UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

DIPARTIMENTO DI SCIENZE STATISTICHE

CORSO DI LAUREA IN ECONOMIA E MANAGEMENT

PROVA FINALE

"I BIG DATA PER COMPRENDERE IL PRESENTE E PREVEDERE IL FUTURO"

RELATORE:

CH.MO PROF. TOMMASO DI FONZO

LAUREANDO/A: SERENA COSTANTINO

MATRICOLA N. 1141315

ANNO ACCADEMICO 2020 – 2021

SOMMARIO

INTRODUZIONE

"Hiding within those mounds of data is knowledge that could change the life of a patient or change the world" affermava Atul Butte, MD, PhD, un professore associato di pediatria della Stanford University.

I dati ormai fanno parte della nostra quotidianità: ogni volta che la nostra posizione viene tracciata dai telefoni cellulari, il nostro stile di guida registrato dai computer di bordo presenti sulle auto; ogni volta che facciamo una ricerca su Google, acquistiamo da un sito, inviamo un messaggio vocale o semplicemente postiamo sui social, diamo vita a dei dati che possono essere raccolti, analizzati e sfruttati, anche economicamente. La nascita di nuove tecnologie, quali smartphone, sensori intelligenti e molto altro, e la loro utilizzazione sempre più frequente sia tra le persone che in ambito economico e aziendale, hanno dato vita ad un flusso continuo e quotidiano di dati economici e sociali.

L'avanzamento tecnologico degli ultimi decenni ha permesso l'immagazzinamento, la manipolazione e l'analisi di questa grande quantità di dati provenienti da fonti molto diverse tra loro. Tra gli anni '80 e '90, infatti, il 95 percento delle informazioni erano archiviate in modo analogico, ma nel giro di poco meno di vent'anni, nel 2007 la situazione si è ribaltata e le informazioni digitali hanno occupato il 94 percento della capacità totale di archiviazione (Tomassi, 2019). Questo processo di digitalizzazione ha permesso l'accesso a una moltitudine di dati in maniera più immediata e a costi ridotti, aprendo la strada a quello che oggi chiamiamo Big Data.

Il presente lavoro ha lo scopo di introdurre il concetto di Big Data e come questo viene usato nei vari ambiti applicativi.

Nel primo capitolo verrà presentata una panoramica generale sul concetto di Big Data. Cercando di dare una definizione che sia la più univoca possibile; tenendo conto che nel tempo sono nate varie definizioni e che a tutt'oggi non ne esiste una di interamente condivisa. Usando come linea guida le definizioni sorte nel tempo, si passerà alla trattazione delle sue caratteristiche, utilizzando il modello a 3 V introdotto da Laney (2001), arricchito di ulteriori V che sono sorte nel tempo. Infine, si porrà un accento sui possibili problemi che nascono dall'uso dei Big Data.

Nel secondo capitolo, invece, verranno introdotti i principali contesti applicativi dei Big Data, con un focus sull'applicazione dei dati nel caso più recente: il coronavirus. Il tutto sarà

supportato da una serie di riferimenti a casi empirici a titolo di esempio e per agevolare la spiegazione.

Nel terzo capitolo, infine, si parlerà di come i Big Data possono essere applicati per la misurazione di alcune importanti variabili economiche. Nella prima parte del capitolo verranno trattati due diversi tipi di classificazioni dei Big Data per l'applicazione nella previsione macroeconomica, la prima basata su dati numerici, la seconda invece suddivide i Big Data in base alla loro fonte. In seguito, verrà introdotto un elenco di alcuni tipi di Big Data utili per la previsione macroeconomica. Nella parte finale del capitolo, si presenteranno i vantaggi e gli svantaggi dell'utilizzo dei Big Data per la previsione. Ed infine, verrà introdotto il concetto di *nowcasting* ed alcuni esempi empirici di come i Big Data sono stati utilizzati per la previsione delle principali variabili macroeconomiche: prodotto interno lordo, disoccupazione e inflazione.

Capitolo 1

I BIG DATA

1.1 DEFINIZIONE

Essendo un fenomeno relativamente nuovo e complesso, la cui nascita si può fare risalire ai primi anni Duemila, non esiste una definizione univoca del termine *Big Data*. Un primo tentativo di definizione risale al 2001 da parte della società di consulenza Gartner, che li definiva come *"high-volume, high velocity and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"*. Per capire la mole di volume e la velocità di cui stiamo parlando ci viene in aiuto un'affermazione, fatta nel 2010 dall'ex amministratore delegato di Google, Eric Schmidt*: "There were 5 exabytes*[1] *of information created between the dawn of civilization through 2003, but that much information is now created every 2 days*." Infatti, proprio in quegli anni il volume annuo dei dati raggiungeva quasi i 2 zettabytes[2]. Grazie ai progressi tecnologici, quasi undici anni dopo, la crescita esponenziale dei dati ha raggiunto la soglia di 50 zettabytes nel 2020 e la sua corsa non sembra rallentare, difatti secondo AGCOM (2020), si pensa che entro il 2025 il volume dei dati arriverà a raggiungere i 163 zettabytes (Figura 1).

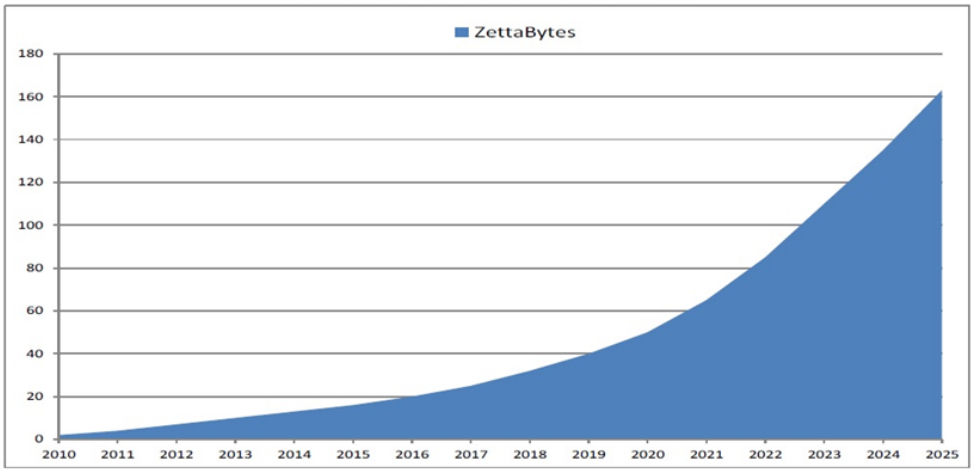

Figura 1: La crescita della datasphere (in zettabyte)

Fonte: IDC Data Age 2025 – aprile 2017

[1] Un exabyte (1EB) corrisponde a 10^{18} byte, equivalente a circa 33.554.432 iPhone 5 con memoria di 32 GB.
[2] Un zettabyte (1ZB) corrisponde a 10^{21} byte, che è l'equivalente di 281.479.977.500.00 file MP3 della grandezza media di 4MB.

Nel corso degli anni sono sorte altre definizioni, tra le più importanti ricordiamo: Apache Hadoop, che nel 2010 definiva i *Big Data* come *"datasets, which could not be captured, managed and processed by general computers within an acceptable scope*" (Chen et al., 2014).

Nel 2011, invece, l'azienda leader mondiale nell'*Enterprise Data Warehousing*, Teradata, affermava che: "Un sistema di *Big Data* eccede i sistemi hardware e software comunemente usati per catturare, gestire ed elaborare i dati in un lasso di tempo ragionevole per una comunità/popolazione di utenti anche massiva". Sempre nel 2011 un'ulteriore definizione di *Big Data* è stata data dal *McKinsey Global Institute*: "Un sistema di *Big Data* si riferisce a dataset la cui taglia/volume è talmente grande che eccede la capacità dei sistemi di database relazionali di immagazzinamento, gestione e analisi" (Manyika et al., 2011).
Anche se non abbiamo una definizione univoca, guardando a quelle che sono nate nel corso del tempo, possiamo delineare una traccia comune nel termine *Big Data*: le sue caratteristiche.

1.2 LE CARATTERISTICHE DEI BIG DATA

Per definire le caratteristiche dei *Big Data* ci basiamo sull'analisi fatta da Doug Laney, analista di Gartner, che nel 2001, descrisse in un report il Modello delle 3V: Volume, Velocità e Varietà.

- **Volume**
  Con il termine volume, si fa riferimento all'ingente massa di informazioni che ogni giorno viene generata e immagazzinata.

- **Velocità**
  La velocità è una caratteristica che ha più di un significato. In primo luogo, rappresenta l'alta frequenza con cui un flusso di dati passa da un punto di origine a un punto di raccolta. In secondo luogo, questa fa riferimento alla necessità di elaborare il flusso dei dati in maniera rapida, molto spesso in tempo reale (AGCOM, 2018). A volte, ottenere un vantaggio competitivo sulla concorrenza può significare identificare una tendenza, problema o opportunità solo pochi secondi o addirittura microsecondi prima di qualcun altro. Inoltre, bisogna considerare che, sempre più dati che vengono prodotti oggi, hanno una durata di conservazione molto breve, quindi le organizzazioni

devono essere in grado di analizzare i dati quasi immediatamente se vogliono ottenere informazioni adeguate per prendere decisioni oculate. (Eaton et al., 2011)

- **Varietà**

  La varietà fa riferimento all'insieme eterogeneo di fonti dalla quale sorgono le informazioni e all'eterogeneità dei dati stessi. Questa ha portato con sé nuove sfide: con l'esplosione dei sensori e dei dispositivi intelligenti, infatti, non ci si limita più ad analizzare dati strutturati, ma anche dati semistutturati e non strutturati provenienti dai social media, forum, dati di sensori attivi e passivi e molto altro (Kumar et al., 2017).

Con il passare degli anni questo modello, seppur ancora valido, si è esteso ad altre caratteristiche, quali:

- **Veridicità**

  Questa proprietà è fortemente dipendente dalle altre tre caratteristiche e si riferisce all'inaffidabilità associata alle fonti dei dati, come ad esempio tutte le analisi fatte sui dati provenienti dai social network (Gandomi and Haider, 2015). Infatti, all'aumentare del volume, della varietà e della velocità, diventa di cruciale importanza per le organizzazioni assegnare un indice di veridicità ai dati, in modo che si possano avere analisi più accurate e affidabili.

- **Variabilità**

  Questa caratteristica riguarda la corretta comprensione dei dati in base al contesto. Infatti, il significato o l'interpretazione di un dato può variare in base al periodo temporale in cui viene fatta l'analisi o semplicemente può variare in base all'ambito in cui viene raccolto e analizzato. È quindi importante trovare dei procedimenti che consentono di dare un significato ai dati in base al contesto al quale fanno riferimento.

- **Valore**

  La caratteristica viene interpretata da AGCOM, in un'indagine conoscitiva del 2018, come: "il valore che i dati assumono allorquando vengono elaborati ed analizzati, così da consentire l'estrazione di informazioni che possono contribuire all'efficienza e alla qualità di processi produttivi "tradizionali" ovvero qualificare intrinsecamente l'offerta di beni e/o servizi, in particolare in termini di innovazione e di personalizzazione".

In questo paragrafo sono state trattate solo le *V* principali, ma si fa menzione che con il tempo al modello generato da Laney, ne sono state aggiunte molte altre. Fino ad arrivare nel 2017, alle *42V's of Big Data* (Farooqi et al., 2019). Tra l'altro, sembra che il processo continui anche se ogni *V* che viene aggiunta risponde a questioni sempre più specifiche, diversamente da come avveniva per le prime *V*.

1.3 BIG DATA TRA PROBLEMI E SFIDE

Per implementare l'uso dei dati al meglio è importante comprendere le sfide che accompagnano i *Big Data*.

Una delle sfide di maggiore importanza è legata alle sue caratteristiche, riguardo la gestione di volumi di dati di grandi dimensioni e in rapido aumento (Almeida e Calistru, 2013). È sempre stata una questione impegnativa da gestire, in passato si è riusciti a mitigare il fenomeno grazie alla creazione di processori sempre più veloci. Ma ora il volume dei dati sta crescendo più velocemente dei processi di immagazzinamento e "eccedono già lo spazio disponibile per l'archiviazione" (Cukier, 2010). Questo significa che le informazioni disponibili vengono generate a una velocità maggiore rispetto allo spazio che serve per immagazzinarle: per questo motivo nel breve e nel medio periodo si dovrà cercare di aumentare la capacità di archiviazione dei dispositivi tecnologici. Collegato al problema della dimensione dei dati, sappiamo che maggiore è il set di dati da elaborare, più tempo ci vorrà per analizzarli. È probabile che l'implementazione di un sistema con uno spazio di archiviazione maggiore permetta anche la scansione più rapida dei dati, ma quando parliamo di *Big Data* dobbiamo far rifermento anche alla velocità di identificare informazioni qualificanti nella moltitudine di dati presenti. La scansione dell'intero set di dati per trovare gli elementi adatti sarebbe poco pratica, per questo motivo si rende necessaria l'implementazione di strutture indice che consentono l'identificazione degli elementi necessari in maniera più rapida (Almeida e Calistru, 2013).

Un aspetto molto dibattuto riguarda la tutela della *privacy:* questa è, molto probabilmente, la questione più delicata da trattare in quanto contiene implicazioni etiche, concettuali, legali e tecnologiche. L'*International Telecommunications Union* (Gordon, 2005) la definisce come *"the right of individuals to control or influence what information related to them may be disclosed"*. I dati che vengono raccolti il più delle volte contengono dati personali e informazioni private, basti pensare alle *query* che vengono digitate su Google o i dati presenti sui profili social. Questo rende la *privacy* una preoccupazione, che ha una vasta gamma di implicazioni per chiunque desideri usare i *Big Data*.

Un ulteriore rischio consiste nella cosiddetta “dittatura dei dati” (Cukier, Mayer-Schonberger, 2013), ossia la tendenza di affidarsi completamente ad essi, raccogliendone sempre di più. Se da una parte è vero, che maggiori informazioni consentono di avere una visione più ampia di un fenomeno, è anche vero però che questa non è sempre la soluzione migliore, in quanto i dati raccolti possono essere fuorvianti o semplicemente di cattiva qualità. Per questo motivo si deve fare una particolare attenzione a non affidarsi solo ed esclusivamente ai dati, ma utilizzarli in maniera complementare ai sistemi tradizionali.

Infine, lavorare con nuove fonti di dati comporta la nascita di un numero significativo di nuove sfide, la cui rilevanza e complessità dipende dal tipo di analisi condotta e dal tipo di decisione per la quale questi dati sono utilizzati (Almeida e Calistru, 2013).

Capitolo 2

I PRINCIPALI CONTESTI APPLICATIVI

Nel 2017, *The Economist* ha pubblicato un articolo intitolato, *"The world's most valuable resource is no longer oil, but data"*. I dati, quindi, diventano le materie prime del business, un input economico, quasi, alla pari del capitale e del lavoro (The Economist, 2017). Avvantaggiati dalla nascita di nuove tecnologie, che permettono la digitalizzazione di molte informazioni a cui prima non si aveva accesso, e dal numero di persone che usufruiscono di dispositivi che trasmettono e immagazzinano dati, si è passati dall'avere poche e scarne informazioni a una moltitudine, che hanno portato forti ripercussioni. Infatti, come afferma Craig Mundie, *Chief Research and Strategy Officer di Microsoft*: *"What we are seeing is the ability to have economies form around the data"*. Sono molte le aree applicative in cui i *Big Data* sono attualmente utilizzati. Essi posso portare valore, potenzialmente, in qualsiasi settore, ma ogni caso deve essere trattato separatamente per capire le opportunità e le sfide, poiché mutano sensibilmente da settore a settore. Diventa, quindi, interessante capire quali sono le differenze nei diversi campi nell'applicazione dei *Big Data*.

2.1 SETTORE MANIFATTURIERO

Come affermavano Kurtz e Shockley, in un report di IBM del 2013: *"Big Data is no longer confined to the realm of technology. Today, leveraging big data is a business imperative, and it is enabling solutions to long-standing business challenges for industrial manufacturing companies around the world. Indeed, industrial manufacturers are leveraging big data to transform their processes, their organizations and, in some cases, entire industries."*.
Nello stesso report i due autori affermano che i dati presentano molte opportunità e sfide per i produttori industriali, la maggior parte dei quali ha dati che provengono dalle operazioni e dal business, in termini di prodotti, *shareholders*, fornitori e partner. I produttori utilizzano, quindi, la tecnologia dei *Big Data* sfruttando questo e molte altre fonti di dati, per ottimizzare la produzione e le operazioni.
Molto spesso i processi produttivi nel settore manifatturiero sono complessi, non solo dal punto di vista del processo produttivo, ma anche per ciò che riguarda la complessa rete di relazioni in cui è coinvolto. Quindi, in questa prospettiva gli errori, i ritardi e il tempo di

inattività sono estremamente costosi, di conseguenza, si ritiene necessaria l'integrazione dei dati e in alcuni casi di sensori, che consentono al proprietario di avere informazioni in tempo reale sulle condizioni del prodotto, e allo stesso tempo creano una serie di dati che possono essere usati per la manutenzione e la qualità predittiva (Kurtz e Shockley, 2013). Grazie a queste informazioni diminuiscono i costi di magazzino e si riduce l'incertezza. Inoltre, si riduce il *time to market*, il tempo dall'ideazione di un prodotto alla sua commercializzazione, grazia alla sincronizzazione dei flussi produttivi con quelli logistici.
Altro aspetto importante è quello che riguarda la previsione dell'andamento della domanda, questo consente di avere una produzione in linea con le esigenze di mercato, sia in termini di qualità sia per quanto riguarda la ricerca di nuovi prodotti, rendendo i processi più flessibili.

2.2 SETTORE BANCARIO E FINANZIARIO

Il settore dei servizi finanziari e bancari è sempre stato caratterizzato da un enorme volume di dati generati e gestiti. La nascita di nuove tecnologie ha permesso alle banche di allontanarsi dai loro precedenti metodi di analisi, e di sfruttare i dati per migliorare il loro processo decisionale. Utilizzando le informazioni transazionali di un cliente, le banche riescono a monitorare continuamente e quasi in tempo reale il comportamento del consumatore; segmentando i clienti in base ai loro profili, il settore bancario sta aumentando le sue prestazioni complessive e di conseguenza la sua redditività, attuando *cross-selling e up-selling* basati sul segmento di appartenenza o grazie alla scoperta dei modelli di spesa, si riescono a realizzare offerte personalizzate per il singolo, aumentando l'esperienza del cliente e di conseguenza la sua soddisfazione (Lecis, 2020).
Un'altra importante applicazione dei *Big Data* nel settore bancario riguarda l'opportunità di identificare in modo tempestivo le frodi su carte di credito, sugli account e su una serie di altri servizi offerti dalle banche. Grazie all'implementazione di un modello che riesce a identificare uno schema di comportamento, come ad esempio la transazione di una grande somma monetaria o comportamenti non in linea con l'usuale atteggiamento del cliente, vengono sospese in completa autonomia del sistema, permettendo di minimizzare le perdite ed eventuali ripercussioni negative sulla soddisfazione del consumatore.

2.3 TELECOMUNICAZIONI

Il settore delle telecomunicazioni offre un buon esempio del perché l'analisi dei *Big Data* è importante per generare valore e rimanere competitivi. Oggi, il mercato delle

telecomunicazioni è caratterizzato dalla ricerca del miglior servizio possibile al minor prezzo di mercato, è frequente che in questo settore ci sia un elevato tasso di abbandono delle forniture, a causa dell'aspra e continua lotta che vi è tra i principali fornitori del servizio. Diventa essenziale quindi in questo campo mantenere le relazioni instaurate con i consumatori, favorendo la fidelizzazione del cliente per preservarne il valore. L'analisi dei dati in questo settore consiste principalmente nel rafforzare le relazioni con i clienti. Attraverso la raccolta delle informazioni sui soggetti e sui loro comportamenti d'acquisto si riescono a sviluppare servizi specifici per il singolo cliente, personalizzando l'offerta si aumenta la soddisfazione e l'*engagement* dei clienti, migliorando la profittabilità. Inoltre, di particolare rilevanza sono i dati che provengono dall'analisi del web e dei *social network*, dalla quale si possono ricavare molte informazioni sulla reputazione aziendale, capire quali sono gli aspetti del servizio maggiormente graditi dai consumatori e quali invece devono essere migliorati.

2.4 SETTORE ENERGETICO

La nascita di sistemi di monitoraggio, come gli *smart meters*, contatori o misuratori intelligenti che immagazzinano dati puntuali di consumo ad intervalli frequenti, ha permesso al settore energetico di implementare modelli predittivi da cui si ricava poi un progetto di pianificazione energetica e di attuazione di strategie di efficienza. I dati di consumo che provengono dalla misurazione intelligente, hanno portato numerosi vantaggi: in primo luogo nella *predictive maintenance*, dove i dati raccolti in tempo reale permettono all'azienda di attutare contromisure in breve tempo, per l'accertamento e le eventuali riparazioni di anomalie nei servizi, questo ha un impatto non indifferente sui costi aziendali e sulla soddisfazione dei clienti. In secondo luogo, i dati hanno avuto un effetto positivo sulla gestione della domanda energetica, infatti, attraverso attività di previsione si riesce a quantificare il fabbisogno del singolo consumatore, ottimizzando la produzione. Infine, un aspetto importante viene dalla gestione della *customer experience*, dove si punta alla trasparenza e alla visibilità, attraverso il *real time billing* (l'accesso alla fatturazione in tempo reale), e all'adozione di tariffe precise che vanno incontro alle esigenze del singolo, sulla base dell'analisi del livello di consumo e di una accurata segmentazione.

2.5 DIDATTICA ED E-LEARNING

L'impatto che hanno i *Big Data* in ambito didattico sia con riferimento all'insegnamento che all'apprendimento, è rilevante non soltanto nella progettazione dei moduli didattici, ma anche in tema di affinamento di obiettivi di apprendimento già predefiniti (Gutierrez-Santoz et al., 2012). Di particolare rilevanza è l'impatto che questi hanno sull'*e-learning*. I *Big Data*, in questo settore, sono dati che vengono creati dagli studenti mentre stanno frequentando un corso di *e-learning* o un modulo di formazione (Pappas, 2014), oltre la 'valutazione di fine-corso', solitamente condotta tramite scheda di gradimento che gli studenti devono compilare con la valutazione generale del corso, nasce l'esigenza di acquisire in tempo reale informazioni sempre più dettagliate ed organizzate sui vari ambiti di valutazione della didattica. Come affermano Giacalone e Scippacercola, nel 2016: "gli accessi ('le visite') alle pagine Web costituiscono basi informative che possono essere acquisite on line assieme ad altri dati, per comporre pattern utili alla valutazione della didattica." Quindi, dai *Big Data* si possono estrapolare informazioni utili per aumentare l'efficacia dell'insegnamento, valutare la comprensione delle informazioni e i comportamenti riguardanti l'apprendimento. Pappas (2014) ha proposto un breve ma significativo elenco di benefici che nascono dall'analisi dei dati relativi alla fruizione di un corso e-learning. Anzitutto, questa consente di evidenziare l'attività didattica che ha più impatto per raggiungere gli obiettivi del corso; permette, inoltre, di indentificare più facilmente i miglioramenti che devono essere applicati, le parti del corso che devono essere riviste o snellite per migliorarne la fruizione e mette in risalto i moduli che hanno attratto maggiore interesse. I dati vengono raccolti in tempo reale, questo consente di adattare la didattica nell'immediato senza dover aspettare la fine del corso. Infine, sulla base dei dati è possibile fare previsioni sui successi e sui fallimenti degli studenti e sviluppare i corsi in maniera tale che gli studenti abbiano sempre la possibilità di ottenere il miglior risultato possibile (Pappas, 2014).

Una interessante opportunità viene offerta dall'utilizzo di appositi programmi che permettono di fare una cernita dei dati, questi insieme all'applicazione di modelli matematici e di metodi statistici sui dati dell'*e-Learning*, una volta organizzati gli stessi in basi di dati, consente di produrre modelli di comprensione o anche di previsione utili all'affinamento o alla semplice valutazione dei metodi didattici (Chatti et al., 2012). Un differente tipo di approccio consiste nel valutare differenti parametri della formazione didattica prefissando per ciascuna variabile opportuni 'valori-soglia' da raggiungere per ritenere conseguiti gli obiettivi-formativi. (Siemens et al., 2011). Riprendendo, infine, le parole di Giacalone e Scippacercola (2016): "il vantaggio principale della raccolta e dell'analisi dei Big Data nell'e-learning, sta soprattutto

nella possibilità di ricavare informazioni utili per personalizzare l'esperienza formativa sulla base delle esigenze e degli stili di apprendimento degli studenti.". Anche se ricco di vantaggi però questo approccio, resta ancora da sviluppare in molti dei suoi aspetti.

2.6 MARKETING

Nel suo manuale "Principi di Marketing", Kotler (2015) afferma che: "Il Marketing è il processo mediante il quale le imprese creano valore per i clienti e instaurano con loro solide relazioni al fine di ottenere in cambio un ulteriore valore". Ed è proprio nella relazione con i clienti, che i *Big Data* danno i maggiori benefici al Marketing, migliorandone l'esperienza per il cliente e di conseguenza la profittabilità per l'azienda. Da questo punto di vista, un aiuto importante arriva dal *predictive marketing*, che attraverso la raccolta e l'analisi di dati acquisiti online e offline, riesce a segmentare il mercato in modo efficiente, identificando i comportamenti d'acquisto e permettendo all'azienda di creare offerte personalizzate alle esigenze del singolo consumatore.
Un caso emblematico è Amazon e il suo *recommendation engine*, come afferma Marr (2016) "*Online retailing relies on making as large a number of products or services available as possible, to increase the probability of making sales. Companies like Amazon and Walmart have thrived by adopting an "everything under one roof" supermarket mode".* Questo però fa sorgere nel consumatore una sorta di confusione, una grande quantità di prodotti offerti potrebbe infatti spostare l'attenzione dell'utente, lasciandolo indeciso su quale sarebbe la migliore decisione di acquisto per soddisfare i suoi bisogni. È per risolvere questo disagio del consumatore che è stato messo a punto "i consigliati in base ai tuoi interessi". Questo strumento che appare sotto la dizione di *recommendation engine* di Amazon, si basa su un filtraggio collaborativo, vale a dire che il sistema realizza un profilo basato sul comportamento d'acquisto e sui precedenti acquisti, sull'orario di connessione, i contenuti visualizzati e molti altri dati. In seguito, abbina questi comportamenti abituali a profili che seguono schemi simili per dare consigli in base a ciò che piace.
Un caso simile ci viene da un altro colosso, Netflix, che utilizza un grande team di specialisti con competenze specifiche nell'analisi dei *Big Data*. Lo scopo principale del servizio di *streaming* è sempre stato prevedere quello che piacerà guardare ai suoi clienti e i dati sono la linfa vitale per raggiungere questo obiettivo. Infatti, già dal 2006, quando ancora Netflix era, principalmente un servizio di noleggio DVD, cercava il miglior modo di creare un algoritmo per la previsione di come i loro clienti avrebbero classificato un film in base alle valutazioni precedenti. All'inizio le informazioni a disposizione consistevano solo in pochi dati relativi al

cliente e al tipo di film visionato, successivamente, quando lo streaming è diventato il metodo di consegna principale, sono diventati accessibili sempre più dati, consentendo a Netflix di creare modelli efficaci nel consigliare film che il singolo cliente avrebbe apprezzato (Marr, 2016)

L'uso dei dati diventa, quindi, un elemento fondamentale per tutte quelle realtà che vogliono restare competitive, fornendo un prodotto/ servizio sempre più personalizzato alle esigenze e agli interessi del singolo consumatore. Questo contribuisce a migliorare il rapporto con l'azienda e di conseguenza la redditività di quest'ultima.

2.7 SANITÀ

"I Big Data in Sanità si riferiscono a grandi set di dati raccolti periodicamente o automaticamente, che vengono archiviati elettronicamente, riutilizzabili allo scopo di migliorare le prestazioni del sistema sanitario", così nel 2016 la Commissione Europea definiva i *Big Data* nella Sanità. Esistono numerose fonti di dati nel settore sanitario, dalle informazioni che vengono dalle analisi del sangue alle più complesse visite specialistiche, che danno vita a un immenso bacino di dati ogni giorno. Purtroppo, ancora oggi l'accesso a queste informazioni è alquanto limitato, innumerevoli sono infatti le sfide che rendono difficile l'uso dei dati sanitari al massimo del loro potenziale. Innanzitutto, i dati provenienti da molti fornitori sanitari, specialmente gli ospedali, sono spesso segmentati e divisi. Infatti, i dati clinici del paziente, che consistono nelle anamnesi, le varie analisi e la registrazione dei segni vitali che vengono memorizzati nella cartella clinica, sono accessibili ai medici, infermieri. Questi però rimangono separati da una serie di altri dati, utilizzati dall'amministrazione, quali i reclami e informazioni sui costi, che vengono utilizzati per svolgere l'attività di assistenza sanitaria, ma non vengono adoperati per informare i protocolli di cura o di trattamento del paziente. Inoltre, i dati sulla qualità dei risultati, come le ricadute o il tasso di ritorno in chirurgia, fanno parte del dominio del dipartimento di gestione della qualità o del rischio, e vengono utilizzati per misurare le prestazioni (White, 2014). L'aggregazione di questi dati porterebbe all'uso ottimale delle risorse dell'ospedale e di conseguenza a una migliore cura del paziente. Infatti, dall'analisi dei dati incrociati, si possono ottenere diagnosi più precise e terapie specifiche per il singolo soggetto, ma permette anche di agire in maniera predittiva e preventiva, l'utilizzo di dati diversi permette di sviluppare un'analisi su campioni di pazienti, permettendo di individuare relazioni tra fenomeni e specifici rapporti di causa- effetto tra le diverse variabili. Bisogna fare attenzione, però, all'uso primario per la quale sono stati raccolti i dati, questo potrebbe infatti limitare e compromettere la validità dei risultati.

Un particolare caso di utilizzo dei dati nel settore sanitario ci viene da un progetto lanciato nel 2008 da Google. Nel novembre di quell'anno il famoso motore di ricerca ha lanciato un algoritmo*, Flu Trends*, in cui si raccoglievano i dati delle ricerche internet degli utenti che contenevano parole chiavi riferibili all'influenza, creando una precisa mappa della diffusione della malattia, su scala globale e quasi in tempo reale. *Google Flu Trends* si rivelò un abile alleato contro la lotta all'influenza stagionale, dove le *query* immesse sul motore di ricerca venivano effettuate da persone che stavano sviluppando i sintomi (Magistroni, 2018). In seguito, l'algoritmo subì una brusca battuta di arresto: durante la pandemia di influenza H1N1, infatti, *Google Flu Trends* fallì le previsioni di diffusione. Si arrivò a un risultato falsato a causa della pericolosità dell'infezione e della continua messa in guardia da parte dei media, infatti sempre più persone, anche chi non avvertiva alcun sintomo, faceva ricerche sull'infezione, portando ad una errata mappatura. Nonostante questo intoppo, l'applicazione della scienza dei dati nell'ambito sanitario continuò e continua tutt'oggi (Mura, 2020). Utilizzando oltre le previsioni che vengono da Google, piattaforme social come Facebook e Twitter, il campo di ricerca si è esteso non solo al controllo della diffusione delle malattie infettive ma in molti campi diversi. Ad esempio, nel 2013, i post su Twitter vennero usati per predire la depressione *post partum* attraverso i cambiamenti di comportamento sui social (Choudhury et., al, 2013). Nel 2015, invece, venne pubblicato un articolo sulla correlazione tra la diffusione dei pollini e *tweets* che riportavano parole che potevano essere ricondotte a sintomi della rinite allergica e nomi di antistaminici (Gesualdo et. al, 2015). Per quanto alla fine l'uso di *Google Flu Trends* non ha avuto il successo sperato, ha il merito di aver spinto la nascita di nuovi sistemi di monitoraggio e mappatura più veloci ed efficaci. Tuttavia, dobbiamo ricordare, che approcci di questo tipo hanno diversi limiti, basta poco perché i dati vengano alterati, come è accaduto per il caso dell'influenza H1N1; bisogna fare attenzione, quindi, alle parole chiave che vengano utilizzate per non rischiare di immettere nella misurazione risultati estranei allo scopo ed infine bisogna tener conto che l'utente non sempre fa ricerche perché coinvolto in prima persona nel fenomeno ma semplicemente per curiosità.

2.7.1 BIG DATA NELLA SANITÀ: IL CASO DEL COVID 19

Alla fine del 2019 a Wuhan, Cina, si è registrato un anomalo aumento di casi di polmonite. Analisi successive a pazienti infetti, hanno rilevato in seguito, che si trattava di un nuovo tipo di virus, conosciuto come *severe acute respiratory syndrome coronavirus 2.* Il numero dei casi è cresciuto rapidamente, fino alla sua diffusione a livello mondiale (Kurian et., al, 2020).

A causa degli alti livelli di diffusione e della rapidità del contagio, si è reso necessario l'utilizzo di metodi di mappatura in tempo reale. Come già detto nel paragrafo precedente, lo studio della distribuzione delle malattie infettive attraverso l'analisi dei *Big Data*, si è rilevata efficace nella diagnosi di eventi infettivi, permettendo la tempestiva preparazione di sistemi sanitari adeguati. Uno degli strumenti di sorveglianza più utilizzati è Google Trends, considerato tra i migliori analizzatori di tendenze nel comportamento di ricerca. (Kurian et al., 2020).

Da un progetto spinto dal Ministero dell'Innovazione, è nato uno studio, firmato da *Digita4good*, laboratorio di ricerca dell'università di Pavia, sull'analisi di parole chiavi e frasi riconducibili ai sintomi da coronavirus, come febbre, tosse, dispnea.

Da una prima analisi, è emerso che già prima del 21 febbraio 2020, giorno in cui l'Italia ha registrato il suo primo paziente infetto, le ricerche con la parola "febbre", effettuate nei mesi precedenti, erano più alte del 33% rispetto alla media dei quattro anni precedenti (2016-19), mentre l'utilizzo della parola chiave: "tosse" sono aumentate del 28%. Le interrogazioni su "palpitazioni" sono cresciute nello stesso periodo dell'86%. Gli incrementi del tasso di utilizzo delle parole chiave danno una indicazione forte e rafforzano l'idea che l'infezione da Covid-19 circolasse prima della scoperta del paziente zero. Un risultato simile è emerso dall'articolo *"Correlations between COVID-19 cases and Google Trends Data in the United State"* realizzato dalla *Mayo Clinic*, in cui si evince una correlazione tra i dati ottenuti dalle ricerche delle parole chiave correlate al coronavirus e i casi di infezione da Covid in America. Correlazione forte, soprattutto 16 giorni prima, del primo caso segnalato. Stessi risultati sono emersi dallo studio cinese dello stesso fenomeno, Chen et al (2020), affermano che: *"the peak Internet searches and social media data about the COVID-19 outbreak occurred 10- 14 days earlier than the peak of daily incidences in China."*

Questi ritardi della sorveglianza tradizionale possono essere ricondotti a vari motivi. In primo luogo, i rapporti ospedalieri possono variare da stato a stato e persino tra regione e regione, anche se si cerca di seguire una linea guida standard per la rendicontazione, le differenze rappresentano un ostacolo non indifferente. Una seconda fonte di ritardo è da attribuirsi alla mancanza di test adatti per la rilevazione dell'attuale pandemia nel suo stato iniziale. Questi problemi mettono in luce, come i modelli di sorveglianza legati a Internet siano un valido strumento per monitorare il diffondersi di malattie infettive e potenzialmente possono essere usati per prevedere nuovi focolai e possibili zone ad alto rischio. Sembrerebbe, quindi, che la soluzione ideale sia quella di affidarsi completamente ai *web data*, ma come nel caso della pandemia di influenza H1N1, bisogna fare attenzione alle ricerche sul tema che sorgono dalla pressione dei media o, come in questo caso, dai vari decreti sorti per il contenimento del

contagio e che portano dati falsati all'analisi. Quindi la chiave per un sistema di sorveglianza efficace è la combinazione tra sistemi tradizionali e sistemi basati sui dati. (Kurian et., al, 2020).

Capitolo 3

BIG DATA PER LA PREVISIONE MACROECONOMICA

La crisi economica scoppiata tra il 2007 e il 2008, ha evidenziato la necessità delle banche centrali e di altre istituzioni, di avere una continua valutazione delle attuali condizioni dell'economia. Il monitoraggio in tempo reale delle variabili economiche e finanziarie diventa, quindi, uno dei principali interessi dei *policy-makers* e altri agenti economici (Kapetanios e Papailias, 2018). Usando i metodi tradizionali non è possibile ottenere un'analisi in *real time*, a causa dei ritardi nella pubblicazione dei maggiori indicatori economici e fiscali, come il Pil e i suoi componenti o variabili fiscali.
Grazie al progresso tecnologico sono nati dei sistemi capaci di organizzare e immagazzinare una grande quantità di dati sulle variabili con una tempistica rilevante rispetto ai sistemi tradizionali. Questo cambiamento ha portato alla ricerca di nuovi metodi di previsioni macroeconomiche basate sull'utilizzo dei *Big Data* e in particolare l'uso degli strumenti come *Google Trends*, principalmente, o gli strumenti di analisi dei *social network*.

3.1 CLASSIFICAZIONE

Come affermato nel capitolo 1, una classificazione dei *Big Data* ci viene data dal Modello delle 3V di Laney (2001), ma in questo contesto di previsione macroeconomica, la suddivisione appare troppo generale.
Viene adottata quindi una classificazione basata per lo più su dati numerici, ideata da Doornik e Hendry (2015). La ripartizione identifica tre principali tipi di *Big Data*:

- *"Tall",* in cui il numero delle osservazioni, T, è alto ma non ci sono molte variabili, N. (T>>N). Un esempio è rappresentato dalle *query* di ricerca.

- *"Fat"* (N>>T), in cui ci sono molte variabili, ma non tante osservazioni. La categoria non è particolarmente interessante dal punto di vista della previsione economica, a meno che T non sia sufficientemente grande o che le variabili siano omogenee a sufficienza da permettere la stima di modelli econometrici.

- *"Huge",* dove il numero di variabili, N, e di osservazioni, T, è enorme. Questa categoria rappresenta il contesto ideale per fare previsioni, ma la raccolta di grandi quantitativi di dati è iniziata recentemente e non consente di fare confronti troppo lontani nel tempo. Basti pensare a *Google Trends* i cui dati più lontani risalgono al 2004.

Una terza modalità di classificazione dei dati ci arriva dall'unità statistica dell'UNECE (*United Nations Economic Commission for Europe*), che identifica tre tipi di data basati sul loro contenuto:

- **Social Networks**
  Questa categoria raccoglie tutti i dati provenienti dalle attività umane *(human-sourced information*), quali l'utilizzo dei *social network*, come Facebook o Twitter, le visite ai blog, i video che vengono visti su Youtube, le e-mail e i messaggi inviati e molti altri. Tutti queste informazioni sono, generalmente, poco strutturate e organizzate a causa delle varietà di tipologie di formati da cui sono composte.

- **Traditional Business Systems**
  Process-mediated data, questi sono processi che registrano e monitorano eventi di particolare interesse per l'istituzione sia pubblica che privata, consistono in transazioni commerciali, movimenti bancari, cartelle cliniche. Rappresentano un insieme strutturato di dati accompagnato spesso da transazioni e tabelle di riferimento per definirne il contesto.

- **Internet of Things**
  Riguardano *machine-generated data*, cioè, informazioni che derivano da sensori e macchine per la misurazione e la registrazione di fenomeni che avvengono nel mondo fisico. Come ad esempio: i dati che vengono dai sensori meteo, dalle telecamere per il traffico e dagli *smart meters,* ma anche dai *mobile sensors* e dalle immagini satellitari. I dati sono strutturati e adatti all'elaborazione dei computer, anche se bisogna fare particolare attenzione al volume e alla velocità con cui questi dati arrivano, che potrebbero mettere in crisi i tradizionali sistemi di gestione.

3.2 TIPI DI BIG DATA UTILI ALLA PREVISIONE MACROECONOMICA

Buono et al. (2017) identificano dieci categorie di *Big Data* importanti per la previsione:

- **Financial markets data**
  Grazie alla raccolta e all'analisi dei dati finanziari infragiornalieri su compravendite e quotazioni, si hanno informazioni dettagliate che potrebbero essere utilizzate nell'analisi dell'efficienza dei mercati, della volatilità, della liquidità e nella scoperta dei prezzi e delle aspettative. Da qui si evince quanto i dati sulle transazioni finanziarie, siano una fonte importante per l'analisi dei mercati e data la loro alta frequenza, questi forniscono notizie molto più velocemente degli indicatori macroeconomici a bassa frequenza, essenziali per il *nowcasting*.

- **Electronic payments data**
  Rappresentano tutti i tipi di trasferimento elettronico di fondi, da quelli più diffusi, cioè le transazioni avviate dal titolare della carta, che secondo il rapporto Capgemini e BNP Paribas del 2016, rappresentano il 65% del mercato globale delle transizioni *non-cash*, all'acquisto di beni o servizi, bonifici e pagamenti elettronici delle fatture sia su negozi online che offline (Buono et al., 2017). I pagamenti attraverso carta sono considerati una importante categoria di *Big Data*, grazie alla loro natura riescono a tracciare l'attività economica e in particolar modo gli acquisti fatti dalle famiglie, dimostrando di essere un utile strumento nel monitoraggio, nel *nowcasting* e nelle previsioni delle vendite al dettaglio dei consumi privati e altre variabili correlate.

- **Mobile phones data**
  I dati di base come il fare e ricevere chiamate o brevi messaggi, che venivano raccolti quasi trent'anni fa, quando sono stati introdotti per la prima volta i telefoni. Avevano portato numerose informazioni in termini di densità di popolazione, sviluppo economico di certe aree geografiche, ubicazione e tra le altre cose, l'uso del trasporto pubblico. I telefoni come li intendiamo oggi fanno molto di più che chiamare e inviare messaggi, la rapida crescita tecnologica ha permesso di raccogliere dati ancora più specifici, fornendo importanti informazioni sul comportamento umano. Da uno studio del 2015 di alcuni ricercatori del MIT emerge che i *mobile phones data* possono provvedere all'analisi in tempo reale dei livelli di disoccupazione. Studiando i dati di una fabbrica che aveva appena chiuso, come base del loro studio, si sono accorti che lo

schema delle comunicazioni cambia quando le persone non lavorano. Come afferma Toole, co-autore dello studio: *"Individuals who we believe to have been laid off display fewer phone calls incoming, contact fewer people each month, and the people they are contacting are different. People's social behavior diminishes, and that might be one of the ways layoffs have these negative consequences. It hurts the networks that might help people find the next job"*

- **Sensor data and the internet of things**

  Oggi giorno molti oggetti di uso quotidiano, come televisori, frigoriferi e molto altro, grazie ad una connessione ad Internet riescono ad interagire non solo con gli esseri umani ma anche tra di loro scambiandosi dati, interpretando segnali e i cambiamenti che provengono dall'ambiente, questo processo viene identificato con il nome *Internet of Things* (Qin et al., 2016).

  L'IoT trova numerose applicazioni in differenti ambiti della nostra vita, a partire dai *wearables*, dispositivi indossabili come *smartwatch* e *fitness tracker* che raccolgono dati sulle abitudini e lo stato di salute degli utenti, passando per i dispositivi intelligenti connessi alle nostre case che ci permettono di spegnere le luci, regolare la temperatura da remoto, rendendo le nostre abitazioni intelligenti e in grado di assecondare le nostre necessità. L'*Internet of Things* viene utilizzato anche nella produzione, come abbiamo visto attraverso gli *smart meters* nel settore energetico ma recentemente ha avuto un forte impatto anche nel campo del *retail* dove attraverso l'analisi dei percorsi seguiti dai consumatori si può migliorare il posizionamento delle merci all'interno dei punti vendita (Marchi, 2020).

- **Satellite images data**

  Le immagini dei satelliti sono state applicate per molti anni, principalmente in meteorologia, oceanografia e geologia. Ma la loro applicazione nell'economia è recente. Principalmente vengono usate per misurare variabili a lungo termine come povertà, disuguaglianza e crescita. Tra i primi ad utilizzare le immagini satellitari per i loro studi economici, troviamo Henderson, Storeygard and Weil che nel 2011 hanno fatto un confronto tra dati relativi all'andamento economico del Pil di molti paesi e i cambiamenti di luminosità notturna nelle diverse regioni trovando una relazione positiva, "Quando il reddito cresce, lo fa anche la "luce per persona" così osserva Henderson.

- **Scanner price data**

  Sin dall'inizio, l'indice dei prezzi di base per i beni di consumo è stato calcolato su un paniere fisso di beni. I prezzi venivano raccolti attraverso rilevazioni mirate e le agenzie statistiche pianificavano le indagini due o tre settimane prima di ogni mese prima della pubblicazione dei dati. Gli anni '70 del secolo scorso segnano un'importante svolta per l'elaborazione dei pagamenti delle merci nei negozi: grazie alla nascita degli *scanner* per i *barcode*, le transizioni vennero registrate elettronicamente, portando miglioramenti non solo nell'amministrazione dei singoli rivenditori, ma aprendo nuove strade per la ricerca accademica (Hanna et al., 2016). I dati che provengono dalla scannerizzazione dei *barcode* contengono importanti informazioni in termini di transizione delle merci vendute, prezzi pagati e quantità vendute per ogni articolo (Białek et al., 2020). Inoltre, contengono informazioni sugli attributi del bene, come il colore o la quantità di beni in un pacchetto, necessari per l'aggregazione dei beni in gruppi omogenei. Pertanto, gli *scanner prices data* possono essere molto utili nella previsione economica, perché grazie alla loro quotidiana scannerizzazione, consentono una misurazione ad alta frequenza della vendita al dettaglio dal lato dell'offerta, permettono inoltre di esaminare diverse regioni nella stessa economia e metterle a confronto, traendone informazioni importanti per quanto riguarda i consumi e i comportamenti d'acquisto.

- **Online price data**

  La crescita esponenziale dello *shopping online* sostituisce, o almeno integra lo *shopping offline*. In questa prospettiva, quindi, i prezzi online possono essere utilizzati come sostituti o integratori dei prezzi offline (Buono et al., 2017). Il *web scraping*, cioè l'attività di raccolta dati su Internet, offre flessibilità e alta automatizzazione, utili per il *nowcasting* e la previsione a breve termine delle vendite al dettaglio. Bisogna tenere a mente però che i prezzi online sono caratterizzati dalla stagionalità che potrebbe portare ad analisi errate.

- **Online search data**

  Consistono nella ricerca di una parola chiave o una frase in un motore di ricerca, che restituirà in seguito informazioni relative alla parola/frase ricercata, sotto forma di mix tra siti web, immagini, video e altri tipi di formato. Già dai primi anni 90 erano nate delle forme embrionali di motori di ricerca, ma è solo con l'avvento di Google, nel 2000, e la successiva introduzione di *Google Trends* che i dati online iniziano ad

essere aggregati e strutturati. Attraverso l'uso dei dati che provengono dalla ricerca online, come abbiamo detto in precedenza, sono nati modelli di previsione in campo medico ma anche in campo economico e finanziario.

- **Textual data**
  Ciò include qualsiasi tipo di *dataset* che fornisce informazioni riepilogative sotto forma di testo. Esempi di dati testuali includono: notizie e titoli dei media, informazioni relative a specifici eventi, ad esempio riunioni del consiglio di amministrazione, informazioni di *Wikipedia* (Buono et al, 2017).

- **Social media data**
  I *social network*, dalla loro introduzione sono diventati, parte integrante della nostra vita, raggiungibili da computer e dispositivi mobili sfruttano la loro accessibilità per mantenere una connettività e un'interazione continua non solo tra utenti ma anche con notizie ed eventi. Le informazioni provenienti dai *social media* descrivono quindi, come nel caso dei dati dalle ricerche online, attività e reazioni umane; dai post sui social si può tracciare la personalità di un individuo, dall'orientamento politico alle preferenze di lettura. Pertanto, è ragionevole presumere che i social hanno una capacità predittiva verso le variabili economiche.

3.3 VANTAGGI E SVANTAGGI DELL'APPLICAZIONE DEI BIG DATA NELLA PREVISIONE MACROECONOMICA

Il vantaggio principale che viene dall'utilizzo dei *Big Data* nella previsione macroeconomica fa riferimento alla natura tempestiva dei dati consentendo un'analisi ad alta frequenza e in tempo reale. Dall'altra parte, però, bisogna fare particolare attenzione ad evitare la cosidetta "superbia dei *Big Data"*, cioè *"the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis"* (Lazer et al., 2014). Anche se pensiamo ai *Big Data* come una grande raccolta di dati, un primo problema sorge dalla disponibilità degli stessi. La maggior parte delle informazioni passa, infatti, attraverso fornitori privati ed è relativa ad aspetti personali, pertanto, non è possibile garantire la continuità del conferimento degli stessi. Ad esempio, Google potrebbe smettere di fornire *Google Trends* in forma gratuita o non fornirlo affatto o potrebbero cambiare le leggi sulla privacy limitando l'accesso ad alcune informazioni. Un altro problema legato alla disponibilità dei dati riguarda la data alla quale si possono fare risalire gli stessi, che per

quanto riguarda i *Big Data* è molto recente, basti pensare che i primi dati disponibili di *Google Trends*, risalgono solo al 2004 (Kapetanios et al.,2018).
Un ulteriore problema da tenere in considerazione riguarda la frequenza con cui la qualità e la dimensione dei dati continua a cambiare nel tempo, derivanti dal sorgere di nuove applicazioni come WhatsApp o Twitter, e dell'eliminazione di molte altre o del loro diverso utilizzo.
Con l'impiego sempre più frequente di dati che provengono da ricerche online e quindi di diverso formato, sorge un ulteriore problema legato all'indisponibilità di dati in formato numerico o in un formato numerico direttamente utilizzabile, molto spesso infatti questi sono sotto forma di testo, immagini o video.
Infine, un problema comune anche ai dati standard ma più pervasivo nei *Big Data* a causa del loro volume, riguarda le irregolarità sui dati: presenza di valori anomali, osservazioni mancanti, modelli stagionali o periodici che richiedono una particolare attenzione per non cadere in trattazioni errate.

3.4 NOWCASTING

Il termine deriva dall'inglese, da *now*, "adesso", e *[fore]casting*, "previsione", questo fa riferimento alla sua caratteristica principale ovvero la capacità di prevedere un evento in tempo reale. Nel 2010, viene definito da Banbura et al.: *"as the prediction of the present, the very near future and the very recent past".* La locuzione, però venne coniata per la prima volta intorno agli anni '80, in ambito meteorologico, per indicare le previsioni a brevissimo termine, generalmente entro le tre ore, riferite ad un'area territoriale ristretta.
Grazie ai già citati miglioramenti nei sistemi di gestione dei dati e alla crescente domanda di previsioni con frequenza più elevata, il *nowcasting* ha trovato applicazione anche in economia. Ha una rilevanza particolare per le variabili macroeconomiche che vengono raccolte a bassa frequenza, generalmente su base trimestrale o semestrale, e rilasciate molto tempo dopo il periodo di riferimento. Per ottenere stime anticipate di questi indicatori economici, i *nowcasters* si affidano alle informazioni provenienti da variabili collegate a quelle target ma che vengono rilasciate con più ampia frequenza e in maniera più tempestiva. Per esempio, il Pil nell'area euro è disponibile solo su base trimestrale e viene rilasciato, generalmente dalle quattro alle sei settimane dopo la chiusura del trimestre. Attraverso l'uso di variabili legate al Pil ma che hanno una cadenza più frequente, si possono costruire le prime stime dello stesso.

Un buon esempio di applicazione di questa tecnica è lo studio "PIL Nowcasting. Il Pil del lockdown del Piemonte" effettuato dalla regione Piemonte nel 2020: le statistiche ufficiali italiane forniscono variazioni trimestrali del Pil italiano che vengono poi rese disponibili trenta giorni dopo la chiusura del periodo di riferimento, mentre le variazioni regionali sono annuali e vengono diffuse un anno dopo il periodo di riferimento. A causa dei lunghi periodi di attesa per avere le statistiche ufficiali, si ha la necessità di conoscere l'andamento economico nel presente. Per fare ciò ed avere stime più tempestive bisogna quindi affidarsi a una strada diversa da quella abituale ed è qui entra in gioco il *nowcasting*. Gli autori, infatti, affermano che: "Il nowcasting è un'esperienza di stima del Pil attraverso le relazioni di questo con variabili economiche rapidamente disponibili: il traffico sulle strade, il consumo di energia elettrica, le esportazioni, l'uso delle reti, indicano tutte un'attività in svolgimento. E il valore dell'attività economica è appunto il significato del Pil… Nuove variabili poi si aggiungono nell'epoca di internet. Le ricerche online di parole chiave, selezionate per capacità di segnalare un'attività con una potenziale ricaduta reale, sono anche esse correlabili al Pil e questa ricerca ha dato un esito positivo. Insomma, la traccia delle persone che cercano su Internet contiene informazioni utili per ricostruire il Pil".
Attraverso l'utilizzo di variabili tradizionali ma a frequenza più elevata, come il traffico pesante e i consumi di energia elettrica e di variabili nuove, quali ricerche su luoghi commerciali e su aziende di brand note in Piemonte, gli studiosi sono riusciti a stimare un modello econometrico. Mettendo in relazione come variabile dipendente, l'andamento trimestrale del Piemonte (fino al 2017) e come variabile indipendente il dataset precedentemente indicato e ripulito di variabili non statisticamente significative. Gli studiosi hanno riscontrato che il modello spiega circa l'86 per cento della varianza del tasso di crescita del Pil del Piemonte, confermando che l'utilizzo di variabili alternative porta a un'analisi adeguata e tempestiva del fenomeno.
Un altro esempio ci arriva da Modugno (2011), che considera come variabile di riferimento l'inflazione. Il documento si focalizza su due gruppi di dati ad alta frequenza: il primo è rappresentato dal prezzo del mercato mondiale delle materie prime con frequenza giornaliera. Il secondo, invece, riguarda osservazioni settimanali sui prezzi del carburante alla pompa*, the weekly oil bulletin price statistics* per l'area euro e *the weekly retail gasoline and diesel prices* per l'area americana. Come nel precedente caso, vengono considerati questi dati in quanto si muovono in contemporanea con l'andamento dell'inflazione delle rispettive aree di interesse e inoltre vengono rilasciate in maniera più tempestiva. L'autore conclude dicendo: "*The results suggest that the chosen weekly and daily data are important to improve the forecast accuracy*

for both the euro area overall HICP and the US total CPI inflation, especially at the shortest horizon, i.e. the current month"

3.5 APPLICAZIONE DEI BIG DATA NELLE VARIABILI ECONOMICHE

In questo paragrafo verranno esaminati brevemente alcuni studi accademici che riguardano l'applicazione dei Big Data per la disoccupazione, il pil e l'inflazione.

3.5.1 DISOCCUPAZIONE

"What would you search for if you thought you might lose your job?" È da questa domanda che sono partiti Choi e Varian, nel loro articolo *"Predicting initial claims for uneployment"*, del 2009. Dallo studio di parole chiave nelle ricerche online, come ufficio di disoccupazione, lavoro, richiesta di disoccupazione e altri, hanno cercato di capire se ci fosse una correlazione tra ricerche online e la crescita di disoccupazione. Attraverso l'utilizzo di *Google Search Insight* e *Google Trends* hanno suddiviso le *query* in categorie, in questo caso due: *"Local/Jobs*" e *"Society/social services/welfare & unemployment*", come affermano i due autori "*it is tempting to think that Google searches in these topics may be related to filings for unemployment benefits."* Come fonte di dati principale, hanno utilizzato l'*"Initial Jobless Claims"*, un rapporto settimanale del Dipartimento del Lavoro degli Stati Uniti che tiene traccia del numero di persone che presentano richiesta per il sussidio di disoccupazione ed è considerato uno dei principali indicatori del mercato del lavoro, ma come tutti i metodi tradizionali questi consentendo di avere una visuale del fenomeno solo tempo dopo l'effettivo periodo di analisi. L'utilizzo dei dati di *Google Trends* ha portato a miglioramenti significativi nell'accuratezza delle previsioni e soprattutto sulla velocità con cui arrivano i dati, in questo caso, quasi sette giorni prima del rilascio del rapporto settimanale. Altri autori, Arkitas e Zimmermann per la Germania e Suhoy per Israele, nello stesso anno, hanno usato *Google Trends* nell' analisi degli *online search data* connessi al tema della disoccupazione e le richieste di sussidio, arrivando alla stessa conclusione di Choi e Varian. Un simile risultato è stato ottenuto anche da Ferreira (2014) nel suo articolo *"Improving prediction of unemployment statistics with Google Trends"*. Nello studio viene considerato un periodo temporale tra il 2006 e il 2014 in termini di numero mensile di disoccupati, in Portogallo. Questi vengono testati utilizzando diversi approcci e dalla comparazione dei modelli è risultato che il modello della previsione della disoccupazione che utilizzava la variabile di

Google Trends, ha ottenuto risultati migliori, in particolare nel periodo in cui si è registrato un brusco cambiamento di tendenza.
Nel 2015 Reis et al., hanno condotto uno studio su come l'uso delle attività sul web possa migliorare le statistiche ufficiali sulla disoccupazione. In particolar modo, gli autori si sono focalizzati su due paesi: Francia e Italia.
Lo studio inizia dalla scelta dei termini di ricerca: per la Francia, gli studiosi hanno considerato termini come "pole emploi" cioè l'agenzia governativa francese che si occupa della registrazione dei disoccupati, li aiuta nella ricerca di un lavoro e fornisce loro aiuti finanziari, "indemnitè" questo si riferisce agli stanziamenti e "être au chômage" che letteralmente significa "essere disoccupato". Mentre per l'Italia, hanno esaminato quattro termini di ricerca: "impiego", "offerte lavoro", "curriculum" e "infojobs". Gli autori poi, hanno preso in considerazione due modelli di analisi: il primo basato sulla previsione della disoccupazione di un determinato mese, utilizzando i dati sulla disoccupazione del mese precedente. Mentre il secondo basato sulle *query* di ricerca, precedentemente considerate. Mettendo a confronto i due modelli gli studiosi hanno notato dei significativi miglioramenti nella previsione di entrambi i paesi attraverso l'uso del modello basato sulle ricerche.
Di particolare interesse è lo studio effettuato da D'Amauri e Viviano (2020) per Banca d'Italia, sull'impatto di breve periodo del Covid sulla ricerca di lavoro. Per analizzare il fenomeno gli autori utilizzano le serie storiche mensili sull'incidenza della ricerca di lavoro tramite Google, tenendo conto delle possibili limitazioni che questo può comportare. Dovute per esempio: alla ricerca di lavoro da parte di persone già occupate o a improvvisi aumenti di notorietà di determinate ricerche online a seguito di notizie dell'ultima ora, che possono incidere sul volume dei dati, portando a trattazioni falsate.
Gli studiosi rilevano una forte riduzione dell'attività di ricerca di lavoro nel marzo 2020, pari al -39% rispetto al periodo precedente. Il notevole calo potrebbe essere dovuto ad un aumento delle attività di ricerca complessiva su Google a seguito di una maggiore disponibilità di tempo libero. Per confutare questa ipotesi gli studiosi hanno esaminato le dinamiche dell'indice di ricerche per parole chiavi particolarmente popolari, come Spotify, YouTube, Facebook, La repubblica e molti altri, non trovando sostanziali differenze con i valori passati.
Infine, gli autori si sono soffermati sull'analisi delle ricerche relative alla mobilità e siti di notizie sportive. Settori che senza dubbio hanno subito un calo di interesse, a causa del blocco degli spostamenti non necessari e della sospensione degli eventi sportivi. Infatti, questo viene riscontrato nel calo delle ricerche registrato rispettivamente al -31% per le attività sportive e -50% per la mobilità, registrazioni non molto lontane da quella trovata per la ricerca di lavoro.
Questo studio mette in risalto una problematica importante: l'aumento dei possibili inattivi.

“La conferma di tale tendenza nei prossimi mesi darebbe un contributo negativo alla variazione del tasso di disoccupazione, mitigandone l’aumento in presenza di un probabile calo marcato dei livelli di occupazione.” così concludono gli autori. Ad oggi, a più di un anno di distanza dal calo rilevato, la tendenza sempre attenuarsi. Grazie alla presunta efficacia dei vaccini e alle riaperture, la fiducia dei cittadini sembra aumentare e con essa la ricerca di lavoro, infatti nell’aprile del 2021, abbiamo assistito a un lieve aumento degli occupati e una crescita più consistente dei disoccupati, a fronte di una diminuzione degli inattivi, ma la ripresa, purtroppo, è ancora lontana (Corriere dell’Università, 2021).

3.5.2 PIL

L’importanza che ha il web nelle nostre vite è indiscutibile, quindi non suscita sorpresa l’idea di studiare il Prodotto Interno Lordo di un paese, attraverso i nostri comportamenti online. In uno studio del 2012 apparso su *Scientific Reports*, si cerca di fare proprio questo. L’idea che sta alla base di questa analisi è, che più le *query online* sono rivolte al futuro più il PIL del paese a cui ci si riferisce è in crescita, mentre più le *query* sono orientate al passato e più il Pil è in difficoltà. La metodologia applicata dagli studiosi, Tobias Preis, Helen Susannah Moat, H. Eugene Stanley & Steven R. Bishop, si basa sul vaglio di *query* di ricerca che includessero il termine “2009”, che in questo caso rappresentava l’orientamento al passato, e il termine “2011”, l’orientamento al futuro. Attraverso la conta delle due, sono riusciti a stabilire un indice di orientamento verso il futuro. Il successivo confronto con la ricchezza di ciascuna nazione, il suo PIL pro capite, ha rilevato una forte correlazione tra le *query* e l’andamento del PIL. Sono riusciti, inoltre, a capire che gli utenti di Internet localizzati in paesi sviluppati digitavano più *query* con il termine “2011”, risultando più propensi a cercare informazioni sul futuro, a differenza dei cittadini dei paesi più poveri.
Già dal 2009, però, autori come Schmidt e Vosen, nel loro articolo *“Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends”* utilizzavano i dati provenienti dalle *query* per fare previsioni sul PIL. Partendo dal presupposto che circa il 70% del PIL statunitense è rappresentato dal consumo privato, avere informazioni tempestive sulla spesa delle famiglie, è importante per valutare e prevedere l’attività economica in generale. I dati sui consumi privati negli Stati Uniti sono pubblicati mensilmente e con un ritardo di un mese, l’uso di indicatori ad alta frequenza sono utili, quindi, non solo per predire il futuro ma anche il presente. Gli indicatori che tipicamente venivano utilizzati per prevedere il consumo, si basavano su sondaggi alle famiglie in cui si chiedeva di valutare le proprie condizioni

economiche e del paese, attuali e futuri. Questi indicatori cercavano di tener conto sia dell'aspetto economico che comportamentale dei consumatori, infatti si era riscontrata da tempo una forte correlazione tra la fiducia dei consumatori e il consumo negli Stati Uniti. Tuttavia, la letteratura empirica guardava con sospetto questo tipo di analisi, in quanto gli indicatori non catturavano accuratamente il legame tra le aspettative e le decisioni di spesa reali, inoltre, si credeva che le informazioni derivanti dai sondaggi potessero essere già acquisite dallo studio del reddito, della ricchezza e dai tassi d'interesse. In base a queste premesse gli autori introducono un nuovo indicatore di consumo, costruito usando i dati dei comportamenti di ricerca provenienti da *Google Trends*, dimostrando l'importanza di quest'ultimo. Infatti, affermano che: *"This study shows that Google Trends is a very promising new source of data to forecast private consumption. In almost all experiments conducted the Google indicators' in-sample and out-of-sample predictive power proved to be better than that of the conventional survey-based indicators."*

In uno studio del 2015, Galbraith e Tkacz valutano l'utilità di un ampio set di *electronic payments data*, comprendenti transazioni con carte di debito, di credito e assegni, come potenziali indicatori della crescita del Pil in Canada. I pagamenti elettronici nel corso del tempo sono diventati sempre più popolari. Basti pensare che nel 2009, l'89% dei pagamenti sopra i 50$ in Canada avveniva attraverso l'uso di carte, mentre solo il rimanente 11% in contanti (Arango et al., 2011). Essendo ogni transazione elettronica osservabile, fornisce una serie di dati utili su un'ampia gamma di attività di spesa, in particolar modo, sulla spesa per consumi, che ci fornisce una fonte di informazione incompleta ma diretta sulle variazioni del Pil. Gli autori procedono quindi all'osservazione dei dati e alla loro successiva aggregazione su base mensile, tenendo conto che i dati potrebbero crescere o diminuire per ragioni diverse dalla variazione della spesa totale. Per esempio: si potrebbe verificare un calo dell'uso delle carte di debito perché gli utenti hanno deciso di affidarsi alle carte di credito. Tenendo conto di queste limitazioni l'uso di questi dati rimane una variabile adeguata delle previsioni di breve periodo nella spesa per consumi. Gli studiosi concludono: "*We assess how electronic payments data, which are in principle available very quickly, can contribute to producing nowcasts. We find some suggestive evidence of an improvement in accuracy of the earliest nowcasts, primarily through the inclusion of debit card payments in the predictive model*"

3.5.3 INFLAZIONE

I metodi di base, usati nella maggior parte dei paesi, nella raccolta dei dati per la misurazione dell'inflazione sono rimasti gli stessi da più di dieci anni: "*A large number of people working*

for national statistical offices visit hundreds of stores on a monthly or bi-monthly basis to collect prices for a pre-selected basket of goods and services" (Cavallo et al., 2016).
I dati poi vengono elaborati e utilizzati per costruire gli indici dei prezzi al consumo e altri indicatori correlati. Il processo è, però, complesso e costoso, con tempi di rielaborazioni elevati. Le crisi economiche recenti hanno spinto *policy-makers* e altri soggetti a richiedere dati sempre più veloci e accurati. Proprio su questo bisogno di tempestività, nasce lo studio di Cavallo e Rigobon, del 2016: *"The Billion prices project: using online prices for measurement and research"*, gli autori utilizzano metodi alternativi, nello specifico gli *online prices data*, per determinare il tasso d'inflazione. Dall'analisi dei prezzi online, Cavallo e Rigobon sono riusciti a ricavare informazioni importanti per l'andamento dell'inflazione e per certi versi l'utilizzo di questi ha migliorato l'accuratezza delle previsioni stesse. L'utilizzo dei dati sui prezzi presenta numerosi vantaggi: in primo luogo, hanno un basso costo di osservazione. In secondo luogo, i dati vengono raccolti con frequenza giornaliera. In terzo luogo, i prezzi vengono registrati dal primo giorno in cui vengono introdotti in un mercato, fino al giorno in cui vengono rimossi. Questo rappresenta una grande differenza con i metodi tradizionali, che iniziavano in genere a monitorare le nuove merci *"only when the goods in the basket disappear from the stores"* (Cavallo et al., 2016). Altri vantaggi consistono nella possibilità di raccolta degli stessi da remoto e applicando metodi comuni che ne consentono il confronto con altri paesi. Infine, i dati online sono disponibili in tempo reale, senza ritardi. Attraverso l'uso di metodi alternativi, quindi, non solo si riescono a fare delle previsioni accurate, in *real time*, ma anche a costi ridotti.
Un metodo alternativo nello studio dell'inflazione ci arriva da Banca d'Italia, che nel febbraio del 2021 (Angelico et al., 2021), pubblica un interessante studio intitolato "Si possono misurare le aspettative di inflazione usando Twitter?".
Lo studio parte dalla selezione di una serie di tweet italiani, pubblicati nel periodo temporale tra giugno 2013 e dicembre 2019, contenenti alcune parole chiave strettamente legate ai termini: 'inflazione', 'prezzi', 'dinamica prezzi' come: "caro bollette", "caro benzina" e molti altri. Dopo un'attenta pulizia dei dati, per eliminare eventuali annunci pubblicitari o tweet che usano le parole chiave ma fuori contesto, gli autori procedono a un confronto tra la misurazione ottenuta attraverso l'applicazione dei dati provenienti da Twitter e il modello che utilizza le aspettative di inflazione dei consumatori da parte dell'ISTAT, "*we find that our new indicators are strongly correlated with them, but they have the advantage of being computed in almost real time. When comparing our Twitter-based measures of inflation expectations with the market-based ones available at the daily frequency, we find that our measures are also highly correlated with the Italian inflation swap rates"* così affermano gli

autori nella loro conclusione. Da questo studio si evince che l'analisi effettuata attraverso gli indicatori basati su Twitter coglie, in maniera adeguata le dinamiche delle aspettative di inflazione dei consumatori e trasmette contenuti informativi aggiuntivi rispetto ai modelli tradizionali.

Nel 2014, Griffioen et al., hanno effettuato uno studio sulla possibilità di utilizzare i prezzi dell'abbigliamento online per l'analisi dell'indice dei prezzi al consumo, cioè il principale indicatore economico utilizzato per monitorare il tasso di inflazione e il costo della vita in un Paese. Lo studio si focalizza sull'attività di raccolta dei prezzi di un singolo rivenditore, durante un periodo di circa due anni, che va dal gennaio del 2012 all'aprile del 2014. Dal documento sono emersi numerosi vantaggi dall'utilizzo del *web scraping*: in primo luogo, la raccolta dei prezzi online è più economica della raccolta dei prezzi nei negozi. Inoltre, la qualità dei dati online tende ad essere molto buona, consentendo l'osservazione di caratteristiche degli articoli in maniera più semplice. Gli autori continuano, affermando che: *"A potential advantage of web scraping is that it could be an effective way, and sometimes perhaps the only way, to collect information on clothing prices and characteristics."*

I siti web dei rivenditori contengono informazioni che vengono costantemente aggiornate ed è molto probabile che molti di essi non conservino i dati storici sui prezzi e caratteristiche. Attraverso l'uso del *web scraping*, le informazioni possono essere viste come delle istantanee e attraverso la loro raccolta è possibile creare una sorta di *data warehouse*, consentendo analisi temporali che altrimenti non sarebbero possibili.

D'altra parte, però, l'uso di questa tecnica ha portato anche degli svantaggi: i continui aggiornamenti sul sito web possono portare a problemi di rilevazione, inoltre la scelta stessa della strategia di *web scraping* può influire sulle informazioni raccolte e sull'oggetto. Ed infine "*the available information on characteristics may be insufficient, depending on the need for quality adjustment."*

CONCLUSIONI

In questa trattazione è stata fornita una panoramica generale sul tema dei *Big Data* e delle sue possibili applicazioni. Come abbiamo visto, il fenomeno è recente ma ha portato numerosi benefici in molti ambiti; grazie alla possibilità di monitorare i dati in tempo reale, le previsioni sono migliorate notevolmente, consentendo di comprendere meglio il presente e gettare le basi per una previsione futura. In campo finanziario, i Big Data hanno permesso di prevedere attività illegali, aiutato ad abbattere i costi e l'impatto ambientale in settori come quello energetico, dove l'analisi dei dati ha portato dei significativi cambiamenti. Ancora, la grande quantità di dati che un'azienda può ricavare in maniere abbastanza semplice e a basso costo, ha permesso di migliorare il rapporto con il cliente consentendo di offrire offerte personalizzate al singolo acquirente. Tutti questi benefici hanno ripercussioni anche sulle persone nella loro vita di tutti i giorni, la maggiore personalizzazione dei servizi soddisfa i bisogni dei clienti che siano questi attuali o futuri. Il tema è di particolare importanza anche in ambito sanitario, poiché dall'analisi dei dati incrociati si può arrivare a una terapia personalizzata incentrata a curare il paziente e non la malattia.

Dall'altra parte però questa enorme quantità di informazioni ha portato anche delle complicazioni, innanzitutto ha richiesto la nascita di nuovi processi e strumenti per poterla sfruttare al meglio. Inoltre, l'implementazione degli stessi non è semplice in quanto non impatta solo sull'aspetto di archiviazione dei dati ma modifica la percezione che era radicata nel tempo, generando problemi nelle routine e nei processi operativi. Ancora, anche se si riescono a sormontare questi ostacoli, una questione delicata sorge dall'utilizzo di dati personali, molto spesso privati, essenziali per migliorare le performance aziendali ma che vanno ad intaccare il concetto di *privacy* dei consumatori.

L'impressione generale che se ne trae è nell'insieme positiva: il miglioramento tecnologico accompagnato dalla mole dei dati che vengono generati ogni giorno, ha messo in movimento un processo di mutamento culturale e sociale che non sembra arrestarsi. Anzi con l'avvento, ormai imminente, del 5G la velocità di elaborazione dei dati subirà una spinta verso l'alto che non sarà limitato al settore delle telecomunicazioni, ma spazierà in diversi campi. Ciò vale in particolare per quanto riguarda la possibilità di dar vita a delle vere *smart city*, città sostenibili, efficienti e innovative, in grado di garantire un'elevata qualità di vita ai suoi cittadini attraverso l'utilizzo di sensori e tecnologie connesse e integrate tra loro.

BIBLIOGRAFIA

- AGCOM, 2018, Interim report nell'ambito dell'indagine conoscitiva di cui alla delibera n.217/17/CONS
- AGCOM, 2020, Indagine conoscitiva sui Big Data
- Almeida F., Calistru C., 2013, The main challenges and issues of big data management. International Journal of Research Studies in Computing. 2. 10.5861/ijrsc.2012.209.
- Angelico C., Marcucci J., Miccoli M., Quarta F., 2021, Can we measure inflation expectations using Twitter?, Banca d'Italia, Temi di discussione number 1318
- Arango, C., K. Huynh and L. Sabetti (2011), "How Do You Pay? The Role of
- Banbura M., Giannone D., Reichlin L., (2011), 'Nowcasting', In Oxford Handbook on Economic Forecasting, Clements MP, Hendry DF (eds). Oxford University Press: Oxford.
- Baronchelli, L. (2021, May 10). *Smart Cities in Europa: 6 esempi a cui ispirarsi*. Lumi. https://www.lumi4innovation.it/smart-city-cose-come-funziona-caratteristiche-ed-esempi-in-italia/ [Data di accesso: 13 maggio 2021]
- Białek J., Beręsewicz M., 2020, Scanner data in inflation measurement: from raw data to price indices.
- Buono D., Mazzi G., Kapetanios G., Marcellino M., Papailias F., 2017, Big data types for macroeconomic nowcasting
- Cavallo A., Rigobon R., 2016, The Billion Prices Project: Using Online Prices For Measurement And Research, Working Paper 22111http://www.nber.org/papers/w2211
- Chatti, M. A., Dyckhoff A.L., Schroeder U., Thüs H., A reference model for learning analytics. International Journal of Technology Enhanced Learning (IJTEL), 4,5-6, 2012, 318-221
- Chen LJ, Chen X, Zhang M, Pang CP, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Euro Surveill. 2020 Mar;25(10):2000199. doi: 10.2807/1560-7917.ES.2020.25.10.2000199. PMID: 32183935; PMCID: PMC7078825.

- Chen M., Mao S., Liu Y., 2014, Big Data: A Survey. *Mobile Netw Appl* **19,** 171-209, https://doi.org/10.1007/s11036-013-0489-0
- Choi H., Varian H., 2009, Predicting Initial Claims for Unemployment Benefits.
- Choudhury M., Counts S., Horvitz E., 2013, Predicting postpartum changes in emotion and behavior via social media. Conference on Human Factors in Computing Systems - Proceedings. 3267-3276. 10.1145/2470654.2466447.
- Corriere dell'Università*, 2021, Lavoro, ad aprile calano i giovani disoccupati e gli inattivi ma la ripresa è ancora lontana*. https://corriereuniv.it/lavoro-ad-aprile-calano-i-giovani-disoccupati-e-gli-inattivi-ma-la-ripresa-e-ancora-lontana/
- Cukier K., Mayer-Schoenberger V., 2013, The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs, 92*(3), 28-40. Retrieved May 2, 2021, from http://www.jstor.org/stable/23526834
- Cukier, K. 2010. Data, data everywhere. The Economist. Disponibile su http://www.economist.com/node/15557443/print, [data di accesso: 19 maggio 2021]
- D'Amauri F., Viviano E., 2020, L'impatto di breve periodo del Covid-19 sulla ricerca di lavoro, Banca d'Italia
- Doornik, J. A. and D.F. Hendry (2015), 'Statistical Model Selection with Big Data', Cogent Economics & Finance, 3(1), 2015
- Eaton C., Deroos D., Deutsch T., Lapis G., Zikopoulos P., 2011, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, IBM Corporation for McGraw Hill
- Farooqi, Muhammad & Shah, Munam & Wahid, Abdul & Akhunzada, Adnan & Khan, Faheem & Amin, Noor & Ihsan, Ali, 2019, Big Data in Healthcare: A Survey. 10.1007/978-3-319-96139-2_14.
- Ferreira P., 2014, Improving prediction of unemployment statistics with Google trends: part 2
- Galbraith J.W., Tkacz G., 2015, Nowcasting GDP with electronic payments data, European Central Bank, Working Paper No 10 / August 2015
- Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35 (2), 137-144.
- Gartner IT Glossary (s.d.). Big Data. Disponibile su <https://www.gartner.com/it-glossary/big-data>. [Data di accesso: 10 maggio 2021]
- Gesualdo F., Stilo G., D'Ambrosio A., Carloni E., Pandolfi E., Velardi P., Fiocchi A., Tozzi A., 2015, Can Twitter Be a Source of Information on Allergy? Correlation of

Pollen Counts with Tweets Reporting Symptoms of Allergic Rhinoconjunctivitis and Names of Antihistamine Drugs. PloSone.10.e0133706.10.1371/journal.pone.0133706.

- Giacalone M., Scippacercola S., 2016, Il ruolo dei big data nelle strategie di apprendimento.
- Gordon A., 2005, Privacy and ubiquitous network societies. Workshop on ITU Ubiquitous Network Societies, 6-15.
- Griffioen, R., De Haan J., Willenborg L., 2014, Collecting Clothing Data from the Internet, Statistics Netherlands Technical Report.
- Gutierrez-Santos S., S. Geraniou, E., Pearce-Lazard, D. & Poulovassilis. A. (2012) Architectural Design of Teacher Assistance Tools in an Exploratory Learning Environment for Algebraic Generalisation. IEEE Transactions of Learning Technologies, 5 (4), 2012, 366-376.
- Hanna J., Willenborgb L., Chessac A., 2016, An Overview of Price Index Methods for Scanner
- Henderson J., Storeygard A., Weil D., 2012, Measuring Economic Growth from Outer Space
- https://innovazione.gov.it/dipartimento/cosa-facciamo/task-force-covid-19/ [Data di accesso: 27 aprile 2021]
  Incentives at the Point-of-Sale", Working Paper 2011-23, Bank of Canada.
- Kapetanios G., Papailias F., 2018, Big Data & Nowcasting: Methodological Review, Economic Statistics Centre of Excellence a collaboration with Office for National Statistics
- Kotler P., Armstrong G., Ancarani F., Costabile M., 2015. Principi di Marketing. Quindicesima edizione, Milano: Pearson.
- Kumar A., Jothimani D., 2017, Big Data: Challenges, Opportunities and Realities.
- Kurian SJ, Bhatti AUR, Alvi MA, Ting HH, Storlie C, Wilson PM, Shah ND, Liu H, Bydon M. Correlations Between COVID-19 Cases and Google Trends Data in the United States: A State-by-State Analysis. Mayo Clin Proc. 2020 Nov;95(11):2370-2381. doi: 10.1016/j.mayocp.2020.08.022. Epub 2020 Aug 20. PMID: 33164756; PMCID: PMC7439962.
- Kurtz J., Shockley R., 2013, Analytics: The real-world use of big data in manufacturing, IBM Institute for Business Value
- Laney D., 2001, 3D data management: Controlling data volume, velocity and variety. Appl. Deliv. Strat. File 949

- Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis.” Science 343 (6176) (March 14): 1203–1205.
- Lecis, N. (2020, May 27). Big Data nel settore bancario: tutto ciò che dovresti sapere. Finance CuE | Close-up Engineering. https://financecue.it/big-data-nel-settore-bancario-tutto-cio-che-dovresti-sapere/13842/
- Magistroni M., 2018, Il web ci aiuta a predire le infezioni, La Repubblica https://www.repubblica.it/dossier/salute/labrevolution/2018/07/16/news/cosi_il_web_ci_aiuta_a_predire_le_infezioni-201906342/ [Data di accesso: 27 aprile 2021].
- Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A., 2011, Big Data: The next frontier for innovation, competition and productivity. McKinsey Global Institute. Disponibile su [Data di accesso: 12 maggio 2021]
- Marchi, M. (2020, July 1). *Internet of Things (IoT): significato ed esempi*. Punto Informatico. https://www.punto-informatico.it/internet-of-things-iot-significato-ed-esempi/ [Data di accesso: 14 maggio 2021]
- Marr B., 2016, Big Data in practice. How 45 successful companies used big data analytics to deliver extraordinary results, Wiley.
- Modugno M., 2011, Nowcasting Inflation Using High Frequency Data. International Journal of Forecasting. 29. 10.1016/j.ijforecast.2012.12.003.
- Morelli A. (2020, July 15). *Sfide e potenzialità di 5G e Big data*. Startmag. https://www.startmag.it/innovazione/sfide-e-potenzialita-di-5g-e-big-data/ [Data di accesso: 13 maggio 2021]
- Mura, M. (2020, March 19). *Coronavirus: Big data e AI a supporto della gestione del rischio*. Riskmanagement. https://www.riskmanagement360.it/risk-technology/big-data/coronavirus-big-data-e-ai-a-supporto-della-gestione-del-rischio/ [Data di accesso: 27 aprile 2021]
- Pappas C., 2014, Big Data in eLearning: The Future of eLearning Industry
- Preis T., Moat H., Stanley H. *et al.* Quantifying the Advantage of Looking Forward. *Sci Rep* **2,** 350 (2012). https://doi.org/10.1038/srep00350
- Qin Y., Sheng Q. Z., Falkner N., Dutstar S., Wang H., Vasilakos A., 2016. When things matter: a survey on data-centric Internet of Things, Journal of Network and Computer Applications, 64, PP. 137-153.
- Regione Piemonte, 2020, PIL Nowcasting. Il Pil del lockdown del Piemonte, Comitato Torino Finanza presso la Camera di Commercio di Torino.

- Reis F., Ferreira P., Perduca V., 2015, The use of web activity evidence to increase the timeliness of official statistics indicators, Eurostat
- Schmidt T., Vosen S., 2009, Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. Journal of Forecasting. 30. 10.2139/ssrn.1514369.
- Shafer, T. (2020, November 11). *The 42 V's of Big Data and Data Science*. Elder Research. https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/ [Data di accesso: 27 aprile 2021]
- Siemens G., Gasevic D., Haythornthwaite C., Dawson S., Shum S.B., Ferguson R., Duval E., Verbert K., and Baker R. S. J. D. Open Learning Analytics: an integrated & modularized platform, 2011.
- The Economist. (2017, May 11). The world's most valuable resource is no longer oil, but data. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data [Data di accesso: 11 giugno 2021]
- Tomassi, A. (2019, August 13). *Big data, che cos'è realmente questo ammontare di dati e da dove vengono*. Key4biz. https://www.key4biz.it/big-data-che-cose-realmente-questo-ammontare-di-dati-e-da-dove-vengono/268413/ [Data di accesso: 13 maggio 2021]
- Toole J., Lin Y., Muehlegger E., Shoag D., Gonzalez M., Lazer D., 2015, Tracking Employment Shocks Using Mobile Phone Data. Journal of the Royal Society, Interface / the Royal Society. 12. 10.1098/rsif.2015.0185.
- White S., 2014, A review of big data in healthcare: challenges and opportunities. Open Access Bioinformatics. 2014. 13-18. 10.2147/OAB.S50519.