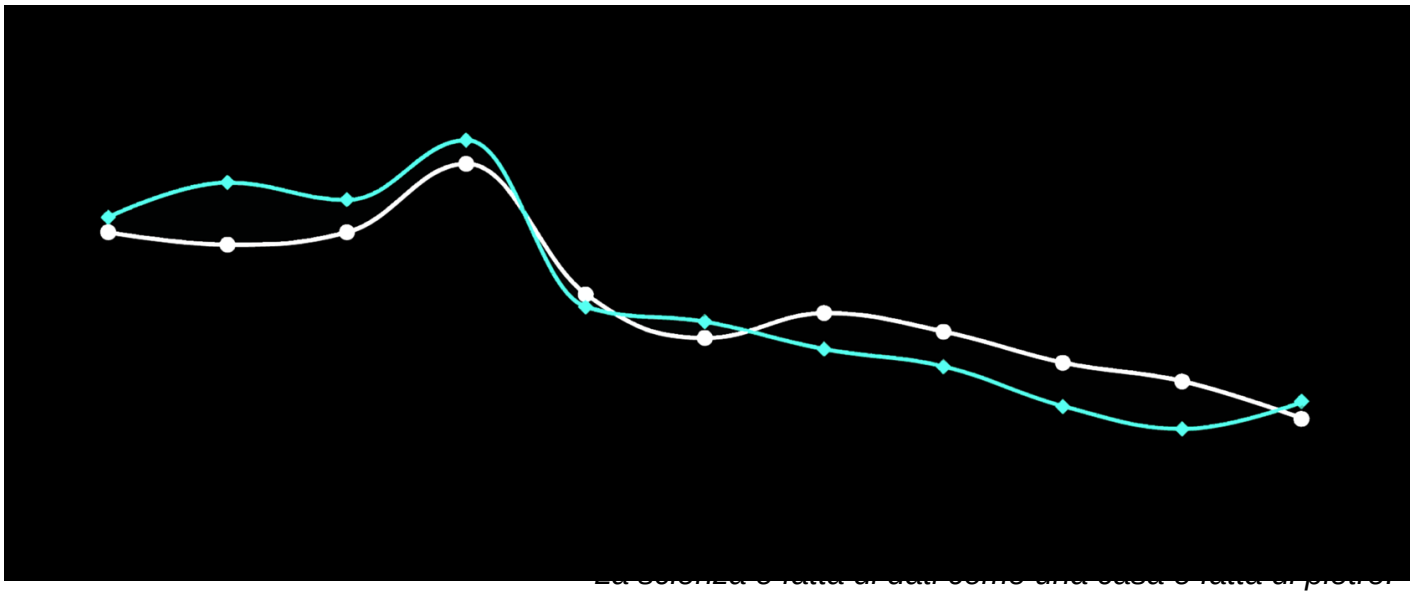
Perché i dati non sono neutrali



La scienza è fatta di dati come una casa è fatta di pietre.
Ma un ammasso di dati non è scienza più di quanto un mucchio di pietre sia una vera casa
Henri Poincaré

Da quando il «fenomeno big data» ha preso piede, si è imposto sempre più nell'ambiente accademico, e non, il dibattito sul rapporto tra dati e ricerca. L'utilizzo massivo e sistematico dei dati ha infatti profondamente messo in discussione il ruolo della scienza e dello scienziato, le tecniche di cui si avvale e il metodo scientifico stesso su cui da Galileo, Cartesio e Bacone in poi (ma forse pure prima) si è fondata la ricerca e, per quanto la definizione stessa di metodo scientifico sia un po' una semplificazione e non corrisponda affatto ad un concetto univoco, l'avvento dei big data ha senza dubbio messo in discussione la nozione di «teoria» per come la si è sempre intesa, aprendo un dibattito su come, quando, dove e perché *produciamo* scienza e, soprattutto, su come tutto questo cambia nel momento in cui abbiamo a disposizione quell'apparentemente inesauribile miniera di informazioni che emergono dai big data.

Ma facciamo un passo indietro e proviamo prima di tutto a ragionare su *cosa* effettivamente rappresenti questa enorme – ed informe? – mole di informazioni e quali contenuti effettivamente questa sia in grado di elaborare e comunicare.

Nel 2011, Teradata, una *public company* statunitense che si occupa di prodotti e servizi legati ai database, affermò che «un sistema di big data eccede, sorpassa e supera i sistemi hardware e software comunemente usati per catturare, gestire ed elaborare i dati in un lasso di tempo ragionevole per una comunità/popolazione di utenti anche massiva». Detta così, effettivamente, sembra una cosa il cui impatto sul modo di fare ricerca – che poi è mettere insieme i dati e unire i puntini – è potenzialmente esplosivo.

In tanti, in effetti, con l'emergere dei big data hanno provato a parlare – (forse) provocatoriamente – di *end of the theory*, affermando l'oggettività del dato di per sé e alimentando quella retorica secondo cui *correlation is enough*, ovvero «la correlazione è sufficiente». Uno degli autori che si è esposto di più in questo senso è senza dubbio Chris Anderson, fisico e giornalista scientifico che, in un articolo pubblicato nel 2008 su *Wired* ottenendo un'enorme visibilità, ha sostenuto che la disponibilità di grandi moli di dati, combinata alle adeguate tecniche statistico-matematiche, sia in grado di soppiantare ogni altro strumento analitico rendendo il metodo scientifico, di fatto, obsoleto.

Il feticcio al centro di questo processo – che si propone come una vera e propria rivoluzione metodologica e culturale – è il *petabyte*, multiplo dell'unità di misura dell'informazione digitale[1]. Se un tempo infatti le informazioni da immagazzinare erano talmente piccole da essere sufficienti dei floppy disk da pochi kilobyte per poterle contenere, presto fu necessario usare *array* di disk da svariati megabyte, e poi array di dischi, la cui capacità si misura in *terabyte*.

Il luogo immaginario – che poi tanto immaginario non è – dove si immagazzinano i petabyte, invece, è il *cloud*, ovvero un'architettura che prevede che l'esecuzione delle azioni sia gestita a livello di rete, alleggerendo in questo modo il carico dei computer locali: la parola cloud, infatti, vuole proprio indicare una massa enorme di singole unità che viste da lontano possono ricordare una nuvola. Dico che tanto immaginario non è perché spesso ci si dimentica che a questo tipo di struttura corrisponde un luogo estremamente «fisico», chiuso in stanze blindate enormi e piene di server e di macchinari che appartengono a qualcuno e la cui proprietà «fisica» corrisponde al diritto di proprietà su tutte le informazioni che in quel luogo arrivano e circolano e il cui consumo energetico è spaventoso. Insomma, come dice un adagio piuttosto famoso per chi mastica questo dibattito: *There is no cloud: it's just someone else's computer!*

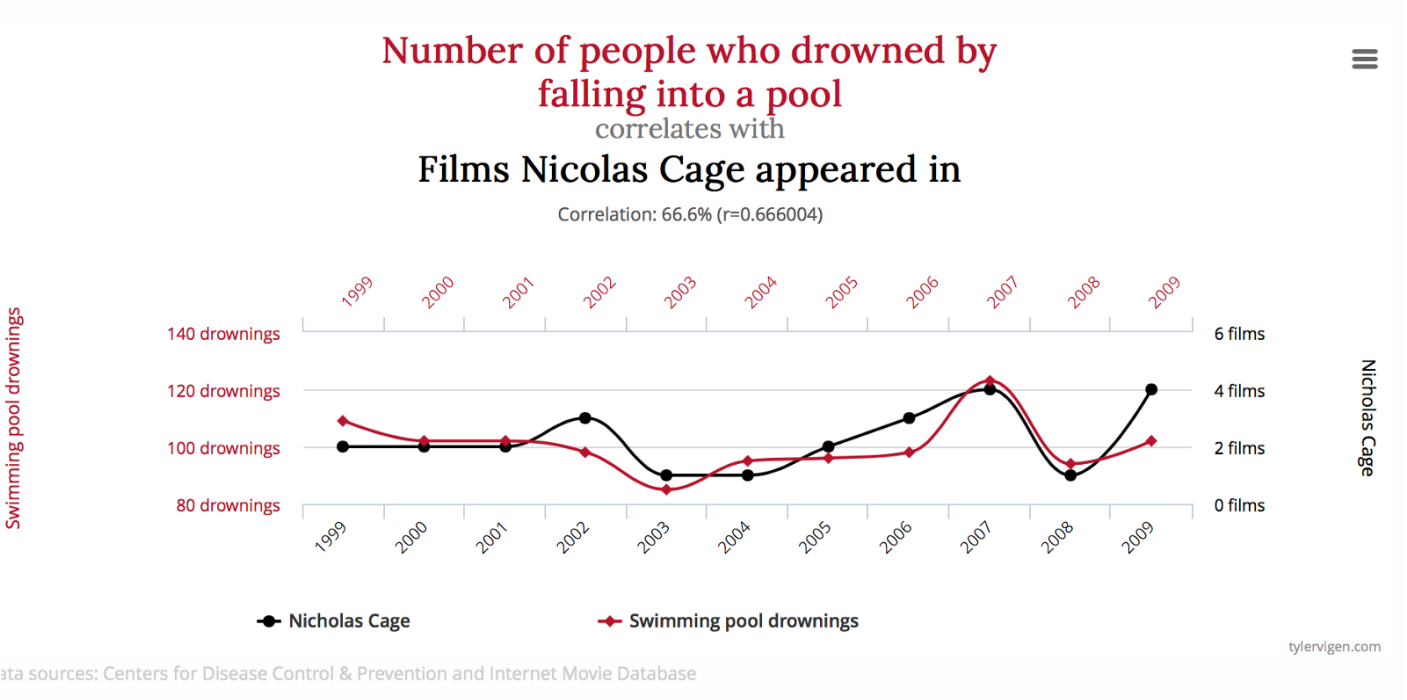
Alla base delle convinzioni di Anderson, che da un positivismo scientifico già sorpassato scivolano pericolosamente verso un incondizionato fideismo tecnologico (che ha poi finito per prendere il nome di *datismo*), vi è l'idea che, nell'era del petabyte, la nozione di correlazione sostituisca quella di causalità, consentendo alla scienza di progredire senza la necessità di confrontarsi con modelli coerenti, teorie unificanti o spiegazioni meccanicistiche: questo, se da una parte è pericoloso nella misura in cui alcune discipline si pongono effettivamente l'obiettivo di indagare determinati rapporti causali, dall'altra diventa deleterio per quelle scienze che nemmeno si muovono in quel campo ma piuttosto mirano all'elaborazione di teorie più generali, che in un

ipotetico mondo «datista» non troverebbero spazio da nessuna parte. Riprendendo lo statistico George E. P. Box che trent'anni fa diceva che «tutti i modelli sono sbagliati, ma alcuni di questi sono utili», Anderson delinea un mondo che «si spiega da solo» semplicemente attraverso i dati che ne catturano le caratteristiche, dove i numeri parlano da sé e dicono tutto quello che c'è da dire. Questa «teoria della non-teoria» rinuncia in partenza – se addirittura non si contrappone – all'idea di costruire un qualsiasi modello in grado di inquadrare tutte quelle informazioni che emergono dai dati all'interno di nozioni e principi generali che si propongono di interpretare la realtà.

In termini di approccio, la proposta alla base del metodo scientifico storicamente prevede la costruzione di un modello teorico, che altro non è se non un'ipotesi (o una serie di ipotesi) da verificare. Il modello viene poi testato attraverso una serie di esperimenti che, a seconda del loro esito, validano o confutano le ipotesi di partenza e determinano in questo modo la coerenza del modello stesso, sia internamente, sia sul piano della coerenza con le evidenze empiriche che emergono dal mondo reale. Il modello teorico che emerge nella sua formulazione (più o meno) definitiva al termine del processo di validazione è, di fatto, un sistema (complesso) che il ricercatore immagina per interpretare le relazioni esistenti tra i singoli elementi che lo compongono: in altre parole, è una semplificazione che lo scienziato propone come schema di riferimento per analizzare il comportamento delle variabili sotto osservazione e produrre degli *statements*, delle considerazioni, che ne determinano l'impatto sul sistema.

Da Cartesio in poi, la scienza ha lavorato secondo questa direttrice per secoli. Se è vero infatti che il dibattito sulla rappresentazione della realtà sotto forma di dati si è alimentato del mito «leibniziano-cartesiano» della descrizione matematica dei processi, questo dibattito è sempre stato filtrato dall'idea che il dato sia «carico di teoria», ovvero espressione tangibile di un fenomeno, che va però inquadrato attraverso i meccanismi della logica deduttiva.

Per questa ragione, lo scienziato è abituato al fatto che la nozione di correlazione non implichi automaticamente quella di causalità e sa che non bisogna trarre conclusioni sulla sola base di una correlazione perché è necessario comprendere le dinamiche che connettono tra loro due dati e, solo una volta che si è costruito il modello, è possibile tentare di connettere gli insiemi dei dati. Tutto questo perché i dati, senza un modello – ovvero senza quella teoria di cui sono carichi, che è in grado di spiegare coerentemente il modo in cui questi sono in relazione tra loro –, sono solo *rumore*, ovvero un segnale indesiderato che si sovrappone all'informazione elaborata all'interno del sistema.



Secondo Anderson, questo approccio alla scienza, basato sul processo ipotesi-modello-test, è ampiamente superato se confrontato con la disponibilità massiva di dati. Gli esempi che porta, dalla fisica alla biologia, bollano i modelli come approssimazioni che nel corso del tempo si sono scoperte essere sempre meno accurate fino a quando non vengono sostituite da altri modelli, più accurati ma pur sempre sbagliati. In sostanza, secondo questo punto di vista, più nozioni apprendiamo su una scienza, più ci troviamo lontani da un modello che la spiega. Logica conseguenza di questo ragionamento è la ricerca di un nuovo approccio metodologico, che Anderson, come già accennato, identifica nel mantra dell'era dei petabyte: *correlation is enough*. In questa prospettiva, la scienza può permettersi di smettere di cercare modelli e finalmente analizzare i dati senza dover prima ipotizzare cosa questi dovrebbero dimostrare. Lo scienziato oggi – racconta Anderson – può comodamente prendere i numeri, lanciarli in cluster di calcolo dalle dimensioni inimmaginabili e lasciare che siano gli algoritmi statistici a identificare i modelli dietro a quei pattern che il vecchio metodo scientifico non era in grado di individuare.

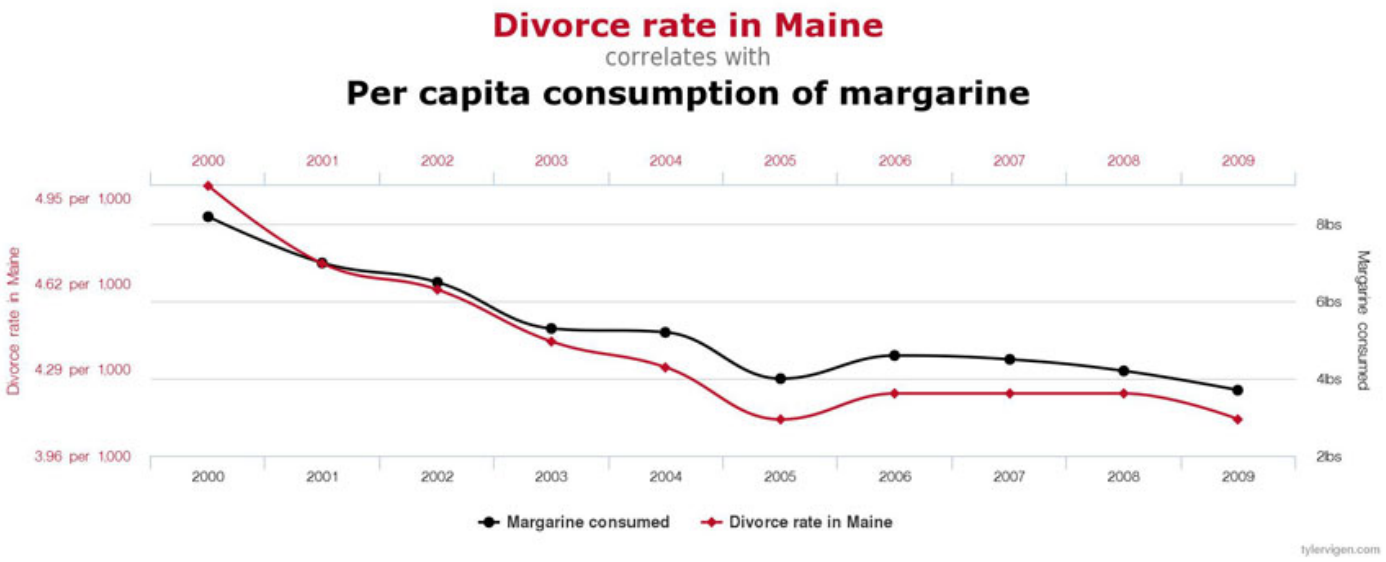
Narrazione dei dati e non-neutralità della scienza

Tuttavia, c'è chi sostiene che se da un lato è vero che i dati siano in grado di comunicare una serie di contenuti, dall'altro però risultano fortemente influenzati dalla lettura che si dà di quegli stessi dati e – contestualmente – dalla narrazione che vi si costruisce attorno. In questo senso, lo scienziato, e in particolare il data scientist, assume un ruolo tutto nuovo da interpretare: quello dello *storyteller* dei dati, per quanto personalmente preferisca la definizione di «narratore». Per far sì che il ricercatore interpreti attivamente questo ruolo senza rimanere marginale rispetto all'evoluzione che lo investe, è utile immaginare che affianchi alle «tradizionali» tecniche fondate sulla data analysis alcuni strumenti di elaborazione teorica, che, attraverso ad un approccio che sia

qualitativo o quantitativo, siano in grado di restituire il quadro di complessità necessario a dare una lettura della realtà che vada oltre al dato fine a sé stesso.

Ha infatti senza dubbio ragione Anderson quando ci ricorda che i modelli sono delle approssimazioni della realtà, ed è per questo logico immaginare che col progresso tecnico degli strumenti a disposizione queste approssimazioni diventino sempre più precise. Ma al tempo stesso non si può dimenticare che i modelli nascono col preciso obiettivo di essere una semplificazione della realtà, uno schema che consente di ricondurre una singola manifestazione della realtà ad una classe di fenomeni che mostrano una serie di caratteristiche in comune in grado di restituire una lettura *sensata* dell'esistente.

Un esempio molto brillante di questa necessità di attribuire un senso ai dati che osserviamo è Spurious correlations, un database online messo a punto da Tyler Vigen che con pungente ironia passa in rassegna una serie di correlazioni più o meno assurde che occorrono nella realtà tra variabili che all'apparenza non c'entrano nulla l'una con l'altra. Il risultato che ne viene fuori è volutamente grottesco: ad esempio, nel periodo tra il 1999 e il 2009 l'investimento della spesa pubblica statunitense in scienza, spazio e tecnologia correlava al 99,79% col tasso di suicidi per strangolamento, il tasso di divorzi nel Maine col consumo pro capite di margarina al 99,26%, il numero di comparse di Nicholas Cage in un film con quello di persone annegate in una piscina al 66,6% e tanti altri ancora. Insomma, il messaggio che ci consegna questo lavoro sembra un invito a non prendere per esauriente tutto ciò che emerge da una correlazione statistica o, per dirla come non la direbbe mai Anderson, *correlation is not enough*!



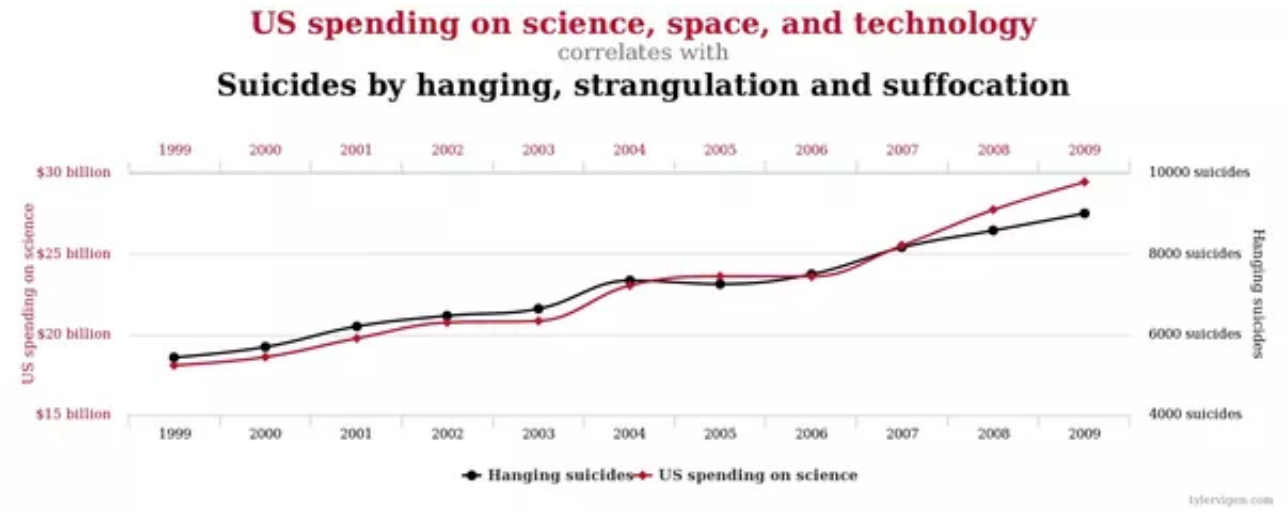
In effetti, se partiamo dalla definizione più scolastica di correlazione, ovvero quella relazione tra due variabili statistiche tale per cui a ciascun valore della prima variabile corrisponde con una «certa regolarità» un valore della seconda, è abbastanza facile immaginare le ragioni per cui la correlazione da sola non risulti essere una spiegazione abbastanza robusta per chiarire la natura della relazione che lega due fenomeni tra loro.

Dando più o meno per scontato che il ruolo dello scienziato sia quello di fornire spiegazioni sufficientemente convincenti su come funziona il mondo a partire dall'osservazione dei singoli fenomeni, tracciando quindi un modello abbastanza flessibile e generalista da riuscire a descrivere una serie di fattispecie cui si riconducono determinate caratteristiche comuni, immaginare il completo superamento della nozione di modello sembra estremamente distorsivo rispetto agli obiettivi che la ricerca si pone o quantomeno dovrebbe porsi.

Questo perché costruire modelli fa in qualche modo parte della natura umana e l'evidenza empirica non può essere altro che una rappresentazione di quei modelli che tanto ci sono utili per leggere (e interpretare) la realtà. In *Why model?*, quel genio pazzo di Joshua M. Epstein[2], sostiene che «chiunque si avventuri in una proiezione, o provi a immaginare come si sviluppa una certa dinamica sociale sta elaborando un modello; ma generalmente si tratta di un modello implicito i cui assunti sono nascosti, la cui coerenza interna non è testata, le cui conseguenze logiche sono ignote e la cui relazione con i dati è sconosciuta».

La scelta quindi non è se scrivere modelli o meno, ma se esplicitarli o lasciare che rimangano impliciti.

Tutto questo dovrebbe bastare a ridimensionare quell'idea per cui le teorie scientifiche emergono dai dati e ne sintetizzano un risultato: sembra infatti più sensato immaginare che le teorie precedano e in qualche modo indirizzino la raccolta dei dati perché cercare correlazioni senza avere un modello a cui riferirsi, di fatto, equivale a smettere di chiedersi il perché delle cose aspettandosi che questo si palesi da solo. Rovesciando il pensiero di Anderson, un mio vecchio professore diceva sempre preoccupato: «Stupidi dati, non vi lascerò rovinare la mia bellissima teoria!», ben sapendo che spesso la scienza ha fatto passi da gigante sulla base di intuizioni che violentavano i dati piegandoli alla teoria. Al netto di questo spunto, però, nessuno nega l'importanza di confrontare la teoria con le evidenze empiriche, ma senza una teoria, per dirla con le parole di Epstein, non sarebbe chiaro quali dati raccogliere.



Se è vero infatti, riprendendo la vecchia massima di Lord Kelvin, che *to measure is to know*, ovvero «misurare è sapere», d'altra parte dobbiamo costantemente confrontarci con l'idea che la scienza, qualsiasi cosa essa sia – perché io onestamente non ne ho idea –, è un sistema costruito da uomini e donne di cui, piuttosto non sorprendentemente, i principali fruitori sono proprio uomini e donne. Tutto questo implica la necessità di individuare un metodo fruibile attraverso cui filtrare ed inquadrare le informazioni che ci vengono trasmesse dai dati – dove per dato intendo ogni manifestazione misurabile e categorizzabile della realtà.

A questo punto, posto che queste manifestazioni misurabili e categorizzabili costituiscono un filtro, ovvero una lente attraverso cui osservare la realtà, credo sia fondamentale puntualizzare che la loro interpretazione non è oggettiva in tutto per tutto, non può esserlo, ma soprattutto non ci è utile in nessun modo immaginarla in questo senso.

Quello che potrebbe essere utile e interessante fare, invece, è ragionare sul senso e sulle implicazioni che ha il fatto di dare una lettura e costruire una narrazione di ciò che i dati ci restituiscono. Questo, lo puntualizzo, non significa farsi ostaggio di qualsiasi mistificazione dei fatti, quanto piuttosto riflettere su tutte quelle contraddizioni che emergono nel momento in cui ammettiamo che la lettura – l'interpretazione – che diamo dei dati è intrinsecamente soggettiva.

Infatti, in questi tempi in cui accarezziamo le sfumature della post-verità (prima o poi dovevo usarla, questa parolina magica!), se da una parte dobbiamo più che mai misurarci con l'idea che esiste una distanza tra un fatto oggettivo – o meglio il dato che lo rappresenta – e la percezione che gli si attribuisce, dall'altra il discorso attorno a questo tema non si esaurisce tutto qui.

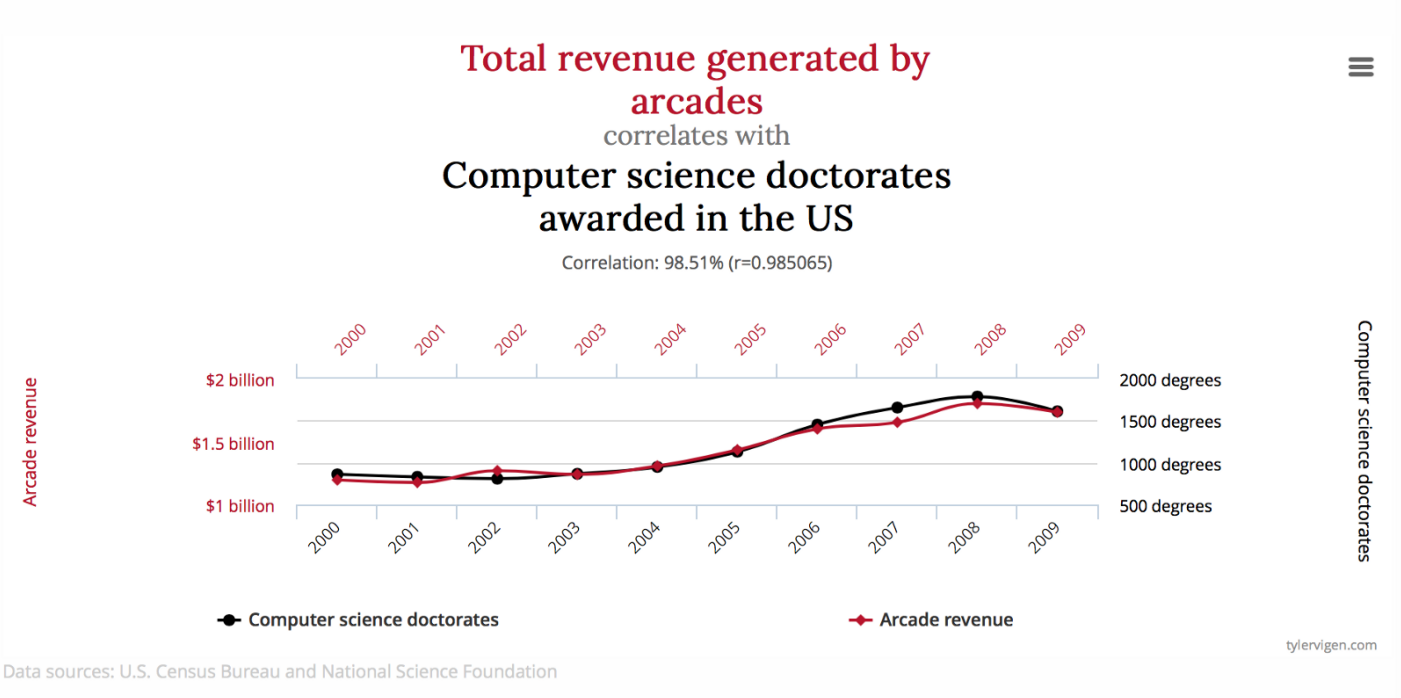
Parlare di interpretazione e interpretabilità dei dati, infatti, significa ammettere che la scienza non è un soggetto neutrale. Per contro, intendere un dato scientifico come qualcosa di oggettivo significa cedere il campo a chi costruisce quella narrazione con l'arroganza di spacciarla per l'unica verità assoluta (e l'esempio più lampante di questo atteggiamento è la violenza con cui la narrazione neoliberista si è imposta nel dibattito economico – e, di riflesso, sulle nostre vite, ma questa è

un’altra storia). Quanto pericolosa si sia rivelata (e possa nuovamente rivelarsi) questa deriva credo sia sotto gli occhi di tutti.

In effetti, prendendo in prestito le parole di Aniello Lampo[3] in *Sulla non-neutralità della scienza* (2011), «nel dibattito pubblico la scienza viene spesso presentata come un’entità super-partes, portatrice di una verità imparziale che trascende le ideologie ed i conflitti. […] In realtà, il sapere scientifico è pesantemente sovradeterminato dal contesto sociale e politico in cui viene elaborato: il lavoro dello scienziato risponde a delle domande, rispetta delle priorità e ha delle applicazioni, in cui si annidano interessi economici e politici, frutto dei rapporti di forza che innervano la società intorno. Sono questi fattori che definiscono le linee di ricerca, plasmano l’organizzazione materiale del mondo accademico e ne dettano i metodi di indagine. Di conseguenza, la scienza finisce inevitabilmente per essere situata politicamente».

Ed è in questo senso che possiamo affermare che la scienza non è neutrale ma intrinsecamente «di parte». Ammettere tutto questo è un passaggio cruciale per chi di mestiere racconta quello che i dati dicono – e, a volte, a prima vista, non dicono – perché significa assumersi le responsabilità del ruolo che in quanto ricercatori abbiamo nella società, che grosso modo è la costruzione dei saperi a beneficio della comunità che ne fruisce.

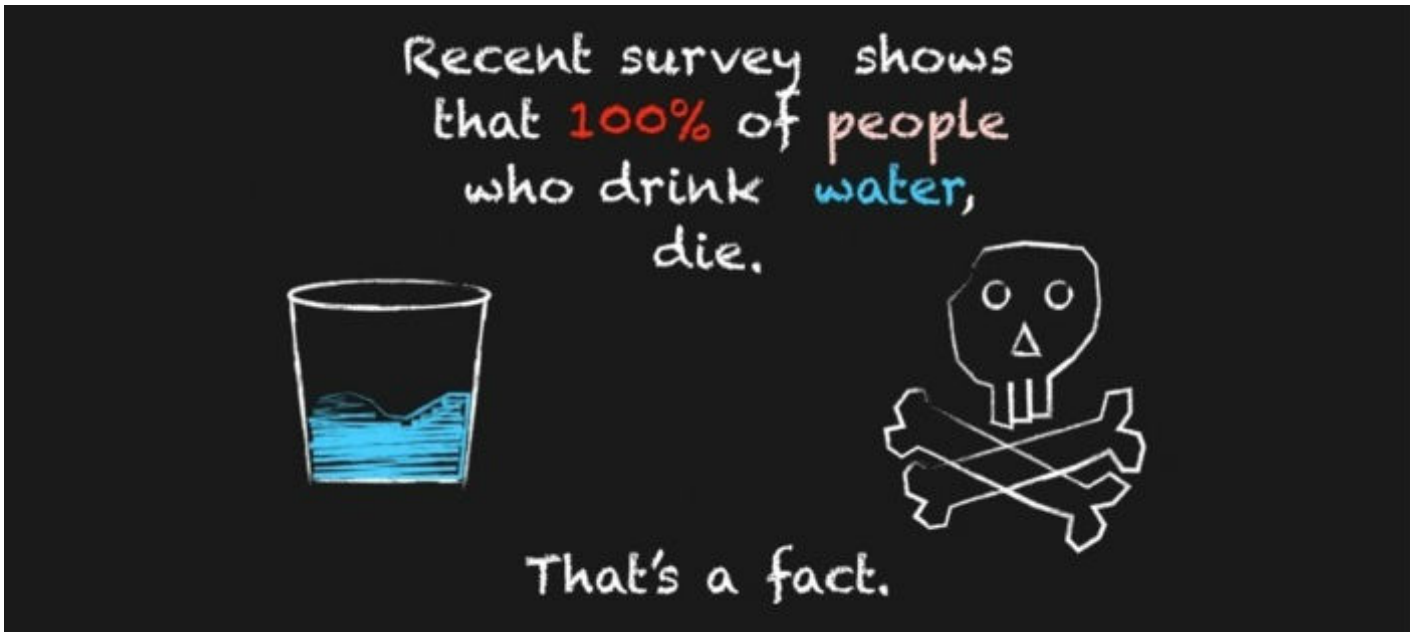
Parlare di interpretazione e interpretabilità dei dati, dunque, ha a che fare col concetto di democrazia nel senso più autentico del termine: quel concetto di democrazia per il quale, attraverso un dibattito orizzontale e accessibile, la costruzione dei saperi diventa un patrimonio collettivo; e lo stesso concetto di democrazia che garantisce che quella comunità che partecipa attivamente alla costruzione dei saperi – mentre contemporaneamente ne fruisce – possa in ogni momento rimettere in discussione i termini di quel dibattito.



Questo quadro idilliaco e petaloso della scienza come una comunità in movimento che «cammina domandando» è senza dubbio estremamente affascinante, ma, come sempre, il diavolo sta nei dettagli, e il confronto con la realtà modifica i termini del discorso. Guardando ai fatti, ci si mette poco a rendersi conto che quella comunità non si compone solo di disinteressatissimi intellettuali che hanno a cuore il solo amore per il sapere, ma che ci sono anche tanti *stakeholders* decisamente di peso, che tendenzialmente coincidono coi colossi dell'informazione che detengono la proprietà dei big data stessi e degli strumenti con cui questi vengono raccolti. In questo senso, sembra allora opportuno interrogarsi su quanto l'interpretazione dei dati dipenda dal palco da cui il ricercatore parla, che tendenzialmente rivela tanti degli interessi che stanno dietro a ogni ricerca, e da come incide sull'interpretazione dei dati il rapporto con quei colossi, a cui la ricerca è legata nei diversi gradini della filiera di produzione della conoscenza. Guardando alle cose con un po' di lucidità, non si può negare che questi rapporti incidano sia nei gradini a monte della filiera – ovvero che chi detiene i «mezzi di produzione», cioè la proprietà di quelle miniere di dati potenzialmente utili alla ricerca, è in grado di condizionare la direzione che prende la ricerca –, sia in quelli a valle – ovvero come il mercato dell'informazione, utile per divulgare al pubblico i propri studi, in realtà riesca a condizionare la ricerca in partenza. Tutto questo, però, non deve svalutare tutte le riflessioni di cui sopra sulla necessità di rendere la scienza uno spazio democratico, quanto piuttosto aiutarci a spostarli da un piano teorico e ideale ad uno sostanziale. In effetti, in una fase storica in cui sono le aziende e i soggetti privati a egemonizzare la gestione e l'utilizzo delle informazioni che i dati registrano, parlare di scienza come di un patrimonio collettivo, capace di smarcarsi da una serie di interessi privati ad oggi non trascurabili per peso, significa restituire sostanza a quei ragionamenti, e per fare questo bisogna, oggi più che mai, ragionare in termini di riappropriazione di quello spazio collettivo.

In questo senso, sono diverse le cose da fare: prima di tutto, bisogna rimettere al centro del dibattito pubblico il tema dell'indipendenza della comunità scientifica come patrimonio di tutte e tutti. Bisogna farlo sia per diffondere un modello culturale che alla logica dell'interesse particolare di pochi contrapponga quella dell'interesse collettivo, sia per incentivare i decisori politici a prendere dei provvedimenti che sappiano rispondere alle questioni che solleva l'utilizzo massivo dei big data nello spazio della ricerca, sia attraverso un serio rifinanziamento del sistema pubblico di ricerca e della ricerca di base, sia attraverso l'introduzione di un quadro normativo e giuridico al passo coi tempi ed attento alla tutela dei diritti digitali in quanto diritti dell'individuo, ma anche in quanto diritti sociali e della collettività. Fare pressioni in questa direzione non significa giocare questa battaglia tutta sulla difensiva delegandola alle istituzioni, ma affiancare ad alcuni strumenti di resistenza altri strumenti di rilancio attivo, che moltiplichino gli spazi di discussione indipendenti ma mettano anche radicalmente in dubbio e decostruiscano quelle modalità con cui si fa ricerca che più fanno emergere la contraddizione tra proprietà dei dati e obiettivi della ricerca intesa come bene comune.

I big data in tutto questo processo svolgono un ruolo fondamentale in termini di tecnologia a disposizione di chi fa ricerca e ne costituiscono uno strumento di supporto utilissimo e validissimo, ma lanciano anche alcune delle sfide che si riveleranno cruciali per il futuro della ricerca come patrimonio collettivo. Se infatti, per la dimensione intrinsecamente umana del *perché* facciamo scienza, questi non riescono ancora a sostituirsi allo scienziato nell'operare quell'attribuzione di senso che è il fine ultimo della ricerca, sta alla comunità scientifica il compito di presidiare la produzione dei saperi come patrimonio di tutte e tutti e di inserire i big-data in questo processo, verso la costruzione di un sapere sempre meno accentrato nelle mani di pochi interessi individuali e sempre più al servizio del benessere della collettività.



Brano tratto da *Datacrazia*, recentemente pubblicato da D editore nella collana Eschaton.

Ringraziamo l'autrice e la casa editrice per la disponibilità.

[1] Petabyte = circa 10^15 byte = 1 milione di gigabyte

[2] Joshua M. Epstein, *Why model?,* in *JASSS – Journal of Artificial Societies and Social Simulations*, 2008.

[3] Aniello Lampo, *Sulla non-neutralità della scienza*, in *SPIN – Scientific and Precarious researchers Independent Network*, 2011.