

L'IRRESISTIBILE AVANZATA DEI *BIG DATA* Intervista a Mauro Campanella

a cura di Roberto Sanvito*

La trasformazione in digitale delle conoscenze ne permette una manipolazione e una elaborazione a livelli di dimensione, complessità e bassi costi che non era possibile solo cinque anni fa.

Si parla molto di Big Data: una rivoluzione nella gestione dell'informazione, con effetti dirompenti in ogni campo, dalla medicina, alla scienza, alla vita privata e pubblica; anche il mondo della scuola ne risentirà.

L'analisi di grandi quantità di dati può semplificare molte attività; ma è urgente sviluppare criteri di gestione responsabile per governare uno sviluppo che procede a una velocità superiore a quella dell'evoluzione culturale umana.

* *Membro della redazione della rivista Emmeci-quadro*

Lo sviluppo impetuoso delle tecnologie informatiche porta alla ribalta nuovi strumenti e nuovi fenomeni che stanno invadendo, in modo sempre più pervasivo, tutti gli aspetti della nostra vita: non solo le attività lavorative ne vengono trasformate ma il panorama culturale, la vita sociale e le relazioni interpersonali sono soggette a mutamenti spesso non ben consapevoli e responsabilmente governati.

Chi opera nell'educazione non solo deve prestare attenzione a tutte queste nuove realtà, ma è sollecitato ad approfondirle criticamente proprio per il ruolo cruciale che può assumere la funzione educativa nel preparare persone in grado di interagire con i sistemi tecnologici in modo efficace e responsabile.

Uno di questi nuovi fenomeni è quello dei cosiddetti *Big Data*: per conoscerlo meglio e coglierne le implicazioni culturali ed educative abbiamo incontrato Mauro Campanella, fisico, responsabile Ricerca & Sviluppo del GARR (la rete nazionale a banda ultralarga dedicata alla comunità dell'istruzione e della ricerca).

Iniziamo da un chiarimento sui termini: cosa si intende per Big Data?

La possibilità di trasformare in numero ogni aspetto della realtà, quali immagini, suoni, esami medici, denaro, ha reso possibile rappresentarla in informazione digitale, o più brevemente, all'inglese, *data*. Lo sviluppo e la diffusione capillare della tecnologia digitale, come i cellulari, e della trasmissione dei dati hanno quasi annullato il costo dell'archiviazione e ha permesso il trasferimento (via *Internet*) in tutto il mondo in modo istantaneo e a bassi costi.

I *Big Data* sono l'insieme di tutti questi numeri che vengono da tantissime sorgenti. Fonti inesauribili legate alle persone sono le applicazioni sociali come *Whatsapp* e *Facebook*, i nostri telefonini (i numeri chiamati, i nostri spostamenti, i dati prodotti dalle *app* che usiamo, le ricerche su *Internet*), ma anche le nostre tessere fedeltà, le carte di credito.

I *Big Data* sono anche dati impersonali, quali il traffico stradale, il numero dei passeggeri in una metropolitana, il numero di mattoncini della LEGO prodotti, i ricambi e la



loro quantità nel magazzino, i numeri di serie dei biglietti del cinema e così via, in ogni situazione in cui sia prodotta informazione digitale.

Questa enorme quantità di informazione, in costante crescita, non ancora aggregata e classificata è chiamata *Big Data* ed è considerata un tesoro grezzo da cui ricavare risultati scientifici, conoscenza e ricchezza.

Quando se ne comincia a parlare? E perché solo oggi sono diventati così importanti?

La conoscenza è sempre stata una ricchezza e una fonte di potere. Dalla metà del ventesimo secolo la tecnologia aveva reso semplice ed economico produrre e memorizzare informazione. Verso il 1995 la memorizzazione digitale diventa più economica della carta e molto più compatta. Dall'anno 2000 quasi tutta l'informazione è prodotta digitalmente e il termine *Big Data* inizia a essere usato diffusamente.

Oggi la trasformazione in digitale della conoscenza ne permette una manipolazione e una elaborazione a livelli di dimensione, complessità e bassi costi che non erano possibili solo cinque anni fa.

Questa rivoluzione nella gestione dell'informazione ha un effetto dirompente in ogni campo, dalla medicina, alla scienza, alla vita privata e pubblica.

Cosa vuol dire quindi usare i Big Data?

L'enorme quantità di dati grezzi prodotta ha una sua utilità intrinseca e per questo è archiviata in memorie digitali: è possibile così, per esempio, ricordare ogni numero di matricola di un oggetto e il suo acquirente per la garanzia o il pagamento della spesa alla cassa. Ma è quando questi dati vengono indicizzati, analizzati, messi in relazione fra di loro che si scoprono la vera utilità e l'interesse dei *Big Data*.

Per esempio, analizzando gli scontrini autostradali si può statisticamente prevedere il traffico nei giorni futuri e il flusso (entrata e uscita) dei veicoli. Tale previsione permette di decidere quanti caselli aprire a ogni uscita per diminuire le code dei veicoli che transitano; gli stessi risultati sono utili ai distributori di benzina per pianificare le scorte, alla polizia per organizzare il controllo.

Maggiore il numero di dati raccolti, maggiore la precisione delle stime. Analizzando le vendite in un supermercato si pianificano i rifornimenti, si comprendono i gusti dei clienti, si crea la strategia per la pubblicità e il cambio dei produttori.

L'analisi può incrociare dati in quantità sempre maggiore e con algoritmi molto sofisticati e che sono in grado di trovare relazioni che non sarebbero ottenibili manualmente.

L'analisi dei *Big Data* produce altri dati, per esempio l'indice dei dati stessi, che sono normalmente detti metadati, che contengono un valore aggiunto ancora maggiore.

I motori di ricerca quali Google sono i principali utilizzatori di *Big Data*. Indicizzano e assegnano un'importanza alle informazioni che hanno, partendo dalle singole parole. Questo compito è titanico ed è svolto automaticamente e costantemente da calcolatori sparsi per il mondo. La rete *Internet* permette a tutti i centri di Google di aggiornarsi vicendevolmente.

La presentazione del risultato di una ricerca dipende quindi da algoritmi e pesi che non necessariamente sono neutri, ma possono fornire importanza maggiore ad alcuni risultati in base ad accordi economici o politici. In questo senso tutti i risultati vanno sempre interpretati e confrontati.

Di fatto, l'analisi dei *Big Data* e lo sviluppo tecnologico stanno crescendo in modo sinergico: tanto più i calcolatori diventano veloci e miniaturizzati, le memorie più capienti e la rete più capillare, tanto più è semplice raccogliere e analizzare grandi quantità di informazione.

Di quali dimensione di dati si sta parlando?

I dati si creano così facilmente che la loro quantità è ben più di «grande» (*big*) e aumenta costantemente. Si pensi a un solo supermercato e a tutte le informazioni dei magazzini, delle vendite e degli scontrini legati ognuno a una carta fedeltà: milioni di numeri ogni giorno. Una quantità ben superiore ai miliardi in poco tempo e solo per quel negozio.

Oggi un disco di un computer, come quello che usiamo normalmente, contiene 500 miliardi di caratteri (*byte*) e un numero decimale ne occupa in media 4. Facilmente i centri dove i dati sono archiviati hanno migliaia di dischi, ciascuno dei quali contiene anche 4000 o 8000 miliardi di caratteri.

Dove «si trovano» questi dati e da chi sono gestiti?

I dati sono ovunque, generati da ogni oggetto che contenga elettronica e che possa comunicare. Ognuno di noi è proprietario di oggetti e gestore di una parte di dati, inclusi film e fotografie.

La tendenza è di raggrupparli, per semplicità di gestione e ottimizzazione dei costi, in grandi centri, dove oltre ai dati ci sono calcolatori e veloci reti di trasmissione. Questi centri sono gli elementi base di sistemi oggi noti come *sistemi di cloud* (nuvola). La loro gestione è spesso di enti privati, come Google, Facebook, Amazon, Dropbox.

Il mondo della ricerca ha creato propri centri di calcolo dedicati alla raccolta e analisi dei dati scientifici. Grandi quantità sono generate per esempio dagli esperimenti di fisica delle particelle al CERN di Ginevra (<http://home.cern/about/computing>), dall'osservazione astronomica, dalla biomedicina.

Quindi c'è una stretta relazione con Internet...

Sì, strettissima. *Internet* è lo strumento alla base della creazione del modello di servizi chiamato appunto *nuvola*; permette di essere sempre «in rete» e di scindere la localizzazione fisica dell'informazione da quella dell'utilizzatore. Oggi si possono consultare via rete i libri di una biblioteca, vedere film e leggere esami medici comodamente seduti da casa per esempio.

La *nuvola* adotta un modello di gestione dei dati fisicamente centralizzato, ma con i suoi utilizzatori fisicamente distanti e collegati ad alta velocità via *Internet*. La *nuvola* sposta anche le capacità di calcolo vicino ai dati, permettendo anche l'analisi e l'applicazione e non solo la consultazione.

Di che tipo di infrastruttura necessitano i Big Data?

Il «dato» è l'elemento base del mondo digitale e viene immagazzinato in dischi magnetici o a stato solido (questi ultimi circa dieci volte più veloci di quelli magnetici). Queste apparecchiature richiedono potenza elettrica, ambienti condizionati in temperatura, puliti e protetti dalle intrusioni, anche fisiche.

Molti *data center* sono alloggiati in edifici ampi, appositamente attrezzati e collegati all'esterno da circuiti multipli a molti miliardi di *bit* al secondo. È comunque possibile e utile fare un piccolo centro dati anche in un liceo, dove ospitare in una stanza chiusa e condizionata sia i calcolatori sia i dischi e collegarli alla rete.

Quali sono attualmente le più interessanti aree applicative?

L'analisi permette di portare innovazione in ogni campo. In prospettiva l'impatto maggiore si prepara a livello sociale: in campo medico attraverso l'incrocio dei risultati di molteplici pazienti, nella ottimizzazione dell'uso dell'energia elettrica, dei trasporti, nella sicurezza.

Il mondo della ricerca sta già proficuamente utilizzando l'analisi dei *Big Data* in varie discipline. Per l'educazione l'impatto sarà il risultato di una evoluzione dei metodi didattici che sappiano sfruttare adeguatamente *Internet* e *Big Data*.

Quali le interazioni con altri segmenti tecnologici?

Anche in questo caso è molto stretta e foriera di cambiamenti. La disponibilità di spazio di archiviazione e calcolo ad alte prestazioni ha già portato a tecnologie quali il riconoscimento vocale e dei comandi come SIRI di Apple e le auto che si guidano da sole.

Simili sistemi, oltre a potersi basare sui *Big Data* e su regole grammaticali, usano logica avanzata per entrare nel significato della frase. L'intelligenza artificiale si applica nella forma di apprendimento attraverso l'elaborazione continua ai *Big Data* e al risultato delle azioni intraprese.

Quali cambiamenti potenzialmente possono essere indotti, a livello sociale, Industriale, culturale?

Siamo solo all'inizio dei cambiamenti che si prefigurano profondi in ogni campo. I frutti dell'uso della tecnologia dei *Big Data* possono essere positivi, ma anche negativi.

Dal lato positivo, è probabile una maggiore automazione e ottimizzazione di tutti gli aspetti pratici della vita come un uso migliore dell'energia, cure mediche più efficaci ed efficienti, una agricoltura più rispettosa dell'ambiente.

A livello sociale, la quantità di dati disponibile alla pubblica amministrazione potrebbe semplificare molti adempimenti e aumentare l'efficienza del sistema.

A livello industriale una maggiore efficienza e automazione dei processi produttivi e velocità nel cambiamento, a fronte di una maggiore invadenza nella vita privata per fornire pubblicità personalizzata.

La pubblicità personalizzata è il risultato di una analisi di dati raccolti dalle abitudini di acquisto della persona. Tali dati vengono ricavati per esempio attraverso l'uso di carte fedeltà e carte di credito, ma anche dagli interessi a prodotti che si sono cercati in *Internet* o da interessi ricavati dai *social network*. Tali informazioni sono scambiate a pagamento fra le ditte interessate.

Sono in fase di prova vetture capaci di guidare da sole, un'applicazione che unisce l'uso dei *Big Data* per le mappe stradali con l'intelligenza artificiale.

A livello culturale, la digitalizzazione di tutte le opere e le informazioni storiche del passato sta lentamente continuando. Tale opera è lenta perché richiede che, manualmente, le informazioni presenti su carta o altri supporti siano digitalizzate trasformandole in immagini e quindi ci sia un eventuale processo di riconoscimento del testo che permetta di completare la digitalizzazione per la ricerca. La facilità di accesso ai dati, le analisi e l'incrocio fra le informazioni sono la sfida culturale dei prossimi anni.

Infine, a livello personale il cambiamento sarà conseguenza dell'uso dei *Big Data* da parte di organizzazioni pubbliche e private, che influenzeranno le nostre vite. Vivremo in un ambiente più ricco di comunicazione interpersonale, ma certamente più invadente.

Ci sono elementi di preoccupazione e si può parlare di potenziali rischi?

L'esistenza dei *Big Data* e il loro uso sono una novità e la comprensione e quindi il controllo della tecnologia sono ancora nella fase iniziale.

La tecnologia ha una potenzialità molto alta di generare informazione per deduzione su ogni aspetto della realtà, sul suo futuro e quindi su ogni persona. Quindi vi sono rischi per la sfera personale ed è necessario che il mantenimento dei dati e il loro uso vadano compresi nelle conseguenze e probabilmente regolati; al momento non sono ancora stati sviluppati un'etica, una cultura e un insieme di regole che permetta di usarla al meglio.

Inoltre spesso i dati e i metadati non sono archiviati nel paese in cui sono prodotti e quindi soggetti alla legislazione locale. Google ha dovuto prendere accordi con vari stati su quali informazioni rendere disponibili ai cittadini di quello stato, rendendo l'accesso all'informazione meno trasparente.

Oltre alla comprensione dell'uso dei dati sul presente e sul passato, va anche compreso come valutare correttamente le previsioni basate su *Big Data* e algoritmi, per non pensare che siano necessariamente certe. Per esempio, le previsioni meteorologiche stanno diventando più precise, ma sono sempre probabili e mai certe.

Cosa si può dire in tema di privacy e di controllo personale?

La *privacy* è un aspetto già parzialmente protetto dalla legge, ma è necessario fare di più e continuare ad aggiornare le leggi e la cultura, soprattutto la nostra comprensione. Il vivere nel modo digitale, crea informazioni che sono raccolte costantemente da molti enti.

Il cellulare che portiamo sempre con noi è tracciato costantemente e fornisce informazioni sulla nostra posizione, le ricerche su *Internet* che eseguiamo forniscono informazioni sui nostri interessi e abitudini, partecipare alle reti sociali di *Facebook* o *Twitter* fornisce informazioni sui nostri collegamenti e naturalmente rende disponibili le nostre opinioni e tutto ciò che rendiamo permanente affidandolo a *Internet*. Conseguentemente vanno tenuti dei comportamenti prudenti sui dati personali anche per evitare i furti di identità, per esempio non buttare nella raccolta della carta i dati bancari.

Noi per primi dobbiamo fornire solo le informazioni che siamo disposti a condividere. Anche perché per ora *Internet* mantiene tutte le informazioni, per sempre. Il diritto e la possibilità dell'oblio devono ancora essere introdotti coerentemente.

I dati del DNA di ciascuno sono un esempio estremo dell'informazione che può essere analizzata e fornire previsioni e informazioni molto personali e delicate. Per esempio la probabilità di sviluppare una patologia.

Data la mole di informazioni che sono disponibili su ciascuno di noi anche indirettamente, per esempio attraverso le riprese di telecamere in luoghi pubblici, è, in effetti, possibile ricavare con molta precisione il comportamento di una persona, avendo accesso ai dati e sufficienti risorse di calcolo. Le esigenze della *privacy* devono essere bilanciate con i controlli che, ad esempio, il terrorismo ha reso necessari.

Si può dire che stiamo andando verso un'organizzazione sociale sempre più dipendente dalla tecnica? Che l'essere umano da «custode del creato» sta diventando «custode delle macchine»?

Probabilmente sì, la nostra è già una società «fragile», per esempio per la sua estrema dipendenza dall'energia oltre che dalla tecnologia. I dati possono semplificare il lavoro di tutti, ma ci vuole anche la volontà di semplificare il più possibile la quantità dei dati che produciamo e che sono richiesti.

Personalmente, come uomo che vuole comprendere il mondo, rispettandolo, ritengo che le nuove tecnologie siano uno strumento splendido per elaborare la nostra conoscenza. In questo senso anche la conoscenza dello strumento deve aumentare, così come la capacità di usarlo in sicurezza. Un compito non facile, perché lo sviluppo tecnologico è molto più rapido dei tempi umani di gestione dell'innovazione attraverso la cultura.

I *Big Data* e la loro analisi sono un'occasione unica di miglioramento sia della vita quotidiana sia della conoscenza. Una morale e un'etica relative all'uso di importanti informazioni, quali per esempio quelle del DNA, devono svilupparsi in fretta per minimizzare gli effetti negativi: queste tecnologie hanno la potenzialità di cambiare profondamente l'essere umano stesso.

*A cura di Roberto Sanvito
(Membro della redazione della rivista Emmeciquadro)*