Big Data: una moda, una necessità o un'opportunità? La differenza tra descrivere e capire.

Da tempo ormai siamo bombardati da una mole di informazioni cui è difficile stare dietro. Difficile estrarre quelle rilevanti ai nostri bisogni, difficile ricordare dove le avevamo trovate o dove le abbiamo archiviate. Le aumentate capacità dei sistemi informatici e della comunicazione hanno permesso una rilevazione e un accesso a dati di ogni genere e al tempo stesso aumentata la difficoltà di "navigazione" degli stessi. Ma c'è grande differenza tra dati, informazioni, conoscenza e saggezza. Ovvero, se da un punto statistico aumentare la mole di dati potrebbe significare aumentare l'accuratezza nelle previsioni o la precisione di una misura, questo è vero solo sotto certe assunzioni di come il sistema si comporta.

Spieghiamoci meglio: complicato non significa complesso, semplice non significa lineare. La complicazione è spesso sinonimo di inconveniente, di difficoltà, e soprattutto di calcolo e di previsione. Ma non significa impossibilità di capire o di risolvere. Inconsciamente si associa alla manchevolezza di preparazione tecnica o alla mancanza di informazioni. Una moltiplicazione tra due numeri molto grandi è complicata se fatta a mano, come il bilancio di una grande azienda. Ma se supportata da dati, procedure e calcolatori, si semplifica.

La complessità è legata invece ad una caratteristica intrinseca del sistema che lega le variabili che lo regolano e che fa sì che che piccoli scostamenti di una variabile, o dei dati, possano influire notevolmente sull'evoluzione del sistema. Non si tratta di non conoscere il sistema ma di non riuscire a descriverlo adeguatamente, a volte perdendo informazioni non ritenute importanti ma che invece influiscono molto. L'esempio famoso dei sistemi complessi è quello dei cambiamenti climatici, riassunto dall'espressione secondo cui un battito di ali di una farfalla ad un estremo del pianeta potrebbe scatenare una tempesta all'altro estremo. Complicato quindi non significa complesso, complesso spesso significa complicato.

Per fare un altro esempio, pensiamo all'approccio tolemaico al movimento dei pianeti intorno al Sole. In linea di principio potremmo ancora usare quella procedura per calcolare la posizione dei pianeti: con grandi calcolatori e adeguate osservazioni saremmo in grado di prevedere con grande accuratezza dove e come lanciare un satellite. Quando Keplero ha introdotto le sue famosi leggi, tutto è diventato più semplice, ovvero tutto si potrebbe calcolare, facendo un'esagerazione, addirittura a mano! Ma Keplero ha semplificato ma non linearizzato il problema, anzi: Tolomeo affrontava il sistema come somma di tanti termini lineari, i famosi cerchi celesti, Keplero ha introdotto le ellissi, più complicate da un punto di vista concettuale. Semplice quindi non significa lineare, e spesso segue la conoscenza del processo che regola il sistema: le leggi di Keplero sono associate alle leggi della dinamica di Newton che stabiliscono le forze tra le masse.

Detto questo, abbiamo alcuni elementi per capire maggiormente la differenza tra avere a disposizione tanti dati e capire cosa sta succedendo. La differenza tra informazione e conoscenza. Con molta licenza letteraria, l'informazione sta allo svago come la conoscenza sta alla bellezza.

La differenza tra correlazione e relazione causa-effetto probabilmente è ben nota, ma spesso la si dimentica. Se un treno di una stazione di provincia parte sempre subito dopo il suono della campana del campanile della piazza del municipio, significa probabilmente che parte ogni quarto d'ora, e non che se non suona la campana, non parte. La correlazione non implica causa-effetto. Dati acquisiti e leggi che regolano il sistema quindi non sono strettamente

legati a meno che non ci siano due condizioni fondamentali: tutte le variabili che influiscono il sistema sono misurate adeguatamente, si conosce la legge che le lega tra loro.

Avere tanti dati però spesso aiuta. Nessuno di noi ha mai visto volare una mela verso l'alto se lasciata cadere. Questo ci fornisce un indizio che esiste qualcosa che la attrae verso la Terra. E' un indizio, non una prova. E qui veniamo alla moda o utilità dei Big Data, ovvero della raccolta e analisi di una notevole mole di informazioni. Tornando a quanto detto all'inizio, va distinto chiaramente se abbiamo tanti dati di poche variabili o di tante variabili. Potremmo infatti avere miliardi di esperimenti su una mela che cade in diverse condizioni oppure miliardi di esperimenti sulla formazione di una nuvola: non è la stessa cosa. E' intuitivo preferire il primo caso al secondo, ma non è sempre così. Vediamo perché.

Statisticamente abbiamo detto che è meglio avere tanti dati per aumentare la cosiddetta robustezza nella descrizione del sistema: maggiore accuratezza, maggiore precisione, maggiore capacità di prevedere il futuro. Se lanciassimo una moneta solo tre volte per capire la probabilità di avere testa o croce al prossimo lancio, avremmo una stima molto distante dalla realtà. Se la lanciassimo mille volte sapremmo che abbiamo il 50% di probabilità, anche se non sapremo se la prossima volta avremo testa o croce. Se potessimo avere tutti i dati del lancio con la mano, le caratteristiche del vento, la distanza da terra, l'elasticità del materiale di cui è composto quando va a rimbalzare ecc., appena lanciata la moneta potremmo prevedere come va a cadere. Ma una moneta che cade non è un sistema complesso. Se avessimo tantissimi dati sull'umidità dell'aria, l'intensità del vento, l'irraggiamento solare, la composizione chimica media dell'atmosfera, tutto in funzione della quota, potremmo stabilire se si forma una nuvola? Forse sì, ma non sempre. Ad esempio ci mancherebbe l'informazione se esistano particelle in atmosfera che possano indurre l'aggregazione di molecole di acqua.

Quindi, avere tanti dati su "tante" variabili aiuta a descrivere il sistema. La complessità viene meglio compresa se si aumenta l'informazione legata a più variabili. E' come giocare a mosca cieca e poter toccare gli oggetti o ascoltare i rumori: aiuta a descrivere l'ambiente, non a capire dove si nascondono gli altri concorrenti, ma forse a prevedere dove si sono nascosti.

Analizzare quindi grandi moli di dati non fornisce la soluzione al problema di capire, conoscere, ma permettere di descrivere con maggiore dettaglio il sistema in gioco ed eventualmente testare alcune soluzioni. A riguardo esistono molti studi facilmente accessibili al pubblico, cercandoli con parole chiave tipo "correlazione e inferenza causale". Questa possibilità di testare il sistema non sempre è possibile però (vedi metodo degli analoghi e lemma di Kac): vedi le problematiche legate alla complessità del sistema socio-economico-politico a livello globale, dove non si possono certo fare esperimenti su diversi pianeti o tornare indietro nel tempo per ripetere l'esperimento e vedere cosa sarebbe successo altrimenti.

L'analisi delle correlazioni permette di prevedere, ma sempre nel senso statistico di probabilità e non di certezza, in quanto non fornisce alcuna comprensione o "legge" che regola il sistema.

La moda dei Big Data potrebbe nascondere due grandi aspetti nascosti: l'uso economico-politico degli stessi, legati ovvero alla possibilità di prevedere/controllare comportamenti sociali, la incapacità di comprendere il sistema complesso attuale, cercando quindi indizi da una descrizione dettagliata dello stesso. Per fare il salto di livello però avremo bisogno di qualcuno che avrà l'intuizione di inquadrare gli indizi in un nuovo paradigma, nuovi Keplero, Dirac, Einstein…