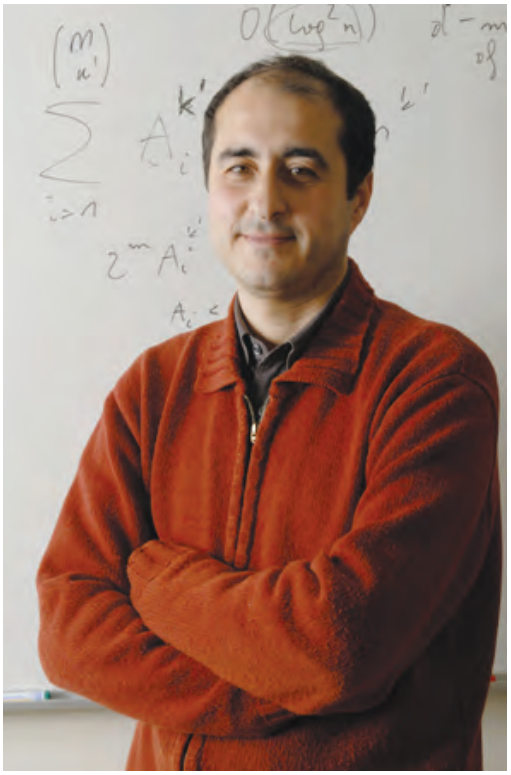
[as] riflessi

Tutto in un click.

di Vincenzo Napolano



a.
Stefano Leonardi, presidente del Consiglio del corso di studi in Data Science della Sapienza.

Si chiamano Big Data. È l'enorme quantità di dati digitali, che grazie alla rivoluzione del web prima e a quella dei dispositivi mobili poi, tutti noi contribuiamo a generare su tutto il pianeta e senza sosta. Ma non solo. Accanto ai dati della miriade di pagine web, email e comunicazioni private ci sono quelli prodotti e memorizzati ogni giorno in tutto il mondo, ad esempio in ambito medico, dai sistemi di sicurezza o dalla ricerca scientifica. Il Large Hadron Collider del Cern è famoso anche per questo: per la strabiliante quantità di dati prodotti ogni giorno dalle collisioni di particelle nell'acceleratore più grande mai realizzato dall'uomo. Pari ogni anno a centinaia di milioni di gigabyte. "La differenza principale tra i dati di Lhc e quelli prodotti in continuazione dalla rete – ci spiega Stefano Leonardi, presidente del Consiglio del corso di studi in Data Science della Sapienza Università di Roma – è che i primi possono essere memorizzati e poi esplorati in modo relativamente più ordinato e rapido". Basandosi sui modelli teorici, i fisici hanno realizzato dei sistemi elettronici, in grado di riconoscere tra i nuovi dati generati ogni milionesimo di secondo nell'acceleratore quelli privi di informazioni utili, scartandoli in modo automatico. Nonostante questo, naturalmente, la mole di dati da archiviare e poi trasmettere ai ricercatori di mezzo mondo resta immensa, e per questo è stata costruita una rete planetaria di calcolo parallelo: la Grid (vd. in Asimmetrie n. 13 p. 21, ndr).
Il volume dei dati generati dalla rete è però ancora più grande. Se consideriamo che ogni minuto gli utenti di Facebook condividono 2 milioni e mezzo di post, quelli di Twitter twittano circa 300 mila volte, su Youtube vengono caricate 72 ore di video, Google riceve circa 4 milioni di richieste e si inviano 200 milioni di email, abbiamo un'idea più concreta della quantità di dati in gioco. Per stimarla si parla di "zettabyte", ovvero mille miliardi di gigabyte: dati destinati a crescere in modo esponenziale nei prossimi anni e che vorremmo continuare a esplorare in pochi istanti, come facciamo oggi. "Il web genera per sua natura in continuazione dati distribuiti su tutto il pianeta, – continua Leonardi – che per altro sono costantemente interrogati ed esplorati da milioni di utenti. Per rendere possibile una ricerca di contenuti istantanea, come quella ad esempio dei motori di ricerca, il problema non è solo la straordinaria quantità dei dati, ma anche il modo complesso e dinamico in cui sono prodotti, i loro mille formati possibili e la diversa qualità e rilevanza". Anche se a noi sembra un semplice click, i motori di ricerca nascondono in realtà algoritmi estremamente sofisticati, in grado di scandagliare contenuti praticamente infiniti, e sempre diversi, in pochi istanti. "Per fare questo – ci spiega ancora Leonardi – si

utilizza quella che si chiama *fingerprint*, l'impronta di una pagina web: una minima quantità di informazioni, a partire da cui si può ricostruire con metodi statistici il suo contenuto integrale. Così gli algoritmi di ricerca possono riconoscere le pagine, ad esempio per determinare contenuti simili o duplicati analizzando una quantità di informazioni molto minore di quella che poi ci forniranno alla fine". Del resto anche ognuno di noi ha una sua impronta digitale in rete, costruita a partire dai dati di navigazione o dai nostri profili sui social network. Anche in questo caso si tratta di una miriade di dati, impossibile da esplorare in modo analitico, e perciò archiviati e utilizzati con un processo statistico. L'efficacia di questo metodo è lampante, se pensiamo alla personalizzazione delle pubblicità online o delle raccomandazioni dei social network, in cui ci imbattiamo ogni giorno.
"La variabilità e distribuzione planetaria dei dati della rete, oltre che la loro quantità, ha finito per trasformare anche lo stesso modo in cui vengono memorizzati e trasmessi i dati. Oggi l'ambiente informatico nel quale vive la rete è il Cloud, dove potenzialmente ogni nodo è in comunicazione con ogni altro. Le informazioni non sono immagazzinate in modo predefinito e gerarchico, ma secondo priorità variabili, così da adeguarsi in tempo reale alle richieste della rete". Anche il mondo della scienza guarda con interesse a questo modello. In un prossimo futuro la Grid potrebbe essere affiancata da un Cloud più flessibile e adattabile a esigenze di discipline diverse, in cui non solo i fisici delle particelle (già attivi con il progetto Indigo Data Cloud, coordinato dall'Infn), ma anche i biologi, gli economisti, gli storici ecc. condivideranno risorse di memoria e si scambieranno dati.
"Qualcosa del genere – conclude Leonardi – è già disponibile, in misura più limitata, per ognuno di noi. Al di là dello straordinario accesso alle informazioni che ci offre la rete, già oggi tutti noi possiamo inventare una app e testarla, ad esempio, sui server di Amazon o Apple. Se funziona, poi, potremmo renderla disponibile a milioni di altre persone. Questo a me sembra semplicemente fantastico".

b.
Il volume dei dati prodotti nel mondo cresce esponenzialmente: si stima che nel 2020 si arriverà a 40 mila miliardi di gigabyte, in gergo 40 zettabyte. La capacità di archiviare tutte queste informazioni in supporti fisici – come gli hard disk e le memorie – è molto più limitata e costosa. La sfida è quella di selezionare e conservare le informazioni sulla base di criteri come qualità e rilevanza, per poter continuare a eseguire la ricerca dei dati di interesse in tempi brevissimi.