I Big data: il punto di vista di uno statistico

Categories : Istituzioni e regole

Tagged as : Menabò n. 62, Monica Pratesi

Date : 14 Aprile 2017

Prima di parlare di Big data, parliamo di dati e informazioni statistiche. Il dato statistico nasce dopo una gestazione più o meno lunga, ma comunque originata dalla griglia concettuale di lettura di un fenomeno reale di interesse, adottata da chi lo progetta. Si, poiché ogni dato statistico, rilevato da indagine o letto da archivio amministrativo preesistente, è comunque sempre pensato prima che rilevato. In genere è pensato insieme al processo produttivo che operativamente lo genera. Esso deve essere standardizzabile e ripetibile, per avere informazioni comparabili nello spazio e nel tempo.

Un esempio semplice: il numero di occupati in Italia. Il dato statistico corrispondente può essere ottenuto con la griglia concettuale che definisce occupato chi ha portato a termine almeno un'ora di lavoro nel periodo considerato e che appartiene alla popolazione residente in Italia. Altro dato otterrei, se contassi i dipendenti per i quali le aziende, con sede in Italia, pagano i contributi. Altro ancora se rilevassi le richieste di sussidio di disoccupazione. E' chiaro che la griglia di lettura varia al variare del fenomeno d'interesse, è frutto della impostazione culturale adottata da chi progetta il dato e in genere condivisa dai futuri utenti del dato. Anche l'organizzazione sociale ha il suo peso nel processo produttivo: le indagini e gli archivi amministrativi, per semplicità spesso si conformano all'organizzazione amministrativa del Paese, dove la rilevazione è effettuata. Nel caso dell'occupazione, l'indagine campionaria sulle Forze Lavoro si avvale della suddivisione amministrativa al fine di pianificare il campione e la sua struttura. Chiariti gli aspetti definitori, impliciti nella griglia concettuale, la qualità del processo produttivo e delle sue singole fasi anche operative, si traduce nella qualità del dato, nel suo significato, nella sua attendibilità e validità.

E' chiaro che l'utilizzatore finale del dato non può ignorare le caratteristiche della griglia, pena l'interpretazione distorta del dato e una rappresentazione errata del fenomeno reale di suo interesse. Responsabilità del produttore è dichiarare esplicitamente le caratteristiche del processo di produzione: proprio per i motivi sopra detti la negligenza sulla comunicazione delle sue caratteristiche sarebbe veramente dannosa.

Il passaggio da dato statistico ad informazione e conoscenza è affascinante, oggi più di ieri. Viviamo, infatti, nella società dell'informazione, immersi in notizie, immagini, numeri che costantemente e talvolta anche insistentemente si candidano a rappresentare i fenomeni reali. Non entro qui nel dibattito epistemologico sul significato d'informazione e conoscenza, non ne ho la competenza. Voglio solo richiamare l'attenzione del lettore sul concetto di qualità dei dati, che forse è la cartina di tornasole che possiamo usare in questa riflessione.

Nell'era del diluvio dei dati sembra che il paradigma di qualità, cui ha fatto correntemente riferimento la statistica, stia cambiando e proprio a causa dell'irrompere delle fonti chiamate con un nome collettivo e generalista "Big data".

Si tratta dei dati provenienti dalla tracciabilità di molti comportamenti umani tramite le tecnologie informatiche che registrano le nostre attività sui cosiddetti social network, i nostri acquisti elettronici, le telefonate, gli spostamenti tramite GPS che rimandano il segnale della nostra presenza al satellite corrispondente. Sensori di traffico automobilistico e droni rilevano dati e immagini senza bisogno della nostra partecipazione attiva.

Volume, Varietà e Velocità sono le ormai famose caratteristiche dei Big data e ne sottolineano l'abbondanza, la diversificazione e l'immediatezza. Oltre a queste caratteristiche, che sono qualificazioni descrittive, alcuni, pochi, aggiungono la Veridicità, la Volatilità, la Validità, che invece mi sembra siano più vicine a dimensioni di qualità statistica dei dati.

Uno studio sistematico della qualità dei dati provenienti da queste fonti però ancora non c'è. E' vero che i tempi forse non sono ancora maturi, ma finora si è parlato molto delle "qualità" dei Big data senza affrontare in maniera esplicita il problema della loro "qualità".

Gli stereotipi si sprecano: "Quality is in the Eye of the Beholder" è il più frequente, a significare che si tratta di un concetto soggettivo.

Cos'è la qualità di un dato statistico? Solo la risposta a questa domanda permetterà che si tracci la strada da dato a informazione e conoscenza non distorte. Come abbiamo già detto, il dato statistico in realtà è ottenuto tramite la definizione di una griglia concettuale e operativa e un definito processo di produzione. La qualità del dato è conseguenza della qualità dei singoli passi del processo. Ci sono delle fasi del processo di produzione del dato che devono essere trasparenti per l'utilizzatore. Pena la vanità e la vacuità (altre due V) dell'analisi e dell'interpretazione che sono necessarie per passare da dato a informazione e conoscenza.

Entriamo nel vivo con un esempio. Le occorrenze delle parole chiave sulla ricerca di lavoro, contate da Google o altri motori di ricerca sono un tipico esempio di "dato" proveniente da fonti Big data. L'occorrenza o la frequenza di tali parole è relativa ad un conteggio qui e ora. Il dato è immediatamente riferito a tempo e spazio, ed è disponibile subito. Se preso come quota di persone in cerca di occupazione presenta tutti i suoi limiti. O, per non dare connotazioni negative, direi meglio tutte le sue "qualità", conseguenti al processo produttivo del dato.

Non sto dicendo che la quota di persone in cerca di occupazione, stimata dall'Istat non abbia limiti di qualità. Sto solo notando che essa è già conosciuta nei suoi limiti tecnici, legati all'accuratezza dell'indagine campionaria da cui deriva, al suo errore per non risposta, tanto per citare solo alcune dimensioni del "profilo di qualità" del dato che ha originato numeratore e denominatore di quel tasso. Esso è già informazione pronta a comporre un quadro di conoscenza del mondo del lavoro secondo lo schema concettuale adottato, quadro che sarà perfezionato da chi lo interpreterà: sociologo, economista, filosofo o forse anche lo statistico stesso.

Certamente l'indagine sulle Forze Lavoro è definita su un target e una griglia concettuale specificati (le forze lavoro), la partecipazione dei rispondenti è controllata e studiata nelle sue determinanti. La copertura della popolazione su cui la popolazione è studiata è pressoché completa, o comunque sotto controllo. Il dato è fornito volontariamente dal rispondente.

Nel caso di Google, la partecipazione alla rilevazione del dato è inconsapevole, l'occorrenza cambia al cambiare delle parole chiave considerate. Il volume dei dati certo supera e di molto le poche migliaia di osservazioni dell'indagine campionaria, ma la popolazione cui è riferito è e rimane occulta. La disponibilità del dato è volatile (altra V). Oggi Google permette il conteggio, domani lo potrebbe negare. Il dato è suo.

Riordinando le idee direi che il processo produttivo dei Big data è ancora da conoscere, o meglio qualcuno lo conosce bene, ma manca ancora la sistematizzazione pubblica del profilo di qualità in un nuovo paradigma condiviso, che porti da dato a informazione e a conoscenza.

Al momento attuale non si può dire che i Big data non portino conoscenza. Si deve solo chiedere: di che

cosa? Di quale popolazione, di quali eventi, con quale accuratezza e validità.

Gran volume non significa non affetto da autoselezione. E l'autoselezione porta a conoscenze distorte. Dichiariamo cosa si può conoscere e cosa no.

Che si tratti del prezzo di un biglietto aereo, dei suggerimenti su nuovi prodotti da acquistare via internet, quando navighiamo, non dimentichiamo che il dato che produciamo è deciso da un algoritmo. Cioè da un procedimento logico matematico composto da una serie di passi che dà una risposta finale determinata, indipendentemente da dove e quando vengano eseguiti e da chi li esegue. Bello, ma per capire il processo produttivo del dato è necessario dichiarare con molto dettaglio l'algoritmo usato e la verifica della "qualità" dei singoli passi dell'algoritmo. Questa è la base di ogni patto di conoscenza tra produttori di dati ed utenti finali, altrimenti si rischia di rappresentare mondi che non esistono o che non si riesce a capire. L'algoritmo non può sbagliare, ma chi lo programma, chi definisce i vari passaggi può sbagliare e contare come ricerca di lavoro una ricerca di altre attività.

E' urgente che la griglia di lettura usata nel processo produttivo dei "Big data" sia rivelata all'utilizzatore finale e che si studi in modo statistico, cioè con processi standardizzati e ripetibili, le dimensioni della qualità di queste fonti. Quanto già sistematizzato sulle tradizionali fonti di dati può servire solo come punto di partenza. I Big data sono veramente nuovi, anche da questo punto di vista.

Anche sotto l'ipotesi di perfetta Veridicità, le caratteristiche di Volume, Velocità e Varietà possono comportare accumulazione di rumori e disturbi nei dati, correlazioni spurie e/o coincidenze, e endogeneità incidentali. Errori di contenuto e dati mancanti non fanno altro che complicare il quadro. L'analisi dei Big data deve tenere conto di questo profilo di qualità (o di errore) per non incorrere in rappresentazioni della realtà del tutto mistificanti. Gli esempi al proposito si sprecano (Fan et al., "Challenges of Big data Analysis", *National Science Review*, 2014). Alcuni di questi sono basati sulla incondizionata fiducia che il volume delle osservazioni sani ogni difetto di misurazione, di attendibilità e di validità.

Un esempio per tutti è l'ormai famosa previsione dell'incidenza di influenza ottenuta tramite le ricerche di termini affini con Google (Lazer et al. "The parable of Google Flu: Traps in Big data analysis", *Science,* 344, 2014). In poche parole, l'andamento dell'incidenza dell'influenza rilevato con dati ufficiali fu previsto bene dal trend delle ricerche Google fino al 2012. Poi l'errore di previsione irruppe improvvisamente e la causa riconosciuta è il propagarsi di ricerche legate al termine "influenza", raccomandate da altri programmi Google, che proposero ricerche su sintomi e rimedi. Tali ricerche gonfiarono artificialmente le occorrenze del termine influenza e dei termini correlati. Il problema fu nella costruzione dell'algoritmo. Certo gli errori si correggono, ma il problema è che raramente divengono "pubblici", come in questo caso.

La riflessione critica sulla tecnica ed il processo produttivo dei Big data deve essere avviata e portare alla definizione in chiaro delle griglie concettuali e dei processi produttivi usati per leggere i fenomeni attraverso algoritmi e motori di ricerca. Solo in questo modo il dato ottenuto potrà chiamarsi statistico ed essere la base per la costruzione d'informazione e conoscenza.

La posta in gioco è alta. Non riguarda solo i confini disciplinari tra statistica, informatica e scienze sociali ed economiche e la necessità di una *data science* interdisciplinare. Il discorso è più ampio e coinvolge la formazione dei giovani, la tutela della privacy e la costruzione di una rappresentazione condivisa dei fenomeni reali.

Che la formazione sia il vero petrolio di domani (non i dati) mi sembra un fatto incontrovertibile. E' necessaria per avere le competenze adeguate a vivere nel terzo millennio. La corretta interpretazione dei dati statistici, tramite un'appropriata *statistical literacy,* deve estendersi anche ai dati provenienti dai nuovi strumenti di comunicazione. Su questo c'è veramente tanto da fare. Si tratta di una nuova ineludibile

alfabetizzazione. Necessaria, altrimenti la tempesta di dati ci renderà incapaci di elaborare le informazioni e di valutarne l'attendibilità. Segnali di disagio nelle generazioni definite *delle reti* (Alleva e Barbieri, *Generazioni*, Donzelli, 2016)già si notano: dipendenze dal Web consumate nella convinzione di essere liberi, ma, mai come oggi omologati e prigionieri di percorsi prestabiliti. La visione dei nuovi media come semplici strumenti di comunicazione risulta ribaltata, per alcuni si rischia il dissolvimento dell'individuo dentro i flussi di trasmissione di dati e di informazioni (Baronciani, *Nella tana del bianconiglio, saggio sulla mutazione digitale*, Effequ, 2014).

In questo ambito la tutela della privacy e la protezione dei dati personali è un tema di grande attualità. Nomi, numeri di telefono ed abitudini hanno un valore di mercato e spesso passano da una lista all'altra senza che siamo consapevoli di fornirli. I metodi per opporsi ci sono ma non sempre sono sufficientemente noti ed efficaci.

Facciamo bene ad essere diffidenti verso l'obiettività dei dati e degli esperti. Facciamo bene soprattutto perché spesso vengono qualificati come statistici, dati che non lo sono poiché non è chiara la loro genesi (griglia concettuale), il loro processo produttivo e quindi la loro qualità. D'altra parte non si può fare a meno di una rappresentazione e descrizione condivisa della realtà, costituita da un insieme di fatti e di dati statistici (Davies, "How statistics lost their power and why we should fear what comes next", *The Guardian*, 19 January 2017).

E' molto difficile commentare i cambiamenti proprio mentre questi avvengono: si rischia di essere catastrofisti oppure eccessivamente ottimisti. Non voglio sembrare né l'uno né l'altro, mi preme solo far notare che sulle modalità di rappresentare e descrivere i fenomeni reali gli statistici e i produttori di dati ufficiali hanno ancora molto da dire.