Jacopo Caratti

Big Data, Big Brother, Big Fear

Quando i dati svelano tutto di noi

Lavoro di maturità - Liceo Cantonale Bellinzona

Anno scolastico 2017 - 2018

Docenti responsabili: Marco Pellegrini e Adriano Martignoni

La realizzazione della copertina è ad opera di Enea Solari, che ha ideato l'illustrazione interpretando il tema del mio lavoro.
A detta di Enea Solari i numeri 1 e 0 corrispondono al codice binario e, più genericamente, alla quantità di informazioni che riguardano il tutto.
Il rosso che li sovrasta parzialmente corrisponde al mio profilo riportato otto volte: quattro volte con lo sguardo interno (costruzione, valutazione, elaborazione dei dati indicati prima) e quattro con lo sguardo esterno (la diffusione dei dati).
Evidentemente è una schematizzazione del tutto; il mio profilo, oltre al richiamo dell'essere umano che opera in questo campo, ha anche il valore di "firma" del mio lavoro di maturità.
Le cifre, nelle loro differenti tonalità di grigio, a dipendenza del loro abbinamento possono avere più o meno rimarco pur restando elementi "freddi".
Il rosso è invece il colore dell'energia e della dimostrata capacità di avere creato il linguaggio binario con le sue fasi di elaborazione, valutazione e diffusione operata dall'uomo.

«Torture numbers, and they'll confess to anything.»
«Tortura i numeri e confesseranno qualsiasi cosa.»

Gregg Easterbrook
Scrittore e giornalista americano.

Un grandissimo grazie a Antonietta Mira, Gianni Cattaneo e Luigi Curini per la disponibilità.

INDICE

1. Introduzione

Già il titolo "Big Data, Big Brother, Big Fear" potrebbe risultare enigmatico. Tradotto in italiano, significherebbe "Grandi Dati, Grande Fratello, Grande Paura". Ho scelto di accostare questi tre termini inglesi poiché mi piaceva l'idea di creare una sorta di filone che collegasse queste parole a prima vista così distanti tra loro. L'ordine secondo il quale sono riportati è causale. Infatti, secondo l'impostazione del mio Lavoro di Maturità, il primo è causa del secondo; così come il secondo termine è – o potrebbe essere – causa di quello successivo. Inoltre queste tre parole sono in grado di riassumere molto bene le tematiche da me affrontate.

Precedentemente alla scelta del tema del mio lavoro, confrontandomi con alcune delle tematiche trattate nei LAM degli ultimi anni, mi ero posto come obiettivo quello di concentrare le mie ricerche in un ambito abbastanza tecnologico, così da approfondire la mia conoscenza nel ramo dell'informatica e della tecnologia. Infatti sono appassionato di informatica e, anche grazie alla miriade di informazioni di cui sono venuto a conoscenza effettuando le mie ricerche, ho potuto espandere i miei orizzonti e probabilmente intraprenderò una carriera professionale improntata su questa nuova dimensione digitale.

In cerca di una tematica relativa all'informatica che potesse in qualche modo incutere timori nelle persone, sono incappato in questa dei Big Data e della Data Science. Ne avevo già sentito parlare, in modo particolare delle continue informazioni che rilasciamo in rete senza neppure rendercene conto. Queste dinamiche erano sempre accompagnate da un alone negativo.

Ho quindi avuto modo di confrontarmi più direttamente con questo ambito, di cui avevo già sentito bisbigliare, partecipando ad attività mirate durante le giornate autogestite organizzate dal Liceo Cantonale di Bellinzona, così come durante la giornata annuale dedicata alla tecnologia. Ricordo di avere anche assistito ad una conferenza incentrata proprio esclusivamente sui Big Data, tenuta dalla professoressa Antonietta Mira, che ho anche poi avuto modo di intervistare.

Di fronte alla vastità di questo – per me – nuovo mondo mi sono ritrovato piccolo, molto piccolo; e mi sono meravigliato di quanto esso tutt'oggi sia parte integrante delle nostre vite, senza che noi ce ne accorgiamo.

Confrontandomi con coetanei e famigliari su dubbi e perplessità che mi turbinavano in testa relativi alla mia scelta, mi sono accorto di quanto essi risultassero ignoranti su questo tema. Scoprendo la grande ignoranza della società – che ho avuto modo di sondare con il mio questionario – dinnanzi a queste dinamiche, mi sono posto come obiettivo anche quello di sensibilizzare le persone rispetto a questa nuova dimensione dei Big Data, illustrando le paure che essa suscita ma anche gli immensi orizzonti che è capace di aprire.

Lavorando a questo Lavoro di Maturità “Big Data, Big Brother, Big Fear” mi sono ritrovato catapultato in una nuova dimensione di cui, a priori, conoscevo ben poco.

Ora mi ritrovo a consegnare un Lam, a mio modo di vedere, completo in quello che mi ero prefissato di mostrare. Costruito secondo una determinata logica, ritengo possa portare il lettore a conoscere il funzionamento di questa realtà, riconoscendola, seppur mascherata, nelle più personali e disparate azioni di vita quotidiana.

Mi auguro che la lettura sia di tuo gradimento. Se per eventuali dubbi o perplessità dovessi avere qualche domanda ti invito a scrivermi alla mail jacopo.caratti@bluewin.ch.

Buona lettura.

Jacopo Caratti

2. La società digitale: nascita e sviluppo

La società digitale è nata con l'avvento di internet e dei servizi che ne derivano. Non dobbiamo però dimenticare che, alle sue origini, la rete era estremamente diversa da quella che noi tutti conosciamo oggi, nel 2017, e non permetteva di svolgere le medesime azioni. Infatti, con il termine "società digitale" si indica la pressoché totalità della società odierna, che a partire dagli anni Novanta ha cominciato a servirsi di internet non più solamente come un mezzo di comunicazione, bensì come uno spazio di condivisione di materiali e conoscenza tra tutti gli individui di questa nuova società mondiale.

2.1. Dalle reti a internet

Le origini di internet risalgono alla guerra fredda. Negli anni '50, sotto la presidenza di Eisenhower, l'esercito degli Stati Uniti in conflitto con l'Unione Sovietica cercava in tutti i modi di imporsi per superiorità tecnologica. Il culmine della sfida fra le due superpotenze fu la conquista della Luna che vide gli americani superare i sovietici.

Ciò portò il Dipartimento della Difesa degli Stati Uniti a ideare e finanziare molteplici programmi di sviluppo e ricerca e, per rilanciare il progresso tecnologico in campo aerospaziale, nel 1958 venne fondato il progetto ARPA (Advanced Research Projects Agency), approvato dal Congresso. Uno dei tanti scopi della nuova agenzia era quello di ideare e realizzare un nuovo sistema di comunicazione, capace di funzionare anche sotto attacco nucleare. Venne perciò creato il progetto ARPANET, che entrò in fase esecutiva nel 1969.

Alla sua ideazione, ARPANET non differiva molto quanto a tecnologia rispetto alle tante reti già presenti allora. L'unica differenza risiedeva nel fatto che tale progetto serviva a tessere una rete condivisa tra le sedi dell'agenzia e dei vari gruppi di ricerca esterni. Per rendere ciò possibile furono messe in collegamento tra loro più reti, sfruttando una tecnologia d'avanguardia: la commutazione a pacchetto (in inglese *packet switching*). Questo nuovo strumento consentiva e consente tuttora di scomporre le informazioni che vengono condivise in piccoli frammenti, che sono poi ricomposti una volta giunti a destinazione.[1]

Perciò la rete ARPANET viene considerata il precursore del moderno internet, il cui funzionamento è basato sulla possibilità di dialogo tra tante piccole reti.[2]

Col passare degli anni, i nodi di collegamento di questa rivoluzionaria rete, ovvero

[1] CASTELLS Manuel, *Galassia Internet*, Oxford, Feltrinelli, 2001, pag 28.

[2] dal documentario *Internet Revolution 1*, 2011, Bbc.

i punti dai quali è possibile connettersi, aumentano. Lo fanno, però in modo molto lento se paragonati ai numeri odierni. In due anni dalla sua creazione, nel 1971, i punti di collegamento di ARPANET risultavano essere solamente 15; negli anni successivi il ritmo di nascita di questi punti è cresciuto in maniera vertiginosa. Infatti, dopo la chiusura di ARPANET (vedi sotto) e con l'avvento dell'internet che oggi utilizziamo, il numero di nodi di collegamento arriverà entro il 2020 a circa 50 miliardi tra computer e oggetti intelligenti connessi ad internet, con un incremento di circa 200 nuovi nodi al secondo.[3]

Nel 1975 ARPANET non è più un'esclusiva dell'ARPA, ma viene resa disponibile anche ad altri settori delle forze armate statunitensi. Questo non è altro che l'inizio della condivisione della rete tra i più svariati soggetti, perché negli anni seguenti, già nel 1984, l'utilizzo di questa rete viene concesso anche alla NSF (National Science Foundation).

Così, in quel periodo comincia a serpeggiare l'idea di liberare queste importanti reti dagli esclusivi ambienti militari. Ben presto diventano così affare di dominio pubblico.

ARPANET, considerato ormai un network superato, date le miriadi di altre scoperte tecnologiche di questi anni, viene definitivamente abbandonato nel 1990.

Il governo statunitense cede il controllo di internet alla NSF che, accorgendosi della totale mancanza di leggi e regole per quanto riguarda l'utilizzo di internet, prende ben presto la decisione di privatizzarlo, fermando parzialmente il processo di sviluppo frenetico che ne era in corso.

Già allora si era capito che un bene della portata di internet non poteva certo essere reso privato da nessuno. Infatti, già nel 1995 la NSF smantella il progetto rendendo finalmente internet pubblico, perciò disponibile a tutti.

Negli anni Novanta molte persone attratte da questa "nuova" tecnologia cominciano a costruire le proprie reti, spesso su base commerciale. Reti che, seguendo nuovamente le direttive iniziali del progetto ARPANET, vengono messe in collegamento tra loro.

Grazie ad ARPANET internet è strutturato con un'architettura decentralizzata, cioè non mediata da nessuno: essa è dunque libera ed ognuno è libero di farne ciò che vuole.

C'è però una cosa importante da tenere in conto: in quegli anni vigeva una legge che vietava l'utilizzo di internet per qualsiasi scambio di natura commerciale. Questo ovviamente risultava di grande impedimento per gli imprenditori, che vedevano in internet un grandissimo potenziale di guadagno, ed è proprio per questo che nel 1991 la legge viene cambiata.

[3] tratto dal sito https://gblogs.cisco.com/it/2013/08/01/quante-connessioni-a-internet-ci-sono-nel-mondo-quando-proprio-adesso/?doing_wp_cron=1502105312.2184159755706787109375 (19 luglio 2017, 16:18)

Internet apre le sue porte alle attività commerciali; da quel momento è possibile effettuare transazioni finanziarie anche a livello informatico.
Una delle prime aziende che decise di espandere i propri interessi commerciali su questa nuova piattaforma fu il celebre Pizza Hut[4]: correva l'anno 1994.[5]
Non bisogna però dimenticare che internet, come servizio, non è stato pensato e progettato per essere un'infrastruttura legata al commercio e alla messa in circolo di beni finanziari. Per questo, fin da subito, la sua architettura ha dovuto conformarsi con le nuove dinamiche finanziarie che cominciavano a prendere piede.[6] È esattamente da questo momento che la notorietà e l'uso di internet subiscono un'accelerazione e internet inizia ad essere utilizzato da sempre più persone e pian piano assume la forma che tutti noi conosciamo.[7]
Ogni anno gli utenti di internet crescono considerevolmente. Basti pensare che solo dal 2016 l'utenza di internet nel mondo è cresciuta del 10%. Attualmente (2017) poco più del 50% della popolazione mondiale ne usufruisce quasi quotidianamente, ovvero circa 3.77 miliardi di persone.[8]

Oggi, internet con tutti i servizi che ne derivano, sta diventando sempre più uno strumento sociale e al servizio dell'ente pubblico, cioè dello Stato. I governi se ne servono – o meglio, se ne dovrebbero servire – per rendere il mondo un posto migliore nel quale vivere. Per questo, e per il fatto che il mezzo internet è comunque il più utilizzato nella comunicazione odierna, il suo uso risulta molto conteso tra i più diversi gruppi delle società, spesso rivali.[9]

2.2. Internet: il servizio più democratico

Con la nascita di internet e con le prestazioni che offre a tutti i suoi utenti il mondo è cambiato. Infatti, nonostante nel mondo esistano anche stati che limitano l'utilizzo di internet praticando la censura per esercitare il totale controllo politico, questo servizio resta probabilmente la più democratica possibilità di esprimersi liberamente e di ottenere informazioni mai esistita nella storia dell'Uomo. Infatti, se internet rispecchiasse veramente le idee e le utopie di chi fra i primi lo ideò e lo usò (spirito californiano), sarebbe davvero lo strumento ideale per garantire la più ampia

[4] conosciuta catena di ristorazione statunitense.

[5] PALMER Kimberly, *News & World* Report, 2007.

[6] dal documentario *Internet Revolution 1*, 2011, Bbc.

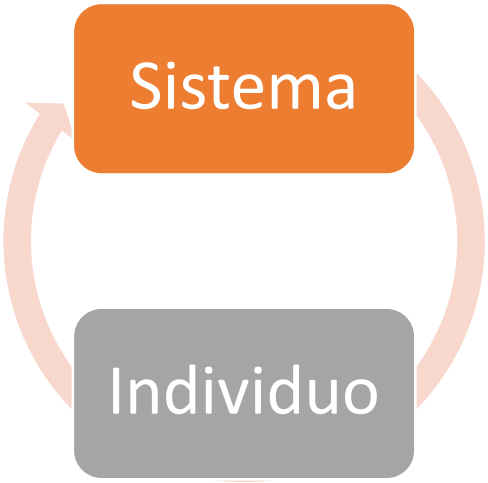
[7] cfr. CASTELLS, pag. 28.

[8] tratto dal sito https://wearesocial.com/it/blog/2017/01/digital-in-2017-in-italia-e-nel-mondo (11 dicembre 2017, 17:59).

[9] cfr. CASTELLS, pag. 134.

democrazia.[10] Nella rete si può infatti fare di tutto: richiedere informazioni, dire la propria opinione sugli argomenti che più ci aggradano o più ci ripugnano; contattare istantaneamente mezzo mondo, penetrare in archivi, biblioteche, giornali… .
Teoricamente il potere con internet è atomizzato e parte dalla base: le vecchie gerarchie traballano e le comunità nascono orizzontali e non più verticali, ma è proprio a questa nuova impostazione della società che il filosofo francese Gilles Deleuze dà il nome di "società del controllo".[11]

Come afferma nel suo libro Manuel Castells, sociologo di origine spagnola e professore in Comunicazione alla University of Southern California (USC): "Internet fornisce il materiale di base necessario alla produzione di una nuova società. Così facendo, i movimenti trasformano anche internet […]."[12]. Quindi, nella società che respira con internet, la grande messe di informazioni, connaturata al sistema, ha effetti diretti sull'individuo. L'individuo influenza poi il sistema che, a sua volta, ri-influenza l'individuo. Ciò fa capire che siamo attori di un sistema più grande, una sorta di enorme cane che si morde la coda (Schema 1).



Schema 1 – L'individuo alimenta la raccolta di informazioni del sistema. A sua volta, ciò che viene raccolto avrà influenze sull'individuo.

Per comprendere al meglio questo schema e la correlata frase di Castells, dobbiamo tenere bene a mente che nella nostra epoca, la cosiddetta "Età dell'informazione" [13], una buona parte della vita quotidiana – dal tempo libero al lavoro, alle relazioni personali - ha luogo in rete.

[10] cfr. CASTELLS, pag. 149.

[11] BAUMAN Zygmunt, LYON David, *Sesto potere: La sorveglianza nella modernità liquida*, Cambridge, Laterza, 2013, pag. 9.

[12] *ibidem*, pag. 139.

[13] cfr. CASTELLS, pag 139.

2.3. L'*Internet of Things* e la fame di dati

Gli apparecchi che oggi sono connessi alla rete sono sempre più numerosi e si suddividono nelle più diverse categorie. Si parte dall'orologio *smart*, passando dall'assistente virtuale, per poi arrivare al frigorifero intelligente, capace di riconoscere se un dato tipo di alimento stia per finire e, in base alle nostre preferenze di consumo, per poi essere in grado di ordinarlo.
In tutti gli ambiti della quotidianità, siamo sempre più connessi ad internet. Bisogna ammettere che la nuova realtà è molto comoda. Rende tutto più semplice e automatico. Si ha l'impressione di poter vivere la propria vita senza le piccole preoccupazioni che puntualmente ci presentava solamente qualche anno fa. Come cucinare, fare la spesa, scrivere testi e molti altri compiti che ormai non ci rendiamo nemmeno più conto siano basilari nelle nostre vite. Però dico appunto "si ha l'impressione", perché infatti di preoccupazioni se ne sono create altre. Di ben altro genere. Esempio: lo stress di avere un dispositivo intelligente ma di doverlo costantemente ricaricare, per non far sì che ci abbandoni quei pochi attimi prima di poter trovare una presa elettrica a cui poterlo collegare. Questo per evitare problemi che – essendoci ormai abituati alla comodità dei moderni ritrovati della tecnologia – non siamo più abituati ad avere e, perciò, ci sembrano estranei e improponibili.
Spesso si rincorre l'ultimo modello uscito, che offre funzioni innovative rispetto al precedente ed è quindi in grado di "semplificare" ulteriormente la nostra esistenza quali esseri umani. Accrescendo in tal modo la nostra dipendenza dalla tecnologia. Inoltre bisogna fare attenzione ad aggiornare tali dispositivi non appena è disponibile una nuova versione del sistema operativo, per non incorrere in *bug* (ovvero in errori nella costruzione del programma) che possono creare problemi di utilizzo ai consumatori, o, peggio, falle nella sicurezza.

Con questo nuovo termine: Internet delle Cose (dall'inglese *Internet of Things*), si vuole anche sensibilizzare l'utenza sulla quantità dei dati che questi oggetti intelligenti raccolgono su di noi per permetterci un'esistenza, all'apparenza, meno faticosa. Approfittando delle miriadi di comodità che tali apparecchi regalano, spesso pur volendolo evitare, rilasciamo bene o male un'infinità di informazioni sulla nostra sfera più privata. La cosa preoccupante è che questi dati vengono sì registrati dagli oggetti che utilizziamo, ma vi passano attraverso, come se questi oggetti *smart* fossero un canale. Queste informazioni, riguardanti la nostra persona e i nostri gusti, entrano quindi legalmente in possesso delle *major* informatiche produttrici (questo punto verrà approfondito nel capitolo: 8. Leggi).

3. I Big Data

«Quando milioni di persone diventano utenti, le loro discussioni si trasformano in dati. Interpretarli ed integrarli fra loro è la sfida che i Big Data pongono al mondo di oggi: dalla politica, all'economia, alla società.»

Voices from the Blogs (2012)
Spinn-off dell'Università degli Studi di Milano e prima azienda ad occuparsi di Big Data in Italia.

3.1. La preistoria: i *database*

Per parlare di Big Data bisogna prima comprendere il significato del termine inglese "database" (in italiano "banca dati" o "base di dati"), poiché sotto alcuni aspetti i Big Data possono essere considerati come un'evoluzione dei *database*. I loro "antenati".

La parola "database" viene utilizzata nel mondo informatico per descrivere dati suddivisi per caratteristiche comuni di contenuto e formato, raggruppati tra loro in un'entità, come una piccola parte di un calcolatore.[14]

Per esempio, aprendo un browser (Google Chrome, Safari, Internet Explorer, …) e digitando una parola a scelta, verrà stampata sullo schermo una lista di siti internet il cui contenuto è inerente alla parola digitata. Cliccando su uno dei siti elencati si accederà pure al suo database di informazioni, del quale possono fare parte foto, video, articoli o quant'altro, ma sempre inerenti alla ricerca effettuata dall'utente.

Queste basi di dati sono essenziali per dare agli utenti della rete la possibilità di accedere ai dati informatici ricercati anche per una modifica, oltre che per una semplice consultazione.

Essendo i dati di natura informatica in fortissima crescita negli ultimi anni, queste banche dati sono di primaria importanza, poiché sono l'unico sistema per poterli catalogare secondo un certo ordine e poterli rendere reperibili con pochi clic dai consumatori che navigano nella rete.

3.2. Cosa sono

I Big Data sono quindi i successori delle banche dati, ma, a differenza di questi "vecchi" sistemi di stoccaggio delle informazioni digitali, i Big Data sono una raccolta di dati su scala globale, tanto grande da non poter essere elaborata con convenzionali sistemi e strumenti di analisi.

[14] tratto dal sito http://it.ccm.net/contents/2-database-introduzione (20 luglio 2017, 13:21).

«Non esiste una definizione univoca dei Big Data […]» dice Antonietta Mira – professoressa di statistica all'Università della Svizzera Italiana (USI) e cofondatrice e ora direttrice dell'Istituto Interdisciplinare di Scienza dei Dati dell'USI (IDIDS) – in un articolo apparso sul Corriere del Ticino il 21 aprile 2017. La prof. Mira afferma che a suo modo di vedere «[…] questa loro denominazione è fuorviante.». Secondo l'esperta il termine "Big" indica qualcosa di "enorme", ma rifacendoci alla sua etimologia latina muta di significato: "e norma", cioè "fuori dalla norma". Secondo l'esperta quindi i Big Data non sarebbero altro che "dati fuori dall'ordinario".[15]
Alcuni esperti preferiscono definire questo tipo di dati *Complex Data* o *Smart Data*, poiché ritengono tale terminologia più appropriata. *Smart Data* non nel senso che questi dati siano in qualche modo intelligenti, ma che dopo il processo di analisi e lavorazione lo potranno diventare.

È importante precisare che, quando si parla di dati in questo nuovo ambito della tecnologia, non ci si riferisce alla tradizionale tipologia di dati, come numeri o caratteri, ma a tutto ciò che può essere condiviso e registrato. Questa nuova concezione eterogenea di dati tratta di testi, immagini, suoni, video, coordinate GPS, dati relazionali, metadati[16]. Ce n'è una quantità davvero impressionante.[17]
Qui di seguito è riportata un'immagine nella quale è indicata la quantità di dati che ogni singolo minuto di ogni giorno gli utenti pubblicano sul web (Figura 1).

[15] «Corriere del Ticino», Un universo di informazioni che può svelare tutto di noi, 21 aprile 2017.

[16] Sono informazioni che descrivono un insieme di dati. Vengono utilizzati per riassumere le nozioni di base che possono facilitare il monitoraggio, la ricerca e il lavoro con dati specifici. Un esempio tipico di metadati è costituito dalla scheda del catalogo di una biblioteca, la quale contiene informazioni circa il contenuto e la posizione di un libro, cioè dati riguardanti più dati che si riferiscono al libro. Tratto dal sito https://it.wikipedia.org/wiki/Metadato (8 novembre, 18:56).

[17] dal materiale video degli archivi della Fondazione Milano: *Cosa sono i Big Data?*, *Il progetto Urbanscope*, *Nuovi paradigmi per ricerca e business*, *Pericoli dei Big Data e impatto economico e sociale*.

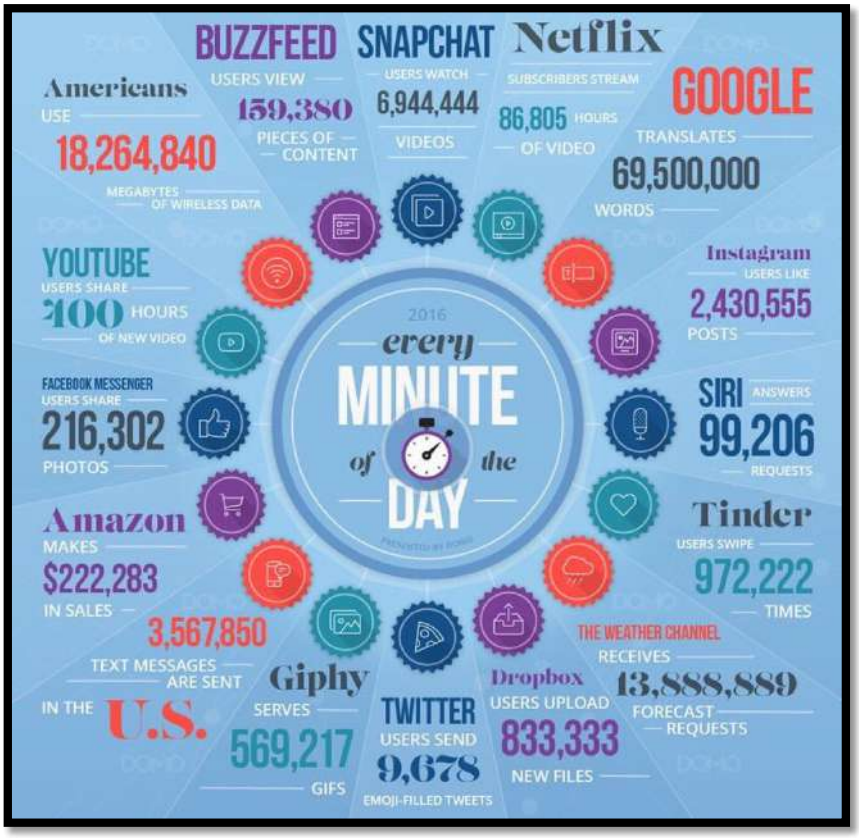


Figura 1 – La mole di dati che ogni minuto viene caricata dagli utenti sul web.

Proprio per questo motivo si necessita di appositi sistemi ed esperti, che si occupano di raccoglierli e analizzarli. Tutto questo fa parte di una nuova scienza che porta il nome *Data Science* (in italiano "scienza dei dati").

Parlando in termini di Big Data, ciò che può essere *big* per una persona non deve necessariamente esserlo per un'altra, che magari nella sua ricerca necessita la raccolta di molti più dati. O ancora, ciò che può essere definito *big* in questo esatto momento, con molte probabilità, non lo potrà più essere nell'arco di pochi anni grazie al continuo miglioramento della memoria dei grandi calcolatori e della loro capacità di calcolo. Il termine assume quindi una prospettiva soggettiva.[18]

A questo proposito, così osserva Castells nel suo libro "Galassia Internet": "Nell'attuale ambiente tecnologico, ogni informazione trasmessa elettronicamente viene registrata e può essere eventualmente processata, identificata e combinata, in unità di analisi individuali e collettive."[19].

Figura 1: tratta dal sito https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

[18] MIRA Antonietta, *La Scienza dei Dati: una Nuova Sfida Multidisciplinare*, Rendiconti della Classe di Scienze Morali: Scienze Economiche e Statistiche, SECS-S/01 – Statistica.

[19] cfr. CASTELLS, pag. 164.

3.3. Da dove provengono

Una domanda sorge certamente spontanea: "Come mai tutto a un tratto, negli ultimi anni, i dati creati sono nettamente più numerosi rispetto a tutti quelli prodotti sin dall'inizio della storia dell'Uomo?". I Big Data nascono al momento in cui l'essere umano comincia ad utilizzare la tecnologia digitale e la scienza che ne deriva impara ad analizzare le tracce digitali da noi create, per estrarre conoscenza.

Solamente nel 2015 l'intero pianeta era in grado di generare una cifra come 500 Terabytes al giorno, ovvero un numero di Bytes a 14 zeri![20]

I grandi numeri sono sempre qualcosa di difficile da capire. Se, per esempio, scrivo 1 miliardo o 1 milione, per la nostra mente cambia poco. Facciamo caso solo al fatto che si tratta di grandi numeri. Nella Figura 2 vi fornisco un chiaro esempio dei numeri di cui sto parlando: se un Byte corrispondesse a una banconota, tutte quelle costituenti 1 miliardo rappresenterebbero 1 Terabyte, mentre la metà di quelle che rappresentano 1 bilione illustrerebbero 500 Terabytes.

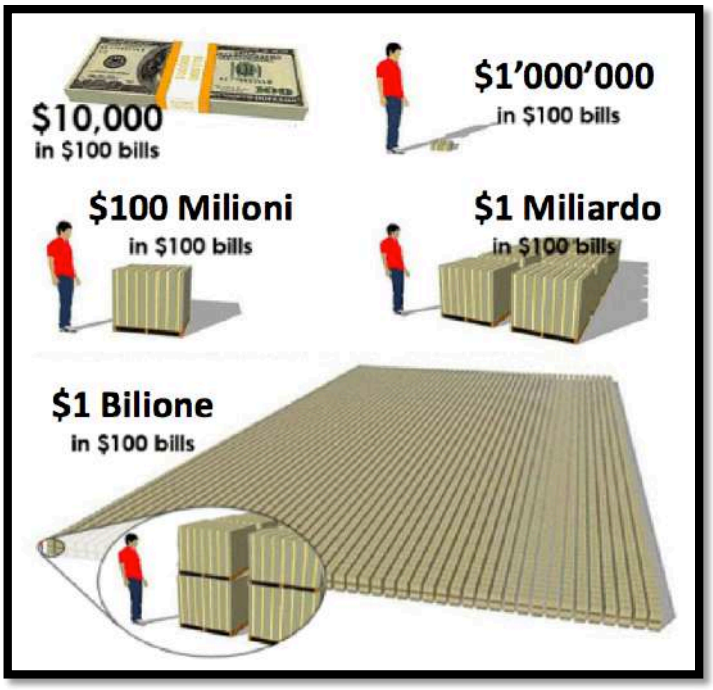


Figura 2 – Paragone fra 100 banconote da 100 dollari (= 10'000 dollari = 10'000 Bytes) e 1 bilione di dollari (= 1'000'000'000'000 bit). L'immagine rende l'idea di quanto sia 1 Byte rispetto a 500 Terabytes.

Figura 2: tratta dal sito https://medium.com/@25stories/julian-s-dab-daily-audio-blog-session-37-million-vs-billion-vs-trillion-1ff8a17980bc e leggermente modificata.

[20] tratto dal sito http://www.rsi.ch/news/oltre-la-news/Prigionieri-dei-Big-Data-8434276.html (13 settembre, 15:26).

…e questa quantità spropositata di dati veniva prodotta nel 2015 ogni singolo giorno!

Le mie letture sull'argomento mi hanno portato a individuare tre principali cause della repentina ed inarrestabile crescita dei dati, che voi lettori potrete facilmente comprendere osservando la "Figura 3". Figura che ho costruito accostando i diversi scenari presenti dal 1994 al 2015.

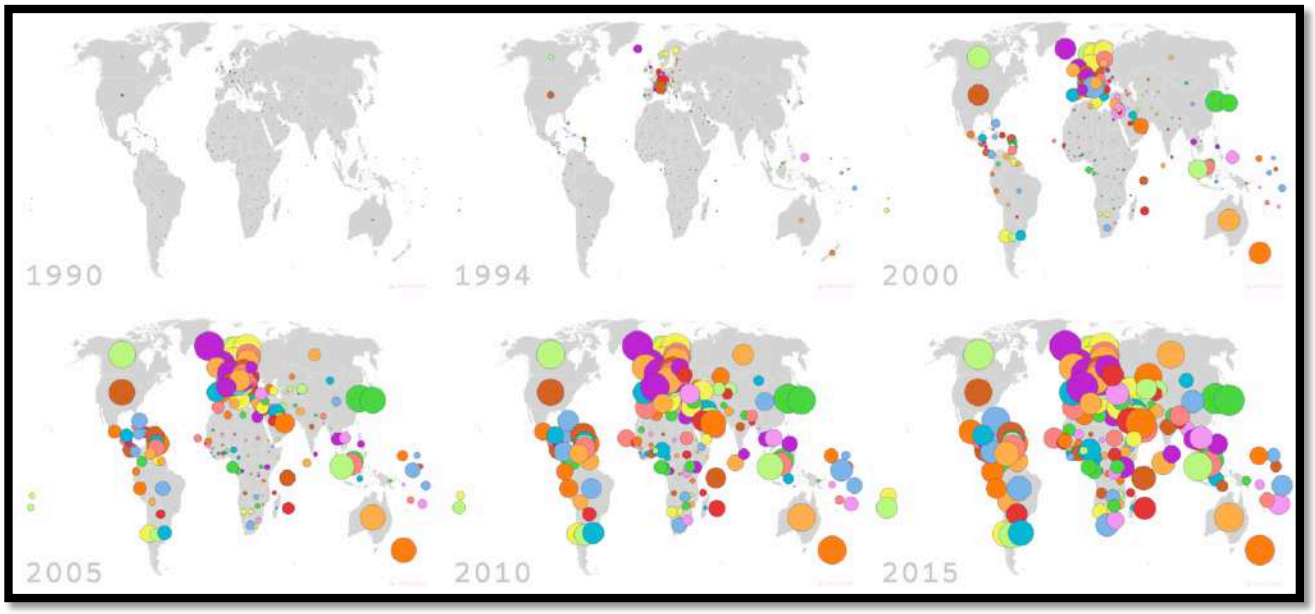


Figura 3 – Immagine che mostra lo sviluppo degli accessi ad internet ogni 100 persone in ogni Paese dal 1990 al 2015 (il 1994 è l'anno della prima compravendita avvenuta sul web).

La prima causa di tale aumento è chiaramente il grande **sviluppo tecnologico** occorso in questi ultimi decenni, partendo proprio dall'invenzione di internet. Le ultime invenzioni e scoperte sul piano della tecnologia hanno radicalmente cambiato il nostro modo di vivere e di relazionarci con gli altri, spingendo milioni di persone a forzarsi ad avere gli ultimi ritrovati in questo campo. Tale punto è a sua volta la 'causa madre' degli altri due fattori che influenzano fortemente la presenza tanto massiccia di dati.

Il secondo fattore scatenante tale boom di informazioni è infatti sicuramente la **diffusione del digitale**. Tutto ciò che riguarda internet è oggi diffuso globalmente, così come i servizi che la tecnologia – e più in generale il digitale – offrono alle persone. Praticamente in ogni dove e in tutte le società umane sono presenti, seppur in minima quantità, influenze di tale digitalizzazione. Mi

Figura 3: ho composto l'immagine accostando sei differenti foto scattate dal sito https://www.gapminder.org/tools/#_state_marker_size_which=internet/_users/_per/_100/_people&domainMin:null&domainMax:null;&color_which=country;;;&ui_presentation:true;&chart-type=map&locale_id=en

riferisco specialmente ad apparecchi tecnologici volti a ottimizzare ogni sorta di azione, che vi ho già esposto nel capitoletto "2.3. *L'Internet of Things*". E veniamo al terzo fattore: sono proprio gli oggetti *smart*, che citavo sopra, e con i quali la gente è sempre più in **costante collegamento**, che portano alla produzione di informazioni sui più celati angoli della nostra sfera intima. Questa è quindi il terzo fattore scatenante da me individuato quale origine dell'esponenziale crescita dei dati a disposizione. La gente è ormai (col)legata alla rete praticamente ovunque.

3.4. Come i dati diventano conoscenza

Grandissime quantità di dati vengono immagazzinate nella loro completa eterogeneità, per poi essere spostate in calcolatori con potenza e capacità enormi. Si parla infatti in termini di Terabytes, ovvero dai 1024 Gigabytes in crescendo. Per rendervi più chiare le cifre sappiate che la dimensione del formato elettronico della "Divina Commedia" di Dante Alighieri, opera mondialmente riconosciuta, pesa poco meno di 0,001 Gigabytes (882 Kilobytes, ovvero 0.00084 Gigabytes).[22]
Dei dati informatici per essere considerati Big Data devono raggiungere una certa dimensione. Esiste infatti una classificazione anche in questo ambito, che va dalla classe *small* alla *medium*, fino ad arrivare a quella di *big*. Come potete vedere dalla "
Tabella 1", se i dati raccolti superano il TB di dimensione, risulta corretto parlare di Big Data.

Classificazione	Dimensione	Gestiti con[23]	Salvati su	Esempi
small	< 10 GB	Excel	Una memoria di una macchina	Migliaia di dati di vendita
medium	10 GB – 1 TB	Monolithic db	La memoria intera di una macchina	Milioni di pagine web
big	> 1 TB	Hadoop	La memoria di più macchine	Miliardi di clic

Tabella 1 – La tabella raffigura la classificazione dei dati informatici in relazione alla loro dimensione. Presenta inoltre alcune curiosità collegate al tipo di dati informatici trattati.[24]

[22] tratto dal sito https://www.amazon.it/Divina-Commedia-Dante-Alighieri-ebook/dp/B003MAK5FK (20 luglio 2017, 11:45).

[23] Software di analisi e gestione dei dati.

[24] tratto dal sito https://www.quora.com/How-much-data-is-Big-Data (22 luglio 2017, 20:34).

Dopo essere stati immagazzinati in più calcolatori, i dati vengono passati al setaccio e analizzati alla ricerca di correlazioni e informazioni utili sulle quali disegnare una sorta di filo che li possa legare per farne qualcosa di sensato. Questo perché i dati raccolti possono trattare milioni di tipologie di tracce digitali – notizie reali, pettegolezzi, fantasie, dati personali,… – generando per finire un confuso brodo di informazioni che va riordinato. Un individuo non potrebbe svolgere un simile compito perché richiederebbe migliaia di anni di lavoro e dispendi di energia enormi. Per questo motivo i *computational scientist*[25], gruppi di analisti che operano in questo campo, ricorrono all'aiuto di algoritmi specializzati e modelli statistici.
Nella definizione della Treccani un algoritmo nel mondo dell'informatica è: "[…] una sequenza finita di operazioni elementari, eseguibili facilmente da un elaboratore che, a partire da un insieme di dati, produce un altro insieme di dati che soddisfano un preassegnato insieme di requisiti.".
Spesso viene impiegato anche quello che in gergo è conosciuto come *machine learning*, algoritmi capaci di imparare dai dati che stanno lavorando, così da migliorarsi nel tempo in quello per cui sono stati concepiti. L'intelligenza artificiale, così come la si intende oggi, si rispecchia esattamente in questa tipologia di algoritmi. "Non significa quindi sviluppare un algoritmo intelligente *dotato di* autocoscienza. […] questi algoritmi sono completamente stupidi e non sanno perché stanno facendo quella cosa. Ignorano il processo causale che sta dietro.", così risponde Luigi Curini nell'intervista allegata, riferendosi al *machine learning*. Curini è professore di Scienze Politiche all'Università di Milano (UNIMI) e cofondatore di un istituto che si occupa di analizzare il flusso di informazioni derivanti dai social network.
La pecca di questi algoritmi – specialmente di quelli legati alla previsione di accadimenti futuri – è che, una volta resi funzionali, non possono lavorare indisturbati ma devono essere puntualmente ricalibrati. Così da non incorrere in errori di interpretazione dei dati.

Conclusa la parte di analisi, si ottengono risultati che devono però essere visualizzati nel modo corretto per essere compresi senza incorrere in errori statistici (ne illustro qualcuno nel capitolo "3.6. Quando i dati mentono"). Qui entrano in gioco i *computational scientist* esperti in grafica, che si occupano di rendere visibili i risultati ottenuti precedentemente tramite programmi di visualizzazione e elaborazione visiva delle informazioni. È solo da questo momento che i dati possono essere considerati informazioni utili: inizialmente non erano altro che un ammasso di informazioni senza alcun legame, né significato (Figura 4).

[25] «Corriere del Ticino», Un universo di informazioni che può svelare tutto di noi, 21 aprile 2017.

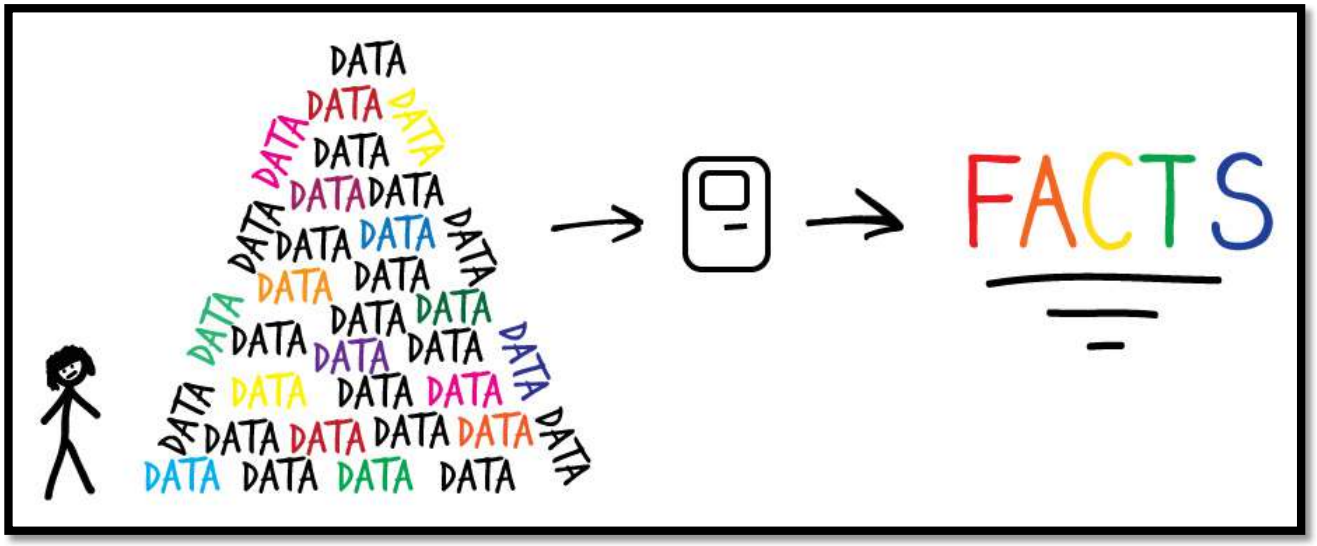


Figura 4 – I dati vengono analizzati dai calcolatori e restituiti. Solo a questo punto diventano informazioni e quindi conoscenza.

I dati, una volta raccolti, possono essere riutilizzati più e più volte senza che essi perdano affidabilità: un dato soggetto ad analisi resterà pur sempre il medesimo.

Parte di questi dati sono poi distribuiti gratuitamente dalle compagnie che li hanno resi "conoscenza", al fine di rendere disponibili interessanti informazioni a chi ne ha più necessità per i propri scopi. Altri, invece, distribuiscono quanto ricavato a pagamento. In questo modo il potere informativo è concentrato in gruppi ristretti di persone, che ne limitano l'accesso ad un certo numero di individui. Alessandro Mantelero, professore di diritto e di diritto di internet al politecnico di Torino, afferma che "[Queste meccaniche portano a] concentrare il potere informativo in capo a gruppi di soggetti via via più circoscritti, sino a quelli che potremmo definire i «signori dei dati»."[26].

Nella storia abbiamo assistito prima all'imporsi di dotti eclettici, persone che possedevano una grande cultura generale, che conferiva loro grande importanza. Persone come Galileo Galilei o Leonardo Da Vinci. Poi, pian piano, all'inizio del '900 è avvenuta la specializzazione dei diversi campi della conoscenza. Da questo momento si sono create delle barriere tra le diverse discipline e i diversi ambiti del sapere.

Figura 4: tratta dal sito http://jiffyclub.github.io/digital-demography-2014/#/ (15 agosto 2017, 18:51).

[26] «Il diritto dell'informazione e dell'informatica», *Big Data: i rischi della concentrazione del potere informativo digitale e gli strumenti di controllo*, Anno XXVIII, Fascicolo 1 - 2012.

Tutto questo rende oggi necessaria la collaborazione di questi molteplici esperti, al fine di poter analizzare nel modo corretto, nonché più accurato possibile, i Big Data a loro disposizione. Ognuno svolge la sua parte e si impegna nel lavoro del quale è più istruito.
Con ciò l'accuratezza delle informazioni estratte mediante l'*analysis* è in costante crescita, regalandoci un pozzo di conoscenza sempre più grande.

La statistica e le scienze computazionali sono nati come due ambiti a sé stanti, che si sono sviluppati seguendo sentieri ben distinti. Queste due discipline, sotto l'aspetto della Scienza dei Dati, si stanno invece accorpando in una sola, cosicché possano trarre benefici l'una dall'altra, senza spreco di risorse. Difatti, di fronte a un problema, entrambe perseguono il medesimo obiettivo: riuscire a trovare una soluzione nel minor tempo possibile, col minore impiego di risorse. Unendo le forze si arriva quindi più velocemente ad una soluzione e spesso e volentieri con minore impiego di capitali. Proprio per questo motivo è importante formare i *computational scientist* alla nozione di rischio statistico e istruire gli statisti che lavorano con grandi quantità di dati alle pratiche computazionali.[27]

Un problema che presenta la "Data Science" è che, se la potenza di calcolo diventa di giorno in giorno più grande grazie alla tecnologia che avanza, la memoria di questi calcolatori non cresce così velocemente. In questo modo si crea una sorta di collo di bottiglia che frena lo sviluppo dell'analisi dei Big Data.[28]

3.5. Metodi di classificazione

Con la nascita dello studio dei Big Data si è improvvisamente avuto bisogno di nuove tecniche di analisi e di elaborazione; questo perché i sistemi precedenti venivano spinti ai limiti dalla quantità di dati raccolti. I Big Data arrivano da più fonti ad alta velocità, volume e varietà. Seguendo questa traccia, il noto analista americano Douglas Laney ha coniato nel 2001 le *V3s*, ovvero tre proprietà (*Volume*, *Variety*, *Velocity*), che per una decina d'anni sono state sufficienti a definire e suddividere i Big Data nelle loro caratteristiche. La classificazione *V3s* aiutava a caratterizzarli e differenziarli nella loro moltitudine ed è ancora oggi un perno nell'analisi dei Big Data.

- *Volume* (quantità)

[27] cfr. MIRA.

[28] «Corriere del Ticino», *Un universo di informazioni che può svelare tutto di noi*, 21 aprile 2017.

Si intende la grande quantità di dati contenuti e derivanti da un processo.

- *Variety* (varietà)
  Ci si riferisce alla loro diversità: "I dati di cui siamo a disposizione riguardano solamente un certo settore oppure sono una miscela di più ambiti?". Differiscono per tipologia, provenienza e struttura, che può esserci interamente o parzialmente o non esserci.
- *Velocity* (velocità)
  Riguarda la velocità con cui nuovi dati sono prodotti, nonché la velocità di lavorazione e analisi di questi enormi gruppi di dati.

Fin da subito ci si è accorti dell'ammontare del valore inestimabile di questo patrimonio di dati., troppo attraente per essere lasciato nell'ombra. Per questo motivo è nato il bisogno di passare dalle *V3s* alle *V4s*.

- *Value* (valore)
  Il valore che possiedono questi Big Data è elevatissimo. Basti pensare che Walmart, multinazionale statunitense attiva nel settore della vendita al dettaglio, impiegando l'analisi dei Big Data è stata in grado di incrementare del 10% - 15% le proprie vendite online. Ciò ha portato ad un incremento dei ricavi pari ad un miliardo di dollari.[29]. Questa è forse la V di cui sentiremo parlare più spesso nei prossimi anni.

Questo argomento viene affrontato più da vicino nel punto "3.5.1. Valore: ciò che li rende così appetibili".

Col passare del tempo, però, queste tre proprietà da sole non sono più bastate per descrivere e caratterizzare tutta questa enorme miriade di dati sempre in espansione; da qui la necessità per gli esperti del settore di aggiungerne una quarta, una quinta, una sesta e addirittura una settima. Le quattro più recenti sono considerate dai più come le caratteristiche secondarie nella qualificazione dei Big Data (ciò non esclude che in futuro, anche mentre qualcuno sta leggendo questo Lavoro di maturità, le "V" da sette siano aumentate).

- *Veracity* (attendibilità)

[29] «Wired», *Putting a Dollar Value on Big Data Insights*, 2017.

È una caratteristica legata alla veridicità dell'informazione estrapolata. Infatti con l'analisi di questi immensi e sempre più disparati gruppi di dati si può incorrere in errore e giungere a conclusioni sbagliate. In gergo queste correlazioni errate vengono chiamate "correlazioni spurie".

- *Variability* (variabilità)
  I dati di cui si è in possesso possono risultare incoerenti.
- *Visualization* (visualizzazione)
  Dopo che i dati sono stati analizzati si deve trovare un modo per renderli leggibili, presentabili e comprensibili. Ciò non è semplice perché ciò che è risultato dall'elaborazione dei dati può contenere decine di variabili diverse, ognuna delle quali deve essere tenuta in conto al fine di non incorrere in errori di analisi.

3.5.1. Valore: ciò che li rende così appetibili

Il valore dei Big Data è riconducibile proprio al capitale immenso di informazioni celate in essi. Coloro che detengono questa incalcolabile quantità di dati sono un passo avanti a tutti. Come vedremo con dei mirati esempi nel prossimo capitolo "
4. Big Data e opportunità: la statistica e le scienze", sono in grado addirittura di prevedere avvenimenti futuri sulla base di dati raccolti in precedenza. Questi colossi dell'informazione sono quindi una sorta di "stregoni del ventunesimo secolo", in grado di prevedere il futuro basandosi sui dati forniti dal nostro comportamento. Non per niente il settimanale The Economist, che tratta di informazioni provenienti da tutto il globo, li ha definiti "La risorsa più preziosa al mondo" e i dati vengono definiti il nuovo petrolio dell'economia.[30]

Ma i miliardi di fatturato derivanti dalla *Data Science* non provengono esclusivamente dal fattore "previsione", bensì anche da quello della "conoscenza". Se volessimo apprendere qualcosa di nuovo dovremmo iscriverci a un corso, prendere parte a seminari, leggere libri,… . Ma come tutti voi ben sapere, ciò comporta un investimento finanziario. Questo per dirvi che la conoscenza ha un prezzo, in molti casi anche alto, e, come presentatovi nel capitolo "3.4. Come i dati diventano conoscenza", l'obiettivo primario della *Data Science* è quello di estrarre conoscenza dal grande ammasso informe di dati disponibili. Potendo registrare suppergiù legalmente[31] la marea di dati fornita da ogni utente, potete ben immaginare che per i *Lord digitali*[32] non approfittare di ciò

[30] «Altroconsumo», *Che potere, i dati*, novembre 2017.

[31] Affronto l'argomento giuridico nel capitolo "8. Leggi" e riporto l'intervista con Gianni Cattaneo negli allegati.

[32] termine introdotto da CASATI Roberto, *Contro il colonialismo digitale*, Laterza, 2013, pag. 119.

sarebbe sciocco, se si pensa agli ingenti introiti che in particolare questo tipo di business può portare.

Una riprova del giro di affari che attornia la Data Science è quello dei premi per le innovazioni e le scoperte. Per esempio, l'azienda di distribuzione televisiva internazionale Netflix nel 2006 aveva offerto un milione di dollari a chiunque fosse stato in grado di migliorare l'algoritmo alla base delle raccomandazioni di nuovi film agli utenti secondo i loro gusti.[33]

3.6. Quando i dati mentono[34]

Così riporta la rivista Focus del maggio 2016, riferendosi alle numerose scoperte scientifiche e pseudoscientifiche occorse negli ultimi anni: "In parte è colpa del progresso tecnologico, che ci ha dato computer superveloci capaci di trovare nelle banche dati ogni tipo di rapporto causa-effetto… compresi quelli che non esistono.". Infatti, sulla base di grandi quantità di dati, abbastanza vasti da poter già essere definiti Big Data, moltissimi ricercatori hanno portato a termine studi nei più disparati ambiti di ricerca. Essi sono giunti a conclusioni comprovate dai dati analizzati e, perciò, ritenute corrette. Senonché, qualche tempo dopo, hanno scoperto che altri ricercatori erano giunti a conclusioni discordanti, utilizzando banche dati e metodologie empiriche differenti.

Quella che ho appena descritto non è finzione, ma pura realtà: da una decina d'anni, nell'ambiente scientifico, si cerca di capire quanto questi sempre più frequenti studi, basati sulla raccolta e analisi di una certa tipologia di dati, possano effettivamente essere attendibili. Oltre la metà dei casi sottoposti a verifica hanno infatti portato a conclusioni sperimentali errate.

Nel 2005, John Ioannidis, docente a Standford e esperto mondiale sulla credibilità delle ricerche mediche, ha riproposto le 49 scoperte in campo medico più importanti dei precedenti 13 anni. A fine studio ha concluso che 14 scoperte erano arrivate a conclusioni errate o esagerate.

Tutto ciò può apparire ai nostri occhi come una novità, ma in verità non lo è. Lo si ha sempre fatto: si raccolgono un certo numero di informazioni e le si analizza, così da ottenere dei risultati.

Già sul finire del 1800, in seguito ad uno studio statistico effettuato nei Paesi Bassi, si era giunti alla conclusione che, se in città arrivavano più cicogne, nascevano anche più bambini. Da qui la ben nota leggenda. Di primo acchito qualcuno potrebbe concludere – come immagino sia stato fatto – che la presenza delle cicogne influenzi in qualche modo la nascita di bambini. La realtà è però ben diversa. Le case nelle quali era presente un neonato, tendevano ad essere riscaldate maggiormente

[33] Cfr. MIRA, 9.

[34] Per questo titolo ho preso spunto dal nome dell'articolo: «Focus», *Quando i dati mentono*, maggio 2016.

per non fare ammalare il bambino. Questo calore emanato dai camini delle case attirava poi le cicogne.

In questo caso il rapporto c'è, ma **indiretto**. Il ricercatore deve infatti tenere in considerazione tutti gli elementi capaci di influenzare in un qualunque modo l'esito dell'esperimento.

Un altro esempio di curiosa conclusione, che a prima vista si potrebbe trarre, la è propongo qui di seguito. Nel grafico di Figura 5 vengono messi in relazione gli annegamenti in piscina negli USA e i film interpretati dall'attore americano Nicolas Cage.

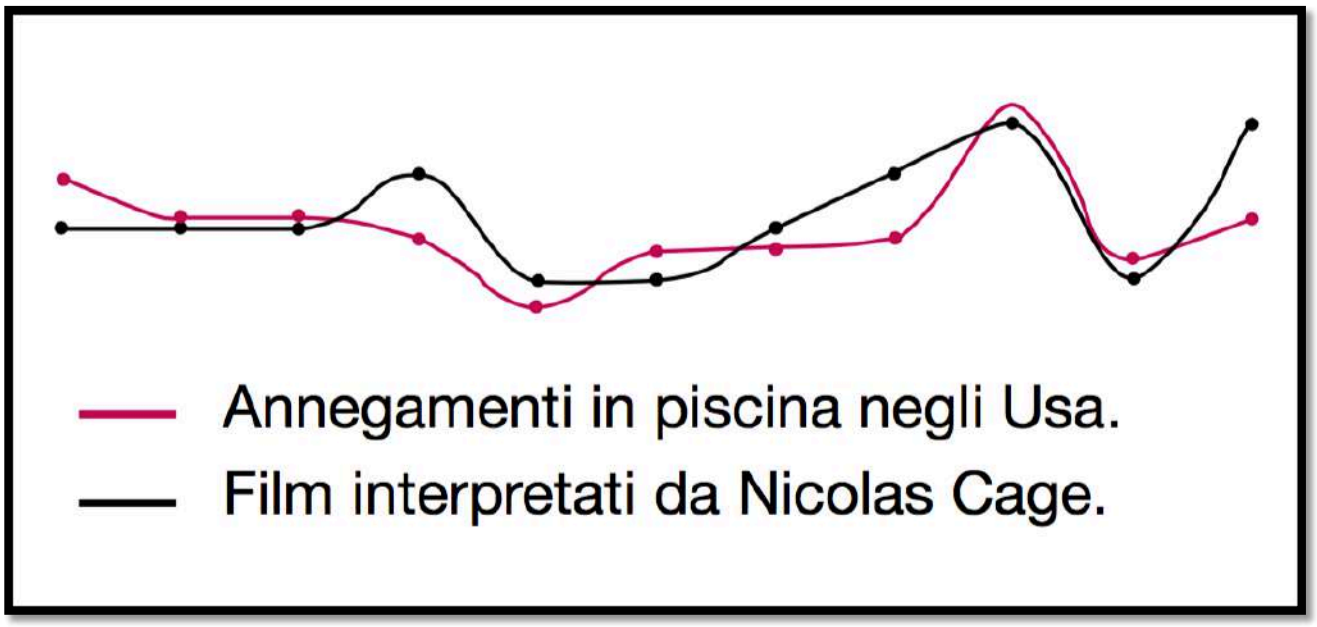


Figura 5 – Il numero di persone affogate in piscina negli USA, insensatamente correlato al numero di film interpretati da Nicolas Cage dal 1999 al 2009.

Questo caso è differente dal precedente, dove era riscontrabile un rapporto causa-effetto indiretto. Qui non c'è alcun tipo di legame, perché, nonostante gli andamenti dei grafici siano pressoché identici, essi sono totalmente **scollegati**. Questa tipologia di errore, nella quale si mettono in relazione due elementi completamente disgiunti, è chiamata "correlazione spuria[35]".[36]

Potete immaginare come, di fronte a quantità di dati enormemente più grandi e complessi, possa risultare estremamente facile incorrere in errori di interpretazione. Per questo esistono i già citati

---

Figura 5: tratta dall'articolo «Focus», *Quando i dati mentono*, maggio 2016.

[35] In statistica, *dati spuri*, dati non appartenenti alla popolazione statistica che si vuole analizzare, e che pertanto devono essere eliminati, con opportuni criteri di selezione, prima dell'elaborazione.
Tratto dal sito http://www.treccani.it/vocabolario/spurio/ (1 ottobre 2017, 16.45).

[36] «Focus», *Quando i dati mentono*, maggio 2016.

computational scientist, in grado di svolgere nella maggior parte dei casi un'accurata e corretta analisi dei dati a disposizione.

3.7. Voices from the Blogs

Eccovi ora un esempio concreto di un istituto che lavora analizzando maxi-flussi di dati. Il suo nome è Voices from the blogs (VfB). Cosa sia esattamente ci viene spiegato dal suo cofondatore Luigi Curini, nell'intervista da me realizzata: "Voices from the Blogs è nato nel 2011 quale centro di ricerca su iniziativa mia e di Stefano Iacus. Il suo obiettivo […] era quello di sviluppare nuovi metodi per l'analisi del flusso di informazioni presenti sui social media. Poi […] l'Università degli Studi di Milano ci ha chiesto: 'Perché non creiamo insieme anche uno spin-off[37]?'.". L'intervista integrale la trovate quale allegato al mio lavoro.

Uno dei principali progetti di cui si occupa VfB verte sul livello di felicità degli abitanti dello Stato italiano. L'indice di felicità iHappy è costruito a partire dalle reazioni istantanee degli utenti di Twitter-Italia agli eventi che caratterizzano la loro vita, pubblica e privata. Questi eventi possono avere un'influenza positiva o negativa sull'individuo, che lo potrà comunicare destreggiandosi nei 140 caratteri messi a disposizione da Twitter per ogni singolo *tweet*.
Il social, tramite algoritmi di *machine learning*, categorizza la totalità dei commenti postati in ogni momento. Sarà etichettato come "felice", se il post conterrà messaggi di gioia o allegria, o "infelice", se conterrà rabbia, paura o ansia. Qui entra in gioco VfB, che, attraverso questa prima analisi effettuata da Twitter, ricostruisce il tasso di felicità italiano. Alla "semplice" analisi di Twitter, VfB impiega un algoritmo che porta il nome di Integrated Sentiment Analysis (ISA), aggiungendo una terza caratterizzazione dei *tweet*, che corrisponde alla categoria "altro". Questi *tweet* per difficoltà di analisi non vengono considerati nel calcolo dell'indice iHappy.
Qui di seguito potete visionare la formula utilizzata dai computer per calcolare l'indice di felicità. [38]

$$\text{iHappy} = \left(\frac{numero\ di\ post\ felici}{numero\ di\ post\ felici\ \&\ infelici}\right) * 100\%$$

[37] Impresa nata per scorporamento da un'altra, la quale mantiene tuttavia un ruolo fondamentale nei confronti della nuova realtà imprenditoriale, esercitando su di essa una significativa influenza soprattutto in termini di competenze e di attività svolte.
Tratto dal sito http://www.treccani.it/enciclopedia/spin-off_%28Dizionario-di-Economia-e-Finanza%29/ (8 novembre 2017, 17:41).

[38] CERON Andrea, CURINI Luigi, IACUS Stefano M., *iHappy 2016*, Corriere della Sera, Milano.

4. Big Data e opportunità: la statistica e le scienze

«Studia il passato se vuoi prevedere il futuro.»
Confucio (551 a.C – 479 a.C)
Filosofo cinese.

4.1. Il bello dei Big Data

I Big Data permettono l'analisi in tempo reale in molti campi della quotidianità, come vi mostrerò fra poco, portando coloro che stanno ai vertici della società a dover prendere decisioni importanti.
I Big Data sono diventati anche un fattore di produzione, ovvero uno strumento grazie al quale le aziende sono in grado di produrre beni e servizi da offrire ai propri acquirenti. In questo modo si crea un circolo virtuoso, nel quale noi tutti guadagniamo qualcosa: le aziende in beni finanziari, noi utenti in servizi migliori e mirati, magari anche a un prezzo più vantaggioso.
Avendo acquisito questo enorme valore è molto importante rendersi conto quale posizione abbiamo in questo grande universo dei dati. Noi, siamo sia i produttori dei dati, sia i consumatori della conoscenza derivante dal processo di analisi.
Proprio per queste ragioni – la loro immediatezza e la loro di capacità di generare beni e servizi – i Big data stanno determinando dei cambiamenti a livello economico e sociale. Infatti, le imprese che negli ultimi anni si sono interessate a questo nuovo universo e hanno cominciato a fare incetta di dati, ora si trovano con le tasche piene e con i prodotti più richiesti e all'avanguardia.

Secondo alcuni esperti in materia informatica, i Big Data somigliano un po' al linguaggio scritto: infatti, come per la scrittura, si arriva al punto in cui per capire le cose si ha la necessità di scriverle, perché troppo complesse. Con i Big Data si è raggiunto il medesimo punto: infatti, per poter comprendere questa immensa mole di dati di cui sono costituiti, occorre che vengano "scritti": ovvero che i dati già analizzati vengano composti nel modo corretto, così da poter essere visualizzati nel modo giusto, senza incorrere in errori di interpretazione.
Gli esperti affermano che in questo preciso momento della storia possiamo evolverci solamente se scriviamo e analizziamo tutte le informazioni di cui siamo in possesso. Infatti, più dati si avranno da analizzare e con i giusti mezzi, che giorno dopo giorno saranno sempre più performanti, più tramite una corretta visualizzazione otterremo informazioni utili e più problemi si potranno risolvere. In poche parole: più dati saranno stoccati, più difficoltà si potranno affrontare e superare.
C'è però una cosa da tenere ben presente: fino a che non si combini la quasi totalità dei dati raccolti negli ultimi anni con una qualche interpretazione e non le si crei attorno una sorta di storia, tutti

questi dati rimarranno perfettamente inutili. Perciò è scorretto parlare di Big Data come se si parlasse di conoscenza. L'ammasso di dati non lo è. Si può parlare di conoscenza solamente una volta che sono stati analizzati e resi in qualche modo visibili tramite appositi programmi, diventando così utilizzabili.[39]

C'è però un "ma". La paura derivante dai recenti scandali che hanno scosso il villaggio globale non ha fatto altro che mettere in cattiva luce una nuova e straordinaria tecnologia e presentarla all'umanità come qualcosa di sospetto o addirittura di cattivo, utilizzata solamente da entità governative e criminali per spiarci da vicino senza il nostro consenso, o per spillarci quattrini attraverso offerte calibratissime sui nostri profili. Qui si ha un chiaro esempio di cattiva pubblicità: infatti la maggior parte dell'opinione pubblica, pur alimentando di giorno in giorno il grande magazzino mondiale delle informazioni, guarda ai Big Data come a qualcosa di oscuro. Qualcosa di cui diffidare. Questo è un peccato perché, come vedremo in questo capitolo, la scienza dei dati può essere usata per migliorare le condizioni di vita del mondo intero, così come potrà risolvere problemi che fin qui appaiono insormontabili. Paura e diffidenza sono anche un segnale da non sottovalutare, come se chi diffida non capisse: no, il nuovo tesoro nella caverna dei dati di Ali Babà va gestito con regole all'altezza dei nuovi tempi tecnologici e va gestito da chi pensa effettivamente ai benefici per la collettività. In altre parole, va protetto dalle mani di potenziali briganti della rete ai quali fa indubbiamente gola.
In proposito ho voluto anche svolgere un sondaggio sugli aspetti della conoscenza e/o diffidenza verso la Data Science per capire se davvero l'opinione pubblica sia stata influenzata solamente in negativo dagli scandali scoppiati negli ultimi anni. Al capitolo "6. Questionario" vedrete quanto realmente l'opinione pubblica conosca poco sul tema Big Data (Domanda 1) e anche quanto questa ignoranza causi paura (Domanda 5).

4.2. Prevedere il futuro: ora è possibile

4.2.1. Anticipare il diffondersi di epidemie

I Big Data sono anche in grado di aiutare a prevedere l'arrivo di epidemie di influenza. Questa non è una novità, perché, puntualmente ogni anno, siamo stati abituati a ricevere avvisi mirati sull'arrivo di una nuova ondata di influenza, pure prima della comparsa dei Big Data. Fino a qualche anno fa, la diramazione di allerte d'epidemia, dalla comparsa dei primi casi di influenza, richiedeva circa

[39] dal documentario *The human face of big data*, 2014, Sandy Smolan, Bbc.

due settimane prima di arrivare alle orecchie della popolazione. Ciò avveniva grazie all'accumulo di casi simili da parte dei medici, che poi inviavano le cartelle cliniche al Centro Controllo Malattie.
Oggi non è più così. Con l'impiego dei Big Data è avvenuto un cambiamento sostanziale: gli esperti, analizzando le ricerche fatte sul web degli utenti di una zona colpita da influenza, si sono resi conto che è possibile prevedere in tempo reale il diffondersi di un'epidemia tramite proprio queste ricerche. La ricerca sul web di determinati termini e frasi da parte di più componenti della società, indicano infatti il diffondersi di un'epidemia.
Tutto ciò è immediato. Basta infatti applicare algoritmi di *machine learning* alla moltitudine di dati presenti. Se verrà superata una soglia scelta a priori del numero di ricerche di quei termini identificati dagli esperti come sentori di influenza, ecco che scatterà l'allarme. Tutto ciò porterà, con ben due settimane di anticipo rispetto al metodo "tradizionale", ad individuare la malattia, spingendo la popolazione a comportarsi di conseguenza, in modo da fermare il contagio.
Il problema dell'influenza è molto serio: causa circa 500 mila morti all'anno
Con questo esempio si può facilmente capire come l'impiego della *Data Science* nei tradizionali settori della società porta a risultati **immediati** e non più **mediati**.
C'è un'abissale differenza tra questi due termini, che spesso non vengono visti come contrari. Con "mediato" si intende qualcosa che viene mediato da qualcuno o più persone (dal verbo mediare), mentre con "immediato" si intende invece qualcosa che non viene in alcun modo mediato da terzi ma accade di colpo, in maniera fulminea.

L'unica pecca di questo sistema è che, se per un qualche motivo i media cominciano a parlare e a condividere informazioni sull'epidemia in questione, l'interesse mediatico aumenterà, spingendo gli internauti a compiere ricerche in merito e portando l'algoritmo (di cui ho parlato prima) ad individuare l'avvento di una nuova epidemia di influenza.

Le informazioni che seguiranno provengono dalla mia intervista effettuata alla professoressa Antonietta Mira. Trovate la versione integrale nel capitolo "

11. Allegati".

Un esempio dell'applicazione di questo sistema è Google Flu Trend, lanciato nel 2008. È in grado di stimare il rischio di epidemie, basandosi sulla quantità di ricerche di determinate parole chiave relazionate all'influenza effettuate in rete dai navigatori. Ricerche come sintomi, medicamenti anti-influenzali, farmacie o ospedali.

Questo sistema offre vantaggi rispetto al sistema tradizionale. Innanzitutto è geolocalizzato, perciò si può conoscere l'evoluzione territoriale dell'epidemia ed è appunto immediato.

GFT si è dimostrato accurato nelle previsioni dagli anni precedenti al suo lancio sino al 2008. In questo periodo la sua accuratezza si aggirava intorno al 97%, risultava quindi decisamente accettabile. Passo per passo ci si è accorti che però, oltre all'arrivo dell'influenza, esso prevedeva l'arrivo della stagione invernale. È probabilmente a causa di questo fattore, e degli accorgimenti attuati da Google, che nel 2009 GFT non è stato in grado di considerare con dovuta rilevanza la tanto mediatizzata influenza suina, che dall'OMS era pure stata bollata come epidemia.

Ha poi lavorato adeguatamente sino al 2012, anno dal quale l'algoritmo ha ricominciato a generare problemi. Questa catena di disguidi ha poi favorito la sottostima dell'epidemia del 2013.

Il motivo delle oscillazioni è stato ricondotto al cambiamento delle abitudini di ricerca delle persone, che suggestionate dalla frequenza e prepotenza con cui i media parlano della natura pandemica di queste influenze, hanno ricercato questi termini chiave solamente a fini informativi. Per esempio per venire a conoscenza di eventuali metodi di prevenzione.

Curiosamente, Google non ha mai rivelato la propria lista di termini chiave studiati per individuare l'arrivo dell'influenza, nel suo caso 45 termini. Così facendo viola uno dei principi fondamentali della ricerca scientifica: il fatto che altri possano fare altrettanto e confrontare. Il termine tecnico è "riproducibilità". Ma per gli istituti e le aziende di indirizzo commerciale attivi nel settore Big Data è una cosa normale. Essi agiscono esclusivamente nella logica del profitto. A detta di Antonietta Mira è dunque di stragrande importanza che "[…] le università si riapproprino di questi ambiti.".

Esistono altri sistemi che si ripropongono il medesimo obiettivo di quello appena descritto. Per esempio Flu Near You, anche se il suo funzionamento diverso: è basato su un network di 46'000 persone che riportano i casi di influenza, ricoprendo un numero di circa 70'000 persone.

Questo sistema è però legato maggiormente al metodo tradizionale invece che all'utilizzo dei Big Data.[40]

[40] Intervista a Antonietta Mira

4.2.2. La sicurezza urbana: *Predicting Crime*

Le autorità hanno capito di poter sfruttare i Big Data anche per monitorare il livello di sicurezza di una determinata zona, calcolando il suo tasso di criminalità. Ciò avviene specialmente negli Stati Uniti, dove in certe aree la malavita dilaga incontrastata, ma anche alle nostre latitudini, per poter vivere una vita più sicura.

Alcuni studiosi della University of California in collaborazione con il Los Angeles Police Department hanno sviluppato un progetto per raccogliere e analizzare tutti i dati criminosi presenti negli archivi del dipartimento di polizia. L'analisi è avvenuta su circa 13 milioni di crimini, commessi negli ultimi 80 anni. Il progetto originale era stato pensato per determinare matematicamente la locazione delle scosse di assestamento dopo un terremoto, ma si ha poi cambiato strada, puntando al crimine con questo nuovo progetto che porta il nome di PredPol, da *predicting police*.

L'interesse degli studiosi era di esaminare la quantità e il tipo di crimini avvenuti nelle diverse zone della città. Così facendo hanno aperto le porte del mondo dei *data* anche alla dimensione della sicurezza urbana, che ben presto ne ha capito l'importanza e come poterla sfruttare al meglio.

Interi *team* di matematici, statistici, informatici, antropologi e criminologi con l'impiego di complessi algoritmi di analisi, chiamati *Robocop*[41], hanno architettato un *software* in grado di fornire interessanti e utilissime informazioni sul dietro le quinte della metropoli di Los Angeles. Tramite la visualizzazione dei dati analizzati si è potuto costruire la mappatura dei crimini avvenuti nella città, correlata a molteplici fattori che a prima vista apparivano totalmente scollegati. Sfruttando poi anche l'intelligenza artificiale ci si è accorti di poter prevedere con una certa precisione le aree della città nelle quali i futuri crimini avrebbero potuto avere luogo.

Figura 6: tratta dal sito http://news.wabe.org/post/concerns-arise-over-new-predictive-policing-program (23 luglio 2017, 11:22).

[41] Tratto dal sito http://www.rsi.ch/news/oltre-la-news/Prigionieri-dei-Big-Data-8434276.html (22 luglio 2017, 17:55).

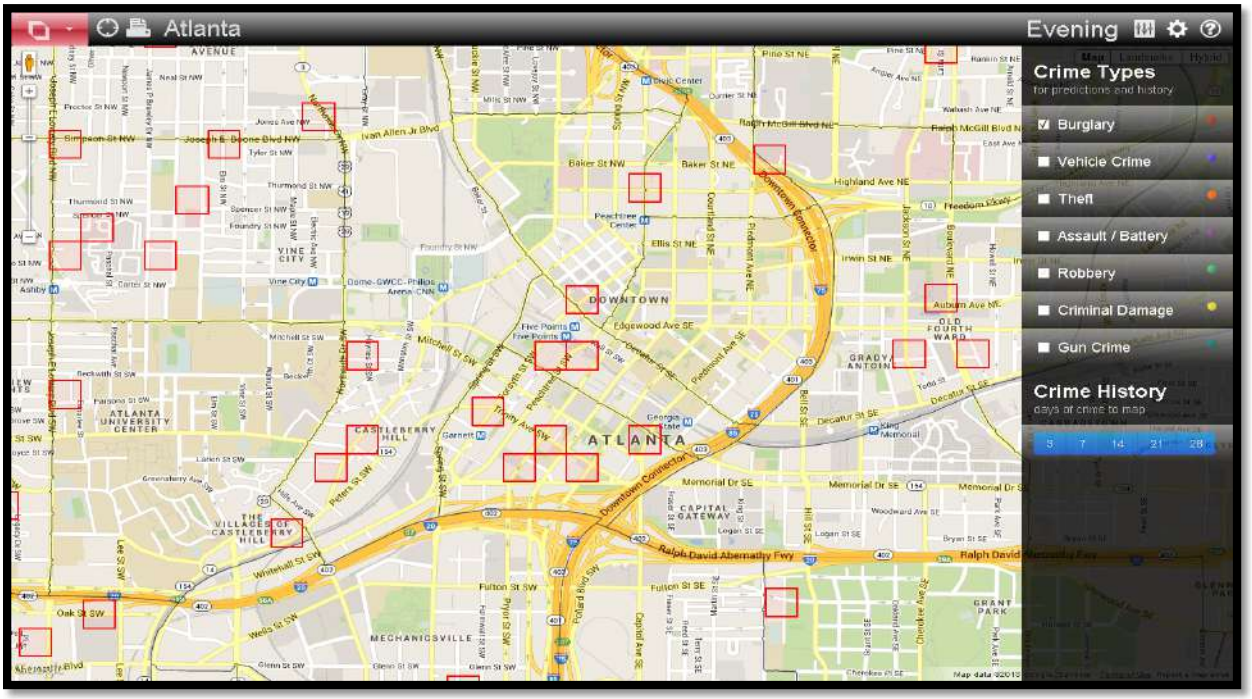


Figura 6 – Immagine che mostra le zone della città di Atlanta nelle quali è probabile si verifichi un "burglary", ovvero un furto con scasso.

In questo modo il comando di polizia ha potuto agire di conseguenza, indirizzando il dispiegamento delle sue forze dove e nel momento in cui la probabilità che alcuni crimini venissero commessi risultava essere maggiore.

I risultati sono netti: in poco tempo i furti in proprietà private sono diminuiti del 12%, mentre i furti con scasso negli appartamenti sono diminuiti addirittura del 28%.

Per rendere sempre più efficiente il *software*, le informazioni relative ai nuovi crimini vengono costantemente inserite nel programma. Le previsioni estrapolate sono così sempre più precise e accurate, permettendo alle forze dell'ordine di compiere un lavoro egregio nel difendere la città dai crimini.[42] [43]

Questo tipo di sicurezza è molto diverso da quello conosciuto fino a qualche anno fa, dove i dispiegamenti delle forze di polizia in una città erano sì influenzati dal tasso di crimini di una determinata zona, ma solo minimamente. Ora la sicurezza urbana ha assunto la forma di qualcosa di futuristico: il nuovo sistema permette di anticipare ciò che probabilmente accadrà mediante l'utilizzo di statistica e nuovi strumenti di origine digitale.[44]

[42] «Corriere del Ticino», Un universo di informazioni che può svelare tutto di noi, 21 aprile 2017.

[43] tratto dal sito https://www.predpol.com (23 agosto 2017, 23:22).

[44] cfr. BAUMAN, LYON, pag. 11.

4.3. Truffe con carte di credito (*Fraud detection*)

Passo ora ad un altro esempio dell'utilizzo dei Big data. Anche dai dati relativi alle sole transazioni effettuate con carte di credito è possibile estrarre importanti ed interessanti informazioni.

Viseca è una delle maggiori aziende svizzere impegnate nella fornitura di carte di credito[45], gestisce addirittura il 30% delle transazioni con carte di credito in Svizzera. Questa sua posizione nel mercato bancario la porta inevitabilmente a possedere una quantità massiccia di dati sui consumatori. Infatti, tutto ciò che può essere fatto con una carta di credito viene abitualmente registrato.

Una piaga che ha sempre accompagnato i possessori di carte di credito è quella delle frodi. Ci sono infatti svariati sistemi e tecniche per poter clonare o rubare la carta a qualcuno. Questo porta inevitabilmente allo svuotarsi del conto del cliente e, una volta accortosi del danno, al blocco della carta.

Tutto questo è veramente un problema. Oltre ad essere uno spreco enorme di tempo sia per l'istituto, sia per il possessore della carta, è anche un continuo dover contattare l'istituto bancario per bloccare e rifare la tessera.

Viseca allora si è domandata: "Come possiamo risolvere questo problema in modo più efficiente?", ed è proprio qui che entra in campo la Data Science, che offre un sistema parallelo a quello tradizionale e – nella maggior parte dei casi – istantaneo.

Per questo motivo Viseca si è rivolta all'Istituto Interdisciplinare di Data Science (IDIDS), che, finanziato dall'AXA Foundation[46], ha costruito un algoritmo basandosi sulle analisi di tutti i dati comprendenti transazioni fraudolente. In questo modo l'algoritmo è in grado di identificare transazioni fraudolente, con ottime probabilità di successo, prendendo in considerazione le caratteristiche che ha ritenuto opportuno associare a questa tipologia di reato.

Infatti, spesso, essi presentano le stesse dinamiche e caratteristiche. Per esempio, una volta rubata o clonata la carta, il ladro tende sempre a prelevare la quantità di denaro più elevata possibile nei giorni seguenti al furto. Questo potrebbe essere uno dei sentori di allarmi che l'algoritmo tiene in considerazione durante la sua analisi. In molte dinamiche, è facilmente riscontrabile una evidente ricorrenza, che in questo caso rende possibile il riconoscimento di prelevamenti e transazioni illecite.

[45] tratto dal sito https://www.viseca.ch/it/chi-siamo (7 dicembre 2017, 16:25).

[46] Fondazione dell'assicurazione svizzera AXA, operante in tutto il mondo.

Con questo sistema, Viseca è stata in grado di minimizzare i falsi allarmi, con conseguenti blocchi ingiustificati di carte, così come di minimizzare i tempi che incorrono tra la frode e il blocco della carta.[47]

4.4. Un progetto tutto ticinese

Se vivete in Ticino avrete già sicuramente sentito parlare della Fondazione Ticino Cuore e del Cardiocentro Ticino. Per chi non lo sapesse, entrambi, come intuibile dal nome, trattano di problematiche legate a patologie cardiache.

La prima è una fondazione istituita nel 2005, con "[…] *lo* scopo principale *di aumentare* la sopravvivenza delle persone colpite da arresto cardiaco." [48], come riportato sulla loro pagina web. Mentre il Cardiocentro Ticino è una clinica universitaria associata all'Università di Zurigo e fin qui gestita da una fondazione privata. È anche riconosciuta a livello internazionale per la sua elevata specializzazione in cardiologia, cardiochirurgia e cardioanestesia.[49]

Qualche anno fa, per perseguire la sua missione, la Fondazione ha raggiunto l'obiettivo di acquistare un certo numero di defibrillatori da poter piazzare sul territorio ticinese. In questo modo, se qualcuno fosse stato colto da un malore mentre si trovava in giro, o in seguito ad un incidente, non avrebbe dovuto attendere l'arrivo dell'ambulanza per ricevere soccorso. Spesso i minuti *post factum* sono decisivi nel salvataggio del malcapitato, che proprio in questi minuti, in caso di un tardo soccorso, rischia la vita o danni permanenti.

La presenza di un defibrillatore nelle vicinanze contribuisce già molto ad assistere in modo migliore chi ha bisogno, ma ciò non basta: le morti sono sì diminuite, ma la Ticino Cuore non si è fermata. Per continuare secondo i propri ideali, i membri della Fondazione si sono chiesti se ci fossero differenze probabilistiche che un dato incidente, necessitante la presenza di un defibrillatore, si verificasse in un preciso punto del Cantone. Se questa probabilità fosse realmente esistita, invece di comprare nuovi defibrillatori – che oltretutto costano diverse migliaia di franchi – si sarebbero potuti riposizionare in punti strategici quelli già in loro possesso.

Per rispondere a questa domanda si sono rivolti all'Istituto Interdisciplinare di Data Science (IDIDS) – cofondato e ora diretto dalla professoressa Mira – che si è occupato di analizzare tutte le telefonate provenienti al servizio ambulanze ticinese legate a patologie cardiache. Dati gentilmente concessi dal Cardiocentro.

---

[47] Intervista a Antonietta Mira

[48] tratto dal sito http://www.ticinocuore.ch/it/chi-siamo (9 dicembre 2017, 16.42).

[49] tratto dal sito https://www.cardiocentro.org/it/cardiocentro/il-centro/ (9 dicembre 2017, 16:54).

Attraverso queste informazioni hanno individuato i punti migliori dove spostare gli apparecchi, ma l'analisi non è finita qui. L'IDIDS ha fatto richiesta al Cantone Ticino per poter ottenere la planimetria di tutti i palazzi ticinesi. Questo per evitare che nella mappatura finale non risultasse magari conveniente posizionare un defibrillatore in mezzo a un lago o sulla cima di una montagna. Mediante l'accorpamento di tutte queste analisi è stato possibile redigere una mappa del Ticino che mostrasse i punti migliori nei quali riposizionare gli apparecchi.[50]

[50] Intervista a Antonietta Mira

5. Big data e rischi: la privacy e il Grande Fratello

«L'aspetto più triste della vita in questo momento è che la scienza raccoglie conoscenza più velocemente di quanto la società non raccolga saggezza.»

Isaac Asimov (1920 – 1992)

Scrittore di fantascienza e biochimico americano.

5.1. Il brutto dei Big Data

Nello scorso capitolo "4. Big Data e opportunità: la statistica e le scienze" ho esposto alcune opportunità che presenta la nuova Data Science e come nei prossimi anni potrà essere sfruttata al fine di migliorare le condizioni di vita del mondo intero nella sanità, nella sicurezza e altro ancora. Il cambiamento avverrà di sicuro – anzi, è già in atto –, vivendo in una società basata sull'utilizzo di questo nuovo e importante sistema di analisi dei dati. Ciò sarà certamente avvertito da tutti, ma non dobbiamo dimenticarci che, se una cosa può cambiare il mondo, lo può anche cambiare in peggio.

Come tutte le cose di grande valore – e i dati lo sono - anche i dati sono oggetto di grande interesse da parte di moltissime persone e aziende. Alcune con intenti buoni, che possano giovare ad un ulteriore sviluppo della società mondiale; altre senza scrupoli e con intenzioni di natura malvagia e criminale. Spaventa e ci fa rendere attenti pensare che i nostri dati, che ricordo, spesso e volentieri, noi non forniamo in maniera totalmente spontanea, possano venire rubati da questa seconda tipologia di individui. La domanda sorge spontanea: cosa può fare un malintenzionato (se non addirittura ramificate organizzazioni criminali) entrando in possesso dell'identità digitale, parziale o completa, della nostra persona?

Questa paura, che riguarda le più recondite parti della nostra sfera privata, spinge il consumatore del servizio internet a rendersi conto di essere parte di un circolo vizioso creato *ad hoc* per favorire la sorveglianza del singolo individuo. Individuo che cerca di tutelarsi come può con i mezzi a sua disposizione, spesso rivolgendosi ad esperti del settore, quali avvocati o informatici, per cercare di frenare questa folle e inarrestabile emorragia di dati. Anche se, così facendo, lo squarcio da cui fuoriescono le informazioni si fa sempre più grande. Come ha scritto Bauman nel suo libro "Sesto potere: La sorveglianza nella modernità liquida": "*Spesso* l'eccessiva sicurezza danneggia la sicurezza effettiva […]." [51].

[51] BAUMAN, LYON, pag. 118.

A ciò si aggiunge che, pur trovandosi a disporre di una quantità sempre maggiore di informazioni, paradossalmente la persona non è in grado di aumentare la sua conoscenza. Questo perché non siamo più spinti a ricordare ciò di cui abbiamo bisogno, bensì siamo portati a memorizzare il luogo in cui esso si trova. Difatti, l'inconscia consapevolezza di poter trovare le informazioni di cui si necessità online, porta la persona a memorizzare molto meno di quanto presente nel database trovato, inducendo l'individuo a inquadrare con più precisione dove questi dati di interesse siano a disposizione per un'eventuale futura ricerca.[52]
Nel Terzo Millennio la concezione del sapere è mutata ed è totalmente differente da quella tradizionale (a tale cambiamento ho già brevemente accennato sul finire del capitolo "3.4. Come i dati diventano conoscenza", parlando di "dotti eclettici").
Ovviamente la società ha subito dei mutamenti, a posteriori per certi versi impensabili, ma questo cambio di veduta è in gran parte riconducibile alla quantità e alla fruibilità delle informazioni.

"Dovremmo riflettere. Quando firmiamo un consenso – sottolinea la professoressa Mira – rinunciamo al possesso dei nostri dati e perdiamo una parte più o meno grande della nostra privacy. Oggi il pericolo non è avere un unico Grande Fratello, ma piuttosto trovarci di fronte a tanti piccoli Grandi Fratelli, che sanno molto di noi. Tutte le aziende che raccolgono dati (ad esempio Google, Facebook, Twitter, Amazon, Microsoft,…) diventano proprietari di tracce digitali che raccontano non solo il nostro passato, ma anche quello che faremo nel futuro."[53]
Inoltre, questi frammenti di dati personali vengono spesso raccolti per un determinato scopo ma, proprio dai nuovi proprietari, vengono utilizzati con altri fini, sempre più spesso volti al guadagno.[54]
Tutto ciò può sembrare strano, se si pensa al fatto che internet, inizialmente, era stato adottato dai governi per condividere con la popolazione le informazioni governative più importanti. E veniva utilizzato esclusivamente per questo. In origine era quindi il popolo a controllare il governo, mentre oggi ci ritroviamo catapultati nella situazione opposta: una dimensione (purtroppo) caratterizzante i totalitarismi, ma molto più silenziosa ed evoluta. Oltre che alle grandi aziende, oggi è il governo a controllarci.

---

[52] SPARROW Betsy, LIU Jenny, WEGNER Daniel M., *Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips*, 2011.

[53] stralcio di intervista tratto dal sito http://www.rsi.ch/news/oltre-la-news/Prigionieri-dei-Big-Data-8434276.html (13 settembre 2017, 15:02).

[54] cfr. BAUMAN, LYON, pag. 8.

Intimorisce pensare che interi calcolatori vengano riempiti di informazioni (private e di dominio pubblico) sulla nostra persona, proprio per conto dello Stato. Stato al quale noi stessi, attraverso un "contratto sociale" [55], abbiamo affidato parte della garanzia dei nostri diritti e accettato doveri e che noi stessi abbiamo in qualche modo contribuito a rendere tale. Su questo argomento al "6. Questionario" mostrerò quanto gli interpellati si dicano spaventati dalla raccolta dati (Domanda 5).
Recenti scandali hanno portato alla luce programmi governativi non ufficiali, aventi come unico scopo quello di sorvegliare e prevedere gli spostamenti e le azioni di milioni di persone. Il mondo ha tremato e ha cominciato a riflettere sui rischi di questa nuova e perentoria evoluzione della tecnologia e dei sistemi di stoccaggio e analisi dati.
Porto come esempio il caso che forse ha fatto più scalpore: quello che in tutto il mondo è ormai conosciuto come *Datagate*.

5.1.1. *Datagate*: sorveglianza di massa

Era il 6 giugno del 2013 quando il quotidiano inglese Guardian e quello statunitense Washington Post cominciarono a pubblicare documenti *top secret* riguardanti programmi segreti dell'Agenzia di Sicurezza Nazionale (NSA). In quel giorno, che può non dirvi nulla, fece grande scalpore la notizia che l'NSA portava avanti sin dal 2001 un programma di sorveglianza di massa su ignari cittadini statunitensi e esteri. Programma che arrivava persino a spiare l'operato e le decisioni più segrete dei governi di altri paesi.
Il protagonista di questo scandalo globale è lo statunitense Edward Snowden, ex-tecnico informatico presso la CIA[56] e collaboratore dell'azienda di tecnologia informatica Booz Allen Hamilton, a sua volta consulente della NSA. Infatti è stato proprio lui, Snowden, a ribellarsi al sistema, decidendo di diffondere le informazioni riguardanti il suddetto programma di sorveglianza.

I media battezzarono lo scandalo col nome di *Datagate*. La traduzione italiana risulta fuorviante ("Datagate" significa letteralmente "cancello-dati"), ma in realtà ogni scandalo che negli *States* raggiunge un clamore di un certo livello, nazionale o globale, viene battezzato dalla stampa accostando il nome del tema in questione al suffisso "gate". Un'abitudine derivante da Watergate, il

[55] "Patto ipotetico attraverso cui individui appartenenti a una stessa società decretano leggi che sottendono al suo fondamento."
Tratto da http://www.treccani.it/enciclopedia/contratto-sociale-teoria-del_%28Dizionario-di-Economia-e-Finanza%29/ (3 dicembre 2017, 15:00).

[56] Central Intelligence Agency

nome dell'albergo teatro di uno scandalo che nel 1974 portò il presidente statunitense Richard Nixon alle dimissioni.[57]

Diversamente da come accadeva in passato, quando lo spionaggio si basava sulla pesca di precise informazioni, oggi lo spionaggio si basa sulla raccolta massiccia di dati. Una raccolta indistinta, alla quale si applicheranno poi, in un eventuale futuro, algoritmi progettati appositamente per estrarre determinati tipi di informazioni. Oggi chi spia detiene tutte le informazioni. Poi, a dipendenza dell'esigenza del momento, potrà semplicemente fare un'analisi per estrarre le informazioni necessarie.

La NSA era in possesso di innumerevoli "porte sul retro" dei maggiori fruitori di servizi tecnologici e digitali. Con "porte sul retro" intendo "possibilità di venire in possesso di dati non ufficialmente, ma con un tacito accordo con questi principali fornitori di servizi. Poniamo che c'è un accordo con Apple – azienda realmente invischiata nello scandalo – ecco che posso facilmente accedere alle informazioni che la persona da me spiata produce col suo *melafonino*. Teoricamente, il programma è stato iniziato dalla NSA per proteggere gli Stati Uniti dal terrorismo, in seguito agli attacchi subiti nel 2001. Poi, in sordina, si è allargato ad una sempre più vasta sorveglianza di persone, molte delle quali totalmente estranee a sospetti di terrorismo. Persino il telefono della cancelliera tedesca Angela Merkel è stato posto sotto sorveglianza per un periodo di dieci anni, così come i telefoni di altri 34 capi di Stato, per durate differenti.

È così che Edward Snowden nel 2013, una volta venuto in possesso di decine di migliaia di documenti "inaccessibili" e segreti, ha deciso di ribellarsi al sistema. I suoi ideali di giustizia e di morale erano in disaccordo con quanto scoperto. Rifugiatosi a Hong Kong, ha consegnato diversi fascicoli al Guardian e al Washington Post, che hanno rivelato al mondo intero lo scandalo. [58] [59] [60]
Dopo richieste d'asilo negate in numerosi Stati, come l'Ecuador, Snowden si trova ora a Mosca, dove possiede un permesso di residenza fino al 2020.[61]

---

[57] tratto dal sito https://it.wikipedia.org/wiki/Richard_Nixon (11 ottobre 2017, 13:21).

[58] tratto dal sito http://www.lastampa.it/2013/10/27/esteri/cose-da-sapere-sullo-scandalo-oSs4F1uOX5VuWvtIBx29YP/pagina.html (10 dicembre 2017, 22:13).

[59] dal film *Snowden*, 2016, Oliver Stone.

[60] dal documentario *Citizenfour,* 2014, Laura Poitras.

[61] tratto dal sito http://www.repubblica.it/esteri/2017/02/11/news/putin_pronto_a_consegnare_snowden_a_trump-158064266/ (8 dicembre 2017, 22.30).

Figura 7 – Immagine che riprende ironicamente il famoso slogan "Yes we can", utilizzato durante la campagna presidenziale del 2008 da Barack Obama, trasformandolo in "Yes we scan". Ciò a fronte delle rivelazioni di Snowden del 2013.

5.2. Così regaliamo i nostri dati…

5.2.1. La trappola dei social network

Come detto, abbiamo paura che qualcuno sappia tutto di noi. Eppure siamo noi a spogliarci nella rete. In particolare sui cosiddetti social network, dall'apparenza tanto innocua e attrattiva. Nel nostro universo i social network (in italiano "reti sociali") sono sempre più utilizzati dai più disparati componenti della società, così come nei più diversi ambiti. Sempre più spesso vengono utilizzati per pubblicare e condividere informazioni di natura sensibile e riservata. In questo modo una grandissima quantità di informazioni private ci sfugge di mano, quasi senza che noi ce ne rendiamo conto.

Nella vita reale tutte le volte che sottoscriviamo un contratto è per stipulare un accordo, nel quale cediamo qualcosa di nostra proprietà a terzi in cambio di un servizio. Ciò che l'utente medio ignora quando si iscrive ad un qualsiasi social network è tanto semplice, tanto evidente. L'utente diventa infatti partner di una transazione nella quale dona i propri dati personali al social in questione.

Figura 7: tratta dal sito http://morrisjfwong.com/blog/yes-we-can-yes-we-scan/

Questa maglia di rete sociale trarrà poi grandi benefici rivendendo i dati ceduti ad aziende e compagnie esterne. In cambio si ottiene il servizio di poter postare foto e quant'altro, per poter poi condividere il tutto con gli amici. Amici reali, virtuali o semi-virtuali che hanno a loro volta concesso al social il completo e libero utilizzo delle proprie informazioni personali.
I social network esistono infatti esclusivamente grazie al monitoraggio degli utenti e alla vendita di tali informazioni a terzi.[63]
Insomma, i benefici del web hanno un costo, anche se non ce ne rendiamo sempre conto, che paghiamo con un bene estremamente prezioso: la privacy.
Nell'ambito privacy e social network c'è stato un cambio di mentalità negli ultimi anni, forse dovuto ai recenti scandali già citati all'inizio di questo capitolo (*Datagate*). Infatti i cittadini iniziano a prendere coscienza delle informazioni che pubblicano in rete e sono più cauti.

5.2.2. I famigerati *cookies*

Galeotti in questo processo "racimola dati" sono i famigerati *cookies*. Galeotti perché ci fanno dimenticare la paura che qualcuno nell'ombra ci stia osservando e ci freghi i dati. Sicuramente durante qualche ricerca su internet sarete capitati su almeno un sito che vi abbia chiesto di accettare i *cookies*. Voi avrete mosso il mouse e cliccato sul "Consentire" o sul "Non consentire"; ma vi siete mai chiesti cosa effettivamente sono questi *cookies*?
Il termine è ingannevole. Infatti, tradotto in italiano, significa "biscotti" che con i *cookies* non hanno proprio niente a che fare.
Per facilitare la navigazione sul web la maggior parte dei siti che si visitano chiede di poter inserire negli hard disk dei visitatori dei cosiddetti "marcatori digitali", altrimenti conosciuti come *cookies* o *markers*.
La funzione svolta da un marcatore digitale è la seguente: una volta impiantato nel disco rigido del computer in questione, tutte le azioni che si svolgeranno con tale macchina verranno registrate e prontamente inviate al server del sito del quale sono stati accettati i *cookies*.
Questo processo serve a velocizzare tutte le azioni che l'utente compie nella rete. Infatti i principali utilizzi dei cookies sono ricordare username e password, evitando ogni volta di dover reinserire i dati per effettuare l'accesso ad un sito; ricordare le preferenze impostate su una determinata pagina web, come argomenti che aggradano maggiormente; tenere traccia dei prodotti marcati come interessanti per un eventuale acquisto futuro in un sito di vendita.[64]

[63] cfr. BAUMAN, LYON, pag. 13.

[64] cfr. CASTELLS, pag. 163.

Così facendo i siti internet sono in grado di formare quella che potrei definire *identità digitale*, ovvero la costruzione digitale della nostra persona sulla base del nostro comportamento sul web.

Per riuscire a "impiantare" questi *markers*, spesso, la richiesta di accettazione viene formulata utilizzando una sintassi appositamente preparata: un po' ingannevole, poco chiara, che trascura informazioni che forse sarebbe corretto fornire.
Nella Figura 8 illustro tre differenti tipologie di richiesta di accettazione dei cookies da parte di tre diverse piattaforme online. La prima proviene dal *provider* internet Zalando Svizzera. In tal caso la richiesta è concisa e diretta. Vengono date scarne informazioni e viene sottolineato esclusivamente il fatto che accettare i *cookies* serve solamente a migliorare il servizio fornito.
La seconda striscia è tratta dalla pagina web di Ford Svizzera e risulta in buona parte illusoria. Infatti, troviamo molteplici formulazioni sintattiche che, con una piccola analisi, risultano essere ben studiate per portare l'utente a cliccare su "Consenti". Frasi e termini come: "I *cookies* sono fondamentali", "quasi tutti ne fanno uso", "per ottenere il meglio", o ancora "non c'è da preoccuparsi".
La terza e ultima immagine è presa invece dal sito de La Gazzetta dello Sport ed è la richiesta più corretta nei confronti dell'utente. Fornisce infatti sia le informazioni "belle" sia quelle "brutte"; per esempio, confessa di utilizzare questo sistema anche per inviare all'utente pubblicità mirate.

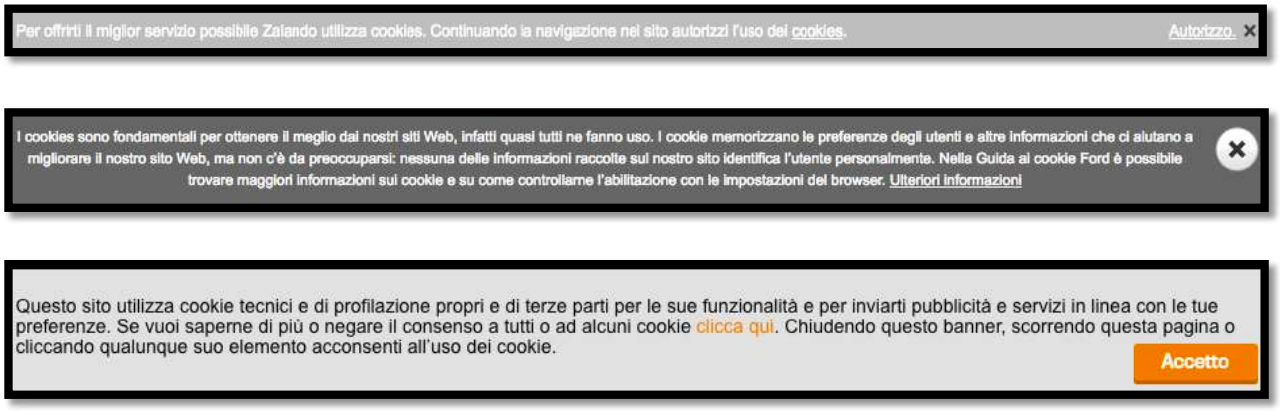


Figura 8 – Tre differenti richieste di accettazione cookies *da parte di tre* provider *internet diversi: Zalando Svizzera, Ford Svizzera e La Gazzetta dello Sport.*

A meno che non si sappia a priori cosa siano i cookies, le formulazioni presentate in Figura 8 e la definizione di "marcatore digitale" risultano essere abbastanza diverse. Se i siti internet usassero

Figura 8: tre strisce di una porzione di tre pagine web scattate da me: https://www.zalando.ch/herren-home/, http://www.it.ford.ch/Auto e http://www.gazzetta.it/ (7 agosto 2017, 11:05, 11:06 e 11:09).

esattamente la corretta definizione di *markers* nel formulare la richiesta, essa suonerebbe troppo invasiva. Come visto, per Zalando Svizzera e Ford Svizzera viene formulata così da mettere in luce unicamente gli aspetti positivi di questa azione: "migliorare l'esperienza di navigazione sulla nostra piattaforma online".

Spesso non si pensa nemmeno a cosa possano essere questi *cookies* e si clicca su una delle due opzioni senza riflettere troppo. Però, se si leggessero le clausole di accettazione dei marcatori digitali o semplicemente se ci si informasse su cosa effettivamente sono, scopriremmo che in questa maniera i siti internet entrano in possesso di una miriade di dati informatici sulla nostra persona, comunicati e donati dall'utente quasi sicuramente involontariamente ma legalmente, perché, come spesso accade, non ci si è degnati di informarsi sui *cookies*, né di leggersi le clausole di utilizzo.

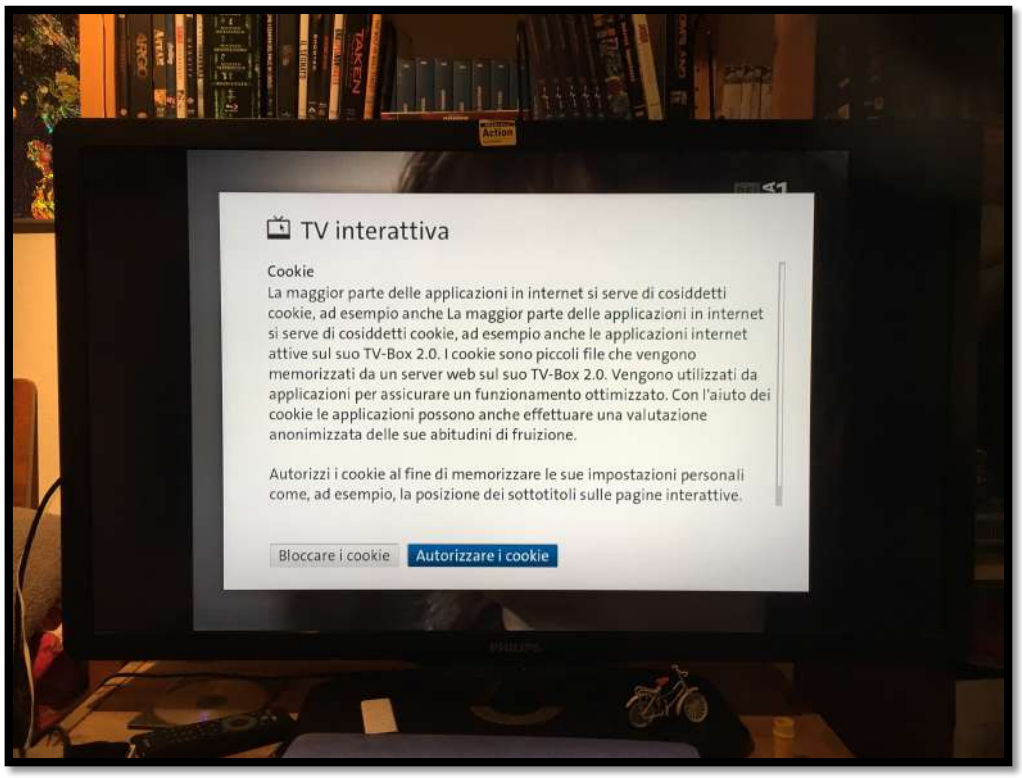


Figura 9 – Richiesta di accettazione cookies *da parte della mia televisione*

Durante la stesura di questo lavoro ero intento a guardare la televisione, quando sullo schermo è curiosamente apparso un messaggio. La comunicazione è rivolta a tutti i detentori del pacchetto di abbonamento Swisscom Tv 2.0 e chiede la nostra autorizzazione per poter abilitare l'utilizzo dei *cookies* sulla nostra televisione. Qui le ragioni sono leggermente diverse rispetto a quelle affrontate

Figura 9: immagine scattata da me allo schermo della mia televisione.

in precedenza, ma la raccolta dati è simile. In questo modo Swisscom è in grado di registrare tutte le azioni da noi compiute attraverso il servizio offerto: trasmissioni che abbiamo registrato; film che abbiamo ricercato; o semplicemente il canale visionato l'ora X del giorno Y.
Da questi dati si possono però ricavare anche informazioni più personali oltre alle preferenze televisive, che già di per sé potrebbero risultare compromettenti per alcune persone. Mi riferisco a quando la nostra televisione è accesa o spenta: non la guardiamo da tanto tempo potrebbe significare che non siamo in casa. Vi lascio immaginare quali grandi opportunità si aprono per un malintenzionato con in mano dati simili.
Poi, "Chi potrebbe mai pensare che una delle tantissime semplici televisioni collegate ad un servizio come quello offerto da Swisscom stia registrando ogni nostra mossa?". Quando parlavo di *smart devices* nel capitolo "2.3. *L'Internet of Things*" mi riferivo proprio a queste nuove dinamiche. La gente si dimostra scettica a riguardo, ma poi si ritrova a guardare la televisione con il telefono in mano e il portatile lì vicino.

La prossima volta che digitate qualche parola sulla vostra tastiera, la prossima volta che scrivete un messaggio a qualcuno o la prossima volta che interagirete con un possibile oggetto intelligente, riflettete su ciò che state facendo e su quante e su che tipo di informazioni state fornendo a persone che vi osservano e che cercano di conoscervi, ma delle quali voi ignorate o volete ignorare l'esistenza.

5.3. Occhio a cosa acquisti!

Come già detto, una volta raccolti, i dati vengono interpretati e se opportunamente collegati fra loro possono fornire un'accurata previsione sugli eventi futuri: legati al singolo utente di internet o addirittura ad un'intera società.[67] Un chiaro esempio di questa situazione ce lo fornisce il documentario della Bbc "The human face of Big Data", diretto da Sandy Smolan nel 2014. Il filmato ci illustra la vicenda del gestore di un supermercato, che ebbe l'idea di assumere Andrew Pole, un esperto nell'ambito Analysis, per osservare i generi di prodotti acquistati da donne incinte nell'arco dei 9 mesi di gestazione. Attraverso ciò, Pole, era infine in grado di stabilire solamente osservando lo scontrino della spesa quali donne fossero in attesa di un figlio e, con una certa precisione, a quale mese di gravidanza si trovassero.[68]

[67] «Corriere del Ticino», Un universo di informazioni che può svelare tutto di noi, 21 aprile 2017.

[68] dal documentario *The human face of Big Data*, 2014, Sandy Smolan, Bbc.

Analogamente è possibile determinare con diversi mesi di anticipo se un uomo divorzierà, studiando solamente ciò che acquista. Le sue spese magari un po' particolari o diverse dagli standard sono sintomo di un malfunzionamento all'interno della coppia.
Perciò certi comportamenti d'acquisto possono essere, con molte probabilità, predittivi di qualcosa che sta accadendo nel consumatore.
Possiamo ovviamente chiederci se ciò sia corretto o invada la nostra sfera privata. Infatti, lo scopo ultimo dell'amministratore del supermercato in questione era quello di spedire sconti e pubblicità mirate alle donne in attesa, così da aumentare i profitti delle vendite. La questione intimorisce, pensando a quante e a che genere di informazioni uno fra i tanti esperti in Data Science possa estrapolare da un semplice scontrino, elemento con il quale tutti noi quasi giornalmente abbiamo a che fare e del quale spesso tralasciamo l'importanza.

In un mondo come questo, capace di trovare informazioni ovunque, anche in posti dove mai ci aspetteremmo, bisogna stare attenti se non si vuole essere bersaglio centrato per esempio di queste nuove strategie commerciali. Strategie che ci toccano e ci influenzano e che vengono ad intaccare la nostra sfera privata, quella più intima. Per questo motivo diventa indispensabile equipaggiarsi delle giuste metodologie di difesa.
Da un semplice scontrino, un pezzo di carta di primo acchito privo di valore, un'attenta analisi è stata in grado di rivelare la gravidanza di più donne: l'inquietudine la fa allora da padrona pensando alla marea di informazioni che quotidianamente rilasciamo consciamente o inconsciamente nel mondo informatico.

Ciò che sono riusciti a fare alcuni ricercatori del MIT[69] è davvero sensazionale e fa emergere zone d'ombra su quello che si può effettivamente fare disponendo di una banca dati abbastanza grande.
Hanno cominciato con l'analizzare una gigantesca banca dati di transazioni effettuate con circa un milione di diverse carte di credito. A studio compiuto erano in grado, con una probabilità del 90%, di identificare in maniera univoca il possessore della carta di credito, essendo a conoscenza solamente di luogo e data di quattro transazioni effettuate. Potevano in seguito venire a conoscenza di cos'altro avesse acquistato, in quale luogo e quando.
Da tale studio si evince che per acquisire informazioni, che per certi versi possono essere definite "compromettenti", non è nemmeno più necessario visionare gli scontrini di acquisto.[70]

[69] Massachussetts Institute of Technology
È una delle più importanti università di ricerca del mondo con sede a Cambridge, nel Massachussetts (Stati Uniti), tratto dal sito https://it.wikipedia.org/wiki/Massachusetts_Institute_of_Technology (10 settembre 2017, 16:16).

5.4. Chi ha in mano questa marea di dati

I responsabili della raccolta dati sono i *big* dell'internet di oggi, i cosiddetti *Lord digitali*. Sto parlando ovviamente dei giganti come Google, Facebook, Twitter, Youtube,… . Queste aziende perseguono però i loro interessi, raccogliendo più dati possibili sugli utenti per poi analizzarli a fini commerciali, in modo da aumentare il proprio profitto.[71]

Perseguono i loro interessi senza che finora nessuno sia efficacemente riuscito a contenere la loro ingordigia di dati informatici. Infatti, tramite per esempio i termini di utilizzo di cui ho già ampiamente parlato nel capitolo "5.2.2. I famigerati *cookies*", con le loro richieste ingannevoli e da noi spesso trascurate, i nostri dati sono a completa disposizione del sito in questione. A volte capita che questi siti internet non siano del tutto onesti e neppure completamente trasparenti con i loro utenti sullo "stoccaggio dati" e, proprio per questo, ricevono multe di centinaia di milioni di dollari. Queste cifre sono da capogiro per gente comune, ma per questi colossi sono solamente un piccolo prezzo da pagare per aver contravvenuto ad una legge, che ha comunque permesso di entrare in possesso di più dati possibili.[72] Per rendervi conto dei loro guadagni basti pensare che Google nel 2010 fatturava all'incirca 250 dollari al secondo, per un fatturato annuo di circa 8 miliardi di dollari.[73]

Come visto in precedenza, le tracce digitali da noi lasciate sul web, se opportunamente analizzate e visualizzate, formano la nostra identità digitale, che permette ai colossi dello stoccaggio dati di conoscere le nostre preferenze e i nostri gusti. Spaventa il fatto che i dati informatici raccolti sulla nostra persona, una volta accettati i termini di utilizzo o semplicemente continuando ad utilizzare il servizio fornito, diventano di proprietà di queste aziende. In questo modo seguiranno pubblicità e offerte mirate dalle più disparate ditte.

Come vedrete nel capitolo "8. Leggi" manca una regolamentazione in questo ambito. Il pericolo è perciò che ne venga imposta una che sia architettata dalle lobby della digitalizzazione. Queste regole sarebbero però ovviamente a loro favore e perseguirebbero obiettivi di natura commerciale.[74]

---

[70] tratto dal sito https://www.rsi.ch/news/oltre-la-news/Prigionieri-dei-Big-Data-8434276.html (11 settembre 2017, 19:36).

[71] «Corriere del Ticino», Un universo di informazioni che può svelare tutto di noi, 21 aprile 2017.

[72] Il tema delle multe e penalità nei confronti dei giganti dell'informazione è affrontato nel capitolo "8. Leggi" e nell'intervista con l'avvocato Gianni Cattaneo.

[73] dal documentario *Internet Revolution 1*, 2011, Bbc.

[74] «Corriere del Ticino», Un universo di informazioni che può svelare tutto di noi, 21 aprile 2017.

Tutto ciò sarebbe quindi completamente legale e regolamentato dalle leggi da loro stessi create, di conseguenza diverrebbe molto difficile da fermare.

In questi ultimi anni, con l'avvento di internet e con l'affermarsi di innumerevoli colossi della digitalizzazione, il potere e la politica si stanno pian piano slegando. Coloro che posseggono più informazioni, hanno anche più potere. Mentre chi rivendica potere politico ha spesso minore influenza.
Questo è un cambiamento epocale, perché sempre nella storia, colui o coloro che detenevano il potere, possedevano di conseguenza anche potere politico.[75]

[75] cfr. BAUMAN, LYON, pag. 11.

6. Questionario

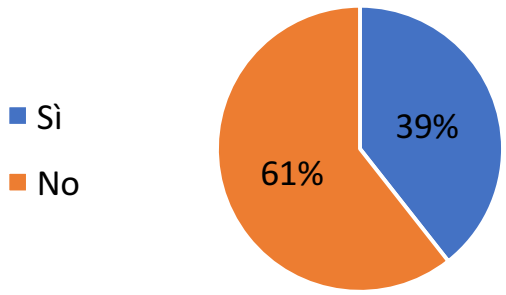
Durante la lavorazione di questo lavoro di maturità mi sono sorte molte domande e riflessioni. In modo particolare volevo confrontarmi con altre persone sui temi trattati, specialmente per quello dei Big Data. Volevo conoscere il loro punto di vista. Perciò, sfruttando gli strumenti che internet ci offre, sono entrato sulla piattaforma di Google che permette di creare dei moduli personalizzati, per sondaggi, quiz, votazioni,… .[76]

Il 1 ottobre 2017 ho quindi inaugurato il mio sondaggio, composto da 12 domande, condividendolo con più persone possibili, in modo da avere un riscontro maggiore. Come mezzo di condivisione ho impiegato la ormai onni-utilizzata *app* di messaggistica per Smartphone Whatsapp. Dopo un mese esatto, il 1 novembre 2017, ho chiuso il sondaggio con la bellezza di 297 risposte.

Qui di seguito potrete osservare i grafici ricavati dalle informazioni raccolte. Ho scelto di inserire anche un commento con un tentativo di interpretazione oggettiva per ogni grafico. Spesso ho anche intravisto la possibilità di incrociare i risultati ottenuti da una domanda con quelli di un'altra. Per cui alcune domande sono state costruite volutamente per poter ricavare informazioni che a prima vista, o osservate su due grafici distinti, non sarebbero visibili.

La maggior parte di coloro che hanno risposto al mio sondaggio sono nati tra il 1997 al 2002, sono quindi ragazzi compresi in una fascia di età tra i 14 e i 20 anni.

Le risposte vanno però prese con le dovute precauzioni. Il campione di persone che ha partecipato al sondaggio non è chiaramente significativo della popolazione ticinese. Per un sondaggio davvero significativo la raccolta dati dovrebbe essere effettuata da un istituto di demoscopia specializzato. Non è evidentemente il mio caso.

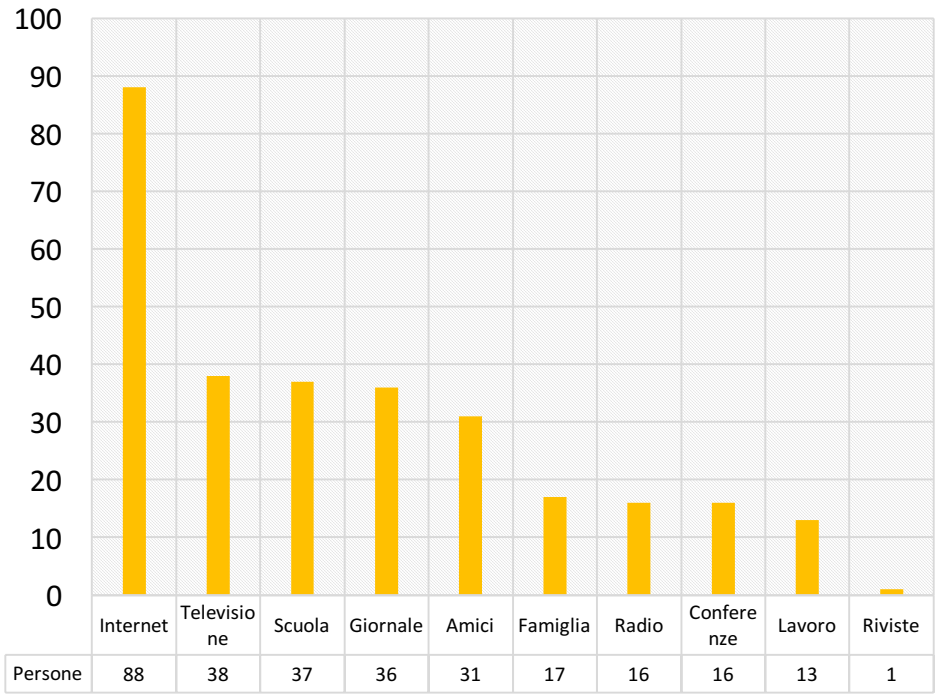
Domanda 1: **Hai già sentito parlare di Big Data?**	

Il 61% delle persone che hanno risposto non hanno mai sentito parlare di Big Data. Ciò è molto rilevante perché mostra quanta disinformazione accompagni questa tematica in esponenziale

[76] Google Forms: https://www.google.com/forms/about/

evoluzione. Non avendone mai sentito parlare, non si rendono probabilmente conto della quantità di informazioni che rilasciano in rete ogni giorno, per esempio pure svolgendo il mio sondaggio.

Domanda 2
Se sì, dove?
(risposta multipla)



	Internet	Televisione	Scuola	Giornale	Amici	Famiglia	Radio	Conferenze	Lavoro	Riviste
Persone	88	38	37	36	31	17	16	16	13	1

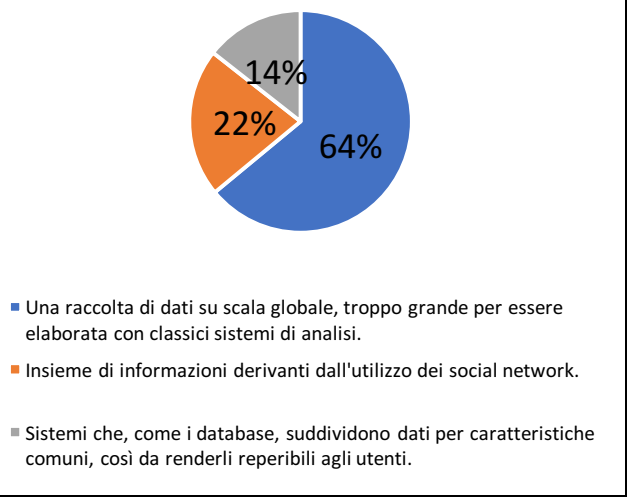
A questa domanda avremmo forse risposto nello stesso modo a priori. Infatti, essendo nati i Big Data con la rivoluzione digitale che ha accompagnato l'ultimo ventennio abbondante (*in primis* a causa dell'avvento di internet), è naturale che le tematiche a esso legate siano state discusse con maggior fervore proprio in internet. Ad aver risposto "Internet" come luogo di discussione dell'argomento sono state addirittura 88 persone tra le 116 che hanno risposto "Sì" alla prima domanda.

Al secondo posto troviamo invece con 38 persone la "Televisione", risultato che si conferma importante veicolo per la diffusione di questo tipo di informazione, anche questo, a mio avviso prevedibile. La Data Science, pur non essendo un argomento così noto e al quale la gente si interessa in modo particolare, è indubbiamente gettonato dai media. La televisione concede spazio a temi di una certa importanza – parlo soprattutto delle trasmissioni di informazione – ne consegue che anche il tema dei Big Data è abbastanza centrale nella società del giorno d'oggi.

Quasi al pari della televisione troviamo poi "Scuola", "Giornale" e "Amici", scelti da rispettivamente 37, 36 e 31 persone. La scuola appare dunque presente in queste tematiche emergenti.

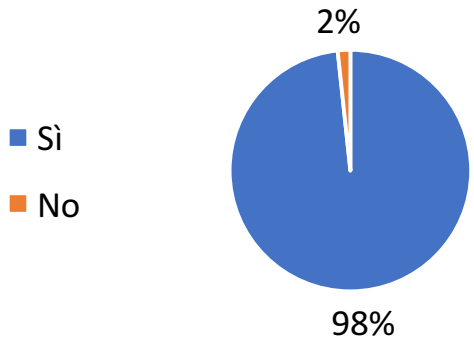
Ci tengo ad aggiungere una nota di carattere personale in proposito: il dato di 37 persone che hanno indicato la scuola è riconducibile, a mio parere, alle numerose manifestazioni di carattere multiculturale organizzate appunto dagli istituti scolastici durante l'arco dell'anno. La mia esperienza al liceo è stata accompagnata dalle giornate autogestite, che, a dipendenza delle nostre scelte, ci hanno portato a conoscere aspetti legati più ad un certo ambito rispetto che ad un altro (nel mio caso spesso tecnologico). Oltre a queste concessioni del Liceo Cantonale di Bellinzona, mi riferisco in particolar modo alla giornata annuale dedicata alla tecnologia in generale. Suppongo che molti allievi ne abbiano sentito parlare proprio in questi due ambiti.

Passando al dato sui giornali, come per la televisione sono un importante vettore di informazione, che riserva ampio spazio anche a tematiche di questo tipo. Perché attuali e interessanti, comunque, una buona parte dei consumatori.
C'è poi il passaparola, che sembra essersi classificato nella seconda fascia di consensi. Passaparola di cui fanno parte le categorie "Amici" e "Famiglia". Si intende infatti il classico "per sentito dire", ovvero il parlare e discutere tra conoscenti e famigliari degli argomenti più in voga o interessanti del momento. Questo è indice di un certo passaggio di informazioni tra i membri della società. La differenza che intercorre tra queste due classi è puramente di carattere protettivo. In famiglia possono presentarsi più frequentemente possibili ammonizioni o consigli da parte dei genitori, che possono mettere in guardia i figli sulle regole di comportamento a cui attenersi in rete. Cosa che tra amici normalmente non succede.

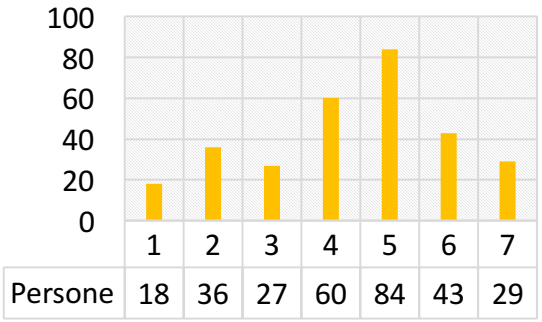
Domanda 3: **Se sì, quale definizione di Big Data ti sembra più appropriata?**	

Da questa domanda si evince che la maggior parte di coloro che hanno rivelato di aver già sentito parlare di Big Data, sanno anche scegliere una corretta definizione per Big Data, ovvero la prima, quella indicata in blu sul grafico.

A fornire due definizioni sbagliate è però stato addirittura il restante 36%. Questo fa riflettere, perché significa che questi ultimi hanno già sentito parlare di Big Data, però non hanno idea di cosa possano essere. Ciò è quini indice di una relativamente grande disinformazione sociale sul tema.

Domanda 4: **Ritieni possibile che le azioni che - direttamente o indirettamente - svolgi via internet vengano registrate?**	

Quasi la totalità dei contatti (98%) hanno risposto che sono coscienti del fatto che le proprie azioni svolte sul web vengano costantemente registrate. La cosa curiosa è che quasi la totalità delle persone conoscono la raccolta dati, ma non conoscono il tema Big Data, al quale la raccolta dati è strettamente legata. Ma allora sorge spontanea una domanda: "La gente cosa pensa venga fatto con le informazioni derivanti dalle loro navigazioni sul web?".

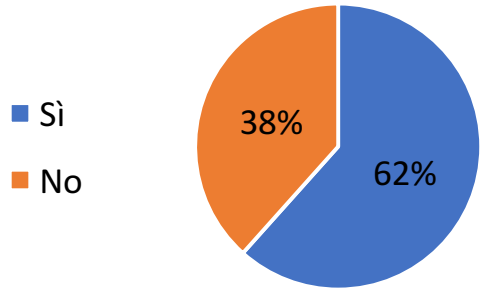
Domanda 5: **Il fatto ti spaventa?** **(risposta multipla)** **Scala: 1 (Per niente) – 7 (Moltissimo)**	

	1	2	3	4	5	6	7
Persone	18	36	27	60	84	43	29

Le informazioni che si possono estrarre da questo grafico sono molto interessanti: ben il 52% delle persone (156 su 297) ha rivelato di essere intimorita dalla raccolta dati. Questa categoria di pensiero si situa da il punto 5 al punto 7 (compresi) della scala. Quindi poco più della metà delle persone rivela di avere paura.

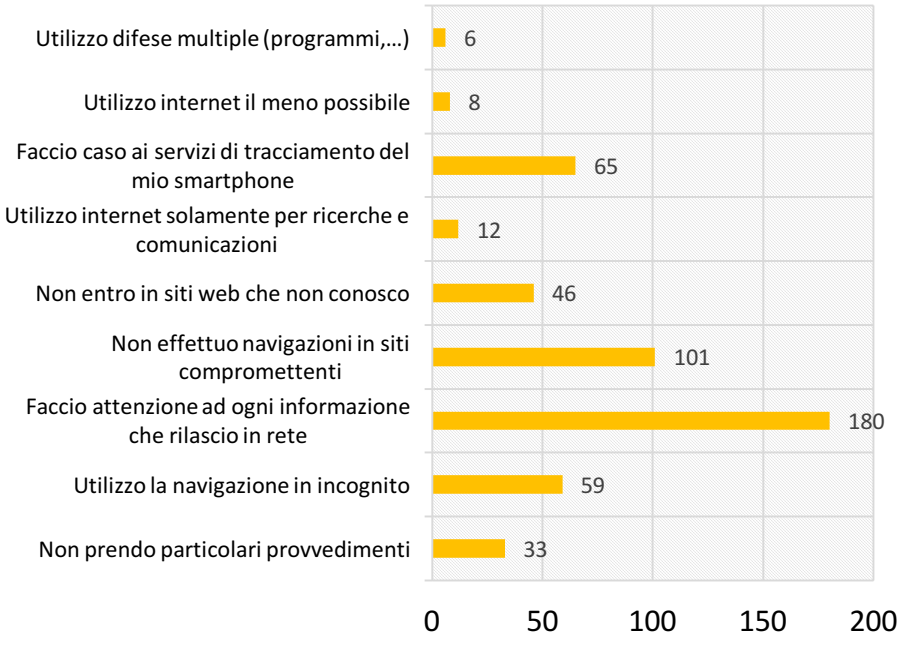
Di mentalità opposta, chi ritiene quindi di non avere paura, è solamente il 27% (81 su 297), cioè meno della metà. Tali risposte si situano tra il punto 1 e il punto 3 del grafico compresi.

Le persone che a questa domanda hanno ritenuto opportuno rispondere in modo neutro sono invece la categoria con meno consensi, ovvero il 20% (60 persone su 297).

Domanda 6: **Questo fatto modifica il tuo comportamento con il web e i nuovi dispositivi digitali?**	

Queste risposte indicano una certa presa di coscienza degli utenti di internet da me sondati. Infatti, addirittura il 62% risponde di preoccuparsi delle azioni che svolge sul web, prendendo di conseguenza accorgimenti e modificando il proprio comportamento.

È curioso incrociare queste risposte con quelle ottenute in altre domande. Praticamente la totalità delle persone, il 98% (Domanda 4), è a conoscenza dell'esistenza della raccolta dati; di queste persone cognite, solamente il 48% (Domanda 5) ha rivelato di esserne intimorito. Tuttavia, ben il 62%, quindi più di quelli che dicono di avere qualche timore, decidono di modificare il loro comportamento sul web. Questa categoria di persone non risulta perciò spaventata, però si rende conto della potenza di questa nuova dimensione e di conseguenza sceglie la prudenza.

Domanda 7: **Se sì, in che modo di tuteli?** **(risposta multipla)**	

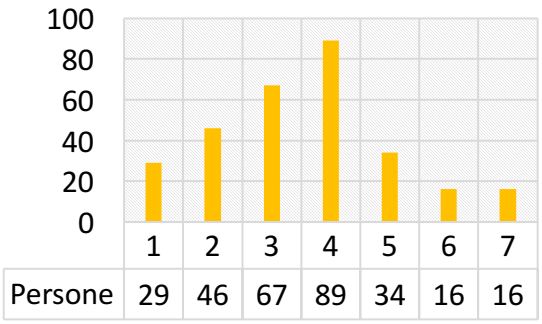
Il grafico mostra che la maggior parte della gente (180 persone) si preoccupa anche delle informazioni che rilascia in rete, "anche", perché questa domanda è a risposta multipla. Quindi la maggior parte delle persone che modificano il loro comportamento sul web a causa della raccolta

dati, tende a pensare prima di condividere qualsiasi informazione. Un'altra tutela che mi premeva evidenziare è quella relativa ai servizi di tracciamento degli smartphone, che si trova addirittura al terzo posto, scelta da ben 65 persone. Come avete avuto modo di leggere nel mio lavoro lo smartphone, con anche i servizi di localizzazione, è uno dei sistemi attraverso i quali i dati fuoriescono più frequentemente e in più grande quantità. Sono perciò "sollevato" nel sapere che c'è comunque una bella fetta di persone che si rende conto di ciò e agisce di conseguenza.

Domanda 8:

Come consideri la raccolta dati?

Scala: 1 (negativa) – 7 (positiva)



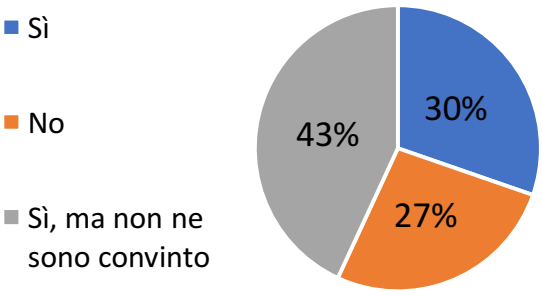
Le informazioni estraibili da questo grafico sono molto intriganti. Addirittura il 48% delle persone (142 su 297) indica di considerare la raccolta dati come qualcosa di negativo, mi riferisco alle persone che hanno risposto 1, 2 e 3.

A identificare la raccolta dati come qualcosa di positivo è solamente il 22%, cioè 66 persone su 297, mentre a restare neutro è in questo caso il 30%.

Curioso è che il grafico della Domanda 5 – di impostazione identica – presenta una forma inversa ma molto simile. È perciò naturale pensare ad una correlazione, il che vorrebbe significare che spesso chi ritiene che la raccolta dati sia una cosa negativa, ha anche paura della stessa.

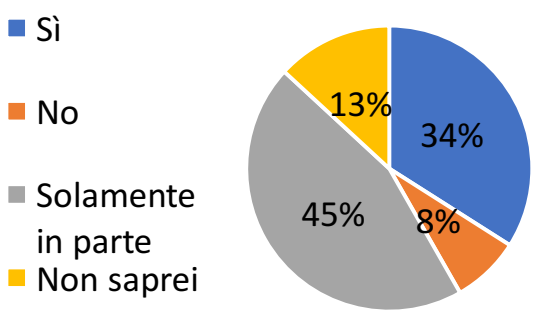
Domanda 9:

Pensi che tramite la raccolta dati sia possibile acquisire nuove conoscenze per migliorare la vita dell'Uomo?



La maggior parte afferma di essere a conoscenza del potenziale di questa nuova Scienza dei Dati, rivelando però più che una punta di sfiducia nei confronti dei detentori dei dati. Questi, risultano essere addirittura il 43% dei contatti. Se la giocano poi quasi al pari coloro che presentano una

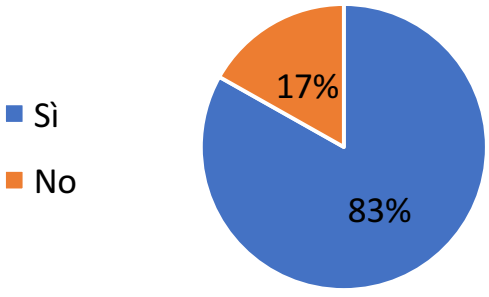
visione positiva (30%), rispetto al pensiero che la raccolta dati possa portare a nuove conoscenze atte a migliorare la vita dell'Uomo, e coloro che hanno una visione negativa (27%). Anche se togliendo la terza opzione, le scelte ricadrebbero sul "Sì", elevandolo alla ben più alta cifra di 77%.

Domanda 10: **Se fossi certo che i tuoi dati venissero raccolti per svolgere azioni positive, lo permetteresti?**	

Dal grafico si evince che, pur essendo certi che i dati concessi finirebbero per svolgere azioni positive, la maggior parte della gente non darebbe piena disponibilità delle proprie informazioni. È intuibile un certo scetticismo da parte di questa categoria di persone e dei "Non saprei" nei confronti di questa dimensione. Benché io nel porre la domanda abbia specificato il "Se fossi certo", queste provano una sorta di sfiducia.

Chiaro, concedere i dati "Solamente in parte" è già un segnale di non-diffidenza. Sommando questa percentuale di individui parzialmente aperti con i "Sì" di chi concede i dati, raggiungiamo il 79%: non è poco.

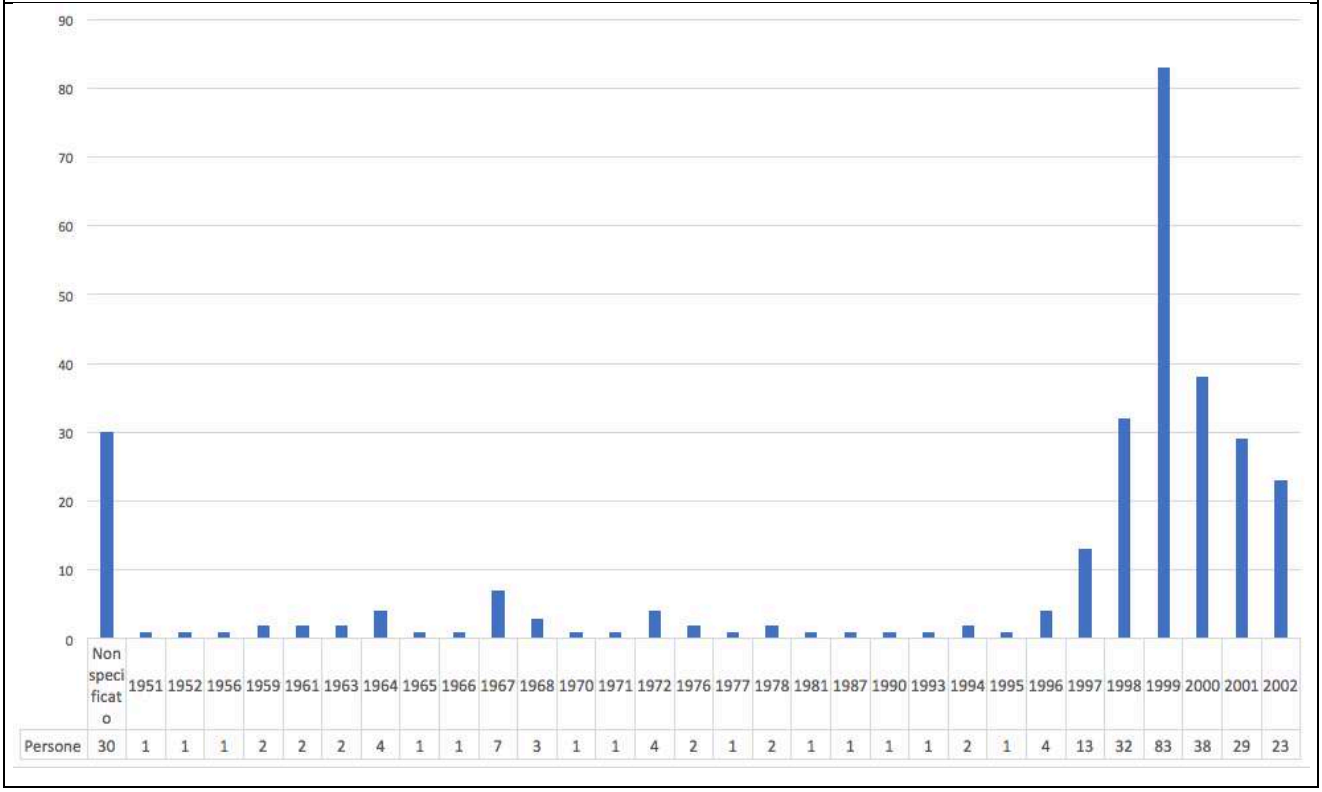
Sono veramente pochi quelli che invece, a dispetto del progresso, non concederebbero proprio nulla delle proprie informazioni personali.

Domanda 11: **Sei un "nativo digitale"? Ovvero, sei nato dopo il 1985?**	

Delle 297 persone che hanno risposto al mio sondaggio, ben l'83% sono dei nativi digitali. Questo dato è per me molto significativo per poter sondare quale fascia della popolazione è riuscita a raggiungere il mio sondaggio, nonché quale fascia della popolazione sia maggiormente a conoscenza delle tematiche da me affrontate.

Domanda 12:

Indica il tuo anno di nascita (se non vuoi puoi lasciare il campo in bianco)



	Non specificato	1951	1952	1956	1959	1961	1963	1964	1965	1966	1967	1968	1970	1971	1972	1976	1977	1978	1981	1987	1990	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Persone	30	1	1	1	2	2	2	4	1	1	7	3	1	1	4	2	1	2	1	1	1	1	2	1	4	13	32	83	38	29	23

Ero curioso di sapere l'età dei partecipanti al mio sondaggio.

Analizzando le informazioni anagrafiche, spicca il picco nella fascia 1997-2002, proprio la fascia d'età a cui appartengo. Sono nato nel 1999 e la stragrande maggioranza dei miei amici e compagni di classe è nata in questo anno: per questo motivo spiego le 83 risposte dei nati nel 1999. A scuola poi ci conosciamo quasi tutti, almeno di vista: l'amico del tuo amico è di sicuro mio amico. Quindi immagino che la presenza degli anni di nascita 1998-2002 sia influenzata dallo spirito di solidarietà tra studenti.

Ho anche qualche risposta dal 1997 e dal 2002 che riconduco alle amicizie di mia sorella e di mio fratello, nati proprio in quegli anni.

Attraverso questo grafico, non si può comunque affermare che la gioventù sia più allacciata ai nuovi mezzi di comunicazione rispetto ad altre generazioni. Semplicemente ho proposto il mio sondaggio principalmente a coetanei, che lo hanno riproposto a loro volta a coetanei, arrivando a toccare qualche adulto.

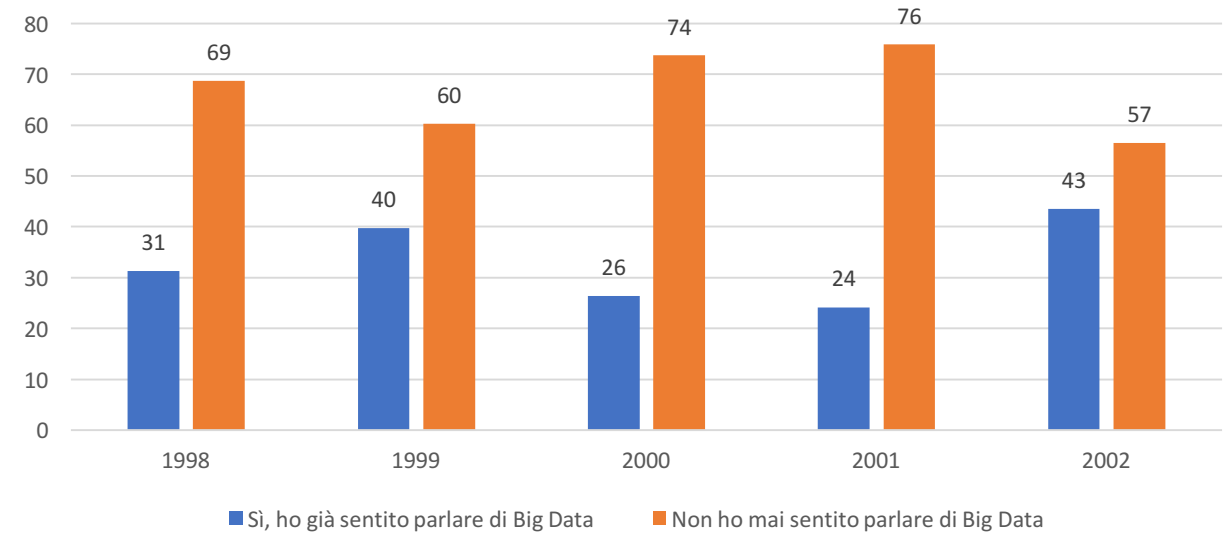
Quando stavo costruendo questa domanda pensavo: "Ma a fronte di domande legate alla paura della raccolta dati e alla dimensione dei Big Data, quante persone, dopo magari aver risposto di aver paura, riveleranno un dato relativamente sensibile come l'anno di nascita?". Mi sono stupito notando che solamente 30 persone su 297, ovvero circa il 10%, ha rifiutato di digitare il proprio anno di nascita. Strano anche dopo quanto rivelato nelle domande 5 e 8.

6.1. Incrociando i dati

I seguenti grafici sono stati costruiti incrociando le informazioni ottenute in diverse risposte. Ho utilizzato specialmente i dati ricavati dalla Domanda 12 sull'anno di nascita. Ero curioso di osservare la relazione che intercorre tra l'anno di nascita e alcune altre risposte. Osservando questi dati ho deciso di escludere dai grafici riportati qui di seguito le persone nate in un determinato anno, se unici rappresentanti della loro categoria. Il 1956 e il 1971 sono due esempi. Nei grafici riportati ho quindi inserito le annate per le quali ho ottenuto più risposte, vale a dire la fascia che va dal 1998 al 2002. Le risposte ottenute per questi anni rispecchieranno maggiormente la reale situazione.
Ho calcolato in percentuale i dati con i quali ho costruito i grafici, così da poter confrontare al meglio le risposte dei diversi anni.

Domanda 12 – Domanda 1

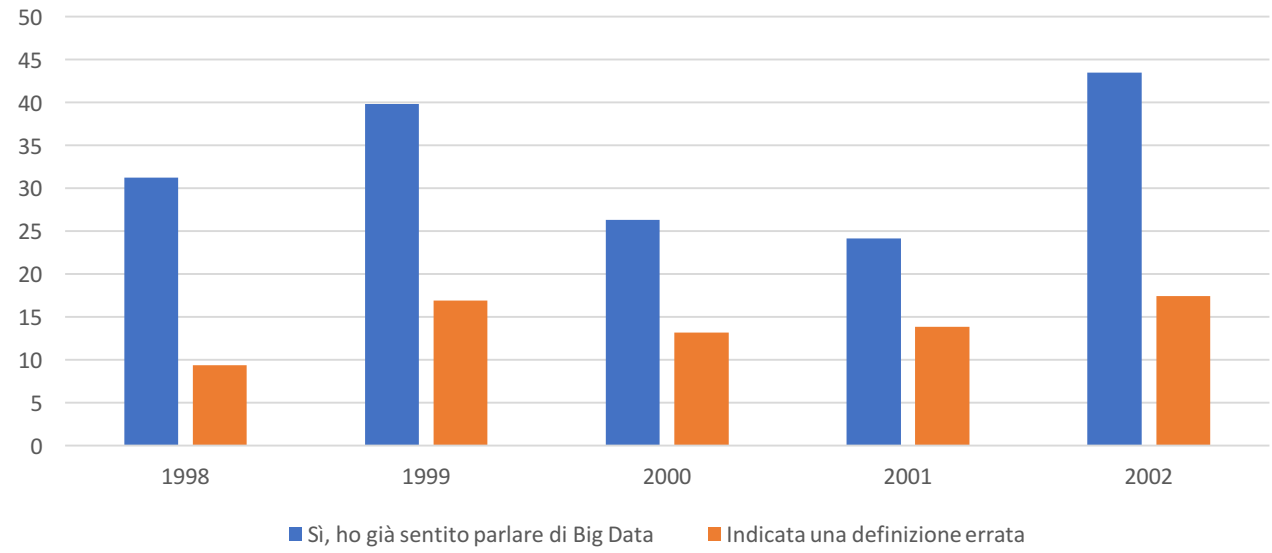
(Dati in percentuale)



Questo grafico mostra in maniera chiara come la maggior parte delle persone appartenenti a tutte le fasce d'età prese in considerazione non abbiano idea di cosa sono i Big Data. Le persone nate nel

2000 e nel 2001 sembrano avere meno dimestichezza con il tema rispetto a tutte le altre; infatti, in entrambi i casi, solamente circa un terzo delle persone sondate hanno rivelato di avere già sentito il termine Big Data.

Domanda 12 – Domanda 3 – Domanda 1
(Dati in percentuale)



Questo grafico mette in luce degli aspetti interessanti sul tema. Infatti ho intrecciato i dati di coloro che hanno già sentito parlare di Big Data in qualche ambito, con la definizione che hanno ritenuto opportuno scegliere come corretta, ma che in realtà si è rivelata errata.

Due annate, 2001 e 2002, quelle che dal grafico precedente risultavano essere le più ignoranti riguardo al tema Big Data, anche qui lo riconfermano. Infatti, la metà delle persone (un po' di più per il 2001) hanno riferito di aver già avuto in qualche modo a che fare con i Big Data, dandone però poi una definizione errata.

L'anno che si piazza meglio è il 1998, nel quale solamente 9 persone su 100 affermano di avere già sentito il termine Big Data, definendoli poi in modo scorretto. Le altre dicono di conoscere e danno l'esatta definizione.

7. Giusto o sbagliato

«Nessuno è più schiavo di colui che si considera libero senza esserlo.»

Goethe (1749 - 1832)

Poeta, scrittore e drammaturgo tedesco.

7.1. L'etica nel nuovo mondo dei dati

Come in molti ambiti della vita, anche esprimere un verdetto definitivo positivo o negativo sulla raccolta dati Big Data è molto complicato, se non addirittura impossibile. A causa della sua vasta portata, non si può in effetti affermare che questo sistema sia solamente bianco o nero, giusto o sbagliato. Dobbiamo, invece, prendere atto della sua duplice natura, positiva e negativa, per poi farci un'idea soggettiva sulla loro nascita e il loro progressivo sviluppo.

Attraverso la conoscenza sia dei rischi, sia delle opportunità, derivanti da questa nuova scienza, la visione dell'opinione pubblica mondiale si è spaccata in tre. Una parte vuole che i dati non vengano raccolti (e a dire la verità inizialmente anche io stavo schierato con loro, ma lavorando a questo Lam e trovando una miriade di informazioni interessanti, che ho provato ad inserire seguendo un filo logico, mi sono ricreduto). Infatti la disinformazione sociale in questo campo e soprattutto l'ignoranza sul tema, sono molto diffuse. Molti di noi quando pensano ai Big Data (già sapere cosa siano è tanto), pensano a una specie di Grande Fratello, che ci spia costantemente, prevendendo ogni nostra mossa e conoscendoci molto intimamente.

La seconda parte dell'opinione pubblica mondiale desidera invece l'opposto: ovvero che i dati vengano stoccati (ora non ci focalizziamo sul fatto che i dati possano essere stoccati in modo trasparente o meno, ma sul fatto che la gente acconsenta che vengano raccolti). Io mi ritrovo a pensarla come queste persone per un semplice motivo. Come visto nel capitolo "Big Data e opportunità: la statistica e le scienze", questa immensa mole di dati porta a benefici e contribuisce allo sviluppo dell'intera società umana. Tocca solamente a noi – ma mi rendo conto che non è poca cosa – decidere di fare la cosa giusta. Perciò non raccogliere tutti questi dati non sarebbe ragionevole: abbiamo la possibilità di migliorare nettamente le condizioni di vita di tutti e di operare per il bene dell'umanità sulle ali della conoscenza che passo dopo passo ci ha, in fondo, fatti scendere dagli alberi e poi portati fuori dalle caverne, dove abbiamo iniziato la nostra avventura di esseri umani.

Poi c'è la terza, e forse più ampia, parte dei fruitori della rete che non sanno, non si informano, non si pongono il problema e non capiscono di essere – nel peggiore dei casi – "manipolati" da chi raccoglie i dati personali, i suoi gusti, idee, tendenze, a scopo commerciale.

Quello che è certo – e che pone non meno interrogativi di quanto già esposto – è che siamo diventati sia produttori sia fruitori dei dati, mentre la produzione è generalizzata, l'accesso ai dati rimane purtroppo spesso selettivo.[77]

7.2. Perché acconsenti o non acconsenti alla raccolta dei tuoi dati

Specialmente per i nativi digitali, diventa sempre più difficile ricordare un numero di telefono perché ormai è lo smartphone a farlo oppure scattare esclusivamente quella foto perché significativa oppure ancora recarsi personalmente in biblioteca per quel libro che tratta proprio ciò che ci interessa, invece di averlo con un clic.

La società digitale – la nostra società – è oramai abituata a vivere nel lusso delle comodità, grazie a tutta una vasta gamma di servizi offerti dalle nuove tecnologie e, in maniera sempre più marcata, dagli *smart devices*.

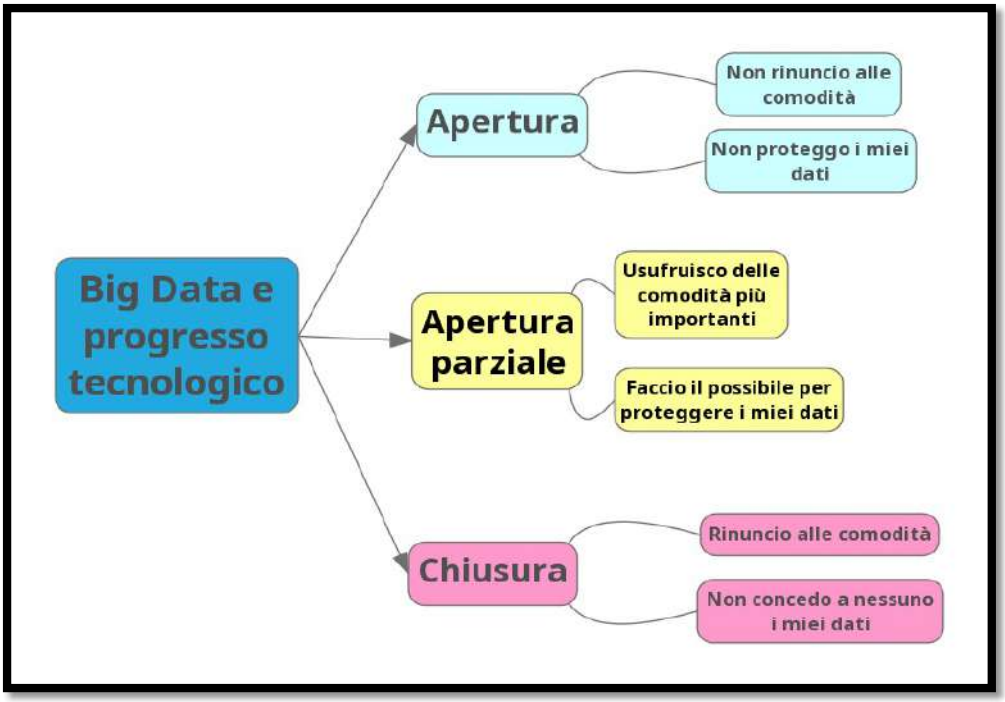
"Tutti siamo contenti di avere l'iPhone o il Samsung […]" dice Luigi Curini, rispondendo ad una domanda. Però, possedere questi nuovi e sempre più performanti apparecchi digitali, è in prima istanza riconducibile proprio alla raccolta dati. Grazie a questa è infatti possibile venire a conoscenza delle preferenze dei consumatori, così da lanciare sul mercato prodotti sempre più rispondenti alle loro esigenze e, più importante, alle loro comodità.

Per cui, dobbiamo accettare che questi apparecchi raccolgano dati sulla nostra persona e che, poi, essi, vengano analizzati. In questo modo le comodità con cui – negli ultimi anni, giorno dopo giorno – si è sempre più abituati a convivere continuano ad essere presenti e ad essere costantemente aggiornate. La domanda da porci è una soltanto: "Vogliamo vivere nel mondo delle comodità di cui ormai facciamo parte, imparando a convivere con quella che è la nuova dimensione della raccolta dati? O vogliamo limitare questa nuova perentoria evoluzione, ponendo regolamentazioni (un po' troppo) restrittive?". Rispondere a questo quesito è complesso e tocca ognuno nella propria sfera più personale. È il singolo individuo che, valutando le proprie esperienze e il proprio stile di vita, è portato a prendere una posizione di fronte a questa nuova dimensione.

Esistono, a mio parere, tre diverse mentalità, comprovate dalla professoressa Antonietta Mira durante l'intervista. Queste sono: **apertura**, che implica il completo assoggettamento da parte dell'individuo a un certo aspetto di questa nuova natura; **apertura parziale**, limitata quindi da alcuni paletti; e da ultimo **chiusura**, che, sotto forma di una sorta di luddismo digitale, conduce la persona ad un rifiuto categorico nei confronti di una parte o di tutta questa dimensione. Nella

[77] cfr. MIRA, pag. 9.

seguente *mappa mentale* avete modo di osservare generalmente queste tre categorie e le sostanziali differenze che le caratterizzano.



Schema 2 – Schema che illustra le tre mentalità che si possono assumere nei confronti dei Big Data e della raccolta dati.

In tutti e tre i casi risulta comunque migliore avere la certezza che i propri dati, accaparrati da enti governativi o privati, siano trattati in maniera trasparente e con nobili scopi. Con "nobili scopi" intendo, da parte di questi istituti, impegnarsi per raggiungere traguardi atti a perseguire esclusivamente i nostri bisogni e che non possano nuocere in alcun modo alla nostra persona. Capite però che ne risulta un controsenso. Porto l'esempio di Google: questo colosso fornisce a tutti i membri della società una miriade di servizi che semplificano di gran carriera molti aspetti della quotidianità. L'obiettivo è però quello di guadagnare e Google nel 2016 ha avuto un fatturato annuo di 8 miliardi di dollari. Quindi: o converte i suoi servizi a pagamento, o rivende i nostri dati di navigazione a terzi, lucrandoci sopra. Anche nell'era digitale nulla è gratuito, e se lo è, bisogna chiedersi come mai.

Ne consegue il bisogno di informare e sensibilizzare la società su queste due facce della stessa medaglia. "[…] Non c'è minimamente coscienza del fatto che siamo ormai arrivati a un punto in cui le informazioni disponibili su ciascuno di noi possono essere utilizzate in modo proficuo." afferma Luigi Curini, nell'intervista. Risulta perciò necessaria una presa di coscienza collettiva, di una

portata tale da obbligare a sostanziali cambiamenti nelle leggi. Cambiamenti che stanno già avvenendo e che approfondisco al punto "8. Leggi".

7.3. Standardizzazione dell'individuo

Le nostre scelte non sono più libere perché la raccolta dati e la conseguente analisi pilotano le pubblicità e gli acquisti verso le preferenze che, secondo la nostra identità digitale, dovremmo avere. In questo mondo nuovo non siamo più noi a scegliere, bensì i dati che noi stessi abbiamo creato e fornito all'universo digitale. Abbiamo tutta la conoscenza a portata di mano – o di clic – ma rischiamo di non essere più capaci a sfruttarla ed utilizzarla.

Questi meccanismi rischiano di appiattire il comportamento dell'utente e di guidarlo sempre più verso quello che è un comportamento medio. Vedrà, comprerà e farà cose che si accordano in modo perfetto con le sue preferenze. Questo rischia di offuscare la fantasia e di non permettere più di vedere oltre quello che già piace e che continuerà a piacere sempre più.[78] Siamo esseri umani e come tali abbiamo la propensione a ricercare conferme di quello in cui crediamo, per rassicurarci quanto alla bontà e alla correttezza dei nostri ideali.

Con queste nuove dinamiche appare evidente il rischio di restare prigionieri sotto valanghe di informazioni che non portano nulla in fatto di conoscenza. Eppure gli algoritmi continuano a riproporci quello che dallo studio della nostra identità digitale emerge possa piacerci, o, cosa anche peggiore, quello che loro ritengono sia di nostro gradimento.[79]

Magari ci accorgiamo di tali meccanismi quando visitiamo un sito commerciale, come Amazon o Ebay, ma non bisogna pensare che tali meccanismi siano esclusivi dei siti commerciali perché magari, per esperienza, avete già notato su questi siti l'esistenza di gallerie di "prodotti-che-ti-potrebbero-interessare". Questa funzione è adottata anche da Google quando svolgiamo una ricerca in rete. Come scrive Roberto Casati, filosofo italiano e direttore del *Centre National de la Recherche Scientifique*: "Google analizza le tue ricerche passate per costruire un modello del tuo io online cui offrire dei risultati di cui anticipa che saranno pertinenti per te. […] ci nasconde una fetta di realtà, rimandandoci di continuo l'immagine delle nostre preferenze."[80]. Oltre a Google, anche molti siti di informazione implementano questa funzione nella loro interfaccia, presentandoci "notizie-che-ci-potrebbero-interessare" in gallerie costruite appositamente ma – perché no? – anche come notizie più importanti.

---

[78] tratto dal sito http://www.rsi.ch/news/oltre-la-news/Prigionieri-dei-Big-Data-8434276.html (13 settembre 2017, 16:09).

[79] cfr. CASATI, pag. 117.

[80] *ibidem*, pag. 118.

L'unico elemento che può essere considerato positivo dentro questo marasma di consigli e algoritmi è che il nostro dispendio di tempo su internet si riduce di molto, anche se, ad essere sinceri, ciò comporta un incalcolabile impoverimento degli orizzonti a cui possiamo ambire.[81]

7.4. Automatizzazione dei processi

Al giorno d'oggi l'analisi dei dati raccolti sui singoli utenti intrecciata a potenti algoritmi e – perché no? – con un po' di intelligenza artificiale, rende possibile l'attuazione di processi automatizzati di accettazione o rifiuto di richieste di servizi. Il mediatore umano è viepiù considerato obsoleto.

Se dalla mia identità digitale rientro nella categoria di persone con scarsa disponibilità economica, e che probabilmente, sempre secondo l'operato di intelligenza artificiale e degli algoritmi, avrà difficoltà nel restituire un prestito, esso potrebbe essermi negato. Si valutano ad esempio i tempi per saldare le fatture o se abito in una zona economicamente forte o depressa. In questi casi l'intervento umano può rivelarsi determinante: la conoscenza diretta della persona, se questa è insolvente o ha un potenziale, potrebbe modificare la scelta.

Negli ultimi anni stiamo contribuendo allo svilupparsi di un nuovo tipo di società, basata sull'impiego della tecnologia nei più diversi settori. Contemporaneamente stiamo però combattendo una battaglia contro la stessa digitalizzazione. Mi spiego meglio: si pretende che tutto sia più *smart* – *smart city,* smartphone –, perché come già visto rende spesso tutto più comodo, ma al contempo lottiamo perché la digitalizzazione non prenda il posto delle persone. Tutti noi abbiamo notato la sostituzione delle tradizionali cassiere con casse automatiche, o il rimpiazzamento dei caselli o dei *check-in* all'aeroporto con delle macchine. Questo rende tutto più veloce, ma anche più anonimo e artificiale: chissà che un giorno in alcuni servizi sarà un costoso lusso avere un contatto umano. Questo è un vero stravolgimento dell'attuale sistema, sempre definito fino ad oggi dal rapporto uomo-uomo.

Davanti a tale perplessità ho compreso di non poter rispondere semplicemente con una riflessione personale. Ho perciò girato la questione a degli esperti che ho intervistato per sapere cosa credono possa accadere in un prossimo futuro, ormai dietro l'angolo.

Antonietta Mira identifica la seguente come sostanziale differenza tra uomo e macchina: "La nostra mente è dotata di fantasia e questa è una caratteristica bellissima dell'essere umano che un algoritmo non sarà mai in grado di prevedere." [82]. La macchina resta quindi pur sempre una

[81] cfr. CASATI, pag. 120.

[82] Intervista a Antonietta Mira

macchina. Qualcosa che è meccanico per definizione, costituito da ingranaggi che non posseggono vita, neppure se combinati tra loro.

Luigi Curini si pone invece così dinnanzi a questo pensiero: "L'idea che si possa fare a meno dell'intervento umano […] è completamente farlocca. […] La migliore tecnologia, come ci ricorda sempre Gary King[83] è "*Computer-assisted and human empowerment*"[84]. […] fare a meno dell'interpretazione umana è un'idea che non è nemmeno affascinante e che, in questo momento, non esiste.". Secondo il suo pensiero è semplicemente impensabile che le macchine in generale possano in alcun modo ergersi all'altezza dell'uomo. La sola convivenza dei due è in grado di portare a dei risultati, senza la parte umana tutto risulterebbe infattibile.

7.5. Categorizzazione delle masse

Attraverso la scienza dei dati ci si è resi conto di poter suddividere le persone secondo le informazioni che producono. La massa viene quindi divisa in classi di persone con caratteristiche comuni, seguendo alcune determinate tracce fornite dalla nostra identità digitale.

Ci si è subito accorti della comodità di tale sistema. Con la sola lettura delle nostre informazioni, veniamo costantemente inseriti in corti di persone che presentano profili sempre più simili ai nostri. In questo modo ci vengono forniti servizi che non sono più **personalizzati**, ovvero costruiti su misura per ognuno, ma di **precisione**, che sono pensati partendo dalla categorizzazione delle persone.[85]

Questo epocale passaggio è avvenuto in sordina. La persona interessata da queste dinamiche, oserei quindi dire tutti i membri di una società, non si è inizialmente accorta del cambiamento. Solo oggi cominciamo a rendercene conto: un esempio lampante di questa nuova realtà sono le pubblicità mirate. "Avete mai cercato uno specifico prodotto su un motore di ricerca come Google e qualche tempo dopo, controllando la mail, visitando un sito o un social network, vi siete ritrovati esattamente una pubblicità di quel prodotto?".

Rispetto al passato tutto è cambiato, ma ciò non appare, perché ciò che intercettiamo con i nostri sensi sulla superficie resta quasi sempre uguale. Prima dell'avvento dei Big Data ricevevamo offerte pubblicitarie secondo i dati che avevamo rilasciato, per esempio in un concorso di una società, o essendoci iscritti ad uno specifico servizio. Conseguentemente ricevevamo opuscoli con offerte dei loro prodotti pubblicizzati ma non obbligatoriamente necessitavamo di acquistarli. Oggi

[83] esperto di Scienze Politiche statunitense

[84] dall'inglese "*Computer-assisted and human empowerment*", "Assistenza dei computer e sviluppo umano".

[85] Intervista a Antonietta Mira

riceviamo invece offerte da aziende che non conosciamo nemmeno, che rispecchiano però molto spesso esattamente le nostre attuali esigenze.

Questa categorizzazione degli individui, secondo gusti, bisogni, stato di salute e un'altra infinità di aspetti, torna molto utile a tutti gli istituti e le aziende che offrono prodotti e servizi. Infatti, ora possono limitarsi ad inviare pubblicità solamente ai destinatari – parzialmente o totalmente – interessati, invece che a tutti o molti, ma senza alcun criterio, con la speranza di guadagno. Tali compagnie ottimizzano i profitti direzionando offerte mirate esclusivamente a possibili acquirenti, senza sprechi di risorse. In conclusione otterranno un riscontro maggiore effettuando meno promozione.

"Sorge il rischio di venir *intrappolati* in una bolla di informazioni che, invece di generare conoscenza, non fa altro che presentarci in modo suadente quello che si pensa che potrebbe piacerci. Non la realtà, ma l'immagine che vorremmo di essa. O, peggio, l'immagine che un robot informatico ha creato per noi a scopi pubblicitari." [86] scrive Roberto Casati nel suo libro "Contro il colonialismo digitale".

Questa nuova metodologia sarà resa sempre migliore dall'evoluzione compiuta dal costante sviluppo tecnologico. Come avete potuto modo di vedere nel punto "3.3. Da dove provengono", lo sviluppo tecnologico implica una maggiore diffusione degli *smart devices* – o della tecnologia in generale – che sarà fonte di un numero crescente di informazioni su di noi. Ciò porterà questi sistemi ad essere di giorno in giorno perennemente più precisi, grazie alla possibilità di caratterizzare maggiormente la società secondo sempre più svariate informazioni.

[86] cfr. CASATI, pag. 117.

8. Leggi

8.1. Internet e regole

Internet è un'infrastruttura che opera al di fuori della giurisdizione di qualsiasi Paese ed è strutturata in modo da rendere quasi impossibile qualsiasi tentativo di controllo centralizzato. Sottolineo il "quasi" perché, anche se un controllo totalmente centralizzato risulta impossibile da attuare dai governi, i motori di ricerca più utilizzati e le case produttrici dei prodotti informatici più all'avanguardia impongono una sorta di controllo tramite la raccolta dati. Tutto ciò, una volta che l'utente viene informato su tali dinamiche dietro le quinte, fa paura. Perché egli utilizza internet per i più svariati scopi: lavoro, transazioni di ogni tipo o *entertainment*. Infatti, il consumatore medio, che non possiede conoscenze specifiche, non sa tutelarsi in alcun modo da questo boom tecnologico in continua evoluzione e continuo perfezionamento. Questo *deficit* informativo solleva molteplici interrogativi quanto alla necessità di una regolamentazione condivisa che vada a tutela dei consumatori.

A complicare maggiormente le cose è il fatto che ogni Paese possiede sue leggi che spesso si scontrano con quelle degli altri. È proprio per questo motivo che la Commissione Europea, istituzione esecutiva dell'Unione Europea (UE), ha intravisto la necessità di introdurre una legislazione più robusta riguardante la protezione dei dati dei cittadini europei. Sia all'interno, sia all'esterno del continente, dimodoché il fattore "ogni-paese-ha-le-proprie-leggi" possa essere in qualche modo superato.

Riguardo alla Svizzera, una legislazione di questo genere esiste già e porta il nome di *Legge Federale sulla protezione dei dati*, perciò non siamo legalmente scoperti da questi reati.

8.2. La *Legge federale sulla protezione dei dati*

Riprendo l'ultima parte dell'introduzione di questo capitolo: "non siamo legalmente scoperti da questi reati", ed è vero, ma non possiamo nemmeno definirci adeguatamente coperti. Ciò perché questa legge, che dovrebbe essere a tutela dei dati dell'individuo, risale al lontano 1992.[87] Se si pensa che internet e la tecnologia degli *smart devices* sono in continuo aggiornamento, questa legge – incentrata su principi generali – è da considerarsi decisamente obsoleta, perché, a differenza di ciò che deve controllare, non è stata aggiornata dal momento del suo concepimento. Nel periodo in cui

[87] Legge federale sulla protezione dei dati

questa legge è stata emanata, per esempio, Google non esisteva ancora[88] e con ciò il tema sempre più presente dei Big Data e della Data Science. Come è possibile difendersi da qualcosa di nuovo, continuando ad utilizzare armi vecchie?

Passato il primo ventennio dell'avvento planetario di internet, è ormai necessario disporre di leggi e regolamentazioni che possano essere considerate al passo coi tempi e possibilmente capaci anche di guardare al futuro. Queste regole eviterebbero, per quanto possibile, tentativi di sorveglianza-controllo mirata da parte di entità governative (o peggio ancora private) che, con il loro operato, vanno contro ogni principio fondamentale e ideale di rispetto della sfera privata. È pure auspicabile un inasprimento delle sanzioni. Purtroppo, a violare tali regole sono spesso e volentieri i *big*, i quali incorrono poi in sanzioni economiche – da considerarsi ridicole rispetto al loro fatturato – o ci si limita a far cadere qualche testa. Provvedimenti che non potranno mai cambiare le cose.

Bertil Cottier, professore dell'USI esperto di legislazione di internet, in un'intervista ha asserito che: "[In questo ambito la Legge Federale sulla protezione dei dati] *è* indietro non di una ma di due generazioni. […] è come se le regole della circolazione stradale, in ambito legale fossero riassunte in tre articoli, di assoluto buonsenso e assolutamente condivisi, ma totalmente generici: guidate con prudenza, date la precedenza e tenete accese le luci di notte." [89].

"A mio avviso la legislazione in vigore, pur essendo "generica", ossia basata su principi generali, è una buona regolamentazione di base poiché si adatta alle nuove tecnologie. Naturalmente, presuppone uno sforzo di interpretazione che a volte può sfociare in una certa insicurezza giuridica.", questo pensa Gianni Cattaneo, avvocato specializzato in diritto di internet e professore di diritto informatico alla SUPSI, sulla legislazione attualmente in vigore in Svizzera. Esiste ed è fondata a suo modo di vedere su principi ben saldi. L'unica pecca che presenta, però di grande spessore, è quella della messa in pratica della stessa dinnanzi a reati concreti. Risulta sempre necessario doverla interpretare diversamente da caso a caso. Chiaramente, ciò non suscita chiarezza e confonde tutti coloro che ad essa si appellano.

Il giudice è sempre portato ad esprimersi soggettivamente in merito ad un reato a processo. Ma lo fa restando più oggettivo possibile, restando conforme a quanto scritto nel codice. Il sistema giudiziario è costruito appositamente in questo modo. Così facendo si evitano grossi problemi, dovuti a scelte discutibili da parte di giudici, che, sul finire, sono persone come noi. La nostra legislazione in merito alla protezione dei dati è scritta secondo principi giusti, ha richiesto sicuramente numerosi interventi dagli angoli più remoti per essere completata. Eppure mancano le

[88] […] il *suo* dominio è stato registrato il 15 settembre 1997, tratto dal sito https://it.wikipedia.org/wiki/Google (29 agosto 2017, 20:51).

[89] «Corriere del Ticino», Un universo di informazioni che può svelare tutto di noi, 21 aprile 2017.

istruzioni per poterla mettere in pratica. Cosa che, come visto prima, è essenziale, per non spingere un giudice azzardato a giudizi troppo soggettivi di fronte a reati molto complessi.

8.3. La *General Data Protection Regulation*

È quindi chiaro che l'utilizzo dei Big Data ha importanti implicazioni nella *privacy*, nella protezione dei dati e nella correlata salvaguardia dei diritti della persona. Il 4 maggio del 2016 è stata pubblicata sulla Gazzetta Ufficiale Europea la *General Data Protection Regulation* (GDPR), che va proprio in questa direzione.

Questa regolamentazione, il cui nome italiano è *Regolamento Generale sulla protezione dei dati*, entrerà in vigore nella sua totalità il 25 maggio del 2018 e dovrebbe introdurre sostanziali cambiamenti alle legislazioni di tutti i paesi europei nell'ambito della protezione dei dati.[90] Sul suo sito web, la SUPSI definisce questa innovazione legislativa come "un passaggio epocale in materia di protezione dei dati" [91], poiché verranno introdotte regole molto più restrittive rispetto a quelle vigenti tuttora.

La sua pubblicazione risale però, come già detto, al 24 maggio del 2016. Ci si può domandare perché tanta attesa? La risposta è semplice: si è voluto lasciare del tempo a tutte le aziende e le pubbliche amministrazioni per organizzarsi e mettersi in regola in vista della data scelta per la sua messa in vigore. Tutte le istituzioni e realtà operanti su suolo europeo aventi a che fare anche solo in minima parte con la raccolta di dati dei consumatori, hanno dovuto conformarsi alle nuove normative, per non farsi trovare impreparate e risultare perciò fuorilegge e perseguibili.

Le innovazioni di questo trattato non sono poche e non vanno sottovalutate.

In primo luogo le pubbliche amministrazioni e ogni azienda avente a che fare con dati sensibili, dovranno assicurare al loro interno la presenza di una nuova figura professionale, addetta alla corretta amministrazione delle suddette informazioni. Verrà perciò introdotto il *Data Protection Officer* (DPO).[92]

Ogni qualvolta che una nuova figura professionale rilevante si è affacciata alle porte delle nostre comunità, ha segnato l'avvio di una nuova "rivoluzione", questa volta tecnologica che ci porta a subire pesanti cambiamenti. Vediamo qualche esempio ripercorrendo la storia. Penso all'artigiano nel periodo classico; o all'artigiano superiore – così come lo definisce il professore in lingua e

[90] tratto dal sito https://it.wikipedia.org/wiki/Regolamento_generale_sulla_protezione_dei_dati (11 ottobre 2017, 17:42).

[91] tratto dal sito http://www.supsi.ch/home/comunica/eventi/2017/2017-10-24.html (11 ottobre 2018, 17:49).

[92] *ibidem*

letteratura araba all'Università di Roma Mario Casari nel suo libro "Amore dei Luoghi" – per il Medioevo, figura che era una sorta di quasi meccanico, o al periodo della cibernetica con l'informatico e, appunto, all'attuale periodo con il nostro nuovo *Officer*.[93] Periodo che, forse, potrà essere superato di qui a pochi anni, per dare spazio a qualcosa di ancora più innovativo, che avrà a che fare con il valore intrinseco dei dati e della conoscenza.

L'importanza della correttezza nella raccolta dati è messa in evidenza in un particolare articolo della GDPR (Articolo 5, Paragrafo 1, Lettera A). Questo afferma che i dati personali devono essere "trattati in modo equo, legale e trasparente in relazione all'oggetto interessato".
Quello che avviene oggi, come abbiamo visto, è quasi sempre proprio il contrario: le grandi analisi dei dati vengono talvolta additate come minaccia per la privacy o addirittura viste come sinistre macchinazioni[94]. Questo perché l'analisi spesso comporta la reinterpretazione dei dati in modi inaspettati, utilizzando algoritmi complessi e traendo conclusioni su persone con effetti imprevisti e talvolta sgraditi.[95]

Questa nuova regolamentazione, però, non sarà e non dovrà essere intesa come una "barriera" allo sviluppo del mondo della Data Science. "Non sarà Big Data o protezione dei dati, oppure Big Data contro protezione dei dati." scrive nel suo trattato "*Big data, artificial intelligence, machine learning and data protection*" l'Information Commissioner's Office (ICO), organismo pubblico che fa capo direttamente al Parlamento inglese. Continua dicendo che, in poche parole, la GDPR permetterà che i Big Data vengano impiegati per fare tutto il bene di cui sono capaci.[96]

Per rimostrarvi l'attualità dei temi che sto affrontando, vi presento una mail arrivatami il 16 ottobre 2017, mentre scrivevo queste righe. Il suo contenuto è prova dell'imminente e decisivo cambiamento che la GDPR attuerà anche in Svizzera. Nella Figura 10, Medtronic, un'azienda svizzera attiva nel settore farmaceutico-sanitario, invia una mail di avvertenza ai propri utenti con l'intento di avvisarli sui cambiamenti che subirà la piattaforma web della compagnia per conformarsi alla nuova regolamentazione europea.

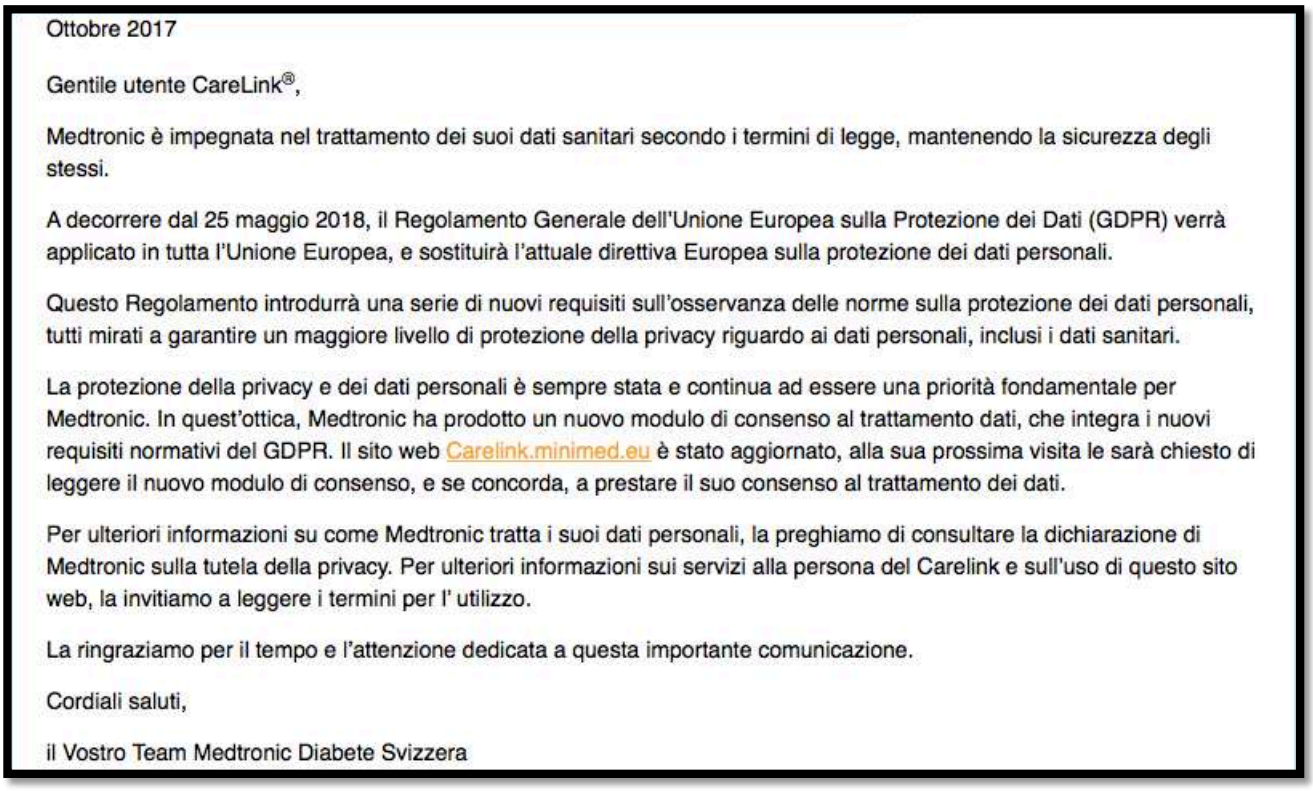
---

[93] CASARI Mario, *Amore dei luoghi,* Aracne, 2013, pag. 46-54.

[94] "Infausto, sfavorevole, avverso (per il prevalere, nelle antiche tradizioni popolari, della credenza che gli auspici provenienti da sinistra fossero di cattivo augurio) […] bieco, torvo, minaccioso […].".
Tratto da http://www.treccani.it/vocabolario/sinistro/ (11 ottobre 2018, 19:02).

[95] Information Commissioner's Office (ICO), *Big data, artificial intelligence, machine learning and data protection*, England, 2014, pag. 19.

[96] *ibidem*, pag. 3.



Ottobre 2017

Gentile utente CareLink®,

Medtronic è impegnata nel trattamento dei suoi dati sanitari secondo i termini di legge, mantenendo la sicurezza degli stessi.

A decorrere dal 25 maggio 2018, il Regolamento Generale dell'Unione Europea sulla Protezione dei Dati (GDPR) verrà applicato in tutta l'Unione Europea, e sostituirà l'attuale direttiva Europea sulla protezione dei dati personali.

Questo Regolamento introdurrà una serie di nuovi requisiti sull'osservanza delle norme sulla protezione dei dati personali, tutti mirati a garantire un maggiore livello di protezione della privacy riguardo ai dati personali, inclusi i dati sanitari.

La protezione della privacy e dei dati personali è sempre stata e continua ad essere una priorità fondamentale per Medtronic. In quest'ottica, Medtronic ha prodotto un nuovo modulo di consenso al trattamento dati, che integra i nuovi requisiti normativi del GDPR. Il sito web Carelink.minimed.eu è stato aggiornato, alla sua prossima visita le sarà chiesto di leggere il nuovo modulo di consenso, e se concorda, a prestare il suo consenso al trattamento dei dati.

Per ulteriori informazioni su come Medtronic tratta i suoi dati personali, la preghiamo di consultare la dichiarazione di Medtronic sulla tutela della privacy. Per ulteriori informazioni sui servizi alla persona del Carelink e sull'uso di questo sito web, la invitiamo a leggere i termini per l' utilizzo.

La ringraziamo per il tempo e l'attenzione dedicata a questa importante comunicazione.

Cordiali saluti,

il Vostro Team Medtronic Diabete Svizzera

Figura 10 – Le compagnie svizzere aventi a che fare con dati sui fruitori del proprio servizio, in questo caso Medtronic, azienda sanitaria internazionale, sono obbligate a confrontarsi con le nuove regolamentazioni decise dalla GDPR.

Nella mail viene sottolineato come la GDPR porterà cambiamenti nel trattamento delle informazioni personali, con l'obiettivo di garantire una maggiore sicurezza alla tutela della privacy dei dati personali.

[97] Figura 10: foto di una mail da me ricevuta relativa alla GDPR.

9. Conclusioni

In questo viaggio nell'affascinante universo dei Big Data ho capito molte cose. E soprattutto ho cambiato la mia opinione su questo nuovo mondo.

Sono partito dall'idea che la conoscenza delle persone sui Big Data fosse estremamente ridotta e che ci fosse una certa ignoranza riguardo l'uso e l'abuso dei dati personali in rete. Dopo le ricerche da me effettuate e sulla base di una miriade di informazioni trovate, posso affermare che questa dimensione, pur essendo – come avete letto – parte integrante della nostra vita, è spesso una presenza di cui non siamo consci. In altre parole: la maggior parte di noi non ha coscienza della sua esistenza, anche se parecchie delle cose che facciamo tutti i giorni ne sono un'emanazione. Questa mia visione è scaturita, oltre che dalle informazioni raccolte, anche dall'intervista a Luigi Curini e pure dal mio sondaggio. Navighiamo ogni ora, ogni giorno nella rete, lasciamo in ogni momento tracce in un brodo di informazioni che qualcuno scruta, raccoglie, analizza, e usa non sempre per il bene collettivo. Molti camminano nella rete bendati, senza vedere chi ci osserva e per quale fine. Siamo poco tutelati dalle leggi che sono in perenne ritardo su una tecnologia che cambia ad un ritmo vertiginoso. Solo la conoscenza dei meccanismi di questi piccoli grandi fratelli può dare a ciascuno la capacità di scegliere in autonomia e uscire dalla paura. Ognuno deve dunque imparare a chiedersi ogni volta che qualcosa è gratuito perché lo è e che cosa sta regalando in cambio. Ci vuole una maggiore coscienza e responsabilità di tutti, finché le norme non saranno adeguate ai rischi.

La diffusa ignoranza relega quindi i Big Data e la Data Science in un oscuro angolo di mistero, dove mille e mille occhi guardano con diffidenza. I media hanno contribuito ad alimentare questa relazione, mostrando spesso soltanto la parte negativa, cioè dando particolare risalto ai grossi scandali (esempio il caso Snowden), meno alle positività. Hanno così spinto milioni di persone a sviluppare un'accezione negativa. Per me il concetto chiave in definitiva e in sintesi è questo: "La gente teme questa dimensione perché non la conosce", come del resto accade con tutto ciò che non conosciamo e di cui abbiamo sentito parlare solo negativamente. Ho potuto riscontrare questa paura anche nel mio sondaggio. Non mi si fraintenda, ritengo sia comunque corretto avere timore della grande novità, ma bisogna altresì imparare a conoscere la nuova scienza nella sua vastità, così da poterla anche apprezzare e utilizzare nella sua parte buona. L'idea che mi sono fatto è che esiste un Grande Fratello, o meglio – come afferma Antonietta Mira – che esistono tanti piccoli grandi fratelli. Il pericolo è perciò reale, ma limitato.

Questi grandi fratelli, ovvero i detentori dei dati, sono aziende governate frequentemente da logiche commerciali, con ultimo scopo quello del profitto. Trafficare quantità di dati di tale dimensione porta inevitabilmente questi privati a guadagni immensi. Un dato che mi aveva fatto rizzare i capelli era quello che Google nel 2010, tramite i suoi numerosi servizi "gratuiti", guadagnava 250 dollari al

secondo, per un fatturato annuo netto di 8 miliardi di dollari.[98] L'astronomico guadagno dietro i Big Data – una delle ipotesi da verificare – non ha fatto altro che confermarsi e riconfermarsi ogni volta che mi ritrovavo sottomano dati relativi alle grandi compagnie raccoglitrici.

L'ultimo punto da me affrontato è quello della tutela legale a livello svizzero e prossimamente a livello europeo. Quello che mi sento di dire, malauguratamente, è quello che temevo e che speravo di non dover confermare. Lo scudo legale, fornito dalla Confederazione con la Legge Federale sulla protezione dei dati, non è adeguato per proteggere in modo soddisfacente l'individuo, che si ritrova solo dinnanzi a questa nuova immensa dimensione L'insufficiente tutela legale è sottolineata dall'esperto di diritto Bertil Cottier, il cui pensiero è stato esposto nel capitolo "8.2. La Legge federale sulla protezione dei dati".

Per ovviare a questa mancanza, la Confederazione ha deciso di sottoscrivere un accordo con la maggior parte degli stati europei, nella speranza che la tutela legale possa risultare migliore. Nel maggio del 2018 verrà così introdotta la Global Data Protection Regulation. "Cambierà le cose?" mi sono chiesto non appena conosciuta la novità. La risposta che più mi ha convinto è quella del professore Luigi Curini: "Permetterà sì una maggiore difesa della privacy, ma per chi vuole ci sono una quantità enorme di modi per girarci intorno. […] Non la vedo fattibile, la vedo su carta.". Non aiuta inoltre l'enorme abisso presente tra i tempi della politica e quelli della tecnologia: le norme sono in ritardo di diverse generazioni.

A fronte di questa problematica, ritengo che la disponibilità dei nostri dati non debba essere circoscritta ai grandi colossi digitali, i cui fini – come abbiamo visto – sono orientati al guadagno. Diversamente, le analisi dovrebbero essere maggiormente condotte da istituti pubblici, indirizzati ad altri fini, quali università o istituti facenti capo direttamente al Governo.

In questa ricerca per me, più importante di ogni informazione appresa, è stato il capovolgimento della mia opinione sul tema "Big Data" e "Data Science". Ho avviato il lavoro nella convinzione che fossimo costantemente spiati/sorvegliati da entità governative e private. Non conoscevo assolutamente nulla delle possibilità fornite da questa moderna scienza e degli orizzonti che essa apre. Veramente poche persone si rendono conto che molto di ciò che fanno è già concretamente parte di questa dimensione, di cui dicono di diffidare. Io ero tra loro. Inoltre in futuro sarà sempre più difficile tenersi fuori da questa società digitale, dove tante azioni quotidiane (pagare le fatture, fare la spesa, prendere l'aereo,…) vedono prevalere l'uso della tecnologia. Pian piano dovremo tutti adeguarci e chi disporrà di maggiori informazioni potrà farlo nel migliore dei modi.

[98] dal documentario *Internet Revolution 2*, 2011, Bbc.

Grazie alla possibilità di approfondire offertami dal Lam, ribadisco in queste ultime righe che la mia visione è totalmente cambiata: da “I Big Data sono esclusivamente una cosa negativa” a “I Big Data non sono esclusivamente una cosa negativa”. Più lentamente, chissà che la mia posizione non muti di nuovo, fino, forse, a farmi affermare che “I Big Data sono una cosa positiva”.

Grazie per la lettura
Jacopo Caratti

«Non esistono reali scoperte né reale progresso finché sulla terra esiste un bambino infelice.»
Albert Einstein (1879 – 1955)
Premio Nobel per la fisica e filosofo tedesco, naturalizzato svizzero e americano

10. Fonti

10.1. Bibliografia

BAUMAN Zygmunt, LYON David, *Sesto potere: La sorveglianza nella modernità liquida*, Cambridge, Laterza, 2013.

CASARI Mario, *Amore dei luoghi*, Aracne, 2013.

CASATI Roberto, *Contro il colonialismo digitale*, Laterza, 2013.

CASTELLS Manuel, *Galassia Internet*, Oxford, Feltrinelli, 2001.

10.2. Articoli su riviste e giornali

Cartaceo:

«Altroconsumo», *Alla salute di Big Data*, maggio 2016.

«Altroconsumo», *Che potere, i dati*, novembre 2017.

«Corriere del Ticino», *Un universo di informazioni che può svelare tutto di noi*, 21 aprile 2017.

«Diritto dell'informazione e dell'informatica», *Big Data: i rischi della concentrazione del potere informativo digitale e gli strumenti di controllo*, Anno XXVIII, Fascicolo 1, 2012.

«Focus», *Quando i dati mentono*, maggio del 2016.

«20 Minuti», *Truffe con carte di credito, dal Ticino parte la guerra*, 25 ottobre 2016.

«La Regione Ticino», *Il dato più grande è la privacy*, 27 giugno 2017.

«La Regione Ticino», *Sore, Google batte la ritirata*, 1 luglio 2017.

Online:

«Il Post», *Perché gli scandali si chiamano "-gate"*, 17 giugno 2012.
http://www.ilpost.it/2012/06/17/perche-tutti-gli-scandali-si-chiamano-gate/

«La Regione Ticino», *Utenti internet, anche gli editori svizzeri raccolgono dati*, 14 agosto 2017.
https://www.laregione.ch/articolo/utenti-internet-anche-gli-editori-svizzeri-raccolgono-dati/48805

«Tages Anzeiger», *Google schaut weg*, 26 giugno 2017.
http://www.tagesanzeiger.ch/digital/daten/google-schaut-weg/story/17743660

«Tio», *Urbanscope: un "macroscopio urbano" per sentire la città che pulsa*, 6 maggio 2016.
http://www.tio.ch/News/Partner/USI/1084152/Urbanscope-un--macroscopio-urbano--per-sentire-la-citta-che-pulsa/

«Wired», *Ecco dove finiscono i tuoi dati*, 24 febbraio 2015.
https://www.wired.it/internet/web/2015/02/24/ecco-dove-finiscono-i-tuoi-dati/

«Wired», *Putting a Dollar Value on Big Data Insights*, 2017.
https://www.wired.com/insights/2013/07/putting-a-dollar-value-on-big-data-insights/

10.3. Audiografia

Offline:

«Rete tre (RSI)», *Baobab*, 23 novembre 2016.

Online:

«Rete uno (RSI)», *C'era una volta… oggi – Un po' di Uri e altri dati*, 13 dicembre 2016.
http://www.rsi.ch/rete-uno/programmi/intrattenimento/c-era-una-volta-oggi/Un-poi-di-Uri-e-altri-dati-8453209.html

«Rete uno (RSI)», *Millevoci – Raccontiamo i numeri*, 6 ottobre 2015.
http://www.rsi.ch/rete-uno/programmi/intrattenimento/millevoci/Raccontiamo-i-numeri-6144028.html

10.4. Filmografia

Citizenfour, 2014, *Laura Poitras*

The human face of big data, 2014, Sandy Smolan, Bbc

Internet Revolution 1, 2011, Bbc

Internet Revolution 2, 2011, Bbc

Snowden, 2016, Oliver Stone

10.5. Videografia

Cosa sono i Big Data? – Antonietta Mira, Fondazione Fiera Milano, 2017.
https://www.youtube.com/watch?v=kItlFRDF5XU&t=1s

Nuovi paradigmi per ricerca e business – Antonietta Mira, Fondazione Fiera Milano, 2017.
https://www.youtube.com/watch?v=abZJdXFgZ0k

Pericoli dei Big Data e impatto economico e sociale – Antonietta Mira, Fondazione Fiera Milano, 2017.
https://www.youtube.com/watch?v=RVRrmSKA4G8

Prigionieri dei Big Data – RSI, 2016.
https://www.rsi.ch/news/oltre-la-news/Prigionieri-dei-Big-Data-8434276.html

Il progetto Urbanscope – Antonietta Mira, Fondazione Fiera Milano, 2017.
https://www.youtube.com/watch?v=PwB2C01lTqw

The human face of Big Data – Rick Smolan TED, 2014.
https://www.youtube.com/watch?v=h8FLkRK-wF4

10.6. Sitografia

Amazon (sito di compravendita): http://www.amazon.it

Cardiocentro Ticino (sito della clinica Cardiocentro Ticino): http://www.cardiocentro.org

CCM (uno dei più grandi siti di tecnologia francese): http://it.ccm.net

Cisco (fornitore di apparati di networking): http://gblogs.cisco.com

Corriere della Sera (sito dell'omonimo quotidiano di informazione) http://www.corriere.it

Fondazione Ticino Cuore (sito della Fondazione Ticino Cuore): http://www.ticinocuore.ch

Ford (sito della marca automobilistica): http://www.it.ford.ch

Gapminder (sito dell'omonima fondazione che promuove lo sviluppo globale sostenibile e offre numerose statistiche): http://www.gapminder.org

Github (sito di condivisione di software): http://jiffyclub.github.io

Google (motore di ricerca): http://www.google.com

Il Post (quotidiano online italiano): http://www.ilpost.it

Infodiritto (sito di Gianni Cattaneo dove ci introduce al Diritto svizzero dell'informatica e di internet): http://www.infodiritto.net

La Gazzetta dello Sport (sito dell'omonimo giornale di informazione sportiva):
http://www.gazzetta.it

La Stampa (sito dell'omonimo giornale): http://www.lastampa.it

Medium (piattaforma di pubblicazione): https://www.medium.com

Microfocus (azienda globale produttrice di software): https://www.microfocus.com

Morrisjfwong (blog cinese): http://www.morrsjfwong.com

PredPol (sito che offre sistemi di prevenzione di crimini): http://www.predpol.com

Quora (spazio di condivisione di domande): http://www.quora.com

Repubblica (sito dell'omonimo quotidiano italiano): http://www.repubblica.it

RSI (sito dell'emittente radiotelevisiva ticinese): http://www.rsi.ch

SUPSI (sito della Scuola universitaria della Svizzera italiana): http://www.supsi.ch

Tio (sito di informazione principalmente a livello ticinese e svizzero): http://www.tio.ch

Treccani (enciclopedia online): http://www.treccani.it

Viseca (sito dell'azienda svizzera di servizi finanziari): http://www.viseca.ch

Wabe (sito dell'omonima stazione radiofonica statunitense): http://news.wabe.org

Wearesocial (istituto che analizza le conversazioni sui social): http://www.wearesocial.com/it

Wikipedia (enciclopedia libera online): http://it.wikipedia.org

Zalando (sito di vendita di capi d'abbigliamento): http://www.zalando.ch

10.7. Pubblicazioni

CATTANEO Gianni, *Il campo di applicazione territoriale del Regolamento generale sulla protezione dei dati dell'UE (GDPR) nell'ottica delle aziende che operano in o dalla Svizzera*, 2016.

CERON Andrea, CURINI Luigi, IACUS Stefano M., *iHappy 2016*, Corriere della Sera, Milano.

Information Commissioner's Office (ICO), *Big data, artificial intelligence, machine learning and data protection*, England, 2014.

MIRA Antonietta, *La Scienza dei Dati: una Nuova Sfida Multidisciplinare*, Rendiconti della Classe di Scienze Morali: Scienze Economiche e Statistiche, SECS-S/01 – Statistica.

PALMER Kimberly, *News & World* Report, 2007.

SPARROW Betsy, LIU Jenny, WEGNER Daniel M., *Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips*, 2011.

10.8. Testi di legge

Legge federale sulla protezione dei dati

Guida ai provvedimenti tecnici e organizzativi concernenti la protezione dei dati

Guida per il trattamento di dati personali nel settore privato

11. Allegati

11.1. Intervista a Antonietta Mira

Antonietta Mira è professoressa di Statistica all'Università della Svizzera Italiana (USI) e cofondatrice e ora direttrice dell'Istituto Interdisciplinare di Data Scienze dell'USI (IDIDS), che si occupa proprio di analizzare grandi quantità di dati.

I Big Data sono una realtà recente. Quale nuovo universo di conoscenze aprono?

È un universo che spazia dalle città intelligenti, ad algoritmi che ti suggeriscono comportamenti: cosa leggere, cosa mangiare, in che ristorante andare, dove andare a dormire. Ovviamente con vantaggi e svantaggi, perché in questo modo si appiattiscono i comportamenti delle persone verso il comportamento medio della gente. Così perché tali algoritmi imparano a suggerire attingendo dai gusti di altre persone a cui sono piaciuti gli stessi film o gli stessi libri che sono piaciuti a te. Il rischio è che continui a vedere film che risuonano con le tue preferenze. Capiterà magari che un film completamente diverso – che forse ti sarebbe piaciuto tanto – non ti verrà mai suggerito.
C'è per esempio questo mio collega di Harvard – con cui ero in videochiamata un attimo fa – che sta facendo delle *app* usando tutte le potenzialità di uno smartphone, per monitorare, ad esempio, pazienti con problemi mentali. In particolare pazienti bipolari: euforici e poi depressi. Questi pazienti prendono farmaci e periodicamente vedono il loro medico. Ogni tanto lo vedono quando stanno bene, ma magari quando stanno male no. Ecco, sul telefono ci sarà un'applicazione in cui si monitorano in modo passivo quello che la persona fa. Per esempio, il medico vede quanto la persona si sposta con il conta passi, a che velocità cammina con l'accelerometro, se la persona risponde o meno alle telefonate o ai messaggi, quindi quanto è attiva socialmente, se al mattino continua spegnere la sveglia oppure si alza subito attraverso un sensore della luce. Questi sono tutti dati passivi, cioè la persona non deve fare niente, semplicemente si porta in giro il telefono, cosa che comunque normalmente fa.
Poi ogni tanto, qualche volta alla settimana, le viene chiesto di registrare un certo messaggio vocale. E di nuovo, dal timbro della voce e da come pronuncia la frase, ci sono algoritmi che capiscono se è la voce di una persona euforica o depressa. Questa è l'unica cosa attiva che viene richiesta di fare. Chiaro è che, se una persona non lo fa, questo è indicativo del fatto che probabilmente ha qualche problema.
Questo è un esempio di quello che si chiama *digital-phenotyping*, cioè capire il comportamento di una persona attraverso gli strumenti digitali. I dati vengono analizzati automaticamente e ci sono

modi per capire e comunicare al malato "Attenzione sei in una fase depressiva, vai dal medico!". Quindi la persona andrà dal medico non a caso una volta ogni tre mesi, ma quando effettivamente ne ha bisogno.

Naturalmente il malato deve acconsentire a concedere i dati. Però in questo caso, se io fossi un paziente bipolare, darei i miei dati, poiché ne trarrei un beneficio personale. Perché, ovviamente, questi algoritmi sono modellati sui miei dati, ma anche sui dati di tutti pazienti bipolari. È vero che ognuno ha il suo comportamento tipico, però ci sono classi di comportamento che sono tipiche di chi è depresso e di chi è euforico.

Esiste un algoritmo anche per pazienti con tumori al cervello, per segnalare quanto è elevato lo stato di sofferenza. Qui è pazzesco, L'applicazione chiede ai pazienti con diagnosi di tumore al cervello di segnalare quanto stanno male su una scala da 0 a 10, ogni giorno. Poi a un certo punto c'è l'intervento, dopo il quale vengono seguiti per un certo periodo.

La soglia del dolore di questi pazienti è praticamente uguale. È abbastanza alta in fase diagnostica, poi prima dell'intervento cresce e allora cominciano dare farmaci, perciò comincia a diminuire sempre di più. Più avanti, quando però tolgono i farmaci, ricomincia salire, stabilizzandosi tipicamente a un livello più basso.

È incredibile vedere come queste curve abbiano più o meno tutte lo stesso andamento. Prima questi pazienti venivano monitorati solamente per la settimana in cui erano in ospedale, poi, a casa, nessuno veniva più ascoltato. Nessuno aveva il tempo e la pazienza di monitorare lo stato di dolore prima e dopo l'intervento mentre con questa app ciò è possibile. Questa è ancora medicina di precisione.

Altri ambiti sono poi la medicina di precisione e quella personalizzata. Quest'ultima cerca di trovare la terapia giusta per te. In un discorso Obama ha detto: "Quello che vogliamo fare è trovare la cura giusta, per la persona giusta, al momento giusto.". Ma personalizzare a questo livello è molto difficile e sono necessari investimenti enormi. Perciò, si passa dalla medicina personalizzata a quella di precisione: non si fa la medicina per me, ma io – come soggetto – appartengo a una certa categoria di persone che hanno fattori di rischio simili ai miei. Per esempio io non fumo, non bevo, ho una certa età e sono donna. Quindi si studia la corte di persone che ha caratteristiche simili alla mia, non solo comportamentali ma anche genetiche, quindi socio-comportamentali. O magari anche a quali fattori di rischio ambientali sono esposto: esempio se vivo in una certa città piuttosto che in campagna. Quindi si cerca di creare delle corti di persone che hanno caratteristiche simili rispetto a dei fattori che sono ritenuti determinanti per la prognosi o per l'efficacia di una certa cura. Poi si studia come essa funziona o non funziona in questa categoria di persone e, in conclusione, a me la si dà o non la si dà, a seconda che abbia funzionato o meno in tale gruppo di persone. Non è quindi

una cura personalizzata, ma una cura che viene studiata su una classe di persone omogenee rispetto a certe caratteristiche.

Negli Stati Uniti sono partiti con questo progetto della *precision medicin*[99], per il quale stanno seguendo diverse migliaia di americani, divisi in gruppi omogenei rispetto a queste varie caratteristiche. Vengono seguiti nel tempo. Questi non sono necessariamente malati oggi, però di queste persone si conoscono il patrimonio genetico, le abitudini. Se in seguito dovessero ammalarsi si studiano certe cure, gli effetti. Poi se una terza persona ha questa stessa patologia, si vede a quale gruppo assomiglia di più; quindi la si assimila a quel gruppo e si fanno su di lei le cure risultate efficaci.

Quanto impiega un sistema del genere ad essere pronto?

Dipende da cosa si vuole studiare. Ci sono alcune patologie di cui già si conosce molto, perciò lo stato di ricerca è già molto avanzato, quindi c'è un aiuto in più. Però, per esempio, su patologie rare o di cui si conosce poco del rapporto che c'è tra patologia e comportamento della persona, o fattori ambientali o esposizione a certe criticità, i tempi sono abbastanza lunghi.

L'idea è comunque che queste persone siano seguite nell'arco del tempo.

"Giusto impiego" cosa significa in questo ambito? Ci sono "regole d'arte" o codici etici?

Ecco, questo è un problema, perché in alcuni paesi si stanno sviluppando codici per legiferare sui Big Data, sia dal punto di vista giuridico, sia dal punto di vista etico. Ma in questo momento quanto a legislazione e codici etici non c'è coordinamento fra i diversi paesi. Perciò non esiste un riferimento a livello internazionale, che addirittura potrebbe essere mondiale. Quindi non sono coordinati e sono molto indietro rispetto alle necessità. È come se il codice della strada, che conta circa un centinaio di articoli, venisse ridotto a tre di buon senso.[100] In questo momento la legislazione sui Big Data è di questo tipo. Ci sono dei principi molto generali e assolutamente condivisi però non sono declinati per renderli pratici e utili.

E la General Data Protection Regulation porterà dei cambiamenti?

Settimana scorsa ho sentito parlare un giurista coinvolto in un progetto della Comunità europea basato sugli aspetti legislativi dei Big Data. Ci sono tantissime problematiche irrisolte. Innanzitutto c'è la visione che fra qualche anno circa una decina d'anni, ognuno di noi avrà un robottino che sarà la nostra interfaccia con il mondo esterno e a cui daremo ordini: "dimmi che tempo fa", "accendi il

[99] medicina di precisione

[100] citando Bertil Cottier: «Corriere del Ticino», *Un universo di informazioni che può svelare tutto di noi*, 21 aprile 2017.

riscaldamento", "fai partire la lavastoviglie", "accendi il fuoco". Però il robottino comincerà a conoscerci, perché studiando mimica facciale e timbro vocale saprà leggere le nostre emozioni. Quindi per esempio saprà vedere quando io passo davanti a un negozio e vedo un nuovo prodotto, la reazione che ho di curiosità. E probabilmente sarà in grado di dire se lo comprerò o no. E sarà talmente informato dai dati che noi produciamo che ci conoscerà probabilmente meglio della persona che ci sta a fianco. Questo perché la persona che ci sta a fianco non riesce a decifrare tutte le nostre inflessioni della voce o non coglie certe sfumature.

La domanda è: a chi appartengono questi dati? Perché questo robottino conterrà tutta una serie di dati su di noi. Saranno nostri, perché noi abbiamo comprato questo robottino e lo abbiamo istruito, permettendogli di stare nostro fianco e di leggere il nostro comportamento? O saranno del robottino e quindi di chi ce l'ha venduto ha ideato l'algoritmo per leggere espressioni facciali, voce, eccetera? E chi potrà usare questi dati? Certamente ci sarà un contratto con chi ci vende il robottino, che non sarà come vendere un'aspirapolvere. Quindi: di chi saranno tutti questi dati e chi vi potrà accedere, siccome diranno molto di noi? Insomma, cominciamo ad andare sul pericoloso.

Quindi nel futuro prossimo ci saranno cambiamenti fondamentali?

Sì, stiamo veramente vivendo una rivoluzione. Ci troviamo proprio nel mezzo di una rivoluzione digitale, forse la quarta rivoluzione. Vivere su questa cresta dell'onda è da una parte affascinante ma, dall'altra, proprio perché ci siamo dentro, non abbiamo la visione di come tutto sta cambiando e ci sentiamo anche un po' spaesati.

Come ci si accorge quando i risultati di un'analisi sono sbagliati o non del tutto corretti? Come nel caso della previsione delle epidemie, che può essere influenzata in modo decisivo dalle ricerche effettuate sul web dall'utente allarmato.

Purtroppo ce ne accorgiamo soltanto *ex post*. Diventa difficile accorgersene *ex ante*. Facciamo un esempio: i dati su Google Flu Trend (GFT), un sito sul quale si cercava di prevedere le epidemie di influenza. Come faceva Google? Andava a vedere tutte le ricerche online di termini legati all'influenza che le persone facevano: tipo aspirina, tipo raffreddore, tipo farmacia, eccetera. Aveva scelto una quarantina di termini legati all'influenza e, siccome le ricerche su Google sono geolocalizzate abbastanza precisamente attraverso l'IP del computer, sapeva che stavo facendo una ricerca qui piuttosto che in Italia o in un altro paese. Ecco, quando loro vedevano un picco di ricerche di queste 40 parole chiave, prevedevano che in quella zona era in corso un'epidemia di influenza. Lo hanno fatto abbastanza bene questo lavoro, ma fino ad un certo punto.

Io non vado a cercare un rimedio per l'influenza se non ho l'influenza, a meno che si parli tanto per esempio dell'aviaria. E allora tutti leggono e cercano l'influenza aviaria per capire cosa sia. L'algoritmo di Google non era in grado di distinguere se le ricerche erano fatte per cercare effettivamente un rimedio, perché lo avevo bisogno, o semplicemente perché ero curioso di sapere dell'influenza aviaria.

Come si allena, come si fa il *training* di un algoritmo di questo tipo? Dunque si dividono i dati in *training-set* e *test-set*. Se abbiamo i dati degli ultimi – diciamo – 20 mesi, se ne prendono solitamente il 75% per allenare l'algoritmo (*training-set*). Poi si prendono i rimanenti 25% dei dati per fare il test. Su quest'ultimo, ovvero sul *test-set*, si vede se l'algoritmo funziona bene o male e se non funziona bene lo si ricalibra. Per finire lo si mette online e si vede se effettivamente funziona bene nella realtà.

I vantaggi di GFT erano e sono tuttora che è geolocalizzato e immediato, perché se tu hai l'influenza immediatamente cerchi rimedio, non è mediato, nel senso che non cerco rimedio per l'influenza se non ho effettivamente l'influenza.

Non dobbiamo dimenticarci che l'influenza è un problema serio. Causa in media 500.000 morti l'anno. GFT si è rivelato abbastanza accurato fino 2008, l'anno in cui stato messo a punto. L'accuratezza era circa del 97%. Il problema era che l'algoritmo prevedeva in parte sì l'arrivo dell'influenza, ma anche l'arrivo dell'inverno. E forse per questo nel 2009 ha pesantemente sottostimato l'influenza suina, che era iniziata ad aprile ed era stata dichiarata pandemica in giugno dall'Organizzazione mondiale della sanità. Per cui era un po' in controtendenza: un'influenza in un periodo dell'anno dove tipicamente non è presente. Qui GFT ha sbagliato pesantemente. Allora in quel momento ci si è accorti che aveva un problema ed è stato ricalibrato, così ha funzionato abbastanza bene fino 2011. Nel 2012 ha ricominciato a dare problemi, fino al 2013, anno in cui ha di nuovo fatto un grossolano errore: ha sovrastimato pesantemente i casi di influenza. Il motivo principale di questo problema è stato che sono cambiate le abitudini di ricerca online delle persone. La natura pandemica di talune influenze aveva fatto sì che la gente ricercasse il termine online a fini informativi e non perché cercava rimedi contro l'influenza. Tecnicamente il problema di GFT si chiama un problema di *over-fitting*[101], cioè sei molto aderente al passato e ne *fitti* molto bene i dati, poi però la capacità predittiva del futuro diventa praticamente nulla.

Allora, come la matematica insegna, se tu hai due punti di lì passa una sola retta, quindi un polinomio di ordine 1 calza perfettamente. Se tu hai tre punti, c'è un'unica curva quadratica che ci passa, quindi un polinomio di ordine 2 ci passa di nuovo perfettamente. Se nei quattro di punti è una cubica che ci passa perfettamente. In GFT i punti sono i casi di influenza. L'algoritmo che aveva

[101] da *fitting*, in italiano *fittare*, ovvero "adattare una curva ad un insieme di punti".

fittava molto bene il passato perché aveva troppi gradi di libertà il polinomio con cui si *fittavano* i dati. Quindi se tu hai 10 dati e prendi un polinomio di grado 11 ci passa esattamente per tutti questi dati, però se poi chiedo di farmi la previsione di domani per i casi di influenza, probabilmente sbaglierà. Il polinomio era troppo perfetto per il passato, per cui non prevedeva nulla per il futuro.
Del problema se ne sono accorti solo quando hanno cominciato a fare danni. In realtà, probabilmente, se si fossero affidati a uno statistico serio, l'avrebbe detto in anticipo. Questo è uno dei problemi di questi algoritmi: vengono elaborati da colossi che perseguono finalità commerciali. Se invece collaborassero con università o ricercatori che usano modelli seri e che conoscono i limiti di questi modelli, magari alcuni errori grossolani, come questo di GFT, li avrebbero previsti. Uno statistico ti avrebbe detto che questo è un problema classico e avrebbe quindi evitato brutte figure a GFT. Se tu *fitti* troppo bene il passato la tua capacità predittiva del futuro è inesistente. Non prevede mai i cigni neri (anomalie).

Grazie alla *Data Science* si è in grado di "prevedere" alcuni comportamenti dell'Uomo, ciò significa che l'Uomo è sempre più prevedibile? E con sempre maggiore precisione?
Con la Data Science siamo in grado di prevedere sempre meglio e una sempre maggior quantità di comportamenti umani. In realtà, per fortuna, il comportamento umano, così come la fantasia e certe caratteristiche, è difficilmente prevedibile. La nostra mente è dotata di fantasia e questa è una caratteristica bellissima dell'essere umano che un algoritmo non sarà mai in grado di prevedere. Però certi comportamenti abitudinari diventano sempre più prevedibili.
Il problema è che nel momento in cui noi utilizziamo queste tecnologie digitali acconsentiamo ad uno scambio di valori. C'è un abuso del pensiero che ciò che ottieni è molto meno caro di ciò che offri. I miei dati rendono molto di più del benessere che ottengo. Quindi dovremmo essere consapevoli che noi usufruiamo di un servizio e in cambio regaliamo dati. Ma forse non tutti capiamo il valore di questi dati, per cui ben vengano conferenze, tesine e divulgazioni sul valore dei nostri dati. La gente deve essere sensibilizzata.
Io nel mio lavoro ho identificato tre categorie di pensiero riguardo alla raccolta dati e all'utilizzo della tecnologia: apertura, apertura parziale e chiusura. Cosa pensa di questa divisione?
È chiarissima la tua distinzione, quello che cerco di fare io – e che secondo me sarebbe idealmente ipotizzabile che avvenga – è non avere queste persone che hanno una di queste tre categorie mentali ma che il mio atteggiamento di fronte al "Do i dati" / "No" / "Sì", in parte, dipenda dal fatto che io sia informata sulla finalità per la quale la gente vuole i miei dati. Esempio: se viene l'esperto medico che mi dice: "vorrei studiare il tuo patrimonio genetico per capire l'evoluzione di questa

patologia che sappiamo esserci nella tua famiglia, a fini terapeutici", io dico "benissimo" perché la finalità la condivido: ricerca medica, scientifica. Altro esempio: se qualcuno viene a dirmi "voglio i tuoi dati su come ti sposti durante la settimana per andare al lavoro, su che strade fai, eccetera, per migliorare la gestione del traffico nella città di Lugano", io dico: "è una finalità interessante", però magari non voglio far sapere a tutti a che ora esco da casa, che ora arrivo l'università, che strada faccio, ogni tanto faccio un senso vietato in bicicletta, vado in zona pedonale. Per cui ti dico: "Ti do i miei dati però in forma anonima e magari non tutti", quindi ti do i miei dati, ma è come se avessi una manopola. Se c'è qualcuno che mi dice "voglio i tuoi dati per ottimizzare la pubblicità che ti mando sul computer per l'albergo dove devi andare o per proporti il libro che devi leggere", ti dico: "No grazie, non mi interessa, i miei dati non te li do".

Vanno benissimo le tue tre categorie. L'ideale sarebbe che ciascuno di noi assumesse una di queste tre categorie mentali a seconda della finalità che chi ci chiede i dati deve comunicarci. Ma non è tutto qui: siccome i dati una volta dati sono dati – lo so, è un gioco di parole – e i dati possono essere riutilizzati infinite volte con finalità diverse, tu magari in prima istanza mi chiedi i miei dati sulla mobilità per migliorare il traffico, ma una volta che ho detto "va bene mi interessa migliorare il traffico la mattina", tu magari i miei dati li usi anche per altro, perché non ho visto una clausola. Forse per mettermi una pubblicità all'incrocio dove tu sai che io mi fermo solitamente a causa del traffico. Perciò è importante non solo sapere la finalità per la quale i dati vengono utilizzati, ma anche sapere che ciascuno di noi possa seguire il percorso che questi dati fanno.

Questa è una caratteristica della scienza dei dati: i dati esistono a prescindere da domande che ci siamo posti. La statistica fino a anche solo cinque anni fa funzionava così: tu, Data Manager, venivi da me dicendo "voglio fare un'indagine su questa nicchia di mercato, questo prodotto", allora io raccoglievo dei dati per rispondere a una certa domanda di ricerca. Adesso non è più così: i dati esistono a prescindere dalle domande che qualcuno si è posto. Oppure sono stati raccolti per certe finalità, ma siccome poi costa poco tenerli e renderli disponibili, ci sono. Allora il vero scienziato è quello che sa porre le domande giuste ai dati, colui che si inventa con la fantasia le domande giuste per interrogare dati che magari sono stati raccolti con finalità completamente diverse. Un esempio tipico che faccio sono le camere di sicurezza dei *drive-in* dei fast food: prendono immagini per sicurezza, che poi però rimangono. Ora qualcuno si è inventato di usare queste immagini per ottimizzare il servizio fornito ai clienti. Come? La camera di sicurezza vede se al *drive-in* ci sono una macchina, due macchine o 10 macchine. Se c'è poca gente, ti faccio vedere sul display digitalizzato prodotti che ci metto un pochino di più a prepararti, che hanno un valore aggiunto, che costano anche un pochino di più. Se invece la coda è lunga, ti mostro soltanto prodotti che ho già lì pronti, per cui tu me lo chiedi e io te lo do. I dati li avevano comunque, perché li raccoglievano con

finalità di sicurezza. Le telecamere ti dicono addirittura se sull'auto c'è un bimbo e allora ti offro l'Happy Meal[102], o l'anziano o la donna. Siccome conosciamo le preferenze in base all'età, al sesso,…, allora ti faccio una "offerta speciale".

Sì, ottimizzo il servizio, però chiaro che magari un po' di privacy non guasterebbe. Forse sono al *drive-in* con una persona con cui non dovrei essere la sera.

Al di la dell'Happy Meal, nessun pranzo è gratis.

Se quello che ti offro è un servizio gratuito, significa che paghi da un'altra parte. Lì è un po' il problema.

Pensa il caso del Canton Ticino nel suo piccolo: esiste un ufficio per la protezione dei dati che è regolato dalla legge X. Questo ufficio si occupa dei dati raccolti dagli enti pubblici, non ha nessun potere sui dati raccolti da chiunque altro. Quindi se la Coop passa i suoi dati alla Denner, questo ufficio non ha niente da dire. Ma se io, direttore della scuola tal dei tali, passo i dati del mio allievo alla scuola Y, per legge non potrei. Si suppone che il cantone non faccia commercio dei suoi dati, mentre gli altri sicuramente lo fanno. Si vuole controllare la seconda categoria, che si suppone sia sicuramente etica, mentre non si controlla chi oltre all'etica, giustamente fa suo mestiere. L'azionista vuole i soldi, perciò adopera i dati per fare altre cose. Ecco, queste sono le cose che danno da pensare. Proporre una legge che protegga i dati non solo concessi all'ente pubblico, ma anche all'ente privato. Ma una legge così non passerà mai, a meno di un cigno nero, che può sempre capitare.

Ha mai pensato ai rischi legati agli abusi possibili nella manipolazione di grandi quantità di dati? Mettiamo che chi li maneggia non abbia nobili scopi (criminalità informatica).

La scorsa settimana c'era in USI un interessante conferenza sugli *spin-doctors*. Gli *spin-doctors* sono esperti di comunicazione che lavorano come consulenti per conto di personaggi politici, per esempio Obama e Trump li hanno. Il loro compito è elaborare l'apparenza del politico e come si pone attraverso i media all'opinione pubblica per ottenere un consenso elettorale. Nei dibattiti pubblici il politico degli Stati Uniti non ha un discorso cartaceo, ma uno scritto sul tablet, poiché i suoi *spin-doctors* monitorano in tempo reale le reazioni che le persone hanno sui social al suo discorso. E pure in tempo reale gli cambiano certe parole del discorso a seconda di come la gente *twitta* o ri-*twitta*. Modificano i termini e le frasi da usare, piuttosto con una sfumatura invece che un'altra. Addirittura gli dicono come porsi fisicamente di fronte al pubblico, e di compiere determinate azioni. Esempio: "Vai sul tram e attaccati a quella maniglia, perché questo dà l'idea di una persona vicina al popolo".

[102] pasto per bambini caratteristico della catena di fast food McDonald's.

Questi *spin-doctor* possono anche manipolare, perché usano i dati per guidare comportamenti di personaggi pubblici e li guidano verso quello che la gente vuole sentirsi dire o verso quello che loro pensano alla gente piaccia e non piaccia.

Quindi il problema dell'appiattimento di cui diceva prima cresce?

Eh sì.

Ma questo è fatto da persone oppure da algoritmi?

Probabilmente è un mix delle due componenti. Perché i *tweet* e i ri-*tweet* che la gente manda durante un dibattito politico, non li puoi monitorare come persona. C'è un algoritmo che ti legge il sentimento: si chiama *sentiment analysis*. Qui ci sono algoritmi che innanzitutto devono capire il senso di un *tweet*. Se uno scrive "Bella fregatura" mentre un politico dice una certa frase, un algoritmo che in automatico tenta di capire la natura del commento ne concluderà: "Beh, *bella* è positivo e *fregatura* è negativo", per cui non sa come catalogarlo, mentre un umano ti dice subito che è negativo. Alcuni di questi algoritmi funzionano completamente in automatico, perciò da questo tipo di frase sono completamente spiazzati. Un altro esempio è un film con Stallone[103]: "Lui dà il meglio di sé, il cast è incredibile, fantastici effetti speciali,… però la pellicola delude.". Nel commento si contano tre aggettivi positivi, però alla fine *delude*. La macchina cosa fa? Dice: "Beh, tre contro uno è positivo.". Invece no.

Quindi gli algoritmi in automatico ti dicono se il sentimento generato da un certo discorso fino a quel momento è stato positivo o negativo. Io penso a una persona *spin-doctor*, ma potrebbe essere fatto in automatico: da un algoritmo che legge il sentimento e un altro che ti cambia discorso. Però in questo caso c'è comunque sempre una persona che intermedia la *sentiment analysis*.

Questa grande messe di dati fa anche paura. Io ho svolto un piccolo sondaggio che sembrerebbe confermarlo. Il cittadino che può fruire della conoscenza in realtà teme questa nuova dimensione. Insomma Big data è un po' anche Big Brother?

Secondo me non è così. Ma c'è il rischio che diventi così. In questo momento, più che da un unico Grande Fratello, siamo circondati da tanti piccoli fratelli. Ci sono tanti colossi dei dati che possiedono ciascuno una fettina di quelli che generiamo e le logiche concorrenziali fanno sì che i colossi non collaborino fra loro. Quindi ognuno si tiene i suoi dati, non li condivide e cerca di estrarne il meglio per il suo business. Esempio: sappiamo benissimo che Netflix conosce le nostre

[103] Silvester Stallone, attore americano.

preferenze sul cinema, Amazon le nostre preferenze sui libri. Però, fortunatamente, sono concorrenziali fra di loro, per cui non si scambiano le informazioni.
C'è però il rischio che adesso Google cominci a comprare. Ha l'email, ha il *tracker*[104] di dove ci muoviamo, ha Maps: insomma ha tante cose. Perciò Google comincia ad avere più fettine.

Mi può spiegare come lavora il vostro istituto, la sua storia e i suoi progetti? Collabora con altri istituti? Utilizzate il supercalcolatore del Centro Svizzero di Calcolo Scientifico (CSCS)?
Sì, lo utilizziamo: abbiamo parecchi progetti con il CSCS. Per esempio il progetto di cui proprio poco fa stavo parlando in questa chiamata Skype con il collega di Harvard. Lavoriamo per progetti di ricerca perlopiù finanziati dal Fondo nazionale svizzero, ma non solo. Per esempio abbiamo un progetto finanziato dall'AXA Foundation. E su questi progetti io riesco a retribuire *post-doc*. E tipicamente tali progetti sono fatti in collaborazione con altri colleghi, perché la Data Science è una scienza interdisciplinare, per cui io copro la parte di analisi dei dati, di modelli, di previsioni e in parte anche un po' computazionale. Però poi ci sono altri colleghi che hanno altre informazioni e altre competenze. C'è sempre una *knowledge-to-main-expert*[105]; per esempio collaboro con il professor Barone Adesi, della facoltà di Economia dell'Usi, per i progetti di finanza e collaboro con il collega di Harvard - di nome J.P. Honeill - che è un esperto di dati di tipo network. Ovvero di dati di tipo relazionale, che oggi abbondano. Penso ai dati tipo Facebook, Twitter, che sono di tipo network, cioè dove ci sono dei nodi, che possono essere persone, e ci sono dei link fra i nodi, che possono essere per esempio le amicizie. Questi nodi possono anche essere imprese, paesi o proteine; i link in questo caso sono rapporti di tipo commerciale, come fornitore-cliente, di import-export, o di interazione tra proteine.
Ci sono quindi dei link che legano i nodi, e noi studiamo la tipologia di questi network. Per esempio quanto denso è il network delle amicizie su Facebook nei paesi europei, rispetto a quelli americani; e per esempio certi comportamenti di questi network: se si tende a triangolare le amicizie, cioè se A è amico di B e B è amico di C, quanto è alta la probabilità che A diventi amico di C? Queste sono proprietà di questo network che noi studiamo utilizzando dei modelli. Questo è proprio il progetto che abbiamo con il collega della Harvard School of Public Health. Perché poi studiamo anche i processi che si sviluppano su questi network. Per esempio se il network è quello delle persone e i link non sono di amicizia ma di contatto/vicinanza fisica, attraverso lo studio della tipologia del network, noi indaghiamo come si sviluppa e come evolve un'epidemia nello spazio e nel tempo.

[104] dal termine inglese *track*, ovvero *percorso*.

[105] dall'inglese "la conoscenza dell'esperto principale".

Altri processi che si sviluppano sui network possono essere l'espansione delle *fake news*[106] su un network di tipo social. Penso a Facebook: una notizia può essere vera o falsa, a seconda di come questa si muove.

Un altro progetto molto interessante è in corso con il Cardiocentro e la Fondazione Ticino Cuore. Anche qui c'è la *knowledge-to-main-expert*, cioè chi conosce i dati e le problematiche. In questo caso abbiamo i dati su tutte le chiamate di emergenza alle ambulanze. Tra tutti questi abbiamo estratto quelli con eventi di tipo cardiaco: come ictus e infarto. Abbiamo la geolocalizzazione dell'evento, quindi dove è avvenuto il fatto, quanto tempo ci ha messo l'ambulanza ad arrivare, in che stato era il paziente e se vi era un *first-responder*[107], ovvero colui che si accorge che c'è una persona che sta male. Poi, se questa persona è intervenuta per esempio andando prendere un defibrillatore: quale ha preso, quanto tempo ha impiegato ad arrivare, com'era lo stato del paziente quando è arrivate in ospedale. L'obiettivo è quello di ottimizzare il posizionamento dei defibrillatori in Ticino.

Ma è già stato fatto?

Sì, abbiamo un articolo pubblicato. L'obiettivo non è però soltanto posizionare dei nuovi defibrillatori sul territorio, ma anche stabilire se conviene comprarne di nuovi, che costano circa 3'500 CHF l'uno, oppure spendere tra i 700 e gli 800 CHF per prendere un defibrillatore che già esiste sul territorio e spostarlo da un'altra parte, avendo tot. finanziamenti per migliorare l'assistenza sul territorio per problemi cardiaci.

Questo è molto interessante perché è la prima volta che è stata fatta un'analisi di questo tipo: non solo comprarne uno, ma anche riposizionare defibrillatori esistenti.

Quindi l'avete inventato voi?

Sì, l'algoritmo e la metodologia l'abbiamo inventato noi. Ci sono i colleghi del Cardiocentro e della Fondazione Ticino cuore che ci hanno posto il problema. Quelli della fondazione, che posseggono i dati, ce li hanno messi a disposizione; poi abbiamo chiesto all'Ufficio federale di statistica, che ci ha dato le coordinate GPS di tutti gli edifici sul territorio ticinese. Perché se possiedi un algoritmo di ottimizzazione del posizionamento dei defibrillatori, magari lui ti dice di metterlo in mezzo al lago o in un posto dove non puoi metterlo. Allora abbiamo detto all'algoritmo che poteva mettere defibrillatori soltanto dove c'erano edifici. Quindi i dati diventano grandi, perché ci sono tutti gli eventi, tutte le coordinate degli edifici, e ora vogliamo migliorare il modello mettendo anche i dati di Meteo Swiss e costruire una mappa del rischio per problemi di tipo cardiaco sul territorio cantonale.

---

[106] notizie false

[107] primo soccorritore

In relazione alle condizioni metereologiche?

Sì, in particolare alle condizioni di inquinamento da polveri sottili, perché la letteratura dice che esiste una correlazione fra problemi cardiaci e condizioni atmosferiche.

Un altro progetto, a cui accennavo, è finanziato dall'AXA Foundation, che abbiamo in collaborazione con Viseca, che gestisce circa il 30% delle transazioni con carte di credito in Svizzera. Questo progetto tratta delle frodi con carte di credito. Il problema è *early for detection*[108]: ogni tanto ci sono delinquenti che clonano o rubano la nostra carta di credito e fanno transazioni online oppure fisiche. Quindi l'obiettivo è cercare di arrivare ad intervenire in tempo.

Analizzando i dati delle transazioni resi disponibili da Viseca, ovviamente anonimi, cerchiamo di individuare in tempo reale o il più presto possibile, eventuali transazioni fraudolente. Cercando, da un lato, di minimizzare i falsi allarmi, perché se ti blocco la carta di credito per niente non sei molto contento. E, dall'altro, di minimizzare i tempi che trascorrono fra la frode avvenuta e il blocco della carta. Anche questo è un progetto che coinvolge grandi basi di dati.

Da dove provengono i dati che utilizzate nelle vostre analisi?

Alcuni li compriamo. Per esempio adesso abbiamo comprato i dati su tutti i brevetti a livello europeo e mondiale: quelli cinesi, quelli giapponesi, di alcuni uffici brevetto specifici, perché – restando in tema di network – in un brevetto citi altri brevetti. Quindi si crea una rete e noi vogliamo studiare come la conoscenza e l'innovazione nascono e si sviluppano e anche come brevetti di diverse classi tecnologiche sono legati tra loro.

Questi fondamentalmente li abbiamo comprati, anche se i dati sui brevetti sono pure disponibili online. Per cui ci sono i colleghi di informatica che costruiscono dei *crawler*[109], che sono come dei ragni che girano sul web e raccolgono dati. Per esempio tutti i dati di Twitter sono pubblici da oggi fino a un mese addietro. Però, se io decido da oggi in poi di fare un'analisi di tutti i *tweet* su l'Università della Svizzera italiana, Lugano, Ticino, e comincio a monitorare certe parole chiave, li ho gratuitamente. Se invece voglio andare indietro nel tempo più di un mese devo pagare.

Perciò alcuni si comprano, altri, come quelli della Fondazione Ticino Cuore, ce li danno, gratuitamente. Però, in cambio della nostra analisi. Quindi noi paghiamo con la ricerca che facciamo, fornendo poi un servizio alla Fondazione e al Ticino sul posizionamento dei defibrillatori. Quello che ci guadagniamo noi è fare delle pubblicazioni e della ricerca interessante e sviluppare delle metodologie nuove. Quindi in questo caso è un *do ut des*: loro ce li danno gratuitamente, loro

[108] troppo presto per il rilevamento

[109] programmi che analizzano il contenuto della rete

hanno servizio, noi abbiamo il vantaggio di avere dati interessanti e di fare dei casi studio. L'idea, insomma, è che questo progetto sia un prototipo e che venga poi espanso in Svizzera, in Lombardia, o in un'altra regione dove ci sono i registri degli eventi cardiaci.

Per cui: o andiamo dai nostri colleghi informatici e diciamo "Cercate i dati sul web", o li paghiamo, o ci vengono dati gratuitamente. Per esempio Viseca, con il progetto della AXA Foundation, ci fornisce dati gratuitamente. Noi li analizziamo con queste metodologie, però i risultati della nostra ricerca li diamo a loro in maniera privilegiata. Prima di pubblicarli su una rivista scientifica loro hanno diritto di sfruttarli per migliorare la *fraud-detection*[110].

Sono forme di collaborazione fra gli enti che detengono questi dati e la ricerca universitaria su questi dati. Loro ce li forniscono gratuitamente, noi costruiamo modelli e algoritmi che estraggono informazioni, li mettiamo a loro disposizione e pubblichiamo degli articoli su riviste scientifiche.

Perché poi la ricerca che viene fatta in università è gratuita, cioè è chiaro che io sono pagata per fare ricerca, ma poi pubblico un articolo e lo metto gratuitamente disponibile a tutti su una rivista scientifica. A meno che uno si inventi qualcosa di assolutamente innovativo. Per esempio potrebbe succedere che dalla ricerca con l'AXA Foundation, se la metodologia che usiamo e l'algoritmo che troviamo risultano molto migliori rispetto quelli che attualmente loro usano per fare la *fraud-detection*, magari potremmo decidere di brevettarlo. Allora potremmo brevettarlo con loro e forse avere un ritorno economico. Anzi l'università ci motiva a fare ricerche che poi hanno delle ricadute brevettuali, perché è un modo con cui l'università si misura in termini di competitività.

Il brevetto a chi appartiene? All'università o al ricercatore?

Dipende. Io posso brevettare da sola, però, siccome brevettare ha dei costi, devo sostenerli io. Ad esempio, ISA[111] lo ha brevettato VFB come società, non come università. Il collega di cui parlavo prima ha brevettato invece questa applicazione attraverso Harvard, che ha curato tutta la parte legale, le procedure e i costi.

11.2. Intervista a Luigi Curini

Luigi Curini è professore di Scienze Politiche all'Università di Milano (UNIMI) e cofondatore di Voices from the Blogs, spin-off dell'UNIMI di cui ho già parlato nel punto "3.7. Voices from the Blogs".

Ho potuto contattarlo via Skype il professore per porli le mie domande sul tema Big Data.

[110] individuazione di una frode

[111] l'algoritmo brevettato da VFB di cui ho parlato nel capitolo "3.7. Voices from the Blogs".

Per profani che vogliono capire: che cos'è in poche parole *Voices from the Blogs*, un marchio dal nome decisamente intrigante?

Voices from the Blogs (*VfB*) è nato nel 2011 quale centro di ricerca su iniziativa mia e di Stefano Iacus. Il suo obiettivo – un po' sulla falsa riga di quanto stava nascendo negli Stati Uniti in quel periodo – era di sviluppare nuovi metodi per l'analisi del flusso di informazioni presenti sui social media.

Siamo scienziati sociali. Ritenevamo quindi importante recuperare le informazioni presenti nei milioni di siti web. Non sfruttare questa fonte di informazione ci sembrava una perdita intollerabile e ingiustificabile. Perciò l'idea all'inizio era: creiamo la nostra metodologia, volta ad analizzare questo flusso di informazioni. Poi, siccome il progetto ha avuto molto successo e molta eco sulla stampa – in particolare per l'analisi fatta per il Corriere della Sera sulle presidenziali americane del 2012 – ecco che l'Università degli Studi di Milano ci ha chiesto: "Perché non creiamo anche uno spin-off[112]?", in modo da andare sul mercato per intercettare una fetta di domanda che a fine 2012 era ancora agli albori, ma che in questi anni è cresciuta esponenzialmente. La fetta di mercato è quella delle istituzioni pubbliche e private, interessate in senso ampio ai Big Data e in particolare all'*Analytics*[113]. "Data Analytics" è un termine che va molto di moda oggi ma quando avevo iniziato con VfB, non era un concetto molto diffuso. Adesso, la prospettiva di analisi di dati è ormai molto richiesta, sia nel settore privato che pubblico.

Più o meno è questo VfB: ha iniziato come progetto di ricerca che si poneva degli obiettivi ben chiari nel campo delle scienze sociali statistiche e si è trasformato col tempo in spin-off che opera sul mercato.

Da quando siete attivi, vi siete resi conto del repentino aumento dei dati a disposizione. A cosa riconduce il fenomeno della potente crescita?

Allo sviluppo informatico. La crescita esponenziale dei social media è però solamente una delle fonti di informazioni sui dati a disposizione.

---

[112] Impresa nata per scorporamento da un'altra, la quale mantiene tuttavia un ruolo fondamentale nei confronti della nuova realtà imprenditoriale, esercitando su di essa una significativa influenza soprattutto in termini di competenze e di attività svolte.

Tratto dal sito http://www.treccani.it/enciclopedia/spin-off_%28Dizionario-di-Economia-e-Finanza%29/ (8 novembre 2017, 17:41).

[113] intesa come Data Science

I Big Data sono legati alla diffusione cruciale del digitale: si pensi agli *open data*[114], legati all'amministrazione pubblica, resi disponibili ancora una volta dalla rivoluzione digitale. Oppure ancora i dati legati alle transazioni commerciali con carte di credito, con bitcoin[115], con assicurazioni. Tutta una pluralità di fonti di dati che dal digitale sono state rese disponibili. Perché in realtà sono tutte fonti di dati che c'erano già anche prima. Una volta, al posto dei social media c'era il chiacchiericcio del bar, che produceva esattamente lo stesso flusso di informazioni, solo che era più difficile da raccogliere. Gli *open data*, se si pensa in senso ampio, erano sempre presenti, solo che erano presenti su pezzi di carta sparsi in migliaia di uffici. Adesso sono stati riportati a unità grazie soprattutto allo sviluppo del digitale. Ciò che ha fatto il digitale è semplicemente rendere disponibile un flusso di informazioni che era già presente in forma non digitale.

Da un singolo *tweet*, quale tipo di informazioni sull'utente e con quale precisione è possibile estrarre?

Da un *tweet* si può ricavare tutto. Leggendo un *tweet*, specialmente ora che aumenta da 140 a 280 caratteri[116], si è in grado di ricavare tutta una serie di informazioni ad esso connesse. Faccio sempre questo esempio: immaginate un *tweet* che dica "Oh che bello il nuovo iPhone. Ha lo schermo più grande, ma non lo compererò mai perché costa troppo.". La frase contiene un sacco di informazioni: qualcuno sta parlando del nuovo iPhone, gli piace, perché gli piace? Ha lo schermo più grande ma non lo compra comunque perché costa troppo. Il problema è come scalare questa serie di informazioni, che si possono estrarre da un *tweet*, a milioni di dati. Questa è la vera sfida.
Ci sono vari algoritmi che puntano a questo obiettivo. ISA, è il nostro. Senza entrare nella parte statistica, che è un po' complicata, ci sono i codificatori umani, che fanno il ragionamento che abbiamo appena fatto noi. Leggono un sottoinsieme di post che possono essere di Twitter, Facebook, blog o qualunque cosa. La parte di codifica manuale è fondamentale perché permette effettivamente di sfruttare lo *human brain*[117] per associare una serie di lemmi, di parole, a una serie di categorie semantiche.

[114] dall'inglese "open data", "dati aperti", sono dati accessibili a tutti senza restrizioni.

[115] è una criptovaluta elettronica che ha acquistato sempre maggiore valore e popolarità negli ultimi anni. Nata nel 2009 è sempre più utilizzata per scambi di natura commerciale sul web, specialmente per commerci di tipo illegale. Questo perché è irrintracciabile e i traffici non sono mediati da alcun tipo di istituto finanziario.

[116] Prima era possibile digitare solamente 140 caratteri in un *tweet*, oggi 280.

[117] cervello umano

Questa fase è cruciale, perché permette di fare il *training*[118] dell'algoritmo che apprende dalla codifica manuale e la estende sull'intero insieme di dati, che possono essere milioni, producendo nel nostro caso un piccolo errore statistico che è inferiore al 3%. Un errore basso, ma che al tempo stesso ha il vantaggio di permettere comunque di analizzare una massa di dati che sarebbe impossibile leggere (Big Data).

Per analizzare i dati in vostro possesso ricorrete ad algoritmi specifici (*ISA*) e all'intelligenza artificiale. È tutto automatizzato o l'intervento umano è comunque necessario? Sarà così anche in futuro?

È necessario e continuerà sempre ad essere necessario, anche se, in questo momento, si parla tanto di intelligenza artificiale. L'intelligenza artificiale di cui si discute adesso è legata agli algoritmi di *machine learning* e non ha niente a che vedere con l'idea del robot di "2001: Odissea nello spazio"[119]. Non è pensabile fare a meno della parte di interpretazione umana.
L'idea che si possa fare a meno dell'intervento umano per l'interpretazione dei testi in senso ampio, come in questo caso, è completamente farlocca. Nel senso che la migliore tecnologia, come ci ricorda sempre Gary King[120] è "Computer-assisted and human empowerment"[121]. L'idea è che ci debba essere il contenuto umano, sfruttando al tempo stesso la potenza dell'informatica e la precisione resa possibile dalla statistica. Ma fare a meno dell'interpretazione umana è un'idea che non è nemmeno affascinante e che, in questo momento, non esiste.

Nuove conoscenze, ottenute dalla navigazione mirata nell'oceano dei dati, possono migliorare la nostra vita. Ma per ottenerle la contropartita è un possibile più esteso controllo sociale. Non teme che, attraverso la gigantesca raccolta dei dati, potrebbe aprirsi la strada ad un controllo invasivo e pericoloso della vita (dei pensieri) degli esseri umani?

Non "potrebbe aprirsi", togli pure il condizionale. Questa paura c'è ma è la sfida che avviene sempre in ogni tipo di rivoluzione tecnologica. Come quando si parlava dell'evoluzione legata alla scoperta dell'energia atomica, che ha poi provocato la bomba atomica.
Il problema è sempre riuscire a controllare la cosa. Secondo me il vero rischio è di non essere coscienti del rischio. Questa è il vero pericolo. È la coscienza del rischio che ci permette di pensare a possibili contromosse. Pensare a qualcosa che permetta di tenere sotto controllo questo pericolo,

[118] allenamento

[119] film di Stanley Kubrick uscito nel 1968.

[120] Esperto di Scienze Politico statunitense

[121] dall'inglese "Computer-assisted and human empowerment", "Assistenza dei computer e sviluppo umano".

perché non lo si può annullare come in altri casi: lo si può solamente tenere sotto controllo. Il vero rischio è la mancanza di consapevolezza di questo rischio, che è drammatica, perché le persone non si rendono conto di quanto stanno pagando per l'utilizzo ad esempio dei social media. Non si rendono minimamente conto che Facebook, Twitter, Instagram non sono gratuiti, ma che stiamo cedendo loro pezzi della nostra privacy. Pezzi, che possono essere utilizzati in modo positivo o negativo, certo. Vero però è che li stiamo cedendo.

Spesso non siamo abbastanza coscienti della cosa, di questo sono un po' preoccupato: della mancanza di coscienza di quello che noi stiamo pagando in termini di accesso a aspetti ludici come i social network. È un po' questo quindi che mi mette più che altro ansia, perché la disponibilità di un flusso crescente di informazioni può sempre essere utilizzata per un controllo sociale.

C'è qualcosa di nuovo a questo riguardo? No, perché la politica fin dall'antichità ha sempre cercato di produrre un certo tipo di controllo sociale sulla base delle informazioni che aveva a disposizione. Un tempo le informazioni che erano a disposizione, venivano prodotte dal tempio religioso, dai burocrati dell'antico Egitto, dalla chiesa più avanti o dalle informazioni che venivano distribuite a scuola. C'è sempre stato un tentativo di utilizzare informazioni a proprio vantaggio. Questo è indubbio. Peggio, ora le informazioni a disposizione sono cresciute e l'accesso a queste è più facile e il rischio cresce di conseguenza.

Ancora una volta: non è nulla di nuovo rispetto alla storia umana. Quelli che tentano di descrivere l'avvento dei Big Data come un cambiamento sono piuttosto miopi, perché in realtà è la semplice riproposizione di problematiche che sono state sempre presenti all'interno della struttura sociale, semplicemente si sono modificate e sono accresciute dalla maggiore disponibilità di questi dati.

Introdotta la General Data Protection Regulation, cambierà qualcosa per voi e per istituto come il vostro?

Per esempio, ogni azienda che utilizza dati di Twitter deve fondamentalmente sottoscrivere un accordo. In maniera implicita, perché non è mai formale. Chi utilizza dati di Twitter in modo sistematico, deve soddisfare una serie di criteri etici. Tra questi la regola secondo la quale tutti i dati utilizzati devono essere anonimizzati. Questo è quello che si dovrebbe fare ma per esempio, quando noi scarichiamo dati da Twitter o da Facebook, non sono anonimizzati: noi sappiamo da chi li stiamo scaricando. Di fronte a terze persone dobbiamo garantire questo anonimato, però noi, questi dati, li possiamo comunque utilizzare a nostro vantaggio. Se nella tua pagina Facebook ti

compaiono degli *ad*[122], che sono in qualche modo collegati alle ricerche che fai normalmente non solo su Facebook, ma anche su Google, significa che c'è un collegamento tra le varie cose.
Quindi, sì, la GDPR permetterà di garantire una maggiore difesa della privacy. Però… però ci sono molti modi per girarci intorno, capisci? Per chi vuole ci sono una quantità enorme di modi per girarci intorno. La difesa sarà radicale? Secondo me no, solamente quando verranno portati in sede legale alcuni casi, si potrà fare affidamento a questo scudo come modalità di ulteriore protezione.
In azioni concrete mi pare comunque difficile. Non la vedo fattibile: la vedo su carta ma non in azioni concrete.

A chi sono destinate le vostre analisi? Chi ricorre hai vostri servizi? Privati, enti pubblici, ricercatori, scienziati…

Ci sono fondamentalmente quattro grandi aree di agenti interessati alla nostra analisi. La prima grande area è quella del business, generalmente il grande e non il piccolo business. Sono infatti le grandi realtà che hanno maggior consapevolezza dell'importanza della cosa. Nelle realtà piccole-medie, soprattutto quelle piccole, c'è ancora una sostanziale sottovalutazione dell'importanza che questo tipo di fonte di dati potrebbe procurare nelle ditte. Quelle grandi sono ad esempio settori come quelli bancario, finanziario, farmaceutico, alimentare. Poi, seconda area, c'è tutto il settore delle istituzioni pubbliche dove l'interesse pure è enorme. Il terzo è quello delle arene politiche, i partiti politici, i candidati, che sono molto interessati e vedono tali analisi quale strumento sostitutivo dei sondaggi. Un quarto settore è quello delle agenzie di comunicazione per produrre campagne.

Sta pensando ad altri progetti nell'ambito della *Data Science*?

Stiamo operando molti sviluppi: una delle cose che stiamo sviluppando è il nostro algoritmo per l'analisi delle immagini. Io penso alla tua generazione o a quella dopo di te[123]. I nativi digitali non si scambiano comunicazioni attraverso un flusso di dati testuali, quello lo fa la mia generazione e quelli un pochino più vecchi. I nativi digitali lo fanno sempre di più attraverso l'utilizzo di immagini. Quindi diventa importante sviluppare strumenti e algoritmi, perché anche le immagini portano con sé una serie di significati semantici importanti: felicità, rabbia... . Questo è un progetto su cui stiamo lavorando da un punto di vista scientifico e siamo già a buon punto.

[122] abbreviazione del termine inglese *advice*, che sta per avviso, notifica.

[123] la mia generazione è intesa come quella di chi è nato negli anni '90; quella successiva sono tutti i nati dopo il 2000.

Fate uso di intelligenza artificiale?

Gli algoritmi che apprendono sono già intelligenza artificiale. Il nome intelligenza artificiale è altosonante. Il dibattito odierno sull'intelligenza artificiale è molto diverso da quello che c'era prima. Venti-trenta anni fa c'era l'idea di sviluppare un robot che fosse autocosciente e l'obiettivo di sviluppare una macchina che passasse il Test di Turing[124] come prova di autocoscienza. Ma l'intelligenza artificiale attuale non è quella roba lì. È invece legata al *machine learning*, alle reti neurali, al *deep learning*[125], che in qualche modo rientrano anche nel nostro algoritmo. Non significa quindi sviluppare un algoritmo intelligente nel senso che superi i test di Touring e che quindi abbia autocoscienza. No, significa sviluppare algoritmi che sono in grado di apprendere qualche cosa per sostituire gli umani nel fare quella cosa. Magari funzionano benissimo, ma questi algoritmi sono completamente stupidi e non sanno perché stanno facendo quella cosa. Ignorano il processo causale che sta dietro. Non c'è uno sviluppo di coscienza, una mente. Semplicemente apprendono, come *SIRI* sul tuo iPhone. Dopo un po' lui apprende dalla voce e dal tono che si utilizza, ma non è che *SIRI* sia diventato intelligente. Semplicemente ha imparato da una serie di *training* fatti nel tempo. Quindi è l'uso del *training* che permette l'ottimizzazione di quello precedente. Non si tratta quindi di sviluppo del pensiero critico su quanto appreso.
In altre parole, nella discussione attuale sull'intelligenza artificiale, non si pone più il problema della Legge di Turing. Il Test di Turing non esiste più, perché si è capito che non è questo il momento per fare un dibattito sull'autocoscienza. Insomma, tutti gli sviluppi di successo, dal punto di vista dell'intelligenza artificiale, rinviano a un processo che è completamente differente rispetto all'idea di intelligenza artificiale che c'era venti-trenta anni fa. Idea che, se non è morta, quasi lo è. Il robot che va in giro e sostituisce gli essere umani è un'idea – ancora una volta – completamente farlocca. È un'idea che funziona nel senso che un robot capace di imparare a svolgere determinate azioni – facendole anche estremamente bene – non ha coscienza. Il grande pubblico rimane legato all'immagine del robot che arriva e poi prende il posto come fosse Terminator[126]. No, non ci siamo. Ciò accade in letteratura, nell'applicazione concreta siamo lontanissimi da quest'idea.

---

[124] *È una* prova ideata dal matematico Alan Turing nel 1950 *che* consiste proprio nel sottoporre dei computer all'interrogatorio di giudici «umani» che devono capire se si trovano al cospetto di esseri pensanti in carne e ossa o di macchine in grado di formulare pensieri ed esprimerli.
Tratto dal sito http://www.corriere.it/tecnologia/14_giugno_09/computer-intelligente-supera-test-turing-62f684fa-efbb-11e3-85b0-60cbb1cdb75e.shtml (8 novembre 2017, 18:06).
[125] dall'inglese "deep learning", "apprendimento profondo", è uno dei campi di ricerca nell'ambito dell'intelligenza artificiale che si occupa dell'apprendimento automatico di un sistema.
[126] Personaggio della saga di film fantascientifici *Terminator*, nella quale robot umanoidi dotati di intelligenza artificiale prendono il sopravvento sull'Uomo.

Big Data è anche un po' Big Brother?

Sì, sicuramente. Ma questo rinvia al problema della privacy. Le persone non sanno quante tracce digitali facilmente riconducibili a loro stesse stanno seminando ogni giorno. E siccome non sono minimamente coscienti continuano a farlo in modo ampio e, siccome continuano a farlo in modo ampio, esiste la possibilità concreta – ritorno al discorso precedente – di raccogliere e di utilizzare in modo proficuo tali informazioni su ciascuno di noi. Informazioni che ognuno di noi rilascia in modo comodo a disposizione di chiunque. Quindi questo è una sorta di Big Brother, in senso ampio ripeto. Il Big Brother in qualche modo c'è già: quando tu vai su Amazon, e prima di fare una richiesta, ricevi già possibili suggerimenti sui prodotti, è in qualche modo Big Brother. In senso ampio, in senso benefico, benigno più che benefico. C'è già, okay? Le persone non si rendono conto di questa realtà. Non c'è minimamente coscienza del fatto che siamo ormai arrivati a un punto in cui le informazioni disponibili su ciascuno di noi possono essere utilizzate in modo proficuo. Speriamo nel nostro interesse ma possono essere utilizzate anche non nel nostro interesse. Molto spesso questo è un tema su cui c'è poca sensibilità. E un po' in generale tutto quello che riguarda la Data Science. Però questo in sé è solamente il lato negativo perché noi tutti siamo contenti di avere l'iPhone o il Samsung. È sempre una questione di pro e contro. Il mio non vuole essere un luddismo[127] digitale. Lungi da me essere luddismo digitale, tutt'altro. Però, come ogni sviluppo tecnologico, comporta una serie di rischi collegati di cui forse, solamente adesso, cominciamo a rendercene conto. Ma per un certo periodo – è chiaro – non ci abbiamo fatto minimamente caso

11.3. Intervista a Gianni Cattaneo

Gianni Cattaneo è un avvocato specializzato in diritto di internet e professore di diritto informatico alla SUPSI.

Dopo vari incontri a diverse conferenze nell'ambito delle nuove tecnologie, ho avuto modo di recapitargli le mie domande via mail.

Alla prima risposta ho necessitato una precisazione, siccome il parere dell'avvocato Gianni Cattaneo risultava parzialmente in disaccordo con quanto affermato da un suo collega in un'intervista.

[127] Ideologia che si oppone al progresso tecnologico.

La Legge federale sulla protezione dei dati risale al 1992, Google non era nemmeno nato e i Big Data non erano neppure un tema. Che ne pensa della qualità dell'attuale legislazione sulla protezione dei dati in Svizzera?

L'attuale Legge federale sulla protezione dei dati è una legge flessibile e tecnicamente neutrale fondata su principi generali di tutela sostanziale della privacy e dei dati personali. Dunque posso esprimere un giudizio positivo a livello generale.

Bertil Cottier è professore associato di Diritto dell'informazione alla Facoltà di Diritto dell'Università di Losanna e professore ordinario di Diritto alla Facoltà di scienze della comunicazione all'USI. Cottier, in un'intervista rilasciata al Corriere del Ticino, sostiene che:

> *"[Dal punto di vista della legislazione sulla protezione dei dati in Svizzera] siamo indietro non di una ma di due generazioni. […] è come se le regole della circolazione stradale, in ambito legale fossero riassunte in tre articoli, di assoluto buonsenso e assolutamente condivisi, ma totalmente generici: guidate con prudenza, date la precedenza e tenete accese le luci di notte.".*

Cosa pensa di questa osservazione? È in contrasto con la sua espressa nel punto precedente oppure no?

A mio avviso la legislazione in vigore, pur essendo "generica", ossia basata su principi generali, è una buona regolamentazione di base poiché si adatta alle nuove tecnologie. Naturalmente, presuppone uno sforzo di interpretazione che a volte può sfociare in una certa insicurezza giuridica. La stessa non è però tanto imputabile alla norma legale, quanto alle norme di applicazione e alla parziale assenza di direttive e codici di condotta chiari ed esaustivi. A questo problema la nuova legislazione porrà rimedio sostenendo la promulgazione di codici di condotta delle associazioni interessate e di direttive dell'autorità di controllo.

Una revisione totale della Legge federale sulla protezione dei dati è in corso: quali principi/elementi sarà indispensabile introdurre per tornare ad essere al passo coi tempi?

La revisione, svolta sul modello europeo, introduce specifici processi interni aziendali di tutela dei dati personali, nuovi reati penali particolarmente dissuasivi nel caso di violazione della legge, come pure poteri ampi d'inchiesta in capo all'Incaricato federale della protezione dei dati. Si tratta di un

cambiamento totale di paradigma. La protezione dei dati deve entrare nella cultura aziendale di ogni realtà, piccola o grande che sia.

La GDPR (General Data Protection Regulation) entrerà in vigore nell'Unione Europea nel mese di maggio 2018. Quali sono i suoi punti forti e questi potranno interessare anche noi?

I punti forti della GDPR, oltre a quelli menzionati sopra (ripresi dal progetto svizzero), sono legati alla possibilità data alle Autorità garanti di protezione dei dati di emettere sanzioni pecuniarie elevatissime (fino al 4% della cifra d'affari mondiale della società che opera illecitamente). Per le aziende svizzere, la GDPR è potenzialmente rilevante, considerato che essa si applica anche alle persone all'estero nei seguenti casi:

a) offerta di beni o la prestazione di servizi ad interessati nell'Unione, indipendentemente dall'obbligatorietà di un pagamento; oppure
b) monitoraggio del comportamento nella misura in cui tale comportamento ha luogo all'interno dell'Unione.

Il semplice utilizzo di un *cookies* di tracciamento / profilazione sul sito svizzero, ad esempio, può determinare l'applicabilità della GDPR.

Ritiene che la questione delle penalizzazioni economiche, oggi irrisorie per i *giganti della comunicazione* (Facebook, Instagram,…), potrà essere un argomento centrale?

Le sanzioni previste dal GDPR sono tutt'altro che irrisorie e fanno paura ai colossi del web. Per contro, le previste sanzioni svizzere (al massimo CHF 250'000.- di multa) avranno verosimilmente un effetto deterrente molto limitato.