



theFuture ofScience andEthics

Rivista scientifica a cura del Comitato Etico
della Fondazione Umberto Veronesi

Volume 1 numero 2 ■ novembre 2016



**Fondazione
Umberto Veronesi**
– per il progresso
delle scienze

Articoli

Big Data, Sed Data.
L'era degli algoritmi,
dal potere dei dati
al mistero della
narrativa

*Big Data, Sed Data.
The era
of algorithms,
from the power
of data to the
enigma of narrative*

GIANNI RIOTTA
griotta@Princeton.edu

AFFILIAZIONE
Princeton University, NJ, USA

ABSTRACT

Il testo propone un'analisi del ruolo e dell'impatto che la scienza dei Big Data può avere per la società contemporanea. Partendo da una breve ricostruzione storica, l'articolo delinea quali scenari presenti e futuri si aprano oggi per la scienza, l'economia, l'etica e la politica alle porte di una nuova rivoluzione globale che riguarda come i dati possono essere generati, conservati e interpretati.

ABSTRACT

The text analyses the role and impact that Big Data science might have on contemporary society. Departing from a brief historical reconstruction, the article outlines the present and future scenarios for science, economy, ethics and politics that emerge at the dawn of a global revolution that concerns how data are currently produced, stocked, and interpreted.

KEYWORDS

Big Data
Big Data

Rivoluzione dell'informazione
Information revolution

Diritti politici
Political rights

Quando gli amici della rivista mi hanno commissionato un articolo sui Dati ho aderito con entusiasmo, lieto di collaborare con l'amico Umberto Veronesi. Dal Corriere della Sera alla Stampa, dal Tg1 al Sole 24 Ore, Umberto mi aveva sempre offerto totale disponibilità di scienziato, medico, umanista, per affermare diritti alla conoscenza e alla libertà di ricerca. Ma lo studioso non era che una parte di Veronesi, accanto c'era l'uomo, con gli affetti familiari, la verve, l'energia, la cultura, l'allegria, la gioia di vivere. Non dimenticherò mai, quando mia madre ebbe ad ammalarsi, come Umberto si sia prodigato in consigli, interventi, sostegno, salvandole la vita con i suoi collaboratori. Mancherà ora a chi gli voleva bene, alla scienza, all'Italia e a me, tanto. Questo articolo, in questa rivista che gli era preziosa, è dedicato dunque alla memoria di Umberto Veronesi.

Nel 1424, la biblioteca dell'Università di Cambridge conteneva 122 volumi, quanto un uomo, pur nella vita più breve del XV secolo, poteva aspirare a leggere tutti e almeno un paio di volte. Un vescovo italiano, intorno al 1420, impiegò tre tipografi stampatori per creare in circa tre mesi trecento volumi, mentre un monaco amanuense ne copiava a stento un centinaio nell'intera vita, aguzzando la vista e gelando nell'insonnabile *Scriptorium*. Nel XIX secolo lo Zar poteva spedire un protocollo urgente ai suoi dignitari in Siberia e sperare che il suo diktat li raggiungesse in sei mesi, per poi attendere la tremebonda risposta in un periodo di tempo altrettanto lungo, a patto che l'inverno fosse clemente. Per scambiarsi il primo telegramma attraverso l'Oceano Atlantico, la Regina Vittoria a Londra e il presidente americano Buchanan a Washington restarono trepidanti per diciotto ore, finché il 16 agosto 1858 il messaggio non sbucò dal cavo lungo duemila miglia. Durò poco, già in settembre andò fuori uso, ma che le notizie potessero diffondersi in ore e non più in settimane o mesi, stregò i nostri antenati. Oggi gli esperti di Big Data parlano di yottabyte, unità di misura dell'informazione pari a unità bit 8.000.000.000.000.000.000.000.000, e calcolano quante unità servano a immagazzinare Google, quante per Facebook, o per l'intero web, che spesso raddoppia e triplica la sua riserva di informazione in un attimo rispetto agli amanuensi¹. La diffusione dei dati, poco importa

Big Data,
Sed Data.
L'era degli
algoritmi,
dal potere
dei dati
al mistero
della narrativa

Articoli

quanto imponenti, avviene in tempo reale. L'intera Library of Congress americana, la formidabile biblioteca che raccoglie a Washington lo scibile umano, conta su oltre 10 petabytes (10^{16} bytes), ma viene inesorabilmente superata dal web nell'autunno del 2012.

Il lettore non si faccia troppi scrupoli nel cercare di afferrare la vastità di queste cifre, gli studiosi Kenneth Niel Cukier e Viktor Mayer Schönberger ne offrono una immagine grafica: il web ha tanti dati quanti ne potete trascrivere in vecchi CD, da allineare poi in cinque colonne parallele, su e giù dalla Terra alla Luna². Ma sono appunto immagini fantastiche, senza senso diretto nella nostra percezione umana: la rivoluzione dei Big Data non sta nella loro sterminata massa, chiara a chiunque in una giornata abbia avuto accesso al web, usato una carta di credito, mandato una mail, usato un Gps, un Bancomat o un Telepass, banali atti quotidiani generatori di dati. Sta nel filo che noi umani riusciremo, o non riusciremo, a tessere attraverso i dati, per governarli in una narrativa comprensibile, razionale, scientifica. La sfida in corso nel XXI secolo, nella società, nelle aziende, nelle università, in politica, economia e cultura, è questa: chi navigherà, e con quali algoritmi di selezione e narrativa semantica, i dati? Gli algoritmi come saranno organizzati? Che "pregiudizi" riceveranno in eredità dai loro sviluppatori umani³? Chi ha accesso ai dati? In Europa la politica difende i cittadini, negli Usa le aziende chiedono priorità. E lo Stato ha o no diritto ad accedere ai dati del telefonino di un terrorista, come nel caso che ha opposto in California Apple ed Fbi dopo la strage di San Bernardino⁴? Lo scandalo dei metadati dello spionaggio americano Nsa, rivelati dall'ex agente Snowden, parla solo di questa sindrome: chi governa i dati, governa il XXI secolo, come il vapore nel XIX, petrolio e nucleare nel XX.

Big Data è un concetto difficile da definire; se sappiamo con precisione che ad usare il termine "software" per la prima volta fu un matematico di Princeton University, John Tukey, nel 1958 su "American Mathematical Monthly", è in corso una battaglia per l'onore del battesimo di Big Data⁵. Francis X. Diebold, economista dell'Università di Pennsylvania, ha scritto un paper dal titolo "A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline" datando l'uso al 2000⁶, ma gli contendono il titolo John Mashey e altri col-

laboratori di Silicon Graphics già da metà anni Novanta, gli informatici Weiss e Indurkha, il data analyst Douglas Laney di Garter⁷. Non si tratta di una inutile acribia filologica, indagare sulla genesi dei Big Data illustra magnificamente il carattere multidisciplinare della disciplina, dove matematici, fisici, statistici, informatici, economisti, microbiologi, neuropsichiatri, epidemiologi, designer, videografici, webmaster, giornalisti, semantici, sono tutti, volta a volta, indispensabili nel creare dai dati un coerente algoritmo e conseguenti modelli di rappresentazione, analitici e di narrativa.

Per le aziende, da Wall Street, all'economia industriale classica alla *new economy* digitale, i Big Data sono spesso già in *house*, accumulati in anni, se non decenni di lavoro. Serve però ricrearli in un modello razionante, capace di aiutare il *management* nelle decisioni. Consideriamo, solo per fare un esempio, l'industria delle assicurazioni, dove previsioni del tempo, delle epidemie, dei disastri naturali, dei movimenti sociali ed economici, diventano ora assai più cogenti e predittivi di un tempo, permettendo di assicurare non più vaste fasce omogenee di popolazione divise per età, censo, sesso, residenza, ma singoli individui, maschi, femmine, giovani, anziani, con polizze individuali *ad hoc*⁸.

Qui sorge uno dei dilemmi morali che punteggiano il nuovo mondo dei dati: quando le mappature genetiche saranno completate, e quindi ciascun essere umano saprà alla nascita a quali malattie genetiche è esposto, chi dovrà controllare le informazioni? Le compagnie potranno dire, fammi leggere il tuo genoma completo e ti dimezzo la polizza, o sarà vietato? E come useranno i datori di lavoro queste informazioni? Cinquemila dipartimenti di polizia di varie città americane, da Boston a Chicago, usano i dati della compagnia Coplogic (gioco di parole che suona Logica da Sbirro)⁹ per prevenire criminalità e disordini in anticipo, analizzando la conversazione sociale sulla rete, gli eventi programmati nelle aree a rischio, i movimenti di certi individui o gruppi collettivi. I risultati son buoni per l'ordine pubblico ma fanno alzare subito le orecchie di chi difende la *privacy*, come il gruppo Palantir Technologies di Palo Alto, in California, che teme una persecuzione da algoritmo fissata contro certi ceti urbani¹⁰. Perfino l'automobile senza pilota, o il Tir guidato da robot, Gps e sensori che la precederà, pongono dilemmi etici: salvare la vita degli occupanti o di un

pedone, rischiare una manovra che investe un solo passante salvando un gruppo, proteggere il veicolo o un animale, annullare il pilota se innestasse un'azione kamikaze¹¹?

In politica è stato il presidente Barack Obama, con qualche timidezza nel 2008 e con poderoso spiegamento di forze tecnologiche nel 2012, a impugnarne i Big Data con successo. Ad ogni elettore, spiega la studiosa Zeynep Tufekci¹² del Berkman Center for the Internet and Society presso Harvard University, veniva assegnato un certo punteggio, in scala, per calcolare quanto fosse disposto a votare per il partito democratico e quanto fosse solito poi recarsi davvero alle urne. Il "vault", una banca dati custodita a Chicago, indica dunque su quali elettori fare leva e quali abbandonare, specificando l'invio di mail ecologiste a chi è abbonato a una rivista ambientalista, sulle armi ai cacciatori, sulla Cina a chi ha cognome cinese e via dicendo¹³.

La deformazione che questo processo ha indotto sul dibattito democratico è stata però drammatica, e imprevedibile dai suoi tecnocratici registi. Anziché, almeno in astratto, rivolgersi all'intero corpo elettorale come nell'epoca della democrazia di massa, il leader mira solo, grazie a dati ed algoritmi, a mobilitare la sua base, polarizzando inevitabilmente il dibattito e gli umori profondi, già scossi dalla crisi finanziaria del 2008 e dalla crisi della classe operaia e del ceto medio, seguita ai processi internazionali di automazione. I social media poi, grazie all'effetto delle *echo chambers*, comunità di cittadini che la pensano allo stesso modo e che, inevitabilmente, gli algoritmi di Google e Facebook tendono a raccogliere insieme, completano l'effetto tribale. Ciascuno di noi si ritrova online a ragionare solo con chi la pensa come lui, e se nel calcio o nei consumi privati questo può essere divertente in politica e nella cultura ha effetti nefasti nella chiusura mentale, studiati da Nicholas Carr e Cass Sunstein¹⁴. Il fisico italiano Walter Quattrocchi, docente all'Imt di Lucca, è stato uno dei pionieri degli studi sulle false notizie (con Farida Vis del World Economic Forum), i greggi culturali online e le *echo chambers*, notando come il dibattito online si isterilisce, grazie alla manipolazione dei dati, in un muro contro muro senza maturazione collettiva¹⁵.

Le elezioni appena concluse negli Stati Uniti però offrono agli analisti

nuove, contraddittorie evidenze. Il presidente Obama ha offerto alla candidata del suo partito, Hillary Clinton, il supporto della banca dati comune, da cui per qualche tempo era stato escluso, in punizione, il senatore socialista Bernie Sanders alle primarie. E la campagna della ex segretario di Stato ha seguito la falsariga del 2012, puntando alla base identificata degli elettori via dati. Il neo presidente repubblicano Donald Trump ha invece contraddetto il modello dominante negli ultimi anni, sbaragliando gli avversari di partito nelle primarie, e i democratici con una sorpresa straordinaria, malgrado, a lungo, il partito gli abbia negato, o concesso solo in parte, accesso ai dati. Nate Silver, studioso di dati che per il New York Times ha previsto Stato per Stato con precisione i risultati delle ultime presidenziali fino al 2012, dava ancora nella scorsa primavera solo il 2% di chance di nomination a Trump, tenendola appena intorno al 20% nella fase finale della corsa. In un suo libro, Bibbia degli studiosi di dati¹⁶, Silver postulava come il fiuto, il naso, l'istinto delle professioni, venga oggi superato dallo studio dei dati, dal baseball, all'economia, alla scienza. Anziché lavorare su campioni limitati, come da sempre fanno gli scienziati empirici, i Big Data permettono di elaborare teorie e conclusioni su sterminate campionature, fino all'insieme totale dei casi, N=Tutto. Silver insegnava dunque a diffidare dall'esperienza personale, anche nel baseball, nello sport, e anche grazie a "nerd" come lui perfino il calcio oggi si nutre di Big Data¹⁷.

Donald Trump invece è partito contraddicendo ogni dato, "di pancia" come si sarebbe detto un tempo. Ha condotto una campagna senza mai badare agli *spin doctor*, agli uomini dati, alle statistiche, spesso facendo l'opposto di quanto veniva da loro indicato. In una qualche ancestrale, primitiva comunicazione Trump ha raggiunto, via tv, media, giornali, social e comizi, la coscienza di chi non si sentiva rappresentato dalle infinite caselle, tutte uguali, degli algoritmi, scaldandone le emozioni. Solo alla fine si è rivolto a una azienda di studio dei dati, Cambridge Analytica, che però predilige un approccio psicologico, non quantitativo, ricercando gli atteggiamenti culturali e morali degli elettori, e allineandoli in parallelo alla proposta dei candidati. Trump ribalta il modello Obama 2012 non va a ricercare i propri elettori per identificarli grazie ai dati, lascia esplodere il messaggio populista e sanguigno online e poi esplo-

ra chi lo segue nelle *echo chambers*, mobilitandolo¹⁸. La vittoria di Trump non segna, come qualche sempliciotto ha provato a sostenere nell'eterno luddismo italiano, la fine dei sondaggi, dei Big Data, della politica con gli algoritmi¹⁹, che resteranno con noi a lungo. È però un monito solenne contro le cifre senza anima, le catene analitiche senza messaggio, i Big Data senza narrativa, come il World Economic Forum aveva predetto da tempo. Illudersi che, dalla medicina allo sport, la politica e l'economia, i dati cessino di essere il nostro pane quotidiano è follia.

Cadere nel miraggio di dati che parlano senza la nostra fatica di comprensione e rappresentazione è però come pensare di farsi una cultura, comprando l'intera biblioteca di Babel ma senza mai aprire un libro²⁰.

BIBLIOGRAFIA

- Carr, Nicholas (2011), *Internet ci rende stupidi? Come la rete sta cambiando il nostro cervello*, trad. it. Raffaello Cortina, Milano.
- Mayer-Schönberger, Viktor, Cukier, Kenneth N. (2013), *Big data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, trad. it. Garzanti, Milano.
- Riotta, Gianni (2013), *Il web ci rende liberi?*, Einaudi, Torino.
- Riotta, Gianni (2014), *Is the web really a gift from God?*, World Economic Forum, <https://www.weforum.org/agenda/2014/06/digital-media-strategy>.
- Silver, Nate (2013), *Il segnale e il rumore*, trad. it. Fandango, Roma.

NOTE

1. <http://gizmodo.com/5955437/the-internet-archive-has-now-saved-a-whopping-10000000000000000000-bytes-of-data>

2. Mayer-Schönberger & Cukier 2013.

3. Che un algoritmo possa avere “pregiudizi” sorprende di solito il lettore non addetto ai lavori, ma basta riflettere che si tratta solo di un “filtro” redatto da un essere umano per cogliere, o non cogliere, certi dati: riflette dunque sempre personalità, cultura e motivazioni del suo autore, con rischio di discriminazioni e intolleranze: http://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html?_r=0

4. techcrunch.com/tag/apple-vs-fbi/

5. <http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/>

6. <https://economics.sas.upenn.edu/pier/working-paper/2012/origins-and-development-term-%E2%80%99Cbig-data>

7. <http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/>

8. <https://www.ft.com/content/3273a7d4-00d2-11e6-99cb-83242733f755>

9. <https://www.coplogic.com/>

10. <http://www.nytimes.com/2016/11/07/technology/the-risk-to-civil-liberties-of-fighting-crime-with-big-data.html>

11. <http://moralmachine.mit.edu/http://ed.ted.com/lessons/the-ethical-dilemma-of-self-driving-cars-patrick-lin>

12. <http://firstmonday.org/ojs/index.php/fm/article/view/4901>

13. Riotta 2013.

14. Carr 2011; Cass Sunstein, “Echo Chambers” Princeton University Digital Books, <https://press.princeton.edu/sunstein/echo.pdf>

15. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110

16. Silver 2013.

17. Allenando il Manchester City l'allenatore italiano Roberto Mancini modificò, secondo le raccomandazioni del team dati, il modo di tirare i calci d'angolo, migliorando il quoziente di gol. La vicenda sollevò infinite polemiche, come sempre nel soccer <https://soccerlogic.wordpress.com/2014/05/06/analytic-insights-and-dubious-corner-stats/>

18. <http://www.wsj.com/articles/inside-donald-trumps-data-analytics-team-on-election-night-1478725225>

19. Sarebbe utile una volta riflettere sulle remote cause culturali del pregiudizio anti scientifico del nostro Paese, dal “Mathematica sunt non leguntur”, caro a Benedetto Croce, allo scetticismo diffuso nei media per ogni strumento statistico o non di “pancia”. Un qualunquismo cronico che risale a Leo Longanesi, Maccheri, il giovane Montanelli, la cultura provinciale dello Strapaese, gli “apoti”, quelli che non la bevono, e preferiscono sogghignare tra amici in provincia, anziché affrontare il mondo a viso aperto.

20. Riotta 2014.

Big Data,
Sed Data.
L'era degli
algoritmi,
dal potere
dei dati
al mistero
della narrativa

Articoli



**Fondazione
Umberto Veronesi**
– per il progresso
delle scienze