

Data Driven Economy: come il processo di informatizzazione ha portato ad una nuova economia basata sui Big Data

Prof. Guido Tortorella Esposito

RELATORE

Silvio Dimonopoli 241281

CANDIDATO

Anno Accademico 2021/2022

INDICE

INTRODUZIONE	4
CAPITOLO I	5
INFORMATIZZAZIONE GLOBALE: COME I COMPUTER SONO PASSATI DA STRUMENTO TECNICO A FENOMENO DI MASSA	5
1.1 Informatica primordiale: dominio degli USA	5
1.1.1 Microchip, produzione di massa e acquisizioni: USA first mover nel settore	6
1.1.2 BASIC: il primo sistema operativo.....	7
1.1.3 Computer fenomeno di massa: dalle pubblicità ai video giochi, fino ai film	8
1.2 Regno Unito, trait d'union tra i due mondi	11
1.2.1 Dalle teorie e i primi modelli di computer... ..	12
1.2.2 ...agli usi didattici e la concorrenza ai produttori Statunitensi.....	14
1.2.3 BBC Micro.....	16
1.3 Europa: i “first-follower”	18
1.3.1 La Germania: culla della produzione hardware.....	18
1.3.2 Italia: Olivetti S.p.A., gli anni '80 e il fenomeno USA	19
1.4 Conseguenze di una diffusione planetaria	20
CAPITOLO II	22
I BIG DATA	22
2.2 Big Data: quando i dati fanno la differenza	22
2.1.1 Definizione: cosa sono i Big Data.....	22
2.1.2 La prima legge di Moore.....	24
2.1.3 Il Modello delle “3V”.....	25
2.2 Dalle biblioteche ai Big Data: cenni storici	27
2.2.1 Talete e i “Big Data”	27
2.2.2 Dal 1800 sino ad oggi	28
2.3 Ciclo produttivo dei Big Data	29
2.3.1 Big Data Management: fase di Generazione ed Acquisizione	30
2.3.2 Estrazione e pulizia dei dati, fasi di “ETL e Data Cleaing”	32
2.3.3 Immagazzinamento e Integrazione	33
2.3.4 Big Data Analytics: Analisi	34
2.3.5 Interpretazione ed estrazione del valore	36
2.3.6 Focus sul valore: si può calcolare quanto valore è stato generato?	38
2.4 In conclusione...	42
CAPITOLO III	43
APPLICAZIONI DEI BIG DATA: NUOVE FRONTIERE PER L'INDUSTRIA MODERNA	43
3.1 Introduzione: Data Driven Economy	43
3.2 Dalla teoria alla pratica: casi d'uso dei Big Data	45
3.2.1 Sviluppo del prodotto: Customer Demand Planning	45
3.2.2 Manutenzione Predittiva	46
3.2.3 Customer Experience	47
3.2.4 Frode e Compliance: applicazioni nel mondo della finanza e non solo	50
3.2.5 Efficienza operativa: il nuovo che migliora il vecchio.....	51
3.2.5 Promozione dell'innovazione: innovare per crescere.....	52
3.3 Machine learning: applicazioni nei vari settori economici di algoritmi di machine learning e lo “statistical learning”	53
3.4 Il “Recommender System”	57
3.4.1 Approccio collaborativo: il collaborative filtering	58
3.4.2 Approccio basato sul contenuto: content-based filtering.....	60
3.4.3 Approccio ibrido: Hybrid Recommender System	61
3.4.4 Caso studio: La Netflix Experience	61

3.5 Per concludere	69
CONCLUSIONE	70
RINGRAZIAMENTI	72
SITOGRAFIA	73
BIBLIOGRAFIA	74

INTRODUZIONE

I *Big Data* sono un argomento di attualità che affascina ma anche molto controverso. Molti ne parlano, quasi tutti ne sentono o ne hanno sentito parlare però in pochi sanno effettivamente darne una definizione ed essere chiari al riguardo. Spesso si fa ricorso a termini specifici come intelligenza artificiale, algoritmi, analisi predittive, ma per poter comprendere a pieno tali concetti occorre avere compreso l'elemento alla base di questi, i *Big Data*.

Il seguente elaborato verterà sul tema dei dati, il loro utilizzo nell'economia moderna e l'informatizzazione che ha permesso l'implementazione di tecniche in grado di sfruttare al meglio il loro potenziale. L'obiettivo nel corso del testo è quello di rendere il lettore consapevole del potenziale che i dati posseggono. Si vuole far comprendere come i *Big Data* non siano un concetto appartenente al futuro, ma sono una realtà quotidiana che incide in modo deciso su quella che è l'industria moderna. Per fare questo la scelta è stata quella di realizzare un testo composto di tre capitoli, all'interno dei quali verranno analizzati prima il processo di informatizzazione che ha riguardato i principali paesi occidentali, un passaggio necessario per poter parlare di un fenomeno legato per sua natura ai computer e al mondo di Internet, che ad oggi appaiono a noi così scontati ma il loro percorso ha inizio da lontano ed è tutt'altro che casuale. Si è poi proseguito nel secondo capitolo, con un'analisi tecnica dei *Big Data* per fornire una definizione che permettesse al lettore di comprendere in modo chiaro ed esaustivo il concetto, tramite l'illustrazione dei vari modelli relativi alle "V" e di nozioni informatiche come la legge di *Moore*. All'interno del secondo capitolo è inoltre descritto il ciclo di vita dei *Big Data*, raccontato attraverso un dettagliato esame del processo di generazione, gestione e analisi dei dati. Sono state analizzate le varie tipologie di analisi dei dati, definendo quelle che sono le differenze e le situazioni d'uso, per concludere poi con un paragrafo dedicato al calcolo del valore aggiunto generato dall'uso dei dati analizzati. Successivamente, all'interno del terzo ed ultimo capitolo, vengono elencati diversi applicativi, in diversi settori, dei *Big Data*. La scelta all'interno del capitolo terzo è stata quella di introdurre alla nozione di *economia 4.0* e soprattutto al fenomeno dell'"*economia guidata dai dati*". Successivamente viene fornita una descrizione dei diversi metodi di applicazione, passando dal mondo della finanza a quello del retail, aggiungendo quando possibile degli esempi pratici, per poi andare nel dettaglio per quel che concerne il *machine learning*, attraverso un focus sugli algoritmi più utilizzati e la presentazione di un caso di studio riguardante la piattaforma di *streaming Netflix*.

CAPITOLO I

INFORMATIZZAZIONE GLOBALE: COME I COMPUTER SONO PASSATI DA STRUMENTO TECNICO A FENOMENO DI MASSA

“Introducing Apple II”

Frase commerciale inserita in manifesto pubblicitario del home computer Apple II, 1977

1.1 Informatica primordiale: dominio degli USA

Vent'anni fa l'informatica era molto diversa dal fenomeno che oggi giorno viviamo e tocchiamo con mano, non era su scala globale e si focalizzava prevalentemente negli Stati Uniti¹. Un manipolo di aziende produceva gran parte dei computer utilizzati poi in tutto il mondo e questi erano per lo più venduti su suolo americano, una piccola parte veniva invece destinata a paesi industrializzati alleati degli Stati Uniti. Una situazione quindi molto diversa rispetto a quella attuale dove gli stessi stati uniti sono importatori di componenti dall'estero, specialmente dall'Asia.

Questo fenomeno ha però delle motivazioni alla base per cui si è venuto a verificare, non si può semplificare pensando soltanto che gli Stati Uniti in quanto primi inventori si sono mossi per primi e tutti i paesi a seguire si sono mossi per imitare. Le ragioni principali sono lato *hardware*, con il costo delle materie prime produttive che crollò negli anni '60-'70, lato *software*, con la diffusione del linguaggio *BASIC* e l'invenzione dei suoi interpreti, su tutti *“Altair BASIC”*. Ulteriore motivazione della diffusione degli home computer risiede nella comunicazione commerciale e nelle abitudini di acquisto degli americani. Tutto questo ha portato alla diffusione prima dei *microcomputer*, computer prodotti con un singolo microprocessore, dal costo limitato e soprattutto dall'ingombro molto ridotto rispetto alle macchine fino a quell'epoca. Questi erano però destinati a un pubblico ridotto, soggetti dotati di competenze tecniche quali hobbisti e tecnici che avevano le conoscenze per poter assemblare i kit venduti. Di fatto questi computer erano per lo più sprovvisti di componenti quali monitor, tastiera o mouse, il singolo contenitore era inoltre spesso da assemblare personalmente. Successivamente, negli anni '70 si diffusero gli *home computer*, definiti la seconda generazione di microcomputer, destinati all'uso domestico da parte di un singolo individuo non tecnico. Questi venivano commercializzati come computer economici e accessibili, dall'ingombro limitato e dalla potenza di calcolo inferiore solo ai computer destinati al lato business (macchine dalla mole più ingombrante e molto tecniche). Primo Home computer della storia è considerato Apple II, commercializzato nel 1977, che influenzò più di tutti gli home computer successivi. La diffusione di questi computer fu possibile grazie alla enorme quantità di microchip prodotti, possibile per via dei costi ridotti di produzione, che permise la produzione in massa dei vari modelli prodotti nel tempo. Il nome suggerisce ovviamente un uso prevalentemente domestico ma non

¹ Seymour E. Goodman, “Defining a Decade: Envisioning CSTB’s Second 10 Years, National Research Council, 1997 (pp. 18-21)

esclusivo, poiché essi erano in grado di svolgere funzioni che tornarono utili anche in ambito lavorativo, specialmente per i piccoli commercianti che gestivano dell'attività e dovevano tenere le informazioni in modo chiaro ed efficiente. In questi anni spopolarono prevalentemente modelli di aziende quali Apple, Atari, Commodore e Sinclair Research, che vendettero milioni di unità e permisero alle masse la prima familiarizzazione con l'informatica. Contribuì alla vendita di numerose copie la possibilità di video giocare tramite le macchine in questione e che spinse quindi molte famiglie a comprare i computer inizialmente come console da gioco per i bambini che però avevano poi la possibilità di accedere a funzioni ben più articolate del solo video giocare, come ad esempio programmare e sfruttare più avanti, dagli anni '80, le "BBS", ossia le *Bulletin Board System*, delle cosiddette "community online primordiali". Questi ultimi aspetti fecero sì che il computer allargasse il suo bacino di utenza tutti gli appartenenti al nucleo familiare, dai bambini sino agli adulti.

Lato software, il sistema operativo BASIC, sistema non complesso permise la realizzazione di computer in massa in quanto non richiedeva memorie di grandi dimensioni per essere memorizzato, inoltre la lingua utilizzata era l'inglese e questo risultò essere molto intuitivo per i cittadini americani e inglesi che per primi fecero uso di queste macchine.

In quegli anni poi il marketing e la comunicazione stavano evolvendosi ed ebbero un ruolo chiave nel pubblicizzare i prodotti informatici, facendo spesso leva su influenze provenienti dal cinema e dalla televisione, che in quegli anni erano dominati dal genere fantascientifico, per invogliare ad acquistare e ammaliare i consumatori. La presenza nelle famiglie americane di un computer in casa aumentò vertiginosamente quindi tra il decennio '77-'87, anni considerati di massimo splendore per gli home computer, che sul finire degli anni '80 lasciarono il posto ai più moderni *personal computer*, che aprirono la strada a quello che oggi tuttora utilizziamo quotidianamente. Questo ha fatto sì che la generazione che crebbe in quel decennio fosse poi in età adulta più informatizzata e pronta alle future forme di computer che hanno poi visto la luce.

1.1.1 Microchip, produzione di massa e acquisizioni: USA first mover nel settore

Uno dei principali motivi che permise la diffusione dei computer, prima negli Stati Uniti per poi arrivare a gran parte del mondo è stata la produzione in massa di microprocessori, processori interamente contenuti in un circuito integrato. Le sue dimensioni erano estremamente piccole, come il nome suggerisce, questo favorì la riduzione delle dimensioni dei computer e la possibilità di costruire macchine utilizzabili nelle case di tutti. I bassi costi di produzione di questi processori permisero lo sviluppo di diversi modelli e ciò permise l'affermarsi di diversi player, molti dei quali ancora oggi tra i principali produttori di componenti hardware per l'informatica. Intel fu la prima a produrre negli anni '70 processori a 4 e 8 bit, che vennero poi sfruttati per la produzione di micro e home computer. Iniziò quindi una produzione massiccia stimolata dalla domanda sempre crescente di computer che arrivava principalmente dal mercato americano, dove risiedevano quasi tutti i produttori, che a loro volta stimolavano la domanda attraverso la produzione e l'offerta di processori sempre più avanzati e in grado di offrire prestazioni performanti. Iniziò internamente agli Stati Uniti una guerra

commerciale quindi tra i vari produttori che si sfidavano nel produrre processori sempre più potenti, vedeva fronteggiarsi *Motorola, Intel, MOS Technology*. Tutto questo, produzione di massa e prezzi contenuti, permise quindi la produzione di computer dal prezzo economico in larga scala.

Fenomeno che contribuì ulteriormente al crescere dei volumi di produzione e delle innovazioni fu quello delle acquisizioni, nonché le partnership commerciali tra imprese dei vari settori hardware e software. In questo Commodore fu regina. Fondata nel 1962 come *Commodore International Ltd*, iniziò come produttrice di calcolatrici elettroniche, divenendo ben presto uno dei maggiori produttori del mercato. Ma nel 1975 la Texas Instruments, società fornitrice di materiale per la produzione di calcolatrici, decise di entrare sul mercato anche come produttore. L'anno dopo si decise allora di acquisire la MOS Technology. Questo permise di entrare in possesso di progetti già avviati e di personale specializzato come ingegneri elettronici, ma soprattutto permise a Commodore di evitare di essere dipendente da fornitori terzi per le risorse destinate alla produzione dei propri dispositivi. Questa strategia fu portata avanti negli anni avvenire, come ad esempio venne fatto per l'home/personal computer *Amiga*, progetto sviluppato inizialmente dalla società *Hi-Toro* che fu acquisita da Commodore nel 1984, permettendo la produzione e la messa in commercio di tutta la famiglia di computer Amiga negli anni a venire. La strategia di integrare all'interno della propria filiera produttiva tramite acquisizioni e fusioni permise quindi di produrre a costi più bassi e l'accesso a conoscenze e competenze tecniche che permettevano altrimenti alle imprese fornitrici di materiale informatico di avere un potere contrattuale elevato. Esse permisero di proseguire fin da subito i progetti già avviati e di migliorarli ulteriormente invece di dover perdere ulteriore tempo nella ricerca di personale qualificato e competente, evitando rallentamenti di produzione. Inoltre, tramite l'integrazione si eliminava la dipendenza da risorse

1.1.2 BASIC: il primo sistema operativo

La produzione in massa e a basso costo di microprocessori è solo una delle ragioni che permisero l'adozione su larga scala dei computer. Il diffondersi dei primi modelli fece nascere l'esigenza di un sistema operativo in grado di eseguire i programmi necessari a soddisfare le richieste dei soggetti e la distribuzione a soggetti non tecnici richiedeva un uso relativamente semplice. A questo andava aggiunto un costo non elevato e un peso non eccessivo per poter essere racchiuso nelle macchine prodotte. Per questo negli anni '70 e '80 nacquero dialetti e interpreti del linguaggio BASIC, nato qualche anno prima al Dartmouth college². La diffusione dei computer fu quindi cruciale per la diffusione del BASIC. Il linguaggio era inizialmente sviluppato per hobbisti ma con la crescente popolarità degli home computer presero sempre più largo libri di programmi BASIC e giochi BASIC che contribuirono ad alfabetizzare gli individui all'informatica. Ulteriore spinta alla propagazione del BASIC fu lo sviluppo da parte di Bill Gates e Paul Allen di una versione di BASIC per l'allora diffusissimo Altair computer. Molta della sua diffusione fu dovuta alle diverse copie pirata che consistevano in quasi il 90% delle copie in circolazione³. Nacque in quegli anni la Microsoft,

² BASIC Begins at Dartmouth, <https://www.dartmouth.edu/basicifty/basic.html>

³ Stephen Manes, Paul Andrews, Gates, New York, Simon and Schuster, 1994, pp. 90

precisamente nell'anno 1975, la cui versione di BASIC, nominata "*Microsoft BASIC*", fu utilizzata dai principali home computer diffusi. Relativamente ad esso furono prodotti numerosi libri, vista la sua popolarità e diffusione, tra i più famosi quelli editi da *David Ahl*, come "*101 BASIC Computer Games*", che negli anni '80, vista la enorme diffusione di home computer, fu il primo libro sui computer a vendere un milione di copie⁴.

1.1.3 Computer fenomeno di massa: dalle pubblicità ai video giochi, fino ai film

Dalla fine degli anni '70 passando per il decennio degli '80 negli Stati Uniti il fenomeno dei computer si è sviluppato e diffuso in modo repentino ed esponenziale, passando in pochi anni da oggetto inaccessibile alla maggioranza delle persone e appannaggio di pochi per uso lavorativo a oggetto d'uso comune e quotidiano presente in quasi tutte le case. Alla diffusione contribuì anche il fenomeno dell'advertising e della pubblicità commerciale che in quegli anni anche grazie alla diffusione dei computer si evolse ed ebbe un ruolo attivo nel rendere i computer uno oggetto commerciale disponibile per chiunque. In questo campo i contributi più grandi vengono dati da Apple, Commodore ma anche da catene di diffusione di prodotti elettronici come *RadioShack* o *High Technology*. Di fatto il computer era diventato un prodotto acquistabile presso negozi specializzati, i cosiddetti *computer store*, e questo li rendeva ancora più vicini alle persone che prima non avevano possibilità di uscire di casa per acquistare un computer viste le dimensioni colossali che li contraddistinguevano in precedenza. Gli spot commerciali ebbero un ruolo importante soprattutto perché permisero alla massa di entrare a conoscenza dell'esistenza dei prodotti e visualizzarne subito le funzioni che venivano presentate dai protagonisti degli spot. I primi spot erano contraddistinti per lo più dalla presenza di famiglie dove i vari membri a turno usavano il computer, spesso i figli, mentre i genitori ne decantavano le qualità e le funzioni per aiutare nelle attività quotidiane come studio e apprendimento⁵. Anche Apple, famosa per gli spot più elaborati e artistici, pubblicava spot che istruivano il potenziale consumatore sulle funzioni del computer ma in modo più particolare e in pieno stile con quello che sarebbe stato il modo di fare pubblicità della società⁶. Oltre agli spot famosi sono anche i manifesti pubblicitari, gli inserti nelle riviste specializzate, che attraverso immagini, didascalie e articoli dedicati introducevano ai computer e alle loro funzioni i lettori o approfondivano il bagaglio di conoscenze di soggetti, appassionati o meno.

⁴ Anderson, J.J. "Dave tells Ahl-the history of creating computing", *Creative Computing*, Volume 10 (November 1984), pp. 66-8

⁵ Spot Commerciale del RadioShack TRS-80, 1978, <https://www.youtube.com/watch?v=5QFV7eovE7w>

⁶ Spot commerciale Apple II, 1977, https://www.youtube.com/watch?v=FxZ_Z-_j71I



Figura 1.1: Pubblicità commerciale di Apple per il suo Apple II, (Fonte: <https://www.pinterest.cl/pin/684687949576583417/>)

Ma nei primi anni i computer erano un oggetto che divideva per il suo fine ultimo: da una parte i produttori spingevano per un uso variegato mettendo in mostra tutte le possibilità che questi offrivano, c'erano poi le strutture governative che ne vedevano un uso educativo per i ragazzi e poi c'erano i consumatori che, soprattutto nei primi anni usufruivano dei computer per videogiochi e attività ricreative. Questa era la funzione maggiormente diffusa per i computer e ovviamente i produttori ne facevano ampiamente parte dei motivi per cui acquistare un home computer, cogliendo nel tempo il segnale dato dal pubblico consumatore e trasformandolo nel principale cavallo di battaglia per molti dei prodotti lanciati nel decennio tra il 1978 e il 1988. In questo senso la pubblicità si rivelò strumento cangiante perfetto per andare in contro alle volontà dei produttori e rivenditori. Essa fu improntata soprattutto sull'esaltare la capacità dei prodotti di permettere prestazioni elevate di gioco, di possedere una compatibilità ampia con la maggior parte dei videogiochi più diffusi del periodo. Vennero resi più accattivanti gli spot televisivi tramite l'uso di testimonial famosi e spesso provenienti da film per cui venivano realizzati videogiochi o appartenenti al genere fantascientifico⁷. Questo diffondersi per uso ricreativo ha portato a uno sviluppo del settore videoludico esponenziale e quindi a un circolo virtuoso per cui il crescere del settore computer ha fatto sì che nascessero nuovi settori legati al comparto software e hardware, che a loro volta acquisendo importanza e popolarità hanno influenzato il settore iniziale. Ad amplificare questo fenomeno fu la messa in commercio nei negozi, a prezzi nel tempo via via sempre più accessibili, che hanno reso il computer da oggetto di nicchia a oggetto alla stregua componente degli elettrodomestici di casa. La presenza nei negozi e l'uso videoludico lo resero ovviamente concorrente a livello commerciale con le console da gioco, all'epoca unica alternativa ma limitata nelle funzioni al solo atto del video giocare. Ma la maggiore diffusione degli home computer, dovuta in parte a un prezzo spesso più contenuto o giustificato dalla maggiore presenza di funzioni e l'elevata popolarità guadagnata grazie ai numerosi spot commerciali, fecero sì il decennio degli anni '80 vide la maggior parte dei suoi titoli videoludici più famosi prodotti per home computer, dove da re indiscusso fece il Commodore 64, home computer che ancor a oggi fa registrare il numero di unità vendute più alto per un apparecchio informatico e la libreria di

⁷ Spot commerciale Commodore VIC20, 1981, <https://www.youtube.com/watch?v=UK9VU1aJvTI>

videogiochi più vasta esistente⁸. Chiave del successo fu proprio il marketing aggressivo che lo portò ad essere venduto in grandi magazzini, discount e negozi per giocattoli.

I computer divennero fenomeno di massa a tal punto che anche il settore del cinema risentì della loro influenza, non a caso nel decennio in cui vennero introdotti sul mercato gli home computer e si sviluppò il loro mercato il genere più diffuso di film e dai maggiori riscontri commerciali divenne quello fantascientifico. Essi avevano modificato le abitudini di vita quotidiana dei consumatori, i ragazzi passavano sempre più tempo a giocare e in tal modo venivano rappresentati sullo schermo, ma soprattutto permisero la realizzazione degli effetti speciali in virtuale, allora invece spesso realizzati a mano e attraverso l'ausilio di tecniche di ripresa particolari. Ma oltre al modo di realizzare i film, che ebbe la svolta esponenziale dagli anni '90 con la computer grafica che permise la realizzazione di modelli tridimensionali sempre più evoluti, ad essere fortemente influenzato fu la parte creativa del cinema, le storie narrate. Negli anni dell'affermazione dei computer, più precisamente nel 1982, la Disney coglie l'occasione di sfruttare la fama crescente delle macchine, passate da strumento futuristico a oggetto casalingo, per realizzare il film "Tron"⁹, film *cult* in cui il protagonista finisce all'interno di un computer. L'anno dopo esce "Wargames – Giochi di guerra", dove il protagonista, un ragazzo adolescente, iniziando una partita ad un videogioco sul suo computer finisce per "hackerare" il sistema della difesa militare americana. Quest'ultimo in particolar modo fece talmente parlare di sé che l'allora presidente degli Stati Uniti, Ronald Reagan, rimase colpito e si interrogò sull'effettiva possibilità che ciò che il film mostra possa essere messo in pratica e realizzato. I suoi allora collaboratori ne rivelarono l'incredibile verosimiglianza e la presenza di falle di sicurezza nel sistema informatico che portarono a un nuovo ordine esecutivo del presidente, noto come NSDD – 145. Questo rese più elevato il grado di sicurezza dell'arsenale strategico americano¹⁰. Il fenomeno dei computer arrivava quindi, attraverso lo strumento del cinema, fino alla tutela della sicurezza nazionale.

Superato il decennio degli anni '80, ed entrando negli anni '90 i computer subiscono un ulteriore slancio in avanti, che con l'introduzione della rete ha subito un'impennata esponenziale che ancora oggi continua ad essere una delle direttrici dell'innovazione. La rete era passata attraverso i sistemi ARPANET, rete di computer creata dalla difesa americana che fu proprio l'oggetto su cui vennero svolte le ricerche sulla sicurezza mosse dalla curiosità generata dal film Wargames, passando poi per i BBS, *bulletin board system*, erano un sistema di bacheche elettroniche che permettevano la trasmissione di file tra computer. Nati a fine anni 70, permisero di ottenere maggiore popolarità ai computer dell'epoca e vengono considerati come una prima forma di Internet, tra le più vicine a quella a noi più nota. Esse erano alla base delle prime comunicazioni telematiche

⁸ Best Selling desktop computer, Guinness World Records, August 1982; <https://www.guinnessworldrecords.com/world-records/72695-most-computer-sales>

⁹ Film anni '80: Wargames, Tron e il computer al cinema, Pietro Ferraro, 24 settembre 2017, <https://www.cineblog.it/post/762646/film-anni-80-wargames-tron-e-il-computer-al-cinema?amp>

¹⁰ <https://www.corriere.it/extra-per-voi/2016/02/27/i-wargames-cinema-che-ispirono-reagan-resero-onnipotente-l-nsa-42d98750-dd79-11e5-8232-4d06db738c6b.shtml>

ma con la nascita del web qualche anno più tardi persero popolarità e vennero abbandonate o si convertirono in *server provider* o siti *web*. Si arriva poi al 1991, con la nascita del *WWW*, *World Wide Web*, che permise l'uso a pieno regime e la diffusione mondiale, per così dire, di Internet. Da lì a poco nacquero i primi siti Internet, nel 1991 nacque il primo in assoluto, il sito del CERN¹¹. Dal 1993 arrivarono i browser. Da quel momento la storia ha preso il percorso che tutti conosciamo.

In quei 10 – 15 anni tra il 1977 e il 1993 il computer negli Stati Uniti passa da oggetto di fantasia e tecnico a strumento per tutti commercializzato in massa dalle società. Esempio di questa transazione radicale è lo spot commerciale di Apple realizzato nel 1984, noto appunto come “1984”¹², allusione al celebre romanzo di George Orwell, che presenta una ragazza che va a scagliare il computer prodotto da Apple noto come “Macintosh”, prodotto in quell'anno, andando a rompere uno schermo che comunicava alle masse, simboleggiando l'abbandono e “rottura” dello status quo.

1.2 Regno Unito, *trait d'union* tra i due mondi

Il primo paese tra quelli presenti nell'area europea ad andare nella stessa direzione che stavano percorrendo gli Stati Uniti in quegli anni fu il Regno Unito. Come accadeva per molti altri settori, di qualunque tipo, i due paesi erano spesso influenzati in modo reciproco con mode che spesso partivano da un paese e si distribuivano ed evolvevano in nuove forme territoriali o addirittura in forme che tornavano poi indietro compiendo il percorso inverso. Un esempio di quanto descritto sono i generi musicali e gli artisti, le cosiddette *rockstar*. Già dai tempi dei *Beatles*, ma anche prima, le mode iniziavano a diffondersi attraversando l'oceano, grazie alle radio, ma anche ovviamente, grazie alle radici culturali simili che spesso hanno inciso nel corso della storia. Il fatto di utilizzare la stessa lingua è stato chiaramente determinante, soprattutto per quel che riguarda la diffusione degli home computer e personal computer, nonché dei linguaggi di programmazione e dei software che ovviamente erano scritti a partire dall'inglese e non necessitavano quindi di una traduzione o di essere interpretati in altre lingue. Questo ha permesso in oltre lo svilupparsi di nuovi mercati che potevano partire da ciò che era già stato prodotto e arrivava dagli Stati Uniti. Settori che in particolar modo divennero estremamente redditizi e di primaria rilevanza furono quello della grafica e soprattutto quello videoludico, che nel Regno Unito stabilì per molti anni la sua sede principale per lo sviluppo e che tuttora rimane una delle culle di questo settore. Nel decennio partito dal 1980, in particolar modo, per seguire negli anni successivi sono nate e si sono affermate grandi e medio-piccole realtà provenienti dal Regno Unito, alcune in grado anche di contrastare il dominio delle aziende americane che erano, come visto in precedenza, partite in leggero anticipo e possedevano il vantaggio del grande mercato interno. Esempio di imprese che sono poi diventate capisaldi del settore per anni fu la *Sinclair Research*, società fondata nel 1961 e che dalla fine degli anni '70

¹¹ 30 anni fa il primo sito Internet: una rivoluzione giovane che ha già fatto la storia, 6 agosto 2021, <https://consigli24.ilsole24ore.com/30-anni-fa-il-primo-sito-internet/>

¹² <https://www.agi.it/economia/news/2021-01-24/apple-macintosh-1984-11133175/>

inizio a rivoluzionare il settore informatico, ponendosi in particolar modo in Inghilterra come principale antagonista di Commodore e Apple, grazie a modelli di home computer come *Sinclair ZX80*, uno dei primi a vendere numeri di unità più elevati e che contribuì alla diffusione nelle case inglesi dei computer, ma soprattutto il *Sinclair ZX Spectrum*, vero campione della Sinclair. Esso fu il computer britannico più venduto della storia e fu il principale antagonista del già citato Commodore64, battagliando in particolar modo nel mercato europeo dove riusciva a farsi strada grazie alle dimensioni ridotte ma soprattutto grazie al prezzo meno elevato.



Figura 1.2: Pubblicità commerciale del home computer ZX Spectrum, (Fonte: <https://zxspectrum.hal.varese.it/publicita.htm>)

Ma dove il Regno Unito, in particolar modo l’Inghilterra, riuscì a creare un fenomeno unico e di incredibile importanza fu l’ambito educativo. L’Inghilterra fin dai primi anni di diffusione delle nuove macchine moderne, iniziò un processo, graduale in principio e via via sempre più deciso, di inserimento e implementazione dei computer all’interno delle scuole, degli ambiti accademici, integrandoli ai tradizionali metodi di apprendimento e insegnando soprattutto all’uso del computer come macchina e come strumento per apprendere e svolgere funzioni.

1.2.1 Dalle teorie e i primi modelli di computer...

Affermare che l’Inghilterra conobbe i computer e i suoi usi soltanto dopo la diffusione del suddetto macchinario negli Stati Uniti non è del tutto corretto. Di fatti è bene spiegare come sia proprio dall’Inghilterra che la storia del computer prenda inizio, per poi ritornare solo dopo aver avuto una “contaminazione” del “paese a stelle e strisce”. Se pur vero che i computer come fenomeno commerciale si sviluppano e prendono piede negli Stati Uniti, è altrettanto vero che i primi macchinari, definibili come computer, della storia nascono dalle menti degli scienziati, ingegneri e inventori britannici. Si pensi ad esempio ad *Alan Turing* e alla sua omonima macchina (*Macchina di Turing*). A lui dobbiamo molto dell’informatica moderna e contemporanea, in quanto ritenuto uno dei padri della scienza informatica e dell’intelligenza artificiale, da lui già teorizzate nel corso degli anni Trenta del ventesimo secolo.

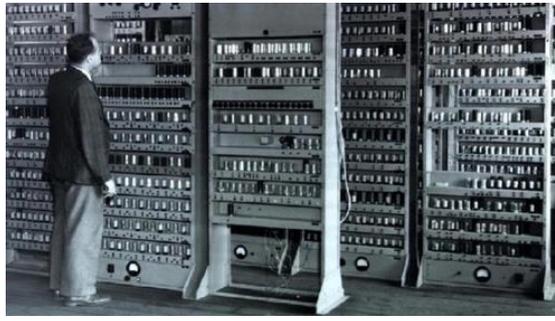


Figura 1.3: Immagine rappresentante la macchina di Turing, (Fonte: <https://novingenio.altervista.org/la-macchina-di-turing/>)

Allora, come mai non avvenne in Inghilterra il processo tecnologico e commerciale che ha portato il computer da oggetto sconosciuto ai più e altamente complesso a passatempo e strumento di lavoro per i giovani e gli adulti? La risposta la troviamo in ciò che è stato descritto nei paragrafi precedenti. L’Inghilterra, rispetto agli stati uniti, viveva un periodo di uscita e ripresa dalla guerra, seppur vinta, completamente differente da quello vissuto dagli USA. Basti solo pensare alle ricostruzioni messe in atto a seguito dei bombardamenti, che richiesero del tempo e che rallentarono i tempi per la ripresa della produzione che nei 15-20 anni successivi alla fine del conflitto portarono ai componenti dei primi prototipi di microcomputer. Inoltre, l’Inghilterra e tutto il Regno unito non avevano la possibilità di mettere mano su materie prime per produrre tali componenti, ne avevano le possibilità di reperirle tanto facilmente quanto gli Stati Uniti, che molto ricavavano dai propri Stati. Risulta quindi lampante come la miscela di situazioni storico-politico-economiche presente degli Stati Uniti rese possibile quello che probabilmente altrove avrebbe richiesto decisamente più anni. Eppure, come in precedenza affermato la storia del computer ha proprio in Inghilterra alcune delle sue tappe più importanti, non solo per il suo percorso evolutivo ma anche per l’intreccio che esso ha da sempre avuto con la storia dell’uomo, sin dalle sue prime forme più rudimentali. In questo senso non si può non citare il *Colossus*, uno dei primi computer programmabili della storia. Esso fu costruito nel 1944 dagli inglesi per essere utilizzata dalla *Royal Navy*, la quale sfruttò le sue potenzialità per riuscire a decifrare i messaggi cifrati dei nazisti, in particolar modo per decodificare la corrispondenza tra Adolf Hitler e i suoi capi di stato maggiore¹³. Il progetto era talmente segreto che a fine della guerra la macchina fu smantellata e i relativi progetti dati alle fiamme su ordine di *Winston Churchill* e dei servizi segreti¹⁴. Della sua esistenza siamo venuti a conoscenza solo negli anni '70 del 1900¹⁵.

¹³ Bozzo Massimo, *La grande storia dei computer*, Edizioni Dedalo, Bari.

¹⁴ Simon Sigh, *Codici & segreti*, Rizzoli Editore, Milano, 1999, pag. 250

¹⁵ Prenel, Bart, “Colossus and the German Lorenz Cipher”, *Advances in Cryptology - EUROCRYPT 2000: International Conference on the Theory and Application of Cryptographic Techniques Bruges, Belgium May 14-18, 2000, Proceedings*, Lecture Notes in Computer Science, Springer, p. 417,

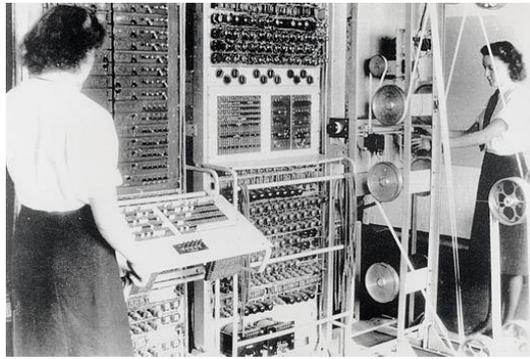


Figura 1.4: Immagine del computer “Colossus”, modello “Mark 2”, (Fonte: <https://it.wikipedia.org/wiki/Colossus>)

1.2.2 ...agli usi didattici e la concorrenza ai produttori Statunitensi

Andando avanti di qualche anno nel tempo arriviamo agli anni '60-'70, periodo in cui come già discusso, la rivoluzione dell'industria dei computer inizia a muovere i suoi passi nei paesi americani, in particolar modo negli Stati Uniti ma anche nel Canada. I Produttori inglesi non rimasero però a guardare mentre la rivoluzione informatica prendeva piede e seppur con qualche anno di ritardo, iniziarono ad arrivare le prime macchine di produzione britannica. Su tutti ricordiamo la già citata in precedenza Sinclair Ltd., una società nata nel 1961 come produttrice principalmente di calcolatrici digitali, tratto caratteristico comune con molte delle aziende statunitensi, ma che sul finire degli anni '70 uscì con un primo modello di microcomputer, venduto in kit da assemblare, noto come MK14. Questo modello ebbe degli ottimi riscontri in termini di unità vendute, dimostrando come esistesse un interesse per questo prodotto e un suo relativo mercato. Questo condusse al citato Sinclair ZX80, modello che fu molto popolare e che per un prezzo di 100£ (230\$) permetteva di entrare in possesso di una macchina dall'elevato quantitativo di tecnologia. Questo modello permise alla società di consolidarsi come concorrente di livello nel mercato degli home computer. La Sinclair si dimostrò inoltre una società all'avanguardia per i tempi in termini di vicinanza al cliente, accogliendo molte delle critiche mosse ai vari modelli aggiornati dei suoi macchinari per produrre quello che sarà forse il loro prodotto più famoso, lo ZX Spectrum. Questo computer ottenne risultati nelle vendite ottimali, specie nel mercato interno dove si impose come il principale home computer delle famiglie del regno unito. Purtroppo, nel corso del decennio '80 -'90 l'evoluzione tecnologica che ebbe luogo negli Stati Uniti permise ai concorrenti di surclassare i computer inglesi, che però giocavano molo della loro concorrenza su un prezzo altamente accessibile al pubblico. Questa enorme diffusione permise la nascita di molte case sviluppatrici di videogiochi che usarono come standard le tecnologie dei modelli prodotti dalla Sinclair Ltd., così come fecero molti altri sviluppatori. Nello stesso periodo, insieme alle imprese private, ad accorgersi del potenziale delle nuove tecnologie informatiche fu il governo inglese. Tra gli anni '70 e inizio anni '80 nel Regno Unito crebbe un senso di preoccupazione per un maggiore allineamento tra istruzione e industria. Ciò era motivato da tre principali fattori: la natura dell'industria era in mutamento e questi cambiamenti dovevano essere compresi sia dagli insegnanti che dagli studenti, le tecnologie sviluppate erano sempre più presenti nelle pratiche industriali e commerciali e i loro usi dovevano essere compresi e considerati, inoltre l'uso del computer nelle scuole doveva

essere considerato da un punto di vista del potenziale commerciale ed economico¹⁶. All'inizio degli anni '70 nel Regno Unito i computer in uso nel settore dell'istruzione erano pochi e per lo più ubicati in strutture universitarie, dove il loro utilizzo da parte di studenti e insegnanti era limitato, per lo più erano strumenti di cui vantarsi di esserne possessori. Nel 1973 arriva però la svolta. Il *National Development Programme in Computer Assisted Learning (NDPCAL)* iniziò la gestione di oltre 30 progetti nel settore dell'istruzione nel Regno Unito per potenziali applicazioni dell'apprendimento assistito dal computer, con un budget di 2,5 milioni di sterline in 5 anni. Si avviò un periodo di presa di coscienza del fenomeno che portò all'avvio di diversi ulteriori progetti e che portò prima, sul finire degli anni '70, alla produzione di microcomputer per le scuole da parte della *Research Machines*, per arrivare poi all'intervento del Dipartimento del Commercio e dell'Industria nei primi anni '80 per rendere possibile l'acquisto di sistemi informatici a costi meno elevati. Quest'iniziativa prese il nome di "*Micros in Schools*". Il primo problema chiaramente era collegato al fatto che gli insegnanti non erano a conoscenza di come utilizzare tali macchine, ma ben presto si interessarono non solo a capirne l'uso ma soprattutto ad esplorarne le possibilità di utilizzi congiunti insieme agli studenti. I primi a farne uso furono ovviamente i matematici in quanto vi era poca distanza tra codifica di software e applicazione di essi. Ma presto anche gli altri insegnanti compresero i benefici che le loro materie potevano trarre da tali macchine.

Come in precedenza indicato, vi era anche un interesse industriale e commerciale nei computer nelle scuole, questo fece sì che un ulteriore dipartimento governativo venisse coinvolto. Nel 1983 l'Iniziativa Educativa tecnica e professionale fu finanziata dal Dipartimento del Lavoro con un ammontare di circa 20 milioni di sterline. Obiettivo era sviluppare i curriculum e l'integrazione di approcci progettuali e tecnologici, inizialmente indirizzati a studenti delle scuole secondarie superiori¹⁷. Questo permise l'acquisto di macchinari quali computer, e strumenti di produzione controllati dai computer, che venivano stipati nelle cosiddette "Aule Informatica". Ma da un iniziale interesse ad applicare tali tecnologie ad aree come l'ingegneria, la grafica e il design, col tempo si arrivò a voler estendere l'applicazione anche a materie delle scuole primarie. Studenti e insegnanti delle scuole primarie dovevano essere coinvolti per renderli consapevoli dei cambiamenti tecnologici e delle loro implicazioni future. Nel 1984 il numero dei computer presenti nelle scuole era molto elevato, tale da creare una domanda e da portare al finanziamento di servizi di formazione degli insegnanti in IT, tecnologia dell'informazione. Si arrivò quindi nel 1989 a istituire nel Regno Unito, una politica che sanciva la creazione di un curriculum nazionale. L'informatica non era indicata come area separata ma era contenuta nella tecnologia del design; tuttavia, l'uso dei computer era fortemente indicato e adoperato in tutte le aree del curriculum disciplinare. Tutto questo processo favorì ulteriormente sotto un punto di vista commerciale i

¹⁶ Passey, D. (2014). Early Uses of Computers in Schools in the United Kingdom: Shaping Factors and Influencing Directions. In: Tatnail, A., Davey, B. (eds) *Reflections on the History of Computers in Education*, Springer, Berlin

¹⁷ Passey, D. (2014). Early Uses of Computers in Schools in the United Kingdom: Shaping Factors and Influencing Directions., pp.135 In: Tatnail, A., Davey, B. (eds) *Reflections on the History of Computers in Education*, Springer, Berlin

produttori inglesi di computer. Tra questi, oltre alla già citata Sinclair c'era anche la *BBC*, famosa emittente radiotelevisiva britannica. Questa produsse diverse macchine che furono molto popolari nelle scuole e in generale sul suolo inglese che contribuirono all'alfabetizzazione informatica dei cittadini. Su tutte, la più conosciuta e forse più importante fu il *BBC Micro*.

1.2.3 *BBC Micro*

In linea con la tendenza del periodo, ovvero accrescere il livello di conoscenze informatiche e le capacità dei cittadini, messo in atto dalle varie istituzioni del Regno Unito, la *BBC*, *British Broadcasting Corporation*, avviò un programma noto come “*BBC Computer Literacy Project*”, traducibile come “Progetto BBC per l'alfabetizzazione informatica”. La BBC decise di avviare questo progetto in risposta a quanto presentato e illustrato in un documentario diffuso dalla stessa BBC noto come “*The Mighty Micro*”, dove il dottor Christopher Evans, studioso del National Physical Laboratory, predisse l'avvento della rivoluzione informatica, la nascita e l'affermazione dei micro e home computer nella società e l'impatto che la rivoluzione avrebbe avuto su economia, industria e stile di vita del Regno Unito. Il progetto prevedeva lo sviluppo e la messa in commercio di un microcomputer che potesse svolgere numerose operazioni, le quali sarebbero state mostrate e insegnate in un programma TV chiamato “*The Computer Programme*”. All'interno del programma venivano illustrate come usare il computer per programmare, fare computer grafica, produrre suoni e musica, accedere ed utilizzare il *Teletext*, in Italia noto come Televideo, ma anche come utilizzare attraverso il computer periferiche esterne o accenni di intelligenza artificiale. Il programma, essendo veicolato attraverso la televisione, permetteva a chiunque fosse in possesso del prodotto di imparare a utilizzarlo e implementarlo nella routine della sua vita. Elemento che permise il successo di questa trasmissione fu il fatto di presentare all'inizio di ogni puntata un episodio di vita quotidiana dove i computer potevano essere o erano applicati. Talvolta venivano rappresentate anche scene possibili sul posto di lavoro. Lo scopo era quindi rendere consapevoli gli spettatori degli usi del computer ma soprattutto rendere il suo uso parte integrante delle capacità in possesso di ogni cittadino in modo che nessuno potesse trovarsi indietro rispetto ai tempi e potesse integrarsi con le nuove realtà lavorative, commerciali e scolastiche.



Figura 1.5: “*The Computer Programme*” condotto da Chris Serle ed Ian McNaught-Davis, (Fonte: <https://www.bbc.co.uk/programmes/p03064m1>)

Successivamente dopo aver deciso come si sarebbe strutturato il progetto, la BBC decise di indire un bando di gara per stabilire chi avrebbe prodotto il computer targato BBC. Inizialmente ascoltò le proposte della ben nota Sinclair, che però si presentò con un'offerta che venne ritenuta non adatta in quanto non all'altezza delle aspettative tecniche che la BBC richiedeva per il suo prodotto. Dopo una serie di proposte scartate, la BBC decise di accordarsi con *Acorn*, società britannica che all'epoca stava sviluppando dei progetti per computer da mettere sul mercato che vennero utilizzati nella produzione della macchina della BBC, in particolar modo su un modello noto come "*Proton*". Il prodotto entrò in vendita con il nome di "*BBC Micro*" appunto, a cavallo tra il 1981 e 1982. Divenne subito molto popolare, a tal punto che la domanda superava di gran lunga l'offerta, questo causò alcuni disagi nei primi mesi di messa in vendita. Esso riuscì perfettamente nella sua missione di posizionarsi come computer per apprendimento e alfabetizzazione digitale, visto l'ottimo risultato ottenuto nelle scuole, dove era molto popolare, circa l'ottanta per cento delle scuole del Regno Unito ne possedevano un'unità. Seppe imporsi anche per la resistenza dei materiali adoperati che erano in grado di resistere ai danni causati dai bambini più piccoli. Il prodotto fu reso disponibile in due configurazioni che differivano in alcune caratteristiche tecniche e hardware, ma soprattutto nel prezzo. Questo era comunque più alto rispetto ad altri per riuscire ad imporsi anche nel mercato retail e Acorn per questo, sviluppò una versione per l'uso a casa, complementare a quello scolastico del BBC Micro, venduto con il nome di "*Acorn Electron*"¹⁸. I risultati ottenuti furono incredibili. I bambini sin dalle scuole primarie iniziarono a adoperare i computer come strumento di lavoro e come passatempo, coadiuvati nello scopo dalla trasmissione televisiva già citata che ottenne un elevato seguito. Grazie alle versioni differenti riuscì a imporsi nel settore e ad ottenere una quota di mercato molto alta che fu mantenuta lungo il decennio degli '80 così come negli anni '90. La sua incredibile fama nonché presenza nelle case e nelle scuole non ebbe modo però di replicarsi al di fuori del Regno Unito. Furono diversi i tentativi per provare ad esportare all'estero, specialmente negli Stati Uniti, il prodotto di Acorn e BBC, ma la competizione agguerrita di prodotti già noti e ben radicati nel mercato come, ad esempio, Apple II nel segmento USA, non lasciarono scampo al home computer britannico. Inoltre, risultava ben più difficile istruire i consumatori ed esportare il prodotto televisivo all'estero, di fatto la strategia era pressoché di natura commerciale e non ebbe fortuna come nel paese d'origine. Il programma di alfabetizzazione informatica di BBC però, grazie anche ai contemporanei sforzi dei vari organi di governo inglesi dell'epoca, ottenne risultati incredibili. Riuscì nell'intento di rendere i computer uno strumento noto e d'uso comune e formò molti dei futuri membri della classe informatica inglese degli anni successivi. Fu il primo fenomeno del genere messo in atto e uno di quelli dalla portata maggiore, tale da essere contestata da molti produttori privati del periodo che non vedevano di buon occhio la cosa. L'accusa era mossa verso il Dipartimento dell'Industria per aver reso la BBC uno strumento per attuare le politiche governative e foraggiare indirettamente l'industria, cosa per cui il governo britannico si mostrava pubblicamente contrario. Ovviamente la mossa dell'emittente

¹⁸ "Electron set for battle at £199". *Acorn User*, pp.5. Giugno 1983

britannica fu copiata e tentata da altri network che però non ebbero lo stesso riscontro in termini di vendite e di successo mediatico. Molti paesi del Commonwealth, osservando il progetto di alfabetizzazione informatica del Regno Unito si dotarono successivamente di un numero di unità discreto di BBC Micro e di programmi propri, sulla falsa riga dell'originale.



Figura 1.6: *BBC Micro*, (Fonte: https://it.wikipedia.org/wiki/BBC_Micro)

1.3 Europa: i “*first-follower*”

Come visto per il Regno Unito, molti paesi dell'Europa sono stati influenti nella storia dell'informatica prima della fase decisiva che permise loro di divenire oggetto d'uso comune. Il “Vecchio Mondo” ha avuto un ruolo di spicco nella nascita di progetti definibili come archetipo del computer. Purtroppo, la distruzione delle due guerre affrontate nei propri confini, con tutte le problematiche economiche affiliate e soprattutto la mancanza di risorse a basso costo e delle conoscenze tecniche adeguate ha reso impossibile quel processo evolutivo della tecnologia che ha caratterizzato come già visto gli USA. Ma questo non significa che alcuni paesi Europei non abbiano rivestito una parte, anche di rilievo, nel decennio che ha visto l'abbandono del computer tecnico in favore del Home computer. Germania e Italia, ad esempio, sono due paesi che hanno recitato parti diverse ma di spessore, con l'Italia protagonista soprattutto grazie a un'impresa capace di risultati straordinari prima di un decadimento dal sapore amaro, la “*Olivetti S.p.A.*”.

1.3.1 *La Germania: culla della produzione hardware*

La Germania come molte, impegnata in entrambe le guerre, fu decisiva nel fornire molte delle tecnologie alla base dei primi grandi computer, oltre a fornire molte delle teste in grado di ideare le invenzioni del futuro del settore. Impossibile non citare *Konrad Zuse*, informatico tedesco che per rendere i numerosi calcoli e procedimenti richiesti nello sviluppo di veicoli aerei, prodotti per la Germania nazista, si dedicò allo studio informatico per arrivare alla produzione, nel 1936, dello storico e famoso *Z1*. Esso era un computer meccanico, famoso per la sua capacità di eseguire istruzioni solo se impartite attraverso schede forate. La macchina fu progettata a casa dei genitori, fatto che curiosamente ricorda la nascita di un altro computer rivoluzionario, l'Apple I. Fu il primo di una serie di computer, tutti denominati con la lettera Z, dal cognome dell'inventore e una cifra ad affiancarla, per dare ai modelli un ordine cronologico. Nonostante un avvio al quanto brillante, la Germania come molti paesi accusò una battuta d'arresto importante, successivamente alla seconda guerra, nel campo dell'innovazione tecnologica che se prima andava ad accompagnare gli sforzi bellici in quanto serviva

peer ottenere vantaggi in termini di migliori armi e veicoli e equipaggiamenti tecnici, nel periodo successivo fu una delle vittime della forte crisi economica. Qualche anno più avanti, più precisamente a partire dal decennio degli anni '80, la Germania tornò ad avere un ruolo nel sempre più importante mondo dell'informatica. In particolar modo la Germania fu uno dei paesi Europei dove i diversi produttori esteri vennero a stabilirsi per la produzione in loco dei propri prodotti da esportare nel mercato europeo. Per lo più si trattava di stabilimenti di assemblaggio di componenti per l'ottenimento del prodotto finale, sfruttati dalle società per evitare di dover esportare un prodotto finito avvalendosi di intermediari e affrontando costi maggiori. Una delle società più importanti a svolgere tale azione fu la Commodore, che in Germania produsse molti dei suoi prodotti esportati in tutta Europa. Non a caso a partire dai modelli "VIC-20", in Germania "VC-20" per evitare pronunce equivocabili, grosse percentuali delle vendite avevano luogo in Europa, specialmente per quanto riguarda il modello "Amiga" che ottenne il 70% della quota di mercato nel Vecchio continente, con il primato non a caso in Germania. Questo processo di espansione globale delle imprese, definibile come internazionalizzazione, permise ai paesi europei di entrare in contatto con i nuovi home computer e di avvicinarsi al mondo occidentale americano.

1.3.2 Italia: Olivetti S.p.A., gli anni '80 e il fenomeno USA

In Italia la situazione di partenza è molto simile a quella in cui si trovavano molti altri paesi europei. Il numero di computer nei primi anni dopo la guerra è esiguo. Le unità sono ingombranti, sono stipate in grandi spazi che solo le università e alcune imprese di medio-grandi dimensioni posseggono e richiedono un uso tecnico che in pochi conoscono. Lo sviluppo come abbiamo visto, coincide con il momento esatto in cui i costi di produzione divengono inferiori al valore produttivo ed economico generato. Questo ha inizio negli anni già dagli anni '60, come in precedenza spiegato, dove gli innovatori hanno come tratto caratteristico quello di essere quasi tutti passati prima per la produzione di calcolatrici. Tra le aziende che in questo settore hanno compiuto maggiori apporti all'innovazione tecnologica troviamo la Olivetti S.p.A., società italiana fondata nel 1908, che nel 1965 rivoluzionò completamente il settore facendo uscire il modello denominato *Olivetti Programma 101*¹⁹. Essa era una calcolatrice programmabile progettata per essere usata da chiunque per differenti mansioni.



Figura 1.7: Calcolatrice Olivetti Programma 101, (Fonte: https://it.wikipedia.org/wiki/Olivetti_Programma_101)

¹⁹ Marcello Zane, "Storia e memoria del personal computer. Il caso italiano", Museo dell'Industria e del Lavoro "Eugenio Battisti", Brescia, pp.3

Purtroppo, da quello che poteva essere l'inizio di un'evoluzione tecnologica "made in Italy" ne deriva invece un periodo di stagnazione, figlio in parte di un atteggiamento refrattario all'innovazione, dovuto per lo più a una non conoscenza del computer che causava un'avversione verso il suo utilizzo. La rivoluzione informatica inoltre correva a una velocità impressionante che stupì chiunque a tal punto, che l'Italia si trovò nel giro di pochi anni drasticamente indietro dal punto di vista tecnologico. Ma sul finire degli anni '70 e con l'avvento degli anni '80 la tendenza a respingere il "Nuovo" cambia. I primi computer erano poco diffusi e visti in modo ambiguo dai lavoratori, con il diffondersi, specialmente nel mondo americano, dei personal e home computer invece si ha un processo inverso. Nel 1978 i personal computer in Italia sono poco meno di mille, per l'esattezza 820. L'anno dopo, nel 1979, saranno più del doppio, ben 2420. Solo nel 1980 il numero dei computer installati si quadruplica, nel corso dei successivi anni la crescita è esponenziale. A fare la differenza è anche il cambio del modello di vendita che deriva a sua volta da quello americano. Attraverso questo i computer vengono indirizzati verso l'immagine di elettrodomestico o macchina d'ufficio, allontanandosi dall'immagine di oggetto misterioso pesante che aveva in precedenza. Essi verranno venduti non tramite reti di vendita dirette e agenti di rappresentanza ma attraverso negozi al dettaglio, esattamente come accadeva negli "States"²⁰. L'inizio è chiaramente in salita, con un numero di rivenditori esigui che andrà via via in crescendo. Ovviamente a questo si accompagna una sempre più crescente stampa dedicata all'alfabetizzazione informatica, se pur contraddistinta da pubblicità non sempre troppo mascherata e da un tecnicismo ancora eccessivo. Questa svolse un ruolo non da poco nell'istruire i cittadini alla programmazione, ma ebbe anche un ruolo di consigliere nel processo d'acquisto fornendo recensioni dei vari macchinari testati dagli esperti. Nel corso quindi del decennio '80-'90, le rivoluzioni in campo informatico che si imponevano nel modello americano venivano, se pur con qualche anno a volte di ritardo, importate e implementate nel modello italiano. Basti pensare alla diffusione dei dischi di lavoro che rendevano compatibili computer domestici e quelli sul post odi lavoro, che permisero una diffusione enorme di unità domestiche negli Stati Uniti, per un uso prettamente lavorativo e non videoludico, che nel corso di qualche anno divennero un prodotto ampiamente diffuso anche in Italia. In Paese si diffonde quindi una tendenza ad importare e fare propri i modelli e le innovazioni che fuoriuscivano dagli Stati Uniti.

1.4 Conseguenze di una diffusione planetaria

Che sia quindi dovuta a fenomeni interni che hanno reso possibile l'avvento e il generarsi di nuove tecnologie e innovazioni, o figlia di una volontà di omologarsi per ragioni tecniche, evitare l'arretratezza tecnologica che avrebbe generato abissi dei livelli produttivi, o per ragioni culturali, la diffusione degli home computer prima e dei personal computer poi è stata repentina e ha coinvolto completamente, o quasi, tutto il mondo occidentale

²⁰ Marcello Zane, "Storia e memoria del personal computer. Il caso italiano", Museo dell'Industria e del Lavoro "Eugenio Battisti", Brescia, pp.45

e gran parte di quello orientale. Questo ha chiaramente determinato degli atteggiamenti di risosta da parte di soggetti privati e pubblici che hanno dovuto adeguarsi alle nuove tendenze, con l'obiettivo di renderle parte integrante del loro vivere e delle loro mansioni. Gli stati hanno re-immaginato quello che era il modo di concepire l'istruzione, il modo di fornire servizi ai cittadini e quindi di svolgere le loro funzioni, così come i privati hanno fatto con il fare impresa. Non è un caso che con il diffondersi delle macchine informatiche anche gli stati abbiano iniziato un processo di informatizzazione delle loro funzioni e delle loro diverse estensioni. L'esempio citato del caso inglese di BBC Micro è un chiaro segnale di come anche le istituzioni hanno saputo cogliere, chi prima chi dopo, un cambiamento che avrebbe radicalmente modificato lo status quo dell'epoca. Ne consegue che sono diversi i progetti che essi hanno nel corso del tempo avviato e diverse sono state le risorse impiegate. Tutt'ora il processo è in corso d'opera e tiene fermamente banco il tema di svolgere programmi di alfabetizzazione informatica che permettano ai cittadini di integrarsi con i nuovi processi e le nuove tecnologie, per evitare che qualcuno possa restare indietro.

CAPITOLO II

I BIG DATA

2.2 Big Data: quando i dati fanno la differenza

L'ondata di innovazione portata dalla rivoluzione informatica ha portato alla luce nuove possibilità, intese come nuovi modi di svolgere attività già esistenti, nuove attività prima non realizzabili o mai immaginate. Si parla quindi di nuovi mercati, nuovi lavori, nuove modalità di svolgimento dello stesso lavoro in modo più efficace ed efficiente. La ricerca di efficacia e di efficienza, che si cerchi una sola delle due o entrambe, è spesso il motivo alla base dell'implementazione delle innovazioni informatiche nel mondo del lavoro e nell'economia.

L'informatica, proprio come fa un essere vivente, nel corso del tempo si è evoluta, è cresciuta e così come trattato nel precedente capitolo, è passata dai computer dalle dimensioni "colossali" ai microcomputer, per poi mutare verso gli home computer e i personal computer, i cosiddetti *PC*. Ma non si è fermata lì. Nel tempo sono arrivate la rete e le connessioni internet, successivamente la realtà virtuale, le intelligenze artificiali, con tutte le loro declinazioni. Tra le innovazioni più importanti, in termini di impatto sulle vite lavorative e non solo delle persone, ci sono sicuramente i *Big Data*. Questi come molti altri, appartengono all'universo di termini di provenienza anglosassone di cui si sente maggiormente parlare, spesso non comprendendo a pieno cosa siano e quali siano le loro potenzialità. Risulta quindi importante capire cosa siano effettivamente i big data, dandone una definizione quanto più corretta possibile prima di sapere come si è arrivati a tale concetto e all'uso che oggi ne viene fatto. Nel corso del seguente capitolo si cercherà quindi di offrire una definizione dei Big Data, analizzando cosa sono, la loro storia e il loro funzionamento.

2.1.1 Definizione: cosa sono i Big Data

Prima di parlare di Big Data, di conoscere la loro storia e il loro utilizzo, occorre necessariamente capire di cosa si tratti. Quando parliamo di Big Data facciamo riferimento a una grossa mole di dati, talmente estesa in termini di volume, velocità e varietà da rendere necessario lo sviluppo e l'implementazione di tecnologie e metodi in grado di estrapolare informazioni e valore da questi. Non si fa quindi propriamente riferimento ai dati, bensì alla capacità di utilizzare, analizzare e mettere in relazione enormi quantità di dati eterogenei per scoprire la correlazione tra fenomeni di natura diversa, ma collegati da una qualche forma di legame, nonché per prevedere fenomeni futuri. Questa definizione ci fornisce gli elementi chiave dei Big Data²¹:

- *Informazione*: essi sono dei dati da cui vogliamo estrarre un contenuto informativo

²¹ <https://www.apogeeonline.com/articoli/impariamo-a-conoscere-i-big-data-andrea-de-mauro/>

- *Tecnologie*: per estrarre valore e informazioni abbiamo la necessità di sviluppare e adoperare nuove tecnologie
- *Metodi*: per raggiungere lo scopo finale implementiamo metodi innovativi
- *Impatto*: essi hanno un forte impatto sulla vita delle persone e sul lavoro, ponendosi come obiettivo la creazione di valore.

McKinsey & Co., famosa società di consulenza strategica su scala globale, nel 2011 ha dato una definizione di Big Data, definendo: «Un sistema di Big Data si riferisce a dataset la cui taglia o volume è talmente grande che eccede la capacità dei sistemi di database relazionali di catturare, immagazzinare, gestire ed analizzare». Da questa definizione capiamo che possiamo parlare di Big Data quando non solo la mole di dati cresce, aumentando quindi la capacità del sistema di archiviare quelle che sono le informazioni già conosciute e quelle che verranno aggiunte, ma migliora anche la velocità di informazioni che il sistema è in grado di gestire, nonché acquisire, ogni secondo. Il volume dei dati è correlato alla capacità di immagazzinamento dei dati del sistema, dati che devono essere archiviati nella forma in cui arrivano dalle differenti fonti che vengono utilizzate. Parliamo di dati eterogenei in quanto le fonti stesse utilizzate sono di natura differente. Possiamo avere dati strutturati, come ad esempio informazioni provenienti da database, ma anche dati non strutturati, esempio di questi sono immagini, dati satellitari o più comunemente le informazioni provenienti dai vari social network. Quando parliamo di Big Data non esiste una soglia minima oltre la quale li distinguiamo dal resto. Oggi giorno la mole di dati è nell'ordine degli *zettabyte*, ossia miliardi di *terabyte*, ma come detto in precedenza non distinguiamo tanto sulla base della quantità di dati immagazzinati e trattati quanto per la complessità dei dati trattati, tale che richiede l'implementazione di tecnologie e metodi per processare le informazioni.²² Basti pensare che tra il 2013 e il 2015 abbiamo prodotto più dati che in tutta la storia dell'umanità sino a quel momento, oggi probabilmente riusciamo a generare una simile quantità di dati anche in un singolo anno²³.

MULTIPLI DEL BYTE			
Nome	Simbolo	Multiplo	byte
Kilobyte	kB	10 ³	1.000
Megabyte	MB	10 ⁶	1.000.000
Gigabyte	GB	10 ⁹	1.000.000.000
Terabyte	TB	10 ¹²	1.000.000.000.000
Petabyte	PB	10 ¹⁵	1.000.000.000.000.000
Exabyte	EB	10 ¹⁸	1.000.000.000.000.000.000
Zettabyte	ZB	10 ²¹	1.000.000.000.000.000.000.000
Yottabyte	YB	10 ²⁴	1.000.000.000.000.000.000.000.000

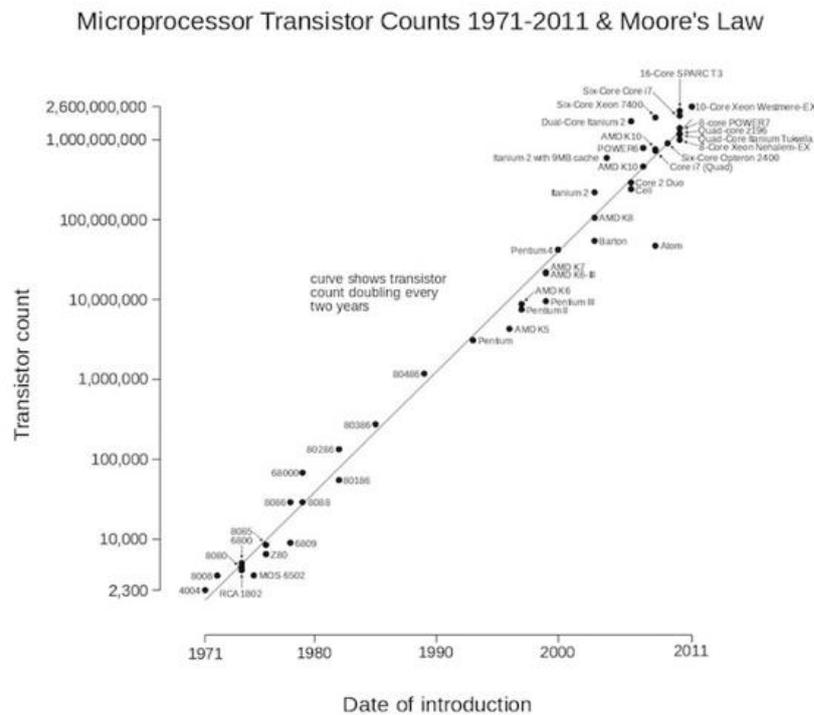
Tabella 2.1: *Multipli del byte*, (Fonte:<https://www.ictpower.it/tecnologia/il-byte-e-i-suoi-multipli-kilobyte-o-kibibyte.htm>)

²² Snijders, C., Matzat, U., Reips, U.-D., "Big Data: Big gaps of knowledge in the field of Internet", International Journal of Internet Science, volume 7, Issue 1, 2012.

²³ <https://www.themarketingfreaks.com/2019/11/big-data-cosa-sono-la-storia-le-caratteristiche-le-analisi-ed-esempi/>

2.1.2 La prima legge di Moore

Parlando sempre della quantità di dati trattati e immagazzinati, è utile citare la *prima legge di Moore*: “La complessità di un *microcircuito*, misurata ad esempio tramite il numero di *transistor* per *chip*, raddoppia ogni 18 mesi”. Questa è la prima di due osservazioni e affermazioni di Moore, uno dei due fondatori di *Intel*, che sono poi divenute leggi dell’informatica nonché principale obiettivo e metro di misurazione delle performance per i principali produttori di microprocessori. Tale legge è dimostrata da casi pratici come i processori *Pentium*, prodotti dalla stessa Intel, che nell’arco di tempo che va dal maggio del 1997 al novembre del 2000, per l’esattezza 42 mesi, sono passati da un numero di transistori pari a 7,5 milioni per il modello Pentium II del 1997 a ben 42 milioni del modello *Pentium 4*. Per l’esattezza la stima che la legge di Moore da del numero di transistori per tale lasso di tempo è addirittura inferiore alla reale crescita, poiché essa stimerebbe per un arco temporale di 42 mesi un numero di transistori pari a circa 38 milioni. Come molte delle leggi dell’informatica, la prima legge di Moore presenta alcuni limiti. Essi consisterebbero prevalentemente in limiti di natura fisica, obbligatori per la riduzione delle dimensioni dei transistori, sotto tali limiti il risultato presenterebbe degli effetti negativi più che positivi. Poiché sono già stati raggiunti per ovviare al problema si è optato per adoperare tecnologie multicore, ossia accoppiamento in parallelo di più processori che quindi operano simultaneamente. Questo ha premesso l’elaborazione di dati sempre più complessi. Questa legge ci permette di affermare che l’evoluzione delle tecnologie consente di immagazzinare e gestire *dataset* le cui dimensioni crescono ininterrottamente nel tempo; quindi, risulta essenziale nel comprendere i Big Data e la loro natura.



2.1.3 Il Modello delle “3V”

Quanto visto fino a questo punto ci permette quindi di dire che si parla di Big Data quando si ha un dataset talmente grande da rendere necessario l’ausilio di strumenti innovativi per estrapolare valore e informazioni. Inoltre, si può affermare che una dimensione di riferimento non esiste in quanto questa, cambia in continuazione in accordo con quanto affermato dalla prima legge Moore.

Nell’ottica di fornire una definizione di cosa sono i Big Data non si può tralasciare il modello delle “3V” fornito da Douglas Laney. Nel 2001 Douglas Laney, analista di mercato, definisce il modello di crescita dei dati come tridimensionale, le tre dimensioni sono appunto le tre V che danno il nome al modello: nel tempo aumenta il *Volume* dei dati, la *Velocità* e la *Varietà* dei dati²⁴. Il modello, nonostante siano passati diversi anni, risulta in molti casi ancora valido, seppur nel 2012 abbia ricevuto un’estensione con l’aggiunta di una quarta variabile prima e di una quinta in seguito, ossia la *Veridicità* e il *Valore* dei dati. Questo è portato a un nuovo modello, non più con tre “V” bensì con cinque. Le tre dimensioni presenti nel primo modello hanno ognuna un proprio significato, che è il seguente:

1. *Volume*: rappresenta la quantità di dati generati per ogni secondo che passa. Questi possono essere strutturati non, come in precedenza già visto. Questo deriva dalle differenti fonti da cui sono generati tali dati. Può trattarsi di dati come, ad esempio, *feed* provenienti da dati di *social network* o da un’*app mobile*. Vista la mole sempre crescente di dati generati oggi, potrebbe sembrare un problema l’archiviazione di questi. Ma ciò non rappresenta un problema reale in quanto *cloud* e virtualizzazione supportano la gestione della grossa quantità di dati disponibili, facilitando quelli che sono i processi di raccolta, l’immagazzinamento e l’accesso a suddetti dati. Una criticità da considerare in relazione al volume è la qualità delle informazioni raccolte.
2. *Velocità*: tale dimensione si riferisce alla velocità con cui si generano nuovi dati; quindi, la velocità con cui si ricevono. Con la crescita del volume dei dati cresce la velocità richiesta per la ricezione e l’analisi di questi; perciò, risulta sempre più costoso e complesso svolgere questi processi. Risulta necessario sviluppare strumenti in grado di svolgere un’analisi rapida che fornisca degli indicatori rilevanti²⁵.
3. *Varietà*: l’ultima dimensione del modello originale fa riferimento ai diversi tipi di dati disponibili. Prima dell’avvento dei Big Data erano considerati solamente i dati strutturati che venivano trattati attraverso l’ausilio di database relazionali. Oggi, per avere analisi più complete e profonde, vengono presi in considerazione anche dati non strutturati e semi-strutturati. Queste tipologie di dati necessitano di un’elaborazione preliminare al fine di riscavare un significato. La differente tipologia di dati rende ancora di più necessaria la formazione di indicatori rapidi e semplici. Un esempio di ciò sono i sistemi di raccoglimenti di feedback degli utenti.

²⁴ <https://www.oracle.com/it/big-data/what-is-big-data/#three>

²⁵ <https://www.themarketingfreaks.com/2019/11/big-data-cosa-sono-la-storia-le-caratteristiche-le-analisi-ed-esempi/>

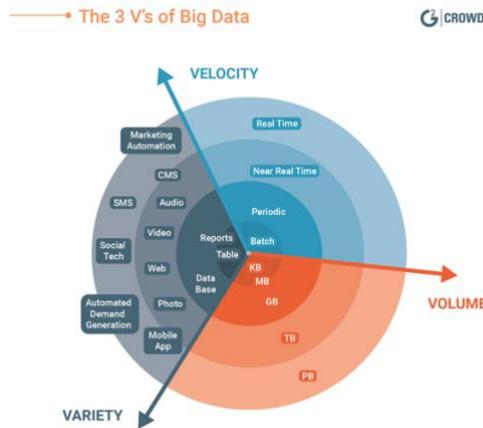


Grafico 2.2: Rappresentazione grafica del modello delle 3 “V”, (Fonte: <https://www.g2.com/articles/big-data>)

Negli ultimi anni il modello originale è stato rivisitato, con la scoperta e l’aggiunta di due nuove “V”, ossia veridicità e valore, queste come le tre originali devono essere interpretate:

4. *Veridicità*: si tratta dell’accuratezza e mutevolezza del significato di un dato a seconda del contesto. Vista la varietà di fonti che possono generare i dati che raccogliamo, oltre che la velocità alla quale essi variano, risulta pressoché improbabile garantire la stessa qualità per i dati in ingresso. Se i dati alla base risultano poco veritieri e accurati, le analisi produrranno risultati che non saranno dei migliori. Poiché questi dati vengono usati per prendere decisioni, talvolta anche di elevata importanza, è opportuno assegnare un indice di veridicità ai dati, affinché si possa avere una misura dell’affidabilità sia di questi che, di conseguenza, dei risultati della loro analisi.
5. *Valore*: rappresenta la qualità dei dati, la capacità di trasformarli in valore. Risulta fondamentale comprendere il valore che i dati possono generare e rappresentare per i business per efficientare le spese di analisi e dei progetti “*data driven*”.

In aggiunta a queste, ulteriori dimensioni, di minore rilevanza, sono state considerate. Generalmente a seconda della situazione e di chi compiva l’analisi, le dimensioni considerate possono essere le classiche tre o cinque “V”, o un numero maggiore con l’aggiunta di dimensioni rilevanti per il singolo caso. Spesso fanno parte di queste:

- *Complessità*: Al crescere del volume dei dati, cresce la loro complessità nell’essere gestiti
- *Variabilità*: rappresenta un possibile problema relativo ai dati, maggiore è la variabilità maggiore è la loro inconsistenza

Come detto però queste sono caratteristiche non sempre considerate, ma che se prese in considerazione permettono di allargare lo spettro di analisi e di andare maggiormente in profondità.



Figura 2.1: Modello delle 5 “V”, (Fonte: <https://www.sadasdb.com/cosa-sono-big-data-quali-vantaggi-per-aziende/>)

2.2 Dalle biblioteche ai Big Data: cenni storici

Quanto scritto nei paragrafi precedenti ci permette di comprendere con precisione cosa siano i Big Data, quanto meno in prima istanza sapere cosa indica il termine. Ora è lecito chiedersi come si è arrivati a parlare di Big Data, quando sono diventati tema d’attualità. Di fatto sono un tema fortemente attuale, che ha generato, grazie le sue innumerevoli applicazioni, ulteriori temi interessanti per le aziende. Ma quando inizia la loro storia?

La raccolta di dati e informazioni è da sempre un tema all’uomo caro, qualunque potesse essere il fine, dalla mera raccolta di conoscenza, alla possibilità di consultare in seguito tali informazioni, sino al tramandare alle future generazioni conoscenza e cultura. Parliamo quindi di una pratica vecchia millenni, che trova riscontro nell’antichità con l’invenzione dell’abaco nel duemila avanti Cristo, utilizzato per compiere calcoli. Ancora, esempio sono le biblioteche dell’antichità come quella di Alessandria, risalente al quarto secolo avanti Cristo.

2.2.1 Talete e i “Big Data”

Si potrebbe dire che i Big Data e la figura del “data scientist” siano nati nel 600 avanti Cristo. Ovviamente l’affermazione così riportata potrebbe sembrare ai più una castroneria, per di più in virtù della definizione vista in precedenza. Eppure, nel 600 a.C. *Talete*, noto filosofo e matematico greco, fu il primo nella storia a compiere un’analisi predittiva e per di più, a trarre guadagno dall’uso dei dati²⁶.

Egli fu uno dei primi a spiegare fenomeni naturali attraverso dati e informazioni e non attraverso miti e leggende su divinità ed eroi. Talete raccolse per diversi anni informazioni sul meteo, che gli permisero nel tempo di riuscire nella previsione di quest’ultimo. Credè quindi una forma primordiale di “database” in cui annotava tutte le informazioni riguardanti il meteo, non solo fenomeni di interesse ma anche elementi definibili di contorno, non di primaria rilevanza. Una prima applicazione dei dati per prevedere un fenomeno. Talete trovò inoltre il modo di guadagnare dalla sua attività di raccolta delle informazioni, unendo le sue conoscenze allo studio degli astri. Egli, per rispondere a chi lo criticava di svolgere un’attività inutile, sfruttò le informazioni accumulate sul clima, insieme alla conoscenza delle stelle, per predire un abbondante raccolto di olive. Conoscendo tale informazione, si mosse e prese in concessione, secondo le forme arcaiche dell’epoca,

²⁶ <https://www.digitalic.it/tech-news/i-big-data-nel-600-avanti-cristo>

i campi di olive di Mileto e dell'isola di Chio. La previsione risultò veritiera e il matematico greco si ritrovò con il monopolio dei campi, che poté quindi sub-affittare a un prezzo che più lo arricchisse.

Tale racconto è, come tutti o quasi quelli riguardanti la figura di Talete, in bilico tra la leggenda e la realtà. Purtroppo, dei suoi testi non è rimasto nulla e le sue vicende sono giunte a noi solo tramite le opere e gli scritti di altri autori ma le fonti che parlano di lui risalgono all'antica Grecia. Questo ci fa affermare che la consapevolezza dell'importanza della raccolta di informazioni per poterli consultare in futuro fosse nota già in tempi remoti.

2.2.2 Dal 1800 sino ad oggi

La storia del personaggio di Talete risulta affascinante e incredibile, ma chiaramente non può essere presa, vista la mancanza di testi precisi al riguardo, come prima fonte storica. Risulta utile però per esporre come, raccogliere informazioni, faccia parte delle attività svolte dall'uomo sin dall'antichità.

Tuttavia, quando parliamo della storia dei big data si fa riferimento principalmente agli ultimi tre secoli circa, partendo dalla seconda metà del 1800.

La prima data che si incontra è il 1865. In questo periodo si parla per lo più di “*business intelligence*”, adoperata per ottenere vantaggi competitivi sui concorrenti attraverso raccolta e analisi strutturata dei vari dati e informazioni rilevanti per il business considerato.

Business intelligence e non Big Data perché i due differiscono nell'utilizzo e nella tipologia di dati. La prima sfrutta la statistica descrittiva per misurare e rilevare tendenze, attraverso dataset limitati e metodi semplici. I secondi invece utilizzano la statistica inferenziale e per dedurre delle leggi, quindi ad esempio regressioni e regressioni non lineari da insiemi molto grandi di dati. Questo per individuare rapporti tra i dati, relazioni e dipendenze, ma anche poter effettuare previsioni di comportamenti e di risultati. I dati trattati in questo caso sono di origine eterogenea e i modelli sono complessi. Il termine business intelligence però venne coniato solo nel 1958, solo oggi osservando il passato siamo in grado di definirla tale.

Il tutto quindi risultava ancora confuso, poco definito. Nel 1880, 15 anni dopo, un dipendente del censimento USA, *Herman Hollerith*, sviluppò un sistema in grado di classificare ed organizzare i dati dei vari censimenti per ridurre le tempistiche per la catalogazione, passando dai circa 10 anni richiesti a pochi mesi. Hollerith fondò una società, la “*Tabulating Machine Company*”, questa dopo varie acquisizioni e nuove denominazioni divenne, nel 1924, la *IBM*.

Passa quasi un secolo, ma nel 1965 viene costruito il primo data center negli Stati Uniti. Pochi anni dopo fu creato da un dipendente di IBM il primo framework per database relazionali. Negli anni '70 l'evoluzione dei computer permise di immagazzinare dati su più sistemi, creando un network indipendente noto come CERNET. Quindi nel 1991 nasce Internet e nel 1999 viene coniata la parola Big Data. Nel 1999 nasce anche la “*Internet of things*”, ossia la connessione di oggetti fisici a Internet, i quali forniscono svariati dati e informazioni. Al tempo le tecnologie non erano però ancora in grado di gestire la quantità di dati e di mettere in relazione dati provenienti da fonti eterogenee e in particolar modo quelli non strutturati. Nel corso dei successivi anni si gettano le basi del marketing digitale, andandosi a definire le regole base. Nel corso degli

anni 2000 poi, l'evoluzione tecnologica ha permesso di archiviare con maggiore facilità e sicurezza quantità di dati sempre più elevate. Si sono sviluppate varie applicazioni nel corso degli ultimi vent'anni che oggi sono alla base delle tecniche di marketing sia più avanzate che più semplici.



Figura 2.2: Rappresentazione grafica dell'IoT, (Fonte:<https://www.pandasecurity.com/it/mediacenter/mobile-news/iot-e-sicurezza/>)

2.3 Ciclo produttivo dei Big Data

Estrapolare informazioni dai Big Data per creare valore e contribuire a decisioni, col fine di migliorare la qualità delle scelte prese, richiede lo sviluppo e l'implementazione di processi che permettono di gestire, convertire e trasformare dataset, i quali crescono in continuazione sia in termine di volume che di varietà. Ogni processo si caratterizza di fasi che modificano lo stato e il contenuto dei dataset, permettendo di convertire quantità di dati grezzi enormi e quindi, di generare valore da questa conversione, che andrà ad arricchire il modello analitico dei dati. I processi fungono da linea guida nella scelta o nella creazione di un'architettura. Con il termine "architettura" intendiamo una struttura logica e fisica che specifica e impone come i Big Data verranno memorizzati, come si potrà avere accesso ad essi e come dovranno essere gestiti all'interno di organizzazioni. All'interno dell'architettura vengono definite inoltre, le soluzioni ai possibili problemi che possono occorrere nel trattamento dei Big Data. Spesso i problemi sono legati alla qualità scarsa dei dati, componenti hardware, ma anche problemi lato software come il flusso dei dati, la privacy e la sicurezza dei dati, temi molto importanti e spesso al centro di diatribe tra aziende e istituzioni.

Il ciclo produttivo dei Big Data, definibile anche come ciclo di vita, si compone di due macroaree, ognuna delle quali si compone di vari processi o fasi, esse sono:

1. *Big Data Management*: prima macroarea, al suo interno si trovano i processi che permettono la raccolta e l'archiviazione dei Big Data, nonché le tecnologie per preparare il loro recupero.
2. *Big Data Analytics*: macroarea che si compone dei processi per acquisire e analizzare le informazioni utili provenienti dai dataset con il fine di compiere analisi descrittive, prescrittive e predittive.

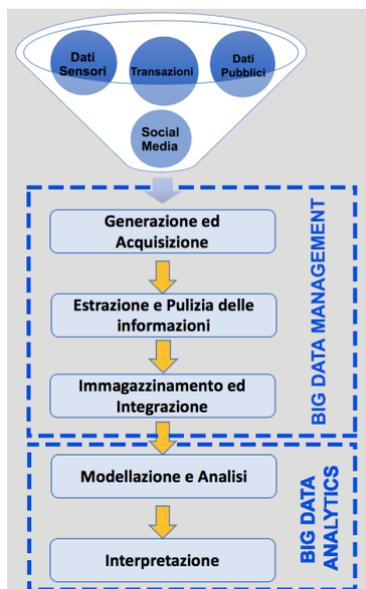


Figura 2.3: Ciclo dei Big Data, (Fonte:https://it.wikipedia.org/wiki/Big_data)

2.3.1 Big Data Management: fase di Generazione ed Acquisizione

Riprendendo quanto affermato nei precedenti paragrafi, occorre ricordare che oggi la quantità di dati generati è immensa, nell'ordine degli zettabyte. Negli anni questa quantità è cresciuta vertiginosamente e seguendo un andamento esponenziale²⁷.

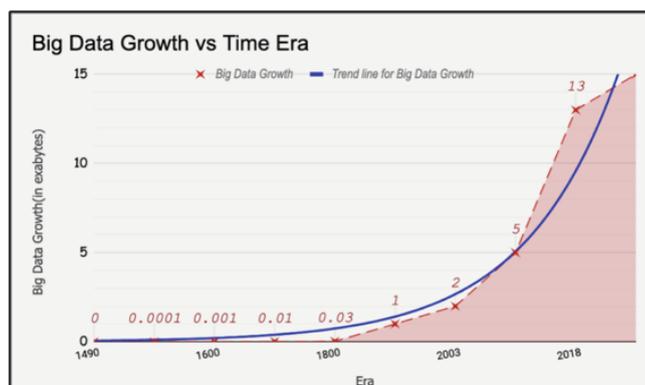


Grafico 2.3: Confronto dati prodotti rispetto a previsioni sulla produzione, (Fonte:https://doi.org/10.1007/978-3-030-14544-6_4)

I dati prodotti non solo differiscono nella loro natura e nella loro struttura, ma sono anche originati da fonti diverse, che permettono di classificare i dati in diverse tipologie, ovvero le seguenti:

- *Human generated*: sono tutti quei dati generati dall'attività umana su social network, come *LinkedIn* o *Facebook*, oppure da attività come il *blogging* o il *microblogging*. Altri esempi sono i dati generati da siti di social news come *Reddit*, siti per la condivisione di materiale multimediale come *Instagram* o *YouTube*, le varie *wiki* come *Wikipedia*, siti per recensioni o di domande e risposte come *Quora*. Anche i siti di *e-commerce* come *Amazon* rappresentano fonti per dati di questo tipo. Quasi tutti i siti web che sono gestiti tramite cookie generano dati riconducibili a questa tipologia

²⁷ Sharma, A., Pandey, H. (2020). Big Data and Analytics in Industry 4.0. In: Nayyar, A., Kumar, A. (eds) A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-030-14544-6_4

- *Machine generated*: sono tutti quei dati prodotti da macchine e strumenti, come sensori *GPS* o *RFID*. Producono dati di questo genere tutte i sistemi di “*H.F.T.*”, *High Frequency Trading*, utilizzate nel settore della finanza. In generale riconduciamo a questa categoria tutti i dati generati da strumenti dell’IoT.
- *Business generated*: appartengono a questa categoria tutti i dati, siano essi machine o human generated, prodotti da aziende che registrano le attività data-driven dei processi aziendali. Gran parte dei dati prodotti di questo genere sono dati storici, vale a dire pagamenti, ordini di acquisto e di vendita, dati di produzione, inventari, dati finanziari. Negli anni, soprattutto con l’avvento di Internet le aziende hanno iniziato a produrre un numero di questi dati talmente elevato che si è reso necessario adoperare tecnologie in grado di eseguire analisi anche in tempo reale per sfruttare il pieno potenziale delle aziende.

Secondo step di questa prima fase è rappresentato dall’acquisizione dei dati da queste tipologie differenti di fonti. Questa fase consiste nella pratica, in una fase di archiviazione e trasmissione dei dati, nonché di una “*data pre-processing phase*”. Occorre utilizzare un sistema di trasmissione dei dati molto efficiente, di modo da permettere un’analisi migliore, evitando ad esempio l’invio di dati che risultano ridondanti e inutili. Questo avviene attraverso l’applicazione di una tecnologia che viene definita “*data compression*”, che opera andando ad eliminare dati superflui. Le modalità di acquisizione dei Big Data sono differenti, spesso specifiche a seconda della fonte che origina i dati e sono:

- *Accesso ad API*: l’accesso ad *API*, permette di interfacciarsi con i vari siti Web per analizzare il contenuto di questi. Ne sono un esempio le *API* di *Twitter*, *Facebook* o le *API* dei vari motori di ricerca come *Google* o *Mozilla Firefox*.
- *Software di web scraping*: l’utilizzo di software di *web scraping*, ossia una tecnica informatica per l’estrazione di dati da siti web tramite l’ausilio di software appositi, permette la messa in pratica di attività di *crawling*, *parsing* ed *entity extraction*, per la raccolta in automatico di Big Data da documenti nel Web.
- *Importare dati*: che sia da database relazionali o non relazionali, o da sorgenti alternative. Questo avviene attraverso strumenti di *ETL*, ossia “*Ectract, Transform and Load*”. Uno dei più noti strumenti di *ETL* è *Apache Sqoop*, ampiamente utilizzato nel settore dei Big Data. Esso permette di esportare grandi quantità di dati da database relazionali e non verso la piattaforma *Apache Hadoop*, non solo può esportare ma anche importare, quindi compiere il passaggio inverso.
- *Acquisizione di flussi continui di dati*: flussi di dati vengono generati in modo continuo e rapido. Questi possono essere acquisiti attraverso sistemi, che sono in grado di catturare degli eventi, elaborarli e salvarli su un database. Strumenti tipici per svolgere queste funzioni sono quelli offerti dalla società sviluppatrice *Apache*, come *Apache Flume*. Altro esempio può essere *Microsoft StreamInsight*.

Una volta che i dati vengono acquisiti dalle diverse fonti, essi vengono trasferiti, con mezzi ad alta velocità, in dei sistemi di archiviazione, che possono essere locali o remoti. Si vengono a formare degli insiemi di dati grezzi, sia strutturati che non strutturati, definiti “*data lake*”²⁸, letteralmente “laghi di dati”, che devono essere ulteriormente precompilati, per eliminare dati superflui e duplicati di modo da aumentare l’efficacia delle successive operazioni. Inoltre, tali operazioni permettono di ridurre lo spazio necessario all’archiviazione rendendo possibile archiviare ulteriori dati.

2.3.2 Estrazione e pulizia dei dati, fasi di “ETL e Data Cleaing”

I dati che sono stati raccolti difficilmente sono nelle condizioni ottimali richieste per poter essere trattati nella fase successiva. Occorre quindi sviluppare un processo in grado di estrapolare i dati dalla fonte e successivamente che sia in grado di rappresentarli in modo standard, utilizzabile immediatamente per l’analisi. Questo processo è definito “*Ectract, transform and load*”, comunemente abbreviato con la sigla “*ETL*” Prima però di tutto questo avviene la pulizia dei dati, meglio nota come “*Data cleaning*”. Questo processo risulta fondamentale, per garantire e certificare la validità dei dati e la qualità di questi. Di norma la qualità dei dati, che possiede differenti definizioni, viene comunemente considerata alta quando questi sono adatti agli usi nelle operazioni e nei processi decisioni che riguardano. Effettuando un lavoro di data cleaning si va a migliorare la qualità dei dati. Occorre quindi, specificare cosa compone la qualità dei dati, ad esempio accuratezza o universalità. Quindi attraverso la pulizia andremo a migliorare di livello tali caratteristiche che compongono la “*Data Quality*”. Questa fase generalmente precede la fase di estrazione, o “*mining*”, ed è realizzata attraverso diverse tecniche, che seguono comunemente due filoni, questi cercano di unificare schemi sottostanti i dati e di rilevare ed eliminare errori non sistematici. Questi sono:

1. *Approccio “schema-level”*, che cerca di identificare una possibile corrispondenza tra database attraverso somiglianze che sono definite ex-ante
2. *Approccio “instance-level”*, in cui si cerca di stabilire delle metriche per misurare le somiglianze. Spesso seguendo questo approccio si ordinano le *tuple* di dati tramite delle chiavi di lettura, le quali permettono di creare dei *cluster* contenenti dati “simili” che si riferiscono alla stessa medesima entità.

Ad oggi i software moderni fanno uso di entrambi gli approcci in modo combinato.

Il processo di estrazione, conosciuta appunto con il termine inglese ETL, si caratterizza per l’estrazione dei dati, la trasformazione quindi il caricamento di questi in un sistema di sintesi, spesso definiti “*data warehouse*”. Il processo si compone quindi di tre fasi la cui prima è l’estrazione dei dati, puliti, da sorgenti che possono essere di vario tipo. In particolar modo i dati grezzi sono solitamente estratti da: database, registri di attività, report su anomalie e prestazioni riguardanti applicazioni, database transazionali vari.

²⁸ https://www.ibm.com/it-it/analytics/data-lake?utm_content=SRCWW&p1=Search&p4=43700068096868228&p5=e&gclid=EAlaIqobChMIwtPJq5ri-AIV2RkGAB1dbAWFEAAAYASAAEgIjs_D_BwE&gclsrc=aw.ds

A questa fase segue quella di trasformazione dei dati, la fase più critica. Ai vari dati vengono applicate regole per soddisfare dei requisiti di segnalazione, che impongono l'applicazione di formati corretti. In questa fase è necessario che i dati siano puliti, altrimenti risulterebbe decisamente complicato applicare tali regole ai dati. Di solito il processo di trasformazione segue questo iter:

- a) Dall'insieme totale dei dati vengono selezionati solo quelli ritenuti interessanti per il sistema.
- b) I dati vengono quindi normalizzati. Questo significa che si va ad eliminare la ridondanza e il rischio di incoerenza. L'obiettivo è decomporre le relazioni che presentano concetti indipendenti in relazioni più piccole, ognuna corrispondente a un singolo concetto. Da segnalare come non sempre questo processo può essere compiuto in quanto, in alcuni casi, può causare una perdita di informazioni e quindi ottenere un effetto negativo.
- c) Successivamente i dati vengono tradotti.
- d) Quindi da questi dati tradotti si va a derivare dei nuovi dati, ma in questo caso dati calcolati.
- e) Segue una fase di "join", vale a dire una fase in cui i dati recuperati da tabelle diverse vengono accoppiati.
- f) In ultima istanza, i dati vengono raggruppati tutti insieme.

Il fine è quindi quello di rendere dati provenienti da fonti diverse simili, quindi confrontabili.

Fa seguito a tale fase, quella del caricamento, la fase di "load". Questa è l'ultima del processo ETL, dove i dati prima estratti e successivamente trasformati, vengono caricati nel sistema di sintesi. Il caricamento nel data warehouse può avvenire attraverso due modalità: caricamento incrementale o completo.

I dati caricati devono evitare di essere eccessivamente dettagliati, perché comporterebbe una perdita di efficacia al momento dell'analisi. Essi devono essere al giusto livello per rendere possibile l'analisi di questi. I processi ETL sono strumenti altamente specifici che richiedono un personale specializzato che in caso di *turn over* o peggio, nel caso di licenziamento, comportano la necessità di saper gestire in anticipo in azienda, la presenza di *back-up*.

2.3.3 Immagazzinamento e Integrazione

La fase di immagazzinamento dei dati possiede un'importanza notevole. Immagazzinare grandi dataset semi strutturati e non strutturati e contemporaneamente rendere essi disponibili e accessibili in modo affidabile e veloce sono due dei maggiori problemi quando si parla di Big Data. Per risolvere queste problematiche, oltre a costruire infrastrutture hardware per archiviare informazioni, si è optato anche per lo sviluppo di meccanismi in grado di gestire i dati in questione, come *file system distribuiti*. Questi permettono di avere una visione d'insieme dell'infrastruttura di memoria sottostante. Essi forniscono operazioni di base per la lettura e scrittura

di moli di dati elevate, garantendo prestazioni elevate. Tra i file system distribuiti più noti ci sono: *Google File System* e *Hadoop Distributed File System*²⁹, sempre della società *Apache*, tra le leader nel settore dei Big Data. I dati sono archiviati in database la cui organizzazione logica è affidata a basi di dati *NoSQL (not only SQL)*, che non fanno uso del linguaggio *SQL (structured query language)* tipico delle soluzioni basate su *RDBMS*, caratterizzato da una staticità in controtendenza con le necessità richieste nella gestione dei Big Data e che proprio per questo si è rivelato fallimentare nel tempo. *NoSQL* possiede tra le tante caratteristiche il supporto per memorie persistenti di grandi quantità di dati.

La fase di integrazione viene effettuata insieme a quella di immagazzinamento. Questa fase permette attraverso ulteriori trasformazioni, di preparare i dati e renderli migliori per la successiva fase di analisi. In genere in questa fase si svolgono operazioni come l'unione di dati provenienti da database esterni.

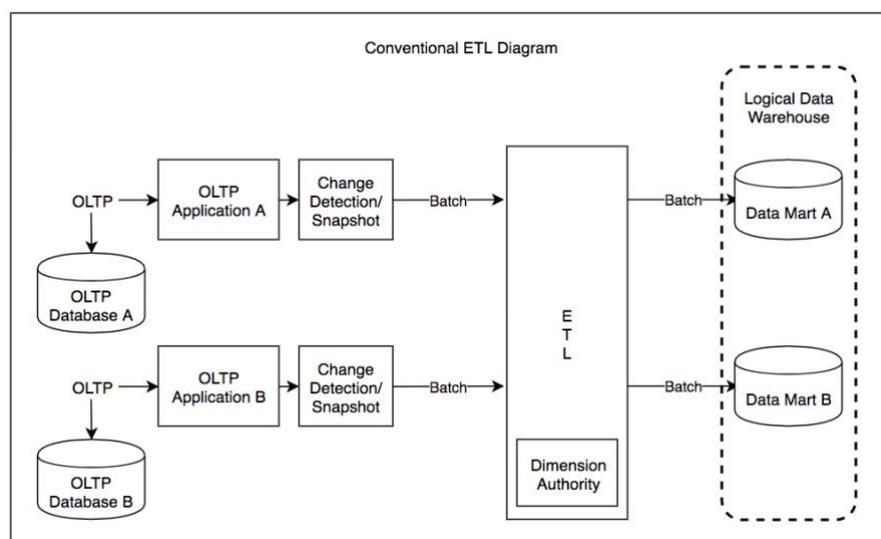


Figura 2.4: Rappresentazione grafica del processo ETL, (Fonte: Ralph., Kimball (2004). *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data.* Caserta, Joe, 1965-. Indianapolis, IN: Wiley)

2.3.4 Big Data Analytics: Analisi

Con la fase di immagazzinamento si chiude il processo di Big Data management. Si dà quindi il via al secondo processo come Big Data Analytics, che ha inizio con la fase di analisi. L'analisi ha lo scopo di estrapolare valore dai dati, analizzando i dataset per trovare correlazioni, possibili tendenze o indici statistici che possono nascondersi nei moltissimi dati che si hanno a disposizione. Le tipologie di analisi sono differenti e implementabili per diversi tipi di dati, siano essi strutturati, semi strutturati o non strutturati.

Per i dati di natura strutturata si fa principalmente ricorso ad analisi tradizionali. Si parla di modelli statistici adoperati per ottenere dai database dati raffinati utili per le decisioni che dovranno essere prese. A seconda del tipo di dato e del fine dell'analisi, si presentano diversi tipi di modelli:

²⁹ <https://www.ibm.com/it-it/topics/hdfs>

- *Factor Analysis*: conosciuta in italiano come “analisi fattoriale”. Attraverso l’uso di fattori si sintetizzano le relazioni tra dati e variabili col fine di identificare il maggior numero di informazioni nei e sui Big Data.
- *Cluster Analysis*: metodo statistico che permette di raggruppare i dati in gruppi, i quali sono caratterizzati da caratteristiche comuni, che permettono quindi una riclassificazione di questi. Spesso utilizzata per raggruppare dati all’apparenza differenti, spesso perché provenienti da fonti diverse. Dati eterogenei vengono quindi riclassificati in “gruppi” contenenti dati omogenei tra loro sulla base di caratteristiche specifiche che li accomunano.
- *Correlation Analysis*: come suggerisce il nome, lo scopo è verificare se tra i dati analizzati sussiste una correlazione. L’intento è inoltre quello di comprendere la forma di questa correlazione tra i dati. Le relazioni che possono esistere tra i dati sono diverse, abbiamo dipendenza unilaterale, dipendenza reciproca, la forma di correlazione più classica, ma anche forme di correlazione complesse, non giustificabili da variabili causali.
- *Regression Analysis*: metodo che si avvale della regressione. L’intento è dimostrare la correlazione tra differenti variabili attraverso numerosi esperimenti su dati osservabili. Limite di tale analisi è l’incapacità di spiegare le correlazioni non giustificate, cosiddette casuali.

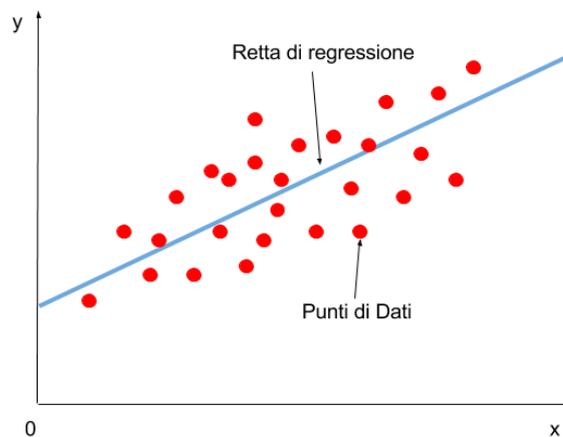


Grafico 2.4: Diagramma di dispersione di regressione lineare, (Fonte:https://www.researchgate.net/figure/Figura-32-Diagramma-di-dispersione-che-rappresenta-un-esempio-di-regressione-lineare_fig3_325477367)

Per i dati di natura semi e non strutturata, sono più comuni forme di analisi differenti da quella tradizionale. I contenuti multimediali sono spesso analizzati attraverso strumenti di Big Data analytics, proprio per la loro natura, la loro eterogeneità e le loro dimensioni. Essi vengono analizzati spesso attraverso algoritmi di machine learning che individuano informazioni di basso e alto livello per descrivere tali dati multimediali. Altra tipologia di analisi molto utilizzata è l’analisi del Web. Attraverso l’analisi di pagine web è possibile ottenere informazioni su uso e struttura del Web stesso. Questa analisi viene effettuata attraverso l’applicazione di tecniche viste in precedenza.

Distinguiamo poi le analisi secondo il fattore tempo in: *real time analysis* e *offline analysis*. Le due differiscono non solo per tempestività ma anche per ambiti d'uso:

- *Real time Analysis*: sviluppata partendo dalla necessità di analizzare dati che mutano molto velocemente nel tempo, che quindi necessitano di essere trattati il prima possibile per poter ottenere analisi utili. I principali utilizzatori di queste tipologie di analisi sono le piattaforme finanziarie o quelle di e-commerce.
- *Offline Analysis*: a differenza delle precedenti non analizza i dati in tempo reale ma li immagazzina prima attraverso dei processi di conversione. Lo storage tool più utilizzato per convertire i dati in queste analisi è tipicamente *Hadoop*. L'intento è di convertire i dati al minor costo possibile.

Le analisi possono essere realizzate su differenti livelli. A seconda dello spazio disponibile all'interno dei dataset analizzati, potremo avere:

- *Memory-level Analysis*: in questo caso i dati archiviati non occupano completamente lo spazio di memoria disponibile. Per i cluster di dataset generalmente si ha a disposizione uno spazio di archiviazione che si aggira intorno alle decine di *Terabyte*. Solitamente questa tipologia di analisi è tipica delle analisi in tempo reale, quindi di *real time analysis*. Lo sviluppo di unità di memoria a stato solido, note più comunemente con il termine inglese "*Solid-state drive*" o "*SSD*", ha dato un enorme slancio alla velocità e alla capacità di questa tipologia di analisi.
- *BI Analysis (Business Intelligence Analysis)*: al contrario della precedente, lo spazio di archiviazione è inferiore al volume del dataset. La mole di dati può però essere importata dal "*BI analysis environment*". In questo caso parliamo quindi di spazi di archiviazione molto più grandi di quelli visti in precedenza.

2.3.5 Interpretazione ed estrazione del valore

I dati e le analisi svolte su di essi sono uno strumento importantissimo e dal potenziale incredibilmente elevato. Tuttavia, non sono tutti in grado di comprenderlo e, soprattutto, di sfruttarlo. Molto dipende da come vengono realizzate tutte le fasi viste in precedenza. Ma trovare valore nei dati, o per creare valore, non basta compiere solo un'analisi o più di una. Il processo è ben più complesso e profondo e richiede specialisti, come analisti, esperti di dati, dirigenti "illuminati" che sappiano porsi e porre i giusti interrogativi, formulando delle ipotesi, oltre ad essere in grado di comprendere i modelli³⁰. Le imprese fanno quindi uso dei dati per incrementare e potenziare le prestazioni dei processi produttivi. L'impiego dei dati può avere diversi fini e scopi, i quali sono principalmente:

1. *Ottimizzare costi e ricavi*: per ottimizzare i costi vengono effettuate analisi predittive per opere di manutenzione, utilizzo dei dati per efficientamento gestionale, inoltre vengono applicati nella *supply chain* e nella logistica. Obiettivo principale è ridurre le spese indirizzando al meglio i processi ed eliminando ciò che è superfluo o inutile. Al fine di ottimizzare i ricavi, i dati possono essere utilizzati

³⁰ <https://www.oracle.com/it/big-data/what-is-big-data/#three>

per comprendere come entrare in nuovi mercati oppure, per segmentare i clienti in modo da poter comprendere al meglio quali scelte possono permettere maggiori ritorni economici. I dati possono essere inoltre utilizzati per andare incontro alle preferenze e alle richieste dei consumatori, agendo quindi sia sulle caratteristiche del prodotto o sui canali distributivi sulla base di ciò che si è compreso dall'analisi dei dati. La differenza tra efficientamento e ottimizzazione risiede principalmente nel fatto che per la prima attività, si sfruttano dati generati internamente, per ottimizzare vengono sfruttati maggiormente dati provenienti da sensori IoT e dati sui comportamenti dei consumatori.

2. *Marketing e Advertising*: si tratta di usare i dati e le analisi di grandi moli di dati per sviluppare campagne marketing più efficienti e mirate. I dati riguardano spesso le transazioni effettuate in passato dagli utenti, dati comportamentali, provenienti da fonti eterogenee come social network. Molto sfruttate sono anche le cronologie di ricerca dei consumatori e le informazioni demografiche. Per questo fine risulta fondamentale la fase di data cleaning, che rimuove ridondanza e permette di ottenere dati più consistenti in termini di valore aggiunto.
3. *Market Intelligence*: si tratta di tutte quelle aziende che forniscono servizi di informazione. Per lo più si tratta di aziende che raccolgono elevate quantità di dati da differenti e innumerevoli fonti. Quindi poi li forniscono a chi ne ha necessità, in modo diretto. Si tratta in sostanza di quelle aziende che forniscono dati per ricerche secondarie³¹.
4. *Market making*: queste tipologie di imprese svolgono una funzione molto importante, ossia far coincidere le esigenze di acquirenti e venditori. Questo avviene tramite lo sviluppo di piattaforme che permettono la raccolta dati, che spesso consistono in segnalazioni, preferenze, transazioni, dati comportamentali. Sono un esempio di imprese simili le applicazioni di car sharing o i siti di incontri. L'obiettivo è quello di mettere in contatto le esigenze dei soggetti in modo efficace ed efficiente. Tale funzione è utilizzata nel settore finanziario, dove le aste per acquisizione di titoli funzionano esattamente nello stesso medesimo modo, mettendo in comunicazione le diverse esigenze per trovare una corrispondenza tra acquirenti e venditori.
5. *“Training data for artificial intelligence”*: con il termine “training data” intendiamo quantità estremamente importanti di dati utilizzati per addestrare o insegnare un “modello predittivo supervisionato”. Machine learning e deep learning fanno un uso copioso di training data. Questi dati sono generati in parte attraverso la ripetizione di simulazioni, in parte sono presi dalla sfera pubblica, altri ancora vengono generati mettendo insieme svariati dati da differenti fonti, come video o foto. Maggiore è la quantità di dati prodotti dalle proprie piattaforme, maggiore sarà il vantaggio di cui si potrà godere, in quanti si avrà maggior tempo per le offerte di imparare e per generare nuovi dati. Sono quindi alla base degli algoritmi di machine learning.

³¹ Philip Kotler, Keller Kevin, Ancarani Fabio, Constabile Michele, “Marketing Management”, Pearson, (2017), pp. 162

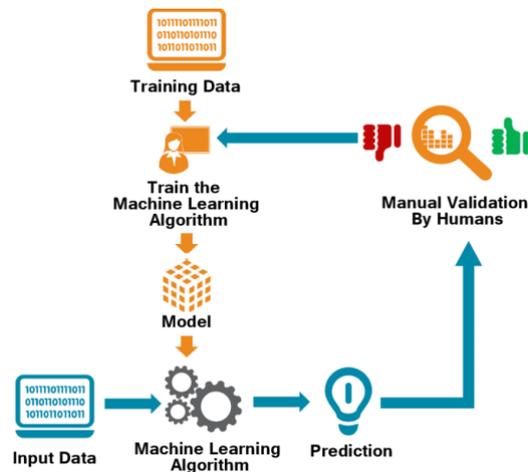


Figura 2.5: *Processo di Training Data*, (Fonte <https://www.cogitotech.com>)

2.3.6 Focus sul valore: si può calcolare quanto valore è stato generato?

Per generare valore è opportuno compiere delle scelte precise in termine di software adottati per la nostra strategia di analisi. Riuscire a definire quanto valore la nostra strategia produce non è così semplice, risulta sicuramente più facile comprendere se genera valore o meno, ma quantificarlo è cosa più complessa. Esistono però diversi metodi per valutare il valore generato dai software scelti. Questi ci aiutano a prendere le decisioni più efficaci al fine di creare valore, permettendoci di comprendere i diversi prodotti e di compiere la scelta più adeguata.

In passato si era focalizzati sui costi, le BI venivano realizzate per produrre report operativi. Questo è ancora oggi uno dei passaggi chiave dell'analisi ma non più l'unico come visto in precedenza. L'analisi moderna si concentra sul valore creato: *“Il valore della business intelligence non si valuta solo in termini di costi e benefici funzionali del software, ma anche di benefici aziendali derivanti dal processo basato sui dati”*³².

Tra i vari fornitori di soluzioni informatiche per la quantificazione del valore prodotto, *Tableau Software* risulta essere una delle più efficaci ed efficienti. Per operare hanno anzitutto stabilito una gerarchia del valore aziendale. Secondo Tableau, a seconda della fase in cui ci si trovi nel percorso che conduce un'organizzazione a diventare un'organizzazione basata sui dati, si può decidere di focalizzarsi su ottimizzare i costi, o di concentrarsi sull'efficienza o ancora, sull'efficacia. Ognuno di questi rappresenta un livello differente di analisi:

- Livello 1: *ottimizzazione dei costi* – Modello *TCO (Total Cost Ownership)*.
 - Si calcolano quelli che sono i costi totali di una soluzione BI e si sceglie a seguito di un confronto quella più economica
- Livello 2: *efficienza* – Modello della *produttività dei collaboratori*.

³² Matt Aslett, 451 Research

- Il confronto avviene sugli incrementi della produttività dei vari collaboratori, ottenuti attraverso varie soluzioni. Si fa riferimento al costo del lavoro per ottenere quelli che sono gli utili sul capitale investito.
- Livello 3: *efficacia – Modello dell’impatto previsto*.
 - Tramite il modello si ottiene una stima del valore ricavato tramite il miglioramento del processo decisionale.

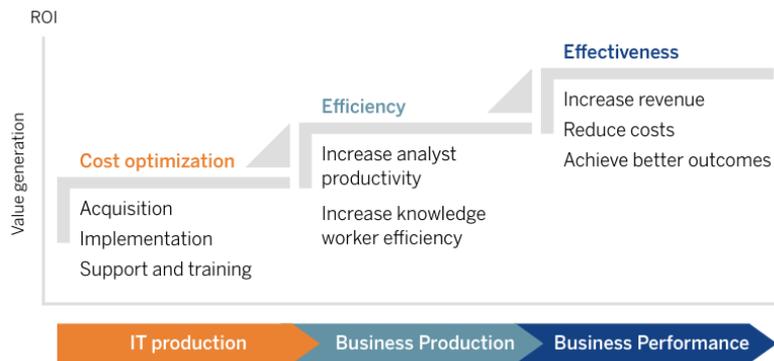


Grafico 2.5: Gerarchia del valore di Tableau, (Fonte: <https://www.tableau.com/it-it/learn/value-of-tableau/measuring-bi-value>)

Il primo modello, ovvero il “*Modello del Costo totale di proprietà*”, noto più comunemente come modello TCO, permette un confronto dei costi delle soluzioni tra cui poter scegliere. Spesso chi usufruisce di questi software compie l’errore di non considerare i costi secondari, valutando solo i costi della licenza e ignorando quelli relativi ad assistenza, formazione e tecnologie varie. Il modello in questione ha come presupposto che il valore generato dalle varie soluzioni BI sia pressoché identico; quindi, la scelta va compiuta secondo la logica dei costi, andando a scegliere quello meno costoso. Ma questo non è sempre vero, in alcuni casi non è così. Nel calcolo del TCO occorre quindi considerare non solo i costi della licenza, bensì anche molti altri costi secondari, che contribuiscono alle spese sostenute per adottare una soluzione tra quelle confrontate. Abbiamo costi per l’hardware, quindi per l’infrastruttura. Costi relativi al software, come licenze o l’accesso ad API. Poi ci sono costi legati al lavoro, ossia quelli relativi a installazione, configurazione, assistenza, personale addetto ai report e la formazione pratica.

Il modello TCO è consigliato quando le analisi sono gestite da consulenti esterni non sono richieste in modo dettagliato e specifico, ma è preferibile l’uso di dashboard. Osservando il grafico della gerarchia del valore, si può notare come in questa fase il valore generato è chiaramente più basso, si fa un uso limitato del potenziale che i dati possono offrire.

Product	Cost		Year 1	Year 2	Year 3	Total
Tableau	Platform	Software	\$228k	\$285k	\$356k	\$1.1M
		Infrastructure	\$146k	\$42k	\$32k	\$220k
	Labor		\$5.3M	\$5.9M	\$6.8M	\$18M
	Cumulative		\$5.7M	\$6.3M	\$7.3M	\$19.3M
Microsoft Power BI	Platform	Software	\$132k	\$165k	\$206k	\$629k
		Infrastructure	\$195k	\$159.5k	\$159.5k	\$514k
	Labor		\$7.5M	\$8.5M	\$9.7M	\$25.7M
	Cumulative		\$7.8M	\$8.9M	\$10.1M	\$26.8M

Tabella 2.2: Esempio di calcolo del TCO, (Fonte <https://www.tableau.com/it-it/learn/value-of-tableau/measuring-bi-value>)

Salendo di livello nella scala del valore troviamo un modello per quantificare il valore generato dalla ricerca di efficienza. Il “*modello della produttività dei collaboratori*” analizza i vantaggi per i collaboratori, analizzando il tempo risparmiato a seconda della soluzione intrapresa. L’ipotesi di base è che se i collaboratori impiegano minor tempo per effettuare un’analisi, possono impiegare il tempo guadagnato per svolgere ulteriori analisi o eseguire attività che vadano a creare ulteriore valore. Il modello prevede che si definiscano delle metriche, per calcolare la produttività dei collaboratori. Queste variano in funzione del business in analisi ma tra le più utilizzate ci sono:

- *Differenziale di creazione*, ossia la differenza tra input necessari e quelli effettivamente utilizzati per creare un risultato di analisi
- *Differenziale di consumo*, ossia il risultato tra la differenza di esperienza dell’utente per ottenere informazioni da fonti eterogenee e l’esperienza offerta dall’attuale soluzione
- *Fattore valore*, l’aspettativa media di valore aggiunto che un *knowledge worker* apporta all’impresa.

Il modello ha lo scopo di massimizzare il valore, perciò, numeri alti corrispondono a risultati migliori.

		Tableau	Anonymized BI Vendor
Number of Content Creators	100		
Time dedicated to BI content creation	30%		
Number of Information Customers	500		
Time dedicated to information consumption	5%		
Creation Differential*		100%	70%
Consumption Differential*		100%	70%
Worker Productivity Created (or Lost)		140,000 hrs	100,000 hrs
Value Factor relative to Labor*	2.0	100%	70%
Baseline FTE Rate	\$50,000/yr	100%	70%
Total Value Created		\$3.5M	\$2.5M

Tabella 2.3: Esempio di modello della produttività dei collaboratori, (Fonte:<https://www.tableau.com/it-it/learn/value-of-tableau/measuring-bi-value>)

L'ultimo modello, il “*modello dell'impatto previsto*”, viene utilizzato per misurare il valore creato nel momento in cui ci si concentra sull'efficacia. Questo modello parte da interrogativo, ovvero: “Se con questa tecnologia un collaboratore ha il X% di probabilità di adottare una decisione ottimale, che valore ricavo da ciò?”. Occorre relazionare la tecnologia utilizzata con i risultati reali che l'azienda ottiene. Proprio per questo il modello risulta quello più realistico, ma anche quello più complesso da quantificare. Risulta utile per migliorare le decisioni prese dai collaboratori e quando le analisi sono indirizzate ad aree specifiche dell'azienda.

L'impatto previsto viene calcolato considerando diversi fattori:

1. *Lo scenario*: se possiedo più informazioni non son automaticamente avvantaggiato, dipende da ciò che cerco di ottenere.
2. *L'ambito di decisione*: ogni settore ha ambiti di decisione più influenti che vanno individuati.
3. *Il miglioramento*: coefficiente di miglioramento che misura il valore aggiunto apportato alle decisioni da un *knowledge worker* che ha accesso ai dati. Misura, inoltre, in che modo le diverse tecnologie influiscono sulle decisioni finali.

Scenario: Property and Casualty insurance claims. Paying out claims quickly, and identifying possible cases of fraud for further investigation. A Special Investigations (SI) team investigates anomalous claims.		
Scope of decision: Paying out claims quickly and accurately both increases the chance of customer retention and helps avoid waste in the form of payouts to fraudulent claims. Avoiding waste also improves customer satisfaction by keeping costs low.		
Average Value of a correct payout decision: \$5,000 Frequency of Claims investigated by SI team: 10 per month		
Improvement: Business Intelligence is used to detect unusual activity compared to regional trends, analyzing billions of records of data to quickly identify anomalies.		
Benefits of Analytics	Tableau	BI Vendor
Improved accuracy of decision*	12.5%	0%
Improved speed of decision making**	10x	3x
Impact: Using the average value of a correct payout decision, improved accuracy, and allowing for more payout decisions made, data-driven decision making results in clear return on investment.		
Impact of Analytics		
Gross value:	\$562,500 per month	\$150,000 per month
Value of previous process:	\$50,000 per month	
Net value:	\$512,500 per month /\$6.15M annually	\$100,000 per month /\$1.2M annually

Tabella 2.4: Esempio di modello dell'impatto previsto, (Fonte:<https://www.tableau.com/it-it/learn/value-of-tableau/measuring-bi-value>)

2.4 In conclusione...

Attraverso i precedenti paragrafi si è quindi cercato di fornire gli strumenti base per poter affrontare il tema dei Big Data, permettendo al lettore di entrare in possesso dei rudimenti base per poter comprendere quanto in seguito verrà trattato. Si è fornita una definizione di cosa siano i Big data, ovvero una mole di dati di enormi dimensioni che necessita l'implementazione l'ausilio di tecnologie innovative per creare valore da essi. Per comprendere al meglio si è esposto il modello oggi alla base della definizione dei Big Data, il modello delle "3V", nonché la legge di Moore. All'interno del capitolo si fa riferimento, inoltre, a come si è arrivati ai big data, indicando le date più significative, fermo restando che il percorso passa per le tappe dell'informatizzazione che hanno coinvolto i diversi paesi, approfondite nel precedente capitolo. L'intento era far comprendere una volta, appresa la definizione di cosa sono effettivamente i Big Data, come questi siano non solo arrivati ad essere centrali nel mondo moderno, ma anche un fenomeno che ha origine da molto lontano. La raccolta dati e il loro utilizzo sono attività che l'uomo compie da millenni, fanno parte delle attività che contraddistinguono l'uomo da sempre e nel tempo si è riuscito a trovare il modo di rendere un'attività all'apparenza fine a sé stessa un mezzo per ottenere vantaggi competitivi. In ultima istanza si è analizzato come essi funzionano. Una volta capito cosa essi siano è necessario, per analizzare le varie applicazioni in ambito lavorativo, sapere come si evolve il ciclo di vita dei dati. Comprendere e fasi di Big Data Management e Big Data Analytics è cruciale per capire al meglio le varie applicazioni sul mercato, dei dati e delle analisi riguardanti quest'ultimi.

APPLICAZIONI DEI BIG DATA: NUOVE FRONTIERE PER L'INDUSTRIA MODERNA

“Per fornire il messaggio giusto alla persona giusta al momento giusto, prima devi raccogliere i dati giusti nel database giusto al momento giusto”

Jhon Caldwell

3.1 Introduzione: *Data Driven Economy*

Se nello scorso capitolo si è intuito quanto complessi ma incredibilmente utili possano essere i dati e le loro applicazioni, si è detto di come i dati possano apportare valore aggiunto alle aziende e ai loro processi produttivi³³, ora si andrà nel concreto ad analizzare diverse applicazioni in ambito lavorativo, passando quindi dalla teoria alla pratica. I dati una volta raccolti possono essere impiegati in vari modi con diversi scopi. Possono essere venduti, a imprese o privati, andando a rappresentare fonti secondarie per ricerche di mercato. Possono essere utilizzati internamente, per migliorare i processi di produzione, andando ad efficientare i costi e migliorando l'efficienza interna. Con efficienza interna parliamo di *ROI* dei *sunk investments*³⁴ in Ricerca e Sviluppo (*R&D*), ma anche quelli effettuati per acquisire piattaforme hardware per l'analisi e l'immagazzinamento di dati. La natura delle imprese ovviamente incide in quelle che sono le modalità con cui le imprese raggiungono questi risultati. Questa maggiore attenzione verso la raccolta e l'analisi dei dati da parte delle imprese, sempre più focalizzate sul creare valore aggiunto, ha portato alla nascita di imprese che fanno dei dati e del loro utilizzo il loro “*core asset*”. Questo ci guida verso una nuova economia, definita “*Data-Driven Economy*”. Si tratta di un'economia in cui i dati e il loro uso sono il fulcro delle strategie, della gestione del business, ma anche la base per lo sviluppo di nuovi prodotti e servizi, nonché centrali nei processi decisionali, nella creazione di nuove innovazioni e nella diffusione di queste e di conoscenza. Gli attori di questa nuova economia sono chiaramente le “*Data-driven companies*”, aziende guidate dai dati nello svolgimento delle loro funzioni. Tra queste il marketing è quella più legata all'uso dei dati e al mondo del *Digital*. Questa nuova forma di economia è testimoniata dalla crescita di *data storage* e *analytics* presentata da tutti i settori. Internet e l'informatizzazione hanno premesso la creazione di numerosi nuovi settori, ma anche il miglioramento di quelli già esistenti, come il settore bancario, il *manufacturing*, la P.A. (pubblica amministrazione)³⁵. I processi delle imprese attive in questi settori sono guidati dai dati, che fungono da bussola per le strategie delle suddette, grazie soprattutto ad algoritmi di *Machine learning*. Questi permettono l'efficientamento dei processi, che conduce al raggiungimento di obiettivi in modo più efficace. La nuova industria guidata dai dati si caratterizza per la produzione di beni, servizi, strategie di marketing e decisioni aziendali realizzate e prese partendo dai dati in possesso delle imprese. A dimostrazione di quanto affermato,

³³Si veda Capitolo II

³⁴ Investimenti fissi sostenuti da un'impresa

³⁵ Si veda Capitolo I

viene citato quanto sostenuto dall'OCSE in un documento del 2015³⁶. Secondo l'OCSE a fare maggior richiesta di dati sarebbero le attività produttive a valle, che stanno facendo dei dati una risorsa fondamentale per la produzione. In questo documento viene stimata una crescita di produttività nelle imprese, correlata alla “*data-driven innovation*”, che si aggira intorno al 5-10%. Inoltre, parlando per gli enti pubblici, la stima relativa alla riduzione dei costi amministrativi è del 15-20% circa. Questo è dovuto all'uso dei dati che ha portato maggiore efficienza e soprattutto un maggiore gettito fiscale, anche grazie all'uso fatto dei dati per prevenire frodi ed errori. Il dato che risalta di più da questo report è chiaramente quello relativo al peso della data economy sul PIL europeo, ossia 1,9% di quest'ultimo. Si stima che tale peso possa crescere nei prossimi anni.

Il più grande “contenitore” di dati esistente è Internet. In esso si trovano quantità di dati incredibilmente elevate, che vengono acquisite attraverso le varie ricerche che compiamo. Attraverso il suo utilizzo condividiamo informazioni come le nostre preferenze, abitudini, generalità e attitudini. Nel tempo si è colta questa importanza dei dati ed è nato un mercato dei dati, che ha attraversato una fase primordiale di scoperta, dove i dati hanno iniziato a prendere posizione all'interno dei vari processi decisionali, sino ad arrivare ad oggi, in una fase di esplosione commerciale, alla quale siamo giunti grazie alla maggiore comprensione dei dati.

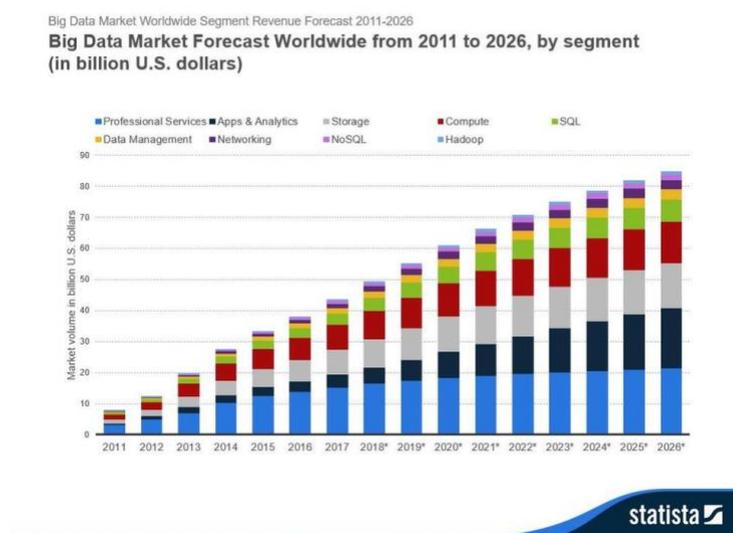


Grafico 3.1: Previsioni Big Data dal 2011 al 2026, (Fonte: <https://www.statista.com/statistics/255970/global-big-data-market-forecast-by-segment/>)

Osservando il grafico sopra riportato si può notare come le previsioni del mercato dei Big Data fino al 2026 sono in crescita in quasi ogni settore analizzato. Come si può notare il volume dei mercati di ogni settore ha avuto una crescita maggiore nel periodo tra il 2011 e il 2016. Questo giustifica la fase iniziale di scoperta dove si è avuta una forte crescita esponenziale. Di lì la crescita ha rallentato, fermo restando che le stime rimangono

³⁶ OECD (2015), Data Driven Innovation: Big Data for Growth and Well-Being, OECD Publishing, Parigi

in positivo e potrebbero aumentare qualora venissero rese disponibili innovazioni in grado di rivoluzionare i settori.

Nel seguente capitolo si illustreranno vari casi di applicazione dei Big Data in differenti settori e con differenti fini. Si vedrà successivamente nel particolare il funzionamento degli algoritmi di machine learning, quindi poi verrà presentato un *case study* su *Netflix*.

3.2 Dalla teoria alla pratica: casi d'uso dei Big Data

I big data sono utilizzati in diversi contesti. L'esplosione di dati prodotti e messi a disposizione delle imprese, insieme all'innovazione tecnologica ha creato nuovi scenari e incredibili opportunità. I sistemi della IoT, Internet of Things, interconnessi tra loro giocano un ruolo chiave insieme a Internet e ai social media. I Big data possono aiutare ad affrontare diverse attività aziendali, quali:

- *Sviluppo del prodotto*
- *Manutenzione predittiva*
- *Customer Experience*
- *Frode e compliance*
- *Efficienza operativa*
- *Promozione dell'innovazione*
- *Machine learning*

In seguito, verranno trattate le seguenti applicazioni, con particolare riguardo al machine learning.

3.2.1 Sviluppo del prodotto: Customer Demand Planning

Le aziende, come ad esempio Procter & Gamble (P&G), usano i Big Data per predire la domanda dei clienti. Esse costruiscono dei modelli predittivi da utilizzare con nuovi prodotti o servizi, classificando gli attributi chiave dei servizi e prodotti passati e attuali. In particolar modo P&G sfrutta dati e *analytics* di *focus group*, social media e dei *rollout* nei negozi per la pianificazione e il lancio di possibili nuovi prodotti. Questo processo di pianificazione aziendale prende il nome di “*Customer Demand Planning*” o “*CDP*”, in italiano gestione della richiesta del cliente. Esso permette ai gruppi di vendita di prevedere la richiesta e risulta un importante aspetto della catena del valore. Il primo passo di questo processo, il *Demand forecasting*, è proprio la previsione della richiesta di un prodotto. Questa può risultare utile per pianificare il *deployment* di risorse, che verrà stabilito proprio a partire dai risultati ottenuti dai modelli predittivi. Questa funzione ha richiesto la nascita di una figura nuova all'interno delle aziende, ovvero il Demand Planner.



Figura 3.1: Processo di Demand Planning, (Fonte: <https://powerslides.com/powerpoint-business/business-analysis-templates/demand-planning/>)

L'utilizzo di questi modelli previsionali permette quindi di ottimizzare la *supply chain* e di sviluppare nuovi prodotti e servizi in grado di soddisfare in modo più efficiente i consumatori, ottimizzando al massimo le fasi di produzione.

Un esempio di impresa, dalle enormi dimensioni, che svolge tale processo e ne ha fatto parte integrante di suo piano strategico è P&G. Attraverso l'uso dei dati P&G sta riuscendo ad affrontare sfide chiave: una supply chain resiliente, migliorare l'esecuzione al dettaglio e fornire prodotti e imballaggi migliori. Attraverso i dati riescono a creare previsioni che permettono di garantire che i prodotti di cui i consumatori hanno bisogno vengano resi disponibili, soprattutto nel momento in cui ne hanno bisogno. Attraverso i dati raccolti con i propri prodotti P&G sta implementando una valutazione che da point in time diventi in tempo reale. Punto fondamentale di ciò è la raccolta automatizzata di dati che permette di scoprire approfondimenti, velocizzare l'innovazione ma soprattutto determinare la producibilità di un prodotto.



Figura 3.2: Come P&G fa uso dei Big Data, (Fonte: <https://www.edureka.co/blog/big-data-analytics>)

3.2.2 Manutenzione Predittiva

Tra i dati possono celarsi i fattori che possono rendere possibile effettuare previsioni su guasti meccanici. Tali fattori possono trovarsi tra i dati strutturati come anno, marca e modello dell'attrezzatura, oppure possono trovarsi in dati non strutturati, come messaggi di errore, dati raccolti dai vari sensori, come ad esempio le temperature dei motori. L'analisi di questi dati permette di prevedere guasti prima che essi si verifichino, permettendo quindi un uso più efficiente delle risorse. Inoltre, il mantenimento delle condizioni ottimali dei vari macchinari permette di non incomberne nell'interruzione della produzione. Nel dettaglio si parla di metodi di conservazione *data-driven* che permettono l'individuazione dell'esatto momento in cui compiere la

manutenzione. Rispetto ai metodi tradizionali di manutenzione, essa elimina il guasto invece di ripararlo andando ad agire in modo mirato nel momento più giusto, cosa che la differenzia dalla manutenzione preventiva. Il passaggio delle industrie verso il digitale, attraverso l'IoT, ha permesso l'implementazione di Big Data Analytics, intelligenza artificiale e algoritmi di machine learning. Quest'ultimo risulta fondamentale nello svolgimento di queste pratiche predittive. Più esso apprende informazioni sui macchinari, più i suoi interventi di manutenzione divengono efficaci. Questo richiede un numero elevato di sensori. Sono inoltre richiesti raccolta dati, rilevamento dei guasti, previsione *time-to-failure*, programmazione. Un'indagine svolta da A. T. Kerney su *Industry Week*³⁷ ha messo in evidenza come 558 imprese che adoperano soluzioni computerizzate per la manutenzione, esse abbiano ottenuto:

- 28,3% di aumento della produttività delle manutenzioni svolte
- 20,1% in meno di inattività dei macchinari
- 19,4% di risparmio per il costo dei materiali
- 17,8% in meno di manutenzioni dell'inventario e di riparazioni
- 14,5 mesi di tempo di recupero dell'investimento

Questa pratica risulta molto diffusa nell'industria ferroviaria, nel *manufacturing* ma anche nell'industria del petrolio e del gas.

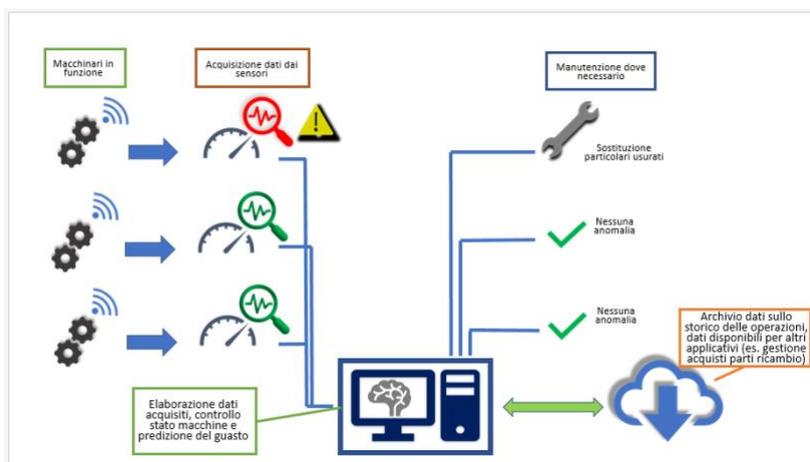


Figura 3.3: Esempio di manutenzione predittiva, (Fonte: <https://www.giardinodellecompetenze.it/manutenzione-predittiva-definizione-funzioni-e-applicazioni/>)

3.2.3 Customer Experience

Oggi più che mai è possibile avere una visione chiara della *customer experience*. I dati raccolti su Web, social media e le altre differenti fonti permettono di migliorare le interazioni dei vari consumatori e di offrire un valore aggiunto massimizzato. Le applicazioni dei Big Data in questo campo permettono di ridurre il tasso di abbandono e di poter gestire in modo proattivo le problematiche che insorgono.

L'applicazione di BDA (Big Data Analytics) in ambito di customer experience è recente e la letteratura a risulta ancora non al pari di quella in altri campi.

³⁷ A. T. Kerney (2021), IBM Maximo: Transform your business with intelligent enterprise asset management, IBM Corporation

Alla base di tutto ci sono i *customer experience insights*. Essi sono definibili come la conoscenza dei consumatori che viene ottenuta tramite Big Data Analytics, al fine di permettere un miglioramento della customer experience. Questi si generano attraverso i dati, sta alle aziende capire quali dati e analisi compiere per generare gli insights. Le aziende possono raccogliere dati strutturati e non strutturati, nonché dati sollecitati, come i sondaggi, oppure non sollecitati. Questi dati possono poi essere trattati attraverso le varie tipologie di Big Data Analytics. Gli insights che si possono quindi generare sono di diversi tipi: attitudinali o psicografici, comportamentali o di mercato. A loro volta gli insights possono essere utilizzati in diverse azioni di customer experience, quali: monitoraggio, “prioritizzazione”, adattamento e progettazione dei vari *touchpoints* del *customer journey*.



Figura 3.4: Framework per il customer experience management, (Fonte: Journal of Business Research 116 (2020), “Customer experience management in the age of big data analytics: A strategic framework”)

Le dimensioni che i dati possono seguire sono quindi due, da strutturati a non strutturati e da sollecitati a non. Attraverso queste si può stabilire che tipologia di dati ottenere nei vari punti di contatto con i clienti.

CX data types and their characteristics.

	Solicited –Structured	Solicited –Unstructured	Unsolicited – Structured	Unsolicited – Unstructured
Typical examples	Customer satisfaction or NPS ratings on surveys	Customer responses to open-ended survey questions	Customer ratings on independent online review platforms, observed behaviour	Online reviews, social media posts, voice recordings, branded vlogs
Degree of application in practice	High	Medium to high	Medium to high	Low
Format of CX data	Numerical	Mainly text	Numerical	Text, audio, image, video
Number of facets to the data	Low (a number is a number)	High (e.g., semantics, syntax, speech acts)	Low (a number is a number)	Very high (e.g., pitch, speech rate, pauses, static versus dynamic, recording quality)
Complexity of capture	Low	Medium	Low to medium	Medium to high
Effort required from customers	Medium	High	None to medium	None to medium
Fixed costs for organization	Low	Low	High	High
Variable costs for organization	High	Low	High	Low
Privacy and legal concerns	Low	Low to medium	Low to medium	Medium to high
Representative studies	Choi, Yang, Yang, and Cheung (2015); Mudambi and Schuff (2010)	Villarrol Ordenes et al. (2014); McColl-Kennedy et al. (2019)	Choi et al. (2015); Mudambi and Schuff (2010)	Proserpio and Zervas (2017); Villarrol Ordenes and Zhang (2019)

Tabella 3.1: Tipologia di dati e caratteristiche per la customer experience, (Fonte: Journal of Business Research 116 (2020), “Customer experience management in the age of big data analytics: A strategic framework”)

Le combinazioni ottenibili sono quindi 4:

1. *Sollecitato-strutturato*, i più comuni ma dal potenziale più basso, poiché spesso sono rappresentati da numeri. La customer experience difficilmente può essere rappresentata attraverso numeri.
2. *Sollecitato-non strutturato*, molto richiesti perché rappresentano una forma dal potenziale elevato per migliorare la customer experience.

3. *Non sollecitato-strutturato*, dati strutturati forniti volontariamente dai consumatori. Questa forma di dati comporta costi fissi elevati.
4. *Non sollecitato-non strutturato*, probabilmente la forma di dati con il potenziale più elevato. Sono esempio le valutazioni effettuate dai clienti sui siti web, *e-mail* o quello che viene scritto sui social. Se da un lato lo sforzo richiesto ai clienti è il più basso, i costi fissi richiesti sono molto elevati e spesso sorge il dilemma della *privacy*.

Su queste forme di dati vengono poi compiute le analisi già trattate nel corso dell'elaborato³⁸, per poi poter intraprendere delle “azioni di customer experience”. Tali azioni sono inquadrare in una logica di lungo periodo, dove le implementazioni di queste azioni nel breve permette nel lungo di ottenere una migliore progettazione.

Le azioni, già in precedenza accennate, sono:

- *Monitoraggio*: la fase di “*Touchpoint journey monitoring*”. Si utilizzano gli insights per fornire indicatori di performance. Le imprese possono inoltre formare gruppi con il compito di monitorare e con responsabilità trasversali tra i vari punti di contatto del percorso. Attraverso la IoT le imprese possono collegare i loro prodotti fisici e monitorare il “*customer journey*”. Questo si ricollega poi alla manutenzione predittiva, descritta in precedenza, che permette di migliorare non solo l'operatività dei macchinari, ma anche l'esperienza complessiva del consumatore.
- *Definizione delle priorità*: nota più comunemente con il termine inglese *prioritisation*. Si tratta di un uso dei dati e degli insights volto a rendere più efficiente l'allocazione delle risorse, siano esse monetarie, umane o tecnologiche. Tale azione permette la modifica dei punti di contatto direttamente nel breve termine, senza dover affrontare una completa riprogettazione del customer experience journey. Tramite dati e insights si possono individuare i touchpoint di maggior rilevanza e i fattori chiave che incidono sull'esperienza complessiva del consumatore.
- *Adattamento del percorso*: definita anche “*touchpoint journey adaptation*”. I vari insights possono essere utilizzati per generare proattivamente dei punti di contatto altamente personalizzati. Un esempio è quello che *Spotify* ha fatto con la campagna nata nel 2019 “*#2019wrapped*”. La società di streaming musicale ha sfruttato analisi predittive e descrittive per sviluppare touchpoint “*customizzati*”, basati sul singolo cliente. Tramite e-mail e notifiche *in-app*, *Spotify* ha informato i suoi abbonati sulle loro abitudini di ascolto durante tutto l'anno. Tali azioni hanno permesso a *Spotify* di fornire ad ognuno dei suoi abbonati playlist personalizzate. Tale campagna ha poi incontrato il favore popolare divenendo immediatamente un fenomeno virale, nonché ad oggi uno degli eventi più attesi dell'anno sui social.
- *Progettazione del customer journey*: azione di lungo periodo, frutto di un'applicazione nel tempo delle azioni già presentate. L'utilizzo dei dati e degli insights permette di disegnare un percorso di customer experience migliore per le aziende che decidono di adoperare tali azioni. In questo senso,

³⁸ Si veda capitolo II paragrafo 2.3.4

rivoluzionario fu quanto compiuto da Jhon Deere, produttore di attrezzature agricole, che dotò di sensori e software le sue macchine per permettere ai consumatori di analizzare i dati delle loro attrezzature. Il modello aziendale passò dall'essere basato sulla produzione ad essere incentrato sulla piattaforma, *myJhonDeere.com*, messa a disposizione dall'azienda.

3.2.4 Frode e Compliance: applicazioni nel mondo della finanza e non solo

Nel tempo i problemi relativi alla sicurezza e i possibili pericoli generati da figure come *hacker* sono aumentati esponenzialmente. I requisiti di sicurezza nella compliance sono in costante mutamento e in questo scenario di perpetua evoluzione, i Big Data risultano uno strumento dal potenziale incredibilmente elevato. Attraverso l'analisi è possibile individuare i modelli nascosti e per identificare frodi. Inoltre, è possibile rendere la burocrazia più veloce tramite l'aggregazione di moli di dati.

Da diversi anni, sta prendendo piede l'uso di pratiche di BDA nel settore *Finance*, ma non solo, perché i diversi report degli ultimi anni hanno visto crescere l'uso di tali prassi e procedure anche in ambiti come il *Telco* o penale.

Il tutto parte dal concetto di "*Pattern Discovery*", ossia la ricerca di correlazioni e similitudini svolta attraverso l'analisi sui dati. Risulta semplice da qui capire come la cosa si leghi al mondo della prevenzione da frodi. Coloro i quali lavorano per prevenire frodi svolgono una serie innumerevole di analisi per identificare possibili situazioni losche ed equivoche. La Big Data Analytics può automatizzare questi processi, seppur in parte, a volte rendendoli anche più efficaci. Principali attori in questo settore sono le analisi in tempo reale, intelligenza artificiale e algoritmi di Machine learning. Esempi di ciò sono gli algoritmi sviluppati per riconoscere truffe assicurative attraverso l'analisi di immagini, oppure algoritmi in grado di suddividere clienti leali da frodatori. Chiaramente l'analisi in tempo reale svolge un ruolo chiave. Minore tempo si impiega per individuare le frodi, maggiore sarà la tempestività delle azioni messe in atto, per fermare sul nascere o prevenirle, riducendo notevolmente i danni economici.

Nell'ambito della pubblica amministrazione e della compliance, così come in quello finance, uno dei principali problemi è la provenienza dei dati. La credibilità delle previsioni ottenute si basa principalmente su questo fattore. L'evoluzione dei software e delle tecnologie ha però permesso di ovviare in gran parte a tale problema. Oggi la capacità dei sistemi di processare dati e informazioni, provenienti da fonti affidabili, è divenuta talmente avanzata da riuscire ad analizzare moli di dati inavvicinabili anche per team composti da numerosi esperti. Questo miglioramento nel corso del tempo, inoltre, ha permesso alle aziende di ridurre gli errori determinati dalle analisi, ovvero i cosiddetti "falsi positivi", ma soprattutto ha premesso la riallocazione di risorse verso altre attività. Il miglioramento di queste tecniche ha dei ritorni anche in termine di *brand reputation* e di formazione del personale. Gli specialisti che nel tempo si sono formati nell'uso di queste tecnologie possono divenire risorse dal valore inestimabile, assumendo anche un ruolo secondario, ma non troppo, di formatori per i nuovi assunti.

Un esempio pratico di machine learning applicato alla finanza sono i vari sistemi di machine learning che hanno come obiettivo quello di individuare frodi in ambito di finanziamenti finalizzati. In passato in tale

ambito, le richieste sospette erano sottoposte al vaglio di attenti esperti dell'antifrode. Oggi il sistema assegna in tempo reale un punteggio rappresentante la probabilità di truffa associata alla pratica in analisi³⁹, in modo simile a come avviene per la valutazione assegnata in relazione al merito creditizio, sempre nel settore finanziario, permettendo un'analisi delle pratiche più rapida.

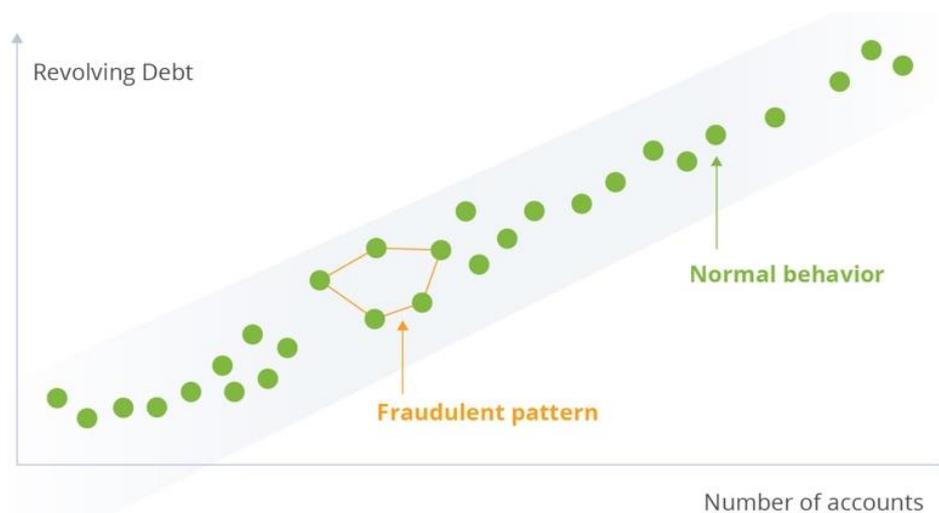


Grafico 3.2: Individuazione di frode attraverso connected analysis, (Fonte: <https://www.infopulse.com/blog/financial-fraud-detection-powered-by-big-data>)

3.2.5 Efficienza operativa: il nuovo che migliora il vecchio

Una delle aree dove i Big Data risultano vere un impatto maggiore è una di quelle forse meno innovative, ovvero l'efficienza operativa. Di solito si associano i dati ad ambiti informatici innovativi, nati negli ultimi anni, ma l'efficientamento e l'ottimizzazione sono un ambito decisamente importante, l'impatto dei Big Data in tale settore è tra i più rilevanti. Attraverso i dati è possibile compiere analisi sulla produzione per ottenere delle valutazioni su di essa, svolgere analisi e valutazione dei feedback e dei resi effettuati dai clienti. Il fine è spesso quello di ridurre le interruzioni della produzione e anticipare quelle che possono essere le richieste future. In questo senso si ricollega a due applicazioni viste in precedenza, ovvero manutenzione predittiva e sviluppo del prodotto. Oltre a questi, vi anche lo scopo di migliorare il processo decisionale, anche e soprattutto per renderlo compatibile con quella che risulta essere la domanda di mercato.

Il processo decisionale diviene quindi data-driven, ossia attraverso dati, metriche ed elementi concreti si orienta il processo decisionale affinché questo sia allineato con gli obiettivi aziendali. Per eseguire scelte migliori e trarre vantaggio da esse occorre comprendere pienamente il valore dei dati. Non si tratta soltanto di scegliere la tecnologia adeguata, occorre creare una cultura aziendale basata sul pensiero critico. Le competenze devono basarsi sulla pratica e sull'applicazione. Questo step, questo gradino, risulta molto più difficile da superare di quanto possa sembrare. Si pensi che solo nel 2018 su un campione numeroso di imprese scelte poiché avevano investito miliardi di dollari per modernizzare e ambire a diventare aziende data-driven,

³⁹ Irene Di Deo, Big Data e Machine Learning per combattere le frodi, 10 marzo 2019, <https://www.bigdata4innovation.it/esperti-e-analisti/big-data-e-machine-learning-per-combattere-le-frodi/>

il 70% circa aveva fallito tale obiettivo⁴⁰, a causa di un'eccessiva focalizzazione sulla tecnologia che ha portato a trascurare lo sviluppo di una cultura dei dati.

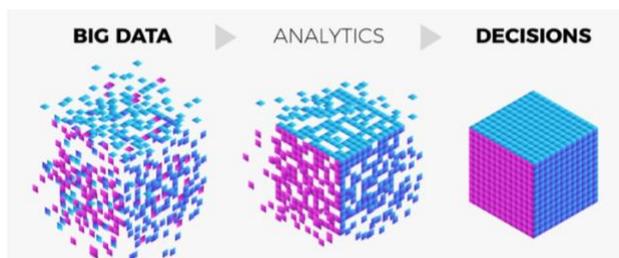


Figura 3.5: Schematizzazione del processo decisionale data-driven, (Fonte: <https://towardsdatascience.com/how-is-the-current-state-of-big-data-analytics-in-controlling-1273c725ac6a>)

L'utilizzo dei dati per migliorare l'efficienza operativa e i processi decisionali permette quindi di ottenere numerosi vantaggi. Esempi pratici di organizzazioni in grado di adoperare i dati per migliorarsi sono:

- *Lufthansa*, che grazie ai Big Data, ha incrementato l'efficienza organizzativa del 30%
- *Providence St. Joseph Health*, che attraverso i dati ha migliorato misure qualitative e costi relativi all'assistenza
- *Charles Schwab Corporation*, che ha reso più rapido il processo di ottenimento delle informazioni aziendali attraverso pratiche di Big Data Analytics e processi aziendali data-driven⁴¹

I dati se sfruttati nel modo opportuno permettono quindi di ottenere vantaggi aziendali determinanti nei contesti competitivi: *“Senza la soluzione di analisi visiva, avremmo continuato ad analizzare enormi quantità di dati nei fogli di calcolo. Le dashboard, invece, forniscono informazioni chiare e concretizzabili, che danno impulso alla azienda”*⁴².

3.2.5 Promozione dell'innovazione: innovare per crescere

I Big Data sono quindi uno strumento in grado di generare grande valore. Possono aiutare ad innovare, attraverso lo studio di possibili correlazioni tra i dati per ottenere nuovi modi di utilizzare i vari insights che essi producono. Le possibilità sono pressoché infinite, possiamo utilizzare i dati per migliorare le decisioni finanziarie, la pianificazione interna. Ancora, possiamo esaminare le tendenze e i comportamenti dei consumatori per anticipare i trend e fornire ciò che più desiderano. I dati generati sono innumerevoli, sia che essi vengano usati o meno. Innovare è fondamentale nella nuova economia basata sui dati per eccellere. Innovazione non è solo un termine utilizzato per dare lustro all'argomento dei Big Data, innovare è ciò che trasforma un'impresa stagnante o indirizzata al fallimento, in un'impresa di successo.

L'innovazione passa attraverso migliori ricerche di mercato, attraverso le ricerche in tempo reale. La necessità in un ambiente competitivo sempre più dinamico come quello odierno è di innovare e implementare ricerche

⁴⁰ Behnam Tabrizi, Ed Lam, Kirk Girard, Vernon Irvin, Digital Transformation Is Not About Technology, 13 marzo 2019, Harvard Business Review, <https://hbr.org/2019/03/digital-transformation-is-not-about-technology>

⁴¹ <https://www.tableau.com/it-it/learn/articles/data-driven-decision-making>

⁴² Donald Lay, Senior Business Intelligence Manager di Charles Schwab Corporation

di mercato che siano in grado di eseguire analisi in tempo reale. Abbiamo visto come l'innovazione passi anche attraverso un processo decisionale più snello e rapido. I dati permettono di prendere decisioni più rapidamente, ma non solo, permettono una qualità più elevata dei processi e di prendere decisioni che siano basate sulle evidenze. Lo sviluppo del prodotto è sicuramente un'altra chiave dell'innovazione. Abbiamo visto come l'applicazione di modelli basati sui Big Data permetta lo sviluppo di nuovi prodotti quanto più vicini a ciò che i riscontri dei clienti evidenziano come loro richieste e preferenze. Inoltre, permettono di rendere il processo di sviluppo più repentino, fattore chiave oggi. Questo comporta un aumento della produttività che quindi risente dell'innovazione che i Big Data possono promuovere.

L'innovazione passa attraverso gli algoritmi di machine learning, che giocano un ruolo chiave nella promozione dell'innovazione e sono una delle maggiori applicazioni dei Big Data, come vedremo più avanti.

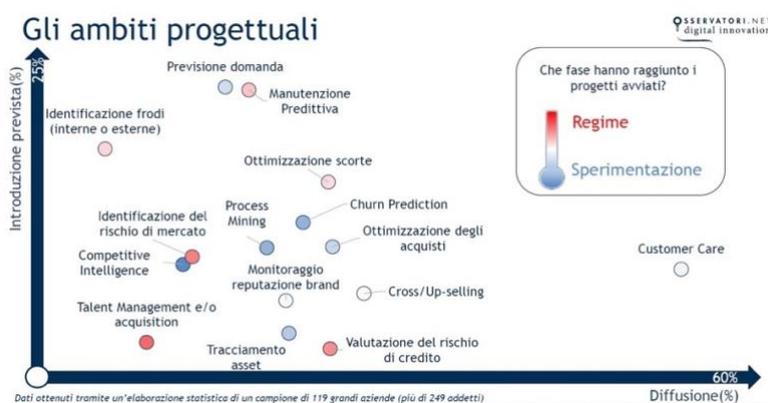


Grafico 3.3: Ambiti progettuali dei Big Data, (Fonte: osservatori.net)

3.3 Machine learning: applicazioni nei vari settori economici di algoritmi di machine learning e lo “statistical learning”

Machine learning e statistical learning trovano applicazione in moltissimi settori. Il loro campo d'azione è vasto e le problematiche che possono essere risolte attraverso il loro ausilio sono delle più disparate.

A favore di ciò viene riportato in seguito un grafico rappresentante uno studio condotto dal *McKinsey&Co Global Institute* dove vengono individuati 120 potenziali casi d'uso degli algoritmi di machine learning in 12 settori diversi. Quindi poi sono stati ascoltati i pareri di oltre 600 esperti riguardo il possibile impatto di queste applicazioni. Dallo studio emergono una vastità di possibili impieghi dell'apprendimento automatico.

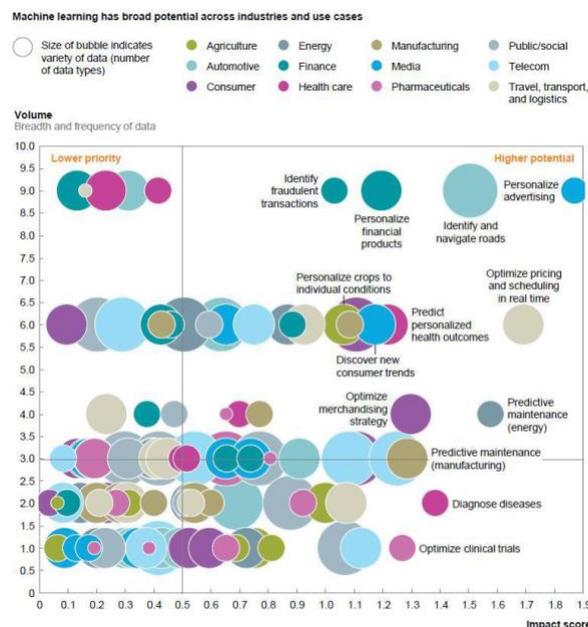


Grafico 3.4: Studio sul potenziale dei possibili impieghi dei Big Data, (Fonte: McKinsey Global Institute, Big Data: The next frontier of innovation, competition, and productivity)

Il grafico presenta sull'asse delle ascisse, quindi l'asse x , l'impatto potenziale sotto forma di un punteggio ottenuto sulla base di sondaggi condotti sui 600 esperti intervistati nello studio. L'asse delle ordinate o asse y presenta invece il volume di dati disponibili, occorre specificare che non si considerano frequenza e varietà.

La grandezza delle "bolle", ossia i vari cerchi colorati dal differente diametro, rappresenta l'eterogeneità delle fonti dei dati che sono disponibili. Più saranno le diverse tipologie d'uso in grado di unire ricchezza dei dati e opportunità, maggiori saranno le dimensioni dei cerchi nel quadrante superiore a destra.

Le bolle che si trovano in questo quadrante rappresentano le aree su cui le imprese dovrebbero focalizzare l'uso delle loro risorse per implementare il machine learning e avviare un processo di trasformazione che le muti in imprese *data-based*. Osservando il grafico si nota facilmente come *personalize advertising*, ovvero pubblicità personalizzata, ottimizzazione dei prezzi e calendarizzazione in tempo reale, veicoli automatizzati, ossia dotati di intelligenza artificiale e sensori, il *Routing*, la predizione di risultati che siano personalizzati sulla salute, l'identificazione di frodi e la personalizzazione di prodotti finanziari, l'ottimizzazione di strategie di *merchandising* e la scoperta di nuove tendenze dei consumatori, siano le applicazioni dalle opportunità più elevate, come dimostra il fatto che tutti questi casi d'uso si trovano proprio nel quadrante in alto a destra.

I casi d'uso appena elencati possono essere suddivisi in diverse categorie, per l'esattezza 4:

1. *Personalizzazione dei prodotti e servizi*: operata in diversi settori, dai prodotti confezionati alla finanza e alla sanità. Questa categoria rappresenta un'area non ancora pienamente esplorata dalle aziende, che non hanno ancora colto a pieno l'opportunità, seppur sia migliorata molto la situazione rispetto al periodo storico in cui il report è stato realizzato, sinonimo della repentinità dell'innovazione nel mondo dei Big Data.

2. *Analisi predittiva*: a questa categoria riconduciamo ad esempio la segmentazione dei consumatori usando come criteri modelli di acquisto o di abbandono, l'uso di algoritmi per prevenire e identificare frodi. Anche la capacità di ottenere diagnosi di malattie dall'analisi di dati e biopsie fa parte di questa classe.
3. *Ottimizzazione strategica*: qui troviamo ovviamente l'ottimizzazione del merchandising, ma anche l'ottimizzazione dei gruppi di lavoro e l'ottimizzazione dei prezzi.
4. *Ottimizzazione in tempo reale*: riguarda gli usi che ottimizzano la logistica in tempo reale, come i macchinari automatici e la calendarizzazione in tempo reale.

Lo studio di McKinsey riporta inoltre una mappa termica, la quale permette di indicare la relativa facilità con cui si riesce a catturare il “value potential” nei vari settori.

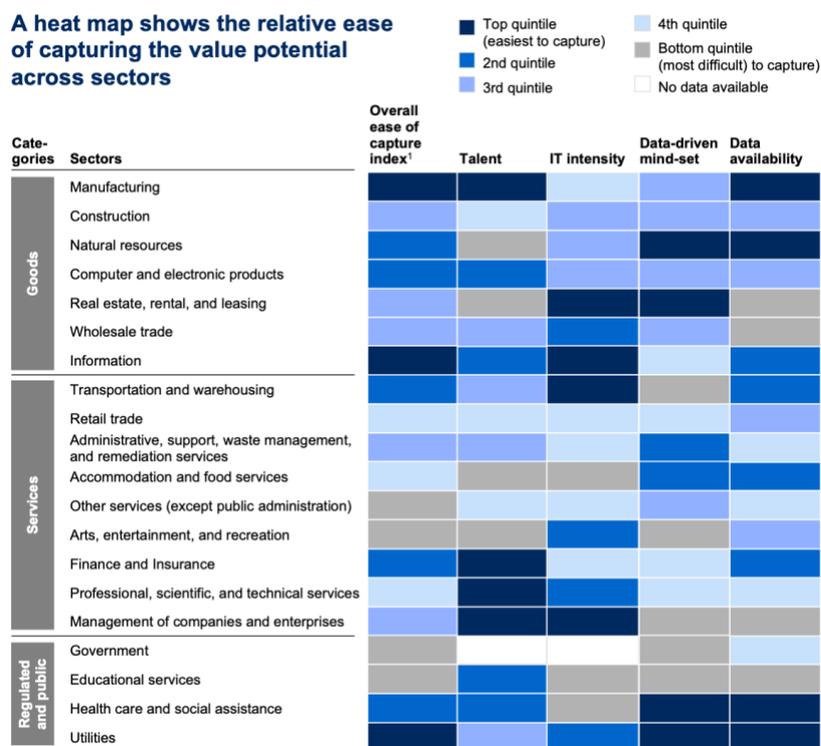


Tabella 3.2: *Mappa termica dei vari settori*, (Fonte: McKinsey Global Institute, Big Data: The next frontier of innovation, competition, and productivity)

Attraverso la leggenda presente in alto a destra comprendiamo come più scura sarà la tonalità del colore più sarà semplice catturare valore potenziale. Inoltre, sono riportati altre analisi compute sui settori che seguono la stessa logica per la loro comprensione, come ad esempio per la disponibilità dei dati, ultima a destra. Non è un caso che quest'ultima sia sempre di colore più scuro nei settori dove la facilità di ottenere valore potenziale sia primo o secondo quintile, vi è una chiara correlazione.

Il machine learning ha quindi svariati casi d'uso, come la prevenzione delle frodi, di cui in precedenza si è già parlato. Essa è una delle più comuni forme applicative dei Big Data e nella sua esecuzione vengono adoperati gli algoritmi di machine learning. Ma non è l'unica applicazione, abbiamo citato anche la pubblicità personalizzata. L'apprendimento automatico permette di implementare campagne pubblicitarie altamente personalizzate, attraverso ad esempio l'analisi della cronologia dei browser, i social network, la cronologia

degli acquisti. Queste permettono di fornire previsioni che surclassano quanto un singolo individuo può ottenere nello stesso lasso di tempo.

Altra forma d'uso del machine learning sono i contenuti raccomandati o *“recommending content”*. Si tratta di algoritmi in grado di fornire e di mostrare al consumatore solo ciò che risulta affine ai suoi gusti e ai suoi interessi. Un po' come accade per il *feed* dell'*homepage* di Instagram, dove il contenuto suggerito, viene scelto, ad esempio, sulla base dei precedenti mi piace o in base alla tipologia dei profili già seguiti.

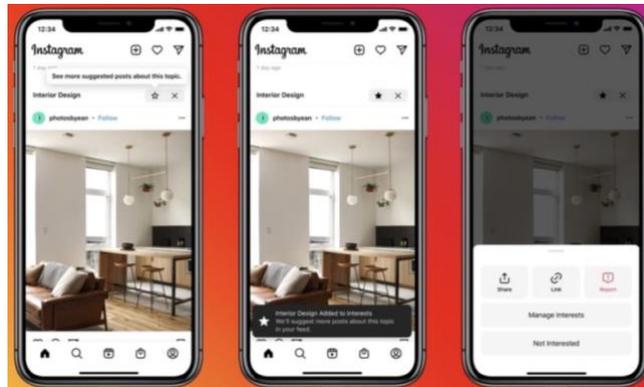


Figura 3.6: Esempio di contenuto suggerito all'interno di Instagram, (Fonte: <https://www.howtosocial.it/news/post-suggeriti-di-instagram-ecco-cosa-cambia/>)

Attraverso l'apprendimento automatico è possibile effettuare una più profonda segmentazione dei potenziali consumatori e avere migliori prospettive per quel che riguarda il targeting. In questo caso gli esempi più classici per quel che concerne segmentazione e targeting sono Google e Facebook. Il primo memorizzando le ricerche che ogni utente compie, è in grado di proporre pubblicità affini alla cronologia di navigazione, anche attraverso l'analisi delle parole digitate, tramite *sentiment analysis*⁴³. Proprio per questo fine Google nel 2000 ha lanciato una piattaforma apposita chiamata *Google AdWords*, che permette di inserire spazi pubblicitari, selezionati sulla base di parole chiave digitate dagli utenti, che andranno ad apparire sopra i risultati gratuiti. Le inserzioni sono spesso accompagnate dalla parola annuncio che precede l'*URL* del sito.

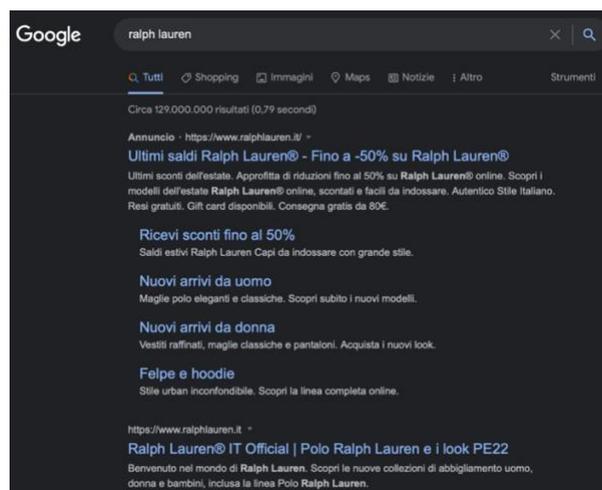


Figura 3.7: Esempio di ricerca in cui appare l'uso del sistema Google AdWords

⁴³ Si tratta di una tipologia di analisi che interpreta il linguaggio per identificare informazioni, opinioni, il cosiddetto *“sentiment”* dal testo

Facebook, oggi *Meta*, dal canto suo ha una piattaforma differente. Attraverso Facebook gli utenti esprimono opinioni, interessi tramite pulsanti di interesse, i famosi “like”, a cui poi hanno fatto seguito i commenti, le condivisioni e ultime in ordine di tempo le *reactions*. Tutte queste informazioni vengono utilizzate per creare profili di interesse dei vari utenti. Questi vengono poi utilizzati per proporre inserzioni attraverso la piattaforma *Facebook Advertising* (oggi *Meta Advertising*). Tramite ciò si è in grado di effettuare inserzioni altamente personalizzate. In questo senso il funzionamento è simile a quanto accade su di un altro social, appartenente sempre “all’universo” Meta, ovvero Instagram.

La differenza tra le due piattaforme risiede nella tipologia di targeting messa in atto. Google mette in atto un targeting basato sui bisogni, Meta e Facebook praticano un targeting basato sugli interessi. Entrambe però funzionano grazie al machine learning e ai dati.

Questi visti fin ora sono alcuni dei più comuni applicativi del machine learning, proseguendo nel capitolo si andrà ad analizzare nel particolare alcuni degli algoritmi citati al fine di comprendere meglio l’importanza dell’argomento e i casi di studio che verranno proposti.

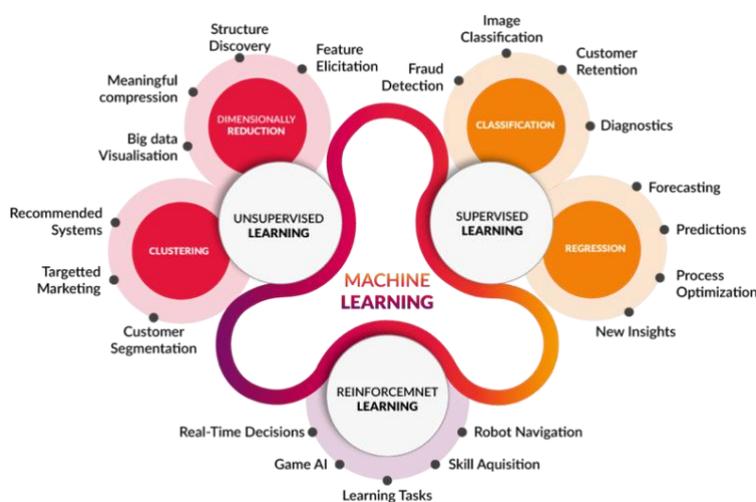


Figura 3.8: Esempi di applicazioni del machine learning, (Fonte: <https://www.developersmaggiori.it/blog/machine-learning-la-scienza-delle-decisioni-automatiche/>)

3.4 Il “Recommender System”

Il *recommender system*, noto anche come *recommender engine*, in italiano sistema di raccomandazione o motore di raccomandazione, è uno degli esempi più noti di algoritmi di machine learning utilizzati per analizzare dati. Esso consiste in un sistema di filtraggio dei contenuti, che opera per creare raccomandazioni personalizzate, come suggerisce il nome, in modo da guidare gli utenti nelle loro scelte. Esso mira ad anticipare le preferenze che un utente potrebbe esprimere nei confronti di oggetti o servizi. I settori dove viene fatto uso di questo sistema sono numerosi, viene utilizzato per musica, film, libri, notizie, nel mondo dei servizi finanziari e delle assicurazioni, sui social media. A farne un uso copioso sono spesso le piattaforme dalle grandi dimensioni, come gli e-commerce, spesso con lo scopo di aumentare engagement e quindi quello che è il tempo trascorso sulla piattaforma e ridurre il tasso di abbandono da parte degli utenti. Per fare ciò si pone come obiettivo quello di fornire ai consumatori suggerimenti su servizi o prodotti che potrebbero essere per

lui interessanti e di cui potrebbe non essere a conoscenza. Si pongono quindi come una valida, oltre che innovativa, alternativa agli algoritmi di ricerca che richiedono un input dato dall'utente. Il recommendation system genera le raccomandazioni sulla base di tre approcci:

- *Approccio collaborativo*, anche noto come *collaborative filtering*. Sotto l'aspetto realizzativo il più semplice, di contraltare produce raccomandazioni poco profonde e meno interessanti.
- *Approccio basato sui contenuti*, ovvero *content-based filtering*. Molto complesso nella realizzazione ma permette di generare raccomandazioni profonde e accurate. Richiede molti dati per essere utilizzato.
- *Approccio ibrido*, anche conosciuto con il termine inglese *hybrid recommender system*. Approccio di livello avanzato, consiste nell'utilizzo combinato dei due precedenti sistemi.

3.4.1 Approccio collaborativo: il collaborative filtering

L'approccio collaborativo parte dal presupposto iniziale che utenti simili abbiano preferenze e gusti simili. Se più utenti hanno mostrato gusti affini nel passato allora questo continuerà anche nel futuro. Sulla base di questo assunto genera raccomandazioni sfruttando la similarità che gli utenti posseggono. Inoltre, a questo aggiunge la creazione di un modello descrittivo basato sul comportamento tenuto dal soggetto. Per fare ciò occorre immagazzinare enormi quantità di dati. Questo particolare modello di collaborative filtering è definito *model-based*.

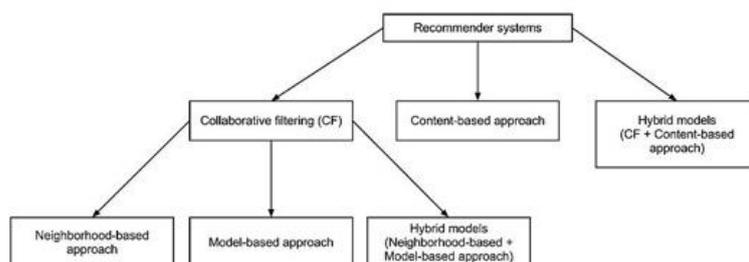


Figura 3.9: Approcci del recommendation system, (Fonte: https://it.wikipedia.org/wiki/Sistema_di_raccomandazione)

I dati per creare tale modello sono raccolti in modo implicito o esplicito. La raccolta esplicita può avvenire:

- Attraverso la richiesta diretta ai soggetti di fornire valutazioni sulla base di una scala predefinita⁴⁴
- Tramite una richiesta diretta all'utente di svolgere una ricerca
- Tramite richiesta di classificazione in ordine di preferenza di elementi
- Ponendo gli utenti di fronte a scelte tra elementi da compiere sulla base delle preferenze

La raccolta indiretta, ovvero implicita, avviene invece in modalità diverse, quali:

- Memorizzazione degli oggetti visualizzati ed acquistati negli *shop online* da parte degli utenti
- Analisi dei comportamenti degli utenti sui social network
- Analisi del numero di visualizzazioni dello stesso oggetto da parte degli utenti

⁴⁴ In genere per queste tipologie di valutazioni viene adoperata la scala di *Lickert*, un esempio è l'assegnazione di un valore numerico da 1 a 5 per un determinato attributo di un ipotetico prodotto

Come detto in precedenza, a adoperare il collaborative filtering sono spesso gli e-commerce, su tutti Amazon. L'algoritmo di Amazon assume che un utente che acquista un determinato oggetto, con ogni probabilità sarà interessato ad acquistare oggetti già scelti da utenti con caratteristiche simili a lui, che hanno compiuto lo stesso tipo di acquisti. Tale meccanica all'interno del sito viene proposta suggerendo una selezione di prodotti che altri utenti hanno acquistato insieme o a seguito dell'acquisto del primo prodotto citato.



Figura 3.10: Amazon's Recommender system, (Fonte: <https://ai-society.github.io/Collaborative-Filtering/>)

Simile è quanto avviene sui social, come Facebook e Instagram, per i profili suggeriti.

Anche Spotify fa uso del collaborative filtering. L'algoritmo adoperato da Spotify confronta la cronologia di ascolti di un utente A , se questo avrà apprezzato le canzoni x , y e z e analizzando gli ascolti di un utente B , questi avrà ascoltato e apprezzato x e y , l'algoritmo provvederà a suggerire all'utente B di ascoltare la canzone z . L'algoritmo è quindi in grado di capire se due canzoni sono simili, se utenti simili le ascoltano, oltre che poter riconoscere utenti simili, se ad esempio ascoltano la stessa canzone. Questo permette all'algoritmo di imparare e quindi perfezionare la sua capacità di suggerire canzoni sulla base delle somiglianze in termini di gusti degli utenti. Questo tipo di funzionamento nel tempo è stato però sostituito con un approccio più efficace, ovvero "due canzoni sono simili se sono inserite da un utente nella stessa playlist". Questo perché l'attuazione dell'approccio collaborativo richiede la presenza di determinati fattori, ossia:

- *Disponibilità di quantitativi di dati elevate*
- *Scalabilità*, questo perché i sistemi hanno a che fare con numerosi prodotti e utenti che richiedono una potenza di calcolo elevata
- *Variabilità*

Il fatto che due utenti ascoltino una canzone A e una canzone B non sempre corrisponde al fatto che i due artisti sono simili. Se però le due canzoni vengono messe all'interno della stessa *playlist*, allora questo è un segnale più forte che le canzoni hanno qualcosa in comune. Lo studio basato su playlist fornisce quindi degli insight più profondi e accurati.

Principale problematica del collaborative filtering risulta essere la *cold start*, ossia la situazione in cui un utente è da poco iscritto alla piattaforma e non potendo questa possedere sufficienti dati, fornirà suggerimenti poco precisi.

3.4.2 Approccio basato sul contenuto: content-based filtering

L'approccio content-based associa la descrizione di un elemento con il profilo degli utenti. Questo per suggerire oggetti simili a quelli già acquistati in precedenza dai consumatori. Ogni oggetto è dotato di una descrizione, ottenuta tramite attributi e parole chiave. Attraverso lo storico dei comportamenti degli utenti è possibile attribuire un valore ai vari attributi che indica un peso relativo attribuito a tale attributo. Il peso è indice dell'importanza attribuita a tale attributo⁴⁵, similmente a come accade nel modello del valore attuale noto anche come modello di *Fishbein*. Per calcolare il peso si possono usare diversi metodi, dalla semplice media ponderata sino a reti neurali o analisi dei cluster. Generalmente i riscontri che i consumatori forniscono di loro spontanea volontà sono utilizzati per attribuire maggiore o minore peso a determinate caratteristiche. Questi consistono spesso in "like" o nelle recensioni positive.

Un sito a far uso di tale approccio è quello di *Pandora Radio*. Il sistema di questo sito possiede una lista di attributi relativi a cantanti e canzoni che gli ascoltatori prediligono e attraverso l'algoritmo riesce a leggere le caratteristiche di canzoni o artisti per individuare quali canzoni e artisti trasmettere col fine di assecondare i gusti degli ascoltatori. Il sistema sfrutta inoltre i like così come indicato in precedenza per proporre suggerimenti ancor più precisi.

Il principale problema del content-based filtering risulta essere applicare le preferenze relative ad un contenuto su diverse tipologie di servizi o prodotti. Un sistema in grado di generare suggerimenti solo relativamente a un tipo di prodotto o servizio ha un valore complessivamente inferiore rispetto a un sistema che utilizza lo stesso approccio per generare suggerimenti su più contenuti. Riprendendo l'esempio del sistema di Pandora Radio, esso è in grado di proporre recommendation solo per quel che concerne quali artisti trasmettere, se fosse possibile applicare lo stesso sistema ad altri contenuti, senza dover riprogrammare ogni volta il tutto, esso risulterebbe avere un valore più alto oltre che un'utilità maggiore. Questo anche in virtù del fatto che tale approccio non richiede una disponibilità di dati al lancio tanto grande come invece pretende l'approccio collaborativo.

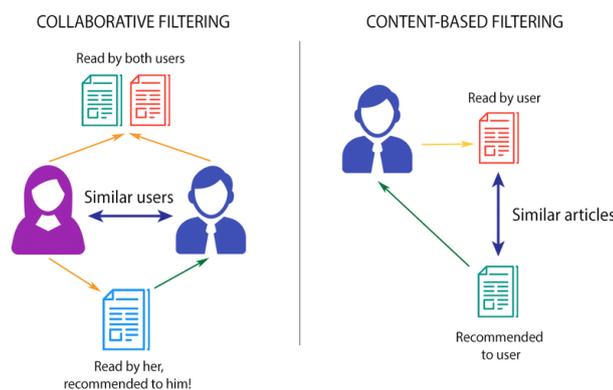


Figura 3.11: Schematizzazione dei due approcci principali, (Fonte: https://www.researchgate.net/figure/Content-based-filtering-vs-Collaborative-filtering-Source_fig5_323726564)

⁴⁵ Van Meteren R., van Someren M., "Using Content-Based Filtering for Recommendation",

3.4.3 Approccio ibrido: Hybrid Recommender System

Il sistema ibrido è la risultante dell'uso combinato dei due precedenti approcci. Le combinazioni possibili sono diverse, solitamente si adoperano nei seguenti modi:

- Utilizzando in modo sperato i due approcci per ottenere suggerimenti. I risultati vengono poi combinati per ottenere un risultato finale unico
- Viene adoperato uno dei due approcci come principale e l'altro va ad arricchire i risultati. Ad esempio, se si sceglie un approccio collaborativo, si implementeranno in aggiunta elementi del modello basato sui contenuti
- Si crea un modello unico. Prendendo elementi dei due approcci sopracitati si va a modellare un sistema ibrido

I sistemi ibridi permettono di ottenere risultati decisamente più accurati e per tale motivo risultano più efficaci. Inoltre, essi permettono di superare problematiche appartenenti agli approcci collaborativo e basato sui contenuti, come la già citata “cold start” o la “sparsità”. Proprio per la loro elevata efficienza e accuratezza, tali sistemi si caratterizzano per una forte complessità. Questa rappresenta la problematica principale per quel che concerne la loro implementazione, in quanto richiedono di forti investimenti, affinché si integrino a pieno delle logiche di business e con le attività di ricerca e sviluppo. Appare quindi evidente come un sistema ibrido implementato in modo efficiente richieda alle imprese determinati fattori, quali:

- Una base dati imponente, contraddistinta da un volume di Big Data analizzabili elevato e da elevata varietà dei dati
- Una base utenti vasta e diversificata, per contrastare il fenomeno della sparsità
- Risorse di calcolo sufficienti per l'implementazione, nonché risorse finanziarie sufficienti per effettuare l'integrazione e risorse tecnologiche, come i *data center*

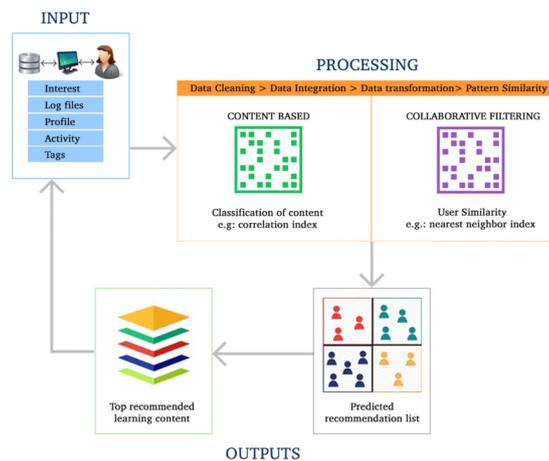


Figura 3.12: Schema Hybrid Recommender System, (Fonte: https://www.researchgate.net/figure/System-architecture-of-Hybrid-recommendation-system_fig3_337938732)

3.4.4 Caso studio: La Netflix Experience

Netflix rappresenta uno degli esempi più famosi di utilizzo dei sistemi ibridi. Di fatto la piattaforma *streaming* americana suggerisce i film sulla base delle abitudini di visione e di ricerca di utenti dal profilo simile,

seguendo quindi l'approccio collaborativo, ma anche sulla base dei contenuti visualizzati in precedenza dall'utente, suggerendo film che l'algoritmo dopo aver analizzato ritiene siano in linea con i gusti del fruitore di contenuti, implementando quindi anche il content-based filtering.

Netflix è stata fin da sempre una delle pioniere nel settore delle internet TV, proprio per questo ha da sempre curato il suo sistema di generazione di raccomandazioni. Ha sviluppato e implementato vari algoritmi sia machine learning che di statistical learning, utilizzando approcci come la regressione o il clustering, quindi in approcci supervisionati e non. Ad incidere in modo positivo fu in questo senso il concorso lanciato da Netflix denominato "*Netflix Prize*", lanciato dalla società americana nel 2006, che metteva in palio un milione di dollari. Il concorso consisteva in una sfida tra vari team, il vincente avrebbe ottenuto il premio in denaro. Per vincere occorreva generare delle recommendations che fossero migliori di quelle generate dal sistema prioritario di Netflix del 10%, sfruttando un dataset di decine di milioni di voti attribuiti ai vari film. La competizione ha richiesto diverso tempo per concludersi e permettere l'assegnazione del premio, che fu consegnato nel settembre del 2009 alla squadra nota come "*BellKor's Pragmatic Chaos*". Ma la competizione lanciata da Netflix permise soprattutto lo sviluppo di numerosi algoritmi che furono poi adoperati in diversi settori e in diversi mercati. Il premio in denaro risultava quindi un *escamotage* per alimentare lo sviluppo di algoritmi dal maggior valore possibile.

Netflix stessa, per il suo recommender system, fa uso di numerosi algoritmi diversi tra loro. Combinati insieme questi restituiscono un'esperienza, nota come "*Netflix experience*", che permette di andare in contro ai gusti dell'utenza e di rispondere alle loro esigenze. Le ricerche condotte dal colosso dello streaming hanno messo in luce come un utente medio perda interesse verso i contenuti disponibili sulla piattaforma dopo un intervallo di attesa che si aggira tra i 60 e i 90 secondi, all'interno del quale ha dovuto scegliere tra all'incirca 10-20 titoli, tra serie televisive e film. La ricerca ha dimostrato come, all'interno di quel lasso di tempo d'attesa, sia auspicabile che il soggetto trovi e scelga un contenuto di suo gradimento, altrimenti il suo interesse diminuirà in modo crescente nel tempo sino a convergere con l'abbandono. Risulta quindi essenziale che in questo periodo di tempo il recommender system generi recommendations altamente efficienti, in modo che l'utente abbia una probabilità alta o certa di trovare un contenuto di suo gradimento, evitando l'abbandono.

Lo streaming di contenuti, per sua natura, risulta un servizio in grado di permettere una raccolta di dati notevole, la quale permette di fornire informazioni su cosa viene visto dagli utenti, quando viene visto e come. I dati permettono quindi di rispondere a domande come: su quale dispositivo viene visto il contenuto? In che giorno della settimana? A che ora? Con quale intensità?

Inoltre, tali dati forniscono informazioni sulla posizione del titolo che viene scelto dall'utente, nonché sulla posizione dei titoli che vengono suggeriti ma scartati⁴⁶. Tutti questi dati vengono sfruttati dal sistema di raccomandazioni di Netflix per generare la "*homepage*" degli utenti, la quale è personalizzata per ognuno dei

⁴⁶ Carlos A. Gomez-Uribe, Neil Hunt, The Netflix Recommender System: Algorithms, Business Value, and Innovation

clienti della piattaforma. Oltre i due terzi del tempo impiegato per la ricerca di contenuti sulla piattaforma da un utente, sono trascorsi sulla homepage, a prescindere dal dispositivo utilizzato. Questo fa sì che sia il luogo con il maggior numero di suggerimenti e raccomandazioni di tutta la piattaforma. La pagina iniziale presenta un *layout* con una struttura a matrice. Ogni istanza rappresenta una recommendation ed ogni riga presenta contenuti relativi al tema del suggerimento. Il tema di ogni riga è indicato attraverso etichette collocate in alto a sinistra rispetto alla riga stessa, di modo da rendere ancora più intuitivo per l'utente l'ambiente in cui si "sposta".

La struttura nella sua impostazione non è predefinita, varia a seconda del dispositivo utilizzato e dalle informazioni raccolte per costruire la migliore esperienza possibile sulla piattaforma; tuttavia, vi sono delle linee guida che suggeriscono la presenza in media di 40 righe e circa 75 contenuti per singola riga, posizionati di modo che possano essere visualizzati attraverso lo scorrimento e non in via generale con vista dall'alto. Come detto in precedenza, un algoritmo si occupa di determinare quali contenuti debbano essere inseriti nelle righe.

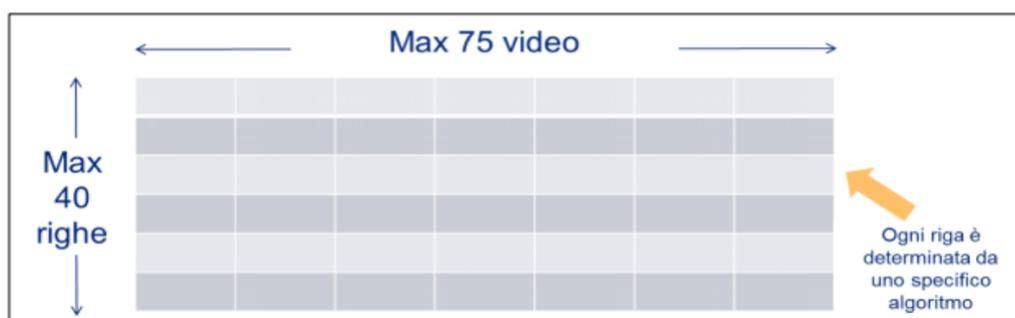


Figura 3.13: *Struttura tipica della homepage di Netflix.* (Fonte: ITMedia consulting)

Gli algoritmi di cui si avvale Netflix sono diversi, 7 però sono quelli più importanti, che vanno a comporre il recommender system di Netflix. Essi sono di seguito descritti:

- *Personalised Video Ranker (PVR)*: questo è l'algoritmo che si occupa di come comporre le righe per ogni genere. Esso svolge la funzione di ordinare l'intero catalogo di contenuti o un sottogruppo individuato sulla base del genere, in modo personalizzato per adattarsi ad ogni utente. Come indicato prima, ogni riga si compone di contenuti appartenenti ad un genere. Il genere ovviamente non è l'unica componente che incide, questo spiega come sia possibile che due utenti possano avere righe con generi simili ma film diversi. Al genere e al comportamento dell'utente viene inoltre abbinata una componente, legata alla popolarità del film. Questo permette di ottenere risultati qualitativamente migliori.
- *Top-N Video Ranker*: è l'algoritmo che si occupa di generare le recommendations per la classifica nota come "Top Picks". Esso ha il compito di individuare le migliori recommendations tra tutte quelle generate, quindi quelle dall'ipotetico punteggio più elevato, per andare a comporre un set limitato di raccomandazioni. Queste sono scelte attraverso l'ausilio di metriche specifiche, che permettono di

scegliere i contenuti video individuati come migliori, tra quelli che sono a loro volta individuati dall'algoritmo PVR.



Figura 3.14: Homepage di Netflix, (Fonte: medium.com)

- *Trending Now*: in italiano noto come “I titoli del momento”. L’algoritmo ha il compito di comporre la riga inserendo contenuti servendosi di criteri di personalizzazione, ma anche della rilevanza temporale dei trend. L’algoritmo suddivide le tendenze in due tipologie:
 - Tendenze dovute a festività, ricorrenze o simili, che di solito si ripetono ciclicamente di anno in anno. Ad esempio Natale o San Valentino.
 - Tendenze momentanee, dovute a eventi improvvisi.
- *Continue watching*: tradotto come “continua a guardare”. Questo algoritmo compone la riga dei contenuti di cui si è già iniziata la visione. Il loro ordine dipende da diversi fattori, come il punto di interruzione precedente, se ad esempio ad inizio contenuto o a metà, il dispositivo utilizzato, l’orario. Il fine ultimo è quello di analizzare le nuove abitudini di consumo dei contenuti. Questo avviene anche attraverso l’analisi di ciò che l’utente vuole continuare a vedere, rivedere o non rivedere.
- *Video similarity*: si occupa di inserire i contenuti della riga “Perché hai guardato...”. L’algoritmo non genera suggerimenti sulla base della personalizzazione ma sceglie contenuti ritenuti simili o identici al film o alla serie tv vista. C’è però un elemento di personalizzazione nello scegliere quali elementi simili al contenuto visto, tra tutti quelli selezionati, posizionare all’interno della riga. In questo frangente incidono elementi che permettono di generare righe differenti per utenti diversi che però hanno visto lo stesso contenuto.
- *Page generation date*: è l’algoritmo incaricato di utilizzare gli *outcome* dei precedenti algoritmi per generare la homepage degli utenti. Le righe utilizzabili per ogni singolo utente sono innumerevoli, questo algoritmo si occupa di scegliere quelle più rilevanti per costruire la homepage personalizzata. Non c’è una linea guida o template da seguire, questo permette all’algoritmo di genare in piena libertà la schermata iniziale proponendo righe in determinate situazioni e nascondendole per mostrane di diverse in altre. Può ad esempio mostrare la riga “continua a guardare” come prima in alto se ad esempio accediamo da computer o mostrarla più in basso se accediamo da *smartphone*.

- *Evidence selection*: algoritmo che si occupa del decidere quali informazioni mostrare di un contenuto, nella zona chiamata *evidence*. Questa si trova in alto in ogni contenuto selezionato. Qui possiamo trovare informazioni come i premi vinti dal film oppure la somiglianza a contenuti che ci sono piaciuti. Questo algoritmo inoltre ha il compito di scegliere quali immagine mostrare come anteprima. Tale scelta è cruciale per attrarre l'attenzione dell'utente ed è fortemente personalizzata.



Figura 3.15: Esempio di outcome degli algoritmi “Page generation date” ed “evidence selection”. Entrando su due account differenti la homepage presenta contenuti in parte uguali in parte diversi, (Fonte: Netflix.com)

- *Search algorithms*: gli algoritmi visti fino ad ora costituiscono circa l'80% dell'esperienza relativa al recommender system di Netflix. Il restante 20% è composto dalle ricerche che ogni utente compie. Anche la ricerca è supportata da diversi algoritmi che la rendono una forma di recommendation. Questi algoritmi si basano su dati dei contenuti già visti e su ricerche effettuate. Se ad esempio si cerca “ita” la ricerca darà come risultati contenuti aventi all'interno del titolo “ita” ma anche, come è possibile osservare nell'immagine in basso, suggerimenti per serie televisive italiane. Questo perché uno degli algoritmi cerca di prevedere cosa andremo a cercare.⁴⁷

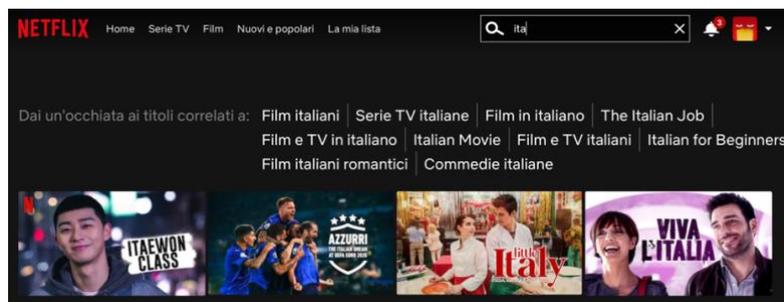


Figura 3.16: Esempio di Search Algorithm, (Fonte: Netflix.com)

Il miglioramento del recommender system permette a Netflix di ottenere numerosi benefici in termini di valore per il business. Il sistema risulta fondamentale nell'affrontare il cosiddetto “*momenth of truth*”, ovvero il momento in cui gli utenti entrano sulla piattaforma e iniziano la ricerca di un contenuto da visualizzare, trovandolo in poco tempo e risultando quindi soddisfatto. Come detto in precedenza, questo evita che si vada in contro all'abbandono, che può comportare in casi estremi anche la cancellazione dell'utente dalla piattaforma.

⁴⁷ Carlos A. Gomez-Urbe, Neil Hunt, The Netflix Recommender System: Algorithms, Business Value, and Innovation

Il sistema di raccomandazioni personalizzate per gli utenti permette inoltre di poter attingere pienamente a tutti i film del catalogo. Questo permette anche ai film di nicchia di poter ottenere maggiore visibilità, soprattutto rispetto a quanta ne avrebbero ottenuta su mezzi come la *TV broadcast*, dove probabilmente sarebbero risultati come degli insuccessi a causa di un'audience bassa, rendendo il ritorno economico tramite pubblicità basso e risultando quindi ingiustificabile l'occupazione dello slot temporale utilizzato per la loro trasmissione. Attraverso il suo recommender system, Netflix riesce a distribuire il consumo di contenuto su un insieme di titoli ampio. Dimostrazione di quanto affermato è la *Effective Catalog Size*, anche nota con l'acronimo *ECS*, del catalogo. L'ECS è una metrica utilizzata per misurare la diffusione di visualizzazione tra i vari contenuti del catalogo. Se il consumo deriva tutto da un singolo contenuto, su cui si concentra quindi tutta la visualizzazione, allora il valore sarà 1. Se invece le visualizzazioni sono diffuse in modo uguale tra i vari contenuti del catalogo, allora il valore dell'ECS sarà simile al numero dei contenuti nel parco titoli.

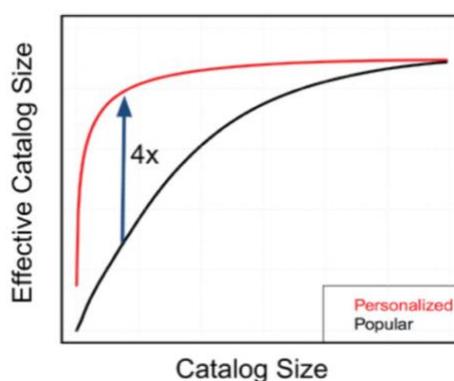


Grafico 3.5: Grafico dell'Effective Catalog Size, (Fonte: The Netflix Recommender System: Algorithms, Business Value, and Innovation)

La curva nera nel grafico rappresenta l'andamento dell'Effective Catalog Size in relazione al numero dei video più visti all'interno del catalogo, spaziando da 1 sino a N sull'asse delle x , dove N identifica l'effettivo numero di titoli del catalogo. La linea in rosso rappresenta invece l'ECS per i primi x numeri individuati attraverso l'algoritmo *PVR*. La personalizzazione ottenibile attraverso tale algoritmo permette di ottenere un rendimento e sfruttamento dell'insieme dei video del catalogo superiore a quello ottenuto considerando lo stesso set senza tale algoritmo, con un picco di circa quattro volte⁴⁸.

Questo viene dimostrato anche attraverso un'ulteriore metrica, denominata "*Take-Rate*" o tasso di accettazione. Questa rappresenta la frazione di raccomandazioni che vengono accolte e si traducono in una riproduzione effettiva del contenuto.

⁴⁸ Indicato nel grafico attraverso la freccia verso l'alto accompagnata dal "4x".

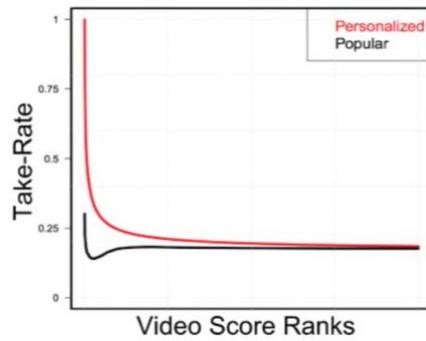


Grafico 3.6: Grafico del Tasso di accettazione, (Fonte: The Netflix Recommender System: Algorithms, Business Value, and Innovation)

Le due funzioni rappresentano l'accettazione di una raccomandazione, in funzione della popolarità e in funzione del "rank" attribuito dal PVR. La personalizzazione permette, come si evince dal grafico, di aumentare notevolmente il tasso di accettazione.

Attraverso quindi un uso ottimizzato del catalogo e alla Netflix experience ottenuta attraverso il sistema di raccomandazioni, Netflix ha ridotto notevolmente il suo *churn rate* mensile, ovvero il tasso di abbandono, giungendo a singola cifra bassa, ad oggi giustificata per lo più da problemi legati ai metodi di pagamento.

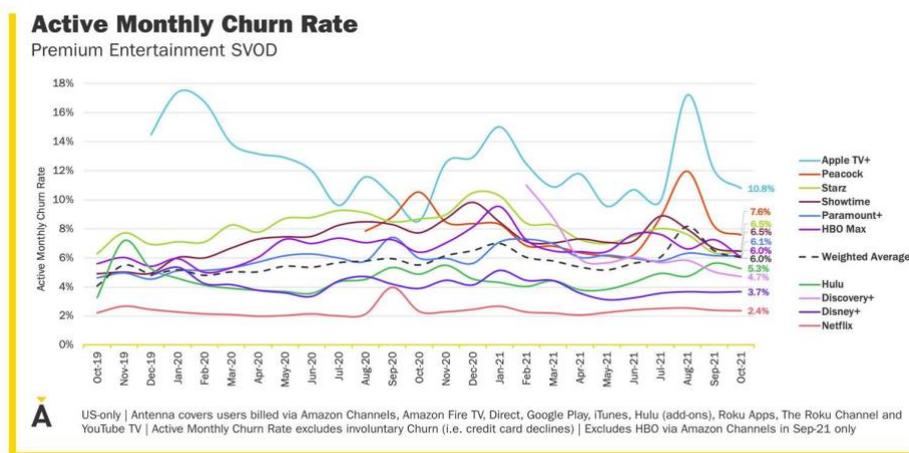


Grafico 3.7: Churn Rate per i principali piattaforme streaming nel 2021, (Fonte: <https://thestreamable.com/news/netflix-disney-plus-show-industry-best-churn-rates-apple-tv-plus-still-bleeding-subscribers>)

La riduzione nel tempo del tasso di abbandono permette di evitare la ricerca di nuovi abbonati che vadano a compensare le perdite ed aumenta il valore degli abbonati nel tempo. Questo comporta dei risparmi in termini di denaro davvero notevoli, nell'ordine dei miliardi dollari.

Il miglioramento del recommender system, quindi dell'esperienza all'interno di Netflix, passa non solo attraverso i vari miglioramenti incrementali dei già citati algoritmi, ma anche attraverso lo sviluppo e l'implementazione di nuovi. Lo sviluppo avviene secondo un processo definito iterativo-incrementale. Questo per adattarsi ad una domanda di mercato molto flessibile e in continuo mutamento. Il processo segue il seguente iter:

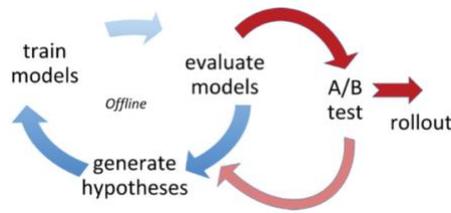


Figura 3.17: Fasi processo iterativo-incrementale, (Fonte: The Netflix Recommender System: Algorithms, Business Value, and Innovation)

La prima fase di questo processo si svolge off-line. Viene generato un nuovo algoritmo che viene addestrato attraverso dei *train model*, come solitamente avviene per tutti i nuovi algoritmi di machine learning. Questi *train models* sono basati su dati storici vengono utilizzati per ottenere delle raccomandazioni che siano il più accurate possibile. Successivamente, quando i test off-line iniziano a fornire i primi risultati positivi, vengono sviluppati dei *test A/B*. Questo tipo di test viene utilizzato in modo sistematico per valutare, in modo casuale, i risultati dell'algoritmo in relazione all'engagement di medio termine di ogni cliente rispetto al churn rate degli abbonamenti. Il risultato positivo di questo test determina l'entrata, in aggiunta o in sostituzione di altri algoritmi, del nuovo algoritmo all'interno dell'insieme di procedure che compongono il recommender system di Netflix. Nel caso in cui, il risultato del test fosse negativo o nel caso di risultati che siano identici o simili a quelli ottenuti con altri algoritmi, le possibilità sono due:

- Scegliere di abbandonare lo sviluppo di tale algoritmo e concentrare risorse di ogni forma sullo sviluppo e l'implementazione di nuovi algoritmi.
- Continuare l'impiego di risorse, intensificando lo sviluppo per ottenere dei risultati positivi

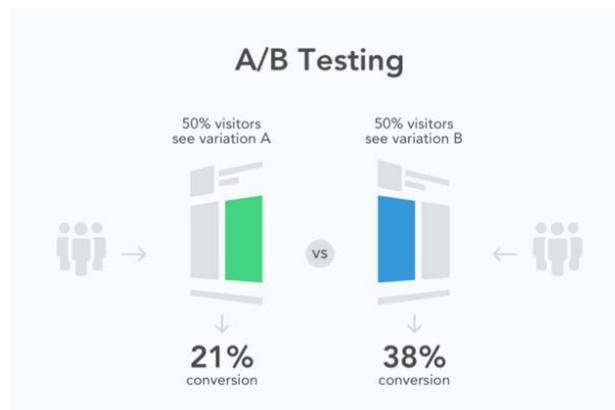


Grafico 3.8: Funzionamento di un Test A/B, (Fonte: <https://devopedia.org/a-b-testing>)

Ad oggi le attività di *R&D*, *research and development*, di Netflix sono messe in atto per diversi fini, tra cui:

- Creare la homepage personalizzata, così come visto nel paragrafo.
- Riuscire a scegliere le *evidence* migliori, che permettano di supportare al meglio le raccomandazioni fornite.
- Limitare le problematiche relative alla *coldstart* che, come visto nel corso del testo, può comportare diversi svantaggi, producendo errori.

- Per ovviare a questa situazione, molte piattaforme hanno iniziato a proporre dei sondaggi esplorativi ai nuovi iscritti, cosicché fin da subito esse possano entrare in possesso di informazioni utili e di valore per poter generare suggerimenti validi e utili. Un esempio è la richiesta di scegliere da un insieme di film noti e di genere vario quelli che sono dall'utente ritenuti di suo gusto. Questo permette sin da subito di conoscere in via generale i generi che attraggono il cliente.
- Riuscire a generare raccomandazioni “globali”, ovvero raccomandazioni valide per un'intera popolazione di un paese, per fornire quelli che sono i contenuti più popolari in quest'ultimo. L'obiettivo è migliorare lo sviluppo internazionale della piattaforma. Molto importanti sono i vincoli posti dalla lingua e dai sottotitoli disponibili per i vari contenuti. Questi ultimi sono molto costosi da realizzare ed è quindi più efficiente suggerire contenuti che possono essere compresi dagli utenti, anche per offrire un'esperienza più rilassante all'utente.

3.5 Per concludere

Il caso di Netflix è un esempio perfetto di come si può raggiungere il vertice nel proprio settore attraverso l'ausilio del machine learning, quindi attraverso l'uso dei Big Data e del loro vasto potenziale. Nel corso del capitolo sono stati forniti ulteriori esempi di applicativi in vari settori, questo con lo scopo di far comprendere come i Big Data siano divenuti uno strumento essenziale, a differenza di quanto si potesse pensare in passato, dove l'idea generale era considerare l'uso delle nuove tecnologie uno step successivo, che richiedeva investimenti possibili solo una volta ottenuti risultati. Oggi per ottenere risultati risulta necessario investire in anticipo sui dati e sull'informatizzazione.

CONCLUSIONE

*"You can't just ask customers what they want and try to give them that. By the time you get it built, they'll want something new."*⁴⁹

Steve Jobs

La decisione di inserire questa citazione per aprire il discorso conclusivo non è casuale. Nel corso dell'elaborato abbiamo visto come i dati abbiano tra le loro potenzialità, quella di poter essere utilizzati per prevedere la domanda di mercato, andando a sviluppare nuovi prodotti. La frase enunciata da *Steve Jobs* si spiega quindi così. Occorre sì produrre e fornire quello che i consumatori vogliono, ma chiedere direttamente al cliente cosa vuole può risultare inutile. I dati in questo senso hanno permesso di ovviare a questa problematica. L'obiettivo dell'elaborato fin dal primo capitolo è stato quello di far arrivare il lettore alla consapevolezza di queste potenzialità, renderlo conscio del valore dei Big Data. Non, quindi, quello di porre un punto, una conclusione, bensì di aprire un discorso che vertà sull'importanza dell'innovazione tecnologica e non solo. Ma raccogliere dati con il fine ultimo di utilizzarli per produrre valore è davvero innovazione? Il concetto di raccolta delle informazioni, come visto all'interno del secondo capitolo non è cosa nuova all'umanità. Le biblioteche esistono sin dall'antichità, eppure *Schumpeter*⁵⁰ affermava "*Fare le cose vecchie in modo nuovo, questa è innovazione*". Qui si spiega la definizione data dei Big Data. Non si tratta soltanto di moli di dati enormi raccolte e archiviate per futuro utilizzo, ma dello sviluppo di nuovi modi di utilizzare e implementare quantità di dati destinate nel tempo a crescere, andando a richiedere una costante innovazione delle tecniche e dei metodi di utilizzo. In questo senso, all'interno del terzo capitolo del presente testo abbiamo la riprova di quanto asserito pocanzi. Esso espone diverse applicazioni di dati, a cambiare è il fine per cui vengono utilizzati e le tecniche adoperate. Ma l'input di tutte queste operazioni sono sempre loro, i Big Data. I dati sono oggi un fattore di successo quasi imprescindibile, specialmente con le nuove forme di economia che, a seguito dell'innovazione, abbracciano quasi completamente il mondo digitale. Il caso studio riportato è un esempio di come oggi interi business si basino quasi esclusivamente sull'utilizzo di dati per fornire servizi eccellenti ai consumatori. Un uso dei dati ottimale permette quindi di eccellere. Si affronta quindi il discorso relativo alla propensione all'innovazione, che accompagna spesso le realtà vincenti. Occorre non aver paura dell'ignoto se si vuole emergere, come fatto dalla società Jhon Deere, produttrice di macchine agricole, eppure in grado di coniugare la sua attività con l'innovazione tecnologica portata dai Big Data⁵¹.

Lungi dal testo però, suggerire che i dati siano l'unico fattore di successo oggi. Ritorna in questo senso la frase di apertura. Occorre sì capire in anticipo cosa vuole il cliente, ma tenendo a mente che è quest'ultimo l'unica vera metrica del nostro successo. Avere i migliori algoritmi ed enormi quantità di dati risulta futile se

⁴⁹ "Non puoi semplicemente chiedere ai consumatori cosa vogliono e poi provare a darglielo. Non appena l'avrai costruito, loro vorranno qualcosa di nuovo"

⁵⁰ Economista austriaco ed ex ministro delle finanze della Repubblica austriaca

⁵¹ Vedasi capitolo terzo, paragrafo 3.2.3

non si ha la capacità di comprenderne il valore e le possibili applicazioni per far sì che il cliente sia soddisfatto. Il fine ultimo è sempre quello, i dati sono uno strumento che ci permette di ottenere tale risultato in modo più efficace ed efficiente. Occorre quindi comprendere che l'innovazione va abbracciata e non temuta, sfruttata per raggiungere i propri obiettivi. Solo comprendendo il valore dell'innovazione si può aspirare al successo, questo perché innovare è essere in grado di vedere i cambiamenti come un'opportunità di successo e non come una possibile minaccia.

RINGRAZIAMENTI

Con la presente sezione mi premeva ringraziare tutti coloro i quali mi hanno permesso di giungere a tale traguardo personale.

In primis ci tengo a ringraziare il Prof. Guido Tortorella Esposito, relatore per la mia tesi, che mi ha dato l'opportunità di studiare in profondità l'argomento, generando curiosità verso lo studio approfondito della materia. La ringrazio per la sua disponibilità e gentilezza.

Ringrazio poi la mia famiglia, senza la quale non sarei la persona che sono oggi, che mi ha permesso di percorrere e superare questa sfida. Grazie a mio zio Marco, per me come un padre, che mi ha insegnato la dedizione al lavoro e mi ha dato la possibilità di inseguire il mio sogno. Grazie a mia madre, che mi ha sostenuto tutti i giorni insegnandomi a non mollare mai davanti alle difficoltà. Ci tengo a ringraziare anche i miei nonni e mia sorella, per tutto il supporto e l'affetto che mi hanno dato in questo percorso e nella vita.

Ringrazio i miei amici, quelli che ci sono sin dall'infanzia, Davide, Camilla, Jules, Alessio e Matteo. Il vostro supporto e la vostra amicizia hanno forgiato una parte di quello che sono e punto ad essere, se sono arrivato sin qui lo devo anche al vostro supporto e al vostro affetto, alle ore perse tra il divertirci insieme e i momenti di sfogo che mi hanno permesso di mantenere la lucidità anche nel mezzo di una pandemia.

Ringrazio poi Cristiano, Matteo, Alberto, Giorgio, Nino e tutti gli altri, vi chiedo scusa se non vi nomino tutti. Quelli che da piccolo erano gli "amici del mare" e che nel tempo sono diventati gli amici di sempre, grazie per tutto quello che fate e avete fatto per me, a voi devo una parte di questo, "che gusto".

Ringrazio Andrea e Claudia, fedeli compagni di avventure in questo percorso. Insieme abbiamo affrontato questi anni di studio, perdendo nel mentre qualche pezzo, ma alla fine siamo riusciti ad arrivare al traguardo e farlo tutti insieme rende la fine ancora più dolce. Vi auguro ogni bene per il futuro e spero di continuare a coltivare insieme questa amicizia, a me molto cara.

Ringrazio poi tutti gli amici che non sono riuscito a inserire. Vi ringrazio per la vostra amicizia e il vostro tempo. Siete parte di me e spero questo basti a perdonarmi.

Mi preme poi ringraziare me stesso, nonostante le tante difficoltà che la vita mi ha posto, sono riuscito a trovare la forza per andare oltre e cercare di essere ogni giorno la versione migliore di me.

Grazie a chi ha lasciato un segno nella mia vita, sia chi è stato di passaggio, sia chi è con me sempre. Questo traguardo lo dedico a me ma anche a tutti voi, grazie.

SITOGRAFIA

<https://www.dartmouth.edu/basicfifty/basic.html>

<https://www.guinnessworldrecords.com/world-records/72695-most-computer-sales>

<https://www.cineblog.it/post/762646/film-anni-80-wargames-tron-e-il-computer-al-cinema?amp>

<https://www.corriere.it/extra-per-voi/2016/02/27/i-wargames-cinema-che-ispirarono-reagan-resero-onnipotente-l-nsa-42d98750-dd79-11e5-8232-4d06db738c6b.shtml>

<https://consigli24.ilsole24ore.com/30-anni-fa-il-primi-sito-internet/>

<https://www.apogeonline.com/articoli/impariamo-a-conoscere-i-big-data-andrea-de-mauro/>

<https://www.themarketingfreaks.com/2019/11/big-data-cosa-sono-la-storia-le-caratteristiche-le-analisi-ed-esempi/>

<https://www.digitalic.it/tech-news/i-big-data-nel-600-avanti-cristo>

https://it.wikipedia.org/wiki/Big_data

https://www.ibm.com/it-it/analytics/data-lake?utm_content=SRCWW&p1=Search&p4=43700068096868228&p5=e&gclid=EAiaIQobChMIwtPJq5ri-AIV2RkGAB1dbAWFEAAAYASAAEgIjs_D_BwE&gclsrc=aw.ds

<https://www.ibm.com/it-it/topics/hdfs>

<https://www.tableau.com/it-it/learn/value-of-tableau/measuring-bi-value>

<https://www.edureka.co/blog/big-data-analytics/>

<https://hbr.org/2019/03/digital-transformation-is-not-about-technology>

<https://towardsdatascience.com/how-is-the-current-state-of-big-data-analytics-in-controlling-1273c725ac6a>

<https://www.tableau.com/it-it/learn/articles/data-driven-decision-making>

<https://www.howtosocial.it/news/post-suggeriti-di-instagram-ecco-cosa-cambia/>

<https://www.developersmaggioli.it/blog/machine-learning-la-scienza-delle-decisioni-automatiche/>

<https://thestreamable.com/news/netflix-disney-plus-show-industry-best-churn-rates-apple-tv-plus-still-bleeding-subscribers>

<https://devopedia.org/a-b-testing>

<https://medium.com/@narengowda/netflix-system-design-dbec30fede8d>

<https://www.bigdata4innovation.it/esperti-e-analisti/big-data-e-machine-learning-per-combattere-le-frodi/>

BIBLIOGRAFIA

- Anderson J.J. (November 1984). "Dave tells Ahl-the history of creating computing", *Creative Computing*, Volume 10, pp. 67
- Bozzo M. (1996), *La grande storia dei computer*, Bari, Edizioni Dedalo.
- Gomez-Uribe C.A., Hunt N. (2016), "The Netflix Recommender System: Algorithms, Business Value, and Innovation", *ACM Transactions on Management Information Systems*, Volume 6, pp. 1-19
- Goodman S.E. (1997), "Defining a Decade: Envisioning CSTB's Second 10 Years", *National Research Council*, pp. 18-21
- Holmlund M., Van Vaerenbergh Y., Ciuchita R., Ravald A., Sarantopoulos P., Villarroel Ordenes F., Zaki M. (2020), "Customer experience management in the age of big data analytics: A strategic framework", *Journal of Business Research*, Volume 116, pp. 356-365
- Kerney A.T. (2021), *IBM Maximo: Transform your business with intelligent enterprise asset management*, IBM Corporation
- Kimball R. (2004), *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*, Indianapolis, Wiley.
- Kotler P., Keller K., Ancarani F., Constabile M. (2017), *Marketing Management*, Milano, Pearson
- Manes S., Andrews P. (1994), *Gates*, New York, Simon and Schuster.
- Manyka J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Hung Byers A. (2011), *Big Data: The next frontier of innovation, competition, and productivity*, McKinsey Global Institute.
- OECD (2015), *Data Driven Innovation: Big Data for Growth and Well-Being*, Parigi, OECD Publishing.
- Preneel B. (2000), "Colossus and the German Lorenz Cipher", *Advances in Cryptology - EUROCRYPT 2000: International Conference on the Theory and Application of Cryptographic Techniques Bruges, Belgium May 14-18, 2000, Proceedings*, Lecture Notes in Computer Science, Springer
- Sharma A., Pandey H., Nayyar A., Kumar A. (2020). "Big Data and Analytics in Industry 4.0", *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*, Cham, Springer.
- Sigh S. (1999), *Codici & segreti*, Milano, Rizzoli Editore.
- Snijders, C., Matzat, U., Reips, U.D. (2012), "Big Data: Big gaps of knowledge in the field of Internet", *International Journal of Internet Science*, volume 7, Issue 1, pp. 1-5
- Tatnall A., Davey B. (2014), *Reflections on the History of Computers in Education*, Berlino, Springer.
- Van Meteren R., van Someren M. (2000), *Using Content-Based Filtering for Recommendation*
- Zane M. (2008), *Storia e memoria del personal computer. Il caso italiano*, Brescia, Jaca Book.