Big Data, opportunità e rischi

di Mario Mezzanzanica

Professore associato di Sistemi informativi,
Università degli Studi di Milano-Bicocca

Nella nostra vita a livello personale, lavorativo e sociale si stanno verificando eventi che fino a pochi anni fa facevano parte solo di alcuni film di fantascienza. Informazioni di interesse, azioni che compiamo, comportamenti e molto altro sono memorizzati digitalmente da una quantità crescente di dispositivi elettronici che pervadono il mondo in cui viviamo e quindi la nostra vita. Il nostro cellulare memorizza la posizione in cui ci troviamo, le telefonate effettuate, permette di acquistare un biglietto del treno, di effettuare operazioni bancarie, di consultare il meteo, di accendere i riscaldamenti della nostra abitazione e regolarne la temperatura, di contare i passi fatti in un giorno, di monitorare i battiti del cuore. Nelle nostre auto le scatole nere memorizzano i dati sui percorsi, le soste, gli stili di guida, lo stato del motore. Nelle abitazioni e negli uffici l'energia elettrica è gestita da contatori intelligenti. Nelle aziende le macchine utilizzate per la produzione sono dotate di sensori che registrano e comunicano informazioni per ottimizzare le attività di produzione e manutenzione. In agricoltura sensori intelligenti acquisiscono dati e governano l'irrigazione delle colture. Nelle città diversi dispositivi elettronici sono utilizzati per la gestione del traffico, l'analisi dell'inquinamento dell'aria, la sicurezza degli accessi.

Questi, e molti altri esempi che potremmo descrivere, testimoniano le profonde trasformazioni che hanno preso avvio nel nuovo millennio e che manifestano, da una parte, la penetrazione dell'innovazione tecnologica nella vita delle persone e più in generale nella società e, dall'altra, la "registrazione digitale" dei nostri interessi, delle nostre azioni, delle cose che utilizziamo, generando una disponibilità crescente di dati digitali.

Innovazione tecnologica e digitalizzazione dell'informazione sono due fattori che evolvono insieme, condizionando lo sviluppo e la crescita l'uno dell'altro e costituiscono l'origine dell'era odierna dei Big Data, un termine utilizzato per rappresentare l'esplosione della quantità e diversità della disponibilità immediata, in real-time, di dati digitali. Martin Hilbert, dell'università della California, ha stimato che nel 2000 il 25% di tutta l'informazione prodotta nel mondo era registrato su supporto digitale, nel 2013 si è arrivati al 98%. Si tratta di un fenomeno emergente che ci

vede tutti coinvolti nel processo di produzione e di utilizzo di informazione digitale; le scelte, le decisioni dei singoli individui e delle famiglie trovano un supporto informativo sempre più ampio dalla disponibilità immediata di dati digitali, che si tratti di scelte legate alla vita privata, lavorativa o sociale.

Tutti noi accediamo continuamente al web per cercare informazioni e rimaniamo stupiti dalla quantità e sempre più accurata formulazione delle risposte che otteniamo. Per capire meglio il fenomeno di cui stiamo parlando occorre sapere che l'informazione digitale pubblica disponibile a cui accediamo liberamente nel web, circa otto miliardi di pagine, rappresenta solo il 4 per cento dell'informazione attualmente contenuta nelle basi di dati dei computer esistenti, mentre il 96 per cento restante è protetta da password (Bergman, Michael K. 2001). Se questa mole di dati venisse raccolta su dei CD-ROM, messi uno sull'altro, arriverebbero alla Luna in cinque pile separate (Shomberger e Cukier, 2013).

Quali sono i principali fattori connessi alla crescita vertiginosa della digitalizzazione e più nello specifico della datizzazione? Quali possono essere le opportunità e i fattori critici o di rischio per le imprese, per le istituzioni pubbliche e per le persone? Quale potrebbe essere l'impatto sul contesto sociale ed economico del mondo in cui viviamo? Queste domande sono certamente rilevanti nella società odierna dei Big data.

I fattori di cambiamento

Gli statistici dalla loro origine raccolgono ed elaborano dati per rappresentare attraverso informazioni quantitative i fenomeni di interesse e, nel tempo, grazie ai progressi della matematica e l'evoluzione delle metodologie e tecniche di analisi statistica, hanno affinato le procedure volte a migliorare le rappresentazioni numeriche in ambiti sempre più ampi del sapere con l'obiettivo primario di migliorarne la conoscenza e supportare le decisioni.

Si possono riscontrare due principali approcci all'analisi numerico-quantitativa: le indagini censuarie e quelle campionarie. Le prime prendono in considerazione l'intera popolazione oggetto dello studio fornendo il valore "vero" dei parametri di interesse come percentuali, medie e totali; un approccio che comporta indagini costose, richiede tempi lunghi di attuazione e restituisce informazioni "non attuali" per supportare molti processi decisionali. Le seconde consentono di superare alcune criticità dei censimenti attraverso la raccolta di dati riferiti ad un sottoinsieme della popolazione di interesse – un campione rappresentativo; le tecniche campionarie sono in grado di inferire l'intera popolazione al netto di un grado di errore, di incertezza, quantificabile se la scelta del campione risponde a determinati criteri di tipo probabilistico. Anche le tecniche campionarie, normalmente attuate attraver-

so la somministrazione di questionari/survey[1], richiedono tempo per la raccolta dei dati, la loro elaborazione e analisi e l'eventuale pubblicazione. Le effettuazioni di indagini campionarie sono spesso gravose (per gli intervistati e nell'organizzazione) e costosa, soprattutto se si volesse renderle scalabili verticalmente[2] cioè aumentarne la profondità (dettaglio) di analisi. In tal senso, la scelta di studiare un sottoinsieme dei dati di una popolazione di interesse comporta un compromesso: si può cercare di scoprire quello che si cerca con maggiore rapidità e con costi ridotti (rispetto ai censimenti), ma non si può rispondere a domande che non siano state poste preventivamente. Entrambe queste tecniche di analisi sono ancora oggi molto utilizzate e continueranno a generare informazioni di rilevante interesse sia a livello personale sia aziendale ed istituzionale. Inoltre, con l'avvento dell'informatica si assiste ad una evoluzione della capacità di analizzare popolazioni empiriche (che rappresentano l'oggetto della statistica), spingendosi al di là dei limiti che fino a qualche tempo fa erano resi invalicabili dall'onerosità delle operazioni (M. Martini 2004).

La traiettoria di evoluzione dei censimenti prende infatti in considerazione l'utilizzo integrato di informazioni presenti di archivi amministrativi delle pubbliche amministrazioni[3], mentre le survey, nelle fasi del loro processo di attuazione, sono supportate da applicazioni informatiche che le rendono più efficienti e performanti[4]. Ciò nonostante, abbiamo visto che queste tecniche sono nate in un momento in cui vi erano forti limitazioni nel trattamento dell'informazione ed in particolare nella capacità di elaborare ed analizzare rapidamente grandi volumi di dati.

Oggi la situazione è cambiata ed in molti campi è in atto un processo evolutivo che procede dalla raccolta di alcuni dati all'accumulo del maggior quantitativo possibile di essi, e, se possibile, di tutti, tale cioè che N= tutti (Shomberger e Cukier 2013). Questo fatto muove nella direzione di poter avere a disposizione l'intero (o quasi) set dei dati di interesse e conseguentemente di effettuare studi da diversi punti di osservazione, entrando nei dettagli e cercando risposte in sottogruppi della popolazione anche là dove un approccio campionario non sarebbe in grado di arrivare, consentendo quindi un aumento vertiginoso della scalabilità o profondità delle analisi.

Un cambio di paradigma

L'utilizzo di volumi di dati di grandi dimensioni, dei Big Data, ha come immediata

1 Oggi sempre più le informazione delle survey vengono raccolte attraverso l'utilizzo di applicazioni informatiche al fine di ottimizzare costi, tempi di raccolta ed errori di digitazione.

2 Scalabilità verticale: dal macro al micro e viceversa – es. ottenere da un campione a livello territoriale nazionale informazioni a livello territoriale comunale.

3 Importanti ed avanzate sperimentazioni in tal senso sono in essere soprattutto nei paesi del nord Europa

4 CATI - Computer Assisted Telephone Interview e CAWI - Computer Assisted Web Interviewing.

conseguenza un cambio di paradigma nell'approccio alle analisi dei fenomeni: da "risposte a domande pre-definite" (dati "precisi", strutturati, raccolti ad hoc e di "piccole" dimensioni) al "lasciar parlare i dati" (grandi volumi di dati, spesso non strutturati, presenza di "imprecisione" e scalabilità delle analisi).

Nel nostro gruppo di ricerca abbiamo iniziato alcuni anni fa a raccogliere gli annunci di lavoro che vengono pubblicati dalle aziende sui principali portali di operatori[5] che offrono servizi di intermediazione tra domanda ed offerta sul web. Dal 2013 a oggi abbiamo raccolto oltre due milioni e cinquecentomila annunci unici[6] di lavoro che giornalmente vengono pubblicati sui portali selezionati e che opportunamente elaborati consentono di analizzare le professioni richieste, le loro caratteristiche in termini di competenze-skill, il settore economico delle aziende richiedenti ed il territorio nel quale l'occupazione è richiesta. Queste informazioni sono disponibili in tempo reale e possono essere utilizzate sia per migliorare la conoscenza dei diversi operatori che si occupano di politiche e servizi sia da parte degli utenti finali (persone e imprese) per osservare dove e quali caratteristiche hanno le opportunità di lavoro offerte. Ma la potenzialità offerta dalla raccolta giornaliera di tutti gli annunci di lavoro consente di andare molto oltre; ad esempio scoprire l'emergere di nuove figure professionali, i cambiamenti in atto in molte professioni (magari legati alle richieste di conoscenze informatiche per lo svolgimento delle funzioni e/o attività), l'emergere di segnali di cambiamento in atto in settori economici dovuti alla richiesta di professioni caratterizzate da competenze particolari (ad esempio il trend verso l'Industria 4.0[7] del manifatturiero).

I principali fattori di diversità di un approccio Big Data si manifestano raccogliendo in real–time la totalità (o quasi) dei dati significativi e seguendo una logica interamente bottom-up che per certi versi è completamente opposta rispetto a quella degli strumenti tradizionali. Le offerte di lavoro scaricate dal web costituiscono infatti un patrimonio informativo totalmente destrutturato da cui è necessario estrarre informazioni attraverso opportuni filtri e meccanismi di classificazione. In questo modo la logica è essenzialmente quella del "lasciar parlare il dato" e di sintetizzare ex post le informazioni che emergono. Questo approccio risulta particolarmente adeguato per intercettare il cambiamento in atto, nella misura in cui le imprese lo esprimono negli annunci online.

Ovviamente anche le *web vacancies* non sono scevre da problematiche. La principale è costituita dal fatto che non tutte le offerte di lavoro vengono pubblicizzate online. Vi sono diversi settori e soprattutto molte occupazioni per cui i canali di

5 Portali specialistici, di operatori di servizi per il lavoro e di testate giornalistiche nazionali

6 Un annuncio di lavoro è composto da titolo, dove viene espressa la professione ricercata, e una descrizione testuale della stessa contenente competenze e altri requisiti richiesti dall'azienda; sono dati non strutturati cioè in linguaggio naturale.

7 Industria 4.0 o quarta rivoluzione industriale è la prospettiva di evoluzione dei settori produttivi in un'ottica di elevata automazione e valorizzazione in tutti i processi aziendali dei Big Data.

reclutamento seguono logiche diverse dalla semplice esposizione online di posizioni di lavoro aperte (si pensi ad esempio alle posizioni aperte dalla pubblica amministrazione che avvengono in Italia attraverso appositi concorsi che seguono logiche e tempistiche totalmente ad hoc). Tuttavia molti studi mostrano che il canale del web sta acquisendo una importanza sempre più crescente nelle modalità di reclutamento da parte delle imprese, seguendo il trend di crescente digitalizzazione della nostra società (Lee, I. 2011).

Nel corso del progetto sono state utilizzate diverse tecniche (machine learning, *information extraction* e *mining*, per citarne alcune) per analizzare la grande quantità di dati disponibile nel web. Queste tecniche sono applicabili attraverso strumenti (algoritmi) prevalentemente di apprendimento automatico; gli algoritmi utilizzati ad esempio per classificare le professioni (descritte in linguaggio naturale e quindi fortemente eterogenee tra portali e nel singolo portale), consentono di "insegnare alla macchina" come riconoscere automaticamente le professioni e, sulla base dei dati utilizzati per l'apprendimento, prendere "decisioni intelligenti" basate sull'esperienza accumulata; il risultato che viene prodotto conterrà sempre un'alea di errore o inesattezza, inversamente proporzionale alla "accuratezza" e "completezza" del set dati utilizzato per l'addestramento. La "minore qualità" o "minore precisione" di un approccio basato sui Big Data (N= tutti o quasi) rispetto a uno basato su basi di dati raccolti ad hoc e quindi di dimensioni ridotte è controbilanciata dall'aumento delle potenzialità di analisi. Lasciar parlare i dati significa favorire un approccio basato sulle domande che emergono dall'osservazione di correlazioni esistenti tra oggetti o fatti "nascosti".

Prendiamo ora in considerazione un esempio nell'area commerciale. Amazon raccomanda libri ai suoi clienti in base alle preferenze d'acquisto di ciascuno, avendo raccolto nel proprio sistema informativo una enorme quantità di dati: cosa acquistano, cosa guardano, quanto tempo dedicano ad osservare un oggetto (che acquistano o solo guardano) e molto altro. Sulla base della grande mole di dati a disposizione, due sono stati gli approcci alle analisi dei dati che si sono succeduti nel tempo: il primo costruito attraverso analisi campionarie volte ad identificare gruppi di clienti con interessi affini e finalizzato ad effettuare proposte "personalizzate"; il secondo basato sull'identificazione delle associazioni o correlazioni tra prodotti ottenuto tramite l'analisi dei dati relativi a tutti i "click" del mouse che i clienti effettuano per l'acquisto dei prodotti. Con l'utilizzo di questo secondo metodo gli analisti di Amazon prendono in considerazione l'intera base dati degli acquisti, elemento che gli ha consentito di migliorare significativamente diversi aspetti della relazione con i clienti: le raccomandazioni/proposte di acquisto diventano quasi istantanee; il metodo utilizzato può essere esportato su qualsiasi categoria di prodotto; si riducono i costi di analisi ed aumentano i casi di successo delle vendite (M. Schönberger e K. Cukier 2013).

Questo secondo approccio basato sulla correlazione tra prodotti sposta apparente-

mente l'attenzione della spiegazione di un fenomeno dal perché accade (principio di causalità) a cosa accade (osservazione delle correlazione tra oggetti o fatti). Se apparentemente e in alcuni casi questo fatto si può verificare, in molti casi la ricerca più approfondita di una connessione causale avverrà dopo che i Big Data avranno fatto il loro lavoro, quando vorremo analizzare il perché, anziché limitarci a scoprire il cosa (Shomberger e Cukier 2013). Questa attenzione al "cosa accade" come elemento che può precedere la domanda del perché ed anzi ampliarla e approfondirla è riscontrabile non solo nei contesti socio economici o commerciali ma anche relazionali, lavorativi e non, degli individui.

In una recente pubblicazione Alex Pentland[8] riporta i risultati di studi che utilizzano i Big Data per analizzare le interazioni tra persone e gli scambi informativi tra gruppi di lavoro in diversi contesti aziendali (call center, gruppi di ricerca, ecc.). Questi dati raccolti tramite l'utilizzo di badge sociometrici[9] e particolari applicazioni su smartphone[10], hanno consentito di studiare cosa accade, in termini principalmente di comportamenti di persone in un dato contesto organizzativo, con l'obiettivo di identificarne le scelte e il perché delle stesse nella prospettiva di migliorare i risultati di un gruppo di lavoro. Le analisi effettuate hanno consentito di verificare che in molti casi il principale fattore di produttività e creatività è la quantità di opportunità di apprendimento sociale che si manifesta generalmente tramite le interazioni faccia a faccia tra colleghi. In altri termini, afferma Pentland attraverso l'osservazione di una enorme quantità di dati, tante idee importanti su come raggiungere il successo e migliorare il rendimento sul lavoro possono emergere durante la pausa caffè o in mensa; ribadendo l'importanza del coinvolgimento e della partecipazione diretta, elementi che favoriscono il flusso delle idee, come fattori primari di una buona organizzazione.

I fattori di criticità

Le opportunità connesse all'utilizzo dei Big Data sono principalmente legate alla possibilità di studiare e analizzare fenomeni con livelli di puntualità, capillarità e flessibilità mai riscontrati fino ad ora. Nel contempo, tali opportunità presentano alcuni aspetti critici che vanno affrontati: chi possiede i Big Data e quale possibilità di accesso è disponibile? Quale impatto sulla concezione della privacy? Esiste un possibile rischio di dittatura dei dati?

8 Fisica Sociale, EGEA S.p.A. , Maggio 2015 – Autore: Prof. Alex 'Sandy' Pentland, direttore dello Human Dynamics Laboratory e il Media Lab Entrepreneurship Program del MIT.

9 I badge sociometrici sono dispositivi elettronici che consentono di raccogliere dati specifici sui comportamenti a livello di comunicazione delle persone che li indossano: tono della voce, linguaggio del corpo (attraverso un accelerometro), con chi e per quanto tempo parlano, ecc.

10 Sistema sensorio per telefonia mobile: applicazioni informatiche per smartphone che consentono di raccogliere dati relativi a: localizzazione, vicinanza, attività di comunicazione, applicazioni installate e attive, file utilizzati e molti altri dati. Con questi dati è possibile ricostruire automaticamente le molteplici modalità interattive dei partecipanti le sperimentazioni.

Nello scenario attuale abbiamo due principali tipologie di soggetti che certamente hanno un grande predominio nella raccolta di informazioni di cittadini e imprese: i grandi social network da una parte e le pubbliche amministrazioni dall'altra.

Facebook, Twitter, Linkedin e Google+ per citarne alcuni dei più noti, sono "servizi web" che consentono agli utenti la creazione di un profilo pubblico o semi-pubblico con opportuni vincoli di visibilità e accesso all'informazione pubblicata, l'articolazione di una lista di contatti e la possibilità di scorrere la lista di amici dei propri contatti. La creazione del proprio profilo comporta la messa a disposizione di informazioni come il proprio indirizzo email, i propri interessi e passioni e le esperienze lavorative. La "rete degli amici" si amplia continuamente e senza limiti prefissati con contatti che si propagano con "gli amici degli amici".

I servizi dei social network sono gratuiti per gli utenti e consentono ai gestori dei siti di trarre vantaggio economico principalmente dalla fornitura a terzi delle informazioni degli utenti e dalla pubblicità mirata che le aziende indirizzano agli utenti stessi e di cui possiedono dati estremamente rilevanti (siti visitati, link aperti, permanenza media, oltre a tutte le informazioni personali che gli utenti stessi hanno inserito). Quello che emerge è una base di conoscenza alimentata gratuitamente da persone e imprese che non ha precedenti e che consente ai proprietari dei social network un "potere informativo" e conseguentemente socio-economico rilevante. Identicamente questo si verifica per i gestori delle reti cellulari o dei motori di ricerca.

Anche se un utente può "limitare" l'accesso ai suoi dati è evidente che ad oggi la capacità di analizzare ed elaborare informazioni personali di così grande portata (Big Data) è "predominio" dei grandi gestori dei social network, della telefonia cellulare e degli stati. Questo fatto apre molte domande sul potere informativo in capo a pochi soggetti e sulla legislazione in materia di trattamento e privacy dei dati personali. Quest'ultima è basata sul consenso informato[11] che nella "sostanza" ha dato origine a pagine web (o cartacee) che non vengono quasi mai lette e/o capite ed accettate a priori per poter fruire di un servizio.

Il tema della privacy è certamente in evoluzione e se da un lato le normative esistenti e le autority che si occupano della materia hanno introdotto regole sempre più stringenti e puntuali sul trattamento dei dati personali, dall'altro, nello scenario dei Big Data, sono gli utenti che decidono, nei fatti, di rendere pubblici i propri dati personali e, in cambio di servizi, permettono ai gestori dei social network di poterli analizzare ed utilizzare per diversi scopi. Uno dei nodi principali sta proprio nel definire gli scopi di utilizzo che non possono essere dettagliati ed enumerati esaustivamente nell'atto di raccolta delle informazioni (consenso informato). Un elemento molto utilizzato nell'analisi di dati personali (con attenzione alla tutela della privacy

11 Il consenso informato si basa sull'essere informati preventivamente su quali informazioni vengono raccolte e per quale scopo.

– riconoscibilità dei soggetti) è rappresentato dalla loro anonimizzazione. Ma l'utilizzo dei Big Data, in molti casi, ha dimostrato che si può arrivare all'identificazione dei singoli e che pertanto l'anonimizzazione non sempre è sufficiente.

Occorrerà studiare nuove "regole" per la protezione della privacy ponendo l'attenzione nell'identificare nuovi fattori per superare gli attuali limiti del consenso informato e dell'anonimizzazione. Siamo di fronte a scenari già vissuti anche nella storia recente (l'avvento di Internet e del web) tipici di una situazione innovativa ed in evoluzione, i cui i confini non sono delimitati e delimitabili facilmente e dove, come la realtà insegna, i tecnicismi e i "punti di equilibrio" da essi derivanti sono quasi certamente falliti nel momento in cui vengono definiti. Infatti le potenzialità di sfruttamento informativo e conoscitivo offerte dai Big Data superano di gran lunga lo scopo primario per cui i dati sono raccolti, aprendo ampie possibilità di analisi su fatti secondari e non preventivamente definibili. Questo fatto ripropone l'attenzione sul concetto di scopo che, evolvendo nel tempo in relazione alla quantità sempre più ampia di dati raccolti e alla loro combinazione (correlazione) possibile, non può che essere associato ad un concetto di responsabilità. Se da una parte occorre assicurare garanzie reali affinché i dati necessari per il bene pubblico siano facilmente reperibili, dall'altra è indispensabile proteggere la tutela della privacy e della libertà (Pentland 2015).

Un rischio latente di utilizzo improprio dei Big Data è riscontrabile nella capacità di elaborare grandi quantità di informazioni personali per effettuare previsioni che potrebbero "condizionare" la vita delle persone. Sono sempre più numerose le città degli Stati Uniti che impiegano la sorveglianza preventiva: si usa l'analisi dei Big Data per selezionare le strade, i gruppi e gli individui da tenere particolarmente sotto controllo, per il solo fatto che un algoritmo ha identificato in essi una più alta propensione alla criminalità. Se da un lato, lo scopo di prevenzione di crimini o rischi per le persone che possiamo attuare con le analisi dei Big Data può contribuire alla sicurezza, da un altro punto di osservazione potrebbe diventare (come ben è stato rappresentato nel film Minority Report del 2002) oltremodo pericoloso l'utilizzo di previsioni estratte dai Big Data per stabilire se qualcuno è colpevole e andrebbe punito per un comportamento che non è ancora stato messo in atto (Shomberger e Cukier 2013[12]). Responsabilità per la costruzione di una società migliore, libertà e tutela della privacy sono fattori che nell'era dei Big Data sono fortemente riproposti e chiedono di essere affrontati andando oltre i pur importanti "tecnicismi" e rimettendo al centro la concezione che abbiamo di persona nella sua dimensione personale e sociale.

Conclusioni

L'utilizzo dei dati e delle informazioni da essi elaborate sono una realtà sempre più

12 In Big Data di V. Shomberger e K. Cukier, Garzanti 2013: citazione di J Vlahos, The department of Pre-Crime, "Scientific american" 306 gennaio 2012.

presente per la conoscenza dei fenomeni e per il supporto dei processi decisionali, nei contesti socio-economici, aziendali e personali. L'evoluzione tecnologica degli ultimi decenni, in particolare con la nuova generazione di internet, del web e delle tecniche di elaborazione ed analisi dei dati, ha dato delle spinte fino a poco tempo fa inimmaginabili nella disponibilità di informazioni ed è facile cogliere che questo fenomeno crescerà rapidamente nel prossimo futuro. Pur se le tecniche di raccolta e di analisi tradizionali (censuarie e campionarie) sono cresciute e in molti contesti daranno ancora un importante contributo, è evidente che i Big Data si stanno affermando come "strumento" innovativo in tutti gli ambiti (professionali e personali) sia per la comprensione di avvenimenti sia per l'identificazione e realizzazione di nuovi servizi.

Il nuovo paradigma con cui possiamo rappresentare quella che viene chiamata l'era dei Big Data, sintetizzabile nel "lasciar parlare i dati", esprime un importante fattore di novità: l'osservazione di ciò che accade, per conoscere lealmente i fatti e favorire un approccio alla conoscenza che allarghi lo spettro delle domande relative al perché accadono. In una celebre frase, Alexis Carrel13 affermava: "Molta osservazione e poco ragionamento conducono alla verità; poca osservazione e molto ragionamento conducono all'errore".

Come abbiamo osservato i Big Data entrano in gioco in moltissimi campi della vita e possono contribuire al progresso della conoscenza e conseguentemente al miglioramento e allo sviluppo della vita sociale, degli scopi aziendali e delle scelte personali. Ma spesso una così ampia disponibilità di informazioni potrebbe creare "confusione" e in taluni casi scelte e azioni in forte contrasto con la libertà e i diritti fondamentali degli individui. È in questo senso che la questione aperta è innanzitutto culturale e quindi educativa e formativa. Le metodologie e le tecniche di trattamento dei Big Data si stanno evolvendo molto rapidamente e sono richieste nel mercato del lavoro nuove professionalità. Diverse sono le iniziative in atto, specialmente a livello dei corsi universitari che in tutto il mondo sono state avviate o si stanno intraprendendo per formare i cosiddetti "data scientist", gli specialisti nel trattamento e analisi dei Big Data. Se un approccio legato alla conoscenza delle metodologie e tecniche matematiche, statistiche e informatiche, che sono alla base del trattamento ed analisi dei Big Data è importante, è nel contempo evidente che occorrerà porre molta attenzione alla creazione di professionalità educate ad una posizione di apertura alla realtà, capaci cioè di tener conto dei diversi fattori in gioco e quindi di operare con responsabilità ed attenzione allo scopo di utilizzo delle analisi derivanti dai Big Data.

I Big Data sono in questo senso una occasione per riproporre quale sia il nesso tra tecnologia e persona, tra tecnologia e sviluppo sociale ed economico, per evitare una concezione in cui la persona debba passivamente adeguarsi all'innovazione tecnologica; come riportato in un famoso slogan dell'EXPO di Chicago del 1933:

13 Alexis Carrel, *Riflessioni sulla condotta della vita*, Editore Cantagalli, Verona 2003.

"La scienza trova, l'industria applica, l'uomo si adegua".

Come la storia ci insegna le invenzioni e le innovazioni tecnologiche sono uno strumento ma la scintilla dell'invenzione è nella persona che per rispondere ai propri (e gli altrui) bisogni sviluppa la sua curiosità, creatività e il suo ingegno che sono sostanzialmente la fonte del progresso. In questo senso occorre affermare che i Big Data sono uno strumento che ci aiuterà a migliorare la conoscenza ma nel contempo dobbiamo ricordare che sono imperfetti (perché gli strumenti che usiamo sono imperfetti) e più in generale ogni tentativo umano è di per sé imperfetto; questo fatto non significa che stiamo sbagliando ma che occorre essere profondamente coscienti dell'imperfezione o incompletezza delle analisi che effettuiamo pur avendo a disposizione una enorme quantità di dati.

"Non bisogna negare le intuizioni che offrono i Big Data, ma occorre inquadrarli per quello che sono: uno strumento che non dà risposte definitive, solo risposte in grado di aiutarci nell'immediato, finché non emergeranno metodi migliori e quindi anche risposte più soddisfacenti. Ciò significa anche che dobbiamo usare questo strumento con una buona dose di umiltà e di umanità" (Schomberger e Cukier, 2013).

Bibliografia

Marco Martini, *La Statistica*, in *Studi in Ricordo di Marco Martini*, a cura di Matteo Pelagatti, Giuffrè Editore, Milano 2004;

Alex 'Sandy' Pentland, *Fisica Sociale*, EGEA, Milano 2015;

Michael K. Bergman, *White paper: the deep web: surfacing hidden value*, in *Journal of electronic publishing 7.1*, 2001;

Alexis Carrel, *Riflessioni sulla condotta della vita*, Editore Cantagalli, Verona 2003;

V. M. Schönberger e K. Cukier, *Big Data - Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, Garzanti, Milano 2013.