Ordine dal Caos: il Tao dei Big Data

Paola Pasca

1. Introduzione

L'evoluzione dei *personal* computer, in termini di memoria e calcolo, ha portato ad una progressiva digitalizzazione dei dati. Informazioni prima presenti in archivi fisici sono state convertite in forma digitale e rese disponibili ad un vasto pubblico. Al contempo, lo sviluppo del *web* e dei *social network* ha creato quello che è stato definito da Cukier e Mayer-Schoenberger (2013) come *datafication*, ossia la trasformazione di molti aspetti della vita delle persone in dati. I *like*, i movimenti rilevati dai dispositivi per il fitness, i messaggi scambiati attraverso le *app* rivelano comportamenti, preferenze, spostamenti ed interazioni. Inizia così l'era dei *big data*, che pervade globalmente il XXI secolo. I *big data*, infatti, hanno saputo affermarsi e trovare applicazione nei più svariati settori: da quello medico a quello aziendale, da quello psicologico a quello della sicurezza informatica.

Esiste una visione condivisa dei *big data* come sinonimo di caos, disordine, diversità. Queste definizioni sono legate ad alcune caratteristiche distintive che li rendono unici e preziosi, le cosiddette *5 v* dei *big data* (Yin & Kaynak, 2015):

- *Volume*, a rappresentare le enormi quantità di dati generati ogni secondo. Tutto ciò che facciamo nel mondo digitale lascia una traccia: di conseguenza, la quantità di dati disponibili è letteralmente esplosa. Secondo Marr (2016), abbiamo creato più dati negli ultimi due anni che in tutta la storia dell'umanità. Si stima, inoltre, che entro il 2020 ogni essere umano genererà 1.7 Megabytes di nuovi dati al secondo e che saranno più di 50 bilioni i dispositivi attivamente connessi a Internet. In sintesi, la quantità di dati disponibile continuerà a crescere in modo esponenziale sino a raggiungere livelli inimmaginabili.
- *Velocità*, ossia la rapidità con cui questi dati vengono generati e si muovono. Si pensi, ad esempio ai post sui social network che diventano virali nel giro di pochi secondi, alle centinaia di ore di video che carichiamo su Youtube oppure alle più di 200 milioni di e-mail che inviamo ogni minuto, ogni giorno, attraverso Google. Il flusso dei *big data* impone la necessità di sviluppare e disporre di piattaforme che permettano di elaborarli, archiviarli e analizzarli in modo più efficiente (Ishwarappa & Anuradha, 2015). In questo contesto, la tecnologia informatica dei *database relazionali* permette di archiviare grandi quantità di

La differenza come risorsa. Atti del Workshop 2018 • ISBN 978-88-8305-143-2 • p. 31-46 •
DOI Code: 10.1285/i9788883051432p31

informazioni in tabelle specifiche, separate, ma comunicanti tra loro attraverso delle *chiavi*. Queste ultime garantiscono la possibilità di accedere alle informazioni in maniera dinamica, permettendo ad un utente di effettuare delle vere e proprie "interrogazioni" dei dati. Di pari passo al flusso dei dati, anche i database relazionali sono stati arricchiti, sino ad arrivare a piattaforme di archiviazione ed elaborazione dei dati in tempo reale (Schuff, Paradice, Burstein, Power & Sharda, 2010).

- *Varietà.* Ad un aumento della velocità corrisponde un analogo aumento nella varietà: oltre a dati strutturati in forma di tabelle (i tipici database aziendali), attualmente si dispone di (e si cerca di utilizzare) dati di tipo non strutturato, quali foto, video, post dai *social network*. Si stima che la parte non strutturata rappresenti, attualmente, dall'80 al 90% dei *big data* attualmente esistenti;
- *Veridicità,* caratteristica più critica che nasce come conseguenza delle prime. L'enorme varietà, il flusso continuo e il disordine che caratterizza questo tipo di dati rende la loro qualità e accuratezza molto meno controllabili: la possibilità di manipolarli porta l'immaginazione oltre il limite delle classiche analisi statistiche, mescolandosi e lasciandosi contaminare dal settore dell'informatica e della programmazione, oltre al settore specifico cui quei dati fanno riferimento. Estrarre conoscenza e significato dai *big* data è un processo fatto di sperimentazioni continue, di operazioni ripetute ed affinate sino a far emergere gli elementi rilevanti e degni di nota (O'Neil, & Schutt, 2013). Ciò è particolarmente vero nel caso di dati linguistici destrutturati (*tweets*, posts, discussioni su *forum, social network*), dove scovare una chiara consistenza semantica può risultare particolarmente difficile. Ci si chiede, dunque, quanto siano affidabili le previsioni o assunzioni provenienti da fonti così diverse. I progressi scientifici e le recenti innovazioni legate all'uso dei *big-data* rivelano che essi non solo rappresentino una miniera di conoscenze prima inaccessibili/inimmaginabili, ma che possano integrare ed arricchire i classici dati in forma strutturata, numerica, precisa.

Va da sé che l'unico modo per fare dei *big data* uno strumento utile al miglioramento e all'evoluzione della società sia la capacità di accogliere, inglobare e trattare in modo oculato la diversità strutturale che li caratterizza. La logica alla base dell'estrazione e creazione di senso da così grandi differenze è simile per tutti gli ambiti di applicazione: i *big data* vengono usati per nutrire algoritmi che imparano a riconoscere le regolarità presenti nella vastità di parole, pixel, numeri (O'Neil, 2017; O'Neil & Schutt, 2013). In termini semplici, se un oggetto ha determinate caratteristiche (ad es. quattro piedi, una superficie piana, uno schienale, un'altezza di tot. cm) allora sarà classificato come "sedia"; se l'altezza è superiore,

e non vi è lo schienale, si tratterà di un "tavolo". Così via per tutte le caratteristiche prese in esame dall'algoritmo e da cui esso apprende (James, Witten, Hastie, & Tibshirani, 2013).

- La possibilità di trovare l'ordine nel disordine, di trasformare il caos in struttura, la differenza in risorsa finisce per produrre rappresentazioni sintetiche della realtà che gettano luce su aspetti psicologici e sociali della nostra epoca. Semplicemente, crea *Valore* in diversi settori. Vale la pena esplorare quelli che esercitano un impatto sostanziale sulla vita quotidiana.

2. Il Valore dei Big Data: complessità come ricchezza

Dal punto di vista psicologico, l'ultima frontiera della ricerca fa ampio uso di *big data* destrutturati per decodificare le caratteristiche di personalità, sia sane che patologiche: si analizza il linguaggio naturale delle persone nelle loro interazioni quotidiane e sul web. La ricerca condotta da J. Pennebaker (2011) e dal suo team ha messo in evidenza il valore sociale di alcune parole. Grazie all'analisi di migliaia e migliaia di scritti personali, *e-mail*, testi di canzoni, discorsi politici, *post*, commenti su forum e social network, è stato possibile scoprire dei *pattern* di parole che, se presenti assieme, caratterizzano in modo preciso gli stili di pensiero e la personalità degli individui: dallo stile di pensiero *formale* tipico di persone adulte, preoccupate per il proprio potere e status sociale, mentalmente sane ma meno oneste con se stesse e con gli altri, allo stile di pensiero *analitico*, tipico di individui cognitivamente complessi che cercano di comprendere il mondo attorno a loro.

Dalle parole che riflettono attenzione rivolta all'interno di sé (come il pronome *Io* e parole che fanno riferimenti a processi di pensiero), tipiche di chi sta vivendo emozioni negative o di condizioni psicopatologiche quali la depressione (Arseniev-Koehler, Mozgai & Scherer, 2018; Rude, Gortner & Pennebaker, 2004), alle parole che uniscono le persone (*noi* e simili), tipicamente presenti in chi sperimenta emozioni positive, o nei discorsi su *social network* immediatamente successivi a tragedie su larga scala, o ancora utilizzate per affermare il proprio potere sociale nelle interazioni. Così le parole diventano marcatori psicologici.

Altri studi (Pennebaker & King, 1999) hanno documentato e confermato la stabilità e l'affidabilità del linguaggio naturale in termini psicometrici, offrendo nuovi spunti allo studio della personalità: sapendo che individui estroversi tendono a utilizzare più parole che fanno riferimento a emozioni positive (*incoraggiare, gioire, piacere*), assieme a parole che fanno riferimento a contesti sociali (*amico, ascoltare, collaborare*) ed ai pronomi in terza persona (*lui,*

lei, parole queste che riflettono un *focus* attentivo rivolto verso il mondo esterno) e che aspetti *nevrotici* della personalità risultano associati al pronome personale *Io*, nonché a parole che fanno riferimento a emozioni negative (*purtroppo, rimpianto, sconforto*), diventa quasi possibile immaginare una tecnologia in grado di valutare le caratteristiche psicologiche degli individui in modo automatico, senza fare appello ai classici test ed inventari tipici della tradizione psicologica. I primi tentativi di previsione automatica della personalità (Argamon, Dhawle, Koppel & Pennebaker, 2005; Oberlander, & Nowson, 2006) danno ancora risultati parziali, ma promettenti: le *performance* migliori si ottengono proprio incorporando informazioni provenienti da diverse fonti (da categorie psicologiche presenti *software* validati a database linguistici aperti e accessibili a tutti) suggerendo che il potenziale dei *big data* risieda, in questo caso, proprio nella loro diversità strutturale (Mairesse, Walker, Mehl & Moore, 2007). Ciò rappresenterebbe una rivoluzione non solo nelle scienze umane e sociali, ma anche in tutti quei settori (aziendale, clinico, militare) in cui la valutazione psicologica e la selezione del personale riveste particolare importanza.

Altre eleganti applicazioni sfruttano i *big data* di tipo testuale sia per identificare lo stile narrativo di potenziali autori di manoscritti anonimi (Boyd & Pennebaker, 2015) che di individuare marcatori linguistici predittivi del comportamento suicidario (Shing, Nair, Zirikly, Friedenberg, Daumé III & Resnik, 2018; Stirman & Pennebaker, 2001).

Largo uso di queste tecniche viene fatto anche nell'ambito della sicurezza informatica: commenti, recensioni e finti utenti possono essere identificati e classificati come *spam* in modo molto rapido utilizzando alcuni *marker* linguistici, in genere discrepanti nella forma e nella coerenza delle frasi, assieme alla velocità di pubblicazione (es. numerosi commenti in un dato e ristretto intervallo temporale) sul *web* (Dewang& Singh, 2018, Mahmood & Afzal, 2013). In modo simile, anche le *e-mail* fraudolente presentano caratteristiche linguistiche specifiche e riconoscibili: parole studiate per creare coinvolgimento e un clima fiducia col ricevente (Alli, Nicolaides & Craig, 2018).

Maggior cautela, rispetto alle applicazioni, si osserva nell'ambito medico, dove i *big data* provengono dalle fonti più diverse (resoconti medici, test diagnostici, banche dati molecolari, spese mediche e farmacologiche, sensori e dati provenienti dai *social media*). La ricerca di regolarità e di *pattern* specifici può partire dai singoli pixel di una TAC o risonanza magnetica, oppure dagli esiti di esami diagnostici (Leff & Yang, 2015). Negli ultimi decenni, i *database* genetici e molecolari, ovvero archivi relativi alla struttura di materiale genetico sano o anomalo, si sono andati affermando sempre più. Ciò ha ampliato le conoscenze e le

potenzialità in campo biomedico, fornendo nuovi spunti per gestire il problema dell'eterogeneità che caratterizza i sistemi complessi, quelli biologici in particolare (Michael, Ma, Fisher, Kreisberg, Raphael & Ideker, 2018). Pur potendo disporre di dati relativi al codice genetico di milioni di pazienti, questi ultimi potrebbero non essere comunque sufficienti: la risposta dell'organismo umano a determinate condizioni è un campo complesso, in cui le variabili in gioco sono numerosissime. Fortunatamente, le tecniche di apprendimento statistico e di *machine learning* vengono in aiuto per creare un modello affidabile che permetta di predire al meglio la probabilità di verificarsi di condizioni patologiche quali cancro o cardiopatie. In quest'ambito, l'impiego dei *big data* permette, da un lato, di risolvere almeno in parte il problema dell'eterogeneità, dall'altro di identificare dei *pattern* biologici che siano identificabili, nonché più chiari e comprensibili (Michael, et al., 2018).

Hood e Flores (2012) sottolineano che lo scopo dell'impiego dei *big data*, in ambito biomedico, sia quello di creare un sistema di apprendimento dai dati continuo, al contempo:

- predittivo, che permetta, cioè, di identificare anomalie patologiche sia attraverso un monitoraggio longitudinale del patrimonio genetico, ma anche di individuare soggetti a rischio di condizioni croniche attraverso il sequenziamento del DNA familiare, al fine di sviluppare misure preventive efficaci;
- preventivo, attraverso la sperimentazione di farmaci e tecnologie che agiscano su più livelli all'interno dell'organismo;
- partecipativo, valorizzando la collaborazione e il contributo di figure professionali differenti, per garantire una presa in carico sanitaria efficace ed efficiente del paziente.

Mehta e Pandit (2018) evidenziano come la differenza nelle fonti di provenienza e nella struttura dei *big data* di tipo medico getti le basi per il passaggio da "una medicina basata sull'esperienza a una medicina basata sull'evidenza" (p. 62).

L'impiego dei *big data* e delle tecniche di apprendimento statistico dai dati rappresenterebbe, dunque, non solo una nuova frontiera per il miglioramento della vita delle persone, ma anche una possibilità di ottimizzare, in termini economici, trattamenti e misure interne alle strutture sanitarie. Bayati, Braverman, Gillam, Mack, Ruiz, Smith, e Horvitz (2014) hanno utilizzato dati relativi a pazienti ospedalizzati a causa di problemi cardiaci per predire la probabilità di una futura ospedalizzazione entro i successivi 30 giorni. Il modello sviluppato per predire il tasso di ricaduta è stato poi inserito all'interno di un sistema di supporto alle decisioni degli operatori esperti (infermieri e operatori

ospedalieri): solo i pazienti che venivano identificati come "a rischio" ricevevano un trattamento preventivo speciale. Si è potuto stimare che lo strumento di supporto alle decisioni, assieme al tipo di intervento preventivo, permettevano di ridurre l'ulteriore ospedalizzazione dei pazienti del 18,2%, diminuendo i costi dei trattamenti del 3,8%. Appare chiaro che più i costi di eventuali trattamenti aumentino, più un intervento mirato permetta un'ottimizzazione delle risorse e dell'investimento finanziario.

Tuttavia, alcuni dei più grandi limiti riconosciuti al trattamento dei *big data* in ambito medico sono proprio la mancanza di analisti che sappiano combinare al meglio dati così destrutturati, complessi e delicati; senza contare le questioni etiche relative, ovviamente, alla *privacy* e alla tutela delle informazioni personali di individui e pazienti.

In ambito militare e rispetto alla sicurezza pubblica, l'uso dei *big data* assume sfumature quasi inquietanti: si pensi alla creazione, da parte del governo americano, di ARGUS, un sistema di *imaging* all'avanguardia in grado di tracciare qualsiasi oggetto in movimento, umano o animale, con precisione centimetrica (Baesystemsinc, 2013). La sua capacità di monitoraggio viene paragonata a quella di centinaia di predatori che osservano un'area grande quanto una città di medie dimensioni, tutti nello stesso istante. Per dare un'idea della precisione del drone, i suoi sviluppatori portano l'esempio un'immagine relativa a un'area di 24 km quadrati, catturata da ARGUS da un'altezza di più di 5000 metri, in Virginia (USA): a partire dalla visione globale del territorio, è possibile selezionare l'area da ingrandire per ottenere un video di ciò che sta accadendo in un quartiere, nelle strade, sino a poter osservare persone o animali, pur mantenendo la visione dell'insieme su uno schermo globale. Curioso è che questa super-tecnologia derivi dai comunissimi sensori fotografici presenti in qualsiasi *smartphone* che, assemblati a formare un mosaico, creano un dispositivo fotografico ad altissima precisione (1.8 Gigapixel). I piani governativi rispetto all'impiego di ARGUS rimangono sconosciuti, ma è possibile pensare che, installato su un drone militare, possa sorvegliare il territorio a migliaia di metri d'altezza, identificando e bloccando catastrofi e minacce terroristiche (Anthony, 2013). Piani così ambiziosi diventano pensabili solo assieme ad altri dispositivi dotati di un'elevata capacità di immagazzinamento dei dati prodotti dai sensori (ben 600 GB al secondo) e di appositi *server* in grado di elaborarli in tempo reale.

3. Paradossi e lati oscuri

Se da un lato queste recenti applicazioni appaiono positive e rassicuranti, dall'altro mettono in luce sfumature paradossali e inquietanti: il traguardo di una società più competitiva, produttiva, efficiente, trasparente e sicura viene raggiunto attraverso "processi che monitorano, controllano, reprimono o persuadono le persone. Esiste una linea sottile tra l'utilizzo dei *big data* per favorire l'emancipazione e l'*empowerment* e l'uso che va a scapito di alcune persone, o del tutto opposto ai buoni propositi che ne avevano animato lo sviluppo" (Kitchin, 2014, p. 206). È come se il positivo e il negativo, la libertà e la costrizione, si manifestassero, in questo contesto, in modo simultaneo. Così, se da un lato i clienti di un negozio risparmiano denaro acquistando con le loro *card*, dall'altro i sistemi che custodiscono le loro informazioni le utilizzano per riproporre loro nuove offerte pubblicitarie e incrementare i profitti. Allo stesso modo, il prezzo per la sicurezza da minacce terroristiche diventa la sorveglianza invasiva di tutti i membri della società: la sicurezza ha il prezzo della *privacy* (Kitchin, 2014). L'immaginario futuro distopico descritto da Orwell (1949) nella sua opera *"1984"* non sembra essere così immaginario, in un'epoca in cui i *big data* e la tecnologia pervadono a 360° la vita delle persone. Di fatto, vivere il quotidiano senza lasciare almeno una traccia digitale è diventato quasi impossibile: anche se non si effettuano acquisti con carta di credito all'interno di un negozio, si viene filmati dalle telecamere di sorveglianza; anche se si accede ai *social network* con uno pseudonimo, gli indirizzi IP e MAC originari vengono comunque registrati nel *web*. Secondo Koops (2011), l'Autorità per la Protezione dei Dati olandese stima che l'individuo olandese medio sia presente in 250-500 *database*, mentre un individuo più socialmente attivo risulterebbe inserito in ben più di 1000 database. Questi archivi di dati riguardano non solo le impronte "digitali" delle persone (i dati cioè che generano di proposito) archiviate in rete, ma anche una serie di informazioni velate (generate cioè da altri, quali dispositivi di sorveglianza, app, *smartphone*). Nell'insieme, le impronte "digitali" e le informazioni velate forniscono un quadro completo e dettagliato della vita quotidiana dell'individuo: le sue abitudini di consumo, i suoi spostamenti, le interazioni e la sua sfera sociale. Mai come ora, tali informazioni sono così facilmente generabili e disponibili (Kitchin, 2014).

Purtroppo, l'esplosione dei *big data* ha anche aperto la strada a fenomeni negativi quali la sorveglianza e lo sfruttamento attraverso essi: da un lato il monitoraggio per prescrivere, persuadere, governare (Raley, 2013); dall'altro, la profilazione psicologica e il *targeting* col solo scopo di trarre profitto, al di là di ogni etica. Chiaro esempio di quest'ultimo *modus*

operandi è il caso *Cambridge Analytica*: attraverso alcune falle del sistema Facebook, nonché violando i termini d'uso di dati che erano stati raccolti a scopo di ricerca, i profili psicologici di milioni di cittadini americani sono stati utilizzati dalla società di consulenza britannica per la campagna di *marketing* politico legata a Donald Trump (Lewis & Hilder, 2018). Nel momento in cui le questioni politiche e sociali che sono particolarmente a cuore ad un certo *target* di popolazione sono note, la costruzione di messaggi politicamente orientati e l'esposizione continua delle masse a tali messaggi ha un effetto manipolatorio e polarizzante delle opinioni. *Profiling* ed esposizione massiccia a siti web e messaggi di parte, in modo ricorsivo, sono stati la base della strategia utilizzata da *Cambridge Analytica* che, dopo l'esplosione dello scandalo, ha dichiarato la bancarotta (Valsania, 2018).

In realtà, non c'è alcun mistero dietro l'attività di raccolta dati e profilazione da parte di Facebook: al momento dell'iscrizione sui social network e *app* di società partner di Facebook, gli utenti forniscono autorizzazioni al trattamento dei dati personali. Così si diventa parte di un *database*. Le informazioni che si forniscono, le pagine a cui si lascia un *like*, assieme alle funzioni di geo-localizzazione degli *smartphone* permettono di costruire la profilazione psicologica più dettagliata al mondo (Mantovani & Consoli, 2018).

In un primo studio, Kosinski, Stillwell, e Graepel (2013) hanno dimostrato che i *like* di facebook sono dei predittori attendibili di orientamento sessuale, etnia, visione politica e religiosa, intelligenza, felicità, nonché di comportamenti patologici quali abuso di sostanze, ma anche della separazione dei genitori, di età e genere. Successivamente, assieme alla rilevazione dei *like* su Facebook, i ricercatori hanno somministrato degli inventari di personalità e creato un algoritmo in grado di predire le caratteristiche di personalità degli individui a partire dalle loro interazioni sui *social network* (Youyou, Kosinski, & Stillwell, 2015): i risultati mostrano, in modo sorprendente, che l'algoritmo produce un quadro psicologico più accurato di quello fornito da un collega di lavoro, semplicemente analizzando 10 *likes*; 70 *likes* sono sufficienti per ottenere un profilo accurato quanto quello fornito da un amico o da un coinquilino, 150 *likes* superano il giudizio di un genitore, fratello o partner. 300 *likes* permettono al computer di superare, in accuratezza, il coniuge. Da un lato gli studiosi, entusiasti, ipotizzano un futuro in cui i computer potrebbero essere in grado di inferire lo stato interiore degli esseri umani e di reagire di conseguenza. D'altra parte, di pari passo all'evoluzione tecnologica si presentano questioni e preoccupazioni relative alla *privacy* degli individui: i ricercatori sottolineano la necessità di creare politiche che garantiscano agli individui il pieno controllo sulle loro impronte "digitali", dal

momento che la tecnologia può essere così pervasiva da influenzare il tipo di lavoro che si svolge, i risultati elettorali e addirittura la ricerca del partner della vita.

Opportunità e sfide etiche si mescolano: se da un lato, ad esempio, messaggi sanitari tarati sulle caratteristiche psicologiche delle persone potrebbero favorire cambiamenti verso stili di vita più salutari, dall'altro messaggi più subdoli, quali pubblicità di casinò *online* dirette a persone che mostrano tratti associati al gioco patologico, sfrutterebbero le debolezze degli utenti a danno degli stessi (Matz, Kosinski, Nave & Stillwell, 2017).

Altro aspetto problematico legato alla diversità e al caos che caratterizza i *big data* è un atteggiamento *naïf* e un po' superficiale che li considera fonte di verità assoluta e sinonimo di qualità superiore solo per la loro natura. Anderson (2008) sostiene che la disponibilità di dati di così ampie dimensioni renda obsoleto anche il metodo scientifico: la scoperta della "verità" si ridurrebbe alla mera osservazione di associazioni che esistono naturalmente e non sarebbe necessaria alcuna competenza settoriale per fare nuove scoperte. Secondo Mazzocchi (2015) quello di Anderson è un articolo provocatorio: da esso si è sviluppato un grande dibattito tra un tipo di ricerca fondata sui dati e quello, tradizionale, fondato sulle ipotesi e sul classico metodo scientifico. Dibattito che merita di essere affrontato anche dal punto di vista epistemologico: sin dall'antichità esiste anche una visione ed una conoscenza empirica e induttiva, sostenuta da filosofi quali Bacone (fautore egli stesso del metodo scientifico) che si contrappone al più classico metodo deduttivo. Secondo Bacone, approcciarsi ad un fenomeno con dei pregiudizi inficerebbe la conoscenza e le scoperte sullo stesso. Il ragionamento deduttivo, sostenuto invece da filosofi come Aristotele e divenuto ampiamente accettato ed applicato nella cultura occidentale e nella ricerca, sarebbe dunque limitato: formulare delle ipotesi in anticipo forzerebbe il processo conoscitivo ad essere coerente con le premesse di partenza.

Lungi dalla pretesa di individuare un'ipotesi più o meno giusta, Mazzocchi (2015) afferma piuttosto che l'era dei *big data* ha portato a riconsiderare il metodo induttivo: un nuovo empirismo basato sulla tecnologia ispira un futuro in cui l'estrazione di informazioni dai *big data* porta direttamente a nuove scoperte. Secondo i sostenitori di questo nuovo approccio, sarebbero gli stessi algoritmi a procedere in maniera induttiva: si tratta di una conoscenza senza fine, che "apprende" continuamente, si affina ed elabora i dati in maniera flessibile a seconda delle possibili applicazioni. L'apprendimento costante, seppur incompleto e imperfetto, produce conoscenza utile. Fondamentale è dunque considerare i *big data* nella loro interezza e trattare in modo oculato la loro complessità.

4. Visioni parziali e danni sociali

Visioni limitate e parziali, in quest'ambito, hanno portato alla creazione di quelle che O'Neil (2017) definisce vere e proprie *Armi di Distruzione Matematica (ADM)*: modelli matematici concettualmente incompleti ed errati che, applicati ai *big data*, finiscono per creare realtà ben diverse da quelle che dovrebbero solo limitarsi a descrivere, giustificate dai modelli stessi. Il risultato è un circolo vizioso che aumenterebbe le disuguaglianze e minaccerebbe la democrazia. Secondo la *data scientist* e matematica, la struttura stessa delle armi di distruzione matematica penalizzerebbe proprio le minoranze. Esistono esempi di clamorosi danni sociali e di vita subiti dalle vittime delle *ADM*: l'adozione, da parte dell'amministrazione comunale di Washington, di un algoritmo di valutazione dell'efficienza degli insegnanti delle scuole pubbliche e centinaia di maestre elementari licenziate per aver valutato in modo onesto i loro studenti, a differenza dei colleghi che avevano preferito "gonfiare" il rendimento degli allievi solo per risultare insegnanti *top* alla luce del modello di *rating*. Oppure il caso della valutazione universitaria da parte di una rivista statunitense, la *U.S. News and World Report*. A partire, nei primi anni '80, da un modello di valutazione dei college estremamente rudimentale (basato sul solo giudizio dei presidi rispetto alla qualità dell'università da loro stessi gestita) ad un'arma a doppio taglio perfezionata per segnare il destino dei college stessi: se pochi studenti conseguono la laurea, se pochi di loro versano ingenti contributi alla propria università, se pochi conseguono un punteggio soddisfacente ad alcuni test di ammissione, il college viene valutato negativamente. Ne consegue che quelle università saranno scartate da professori validi e studenti eccellenti, "condannate", dunque, dal giudizio di un algoritmo su un costrutto complesso come la qualità dell'istruzione. Altro aspetto socialmente rilevante in cui l'effetto negativo dell'ottica riduzionista rispetto ai *big data* si manifesta in modo brutale è l'emarginazione sempre più forte a cui sono destinate quelle persone che, per aver sofferto di problemi psicologici o per essersi trovate in difficoltà economiche in passato, sono etichettate come lavoratori scarsi e inadeguati, o come clienti inaffidabili dagli algoritmi utilizzati da aziende di reclutamento di forza lavoro e di *banking*. Queste minoranze sono condannate a rimanere tali da un circolo vizioso che si auto-alimenta: nessuno vorrà concedere loro un prestito, o dar loro una seconda opportunità di inserirsi nel mondo del lavoro o in un contesto sociale produttivo in cui poter ripartire da zero. Così l'algoritmo segna il destino delle persone, in modo immediato, irreversibile, drastico, come una ghigliottina.

Ciò che O'Neil (2017) mette in discussione è l'approccio riduzionistico: già conosciuto e discusso per i suoi effetti deleteri, nell'ambito dei *big data* esso finisce per produrre danni altrettanto *big* (O'Neil, & Schutt, 2013). Inoltre, le potenzialità delle grandi moli di dati non emergono in modo automatico: non si ottengono con la pretesa un algoritmo mal costruito possa rendere conto di realtà sociali complesse, meritevoli di per sé di considerazione attenta e studio approfondito. Considerare solo parte del tutto come la realtà dei fatti, in tale contesto, finirebbe per penalizzare proprio gli utenti più vulnerabili, aumentando l'indebitamento, il divario socio-economico, e relegando le differenze a rappresentare realtà isolate, destinate a non trarre ricchezza o opportunità l'una dall'altra.

Il quadro relativo alle caratteristiche dei *big data*, ai rischi e potenzialità delle informazioni prodotte direttamente o indirettamente in diversi settori, appare chiaro e completo. Gli effetti pervasivi che ne derivano sono ampiamente documentati nella letteratura: alcuni di essi hanno creato dei veri e propri scandali ma, in generale, stimolano a riflettere sugli aspetti più etici del trattamento dei dati e sulla necessità di creare un codice deontologico per gestirlo.

5. Privacy e implicazioni etiche

Rispetto al trattamento dei dati da parte di aziende e privati di vario tipo, è stato proprio lo scandalo di *Cambridge Analytica* a portare l'amministratore delegato di Facebook, Mark Zuckerberg, a processo e a dare l'avvio alla Normativa Europea in materia di protezione dei dati personali, entrata in vigore già il 24 maggio 2016 ma di fatto attuata solo nel 2018 (Saetta, 2018). La protezione dei dati è diventata un diritto fondamentale dei cittadini; per tale motivo, il *General Data Protection Regulation*, o GDPR mira sia a garantire agli utenti una scelta più attiva rispetto alla diffusione dei propri dati, sia ad uniformare la regolamentazione in materia di dati all'interno dell'Unione Europea. L'utente interessato ha il diritto di essere informato su quali tra i suoi dati vengono utilizzati e come, in piena coerenza con il principio dell'autodeterminazione informativa. L'approccio del GDPR è incentrato su un sistema di valutazione del rischio (*risk-based*) che ogni azienda dovrebbe provvedere come chiaro indicatore della propria responsabilità nel trattamento dei dati, nonché dei rischi per i diritti e la libertà degli utenti. Se da un lato si assiste ad un aumento del potere di utenti e cittadini, in precedenza sfruttati come mere fonti di guadagno o di voti politici, dall'altro, il fatto che siano le stesse aziende a dover effettuare le valutazioni dei propri rischi rende il tutto paradossale e, forse, meno controllabile di quanto non si pensi.

Si spera, in conclusione, che questo insieme di adempimenti, che rappresentano comunque un notevole cambiamento e aggiornamento per le aziende, possa di fatto garantire la tutela dei cittadini, per quanto riguarda le loro impronte "digitali".

Proprio il colosso Facebook è stato pioniere nell'adempimento alle nuove normative che, tuttavia, avrebbe limitato solo all'Unione Europea (Licata, 2018): Zuckerberg avrebbe infatti dichiarato che, nonostante la sua azienda abbia sviluppato una versione di Facebook conforme alle normative europee a livello mondiale, essa sarà utilizzata solo in UE, dove di fatto è entrato in vigore il GDPR. Ciò ha scatenato la polemica dei gruppi americani a difesa della *privacy* quali il *Center for Digital Democracy* di Washington. Mentre gran parte del mondo politico e imprenditoriale americano attacca il GDPR considerandolo un ostacolo all'innovazione e ai modelli di *business*, i gruppi a tutela degli utenti continuano a fare pressione affinché le garanzie dell'UE diventino uno standard globale ed anche altri colossi quali *Silicon Valley* e *Google* si adeguino ad esse.

Rispetto alla necessità di norme etiche relative al trattamento dei *big data*, associazioni *no-profit* si battono affinché essi siano sfruttati in modo etico e a vantaggio della comunità, stabilendo delle linee guida e muovendosi verso l'elaborazione di un codice deontologico (Catlett & Ghani, 2015, Richards & King, 2014). In primo piano c'è anzitutto il concetto di *privacy*, che non si limita semplicemente a ciò che dei propri dati può essere nascosto; piuttosto, si rifà al concetto di informazione, promuovendo una maggior consapevolezza di quali sono le regole d'uso dei dati personali e della loro diffusione. Assieme al concetto di *privacy,* assume un ruolo fondamentale anche la confidenzialità: in un'era in cui un consenso si traduce nel possesso di informazioni personali da parte di aziende partner e affiliati, una maggior fiducia in queste istituzioni è possibile solo garantendo che le informazioni private saranno tutelate dalle stesse leggi sulla *privacy* anche dai terzi a cui vengono diffuse. In ultimo, l'appello alla trasparenza crea una tensione tra apertura e segretezza: se, da un lato, la trasparenza di informazioni aziendali o governative potrebbe compromettere segreti commerciali o la sicurezza nazionale, dall'altro una scarsa trasparenza non permetterebbe agli utenti di affidarsi e fidarsi adeguatamente delle istituzioni.

Occorre pensare, dunque, ai *big data* e alla loro complessità in modo etico. Si tratta di tecnologie non "naturali": esse restano pur sempre il prodotto delle scelte umane, nonostante influenzino gli umani valori. Come emerge da questo *excursus* nel caos dei *big data*, la tecnologia tende a creare il tipo di società nella quale ci si trova a vivere e gli esseri umani che si è destinati a diventare. È responsabilità delle aziende e dei *data analysts*

produrre nuove tecnologie e scoperte nella piena osservanza delle *best practices*: in un periodo di rapidi cambiamenti, essi sono in grado di individuare discrepanze, falle e difetti prima di ogni altro e, ci si augura, di utilizzare questa consapevolezza per creare sistemi che tutelino gli individui sin dall'inizio.

6. Conclusioni

In conclusione, questo viaggio conoscitivo nel mondo dei *big data* apre numerosi spunti di riflessione: la rivalorizzazione di un'antica, seppur marginale, modalità di fare scienza e ricerca più empirica e induttiva; la differenza negli scopi e nelle modalità di utilizzo dei *big data*; la responsabilità sociale ed etica di coloro che li elaborano. Com'è possibile, volendo mutare prospettiva, rendere questo contributo una risorsa per i lettori? Forse la risposta è nella possibilità di poter osservare, a questo punto, aspetti in precedenza sconosciuti, forse considerati scontati o addirittura invisibili, con una consapevolezza nuova.

Riferimenti bibliografici

Alli, R., Nicolaides, R., & Craig, R. (2018). Detecting advance fee fraud emails using self-referential pronouns: A preliminary analysis. In *Accounting Forum*. Elsevier.

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Retrieved from: https://www.wired.com/2008/06/pb-theory/.

Anthony, S. (2013). *DARPA shows off 1.8-gigapixel surveillance drone, can spot a terrorist from 20,000 feet.* Retrieved from http://www.extremetech.com/extreme/146909-darpashows-off-1-8-gigapixel- surveillance-drone-can-spot-a-terrorist-from-20000-feet (last accessed 3 April 2018).

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. (2005). Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Arseniev-Koehler, A., Mozgai, S., & Scherer, S. (2018). What type of happiness are you looking for?-A closer look at detecting mental health from language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 1-12).

Baesystemsinc. (2013, Jan 25). *ARGUS Featured on PBS Nova* [Video file]. Disponible da: https://www.youtube.com/watch?v=13BahrdkMU8&t=30s.

Bayati, M., Braverman, M., Gillam, M., Mack, K. M., Ruiz, G., Smith, M. S., & Horvitz, E. (2014). Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one*, 9(10), e109264.

Boyd, R.L., & Pennebaker, J.W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological science*, *26*(5), 570-582.

Catlett, C., & Ghani, R. (2015). Big data for social good. *Big Data* 3(1), 1-2.

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Aff.*, *92*, 28.

Dewang, R. K., & Singh, A. K. (2018). State-of-art approaches for review spammer detection: a survey. *Journal of Intelligent Information Systems*, *50*(2), 231-264.

Hood, L., & Flores, M. (2012). A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New biotechnology, 29*(6), 613-624.

Ishwarappa, K., & Anuradha, J. (2015). A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia computer science, 48*, 319-324.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Koops, B.J. (2011). Forgetting footprints, shunning shadows: A critical analysis of the right to be forgotten in big data practice. *SCRIPTed, 8*, 229.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 201218772.

Leff, D. R., & Yang, G. Z. (2015). Big data for precision medicine. *Engineering*, *1*(3), 277-279.

Lewis, P., & Hilder, P. (2018). *Leaked: Cambridge Analytica's blueprint for Trump victory*. Disponibile da: https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory.

Licata, P. (2018). *Privacy, Zuckerberg: "La Gdpr varrà per Facebook solo in Europa. Non certo nel resto del mondo"*. Disponibile da: https://www.corrierecomunicazioni.it/privacy/privacy-zuckerberg-la-gdpr-varra-per-facebook-solo-in-europa-non-certo-nel-resto-del-mondo/

Mahmood, T., & Afzal, U. (2013, December). Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In *Information assurance (ncia), 2013 2nd national conference on* (pp. 129-134). IEEE.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research, 30*, 457-500.

Mantovani. R., & Consoli, M. (2018). Noi e Facebook: il social network è sotto accusa per come gestisce le nostre informazioni. *Focus, 307, 22-28*.

Marr, B. (2016). *Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results*. New York: John Wiley & Sons.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 201710966.

Mazzocchi, F. (2015). Could Big Data be the end of theory in science?: A few remarks on the epistemology of data- driven science. *EMBO reports, 16*(10), 1250-1255.

Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics*, *114*, 57-65.

Michael, K. Y., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J., & Ideker, T. (2018). Visible Machine Learning for Biomedicine. *Cell, 173*(7), 1562-1565.

Oberlander, J., & Nowson, S. (2006). Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedins of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*.

O'Neil, C. (2017). *Armi di distruzione matematica*. Milano: Giunti.

O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc.

Orwell, G. (1949). *1984*. London: Secker and Warburg. Print.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology, 77*(6), 1296.

Raley, R. (2013). Dataveillance and countervailance. *Raw data'is an oxymoron*, 121-146.

Richards, N. M., & King, J. H. (2014). Big data ethics. *Wake Forest L. Rev., 49*, 393.

Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121-1133.

Saetta, B. (2018). *Regolamento generale per la protezione dei dati*. https://protezionedatipersonali.it/regolamento-generale-protezione-dati

Schuff, D., Paradice, D., Burstein, F., Power, D. J., & Sharda, R. (Eds.). (2010). *Decision Support: An Examination of the DSS Discipline* (Vol. 14). Springer Science & Business Media.

Shing, H.C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In K. Loveys, K. Niederhoffer, E. Prud'hommeaux, R. Resnik & P. Resnik *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 25-36). New Orleans: CLPsych.

Stirman, S.W., & Pennebaker, J.W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine, 63*(4), 517-522.

Valsania, M. (2018). *Cambridge Analytica travolta dal Datagate: bancarotta e chiusura immediate*. https://www.ilsole24ore.com/art/mondo/2018-05-02/datagate-cambridge-analytica-annuncia--chiusura-immediata205429.shtml?uuid=AEAKYxhE&refresh_ce=1

Yin, S., & Kaynak, O. (2015). Big data for modern industry: challenges and trends [point of view]. *Proceedings of the IEEE*, *103*(2), 143-146.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036-1040.