DATA MINING E *MACHINE LEARNING*: LA NUOVA FRONTIERA DEI *BIG DATA* NEL SETTORE IMMOBILIARE

Maurizio Festa

Data mining e *Machine learning*: la nuova frontiera dei *Big data* nel settore immobiliare

DOI: 10.14609/Ti_1_21_3i

Maurizio Festa *

Parole chiave: *Big Data, Data mining, Machine learning, Predictive analysis, Data science*, settore immobiliare

Abstract

Questo articolo intende apportare un contributo alla comprensione del fenomeno dei *Big Data* fornendo un quadro d'insieme sullo stato dell'impiego di metodologie di *Data mining* e *Machine learning* nel settore immobiliare. In un momento storico nel quale l'enorme mole di dati disponibili sovrasta le capacità umane di gestione, elaborazione e analisi, si delinea una rivoluzione nei processi decisionali, alla base dei quali non vi è più solo esperienza, competenza e intuizione ma soprattutto conoscenza estratta dai *Big data*. Un siffatto insieme di dati ha condotto allo sviluppo di nuove tecniche di elaborazione ed evoluzioni di quelle tradizionali, confluendo in quella che è considerata un'evoluzione della statistica esplorativa, il *Data mining*. In quest'ambito occupa un ruolo chiave la figura del *Data scientist*, oggi presentato come una delle professioni più ricercate nel mercato del lavoro e al quale sono richieste non solo conoscenze tecniche avanzate, quali quelle di statistica, matematica, informatica, ma anche capacità di sviluppare e implementare algoritmi di *Machine learning*, di analisi, creatività e doti comunicative. I dati sono diventati una componente chiave in tutti i settori dell'economia moderna, anche per il settore immobiliare che però, essendo un settore tradizionalmente frammentato ed eterogeneo, ha registrato un ritardo nell'innovazione tecnologica soprattutto per la mancanza di *set* di dati strutturati che ha quindi comportato una maggiore dispersione delle informazioni, asimmetrie informative diffuse e conseguente opacità del mercato. Diversi esempi di applicazioni nel settore immobiliare sono stati sviluppati su modelli di *Data mining* e in particolare nel campo dell'intermediazione, della valutazione, della gestione e dello sviluppo immobiliare. *Big data* e *Data mining*, però, non sono esenti da limiti e criticità. Assicurare il giusto equilibrio tra diritto alla trasparenza e *privacy*, spingere sull'espansione degli *open data*, migliorare le tecniche di selezione dei dati, eliminare l'opacità dei modelli *black box* di *Artificial Intelligence*, certificare l'affidabilità dei modelli di *Machine learning*, gestire il ciclo di vita dei dati, sono solo alcuni dei problemi aperti che bisognerà affrontare per costruire processi di *Data mining* efficienti.

* Capo Ufficio Statistiche e studi sul mercato immobiliare – OMI Agenzia delle Entrate.

I contenuti dell'articolo impegnano la responsabilità dell'autore e non riflettono necessariamente la posizione dell'Ente al quale afferisce.

1. INTRODUZIONE

La stima annuale di MSCI inc.[1] della dimensione monetaria dello *stock* degli investimenti nel settore immobiliare mondiale, gestito da società professionali, restituisce per il 2020 un valore di 10,5 trilioni di dollari (MSCI, 2021). Si tratta di un dato ancora in crescita (+0,9 trilioni di dollari sul 2019, attestandosi a 1,6 volte la dimensione del 2010), mostrando un'inattesa capacità di resistenza all'ondata pandemica di Covid-19 che ha contrassegnato l'economia globale.

In un settore economico di tale rilevanza è massima l'attenzione sullo sviluppo di nuove tecnologie e nuovi modelli di *business,* per i quali l'utilizzo dei dati diventa fondamentale per ampliare le opportunità e creare valore dalla conoscenza.

In questo articolo si proverà a fornire un quadro d'insieme sullo stato dell'impiego di metodologie di *Data mining* e *Machine learning,* con un *focus* sui *Big data* del settore immobiliare. Nuovi *dataset*, caratterizzati dalla granularità delle informazioni, possiedono elevate potenzialità che potrebbero consentire la costruzione di modelli più efficaci per la descrizione di fenomeni socio economici. Il nuovo filone inerente la divulgazione di dati originari non filtrati da elaborazioni statistiche non è però ancora una realtà per il settore immobiliare, ma gli sviluppi nei prossimi anni potrebbero avere significativi riflessi sul *Real estate*. Il *Data mining,* però, non è esente da limiti e criticità, di cui bisogna essere consapevoli per poter trarre il valore conoscitivo che i *Big data* promettono.

L'articolo è strutturato come segue: dopo questa introduzione, nel paragrafo 1, sono richiamate le definizioni associate ai *Big data* e le principali caratteristiche; successivamente, nel paragrafo 2, si descrive la figura dell'esperto che lavora con i *Big data*; nel paragrafo 3 viene descritto il processo di *Data mining* e le principali tecniche utilizzate per elaborare i *Big data*; nel paragrafo 4 viene presentato un approfondimento sui *Big data* del settore immobiliare e alcuni esempi di applicazioni nel *Real estate*; nel paragrafo 5 si pone l'attenzione su limiti e criticità dei *Big data* e dei processi di *Data mining* e, infine, alcune osservazioni conclusive sono fornite nel paragrafo 6.

Per l'esame dell'ampia letteratura accademica disponibile si è fatto riferimento alle dettagliate ricostruzioni di Kitchin (2013, 2014 e 2016), Baldominos (2018) e Conway (2018) per lo stato dell'arte sui *Big Data* e le applicazioni nel settore immobiliare, alle quali si rimanda per ogni approfondimento.

1 Morgan Stanley Capital International.

2. BIG DATA

"*L'universo (che altri chiamano la Biblioteca) si compone di un numero indefinito, e forse infinito, di gallerie esagonali, con vasti pozzi di ventilazione nel mezzo, circondati da ringhiere bassissime. Da qualunque esagono, si vedono i piani inferiori e superiori: interminabilmente. La distribuzione delle gallerie è invariabile. Venti scaffali, cinque lunghi scaffali per lato, coprono tutti i lati tranne due; la loro altezza, che è quella dei piani, supera di poco quella di un bibliotecario normale. Una delle facce libere dà su uno stretto atrio, che sbocca in un'altra galleria... La luce procede da frutti sferici che hanno il nome di lampade. Ve ne sono due per esagono, su una traversa. La luce che emettono è insufficiente, incessante...*" (Borges, 1941).

Così uno dei più grandi visionari del Novecento, lo scrittore argentino Jorge Luis Borges, in uno dei suoi più bei racconti fantastici, apparso nel 1941 nella raccolta "*Il giardino dei sentieri che si biforcano*", sviluppa la metafora dell'universo rappresentato come una biblioteca e teorizza nella Biblioteca di Babele la disponibilità universale delle pubblicazioni, rivelandone al tempo stesso l'irrimediabile utopicità. Avere accesso a tutti i libri che sono stati, sono, saranno e potrebbero essere pubblicati, tutti catalogati e facilmente disponibili in un unico luogo, non può che essere un'illusione. Rileggendo il racconto di Borges nell'era dei *Big data*, la biblioteca di Babele non può non evocare il rincorrere la conoscenza disponendo dell'universo dei dati. La luce insufficiente delle lampade è un'evidente metafora della ragione umana che in modo incessante spinge a cercare risposte a domande alle quali non sempre è possibile rispondere, dati i limiti della ragione stessa.

Nell'ultimo decennio, l'incremento esponenziale del volume dei dati raccolti, e la loro crescente importanza ed utilità nei processi decisionali, ci ha traghettato nell'era dei *Big data*, caratterizzata dall'accesso sempre più diffuso e frequente a *internet*, dall'uso massivo dei dispostivi mobili e dal diffondersi dell'*internet of things* (IoT), cioè dei sensori intelligenti collegati alla rete che generano flussi continui di dati. Siamo in un'epoca, cioè, in cui l'acquisizione e l'uso dei dati è diventato indispensabile a tal punto da poter essere considerato una risorsa economica a tutti gli effetti (AGCOM, 2020). Si calcola che i dati prodotti nel solo 2003, anno chiave nella storia digitale, abbiano superato l'ammontare accumulato fino all'anno precedente e che quella stessa quantità venga attualmente generata in circa due giorni. Si è calcolato che nel 2020 (IDC, 2021) sono stati prodotti, raccolti, archiviati, copiati ed elaborati per diverse finalità più di 64 zettabytes[2] di dati in tutto il pianeta e supereranno i 180 ZB nel 2025[3] (Figura 1). Tuttavia, viene conservata solo una piccola percentuale di questi dati: solo il due percento dei dati prodotti nel 2020 è stato salvato e conservato nel 2021.

2 Un zettabyte è pari a un trilione di gigabyte (1021 byte).

3 Numeri impensabili nel 1986, quando il volume dei dati in circolazione ammontava a solo 281 petabyte (un petabyte è un milionesimo di uno zettabyte). Nel 2000 il volume era quasi dieci volte più grande, raggiungendo 2,2 exabyte (un exabyte equivale a mille petabyte, e a un millesimo di uno zettabyte). Si può quindi stimare che tra il 1986 e il 2010, sulla rete internet sono transitati complessivamente solo circa 5 zettabyte di dati, il volume di dati prodotto nel 2011.

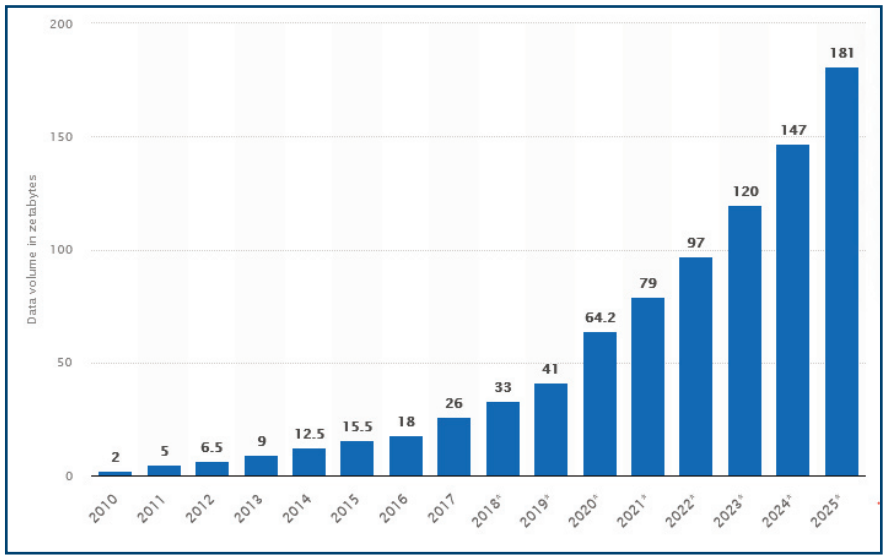

Figura 1 - I flussi di dati prodotti dal 2010, fonte Statista.com[4]

Ci troviamo quindi in un momento storico nel quale l'enorme mole di dati disponibili sovrasta le capacità umane di gestione, elaborazione e analisi, ma nel contempo si delinea, come mai accaduto in passato, una rivoluzione nei processi decisionali delle istituzioni, delle imprese e, sempre più, anche dei singoli cittadini, alla base dei quali non vi è solo esperienza, competenza e intuizione ma soprattutto conoscenza estratta dai *Big data (DDD - Data driven decision making)* (Provost, 2013).

Il termine *Big data*, introdotto come "*the problem of Big data*" nel 1997 da Michael Cox e David Ellsworth, ricercatori della NASA (Cox M., Ellsworth D., 1997), per definire la loro difficoltà nel gestire la crescente mole di dati generati nei loro studi dell'aerodinamica di velivoli, si riferisce quindi a un *set* di dati troppo grande per consentirne un'agevole visualizzazione, memorizzazione e analisi[5].

Alla dimensione si sono poi aggiunte ulteriori complessità, quali la velocità di aggiornamento e la varietà delle fonti che li generano, per giungere alla definizione più comune in base al modello delle cosiddette "3 V", introdotta per la prima volta agli inizi degli anni 2000 (Laney, 2001):

1) *Volume* (*data size*), cioè la dimensione che i dati occupano in termini di *bytes*. Non è mai stata stabilita una soglia perché questa si sposta continuamente in funzione della crescita, come visto esponenziale, dei dati stessi nonché delle capacità di memoria e di calcolo delle macchine.
2) *Variety* (*different forms of data sources*), che può essere intesa come molteplicità di fonti o come eterogeneità

4 https://www.statista.com/statistics/871513/worldwide-data-created/#statisticContainer consultato il 07/10/2021.

5 Va comunque detto che già oltre 60 anni fa, gli statistici si confrontavano con la «maledizione della dimensionalità» (*curse of dimensionality*) riferendosi alle difficoltà che si incontrano quando ci si trova a gestire dati in spazi con tante dimensioni, concetto introdotto nel 1961 dal matematico Richard Bellman. Bellman R. (1961). *Adaptive Control Processes. A Guided Tour*. Princeton University Press, Princeton, NJ. Alla fine degli anni '90 David Donoho contrappone alla maledizione, la benedizione della dimensionalità (*blessing of dimensionality*), tracciando le basi per il futuro del *Data mining*. Donoho DL. (2000), *High-dimensional data analysis: the curses and blessings of dimensionality*. In AMS Math Challenges of the 21st Century, Los Angeles, CA, 6–11 August 2000.

dei formati dei dati. I dati infatti possono essere generati da fonti interne o esterne molto diverse, sostanzialmente classificabili in queste tre tipologie: *human generated*, *machine generated* e *business generated*. Tra le prime si trovano ad esempio le piattaforme di *social network* e di *blogging*, i siti di recensioni, i portali di *e-commerce.* I dati *machine generated* sono invece quelli prodotti da sensori e strumenti scientifici. Infine per dati *business generated* si intendono quelli che vengono acquisiti nei processi di produzione o consumo di beni e servizi come ad esempio le cartelle cliniche, le operazioni bancarie, gli acquisti *online*. Questi dati poi possono assumere vari formati (testo, video, immagini, audio, ecc.). Infine, a seconda che rispettino o meno gli schemi tipici dei *database* si distingue tra dati strutturati (quelli appunto di un *database* che rispetta uno schema predefinito), semi-strutturati (ad esempio gli atti notarili caratterizzati da frasi fisse e frasi variabili) e non strutturati (ad esempio i file di testo generati dalle macchine industriali o i *log di web server* o dei *firewall*).

3) *Velocity* (*speed of change*), che si sostanzia nella rapidità con cui i dati si generano, si raccolgono e si elaborano: i dati confluiscono con frequenza elevatissima e devono essere gestiti e analizzati a volte in *real time* perché non diventino subito obsoleti (si pensi, ad esempio, in ambito sanitario, ai dispositivi medici realizzati per monitorare i pazienti che raccolgono dati per essere analizzati rapidamente).[6]

È importante sottolineare come sia necessario, per impostare una riflessione adeguata sui *Big data*, la contemporaneità dei tre elementi: senza voluminosità (ossia con valori dell'ordine degli zettabyte), dobbiamo parlare più correttamente di *data warehouse*; senza velocità non possiamo avere analisi in tempo reale; senza varietà siamo in presenza di un semplice *database* strutturato.

A queste tre caratteristiche di base se ne sono aggiunte ulteriori due che sono:

- *Veracity* (*uncertainty of data*), che ha a che vedere con quanto può essere accurato o veritiero un *set* di dati, una dimensione, questa, che è particolarmente delicata per l'uso dei *Big data* come fonte statistica ufficiale.
- *Value* (*obtainable information*) è infine la caratteristica che dovrebbe sottendere tutte le altre che riguarda la sfida di estrarre da dati così complessi informazioni che abbiano un significato e siano funzionali e che richiede lo sviluppo di tecnologie sofisticate e di competenze interdisciplinari.

Sono poi state individuate molteplici altre V idonee a caratterizzare i *Big data* (Arockia *et al.*, 2017) tra le quali, degne di nota, *Visualization* (che si riferisce alla complessità di visualizzazione dei dati), *Variability* (che si riferisce all'impatto che deriva dalla mutabilità dei dati), *Virality* (che si riferisce alla velocità di diffusione e di trasmissione dei dati tra gli utenti).

Per comprendere a pieno la portata rivoluzionaria dei *Big data*, bisogna riflettere sul cambiamento di prospettiva che implicano dal punto di vista epistemologico, cioè della creazione di conoscenza.

I *Big data* rappresentano, come si è visto, le tracce digitali che le attività quotidiane di miliardi di persone lasciano con l'utilizzo di sistemi ICT. Quando si parla di *Big data*, infatti, non si ha semplicemente a che fare con dati che aumentano in volume, velocità e varietà. Si tratta di dati con una sfera di influenza sempre più ampia. Questo perché sono accompagnati da un numero crescente di attributi che li descrivono e maggiore è il numero di "descrittori" di

6 Sul sito https://www.internetlivestats.com ci si può rendere conto delle dimensioni coinvolte rispetto a quanto viene pubblicato, caricato, scambiato su Internet in un solo secondo.

un dato, maggiore è il numero di fenomeni che il dato è potenzialmente in grado di spiegare. Ogni attributo aggiunto alla descrizione di un dato diventa, infatti, una nuova dimensione da esplorare per collegamenti tra quel dato e altri dati, tra il fenomeno rappresentato da quel dato e altri fenomeni rappresentati da altri dati.

I *Big data* esprimono un'enorme opportunità anche per la statistica ufficiale, non solo in termini di abbattimento di costi e di riduzione del cosiddetto "fastidio statistico"[7], ma anche e soprattutto ai fini di una produzione statistica più ricca (analisi più dettagliate, scoperta di nuove relazioni) e più tempestiva. Ognuno di noi lascia una traccia in una serie grandissima di registri – l'Anagrafe tributaria, il Catasto, il Servizio sanitario nazionale e regionale, il Sistema scolastico, l'Inps, l'Inail, il Pubblico registro automobilistico, i gestori telefonici e delle reti dati, e così via – per cui basterebbe collegare tutte queste banche dati per avere informazioni accurate, attendibili, dinamiche e aggiornate.

Bruno De Finetti, il grande matematico e statistico, addirittura 60 anni fa, con straordinaria intuizione, aveva formulato lo schema dei collegamenti fra tutti i registri della pubblica amministrazione, in occasione di un seminario all'Istituto di Demografia della Facoltà di Scienze Statistiche dell'Università "la Sapienza" a Roma. Aveva personalmente disegnato e firmato, con gessetti colorati sulla lavagna dell'aula in un paio d'ore il pomeriggio precedente, lo schema in Figura 2.

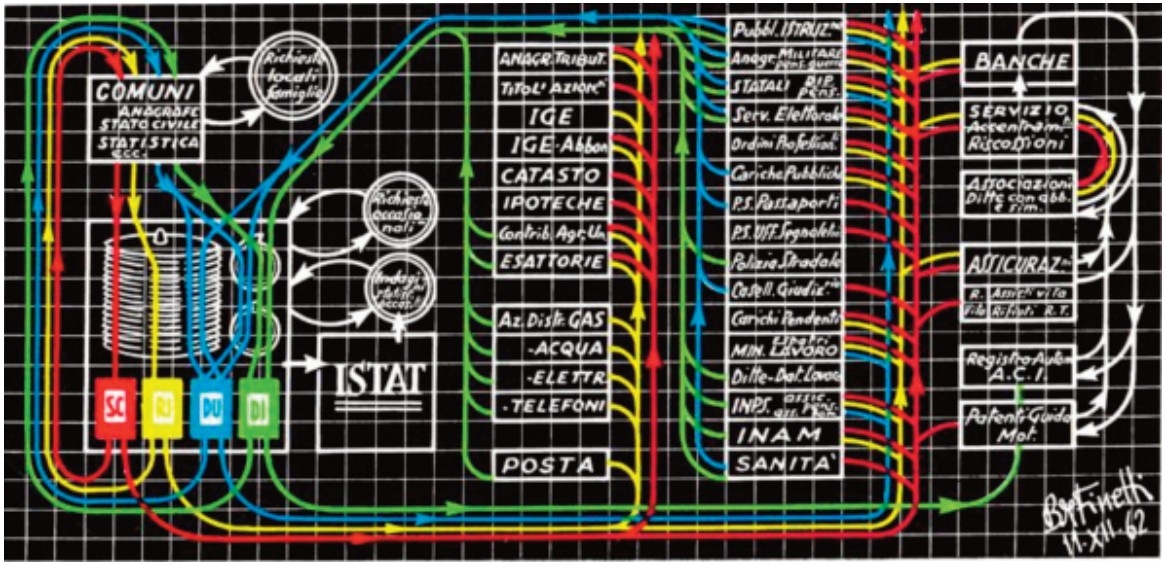

Figura 2 - I flussi di dati tra pubbliche amministrazioni italiane. Da una lezione di Bruno De Finetti del 1962[8]

Al suo schema si potrebbero oggi aggiungere nuovi elementi, come ad esempio i dati di georeferenziazione, così che si possano legare, leggere e interpretare in un continuo divenire gli *stock* e i flussi di popolazione.

7 La disponibilità degli intervistati a collaborare alle rilevazioni statistiche, nonostante gli obblighi giuridici, può infatti incontrare dei limiti che finiscono inevitabilmente per inficiare l'accuratezza stessa dei dati. Il fenomeno del fastidio statistico è molto sensibile alla tipologia (più o meno invasiva) e alla frequenza delle rilevazioni.

8 Tratto da https://centroriformastato.it/iniziativa/presentazione-della-scuola-pop-sui-dati-digitali/.

Non bisogna dimenticare però che questi dati non nascono per essere usati a scopi statistici. Questo vuol dire che, a seconda dei casi, possono presentarsi uno o più di questi problemi: concetti, definizioni e classificazioni non dichiarate o non note, popolazione obiettivo non definita, metadati non disponibili, dati non strutturati, dati non riferibili alle unità della popolazione di interesse, dati che richiedono notevoli *pre*-elaborazioni prima dell'uso statistico, dati che non coprono l'intera popolazione, dati non rappresentativi, variabili di interesse o ausiliarie non direttamente disponibili, dati "sporchi" (Alleva, 2015).

È evidente però che, dopo oltre 20 anni, la definizione di *Big data* non può limitarsi a un *set* di dati così grande da rendere necessarie capacità computazionali di *super* computer. In questa accezione *Big data* rimane un termine generico e qualitativo, che ha trovato un grande impatto mediatico, troppo spesso usato impropriamente.

Con *Big data* va allora identificato l'insieme delle tecnologie e delle metodologie di analisi di una grande massa di dati, estendendolo anche all'insieme degli algoritmi utili a elaborarli e può quindi legarsi alla capacità di acquisire, immagazzinare, gestire, processare e analizzare grandi quantità di dati allo scopo ultimo di estrarne valore conoscitivo: culturale, economico, sociale, tecnologico e scientifico.

Non basta possederli per ottenere conoscenza: solo disponendo di competenze specifiche e tecnologie avanzate è possibile estrarla. Diventano ancora più imprescindibili capacità interpretative di specialisti in grado di fare sintesi nell'enormità dei dati, di selezionare quelli utili, di sviluppare modelli validi, di analizzare propriamente i risultati per aiutare i decisori a comprendere, per effettuare le migliori e più opportune scelte. Quindi, altrettanto fondamentale è lo sviluppo di adeguate competenze professionali quanto lo è la disponibilità dei dati e degli strumenti tecnologici.

3. GLI ESPERTI *BIG DATA*

Nel 2008, DJ Patil, *Chief Scientist* di LinkedIn, e l'allora *Lead of data and analytics efforts* di Facebook, Jeff Hammerbacher, hanno coniato il termine *Data scientist* per indicare una nuova figura professionale in grado di effettuare analisi su grandi moli di dati volte a estrarne informazioni rilevanti che possano essere di valore per l'azienda in cui operano (Leccardi, 2018).

Dino Pedreschi, professore ordinario di Informatica all'Università di Pisa, descrive, nel 2014, il *Data scientist* come: «Una figura che deve avere più competenze. La prima è sapere gestire, acquisire, organizzare ed elaborare dati. La seconda competenza è di tipo statistico, cioè il sapere come e quali dati estrarre, la terza capacità è una forma di *storytelling*, il sapere comunicare a tutti, con diverse forme di rappresentazione, cosa suggeriscono i dati» (Mosca, 2014).

Nel 2018 l'Osservatorio *Big data Analytics & BI* del Politecnico di Milano ha definito il *Data scientist* una figura altamente specializzata che conosce in maniera approfondita le tecniche matematico-statistiche, sa come sviluppare e implementare algoritmi di *Machine learning*, è in grado di programmare nei principali linguaggi (principalmente R o Python) e gestisce gli *Analytics*; sa estrarre dati da *database* NoSQL, usare tabelle pivot in Excel e produrre visualizzazioni chiare e sintetiche per gli utenti *business* (Di Deo, 2019).

Nel *The Future of Jobs Report 2020* del *World Economic Forum* (WEF, 2020) sono stati elencati i ruoli professionali la cui domanda cresce ed è destinata ad aumentare anche nei prossimi anni. Emerge chiaramente che le figure

maggiormente richieste afferiscono i *Big data* (*Data Analyst and Scientists*, *AI e Machine learning Specialist*, *Big data Specialist*), le applicazioni in ambito *Data Science* (come ad esempio *Process Automation Specialists*, *Robotics Specialists*, *Human-Machine Interaction Designers*, *Internet of Things Specialists*) o le tecnologie connesse (come ad esempio *Information Security Analysts*, *Information Technology Services*, *Software and Applications Developer and Analysts*). Ciò deriva dall'elevato interesse delle aziende nell'adottare o di ampliare i propri sistemi di *data analysis*. Nel Report citato viene presentato per la prima volta un modo per misurare e monitorare l'emergere di una serie di nuovi figure professionali utilizzando dati del mercato del lavoro in tempo reale, identificando 99 ruoli la cui domanda è in costante crescita nei diversi settori economici: i primi 3 sono proprio le figure professionali richieste in ambito *Big data* (*Data Analyst and Scientist*, *AI and Machine learning Specialist*, *Big data Specialist*) e tali tre figure compaiono tra i primi dieci ruoli più richiesti in ognuno dei 15 settori economici analizzati.

Al *Data scientist*, quindi, non bastano conoscenze tecniche avanzate, quali quelle di statistica, di matematica, di informatica, ma deve avere anche capacità di sviluppare e implementare algoritmi di *Machine learning*, capacità di analisi, creatività e doti comunicative. E in quasi tutte le principali Università italiane è possibile accedere a specifici percorsi formativi, corsi di laurea magistrale o master post laurea, dedicati alla *Data science*.[9] Servono infine competenze trasversali secondo il settore in cui si lavora, che vanno acquisite con la pratica.

Tutte le grandi imprese stanno investendo cifre rilevanti nelle tecnologie in grado di gestire *Big data* e per questo ricercano giovani *Data scientist* che spesso sono il risultato di un'integrazione tra competenze tecniche nel campo della statistica e l'apprendimento in esperienze lavorative in imprese dotate di strutture di *Analytics*, spesso in sinergia con corsi universitari.

4. *DATA SCIENCE*, *DATA MINING* E *MACHINE LEARNING*

Tra gli impatti più dirompenti dell'avvento dei *Big data* vi è certamente quello di aver scardinato la pratica del campionamento e con essa la ricerca delle relazioni di causalità tra fenomeni a favore della correlazione. Il campionamento statistico, data l'impossibilità dell'osservazione del tutto, è stato fino a qualche anno fa considerato come la metodologia migliore per estrarre informazioni dalla realtà e rappresentarne i fenomeni.

I *Big data*, con la loro ampia estensione, anche se non rappresentano ovviamente l'universo dei dati, contengono molte più informazioni che opportunamente elaborate forniscono notevoli contributi alla conoscenza dei fenomeni studiati. Il passaggio dalla scarsità dei dati nel campionamento all'abbondanza dei *Big data* non significa, tuttavia, ottenere un'informazione migliore.

Se fornire una rappresentazione valida e significativa di una realtà osservata attraverso un campione non è mai stato semplice, anche estrarre conoscenza da una enorme quantità di dati complessi non omogenei non è cosa banale.

9 Solo per citarne alcuni: Corso di Laurea magistrale "*Data Science*" presso la Facoltà di Ingegneria dell'informazione, informatica e statistica dell'Università "La Sapienza" di Roma https://corsidilaurea.uniroma1.it/it/corso/2021/29942/home; Corso di Laurea magistrale "Data science and economics" presso il Dipartimento di Economia, Management e Metodi Quantitativi dell'Università "La Statale" di Milano - https://dse.cdl.unimi.it/it; Corso di Laurea magistrale "Data Science" presso la Facoltà di Fisica dell'Università "Federico II" di Napoli https://www.fisica.unina.it/laurea-magistrale-in-data-science.

Lo sviluppo di modelli relazionali non si è dimostrato più adeguato quando i dati disponibili sono cresciuti esponenzialmente in volume, soprattutto in varietà oltre che in velocità. Un siffatto insieme di dati, non potendo più essere trattato con le classiche relazioni della statistica, ha condotto allo sviluppo di nuove tecniche di elaborazione e analisi più flessibili e scalabili, e, date anche le diverse esigenze di memorizzazione e di utilizzo, di nuovi *database* lontani dal tradizionale approccio relazionale (NoSQL[10]), come ad esempio i *database* a grafo, ampiamente utilizzati nei *social network* dove l'interesse è concentrato sulle relazioni.[11]

Come si è detto, si sono sviluppate e diffuse nuove tecniche, ed evoluzioni di quelle tradizionali, confluendo in quella che è considerata un'evoluzione della statistica esplorativa, il *Data mining*.

Non bastano, quindi, nuovi *database* e nuove tecniche di analisi per estrare conoscenza dai *Big data*, sono necessarie tecniche di selezione delle fonti e dei dati per escludere le informazioni distorsive, tecniche di analisi delle variabili e delle correlazioni e, alla fine, metodologie per interpretare i risultati e riconoscerne la validità. Si tratta di un processo che si può riassumere nel *Knowledge Discovery in Data* (KDD) e che si compone di cinque fasi (Figura 3) che possono condurre dai dati al loro valore conoscitivo:

1) *Selection*: dove sono definite le origini dati.

2) *Data cleaning e Pre-processing*: dove si prova a migliorare il *set* di dati, attenuando il rumore dei dati e individuando gli *outlier*, e dove si prendono decisioni sul trattamento dei dati, dei campi mancanti o incompleti, sull'aggiunta di variabili derivate e valori derivabili dai dati esistenti.

3) *Transformation*: dove si analizzano le variabili, la loro importanza, le loro interazioni e correlazioni.

4) *Datamining*: dove attraverso l'uso di tecniche di analisi evolute si estraggono strutture, correlazioni, modelli e regole all'interno dei dati.

5) *Evaluation/Interpretation*: dove si valuta la validità dei modelli rilevati e se ne fornisce un'interpretazione, con possibile retroazioni verso i punti precedenti al fine di migliorare l'efficacia del processo.

10 *Not only Structured Query Language* è l'acronimo con il quale vengono identificati i database non relazionali. Questi possono essere raggruppati in tre grandi categorie: *key/value store* (nel quale i dati sono archiviati e recuperati utilizzando una chiave che identifica in modo univoco i record); *document store* (memorizzano i dati in documenti e ciascun documento contiene coppie di campi e valori); *wide column store* (memorizzano i dati in tabelle, righe e colonne dinamiche); *graph database* (basati sul modello a grafo, archiviano i dati in nodi e bordi).

11 Esempi di *Database* a grafo sono Neo4j (progettato come modello *open source)*, Amazon Neptun, SAP Hana Graph, OrientDB.

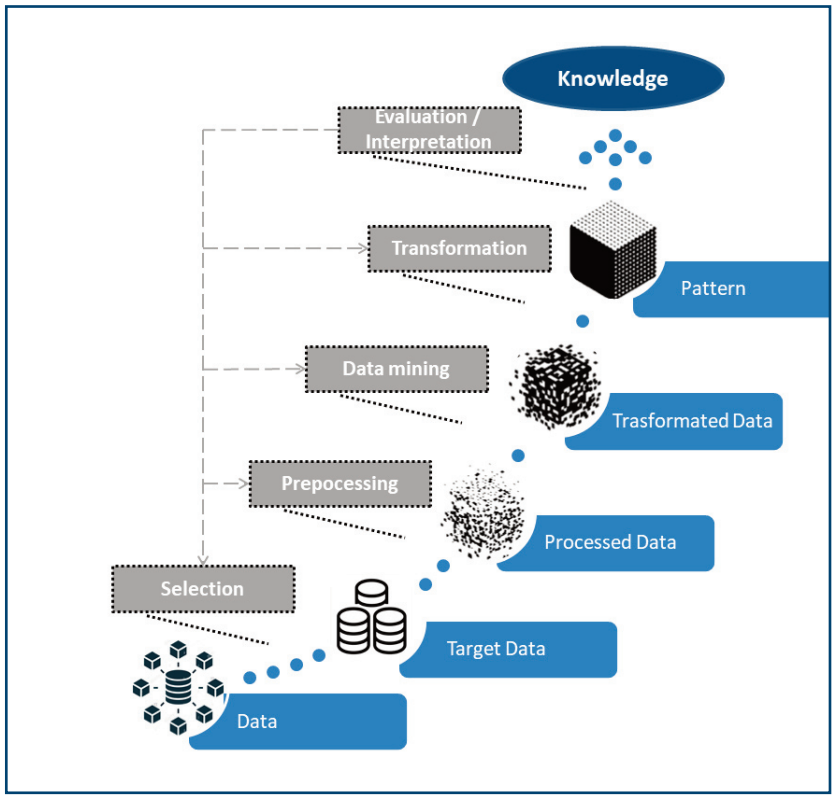

Figura 3 – Il processo di Knowledge discovery in data (KDD)[12]

È quindi evidente che l'approccio ai *Big data* necessita di un passaggio dalla Statistica a qualcosa di più vasto che attiene alla Scienza dei dati (*Data Science*).

L'approccio statistico prevede, ad esempio, una fase di rilevazione che si concretizza nella scelta di un campione in base al fenomeno da analizzare, nella determinazione dei caratteri della popolazione, nella raccolta e classificazione dei dati. La determinazione di un *dataset* nel caso dei *Big data*, come è stato descritto, prevede un'attività molto più complessa e dispendiosa in termini di tempo. In questo caso, i dati sono abbondanti, complessi e variegati, ma possono essere sporchi e ridondanti, e possono diventare velocemente obsoleti.

Un'ulteriore sfida in quest'ambito è inoltre quella di analizzare i *Big data* per ottenere previsioni a brevissimo termine, quasi in tempo reale; si parla infatti oggi di "*nowcasting*", derivato dal concetto di "*forecasting*" (previsione), nato in ambito meteorologico indicando appunto una tecnica per una previsione del tempo a breve termine, usando una stima delle variabili in gioco e assumendo una sostanziale continuità del comportamento osservato grazie proprio al ridotto intervallo di tempo considerato. Passando al campo economico, il "*nowcasting*" trova una sua definizione nella predizione del presente, del futuro immediato e del passato appena trascorso (Bańbura *et al*., 2010).

12 (Hicham *et al*., 2018)

Gli strumenti della *Data Science* fanno comunque parte della scienza statistica, ma devono essere affiancati, come si è già detto, da nuove tecniche di elaborazione e analisi, quali *Machine learning* e *Network Science*. Le modalità di lavoro cambiano nettamente e richiedono nuove professionalità: un team di *Data scientist* che abbia variegate competenze e attitudini.

In linea generale, le tecniche di analisi consistono per lo più in algoritmi tra i quali si distinguono quelli di interrogazione e quelli di apprendimento. Mentre i primi mirano a rispondere a delle richieste precise da parte degli utenti poste in forma di interrogazioni, i secondi invece mirano ad estrarre nuova conoscenza, nuove tesi e si avvalgono di tecniche avanzate di Intelligenza Artificiale (AI), come il *Machine learning* e le *Reti neurali*. Non di rado, poi, le varie metodiche si integrano fra loro.

Il *Machine learning* si colloca nell'ambito dei modelli previsionali. Diversamente dai programmi *software standard*, strettamente codificati attraverso regole severe al fine di eseguire determinati compiti, non adattabili a nuove variabili o richieste a meno che un programmatore non li aggiorni con nuove e specifiche regole, il *Machine learning* utilizza un approccio induttivo per formare una rappresentazione del mondo sulla base dei dati che si esaminano (ITMedia, 2018). È poi in grado di ottimizzare e migliorare la rappresentazione all'aggiungersi di nuovi dati. In questo senso l'algoritmo "impara" dai nuovi dati inseriti e migliora nel tempo. Dunque il requisito fondamentale per l'apprendimento automatico è rappresentato da una grande mole di dati, necessari per guidare l'algoritmo che apprende le relazioni tra le variabili stabilite nel passato per prevedere il futuro. A differenza quindi dei modelli statistici tradizionali, le regole non sono note in anticipo. Lo schema in Figura 4 illustra le differenze nelle metodologie. In un modello tradizionale si utilizzerebbe una funzione che possa descrivere al meglio le relazioni tra più variabili. Osservando e testando queste relazioni si arriva a comprenderne il nesso di causalità approssimato dal modello. In un modello di ML il processo di selezione delle variabili e l'individuazione delle relazioni interconnesse vengono identificate automaticamente in base all'elaborazione progressiva e ottimizzata dei dati fino a individuare un modello predittivo (*predictive analytics*).[13]

13 Analisi predittiva: tecnologia che apprende dall'esperienza (dati) per prevedere il comportamento futuro degli individui al fine di prendere decisioni migliori (Siegel, 2013).

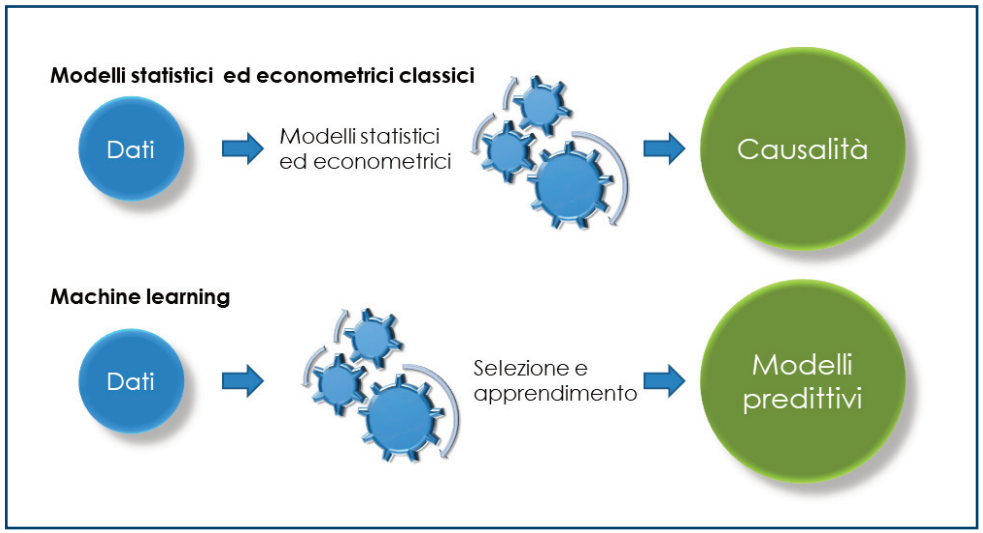

Figura 4 – Modello Statistico/econometrico e Machine learning

Modelli computazionali, ispirati alla struttura e al funzionamento delle reti neurali biologiche (cioè le cellule nervose fittamente interconnesse all'interno di un cervello), sono le Reti neurali, che trovano modelli nei dati. Le reti neurali sono adatte per trovare modelli non lineari. Possono essere utilizzati per il riconoscimento e l'ottimizzazione dei modelli. Alcune applicazioni di rete neurale implicano l'apprendimento supervisionato e altre implicano l'apprendimento non supervisionato. Una rete neurale[14] ad apprendimento supervisionato viene addestrata fornendo una certa quantità di *input* (problemi) ed *output* (soluzioni), in modo che rilevando le associazioni apprenda a produrre autonomamente *output* corretti. Una rete neurale ad apprendimento non supervisionato invece viene addestrata soltanto con *input* costituiti da selezionate tipologie di dati. Esaminando i medesimi, la rete apprende per cogliere analogie e differenze, operando classificazioni. Grazie all'elevata capacità di calcolo parallelo, queste due categorie di reti neurali sono in grado di processare in modo proficuo ed efficiente i *Big data*, effettuando classificazioni, associazioni e *clustering*.

Inoltre tecniche di analisi multivariata, come gli alberi decisionali e i classificatori Bayesiani, consentono di sottoporre i *Big data* a processi di *Machine learning* per costruire modelli predittivi basati sulla classificazione di istanze di grandi quantità di dati.

Un albero di decisione è un grafo in cui, partendo dalla radice (*training set*), si procede ad effettuare una classificazione tramite un percorso che è di volta in volta una scelta fra varie diramazioni, o sottoinsiemi (detti nodi), i cui rami sono le alternative che conducono alle diverse foglie (risultati o classi). Un albero decisionale implementato in modo corretto deve avere dimensioni adeguate, il che significa non eccessive: un numero troppo alto di variabili renderebbe caotico e lento un algoritmo che è invece veloce ed efficiente. Nel *Data mining*, gli alberi di decisione sono utilizzati per operazioni di segmentazione, classificazione, regressione e serie storiche.

14 https://it.wikipedia.org/wiki/Rete_neurale_artificiale

Un *naive bayes* è un particolare classificatore di tipo probabilistico, che determina, cioè, la probabilità di un elemento di appartenere o meno a una certa classe in funzione dei suoi attributi. Peculiarità del *naive bayes* è che ogni caratteristica viene valutata in autonomia: le probabilità di un attributo di risultare attinente a una classe (e quindi il suo contributo alla classificazione dell'elemento) non dipendono dalla relazione con altri attributi. Questa tipologia di classificatori è piuttosto robusta nei confronti del rumore (ossia dati errati, inutili, ecc.) e riesce a essere efficiente anche in presenza di dati incompleti.

5. *BIG DATA* E IMMOBILIARE

Come abbiamo visto i dati sono diventati una componente chiave in tutti i settori dell'economia moderna, anche per il settore immobiliare, che però, essendo un settore tradizionalmente frammentato ed eterogeneo, ha registrato un ritardo nell'innovazione tecnologica soprattutto per la mancanza di *set* di dati strutturati che ha quindi comportato una maggiore dispersione delle informazioni, asimmetrie informative diffuse e conseguente opacità del mercato.
Un elenco, certamente non esaustivo, di *data set* che compongono i *Big data* dell'immobiliare potrebbe essere quello riportato di seguito.

- I dati dello *stock* immobiliare: unità immobiliari urbane, tipologie di fabbricati, caratteristiche, età, classe energetica degli immobili, terreni agricoli, terreni edificabili, unità immobiliari non censite, immobili ipotecati, ecc.
- I dati di georiferimento degli immobili: localizzazione di dettaglio, dati altimetrici, zone urbane omogenee, ecc.
- I dati naturalistici e ambientali: clima, morfologia del territorio, aree protette, qualità dell'aria, qualità delle acque, uso del suolo, ecc.
- I dati delle infrastrutture tecnologiche: acquedotto, fognatura, elettricità, gas, telecomunicazioni, ecc.
- I dati delle infrastrutture di comunicazione: strade, ferrovie, porti e aeroporti, trasporti urbani, ecc.
- I dati dei servizi territoriali: sanità, scuole e università, esercizi commerciali, svago, beni culturali, ecc.
- I dati dell'economia del territorio: industrie, turismo, commercio, logistica, terziario, ecc.
- I dati socio-demografici: popolazione, famiglie, età, occupati, redditi, ecc.
- I dati di sviluppo immobiliare: permessi di costruire, nuove costruzioni, investimenti, ecc.
- I dati delle compravendite immobiliari: prezzi, tipologie e caratteristiche degli immobili compravenduti o offerti in vendita, tipologie e valori di finanziamento per l'acquisto, ecc.
- I dati delle locazioni immobiliari: canoni, tipologie e caratteristiche degli immobili locati o offerti in locazione, ecc.
- I dati finanziari: valori azionari delle società immobiliari, redditività titoli immobiliari, ecc.

Questi sono dati che derivano dalle molteplici interazioni tra i soggetti interconnessi, occasionalmente, periodicamente o continuativamente, nel settore immobiliare: proprietari, costruttori, acquirenti, locatari, intermediari d'affari, intermediari finanziari, agenzie fiscali, periti, enti territoriali, ecc.

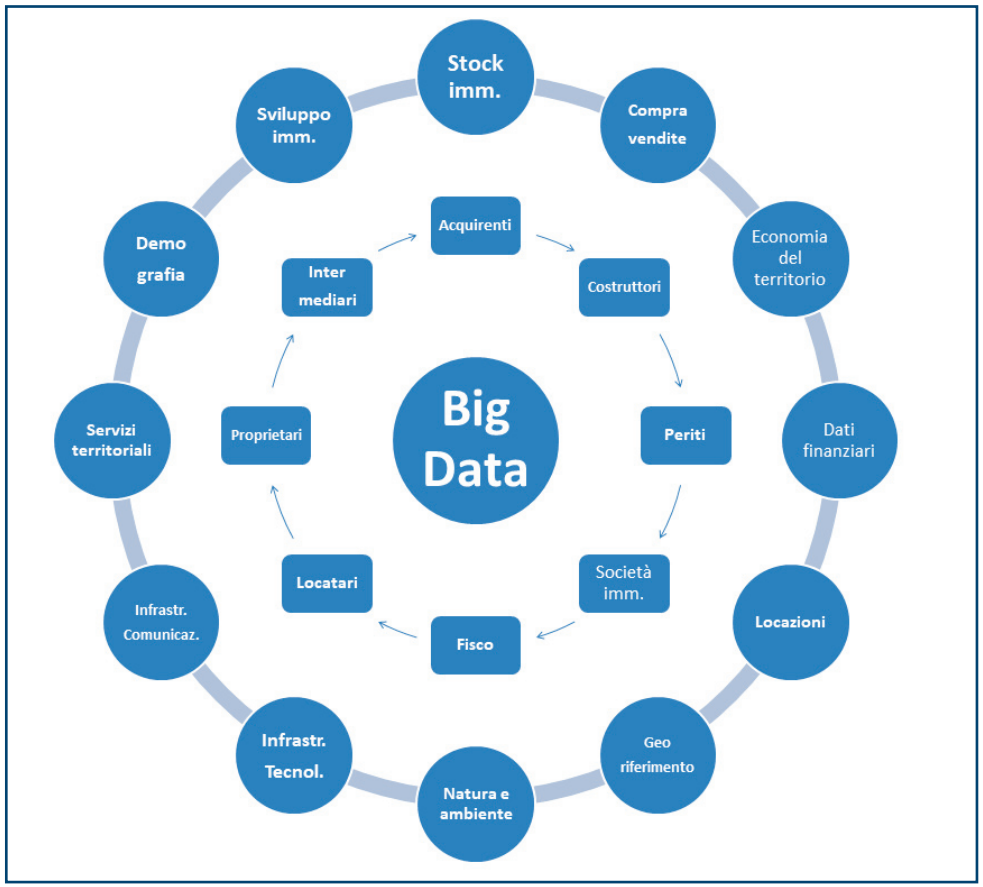

Figura 5 – I *Big data* dell'immobiliare

Si tratta quindi di una gran mole di dati, generalmente non strutturati, residenti in molteplici fonti con elevata disomogeneità, anche nella frequenza di aggiornamento. L'accessibilità a questi dati, però, non è sempre consentita a tutti a causa sia della scarsità di dati pubblici sia della rilevanza di dati proprietari acquisiti nei modelli di *business* di molte aziende. Alcuni dati, inoltre, rimangono alquanto limitati in termini di rilevazione temporale o si riferiscono a un'estensione spaziale ben definita: unità di censimento, area omogenea, distretto.

In Italia non sono note applicazioni di *Data mining* nel settore immobiliare se non ancora al livello di studio e ricerca, sviluppate soprattutto nel campo della valutazione immobiliare.[15]

Diversi esempi di applicazioni nel settore immobiliare, sviluppati su modelli di *Data mining*, sono invece diffusi negli USA e in particolare nel campo dell'intermediazione, della valutazione, della gestione e dello sviluppo immobiliare (Prospetto 1).

15 Si possono citare ad esempio Loberto *et al*. (2018), De Nadai *et al*. (2018), Bergadano *et al*. (2021).

AMBITI DI APPLICAZIONE ML NEL SETTORE IMMOBILIARE	PIATTAFORME
Marketing immobiliare	Zillow, Rex, Redfin, Trulia, Compass, Localize
Analisi predittiva del mercato immobiliare	Skyline AI, Jointer, Citybldr
Valutazione immobiliare	Vero, HouseCanary, GeoPhy, Quantarium
Monitoraggio sistemi edilizi	Enertiv
Sviluppo immobiliare	Deepblocks

Prospetto 1 – Ambiti di applicazione di modelli ML nel settore immobiliare

Tra le applicazioni nel campo dell'intermediazione immobiliare sono diffuse le piattaforme (come ad esempio *Zillow*, *Rex*, *Redfin*, *Trulia* o *Compass*) fondate su modelli di *Machine learning* che mirano a individuare, attraverso l'analisi del comportamento dei potenziali acquirenti, l'immobile desiderato proponendo un *set* di unità che presentano le caratteristiche più ricercate e, se opportunamente profilati, adatte al loro stile di vita. In altri termini, la piattaforma impara i "gusti" dell'utente e i termini di ricerca preferiti, cogliendoli in modo più granulare quanto più l'utente naviga nel sito, proponendogli immobili con le caratteristiche più appropriate alle sue esigenze.

Altre piattaforme, come *Localize*, sfruttano l'intelligenza artificiale anche per verificare gli annunci di vendita o locazione degli immobili, costruendo un *set* informativo spinto sui dettagli più inscrutabili degli immobili (variazione di luminosità nell'arco del giorno, vibrazioni conseguenti al traffico stradale, rumorosità degli affacci, efficienza dei servizi di prossimità, livello di criminalità, ecc.) sulla base di tutti i dati resi disponibili negli annunci, fino a condurre l'utente all'immobile che corrisponde alle sue reali esigenze.

Altre ancora, che utilizzano i *Big data* e tecniche di *Machine learning* per conoscere tempestivamente la dinamica del mercato immobiliare, sono pensate per gli investitori ai quali saranno fornite le informazioni sul momento migliore per comprare o per vendere, come promette *Skyline AI*, o sul migliore investimento immobiliare, come promette *Jointer* o *Citybldr*, valutando, attraverso le potenti basi dati a disposizione, i rischi e le opportunità.

Nell'ambito della valutazione immobiliare sono in rapida crescita modelli di valutazione automatizzati (AVM - *Automated Valuation Models*) che promettono un'analoga accuratezza delle valutazioni tradizionali oltre alla produzione di valori in tempo reale a costi molto contenuti, considerando un numero elevato di fattori non sempre acquisiti dai periti tradizionali, quali la dinamica economica e demografica di dettaglio, l'evoluzione della domanda e dell'offerta, l'accessibilità e il livello di utilizzo dei servizi e delle infrastrutture, ecc. Stime accurate e tempestive dei valori immobiliari sono fondamentali, ad esempio, per identificare opportunità di investimento, eseguire valutazioni del rischio o eseguire una gestione avanzata del portafoglio.

Piattaforme di *Data mining* (come *Vero*, *HouseCanary*, *GeoPhy* o *Quantarium*) promettono agli investitori immobiliari e agli istituti finanziatori valori accurati, sia per gli *stock* sia per i flussi, per minimizzare i rischi, laddove errori sistematici o distorsioni nelle valutazioni possono avere effetti negativi sui capitali impiegati (investimento o debito). Queste piattaforme AVM possono rivelarsi utili, inoltre, per eseguire *stress test* di portafogli immobiliari in scenari economici avversi.

Nel settore immobiliare è in crescente diffusione l'*Internet of Things* (IoT) o l'*internet* degli oggetti. Sono sempre di più, cioè, gli oggetti che ci circondano all'interno nelle nostre case, al lavoro, nelle città, nella vita di tutti i giorni, dotati di sensori connessi alla rete, che lasciano tracce digitali del vivere quotidiano. L'ambizione del *Data mining* è, in questo caso, quello di riuscire, attraverso i dati raccolti, a misurare l'efficienza, la qualità e i costi di gestione degli immobili nei quali tali sensori registrano in tempo reale il consumo di energia, il grado di umidità, lo stato d'uso e tanti altri fattori. Il monitoraggio in tempo reale dei sistemi dell'edificio può essere di grande utilità per fornire una manutenzione predittiva, piuttosto che preventiva o programmata. Aziende come *Enertiv* offrono soluzioni attraverso una piattaforma che acquisisce i dati più granulari possibili sulle prestazioni degli edifici commerciali e quindi esegue analisi per garantire protocolli di manutenzione appropriati.

Infine, modelli di *Data mining* sono implementati per supportare le decisioni su progetti di sviluppo immobiliare. Piattaforme di servizi immobiliari (come *Deepblocks*) utilizzano gli algoritmi di *Machine learning* per monitorare le tendenze demografiche, sociali ed economiche e del mercato, gli investimenti pubblici nelle infrastrutture e le normative urbanistiche, al fine di individuare i siti di costruzione ottimali in una certa zona della città supportando gli sviluppatori a pianificare le tempistiche di costruzione ideali.

6. *BIG DATA* LIMITI E CRITICITÀ

Le attese nei confronti dei *Big data*, come si è visto, sono elevate e riguardano una conoscenza così spinta nei dettagli come mai era stata in precedenza. All'elaborazione e all'analisi di informazioni granulari raccolte in tempo reale si associa non solo la capacità di conoscere fenomeni sociali, economici e demografici, ma anche capacità predittive e in diversi scenari futuri. Di certo i *Big data* rappresentano un nuovo e potente strumento di conoscenza, ma non al riparo da limiti e criticità.

Un primo aspetto critico nell'utilizzo dei *Big data* riguarda l'accesso alle enormi quantità di informazioni prodotte e il rischio che si creino nuove disuguaglianze. Il patrimonio informativo dei *Big data* è disponibile solo in capo ad un numero limitato di soggetti in virtù del ruolo occupato nella generazione dei flussi informativi. Istituzioni pubbliche o aziende private raccolgono elevate quantità di dati ai quali l'accesso è prevalentemente riservato a pochi. L'accesso ai dati grezzi è alcune volte inibito o limitato solamente ad alcune parti o ad alcuni soggetti; a volte viene concesso il privilegio di accesso a pagamento e alcuni offrono piccoli *set* di dati solo a scopo di studio e ricerca. Risulta evidente che solo chi può pagare o lavora all'interno dell'azienda o dell'istituzione che detiene i dati può accedervi per produrre ricerca; chi non ha accesso ai dati oltre ad avere inibita la ricerca non può valutare le affermazioni metodologiche di chi ha un accesso privilegiato. La Pubblica amministrazione è impegnata nel rendere accessibili liberamente *data-set* delle informazioni raccolte, sia in forma grezza che elaborata (*open data*).

Un secondo aspetto riguarda il diritto di ognuno alla protezione dei propri dati personali e della propria sfera privata (*privacy*). In questo nuovo scenario si tende, come visto, a massimizzare la quantità di dati disponibili, a integrare e collegare informazioni che provengono da fonti differenti, nonché a utilizzare e riutilizzare lo stesso insieme di dati per le più diverse finalità di analisi, tutti fenomeni che pongono notevoli sfide a chi si occupa di tutela della *privacy.*

L'anonimizzazione viene vista come la principale strategia per rispondere a queste sfide in quanto strumento con cui si cerca di preservare i vantaggi e attenuare i rischi nell'uso dei dati (D'Acquisto, 2017). Essa, tuttavia, non è esente da problemi, sotto entrambi i profili.[16] Dagli studi di casi e dalle pubblicazioni di ricerca emerge infatti con chiarezza che non è così semplice creare un insieme di dati effettivamente anonimo a partire da un ampio insieme di dati personali e, al tempo stesso, mantenere le informazioni sottostanti necessarie agli scopi per cui sono state raccolte.[17]

Rendere anonimi i dati significa privarli di quegli elementi sufficienti per impedire l'identificazione della persona interessata. Più precisamente, i dati devono essere trattati in maniera tale da non rendere più possibile identificare direttamente o indirettamente una persona fisica ricorrendo a un insieme di mezzi che sono ragionevolmente utilizzabili. Sotto l'altro profilo, il processo di anonimizzazione può ovviamente inficiare l'utilità del dato, rendendolo oltre che incerto (ossia non riferibile ad alcuno), inaccurato e inidoneo a qualsiasi tipo di analisi.

Nel 2016, in tema di *privacy* è intervenuto il nuovo regolamento europeo[18] che ha sancito, oltre quelli già definiti[19] dalle precedenti direttive, ulteriori importanti principi: il diritto alla portabilità dei dati,[20] il diritto all'oblio,[21] e l'obbligo di valutazione d'impatto per il titolare del trattamento.[22]

16 Inoltre anche i dati più anonimi possono comunque essere utilizzati per arricchire i profili esistenti delle persone, determinando quindi nuovi problemi di protezione dei dati. Qui se si entra però nel diverso, anche se collegato, campo della profilazione e delle politiche discriminatorie cui può dare adito con tutte le questioni etiche che queste sollevano.

17 Parere 05/2014 sulle tecniche di anonimizzazione, redatto dal Gruppo di lavoro istituito in base all'art. 29 della direttiva 95/46. Si tratta di un organismo consultivo e indipendente, composto da un rappresentante delle autorità di protezione dei dati personali designate da ciascuno Stato membro, dal GEPD (Garante europeo della protezione dei dati), nonché da un rappresentante della Commissione europea.

18 Il Regolamento europeo in materia di protezione dei dati personali (UE 2016/679) è entrato in vigore il 24 maggio dello stesso anno ed è operativo in tutti i Paesi Ue a partire dal 25 maggio 2018.

19 Vale a dire, solo per citare i principali, l'ampio concetto di dati personali, gli obblighi di correttezza (uso di dati esatti e non all'oscuro dell'interessato), finalità (solo per scopi dichiarati esplicitamente e ben definiti) e proporzionalità (uso dei soli dati strettamente necessari e limitatamente al tempo che serve), cui deve attenersi il titolare del trattamento. Altro principio chiave è che l'uso dei dati è ammesso solo 1) con il consenso degli interessati, che deve essere libero, informato, specifico e inequivoco o 2) se ciò è necessario per l'esecuzione di un contratto o 3) per l'esercizio di un interesse legittimo da parte del titolare o di terzi purché ciò non prevalga sui diritti fondamentali della persona interessata – Regolamento UE 2016/679, art. 5 e art.6.

20 Si tratta in sostanza del diritto dell'interessato di trasmettere tali dati a un altro titolare del trattamento senza impedimenti da parte del titolare del trattamento cui li ha forniti – Regolamento UE 2016/679, art. 20.

21 Questo diritto, che non ha carattere assoluto in quanto dev'essere inevitabilmente contemperato con altri interessi (primo fra tutti il diritto di cronaca), può essere definito come l'interesse di un singolo ad essere dimenticato - Regolamento UE 2016/679, art. 17.

22 La valutazione soggetta a riesame periodico contiene almeno una descrizione sistematica dei trattamenti previsti e delle finalità del trattamento, dell'interesse legittimo perseguito dal titolare del trattamento, la valutazione della necessità e proporzionalità dei trattamenti in relazione alle finalità, la valutazione dei rischi per i diritti e le libertà degli interessati e le misure previste per affrontare i rischi, includendo le garanzie, le misure di sicurezza e i meccanismi per garantire la protezione dei dati - Regolamento UE 2016/679, art. 35.

Il nuovo regolamento è stato impostato secondo la filosofia della *privacy by design* che significa, in sostanza, tutela fin dalla progettazione e per l'intero ciclo di vita dei *Big data*: già nell'ideazione del trattamento occorre di *default* integrare misure tecniche e organizzative per assicurare adeguata protezione ai dati personali (Cavoukian, 2011).

In questa nuova prospettiva, dunque, grande enfasi viene posta sulla prevenzione anche perché la tutela *ex post* della *privacy* risulta estremamente complessa quando non del tutto vana. Ci sono in primo luogo problemi di accesso: non tutte le persone che hanno subito un *vulnus* sono infatti sufficientemente motivate o dispongono delle necessarie risorse e competenze per individuarlo e possibilmente neutralizzarlo. Poi possono presentarsi problemi di giurisdizione, in quanto il contesto tecnologico tende ad annullare le distanze e con esse le tradizionali categorie di legge applicabili. Infine, ci si scontra con problemi legati all'efficacia delle misure in quanto esistono tipi di violazioni difficilmente ripristinabili: non si può infatti ordinare a qualcuno di dimenticare una informazione che sarebbe dovuta rimanere riservata.

Un altro aspetto che merita un punto di vista critico riguarda il nuovo modo di estrarre conoscenza dai dati attraverso i processi di *Data mining*. È sempre più frequente assegnare all'analisi di *Big data* la predominanza su altre forme di analisi riconducibili ai processi classici della statistica economica. Jim Gray, vincitore del premio *Turing* assegnato ai grandi matematici, per primo teorizza un "quarto paradigma" associato ai *Big data* che di fatto sancisce l'inizio di una nuova era epistemologica.[23] Oggi ci è consentito navigare nel mare sconfinato dei dati per produrre nuova conoscenza, attraverso algoritmi capaci di ricercare quello che le teorie scientifiche attuali non vedono e non prevedono.

Si tratta cioè di una nuova forma di empirismo che non prevede la formulazione di ipotesi, la costruzione di modelli, la verifica sperimentale su dati e simulazioni, la formulazione di una teoria. Ai *Big data* a disposizione, l'intelligenza artificiale applicherà una serie di algoritmi che troveranno il modello migliore per descrivere un fenomeno. Algoritmi e modelli che in alcuni sistemi AI di *Deep learning*, presentano meccanismi di funzionamento, i cosiddetti *black box*, imperscrutabili anche agli occhi degli stessi programmatori e sviluppatori.

La rappresentatività dei *Big data*, inoltre, sembra scontata dall'enorme quantità di informazioni di cui sono composti. I *Big data* invece spesso non sono esaustivi, come invece si ritiene nell'immaginario collettivo, e il rischio più frequente è quello di esplorare solo alcuni aspetti della realtà e in modo superficiale.

23 «In origine, c'era solo la scienza empirica, e poi c'era la scienza teorica, con le leggi di Keplero, le leggi del moto di Newton, le equazioni di Maxwell e così via. Poi, per molti problemi, i modelli teorici sono diventati troppo complicati per essere risolti analiticamente e si è dovuto iniziare a simulare. Queste simulazioni hanno accompagnato gran parte dell'ultima metà dell'ultimo millennio. A questo punto, queste simulazioni stanno generando moltissimi dati, insieme a un enorme aumento dei dati provenienti dalle scienze sperimentali. Ora in realtà non si guarda attraverso i telescopi. Invece, si sta "guardando" attraverso strumenti complessi e su larga scala che trasmettono i dati ai *data center*, e solo allora si guardano le informazioni sui computer. Il mondo della scienza è cambiato, e su questo non c'è dubbio. Il nuovo modello prevede che i dati vengano acquisiti da strumenti o generati da simulazioni prima di essere elaborati dal *software* e che le informazioni o le conoscenze risultanti vengano archiviate nei computer. Gli scienziati possono esaminare i loro dati solo abbastanza tardi in questo flusso. Le tecniche e le tecnologie per tale scienza ad alta intensità di dati sono così diverse che vale la pena distinguere la scienza ad alta intensità di dati dalla scienza computazionale come un nuovo, quarto paradigma per l'esplorazione scientifica». (Hey, 2009)

La presenza di errori nei dati, derivanti ad esempio da incompletezza, inattendibilità, infondatezza o ripetitività delle informazioni acquisite, potrebbero condurre a vedere connessioni laddove in realtà non esistono, semplicemente perché l'elevata mole di dati processati possono mostrare correlazioni forti ma spurie, non necessariamente, cioè, associate a un rapporto causa-effetto.[24]

Numerosi sono gli esempi in letteratura di correlazioni spurie,[25] molte di queste fondate su una fiducia incondizionata che disporre di un'elevata numerosità di osservazioni possa compensare ogni difetto di misurazione, di attendibilità e di validità (Fan *et al.,* 2014).

7. CONCLUSIONI

Sono numerosi gli ambiti del settore immobiliare nei quali modelli *data driven* potrebbero fornire la conoscenza necessaria su cui fondare decisioni politiche, economiche e sociali.

La pianificazione urbanistica di un territorio, la programmazione di interventi infrastrutturali per l'attrazione di investimenti nel commercio, nell'industria o nel turismo, misurare il valore sociale ed economico generato da progetti di *social housing*, comprendere la dinamica dei prezzi delle abitazioni di un territorio, pianificare il contenimento di consumo di suolo: sono tutti ambiti nei quali i *Big data* dell'immobiliare potranno di certo espandere le frontiere della conoscenza.

Il processo per l'estrazione del valore dai *Big data* si sta consolidando e team di *Data scientist*, consapevoli della loro complessità, possono sviluppare modelli attraverso i quali abilitare strumenti decisionali basati sull'osservazione effettiva della realtà. L'attenzione, in qualche modo, si è spostata così dalla disponibilità dei dati e dagli strumenti tecnologici necessari alla loro gestione ed elaborazione alle competenze professionali necessarie e all'approccio per estrarne valore conoscitivo.

La proclamazione della "fine della teoria" con l'avvento dei *Big data* trova la sua formulazione più diffusa nell'articolo apparso sulla rivista Wired del 2008, nel quale il guru dell'informatica Chris Anderson annunciava che "[...] *ormai la grande quantità di dati a disposizione rende il metodo scientifico obsoleto... i petabyte ci consentono di dire «la correlazione è sufficiente, possiamo smettere di cercare modelli»*. (Anderson, 2008)

Nell'applicare modelli di *Data mining* solitamente non si mira alla conoscenza del nesso di causalità, perché spesso non interessa necessariamente. L'obiettivo è più prevedere che comprendere i fenomeni e il modo in cui funzionano, perché la causalità è sfuggente: la previsione prevale sulla spiegazione (Siegel, 2013).

24 Esiste inoltre, anche nel *data mining*, il rischio di *overfitting* (o sovradattamento), cioè quando l'algoritmo si adatta troppo bene ai dati di *training* perdendo in generalità. Accade cioè che il modello addestrato sui dati sembra perfetto, tuttavia, quando si applica ai dati di test si registrano molti errori.

25 In statistica, una correlazione è spuria quando due fenomeni risultano statisticamente correlati tra loro ma non vuol dire necessariamente che tra di essi sussista un legame diretto di causa-effetto. Tale correlazione potrebbe essere del tutto casuale ovvero dipendente da una terza variabile in comune (Udny Yule, 1926).

È vero che i processi di *Data mining* mirano a rivelare schemi nascosti trasformando l'analisi tradizionale dei dati basata su modelli a un nuovo paradigma scientifico guidato dai dati. L'interpretazione, tuttavia, rimane al centro dell'analisi dei dati e senza comprendere limiti, errori e distorsioni delle informazioni a disposizione, indipendentemente dalle dimensioni del *set* di dati, eleva il rischio di ottenere risultati errati.

Per giungere ad affermazioni rappresentative di un fenomeno osservato, non può bastare conoscere la provenienza del *set* dei dati processato. È necessario aver compreso, con un approccio alla raccolta e all'analisi dei dati sistematico e scrupoloso, punti di forza e punti deboli di tali dati al fine di poter spiegare ogni possibile distorsione al momento dell'interpretazione dei risultati ottenuti, non inevitabilmente esaustivi. *Big data*, quindi, non può voler dire che i metodi classici alla ricerca di teorie e modelli abbiano perso la loro rilevanza; le nuove tecniche dovrebbero rappresentare un'integrazione o un miglior orientamento dei modelli classici senza sostituirsi a essi.

Dotarsi di una struttura basata sui *Big data* e di un team di *Data scientist* consentirà di ottenere una visione approfondita dei fenomeni sociali ed economici e delle loro dinamiche, delle tendenze di settore in diversi scenari e di orientare le decisioni per migliorare le proprie attività di *business*.

Sebbene però l'uso dei *Big data* sia importante, quando non fondato su rigorose e accurate valutazioni, dovrebbe essere affrontato con opportuno scetticismo. C'è sempre il rischio che un'enorme quantità di dati conduca a un'overdose informativa che possa diminuire la conoscenza o fuorviarla. Produrre conoscenza, come si è cercato di argomentare nell'articolo, necessita di chiavi interpretative, assicurate da adeguati contesti scientifici che tengano conto delle complessità e che possano consegnare risultati affidabili e attendibili.

I problemi aperti però sono ancora numerosi e complessi. Assicurare il giusto equilibrio tra diritto alla trasparenza e *privacy*, spingere sull'espansione degli *open data*, migliorare le tecniche di selezione dei dati, eliminare l'opacità dei modelli *black box*[26] di AI, certificare l'affidabilità dei modelli di *Machine learning*, gestire il ciclo di vita dei dati, sono solo alcuni dei temi che bisognerà affrontare per costruire processi di *Data mining* efficienti.

Difficile affermare se i *Big data* possono ambire a imporsi come il "quarto paradigma" per l'esplorazione scientifica, ma innegabilmente rappresentano un valido e potente strumento di ricerca, se non si pretendesse di fare a meno delle teorie e delle verifiche sperimentali. In un modo o nell'altro, bisognerà compiere ancora nuovi passi in avanti per garantire che le tecnologie e le tecniche di elaborazione e di analisi siano utilizzate in modo efficace e responsabile. Altrimenti si continuerà a rischiare che qualcuno utilizzi i dati «come un ubriaco i lampioni, per sostegno piuttosto che per illuminazione», come disse efficacemente Andrew Lang, scrittore e poeta scozzese, noto per i suoi libri di favole.

26 Il Parlamento europeo, nella risoluzione del 20 gennaio 2021 sull'intelligenza artificiale, pone, tra l'altro, in evidenza i chiari rischi correlati alle decisioni prese dagli esseri umani qualora si basino unicamente su dati, profili e raccomandazioni generati dalle machine e chiede che sia introdotto un obbligo in materia di trasparenza e spiegabilità delle applicazioni di I.A. Il richiamo ai sistemi di Deep Learning - e quindi della Black box che li connota - è espressamente contenuto nella proposta di Regolamento della Commissione dell'U.E. del 21 aprile 2021 secondo la quale gli *standard* di progettazione dei sistemi di I.A. debbano tener conto del principio della trasparenza "rafforzata" che la giurisprudenza amministrativa ha indicato come garanzia minima di legittimità per l'utilizzo di algoritmi nell'attività amministrativa decisionale, anche se privi di "autonomia".

BIBLIOGRAFIA

AGCOM, AGCM, Garante protezione dati personali, (2020). *Indagine conoscitiva sui Big data*

Allan, S., Redden, J. (2017). *Making citizen science newsworthy in the era of Big data*, Journal of Science Communication 16 (02). C05.

Alleva G. (2015). *Il valore dei dati nell'era dei Big Data,* disponibile su https://www.istat.it/it/files/2015/10/Napoli_2mar_v3SN_2017.pdf (consultato il 07/10/2021)

Anderson C. (2008). *The end of theory: the data deluge makes the scientific method obsolete*, Wired disponibile su https://www.wired.com/2008/06/pb-theory/ (consultato il 07/10/2021)

Anenberg E., Laufer S., (2014). *A More Timely House Price Index*, Federal Reserve Board

Arockia Panimalar.S., Varnekha Shree.S., Veneshia Kathrine.A., (2017). *The 17 V's Of Big data*, International Research Journal of Engineering and Technology (IRJET). Volume: 04 Issue: 09 | Sep-2017

Barker, A., Stuart Ward, J. (2013). *Undefined By Data: A Survey of Big data Definitions*

Baker, R. P. (2017). *Big data*: A Survey Research Perspective

Baldominos A., (2018). *Identifying Real Estate Opportunities Using Machine learning*, Applied Science, 2018, 8, 2321

Bańbura M., Giannone D., Reichlin L. (2010). Nowcasting, European Central Bank (ECB) Working Paper Series

Bassa, A., (2018). *Managing a Data Science Team*. *Harvard Business Review*. Disponibile su https://hbr.org/2018/10/managing-a-data-science-team (consultato il 07/10/2021)

Bergadano, F., Bertilone, R., Paolotti, D., Ruffo, G. (2021). *Developing Real Estate Automated Valuation Models by Learning from Heterogeneous Data Sources*. International Journal of Real Estate Studies, 15(1), 72–85.

Bollier D., Rapporteur (2010). The promise and the peril of *Big data*. Eighteenth annual Aspen Institute roundtable on Information technology.

Borges, J. L. (1941). *El Jardin de senderos que se bifurcan*, Traduzione italiana di A. Melis (Finzioni, Adelphi, 2003).

Boyd, D., Crawford, K. (2012). *Critical questions for Big data*. Information, Communication & Society

Beltratti A., Bezzecchi A. (2019). *Dati, modelli e real estate: nuove frontiere per l'immobiliare e le banche*, Bancaria 9-2019

Bussoletti F. (2019). Le criticità dei Big data: Intelligenza umana vs. Intelligenza Artificiale (AI). disponibile su https://www.difesaesicurezza.com/difesa-e-sicurezza/le-criticita-dei-big-data-intelligenza-umana-vs-intelligenza-artificiale-ai/ (consultato il 07/10/2021)

Caligiuri, M. (2016). *Cyber intelligence, la sfida dei Data scientist*. Disponibile su https://www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2016/06/cyber-intelligence-sfida-data-scientist-Caligiuri.pdf (consultato il 07/10/2021)

Castelvecchi D. (2016). *Can we open the black box of AI?*, Nature, vol. 538, 2016 disponibile su https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731 (consultato il 07/10/2021)

Cavoukian A. (2011). *Privacy by Design in Law, Policy and Practice: A White Paper for Regulators, Decision-makers and Policy-makers*

Cedrini M. (2019). *Big data e economia: una riflessione epistemologica*, Menabò di etica ed economia, disponibile su https://www.eticaeconomia.it/big-data-e-economia-una-riflessione-epistemologica/ (consultato il 07/10/2021)

Conway J., (2018). *Artificial intelligence and machine learning: current application in Real estate*, Massachusetts Institute of Technology, September 2018

Cox M., Ellsworth D., (1997). *Application-Controlled Demand Paging for Out-of-Core Visualization*, Report NAS-97-010

D'Acquisto G., Naldi (2017) Big Data e Privacy by Design

De Finetti B. (1952). *Macchine «che pensano» (e che fanno pensare)*. Università degli studi di Trieste

De Nadai M. e Lepri B. (2018). *The economic value of neighborhoods: Predicting real estate prices from the urban environment*. In Proceedings of 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Torino, IEEE, 2018.

Di Deo I. (2019). *Data Science: i must have per la gestione dei Big Data* disponibile su https://blog.osservatori.net/it_it/data-science-must-have (consultato il 07/10/2021)

Di Meglio A., Ferrari A. (2021). *Big Data*, Treccani

Fan et al., (2014) *Challenges of Big data Analysis*, National Science Review

Festa M., Serafini S. (2017). *La trasparenza del mercato immobiliare nell'era dei Big data*, Quaderni dell'Osservatorio, OMI – Agenzia delle entrate

Friedman U. (2012). *Big data: A Short History*, disponibile su https://foreignpolicy.com/2012/10/08/big-data-a-short-history/ (consultato il 07/10/2021)

Geideman K., Krause A. (2016). *Transformations in real estate research: the Big data revolution,* 22nd Annual Pacific-Rim Real estate society Conference Sunshine coast, Queensland, Australia 17-20 January 2016

Gunther McGrath, R., McManus, R. (2020). *Discovery-Driven Digital Transformation*, Harvard Business Review, May–June 2020 Issue.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). *The rise of "big data" on cloud computing: Review and open research issues*. Information systems, 2015, 47, 98-115

Hey T. Tansley S., Tolle K. (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery*, Microsoft research

Hey T., Trefethen a. (2003). *The data deluge: an e-science perspective*, Grid Computing – Making the Global Infrastructure a Reality

Hosni H., Vulpiani A. (2017). Forecasting in light of big data, Philosophy and Technology

Hicham M. S., Bouchra F., Brahim O. (2018). A*ssessing reliability of big data knowledge discovery process* - Second International Conference on Intelligent Computing in Data Sciences (ICDS 2018)

IDC (2021). *Annual DataSphere and StorageSphere forecasts*. Disponibile su https://www.idc.com/getdoc.jsp?containerId=prUS47560321, (consultato il 07/10/2021)

Kim J., Seo D., Chung Y. (2020) An integrated methodological analysis for the highest best use of big data-based real estate development, Sustainability 2020, 12, 1144

Kitchin R. (2013). *Big data and human geography: Opportunities, challenges and risks*. Dialogues in Human Geography, 3(3), 262–267

Kitchin R. (2014). *Big data, new epistemologies and paradigm shifts*. Big Data & Society, 1-12.

Kitchin R., McArdle G. (2016). *What makes Big data, Big data? Exploring the ontological characteristics of 26 datasets*. *Big data* & Society, January–June 2016.

Kok N., Koponen E., Martinez-Barbosa C. (2017). *Big data in Real Estate? From Manual Appraisal to Automated Valuation*, The Journal of portfolio management - Special Real estate issue 2017

IT Media Consulting con il contributo scientifico dell'Università Bocconi (2018), *L'economia dei dati* su http://www.itmedia-consulting.com/DOCUMENTI/economiadeidati.pdf (consultato il 07/10/2021)

Laney D. (2001). *3D data management: controlling data volume, velocity, and variety*, META Group, Tech. Rep. 2001

Leccardi F. (2018). *Big Data Scientist: storia di una professione* su https://blog.osservatori.net/it_it/big-data-scientist-storia (consultato il 07/10/2021)

Loberto M., Luciani A., Pangallo M. (2018), *The potential of big housing data: An application to the italian real-estate market*. Banca d'Italia, Temi di Discussione (Working Paper) No. 1171

Lombi L. (2016). *La ricerca sociale al tempo dei Big data: sfide e prospettive*

Lo Sapio G. (2021). *La black box: l'esplicabilità delle scelte algoritmiche quale garanzia di buona amministrazione*, federalismi.it - Rivista di diritto pubblico italiano, comparato, europeo

Mantelero A. (2012). *Big data: i rischi della concentrazione del potere informativo digitale e gli strumenti di controllo* - Il diritto dell'informazione e dell'informatica- Anno XXVIII Fasc. 1 - 2012

Mayo M., (2016) *10 Must Have Data Science Skills*, disponibile su https://www.kdnuggets.com/2016/05/10-must-have-skills-data-scientist.html (consultato il 07/10/2021)

McKinsey Global Institute (2011). *Big data: The next frontier for innovation, competition, and productivity*, McKinsey & Company

Morcellini M. (2018). *Valore e limiti dei Big data*, disponibile su https://www.agendadigitale.eu/cultura-digitale/valore-e-limiti-dei-big-data-morcellini-ecco-le-questioni-aperte/ (consultato il 07/10/2021)

Mosca G. (2014). *Big data: tutto quello che c'è da sapere sulla professione del Data scientist*- Il Sole24ore, disponibile su https://st.ilsole24ore.com/art/tecnologie/2014-10-27/big-data-tutto-quello-che-c-e-sapere-professione-data-scientist—144450.shtml (consultato il 07/10/2021)

MSCI (2021). *Real Estate Market Size 2021. Annual update on the size of the professionally managed global real estate investment market*

Musacchio M. et al. (2018). *Intelligenza artificiale e Big data in ambito medico: prospettive, opportunità, crtiticità*, The Journal of AMD

Pratesi, M. (2017). *I Big data: il punto di vista di uno statistico*. Il menabò – Associazione Etica ed Economia, Menabò n. 62 14 aprile 2017 disponibile su https://www.eticaeconomia.it/i-big-data-il-punto-di-vista-di-uno-statistico/ (consultato il 07/10/2021)

Provost F., Fawcett T. (2013). *Data science and its relationship to big data and data-driven decision making*, Big Data , Vol. 1, No. 1, March 2013, Mary Ann Liebert, Inc

Rice M. (2019). *21 AI Real Estate Companies to Know*, disponibile su https://builtin.com/artificial-intelligence/ai-real-estate (consultato il 07/10/2021)

Siegel E. (2013). *Predictive analytics*. Wiley & Sons, Inc., Hoboken

Tremolada L. (2016). *A cosa servono i Data scientist?* – Nòva-Il Sole24ore, disponibile su https://nova.ilsole24ore.com/progetti/a-cosa-servono-i-data-scientist/ (consultato il 07/10/2021)

Udny Yule G. (1926), *Why do we Sometimes get Nonsense-Correlations between Time-Series?* A Study in Sampling and the Nature of Time-Series Journal of the Royal Statistical Society, Vol. 89, No. 1. (Jan., 1926), pp. 1-63.

Ward, J.H. (2012). *Managing Data: the Data Deluge and the Implications for Data Stewardship*. Unpublished Manuscript, University of North Carolina at Chapel Hill

World Economic Forum (2020). *The future of jobs report* su https://www.weforum.org/reports/the-future-of-jobs-report-2020 (consultato il 07/10/2021)

Wu L., Brynjolfsson E. (2015). *The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales*, in Economic Analysis of the Digital Economy, University of Chicago Press

SITOGRAFIA CONSULTATA

http://project.sobigdata.eu/ (data accesso 7 ottobre 2021)

http://www.tylervigen.com/spurious-correlations (data accesso 7 ottobre 2021)

https://ec.europa.eu/eurostat/cros/content/essnet-big-data-1_en (data accesso 7 ottobre 2021)

Quest'opera è distribuita con
Licenza Creative Commons Attribuzione - Non commerciale 3.0