato approcci simili. A ciò si affianca il "*data parking*", ossia il salvataggio temporaneo di un sottoinsieme di dati grezzi, necessari in caso di scoperta (e che diversamente potranno essere cancellati). Queste politiche di gestione attiva dei dati, utilizzate nel *run* 2 di Lhc dal 2015 al 2018 per la ricerca di fisica esotica (come lo Z'), rendono percorribili canali di indagine che con i metodi "tradizionali" ("salva subito e analizza in seguito") richiederebbero costi enormi, a fronte di esigue speranze di scoperta. D'altra parte, introducono fattori di rischio, poiché errori di valutazione nelle selezioni o nell'implementazione portano a una perdita secca di dati, e pertanto possono essere usate solo per incrementare il campione, senza potersi sostituire ai trigger convenzionali. L'ottimizzazione della presa dati è da tempo una realtà nel campo dell'astronomia osservativa (vd. p. 31, ndr), e in tempi di astronomia multimessaggera, che richiede l'interazione di telescopi ottici, radio, gravitazionali e per neutrini, operativi in diversi luoghi della Terra, la variazione dinamica dei trigger in conseguenza di allarmi globali è ormai prevista già in fase di sviluppo dei sistemi di acquisizione, come nel caso di Km3net che dev'essere pronto a segnali di supernova e altri transitori. La collaborazione Ska, che costruirà e gestirà la più grande rete di radiotelescopi del mondo, ha estremizzato i concetti di acquisizione adattiva e di elaborazione "al volo": con più di 10 exabyte di dati raccolti al giorno, di cui "solo" 1 petabyte al giorno sarà immagazzinato, la riduzione dei dati deve avvenire più vicino possibile all'antenna, possibilmente utilizzando Cpu economiche ma relativamente potenti e non "avide" di energia. I partner commerciali non perdono l'occasione di contribuire a questi sviluppi: come accaduto spesso in passato, la ricerca pura sta già incubando le tecnologie che cambieranno il nostro modo di vivere domani.

Biografia
Cristiano Bozza, Ph.D., si occupa dal 1996 di acquisizione, ricostruzione dati e database, prima in Chorus e poi in Opera, esperimento che ha scoperto l'oscillazione dei neutrini muonici in neutrini tau. Attualmente lavora in Km3net, sui neutrini da astrosorgenti e atmosferici. Nel progetto europeo Asterics, per cui è stato responsabile nazionale Infn, ha contribuito a simulazioni, supercalcolo e applicazioni di intelligenza artificiale, attività che continua nel nuovo progetto Escape.

Link sul web
https://www.km3net.org
https://www.skatelescope.org/

DOI: 10.23801/asimmetrie.2019.27.5