# Big Data e turismo: una relazione complicata

Rodolfo Baggio

Master in Economia del Turismo

Università Bocconi, Milano

=====

In E. Becheri & A. Morvillo (Eds.), XXIII Rapporto sul Turismo Italiano – 2018/2019 (pp. 91-106). Napoli: CNR-IRISS - Rogiosi editore (2019)

=====

## 1 Dati, dati, dati

L'enorme diffusione di Internet e delle applicazioni sviluppate in questo ambiente, nonché dei dispositivi che si appoggiano alla Rete per il loro funzionamento, hanno, come noto, radicalmente cambiato molti aspetti della nostra vita. Forse l'aspetto più eclatante di questi cambiamenti sta nell'enorme quantità di 'oggetti' prodotti e diffusi in Rete e delle tracce digitali che i milioni di utenti lasciano ogni secondo come 'sottoprodotto' delle loro attività.

Le dimensioni di questo fenomeno sono tali che, almeno fino a qualche tempo fa, ci si poneva il problema di come 'sopravvivere' a questo vero e proprio diluvio e come contrastare quella che è stato definito sovraccarico cognitivo dovuto all'informazione presente online.

Di *information overload* si parla, e ci si preoccupa, da secoli, in pratica da quando lo sviluppo di tecnologie di trattamento delle informazioni ha cominciato a permettere una certa diffusione dei diversi prodotti. Già Seneca (I sec. d.C.) nella sua lettera a Lucilio ammoniva che la moltitudine di libri serve solo a distrarre, e Vincenzo di Beauvais, dotto frate domenicano, nel XII secolo produce una monumentale (80 volumi) proto-enciclopedia per contrastare il fatto che la moltitudine di libri in circolazione non consentiva di apprezzarne il contenuto per via del poco tempo disponibile e della limitatezza della nostra memoria (Eckenrode, 1984). Le allarmate dichiarazioni si intensificano poi dopo l'invenzione della stampa. Conrad Gessner (1516-1565) scienziato svizzero, probabilmente il primo a discutere di sovraccarico informativo, osservava che il problema di una 'ingestibile' quantità di informazioni era nato con l'invenzione della stampa, lamentandosi della stupidità di scritti inutili e dell'abbondanza dannosa e confusa di libri (Blair, 2010). E Barnaby Rich scriveva nel 1613 (Braun & Zsindely, 1985: 529): "una delle malattie di questi tempi è la molteplicità dei libri; sovraccaricano il mondo che diventa incapace di digerire l'abbondanza di materia inattiva che viene prodotta ogni giorno".

Un bell'articolo pubblicato dalla prestigiosa Harvard Business Review nel 2009 (Hemp, 2009) riassumeva la situazione parlando di *morte da information overload* e dando suggerimenti su come sopravvivere. Ma quasi esattamente tre anni dopo, nel 2012, McAfee e Brynjolfsson magnificavano le bellezze e l'utilità dell'avere a disposizione così vaste quantità di dati e parlavano di 'rivoluzione dei Big Data' (McAfee & Brynjolfsson, 2012).

A parte l'indubbia componente promozionale da parte dell'industria informatica, il concetto portava alla luce il fatto che, al di là dei problemi creati, la disponibilità di grandi quantità di informazioni di tutti i tipi aveva suggerito a industria e ricerca un nuovo approccio alla raccolta, selezione e analisi di quanto si può trovare abbastanza liberamente in Rete, e per di più in forma digitale, quindi già quasi pronta per l'elaborazione, e portato alla messa a punto di metodologie più raffinate ed efficaci.

Lo sviluppo tecnologico e l'adozione esponenzialmente crescente di dispositivi che consentono l'automazione di varie attività e la connessione a Internet, infatti, ha indotto la produzione di notevoli quantità di dati, spesso strettamente legati a contenuti generati dagli utenti attraverso social network e altre applicazioni online. Per le loro dimensioni e le loro caratteristiche, i Big Data sono difficili da elaborare con metodi statistici e tecniche software tradizionali. Tuttavia, essi stanno diventando molto diffusi e apprezzati come campo di indagine anche nelle scienze sociali, dove sono stati indicati come fattore prezioso per migliorare la crescita economica e sociale, risolvere problemi (Mayer-Schönberger & Cukier, 2013) e come motore principale per la creazione di valore per aziende e clienti (Mariani et al., 2018; Verhoef et al., 2016).

## 1.1 Big data

Il termine Big Data (BD), nell'accezione corrente, identifica l'enorme volume di dati, strutturati e non, generati dagli utenti di quel vasto e complesso mondo digitale che si è imposto negli ultimi anni raggiungendo rapidamente diffusioni imponenti. Le ultime stime di internetworldstats.com parlano di oltre quattro miliardi e mezzo di utenti Internet nel mondo per il 2019.

Nella storia recente, il termine Big Data compare verso la fine degli anni 90 negli ambienti informatici e accademici (Lohr, 2013), ma diventa noto al grande pubblico quando viene utilizzato da Roger Mougalas di O'Reilly Media nel 2005, un anno dopo l'introduzione dell'idea di Web 2.0, per riferirsi a quei grandi insiemi di dati disponibili (Ross, 2010).

Benché, come spesso accade, non ci siano definizioni formali condivise, gli attributi principali dei Big Data, molto citati sia dai media che dalla letteratura scientifica, sono ormai noti e mettono in luce dimensioni, variabilità nel tempo, varietà di formati e di significati, affidabilità (o inaffidabilità) delle fonti e così via (Baggio & Klobas, 2017: Chapter 8; Laney, 2001).

Molte altre caratteristiche potrebbero essere aggiunte, come quelle riguardanti la completezza di dati utilizzati, che riguardano spesso popolazioni intere anziché solo campioni; la risoluzione, cioè la quantità di dettagli su un qualche elemento; la natura relazionale delle variabili che potrebbe essere comune a fonti diverse; la flessibilità necessaria durante l'analisi della raccolta dei dati per consentire il ridimensionamento o l'estensione con ulteriori variabili e casi (Mayer-Schönberger & Cukier, 2013).

Con questo scenario è difficile raggiungere una semplice (e persino univoca) caratterizzazione dei Big Data. Qui, adottiamo il punto di vista di Boyd e Crawford che definiscono i Big Data come (2012: 663):

> "... un fenomeno culturale, tecnologico e accademico su cui poggia l'interazione di: 1. Tecnologia: per massimizzare la potenza di calcolo e precisione algoritmica per raccogliere, analizzare, collegare e confrontare grandi set di dati; 2. Analisi: attingendo a grandi set di dati per identificare modelli che permettano di fare considerazioni

economiche, sociali, tecniche e legali; 3. Mitologia: la diffusa convinzione che grandi insiemi di dati offrano una forma superiore di intelligenza e conoscenza che può generare intuizioni che prima erano impossibili, con un'aura di verità, obiettività e accuratezza."

I dati 'disponibili' online sono quelli che gli utenti pubblicano sulle varie e note piattaforme (Facebook, Twitter, Instagram, Flickr, Tripadvisor, ecc.) ai quali vanno aggiunti i record registrati dalle compagnie telefoniche che sono, forse, le registrazioni più affidabili per conoscere i movimenti di chiunque possieda un telefono cellulare e quelli, non sempre disponibili, che riguardano transazioni di vario genere, monetarie e non.

Un'altra categoria, poi, riguarda i dati generati dai numerosi dispositivi che oggi possono essere collegati attraverso una rete. Sensori, attuatori, altri dispositivi dotati di capacità di comunicazione e connessi tramite Internet, spesso tramite servizi cloud; quell'insieme noto oggi come l'Internet delle cose (Internet of Things, IoT). Tutti questi generano dati o necessitano di dati per il loro funzionamento, spesso automatizzato, e possono rivelarsi estremamente utili ed efficaci in molte situazioni (Lee & Lee, 2015).

### 1.1.1 Big Data per il turismo e l'ospitalità

Nel settore turistico e dell'ospitalità i Big Data sono stati utilizzati, per esempio, per migliorare la conoscenza di una destinazione (Fuchs et al., 2014), o per comprendere desideri e soddisfazione di clienti (Xiang et al., 2015), ma anche per ottenere previsioni più accurate (Pan & Yang, 2016).

Per quanto riguarda i sistemi IoT, essi sono sempre più utilizzati per diversi scopi. In un hotel, ad esempio, è possibile personalizzare tutte le impostazioni ambientali (temperatura, luci, flussi d'acqua ecc.), controllare gli accessi o verificare lo stato operativo di diversi dispositivi e apparecchiature per riparazioni e manutenzione predittive (Car et al., 2019). In una destinazione i dispositivi IoT possono tracciare i visitatori e aiutare a gestire i flussi o fornire informazioni basate sulla posizione e fornire indicazioni ai turisti su trasporti, attrazioni, tour, shopping, hotel (Wise & Heidari, 2019). Oltre a una pura registrazione, algoritmi di machine learning (vedi par. 2.1) consentono di utilizzare apparati IoT per ottimizzare o gestire automaticamente molti processi, aspetto importante per garantire lo sviluppo di strutture e destinazioni intelligenti (Inanc–Demir & Kozak, 2019; Mahdavinejad et al., 2018).

Per quanto riguarda i dati 'ufficiali' esistono diverse iniziative. Un gruppo di lavoro dell'ufficio statistico della UN Statistics Division (https://unstats.un.org/bigdata/) ha il compito di esplorare l'utilizzo di Big Data per le statistiche ufficiali. Lo stesso fa Eurostat nell'ambito della Collaboration in Research and Methodology for Official Statistics (https://ec.europa.eu/eurostat/cros/content/big-data_en) che contiene documenti e rapporti sulle iniziative nel campo. Inoltre, praticamente tutti gli uffici statistici nazionali hanno mostrato interesse e hanno avviato progetti pilota. Non esistono, però, al momento, realtà consolidate di utilizzo a livello 'ufficiale'. In Italia è stata recentemente attivata la piattaforma "Data intelligence" delle Camere di Commercio. Turismo Big Data (https://turismobigdata.isnart.it), un sistema che offre una panoramica del settore turistico arricchito da informazioni che provengono da molteplici fonti Web.

## 1.2 La raccolta dei dati

Molte applicazioni forniscono e in qualche modo controllano l'accesso, la grande ricchezza dei dati raccolti dai social network online. Raccolta di tweet, Facebook post, recensioni o articoli simili può

però essere un compito piuttosto complicato. Una possibilità è quella di ricorrere a un fornitore di dati (vedi per es. www. bigdatavendors.com o socialmediadata.wikidot.com).

Questa, tuttavia, può essere una soluzione piuttosto costosa, almeno per una piccola azienda o destinazione. La seconda possibilità è quella di utilizzare alcune applicazioni sviluppate per altri scopi, ma che forniscono anche un plug-in per il download di dati per l'analisi. Esempi sono NodeXL (nodexl.com), un componente aggiuntivo di analisi di reti per Excel, Gephi (gephi.org), una piattaforma open source per la visualizzazione e l'analisi di reti complesse, o Rapidminer (rapidminer.com), un software per l'analisi dei dati con sofisticati algoritmi di machine learning.

Anche queste soluzioni non sono completamente soddisfacenti a causa delle loro intrinseche limitazioni nella quantità di dati che possono gestire, o del tipo di informazioni che possono raccogliere; i plug-in, infatti, sono progettati per integrare le funzioni principali dei prodotti e non necessariamente completamente in linea con ciò che potrebbe essere necessario per la raccolta e l'analisi dei Big Data. Quando è richiesto un approccio personalizzato, le competenze e le risorse necessari (in termini di hardware e software) sono relativamente moderati. È possibile equipaggiare un normale PC (o un gruppo di macchine a basso costo) con librerie open source che, con un intervento di programmazione limitato, consentono raccolta dei dati da personalizzare (un esempio è il sistema presentato in (d'Amore et al., 2015).

Va notato qui che, a seguito del noto scandalo di Cambridge Analytica (Cadwalladr & Graham-Harrison, 2018), molte piattaforme hanno limitato fortemente, se non chiuso, le possibilità di accedere liberamente ai loro dati. Restrizioni che però non toccano sensibilmente i dataset venduti dalle aziende attive in questo settore.

Le registrazioni delle compagnie telefoniche sono da sempre disponibili solo a pagamento e, ovviamente, in forma anonimizzata. I 'call detail record' (CDR; in italiano noti anche come cartellini di traffico) sono le registrazioni degli scambi di informazioni tra i terminali (telefoni cellulari, smartphone ecc.) connessi alla rete e contengono tutti i dettagli delle varie conversazioni: numeri di telefono, nominativi, durate, tipo di connessione, scambi e celle alle quali il terminale è connesso durante la chiamata. La loro funzione principale è quella di fare da base alla fatturazione delle compagnie telefoniche.

La validità del ricorso ai Big Data è anche data dal fatto, da tempo noto ma che a volte sembra essere 'dimenticato', che gli strumenti tradizionali di analisi che poggiano su questionari e interviste, sono spesso affetti da due problemi. Il primo è che il ricercatore tende a formulare domande su argomenti che ritiene rilevanti, ma che magari lo sono per un numero ristretto di individui o lo sono in modo molto parziale, col risultato di fare affermazioni, magari ben supportate dall'indagine, ma che sono di importanza molto limitata. Il secondo problema è che (sempre più spesso) le risposte date a questi questionari non sono del tutto 'oneste'. La tendenza a mentire è molto grande, e ciò inficia i risultati che rischiano di mettono in evidenza fattori poco realistici. Un bel lavoro di Stephens-Davidowitz (2017), analizza bene questi problemi, e fa notare pure che questo fenomeno tocca anche, spesso, i vari social network, dove l'essere in pubblico può condizionare fortemente l'*onestà* delle affermazioni che si fanno. Secondo Stephens-Davidowitz l'unico strumento neutro da questo punto di vista è costituito dalle ricerche effettuate su un motore di ricerca (Google, dove l'autore ha lavorato). Purtroppo, questi dati non sono, in generale, disponibili in maniera diretta e bisogna dedurli

interpretando quel poco che viene messo a disposizione pubblicamente (vedi per es. Google Trends, trends.google.it, che fornisce alcune indicazioni sugli argomenti più ricercati).

## 2 L'analisi: metodi e tecniche

La raccolta, l'elaborazione, l'analisi e l'interpretazione dei dati sono state per secoli il compito della statistica. Tutti i metodi che conosciamo e utilizziamo, tuttavia, sono stati progettati per trattare moderate quantità di dati strutturati. I Big Data pongono una serie di problemi che minano, in molti casi, la validità dei metodi statistici standard. Alcuni autori hanno persino suggerito che i Big Data segnino la fine dei metodi statistici (Anderson, 2008), sostenendo che il test di ipotesi, principio di base di molte procedure statistiche, è non più necessario e che i petabyte di dati disponibili online possono fornire tutte le risposte richieste, semplicemente cercando gli schemi e le relazioni presenti 'naturalmente'. Cosa che studi e ricerche approfondite hanno mostrato essere quantomeno debole (se non pericoloso) per quanto riguarda poi l'affidabilità e la validità dei risultati ottenuti (Boyd & Crawford, 2012). Infatti, c'è ancora un forte bisogno di teorie e modelli che si occupino del consolidamento e dell'interpretazione dei dati. Più precisamente, solo una 'teoria' è in grado di dedurre conclusioni da modelli nei dati in modo autoconsistente (Han, 2015). Da un punto di vista epistemologico, senza teorie i Big Data creano semplicemente (genericamente) 'numeri' indistinti disaccoppiati dalle realtà sociali. Secondo il filosofo tedesco Hegel (1830), un fenomeno non è materia caotica, caratterizzata dal caso, ma ha un suo sviluppo logico, poiché è il manifestarsi di una struttura razionale. Pertanto, le teorie sono e saranno sempre necessarie come 'via narrativa' per generare conoscenza, e gli approcci che utilizzano metodi misti e la creazione di significato attraverso vari modelli teorici sono di fondamentale importanza quando si ha a che fare con Big Data. Inoltre, essi potrebbero essere integrati e combinati con dati derivanti dalle tradizionali tecniche rigorose di raccolta e analisi e più familiari alle piccole e medie imprese turistiche (Coleman et al., 2016; Kitchin & Lauriault, 2015).

Quando manchi un obiettivo di ricerca molto chiaro e un piano di raccolta rigoroso dei dati necessari, il rischio di scoprire risultati insignificanti o conclusioni ingannevoli sono piuttosto alti. Un esempio è la revisione dell'indicatore di Google Trend sull'influenza che è stato contaminato da una serie di fattori non correlati (Lazer et al., 2014).

Le grandi quantità di dati soffrono di una nota 'maledizione della dimensionalità' (Keogh & Mueen, 2011), che spingono al limite (e oltre) i metodi analitici convenzionali (Fan et al., 2014). Per limitarci ad alcune delle questioni più importanti possiamo segnalare il concetto di significatività statistica, almeno per come è comunemente inteso. Quando vengono esaminati milioni di casi, la solita significatività statistica può perdere il suo valore e dare una falsa impressione di precisione. Molti test statistici, infatti, usano l'idea di misurare la distanza tra un'osservazione e un valore atteso dividendolo per una misura della variabilità dei dati (varianza), che viene calcolata come differenza media da un punto centrale (per es. la media) diviso per il numero di osservazioni raccolte. È ovvio pensare che se le dimensioni del campione analizzato sono estremamente grandi (dell'ordine di milioni) si possano ottenere significatività artificialmente gonfiate.

Un altro problema si presenta nella costruzione di modelli ricavati da osservazioni e tipicamente espressi in termini di correlazioni. È noto che si possono avere valori elevati di correlazione anche

se esistono scarse concordanze (Bland & Altman, 1986), poiché correlazioni significative possono essere scoperte anche quando vengono prese in considerazione sequenze completamente casuali o quando esistono effetti particolari come una tendenza temporale.

Infine, quando ci sono troppe variabili in un modello di regressione (cioè per es. quando il numero di parametri da stimare è maggiore del numero di osservazioni), il rischio del cosiddetto 'overfitting' è molto elevato. In altre parole, c'è un rischio molto elevato di ottenere errori o risultati casuali 'sporchi' invece di una relazione valida (Granville, 2013).

Gli studi sui Big Data hanno spesso un obiettivo esplorativo. In tali casi, molti dei problemi e delle preoccupazioni relative a possibili errori possono essere mitigati usando un approccio bayesiano (per una semplice introduzione vedi (Zyphur & Oswald, 2015). I metodi bayesiani si basano su un'interpretazione di probabilità come misura di incertezza che può essere modificata in base a regole decise quando si possono avere più informazioni rispetto a una configurazione iniziale. In particolare, essi consentono l'incorporazione di ipotesi nell'analisi (mediante distribuzioni di probabilità) e possono essere più facilmente utilizzati per risolvere problemi la cui struttura è troppo complesso per i metodi convenzionali. Inoltre, i metodi bayesiani possono meglio affrontare situazioni nelle quali i fenomeni continuano a evolvere man mano che si raccolgono i dati. Per loro natura, questi metodi sono pesanti dal punto di vista computazionale e richiedono validi strumenti hardware e software. Per questi ultimi, buone introduzioni con molti esempi sono reperibili, per esempio, nei testi di Kruschke ((2015), basato su R) e Downey ((2013), basato su Python).

Infine, bisogna tener presente che i grandi insiemi di dati provenienti da fonti online sono spesso inaffidabili, e la loro natura dinamica spesso impedisce qualsiasi tentativo di replicare uno studio a scopo di conferma. Inoltre, errori e lacune possono essere amplificati quando più insiemi diversi per origine, forma o dimensioni vengono utilizzati contemporaneamente (Sivarajah et al., 2017).

L'analisi di molti lavori teorici ed empirici che utilizzano i Big Data porta a formulare un percorso metodologico che sembra essere corretto ed efficace e che consiste nel:

- definire un obiettivo chiaro per quanto riguarda i risultati che ci si aspetta e definire quali sono le informazioni necessarie per raggiungere l'obiettivo;
- scegliere fonti di dati adeguate (per es. social media, transazioni online, traffico di telefonia mobile, motori di prenotazione ecc.);
- raccogliere e archiviare dati su larga scala, avvalendosi di tecniche di gestione dei dati appropriate, che vanno oltre le tradizionali strutture di data warehouse verso più flessibili *data lake* e altre tecnologie di archiviazione (vedi 2.2) che consentono di memorizzare grandi quantità di dati di diversi formati (numerici, testuali, immagini ecc.);
- pulire e validare i dati, tenendo conto delle loro caratteristiche di volatilità, replicabilità, variabilità, e di sovrapposizioni e ridondanze;
- combinare diverse fonti di dati per ottenere risultati completi, validi e affidabili a diversi livelli (individuo, impresa, rete aziendale, industria);
- estrarre conoscenze significative da questi grandi volumi di dati (considerando anche modelli teorici adeguati).

Un corretto approccio metodologico consentirà di utilizzare i contenuti generati spontaneamente dagli utenti su Internet e sul Web per acquisire conoscenze su convinzioni, comportamenti e

preferenze dei viaggiatori al fine di superare i limiti dei tradizionali metodi di ricerca basati su sondaggi.

## 2.1 Intelligenza artificiale e machine learning

Come si è detto sopra, le tecniche statistiche tradizionali sono quantomeno inefficaci nel trattamento e nell'analisi dei Big Data. Gli ultimi anni, però, hanno visto un incredibile sviluppo nel campo dell'intelligenza artificiale, che si esplicita, sostanzialmente, nella messa a punto di algoritmi sempre più sofisticati ed efficienti nello svolgere questi compiti.

Intelligenza artificiale (AI : artificiale intelligence) e apprendimento automatico (meglio noto col suo corrispondente inglese machine learning, ML) sono quasi diventati sinonimi, anche se ML è una delle possibili declinazioni dell'intelligenza artificiale, originariamente definita da Minsky (1967) come una tecnologia (o una macchina) in grado di svolgere un compito che, se condotto da un essere umano, richiederebbe 'intelligenza'. Le definizioni successive attribuiscono all'intelligenza artificiale capacità di apprendere, percepire, ragionare e agire, nonché di rilevare, determinare e sviluppare autonomamente decisioni per "scoprire quali elementi o attributi in un gruppo di dati sono i più predittivi" (Buhalis et al., 2019; Sterne, 2017).

L'apprendimento automatico, invece, è (Samuel, 1959: 211): "la programmazione di un computer digitale per eseguire operazioni in un modo che, se svolte da esseri umani o animali, verrebbe descritto come coinvolgente un processo di apprendimento". Esso implica, pertanto, la possibilità per un programma software di modificare il proprio comportamento in base agli eventi (input) e ai risultati, senza essere esplicitamente programmato per gestire ogni specifica situazione. Le applicazioni spaziano dai programmi di data mining che scoprono regole o schemi generali in grandi set di dati, a sistemi di filtraggio delle informazioni che identificano automaticamente gli utenti in base a interessi o preferenze, al raggruppamento di grandi raccolte di oggetti in un numero limitato di classi, al riconoscimento di forme o suoni, e così via.

Molte applicazioni oggi, e per ciò che è prevedibile nel prossimo futuro, forniscono risultati utili ed efficaci per l'analisi dei comportamenti, dei movimenti e dei desideri dei turisti, per la *business intelligence* generale e, per alcuni algoritmi, esistono esempi che rendono più efficaci diverse attività di previsione (Mariani et al., 2018). Inoltre, aree come i sistemi di raccomandazione possono sfruttare la grande quantità di dati che altrimenti sarebbero stati difficilmente impiegabili con i metodi tradizionali (Nilashi et al., 2017; Oussous et al., 2018; Portugal et al., 2018).

L'idea alla base delle tecniche di ML è che un algoritmo può scoprire regole o caratteristiche generali in grandi insiemi di dati utilizzando un approccio generalizzato senza essere esplicitamente disegnato per qualche lavoro specifico (Mitchell, 1997; Witten et al., 2016). Un algoritmo ML non fa ipotesi preliminari sulle possibili relazioni tra le variabili, ma, a volte guidato da alcuni esempi, elabora i dati e scopre configurazioni che possono essere utilizzate per individuare caratteristiche, classificare elementi o fare previsioni.

L'ampia serie di algoritmi ML può essere suddivisa approssimativamente in due famiglie principali: algoritmi supervisionati e non supervisionati. I primi sono quelli che richiedono una sorta di 'addestramento'. Un insieme di dati pre-classificati (o pre-etichettati) vengono dati in input all'algoritmo che deduce le caratteristiche dei vari elementi e crea un modello. Il modello viene poi verificato e validato usando parte dei dati originari e confrontando la classificazione iniziale con

quella derivata dal modello creato. Diversi algoritmi possono essere provati finché non si ottengono i livelli di accuratezza desiderati o necessari. Esempi di algoritmi di apprendimento automatico supervisionato sono i modelli di regressione e quelli di classificazione (naïve Bayes, support vector machine, k-neighbors ecc.).

Nel caso dell'apprendimento non supervisionato il modello viene preparato autonomamente a partire dai dati in ingresso. Questi algoritmi sono in genere ricorsivi, con successive approssimazioni e perfezionamenti che vengono eseguiti fino a quando la verifica mostra che viene raggiunto il livello di precisione specificato o atteso. Appartengono a questa classe gli algoritmi di clustering (gerarchici o k-means) e quelli usati per riduzioni di dimensionalità (analisi fattoriale).

La distinzione, tuttavia, non è sempre così chiara ed esistono forme miste. Inoltre, alcune tecniche prevedono un 'rinforzo' che consiste nel cominciare con una con una procedura di apprendimento supervisionata o non supervisionata e riutilizzare l'output, una volta verificato e validato, aggiungendolo all'input e aggiornando, eventualmente, il modello creato e le conclusioni precedenti e quindi migliorando notevolmente così l'accuratezza e la precisione dei risultati successivi. Questo processo dinamico può essere ripetuto molte volte permettendo, per esempio, un continuo raffinamento delle classificazioni ottenute o ottenibili.

Un'altra classe di algoritmi sono i cosiddetti 'deep learning' (DL, apprendimento profondo). Si tratta di un'area specializzata dell'apprendimento automatico che si riferisce ad algoritmi ispirati alla struttura e alla funzione del cervello chiamati reti neurali artificiali (ANN). Alcuni ritengono che questa che sia la tecnica più vicina agli obiettivi originali del più ampio campo dell'intelligenza artificiale, che mira a sviluppare sistemi e macchine che imitano le funzioni cognitive di un cervello umano (Arel et al., 2010; LeCun et al., 2015). L'architettura DL è stata applicata nella visione artificiale, nel riconoscimento automatico della lingua parlata, nell'elaborazione del linguaggio naturale, nel riconoscimento audio e nella trasformazione o nella creazione di oggetti digitali sintetici (immagini, audio, video) (Pouyanfar et al., 2019).

I sistemi DL utilizzano a cascata vari livelli di unità elementari per eseguire attività di estrazione e trasformazione. Ogni livello utilizza l'output del livello precedente come input. Sono essenzialmente un'applicazione delle reti neurali artificiali ad alcuni problemi di analisi dei dati. Una rete neurale artificiale (ANN) è un insieme di elementi software costruiti in modo simile a una rete neurale biologica. Più che un singolo algoritmo, un'ANN è una raccolta di diversi processi di apprendimento automatico (neuroni) interconnessi, che lavorano insieme e gestiscono input di dati complessi. Ogni connessione, come sinapsi in un cervello biologico, può trasmettere un segnale da un neurone a un altro. Il neurone ricevente può elaborarlo e segnalare ulteriormente ad altri neuroni collegati formando una catena di unità computazionali. Un neurone (nodo della rete) è un piccolo algoritmo che, in modo simile al neurone biologico, si attiva (trasmette un segnale) quando incontra uno stimolo appropriato. In pratica, un nodo combina input dai dati con un insieme di coefficienti, o pesi, che amplificano o riducono quell'input, assegnando un significato all'input in base al compito che l'algoritmo sta apprendendo. In altre parole, i pesi vengono assegnati in modo che l'output calcolato sia il più vicino possibile a un output predefinito (o atteso). L'intera rete neurale, nella sua forma più semplice, è composta da una serie di strati: gli elementi di input, lo strato neurale e l'output. Lo strato neurale è generalmente chiamato strato nascosto. Gli strati nascosti possono essere replicati e sono usati per costruire più livelli di astrazione che migliorano le capacità complessive della rete. I livelli multipli vengono ricalcolati iterativamente usando una tecnica di propagazione all'indietro per aggiornare i

pesi di ogni nodo. Questi livelli multipli si traducono in un apprendimento molto migliore per risolvere problemi complessi di riconoscimento proprio perché ad ogni livello intermedio aggiungono informazioni utili e migliorano l'affidabilità dell'output. In altre parole, ogni livello impara a diventare specializzato e ad attivarsi quando rileva caratteristiche specifiche.

Una rete neurale richiede grandi capacità computazionali, essa migliora le proprie prestazioni quando aumenta la quantità di dati, mentre le normali applicazioni di apprendimento automatico, una volta raggiunta una certa accuratezza, non sono più migliorabili anche aggiungendo ulteriori esempi e dati di addestramento. Nei sistemi di machine learning, infatti, le funzionalità di un determinato oggetto sono selezionate per la creazione di un modello, mentre nei sistemi di deep learning l'estrazione delle funzionalità avviene automaticamente: la rete neurale apprende autonomamente come analizzare i dati grezzi e come eseguire un'attività.

Le grandi capacità computazionali richieste dai sistemi DL sono, di solito, difficilmente accessibili alle singole aziende. Le più grandi aziende tecnologiche (Google, Amazon, Microsoft, IBM ecc.), tuttavia, offrono la possibilità di utilizzare questi sistemi in base a pay-per-use.

Le aree di applicazione usuali per i sistemi DL sono il riconoscimento e la classificazione di immagini, voce e testo. Oltre a ciò, queste capacità di classificazione sono state impiegate per sviluppare traduzioni automatizzate per diverse lingue o sistemi di guida per veicoli senza pilota o per costruire oggetti ‘artificiali’ da pezzi scritti, dipinti, voci o suoni.

L'evoluzione di questi sistemi ha fornito, negli ultimi anni, un'incredibile serie di risultati. Il riconoscimento delle immagini, ad esempio, ha raggiunto un tasso di errore di circa il 2,3% quando il limite umano è di circa il 5% (Hu et al., 2018). Una delle applicazioni più interessanti, tuttavia, è come strumento per attività specializzate. Per esempio, la combinazione di ANN avanzate con la competenza di un medico esperto è stata in grado di ottenere un tasso di errore dello 0,5% nel riconoscimento del carcinoma mammario femminile (Wang et al., 2016).

Un altro bell’esempio recente è il completamento della sinfonia n. 8 in si minore D 759, di Franz Schubert, comunemente detta Incompiuta. Un sistema DL ha ‘dedotto’ la melodia per i due movimenti mancanti che son stati poi arrangiati da un compositore (vedi: https://www.classicfm.com/composers/schubert/unfinished-symphony-completed-by-ai/).

Nel settore del turismo stanno diventando disponibili applicazioni che combinano, per esempio, le capacità di riconoscimento delle immagini con la realtà aumentata per la progettazione di guide sulle attrazioni turistiche (Zhou et al., 2019) e sono reperibili software e app mobili che forniscono servizi di traduzione ragionevoli (per es. www.deepl.com/translator, o Google traduttore). Sistemi DL, e in particolare quelli per il riconoscimento vocale, sono ormai comunemente utilizzati nei dispositivi robotici come quei robot umanoidi che iniziano ad apparire negli hotel di tutto il mondo e che aiutano, in genere, concierge e front-desk a rispondere alle domande e le richieste dei clienti (Bowen & Morosan, 2018). Inoltre, tecniche DL si stanno dimostrando efficaci per migliorare le capacità di previsione della domanda turistica (Law et al., 2019).

## 2.2 Tecnologie per i big data

Le caratteristiche dimensionali e di variabilità e dinamicità dei Big Data pongono anche dei limiti alle infrastrutture hardware e software e alle tecnologie utilizzabili per la raccolta e la conservazione dei

dati. I tradizionali sistemi database, infatti, mal ‘digeriscono’ formati non strutturati e, soprattutto, sono inefficienti quando si tratta di gestire grandi quantità di registrazioni.

L’architettura più conosciuta e utilizzata per operare con i Big Data è Hadoop (White, 2015), progetto open source Java di Apache Software Foundation per l'archiviazione e l'elaborazione distribuita su larga scala di insiemi di dati su cluster di computer di qualunque tipo. Hadoop supporta un numero illimitato di nodi (computer) che possono contenere ed elaborare grandi quantità di dati (dell'ordine dei petabyte: $10^{15}$ byte). L'architettura è progettata per essere altamente modulare e si basa su un file system distribuito (Hadoop Distributed File System: HDFS) originariamente sviluppato da Google e altamente tollerante agli errori. HDFS è progettato per trattare in modo affidabile dataset di dati di grandi dimensioni e per trasmetterli in streaming ad alta velocità per le varie applicazioni.

Altra tecnica utilizzata è Map-Reduce, un sistema sviluppato e brevettato da Google, che consente l'elaborazione parallela di grandi archivi. Le applicazioni Map-Reduce sono divise in due parti e sono eseguite utilizzando un gran numero di nodi collegati in un cluster. La funzione Map divide una richiesta (query) in varie parti ed elabora i dati a livello del singolo nodo. La funzione Reduce combina i diversi risultati nel nodo principale (quello dal quale è partita la richiesta) per ottenere il risultato finale (Miner & Shook, 2012).

Queste tecnologie modulari e integrabili fra di loro sono la base per includere o supportare diverse altre tipologie di strumenti, dai database relazionali tradizionali (RDBMS), ancora ben utilizzati, a strutture NoSQL (non solo SQL) che forniscono un accesso efficiente in tempo reale a dati in svariati formati o di dimensioni per le quali i tradizionali RDBMS hanno problemi di prestazioni (McCreary & Kelly, 2014; Strauch et al., 2011).

Diverse applicazioni per estrarre, trasformare e caricare dati (ETL: extract, transform and load) da e verso fonti eterogenee, e moderni sistemi di data warehouse a di reporting sono poi costruiti su queste architetture (Cloudera, 2014; Russom, 2013).

Va notato qui che, benché in linea di principio sia possibile utilizzare questi sistemi in ambienti limitati (organizzazioni o gruppi di aziende), le risorse, soprattutto hardware, necessarie possono essere notevoli. Il loro utilizzo prevalente avviene quindi attraverso i servizi cloud messi a disposizione dalle grandi aziende del mondo digitale (Google, Amazon, Microsoft, IBM, Apple ecc.).

## 2.3 Una digressione etica

A chiusura di questo contributo non si può non considerare almeno alcuni dei temi di natura etica che riguardano il mondo digitale quando si considerano le numerosissime tracce digitali che vengono lasciate ogni giorno da svariati milioni di utenti nei loro utilizzi e che sono la fonte di tutti i Big Data e le tecniche avanzate messe a disposizione dagli sviluppi dell’intelligenza artificiale e della robotica (Floridi, 2013; Taddeo & Floridi, 2018).

Al di là delle reazioni allarmate di alcuni di fronte ai ‘pericoli’ degli sviluppi contemporanei della tecnologia, che però sono abbastanza naturali quando nuovi strumenti vedono la luce, le questioni etiche più profonde che riguardano AI e robotica possono essere affrontate in almeno due modi. Uno riguarda gli sviluppatori di applicazioni e sistemi che dovrebbero essere consapevoli delle possibili conseguenze, spesso non volute, del loro lavoro e predisporre modalità atte a evitare abusi e consentire, finché possibile, un’ispezione umana delle funzionalità degli algoritmi e dei sistemi. In

secondo luogo, quando ci si muove così rapidamente verso sistemi sempre più sofisticati e, soprattutto, sempre più autonomi, questi dovrebbero essere dotati di funzionalità che permettano di prendere decisioni 'etiche' per ridurre il rischio di comportamenti indesiderati, o comunque di funzionalità che forniscano una 'spiegazione' dei criteri utilizzati per arrivare a quelle decisioni (Torresen, 2018). Inoltre, da definire è il concetto di 'responsabilità' morale e soprattutto legale di un algoritmo o un robot e da chi (sviluppatore del software, costruttore della macchina, utilizzatore, fornitore dei dati ecc.) questa responsabilità viene assunta nel caso di problemi o incidenti (Turner, 2019: 81-132).

I temi sul tappeto sono molti, dalle preoccupazioni per il futuro del lavoro e di come le 'macchine' sostituiranno quanto l'uomo fa oggi e nell'immediato futuro, alle disuguaglianze che si possono generare e acuire fra chi ha a disposizione questi strumenti e chi non se ne può avvalere, ai rischi di incidenti provocati da malfunzionamenti o da errati utilizzi, ai possibili pregiudizi inseriti negli algoritmi e derivanti dall'uso di dati 'parziali' durante il loro addestramento o da posizioni e convinzioni degli sviluppatori, alle capacità di influenzare e modificare comportamenti (Torresen, 2018).

Questi ultimi due temi sono fra i più discussi oggi. I recenti scandali di Cambridge Analytica (Cadwalladr & Graham-Harrison, 2018) e le discussioni generate dalla pubblicazione di alcuni lavori (Kramer et al., 2014) hanno portato alla luce la possibilità concreta che la combinazione di meccanismi automatici e di analisi approfondite dei dati a disposizione possano 'facilitare' azioni che inducono modifiche di opinioni e comportamenti da parte degli utenti di certe applicazioni e che possano poi avere riflessi significativi nelle loro attività. E l'uso di dati di addestramento parziali o poco equilibrati potrebbero portare a decisioni automatiche fortemente pregiudiziali per esempio in campo finanziario (erogazione di prestiti, mutui ecc.) o legale (previsione di inclinazione al crimine di certe classi o categorie di persone (Richardson et al., 2019).

Non è questa la sede per una discussione approfondita di questi temi, ma va qui ricordato che, al di là della sensazione un po' 'magica' delle possibilità dei moderni sistemi di intelligenza artificiale e degli sviluppi rapidissimi delle loro capacità e funzionalità, si tratta comunque di sistemi software, sviluppati da esseri umani (almeno per il momento) e che come tali riflettono idee, filosofie, epistemologie e pregiudizi dei loro creatori. Inoltre, il loro utilizzo si basa su quanto già disponibile e quindi una raccolta parziale o non equilibrata può condizionare fortemente i risultati.

Diversamente da quanto accaduto nel passato, però, un buon numero di organizzazioni e istituzioni pubbliche e private, nazionali e internazionali, si sono già attivate nel proporre soluzioni a questi dilemmi nell'unico modo possibile (almeno per ora): quello di ragionare in termini di linee guida per coloro i quali progettano e sviluppano sistemi a applicazioni di intelligenza artificiale.

Alla base di tutto ci sono le questioni che riguardano la sicurezza, la riservatezza e la protezione dei dati personali (Fang et al., 2017). Su queste molti hanno, già da tempo, attivato leggi e norme che dovrebbero proteggere gli utenti (ricordiamo qui solo il recentissimo regolamento generale sulla protezione dei dati dell'Unione Europea (Voigt & Von dem Bussche, 2017). Nonostante le buone intenzioni, però, va notato che tutte queste leggi sono ancora emanate (e hanno validità) in modo 'geografico' e difficilmente, viste anche le diverse impostazioni e filosofie di base, possono efficacemente garantire persone che usano un ambiente senza confini di alcun tipo.

Per quanto riguarda gli altri temi etici ricordiamo qui le più recenti e importanti iniziative:

- *Future of Life Institute*: Asilomar AI Principles. Documento sviluppato in collaborazione con la conferenza Asilomar del 2017, questo elenco di principi è citato universalmente come punto di riferimento da tutti gli altri framework e standard di etica AI introdotti da quando è stato pubblicato [https://futureoflife.org/ai-principles/]
- *IAPP* (International Association of Privacy Professionals): il white paper Building Ethics into Privacy Frameworks for Big Data and AI descrive e discute gli strumenti disponibili per le organizzazioni che cercano non solo di sviluppare strutture interne ma di rendere operative le politiche di etica dei dati [https://iapp.org/resources/article/building-ethics-into-privacy-frameworks-for-big-data-and-ai/]
- *IEEE* (Institute of Electrical and Electronics Engineers). La più grande associazione mondiale del settore ha da tempo attivato un'iniziativa: l'IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems che ha pubblicato Ethically Aligned Design, un ricco e completo manuale per affrontare 'valori e attenzioni, nonché implementazioni' di questi sistemi [https://ethicsinaction.ieee.org]
- *Public Voice*: coalizione istituita nel 1996 dall'Electronic Privacy Information Center (EPIC) per promuovere la partecipazione del pubblico alle decisioni riguardanti il futuro di Internet. Collabora con istituzioni come OCSE, UNESCO e associazioni e istituzioni di circa 80 paesi (per l'Italia Garante per la protezione dei dati personali). Ha pubblicato le Universal Guidelines for Artificial Intelligence, scritte per essere 'incorporate negli standard etici, adottate nella legislazione nazionale e negli accordi internazionali e integrate nella progettazione dei sistemi'. È un documento che pone l'enfasi sulla trasparenza, l'equità, l'accuratezza e la qualità dei dati e introduce l'obbligo del governo di limitare la profilazione segreta o il punteggio dei cittadini [https://thepublicvoice.org/ai-universal-guidelines/]
- *OCSE*: Gli OECD Principles on Artificial Intelligence sono linee guida e costituiscono una raccomandazione adottata a maggio 2019 dai paesi membri. Sono tra i primi principi di questo tipo sottoscritti dai governi e base per il documento del G20 su *Human-centred AI Principles* del giugno 2019 [https://www.oecd.org/going-digital/ai/principles/]
- *Commissione Europea*: Linee guida sull'intelligenza artificiale e la protezione dei dati. Redatto da un gruppo indipendente di consulenti etici dell'IA e ottimizzato utilizzando più di 500 commenti pubblici raccolti durante un periodo di feedback di cinque mesi, è uno dei più recenti e completi framework pubblici sull'etica dell'IA fino ad oggi. Non è un documento politico ufficiale o un regolamento della Commissione europea, ma piuttosto un insieme di suggerimenti intesi a guidare il discorso pubblico su come si presenta un'AI affidabile [https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai]

Queste linee guida hanno, in generale, lo scopo di aiutare i progettisti e gli utenti di AI a scegliere sistemi che siano corretti, etici e robusti, e forniscono un primo tentativo di rispondere alle tre grandi domande: Chi è responsabile per i danni o i benefici che derivano dall'uso dell'intelligenza artificiale? Dovrebbero i dispositivi (hardware e software) dell'AI avere diritti? E come dovrebbero essere stabilite e attivate regole etiche per l'AI? Esistono, ovviamente, molte somiglianze fra questi documenti e, di fatto, hanno tutti più o meno esplicitamente origine dalle famose tre leggi della robotica pubblicate nel 1950 da Isaac Asimov (Asimov, 1950):

1. *Un robot non può recar danno a un essere umano né può permettere che, a causa del suo mancato intervento, un essere umano riceva danno.*

2. *Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché tali ordini non vadano in contrasto alla Prima Legge.*
3. *Un robot deve proteggere la propria esistenza, purché la salvaguardia di essa non contrasti con la Prima o con la Seconda Legge.*

# 3 Conclusioni

I Big Data sembrano essere sempre più non solo uno slogan promozionale, ma un vero e proprio settore in grande sviluppo e capace di fornire informazioni preziose per l'operatività quotidiana e per la definizione e l'attuazione di strategie e politiche in ogni campo di attività. Ancor più preziosi si rivelano poi per il comparto turistico, ambiente a elevata complessità e dinamicità e che coinvolge un numero e una varietà impressionante di operatori, attività, tecnologie, individui.

Da un punto di vista metodologico, gli approcci basati sui Big Data consentono di superare le difficoltà di lavorare con limitati 'campioni rappresentativi' poiché è possibile, virtualmente, analizzare intere popolazioni (Gerard et al., 2016), e presumibilmente, di rispondere a molte domande relative alle opinioni, idee e comportamenti delle persone utilizzando le loro esternazioni (più o meno) spontanee, superando così anche i limiti e i rischi di indagini troppo legate alle idee (e a volte ai preconcetti) del ricercatore. Allo stesso tempo, essi sembrano essere un potente strumento per affrontare nuove domande di ricerca e per sviluppare progetti innovativi per il progresso della conoscenza, generando in ultima analisi un buon supporto politico e gestionale.

Da un lato, vi è consenso tra imprenditori e studiosi sul fatto che i Big Data rappresentino un elemento necessario per indagare sui complessi fenomeni economici e sociali di oggi attraverso la possibilità di combinare e ricombinare fonti di informazione estremamente diverse (Bedeley & Nemati, 2014). Allo stesso modo, le aziende che li sfruttano possono migliorare il proprio vantaggio competitivo in un mondo in cui i mercati sono globali ed enormi quantità di informazioni sui consumatori sono disponibili su Internet (Verhoef et al., 2016).

D'altra parte, i Big Data portano con sé un numero significativo di nuove sfide, rischi e problemi, che sono stati esplorati e affrontati in una serie di lavori (Boyd & Crawford, 2012; Floridi, 2013; Gerard et al., 2016; McFarland & McFarland, 2015; Verhoef et al., 2016) che discutono ed esemplificano i problemi legati alla condivisione dei dati e della privacy o nuovi dilemmi epistemologici, ma, soprattutto, le sfide relative all'estrazione, raccolta, archiviazione, elaborazione, analisi e visualizzazione e reportistica dei dati (Gandomi & Haider, 2015). Questi processi, che saranno sempre più contrapposti a metodologie più consolidate e tradizionali, richiedono risorse specifiche e competenze specializzate e diversificate, (Kitchin, 2015). Ciò implica che un approccio 'disciplinare' non è sufficiente per l'analisi e l'utilizzo dei Big Data, poiché le competenze e gli strumenti necessari vanno oltre le conoscenze compartimentalizzate e i dialoghi occasionali che sono (quando lo sono) normalmente utilizzati in casi simili.

Nonostante i molti problemi, comunque, esiste la consapevolezza, all'interno dei circoli accademici e aziendali, che i Big Data possono fare la differenza in quanto catturano comportamenti e opinioni in tempo reale su praticamente qualsiasi aspetto della vita umana (Chang et al., 2014).

Va ribadito qui, comunque, che, soprattutto per via delle caratteristiche dei dati e dei metodi di trattamento e analisi (ML e DL), è essenziale adottare un approccio metodologico rigoroso e corretto, basato sulla formulazione di ipotesi o teorie da verificare (o da smentire), e che queste domande di ricerca devono essere dinamicamente modificate al mutarsi delle condizioni dell'ambiente nel quale questi strumenti vengono utilizzati.

Nel mondo del turismo e dell'ospitalità l'interesse verso i Big Data è grande (Heerschap et al., 2014; Wise & Heidari, 2019), e molte applicazioni sono già attive nel campo della previsione (Pan & Yang, 2016), del marketing (Park et al., 2015) e della gestione delle destinazioni (Fuchs et al., 2014; RocaSalvatella, 2014). Interessanti poi le informazioni ricavabili dallo studio multidisciplinare delle registrazioni di attività 'telefonica' ottenute analizzando i CDR che possono evidenziare caratteristiche e problemi legati ai flussi di visitatori e alle loro preferenze in tema di itinerari in una destinazione (Baggio & Scaglione, 2018a, 2018b; Vanhoof et al., 2017).

Tuttavia, come alcune revisioni sistematiche della letteratura scientifica mostrano (Li et al., 2018; Mariani et al., 2018), il mondo del turismo sembra segnare il passo e, anche se si nota un costante aumento dei lavori che applicano tecniche analitiche utilizzando Big Data, questo campo di ricerca ha una portata abbastanza frammentata e limitata nelle metodologie, e presenta diverse lacune. Manca, spesso, un quadro concettuale che possa aiutare a identificare i problemi più critici e collegare i Big Data alla gestione e allo sviluppo del turismo e dell'ospitalità. Sono poi, a dispetto del ricorso al termine, relativamente pochi i lavori che utilizzino davvero grandi quantità di dati e le metodologie utilizzate sono spesso abbastanza 'datate' o limitate all'uso di strumenti software preconfezionati. Qui, anche nel mondo della ricerca, la carenza di competenze avanzate nelle tecniche di raccolta, trattamento e analisi si fanno sentire. Nel turismo (e non solo), viste le risorse necessarie in termini di risorse umane, attrezzature e investimenti, si avverte fortemente la preponderanza di grandi aziende e organizzazioni che sembrano essere molto più avanzate dell'accademia in questo campo (vedi per esempio l'articolo del Guardian sull'argomento: https://www.theguardian.com/science/2017/nov/01/cant-compete-universities-losing-best-ai-scientists). A maggior ragione aziende turistiche piccole e frammentate, con scarsa propensione alla collaborazione e agli investimenti, sono quasi assenti nei discorsi sui Big Data e sull'intelligenza artificiale. Qui l'unica soluzione possibile sarebbe quella di poter contare su aggregazioni collaborative che siano in grado di creare quella massa critica di risorse necessarie.

## Bibliografia


Anderson, C. K. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. *Wired Magazine*. Retrieved December, 2015, from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research. *IEEE computational intelligence magazine, 5*(4), 13-18.

Asimov, I. (1950). *"Runaround", in I, Robot (The Isaac Asimov Collection ed.)* New York City: Doubleday.

Baggio, R., & Klobas, J. (2017). *Quantitative Methods in Tourism: A Handbook* (2nd ed.). Bristol, UK: Channel View.

Baggio, R., & Scaglione, M. (2018a). Destination attractions system and Strategic Visitor Flows: An exploratory study. In F. Sánchez, C. Pautasso & K. Systä (Eds.), *Current Trends in Web Engineering, ICWE2018 workshops* (pp. 227-237). Cham (CH): Springer.

Baggio, R., & Scaglione, M. (2018b). Strategic Visitor Flows and destination management organization. *Information Technology and Tourism, 18*(1-4), 29-42.

Bedeley, R., & Nemati, H. (2014). *Big Data Analytics: A Key Capability for Competitive Advantage.* Paper presented at the 20th Americas Conference on Information Systems (AMCIS), Savannah, GA, 7-9 August.

Blair, A. M. (2010). *Too much to know: Managing scholarly information before the modern age*. New Haven, CT: Yale University Press.

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310.

Bowen, J., & Morosan, C. (2018). Beware hospitality industry: the robots are coming. *Worldwide Hospitality and Tourism Themes, 10*(6), 726-733.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662-679.

Braun, T., & Zsindely, S. (1985). Growth of scientific literature and the Barnaby Rich effect. *Scientometrics, 7*(529-530).

Buhalis, D., Harwood, T., Bogicevic, V., Viglia, G., Beldona, S., & Hofacker, C. (2019). Technological disruptions in services: lessons from tourism and hospitality. *Journal of Service Management,* (doi: 0.1108/JOSM-12-2018-0398).

Cadwalladr, C., & Graham-Harrison, E. (2018). *The Cambridge analytica files*. *The Guardian*. Retrieved April, 2019, from http://davelevy.info/Downloads/cabridgeananalyticafiles%20-theguardian_20180318.pdf.

Car, T., Stifanich, L. P., & Šimunić, M. (2019). *Internet of things (IoT) in tourism and hospitality: opportunities and challenges.* Paper presented at the 5th International Scientific Conference on Tourism in Southern and Eastern Europe, Opatija, Croatia (May 16–18).

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems, 63*, 67-80.

Cloudera. (2014). *Data Warehouse Optimization with Hadoop*. Retrieved September 2018, from https://www.informatica.com/content/dam/informatica-com/en/collateral/white-paper/data-warehouse-optimization-hadoop_white-paper_2609.pdf.

Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort-Martorell, X., & Seabra Reis, M. (2016). How can SMEs benefit from big data? Challenges and path forward. *Quality and Reliability Engineering International, 32*(6), 2151-2164.

d'Amore, M., Baggio, R., & Valdani, E. (2015). A practical approach to big data in tourism: a low cost Raspberry Pi cluster. In I. Tussyadiah & A. Inversini (Eds.), *Information and Communication Technologies in Tourism 2015 (Proceedings of the International Conference in Lugano, Switzerland, February 3-6)* (pp. 169-181). Berlin - Heidelberg: Springer.

Downey, A. (2013). *Think Bayes*. Needham, MA: Green Tea Press /O'Reilly Media.

Eckenrode, T. R. (1984). Vincent of Beauvais: a Study in the Construction of a Didactic View of History. *The Historian, 46*(3), 339-360.

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review, 1*(2), 293-314.

Fang, W., Wen, X. Z., Zheng, Y., & Zhou, M. (2017). A survey of big data security and privacy preserving. *IETE Technical Review, 34(*(5), 544-560.

Floridi, L. (2013). *The ethics of information*. Oxford: Oxford University Press.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations – A case from Sweden. *Journal of Destination Marketing & Management, 3*(4), 198-209.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Gerard, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal, 59*(5), 1493-1507.

Granville, V. (2013). *The curse of big data*. Retrieved June, 2014, from http://www.analyticbridge.com/profiles/blogs/the-curse-of-big-data.

Han, B.-C. (2015). *Psychopolitik – Neoliberalismus Und Die Neuen Machttechniken*. Frankfurt a.M., Germany: Fischer.

Heerschap, N., Ortega, S., Priem, A., & Offermans, M. (2014). *Innovation of tourism statistics through the use of new big data sources.* Paper presented at the 12th Global Forum on Tourism Statistics, Prague, CZ, 15-16 May. Retrieved July 2014 from http://www.tsf2014prague.cz/assets/downloads/Paper%201.2_Nicolaes%20Heerschap_NL.pdf.

Hegel, G. W. F. (1830). *Enzyklopädie Der Philosophischen Wissenschaften Im Grundrisse – Die Wissenschaft Der Logik*. Hamburg: Felix Meiner Verlag.

Hemp, P. (2009). Death by information overload. *Harvard Business Review, 87*(9), 82-89.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT (June 18-22)*, 7132-7141.

Inanc–Demir, M., & Kozak, M. (2019). Big Data and Its Supporting Elements: Implications for Tourism and Hospitality Marketing. In M. Sigala, R. Rahimi & M. Thelwall (Eds.), *Big Data and Innovation in Tourism, Travel, and Hospitality* (pp. 213-223). Singapore: Springer.

Keogh, E., & Mueen, A. (2011). Curse of dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 257-258). New York: Springer.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS, 31*(3), 471-481.

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal, 80*(4), 463-475.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788-8790.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan* (second ed.). London: Academic Press.

Laney, D. (2001). *3D data management: controlling data volume, velocity and variety* (Research Note 949). Retrieved October 2015, from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research, 75*, 410-423.

Lazer, D. M., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(14), 1203-1205.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521(*(7553), 436-444.

Lee, I., & Lee, K. (2015). The Internet of Things (IoT): Applications, investments, and challenges for enterprises. *Business Horizons, 58*(4), 431-440.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management, 68*, 301-323.

Lohr, S. (2013). *The Origins of 'Big Data': An Etymological Detective Story*. *The New York Times*. Retrieved December 2019, from https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., & Sheth, A. P. (2018). Machine learning for Internet of Things data analysis: A survey. *Digital Communications and Networks, 4*(3), 161-175.

Mariani, M., Baggio, R., Fuchs, M., & Höpken, W. (2018). Business Intelligence and Big Data in Hospitality and Tourism: A Systematic Literature Review. *International Journal of Contemporary Hospitality Management, 30*(12), 3514-3554.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review, 90*(10), 60-68.

McCreary, D., & Kelly, A. (2014). *Making sense of NoSQL: a guide for managers and the rest of us*. Shelter Island, NY: Manning.

McFarland, D. A., & McFarland, H. R. (2015). Big Data and the danger of being precisely inaccurate. *Big Data & Society, 2*(2), 1-4.

Miner, D., & Shook, A. (2012). *MapReduce design patterns: building effective algorithms and analytics for Hadoop and other systems*. Sebastopol, CA: O'Reilly Media, Inc.

Minsky, M. L. (1967). *Computation: Finite and Infinite Machines*. Upper Saddle River, NJ: Prentice-Hall.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw Hill.

Nilashi, M., Bagherifard, K., Rahmani, M., & Rafe, V. (2017). A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Computers & industrial engineering, 109*, 357-368.

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences, 30*(4), 431-448.

Pan, B., & Yang, Y. (2016). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research, doi: 10.1177/0047287516669050*.

Park, S. B., Ok, C. M., & Chae, B. K. (2015). Using Twitter Data for Cruise Tourism Marketing and Research. *Journal of Travel & Tourism Marketing*, doi: 10.1080/10548408.10542015.11071688.

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications, 97*, 205-227.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2019). A survey on deep learning: Algorithms, techniques, and applications. (CSUR), 51(5), . *ACM Computing Surveys, 51*(5), art. 92.

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online, 94*(4), 193-233.

RocaSalvatella. (2014). *Big Data and Tourism: New Indicators for Tourism Management*. Barcelona: Telefónica I+D and RocaSalvatella. Retrieved September, 2015, from http://www.rocasalvatella.com/sites/default/files/big_data_y_turismo-eng-interactivo.pdf.

Ross, J.-M. (2010). *Roger Magoulas on Big Data*. Retrieved February, 2019, from http://radar.oreilly.com/2010/01/roger-magoulas-on-big-data.html.

Russom, P. (2013). *Integrating Hadoop into Business Intelligence and Data Warehousing* (TDWI Best Practices Report). Renton, WA: TDWI Research. Retrieved November 2018, from http://www.datascienceassn.org/sites/default/files/Integrating%20Hadoop%20into%20DW%20-%20BI.pdf.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263-286.

Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*. New York, NY: HarperCollins.

Sterne, J. (2017). *Artificial Intelligence for Marketing: Practical Applications*. Hoboken, NJ: Wiley.

Strauch, C., Sites, U. L. S., & Kriha, W. (2011). *NoSQL databases. Lecture Notes*. Stuttgart: Stuttgart Media University. Retrieved December 2018, from https://christof-strauch.de/nosqldbs.pdf.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751-752.

Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI, 4*, art. 75.

Turner, J. (2019). *Robot Rules*. Cham, CH: Palgrave Macmillan.

Vanhoof, M., Hendrickx, L., Puussaar, A., Verstraeten, G., Ploetz, T., & Smoreda, Z. (2017). Exploring the use of mobile phone data for domestic tourism trip analysis. *Netcom. Réseaux, communication et territoires, 31*, 335-372.

Verhoef, P. C., Kooge, E., & Walk, N. (2016). *Creating Value with Big Data Analytics: Making Smarter Marketing Decisions*. London: Routledge.

Voigt, P., & Von dem Bussche, A. (2017). *The EU general data protection regulation (GDPR). A Practical Guide*. Cham, CH: Springer International Publishing.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). *Deep learning for identifying metastatic breast cancer* (arxiv/1606.05718). Retrieved November 2018, from http://arxiv.org/abs/1606.05718.

White, T. (2015). *Hadoop: The Definitive Guide, 4th Edition - Storage and Analysis at Internet Scale*. Sebastopol, CA: O'Reilly Media.

Wise, N., & Heidari, H. (2019). Developing Smart Tourism Destinations with the Internet of Things. In M. Sigala, R. Rahimi & M. Thelwall (Eds.), *Big Data and Innovation in Tourism, Travel, and Hospitality* (pp. 21-29). Singapore: Springer.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Cambridge, MA: Morgan Kauffman.

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120-130.

Zhou, X., Sun, Z., Xue, C., Lin, Y., & Zhang, J. (2019). Mobile AR Tourist Attraction Guide System Design Based on Image Recognition and User Behavior. *Proceedings of the 2nd International Conference on Intelligent Human Systems Integration (IHSI 2019), San Diego, CA (February 7-10)*, 858-863.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A User's Guide. *Journal of Management, 41*(2), 390-420.