Analisi Le quantità immense di informazioni non possono sostituire il metodo come lo conosciamo da Galileo in poi. Nei database generati a caso si trovano correlazioni assurde (tra il consumo di mozzarella e il numero di lauree in ingegneria civile) e la previsione degli eventi non è per nulla assicurata

Attenti, i Big Data non sono la scienza

di MASSIMO PIATTELLI PALMARINI

Per dar corpo alla sempre più frequente e sempre più incalzante espressione Big Data basti pensare che soltanto negli ultimi due anni sono stati generati (da messaggi di posta elettronica, sensori atmosferici, acquisti con carte di credito, foto e filmati digitali, telefoni cellulari, scambi su Facebook, Gps, eccetera), il 90 per cento di tutti i dati attualmente disponibili.

Il ritmo attuale è di due virgola cinque quintilioni di dati al giorno (un quintilione è un miliardo di miliardi). Un vero diluvio. Certamente un tesoro per vari settori scientifici, epidemiologici e ingegneristici, per le tecnologie di analisi dei dati e per la *computer science*. Il rovescio della medaglia, però, è che si sta lentamente diffondendo in alcuni ambienti informatici la stolta idea che l'analisi dei Big Data possa — anzi debba, alla lunga — sostituire la scienza così come noi la conosciamo, da Galileo in poi.

Niente più teorie, leggi di natura, ipotesi, concetti profondi e unificanti. Solamente opportune analisi statistiche di quest'enorme massa di dati, capaci (si sostiene) di rivelare fondamentali correlazioni e prevedere il loro ripetersi.

Già otto anni fa, nel giugno del 2008, la fine della scienza era stata esplicitamente auspicata da Chris Anderson, direttore della rivista specializzata «Wired» in un famoso (ma per alcuni, tra i quali ci sono io, famigerato), articolo significativamente intitolato *Fine della teoria: il diluvio di dati rende obsoleto il metodo scientifico*. Assai più recentemente, a maggio, sul quotidiano britannico «Daily Telegraph», Michael Wilkinson ha sbandierato il successo dei ricercatori di Google nell'aver creato un algoritmo chiamato *Parsey McParseface*, capace di risolvere uno dei più antichi problemi al mondo: catturare globalmente e definitivamente la sintassi delle lingue umane.

Messo a punto, si sostiene, per l'inglese, sarebbe capace, con pochi aggiustamenti e spaziando su Big Data opportuni, di svelare la sintassi di qualsiasi lingua. Il motore di questo algoritmo è chiamato *Network Grammar*.

Nel blog LanguidSlog, i suoi autori, Michael Wilkinson e Richard Harber, espongono i dettagli di questa grammatica statistica, con tanto di foto (a mo' di «spauracchio») di Noam Chomsky, le cui teorie sono da loro dichiarate fallimentari. Robert Berwick, collaboratore di Chomsky, professore di linguistica e di *computer science* al Massachusetts Institute of Technology (Mit), si è preso la briga, meritoria e non lieta, di esaminare in dettaglio questo algoritmo. Ha scoperto enormi svarioni e ha verificato come dia per buone frasi che contengono, invece, numerosissimi e diversificati tipi di errori. Per esempio (traduco dall'inglese, senza perdita di pertinenza) la «Network Grammar» accetta come buone frasi del tipo: «Li hanno persuasi di andare», «Hanno promesso a leggere», «Ha guardato una vecchia bici per riparare», «Ha recensito quel libro piuttosto leggeva» e innumerevoli altri esempi.

Assai più ampia e approfondita è la smentita delle pretese dei Big Data contenuta in un articolo tecnico, pubblicato nel marzo di quest'anno su «Foundations of Science», dal *computer scientist* neozelandese Cristian Calude e dal biofisico e matematico italiano Giuseppe Longo, professore alla École normale supérieure di Parigi e alla scuola di Medicina della Tufts University di Boston.

Calude e Longo, in sostanza, vanno alle radici del problema, mostrando che anche in banche dati generate strettamente a caso si possono trovare correlazioni spurie, prive di qualsiasi significato. Citano un buon numero di correlazioni assurde, come quella tra la frequenza di matrimoni nello stato del Kentucky e il numero di affogati in seguito alla caduta in mare da un peschereccio (oltre il 95 per cento) o come quella tra il consumo pro capite di mozzarella e il numero di lauree in ingegneria civile (96 per cento). Più devastante ancora è il loro calcolo, basato su noti teoremi matematici, del tempo che deve intercorrere prima che due eventi tra loro oggi correlati si ripetano in futuro. Per avere la certezza che eventi paradigmatici ma determinati da un gran numero di variabili — situati in spazi con un gran numero di dimensioni — dovremo aspettare alcuni miliardi di miliardi di volte l'età dell'universo.

Tony Oursler (1957) *Nu*, (2016, installazione mixed media), in mostra alla Galleria Hans Meyer di Düsseldorf fino al 15 agosto: l'artista americano è diventato celebre per i suoi lavori in tre dimensioni che accentuano la carica espressiva del soggetto

Inoltre, fanno vedere che, comunque si fissi, a piacere, una regolarità fra numeri (ad esempio, tutti prossimi per quanto spazio o tempo si voglia), esiste un numero tale che ogni insieme di numeri di quel tipo contiene la regolarità prefissata.

In altri termini, gli immensi archivi di dati, e più sono grandi meglio è, dicono i sostenitori dei Big Data senza scienza, nascondono necessariamente correlazioni del tutto arbitrarie, insensate, che possono benissimo non ripetersi nel tempo e nello spazio. Non a caso questo loro lavoro meritorio si apre con una famosa citazione dantesca: «Fatti non foste a viver come bruti,/ ma per seguir virtute e canoscenza» (*Inferno*, Canto XXVI). Occorre, infatti, che gli studiosi di valore blocchino il folle volo dei puri Big Data se questi vengono gestiti alla cieca, senza avere dietro il sostegno di alcuna vera scienza.

La vendetta di Chomsky
L'algoritmo «Parsey McParseface» prometteva di catturare la sintassi delle lingue umane. In realtà le frasi contengono svariati errori


Attenti, i Big Data non sono la scienza

Il termine
L'espressione Big Data indica la capacità di estrapolare, analizzare e mettere in relazione un'enorme mole di dati con l'obiettivo di scoprire correlazioni tra i fenomeni e prevedere eventi. Nel libro *Big Data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà* (traduzione di Roberto Merlini, Garzanti, pp.306, € 18,60) Mayer-Schönberger e Kenneth Neil Cukier scrivono: «L'era dei Big Data sfida il modo in cui viviamo e interagiamo con il mondo. La cosa più impressionante è che la società dovrà accantonare alcune delle sue ossessioni per i sistemi di causa ed effetto in cambio di semplici correlazioni: interessandosi non ai perché ma solo ai cosa. Questo ribalta secoli di pratiche consolidate e mette in discussione i nostri più basilari approcci a come prendere le decisioni e comprendere la realtà». Eppure le tecniche statistiche e di apprendimento automatico utilizzate fino a oggi per analizzare i database, ancora scarsamente sottoposte a controllo o regolamenti, hanno margini di errore altissimi: la National Security Agency americana nel 2012 ha sorvegliato milioni di chiamate nella città di Washington perché il sistema ha confuso il codice telefonico della città (202) con quello internazionale dell'Egitto (20)

I numeri
Ogni giorno vengono prodotti 2,5 quintilioni di byte di dati (fonte Ibm), il 90% di essi è stato generato negli ultimi due anni. Un quintilione equivale a un miliardo di miliardi

Bibliografia
Oltre al testo di Mayer-Schönberger e Kenneth Neil Cukier, ricordiamo: *Il segnale e il rumore. Arte e scienza della previsione* di Nate Silver (traduzione di Manfredi Giffone, Fandango, pp. 670, € 24,50); *Internet non salverà il mondo. Perché non dobbiamo credere a chi pensa che la rete possa risolvere ogni problema* di Evgeny Morozov (traduzione di Gianni Pannofino, Mondadori, pp. 452, € 19); *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* di Eric Siegel (Wiley, pp. 368, $ 24); *Naked Statistics: Stripping the Dread from the Data* di Charles Wheelan (W. W. Norton & Company, pp. 304, $ 16.95)