Un approccio semantico

Sfide e opportunità per l'epidemiologia 2.0

Big data. Un rivoluzionario approccio per l'epidemiologia, la chiave di volta del futuro sistema sanitario, un nuovo strumento per monitorare e contrastare, in modo tempestivo, il manifestarsi di epidemie? Un capovolgimento di metodologia, l'abbandono dell'inferenza in sostituzione del machine learning? La possibilità, fino a pochi decenni fa inimmaginabile, di poter seguire le fasi di cura del paziente in tempo reale, mediante sensori direttamente collegati con il proprio medico curante? Big data per mappare la sequenza del genoma umano? O per prevedere il diffondersi persino di pandemie sfruttando le ricerche testuali su Google, effettuate dalla popolazione, per ottenere la geo-localizzazione di una determinata patologia? Oppure, i big data sono solo un grande bluff?

Secondo alcuni è solo un'operazione di marketing. O poco più. Una moda, uno dei tanti prodotti transeunti della modernità: tutti parlano di grandi insiemi dati e di nuovi flussi (data deluge), nonostante nessuno li abbia mai visti davvero. Anche in ambito epidemiologico e scientifico, non mancano le critiche sul possibile contributo informativo dei big data, rafforzate dalla consapevolezza che ai grandi numeri non necessariamente corrispondano grandi informazioni o, comunque, informazioni di validità scientifica o di qualità.

Il ricercatore appartenente a questo periodo storico, che possiamo far idealmente partire dal 2010, anno in cui il termine "big data" ha ottenuto una diffusione planetaria, si troverà probabilmente nel limbo dell'indecisione, tra scetticismo e curiosità, come se dovesse scegliere tra il nuovo approccio di indagine che fa uso dei big data oppure quello tradizionale, come se i due ambiti fossero mutuamente esclusivi.

La stampa popolare e accademica ha – con notevole entusiasmo – iniziato a utilizzare il termine "big data" per descrivere la rapida integrazione e analisi su larga scala; tuttavia, una chiara definizione di big data rimane sfuggente. Le modalità con le quali i big data potrebbero influenzare il futuro della ricerca epidemiologica e di intervento sanitario sulla popolazione rimangono, al momento, poco chiare.

Il peso della terminologia

Il corretto approccio alla tematica forse ha senso a partire dall'aspetto semantico, e conseguentemente concettuale, della definizione ormai celebre di big data riassunta nel primo modello teorizzato delle 3 V: volume, varietà e velocità. Un primo passo propedeutico alla formulazione di ipotesi idonee alla ricerca e alla formazione di un background realmente critico risiede nella terminologia: cercare di fare chiarezza su cosa differenzi i big data, nella loro natura, dai dati ottenuti dalle fonti tradizionali. In sostanza si torna alla domanda iniziale: cosa sono i big data?

Big, ad esempio, è di per sé un attributo generico e poco calzante che fa esclusivamente riferimento al *volume*, alla mole di dati, che ha un'accezione soggettiva e non indica nulla di quantitativo. Ad esempio, anche le schede di dimissione ospedaliera della Regione Lazio dovrebbero appartenere ai big data in quanto composte da milioni di record. La percezione del concetto di grandezza è ovviamente soggettiva.

La *varietà*: i dati possono presentare eterogeneità nel tipo, nella rappresentazione e nell'interpretazione semantica. Possono essere di qualsiasi natura (strutturati, semistrutturati o non strutturati). Pertanto, ad esempio, considerando un ipotetico linkage tra sistemi informativi riguardanti farmaceutica, ricoveri e assistenza specialistica, possiamo parlare di big data?

La *velocità*: alle nuove informazioni estraibili dai dati viene spesso associata una funzione di utilità che degrada velocemente con il passare del tempo. La velocità inoltre è anche relativa al tasso di produzione dei dati.

Alessandro Rosa

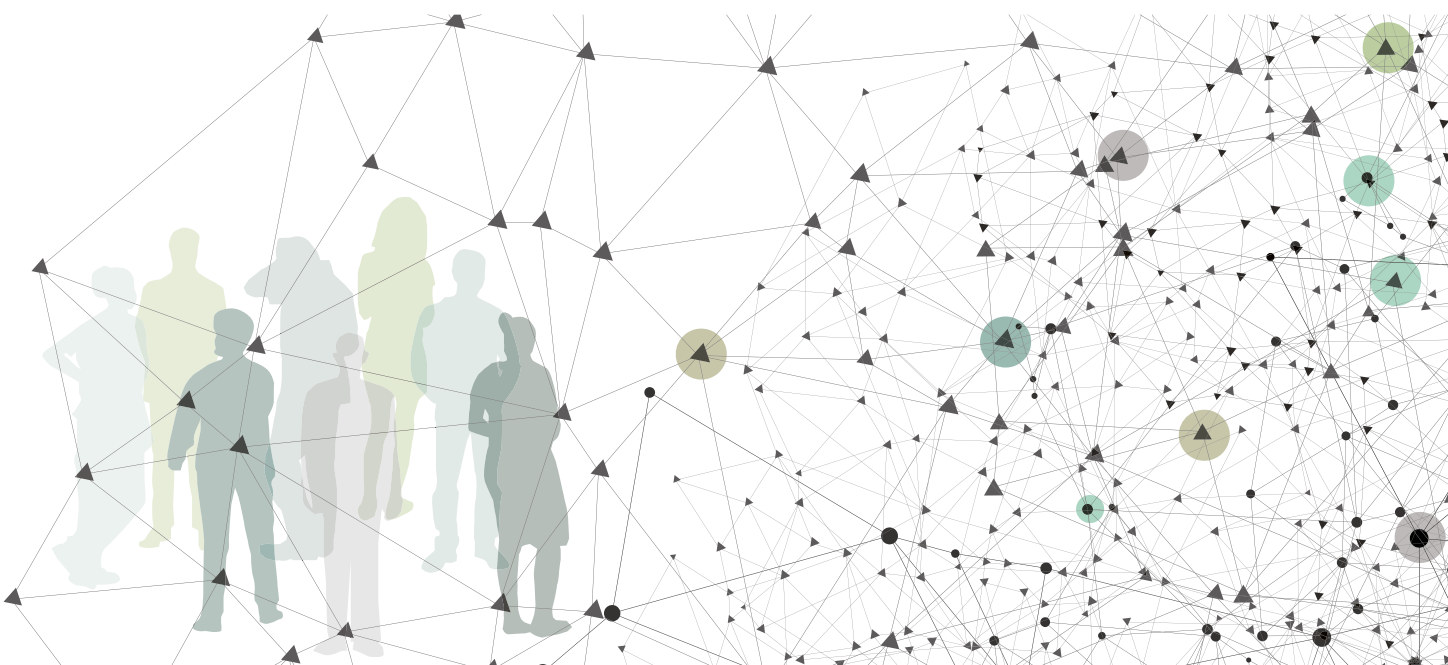
Dipartimento di Epidemiologia Servizio sanitario regionale del Lazio

È importante aggiungere che i big data vengono generati automaticamente da operazioni di interazione persona-macchina (un esempio, in ambito finanziario, sono i dati transazionali), persona-persona (social network) e macchina-macchina (si pensi ai dati inviati dai sensori direttamente ai telefoni cellulari). Nella convenzione universalmente accettata si associano ad enormi moli di volume: si passa dai terabyte (1 tb = 10^{12} b) e petabyte (1 pb = 10^{15} b), fino ad arrivare agli exabyte e addirittura agli zettabyte. Devono presentare un tasso di produzione alto e, inoltre, possono essere di provenienza varia e talvolta non convenzionale: parliamo anche di documenti testuali, immagini, audio, video, dati da sensori o Gps.

I big data, in sintesi, presentano congiuntamente le tre caratteristiche sopra elencate e sono la materializzazione (per usare un ossimoro) dell'*internet of things*, cioè la visione secondo cui gli oggetti nel mondo informatizzato creano un sistema pervasivo e interconnesso avvalendosi di molteplici tecnologie di comunicazione. In pratica, parliamo di dati e flussi continui.

Le caratteristiche sopra elencate differenziano pertanto i big data veri e propri dai dati desunti dalle fonti tradizionali. Tuttavia, a causa di un problema definitorio, l'incompleta espressione "big data" porta a fare confusione con le fonti attualmente disponibili, proprio per il fatto che, anche queste ultime vertono su considerevoli moli di volume.

In ambito sanitario, ad esempio, l'espressione "la tale struttura sta avviando un'iniziativa di big data" dovrebbe essere sostituita dalla più aderente formulazione secondo cui "la tale struttura intende combinare i dati sanitari in formato elettronico e i dati genomici per applicare ai pazienti trattamenti personalizzati", posto che la tale struttura stia davvero portando avanti un progetto sui big data.



	Big data	Analytics tradizionale
Tipologia dei dati	Non strutturati	Ordinati per righe e colonne
Volume dei dati	Da cento terabyte ai petabyte	Decine di terabyte (o meno)
Flusso dei dati	Flusso costante di dati	Pool statistico di dati
Metodi di analisi	Machine learning	Per ipotesi
Scopo principale	Prodotti data based	Servizi, supporto alle decisioni

Fonte: Davenport T. Big data @l Lavoro. Milano: Franco Angeli Edizioni, 2015.

Nuove linguaggi, nuove competenze

Approfondito l'aspetto formale, avendo natura e struttura differenti rispetto all'approccio classico, i big data necessitano inevitabilmente di tecniche di acquisizione, selezione, normalizzazione, storage, calcolo e conseguente analisi statistica del tutto parametrati alle nuove esigenze.

La vastità dell'argomento richiederebbe un approfondimento ma appare evidente come l'approccio globale del ricercatore a dati di questa tipologia risulti rivoluzionato, nel metodo e negli strumenti. La nuova sfida dell'epidemiologo, pertanto, potrebbe consistere nella capacità di cogliere nei big data elementi di confronto con i risultati precedentemente ottenuti. In che modo? Gli esempi potrebbero essere molteplici ma rimane il dubbio posto inizialmente. Ricavare informazioni da enormi database, sovente destrutturati, impone una riflessione sulla qualità del dato. Ovvero: gli algoritmi, utili a identificare pattern e a effettuare previsioni, possono produrre risultati consistenti per la ricerca scientifica? Qual è il punto di incontro tra le tecniche di data mining (l'analisi, da un punto di vista matematico, eseguita su database di grandi dimensioni, senza formulazioni di ipotesi a priori) e la statistica basata sull'inferenza, per cui partendo da un campione si tenta di generalizzare le conclusioni sull'intera popolazione? I due approcci potranno interagire e trovare un punto di incontro nella ricerca epidemiologica?

Il futuro probabilmente richiederà ai ricercatori di abbracciare nuove competenze tecnologiche, in particolare linguaggi di programmazione *ad hoc*. Ad esempio, i dataset analitici possono essere accompagnati da metadati, fruibili dal pubblico, generati tramite programmi di web scraping, tecnica che permette lettura ed elaborazione di dati desunti dal web.

I big data, in farmacologia, sono rappresentabili mediante aggregazioni di informazioni legate alle caratteristiche delle popolazioni che assumono i farmaci, ovvero: dati biometrici (tra questi, ad esempio, peso, pressione, altezza e quantità di grasso corporeo), dati correlati alle abitudini delle popolazioni, dati omogenei sugli obiettivi che si vogliono raggiungere con la terapia, dati di riferimento sull'andamento naturale delle stesse patologie, dati sulla durata della risposta farmacologica nel tempo. Secondo alcune previsioni, crescerà nel prossimo futuro sia lo sviluppo di app sanitarie che potranno essere utilizzate come tramite tra il medico curante e il paziente, sia il numero di coloro che utilizzeranno mezzi tecnologici per memorizzare e trasmettere documentazione sanitaria.

In ambito di sanità pubblica, le decisioni pensate per un rapido intervento possono aver bisogno di sfruttare applicazioni sui cellulari, incamerando i dati in server centralizzati (data warehouse).

Un ulteriore sviluppo, infine, potrebbe riguardare l'ampliamento delle modalità attraverso le quali gli studi epidemiologici possono migliorare, nel concreto, la salute della popolazione. Ad esempio, enti che si occupano di benessere e che incoraggiano stili di vita sani, accumulando big data sulle abitudini di vita della propria popolazione di studio, potrebbero vedere negli epidemiologi una valida opportunità di collaborazione, per beneficio sia accademico sia industriale.

Parafrasando Alan Kay, "il miglior modo per predire il futuro è inventarlo". F

VEDI ANCHE

Occorre separare il segnale vero dal rumore di fondo: non è né facile né immediato ma è la sfida che è necessario vincere per tradurre le informazioni sempre più numerose di cui disponiamo in benessere e salute per i cittadini. Il titolo della rubrica della rivista *Science* nella quale è stata pubblicata la nota di Muin J. Khoury e John P. A. Ioannidis è programmatico: *Insights*[1]. Poco più di una pagina ricca di indicazioni essenziali che giungono da due autori di istituzioni in certo modo complementari: da una parte i Centers for disease control and prevention e dall'altra il Meta-Research innovation center di Stanford. Pubblico e privato, est e ovest degli Stati Uniti.


Muin J. Khoury


John P.A. Ioannidis

"Big error can plague big data", avvertono gli autori citando il caso del monitoraggio dell'andamento dell'epidemia influenzale attraverso Google flu trends (ne parliamo a pagina 9): paradossalmente, la proporzione di falsi allarmi all'interno di quanto registrato si moltiplica quando l'oggetto di misurazione diviene più ampio. Allo stesso modo, la tentazione di correlazioni spurie e bizzarre diventa quasi irresistibile, così che possiamo finire col leggere su riviste anche teoricamente rispettabili che la produzione di miele da parte di colonie di api si correla alla frequenza di arresti per detenzione di marijuana nei giovani delle zone dove si trovano gli alveari… "La forza dei big data è nel trovare associazioni e non nel mostrare che queste associazioni abbiano significato", ricordano gli autori.

Come migliorare il potenziale insito nei big data di migliorare la salute delle popolazioni e prevenire le malattie? I big data sono per definizione di tipo osservazionale e come tali sono esposti alle distorsioni più varie, ma possono essere *embedded* in popolazioni epidemiologicamente ben caratterizzate e rappresentative. È questa la chiave per rendere utili le rilevazioni estese o "spontanee" di dati strutturati e non strutturati.

Un passo in questa direzione è stato fatto dall'Institute for Health metrics and evaluation che ha deciso di affidare al *Lancet* la pubblicazione del **Global burden of diseases, injuries, and risk factors study** 2015, la rilevazione dei dati epidemiologici di 195 paesi e territori, per anno, età e genere[2]. Si tratta di un documento molto rappresentativo della volontà di sistematizzare l'informazione disponibile riguardante la salute, per trasformarla in conoscenza: "This is the science of making data meaningful", afferma l'editoriale della rivista inglese. La pubblicazione di un riferimento di questo tipo in una rivista accademica è una scelta significativa perché rende visibile l'impegno per garantire ai dati il massimo rigore, anche per aver superato il processo di peer review al quale ogni contributo pubblicato dal *Lancet* deve essere sottoposto. •

1. Khoury MJ, Ioannidis JPA. Big data meets public health. Science 2014;346:1054-5.
2. Editorial. GBD 2015: from big data to meaningful change. Lancet 2016;388:1447.