Università
Ca'Foscari
Venezia

Corso di Laurea
magistrale
In Economia Aziendale

Tesi di Laurea

La rivoluzione dei Big Data

Relatore
Ch. Prof. Giovanni Favero

Correlatrice
Ch.ma Prof.ssa Silvia Avi

Laureanda/o
Diana Rocchi
Matricola 743325

Anno Accademico
2019 / 2020

Indice

Introduzione

Ho intitolato questa mia tesi *La Rivoluzione dei Big Data* perché credo che il loro effetto nell'economia, nella cultura e nella società in cui viviamo sarà dirompente, forse anche più della globalizzazione.

Ad un livello mediatico, le tre principali V dei Big Data, Variabilità, Velocità, Volume, danno un'idea di "onniscienza" informativa. In realtà sono prima di tutto tre impegnative sfide in termini, rispettivamente, di capacità di analisi, di tempestività e di potenza computazionale.

I Big Data non sono che dati e quindi di per sé stessi non rappresentano una "realtà". Essendo Big, invece, creano molto "rumore" e ridondanza informativa. Tanto più che sono dati secondari, estratti cioè dalle tracce che lasciamo con l'utilizzo di vari dispositivi e dalle varie attività online che compiamo nel quotidiano. La distanza dei Big Data dai dati primari ottenuti con una survey richiede quindi una forte analisi critica con le competenze, le teorie e le metodologie tipiche della scienza sociale. Senza queste, un algoritmo di analisi può risultare facilmente errato, discriminatorio, pregiudizievole, non efficace. In più ci si sta sempre più rendendo conto che mancano tanti dati ai Big Data, tanto da rendere alcune zone del mondo quasi invisibili. Questa continua raccolta, analisi e utilizzo dei Big Data da parte delle aziende, private e pubbliche, e dei sistemi governativi pone inoltre complicati problemi etici, morali e di privacy.

Tutto questo rende necessario ripensare alla cultura aziendale, ai modelli organizzativi di business e alle nuove figure professionali capaci di valorizzare e coordinare tutte le aree e le competenze, sia tecnico-operative che manageriali, coinvolte dalle differenti fonti di dati, le n V dei Big Data.

Capitolo 1. LA RIVOLUZIONE BIG DATA

1.1. I Big Data

Oggi sicuramente il termine Big Data apparare in articoli, blog, social ed è impossibile non averne sentito, almeno una volta, parlare, avendone cognizione o meno. Altrettanto però sicuramente in molti si domandano esattamente cosa siano. Banalmente sono i dati generati da una popolazione, in crescita esponenziale, tramite cellulari, internet e social network e che non possono che essere definiti Big, senza pretendere di dare una misura in termini di Giga, Tera o Zetta byte. Questo non spiega però l'enfasi che accompagna questo termine. Cosa sono davvero i Big Data? Sono una prospettiva grandiosa o solo un mito? Un'eccessiva esaltazione del *dato* o un angosciante futuro di pericolo per la nostra privacy? O ancora un argomento di moda trattato con leggerezza e superficialità?

Il ruolo dei dati è cambiato: dalla computabilità e precisione degli small data si passa all'incomputabilità e alla ridondanza dei Big Data. Dal dato come fotografia di un sistema passato, al dato come base su cui costruire modelli predittivi dell'evoluzione del sistema stesso. Dal dato che descrive una relazione precisa di causa-effetto al dato che evidenzia una correlazione in base a pattern emergenti.

L'enorme quantitativo di dati aumenta la difficoltà, ma soprattutto l'importanza, di filtrare e di pulire i dati, per trasformare questa grande mole in informazione e in sapere e quindi in conoscenza, attraverso l'applicazione di modelli. L'estrazione del valore dai dati deve necessariamente interagire con un insieme di componenti quali il contesto, le relazioni, l'interdisciplinarietà e quindi con la complessità.

La complessità nella sua sfaccettatura della globalizzazione o dello sviluppo di tecnologie *machine-to-machine* o delle enormi interazioni sociali virtuali o ancora della velocità, varietà e variabilità dei Big Data. In particolare la varietà dei dati produrrà necessariamente un'interazione progressiva fra discipline scientifiche diverse. Alla base dei Big Data c'è la condivisione di un'enorme quantità d'informazioni provenienti da domini differenti della conoscenza. Le correlazioni che ne deriveranno ci permetteranno di avere gli strumenti per fare predizioni in aree diverse, apparentemente lontane.

Nel momento della storia di massima accessibilità ai dati è un errore pensare ancora di prendere decisioni guidate esclusivamente dall'istinto o dall'esperienza. I dati sovrastano materialmente le capacità umane di elaborazione, ma nel contempo diventano ancora più necessari gli esseri umani con loro capacità di interpretazione.

Servono persone capaci di fare sintesi nell'enormità dei Dati, di cogliere quelli importanti e utili dall'immensa ridondanza in cui sono annegati, per aiutare i decisori a comprendere. Big data non vuol dire che i dati si organizzano da soli e che possederli dà automaticamente più conoscenza rispetto a prima.

A partire dai dati, si procede con lo sviluppo di modelli attraverso i quali definire gli strumenti di decisione basati sull'osservazione della realtà. Quindi, pur essendo rilevante la disponibilità dei dati e degli strumenti tecnologici, le competenze professionali sono fondamentali.

La sfida e la rivoluzione dei Big Data, quindi, consiste nel formare nuove figure professionali con competenze trasversali e nel creare nuovi domini del pensiero globali o almeno allargati a tutti gli attori coinvolti.

Servono quindi persone che sappiano dove utilizzare i Dati per risolvere nuovi problemi aziendali o risolvere meglio i vecchi problemi, con percorsi di analisi sempre più sofisticati e sempre meno chiusi in schemi e modelli tradizionali.

L'interdisciplinarità comporta uno sforzo congiunto, condiviso, aperto e trasparente.

Un progetto Big Data è quindi *sociale* e per questo deve avere rigidi principi etici, evitare monopoli o sfruttamento colpevole dei dati: anche l'impegno etico collettivo è Big come i suoi dati, proprio di una dimensione tale mai affrontata fino ad ora!

Come è stato per la globalizzazione, anche per i Big data si stanno delineando crisi e limiti.

Non esiste un unico modo di competere tramite i Big Data. Esistono diversi modelli e strategie possibili, anche in un'economia digitale. I Big Data possono portare con sé competitività come fallimento nei mercati. La discriminante rimane la strategia utilizzata dalle aziende per competere, la competenza professionale e il business management nell'affrontare il progetto.

Inoltre si manifestano elementi di criticità anche solo per l'incertezza e i dubbi che questa pervasiva e repentina evoluzione porta. Questa nuova economia fondata sui dati, raccolti massivamente ed istantaneamente, riutilizzati e trasmessi in continuo, accresce anche la nostra esposizione a nuovi rischi che diventano ancor più elevati, etici e di privacy, visto che riguardano le nostre relazioni, le nostre preferenze, insomma il nostro "avatar" digitale.

Oggi c'è molta propaganda pubblicitaria su Business Intelligence (BI), Big Data, Cognitive Business e Analytics. Questa confusione distrae i manager e gli analisti dalla vera valutazione del valore di questi strumenti innovativi.

Alcune indicazioni, da valutare, possono essere utili per scovare quelle che sono solo montature pubblicitarie. Ad esempio, se gli esperti di business dell'azienda non riescono a spiegare come la BI, i Big Data, l'Analytics o la Cognitive Business possano aumentare i guadagni o ridurre i costi. Oppure c'è un divario troppo grande tra visione futura e gli attuali prodotti e servizi venduti. L'azienda deve essere ad un certo livello della curva di maturità della BI per poter accedere a grandi risultati. Attenzione quindi se l'azienda, l'organizzazione, le competenze sono sovrastimate per attuare un progetto di BI. Attenzione ancora se l'interesse è solo, e non definito, per Big Data, Data Analytics o Cognitive Business.

I dati con maggior valore per la BI sono ancora oggi i dati strutturati come dati delle transazioni, i dati dei clienti e i dati finanziari. I dati non strutturati sono poco utilizzabili e soprattutto molte delle aziende tradizionali non li producono. Ancora se i software risultano troppo generici, adatti ad ogni business, non personalizzabili oppure i prodotti e i servizi venduti sembrano essere da soli capaci di aumentare i guadagni e/o ridurre i costi. La BI, i Big Data e la Cognitive Business devono essere governate dal business e non dalla tecnologia.

1.2. I Big data e la complessità

La rivoluzione scientifica del XVII secolo metteva al primo posto l'esperienza per comprendere la realtà. Esperimenti ripetibili validavano le teorie, dando una spiegazione quantitativa delle ipotesi.

Nel XVIII secolo con la rivoluzione razionale-illuministica la realtà è vista come un insieme ordinato e spiegabile attraverso principi oggettivi e lineari. Il mondo va studiato rilevando i fenomeni che sono sempre riconducibili ad un comportamento lineare. Un fenomeno che produce una trasformazione di stato è sempre reversibile. Ogni studio di un fenomeno può essere scomposto in n parti che possono essere analizzate separatamente.

La scienza classica quindi crede in un mondo semplificabile sempre, deterministico, dove i comportamenti non lineari sono eccezioni che possono, con l'elevarsi della conoscenza, essere ricondotti ad una spiegazione lineare.

Nel XIX secolo si cominciarono ad insinuare alcuni dubbi: la comprensione del mondo tramite l'analisi di microscopiche parti ricondotte tramite le stesse chiavi di lettura all'insieme macroscopico cominciava ad avere serie difficoltà teoriche e pratiche. Questo dovuto soprattutto all'aumentare continuo degli elementi costitutivi del sistema. La rivoluzione industriale fece bene notare, inoltre, la non-reversibilità di alcuni fenomeni fisici, come la termodinamica. Più precisamente alcuni fenomeni non sono irreversibili ma c'è una bassissima probabilità che possano essere reversibili.

Con la teoria della relatività di Albert Einstein vengono rivisti i concetti di spazio e tempo. Si introduce la soggettività cioè il punto di vista dell'osservatore, che influenza il processo cognitivo. L'introduzione della probabilità determina un processo di osservazione di una serie di eventi possibili, non più di un unico evento. Dirompente è stata la scoperta che le particelle della materia hanno due nature diverse, corpuscolare ed ondulatoria, che si manifestano a seconda del metodo di osservazione.

Il concetto di complessità comincia ad essere ben presente nella Scienza e nella Tecnologia: i fenomeni sono connessi e non esiste una chiara e prevedibile riconducibilità tra micro e macro, i comportamenti sono instabili, emergenti e non predicibili.

In questa fase sono di determinante importanza la visione sistemica, le risorse computazionali e il concetto di rete. L'interazione tra le parti determina una configurazione riconoscibile, le variazioni a livello micro causeranno delle modifiche nel sistema, ma la loro comprensione potrà esserci solo con una visione di insieme. L'aumento della capacità computazionale ha permesso simulazioni di complessità, multidisciplinari e sistemiche, non alla portata delle capacità umane. Le reti presentano nodi con connessioni dinamiche simili, a prescindere dall'ambito a cui si riferiscono.

"Abbiamo quindi bisogno di nuovi concetti e nuovi strumenti per descrivere una natura in cui evoluzione e pluralismo sono divenute le parole fondamentali" (Nicolas & Prigogine, 1991, citato da Camiciotti & Racca, 2017, p. 26).

I sistemi complessi hanno le seguenti caratteristiche: sistematicità, emergenza, auto-organizzazione, non linearità, connettività e adattamento.

I sistemi complessi sono fatti di un numero elevato di elementi diversi, fortemente interagenti tra loro e con altri sistemi. Gli elementi, nella loro dinamicità, creano comportamenti, caratteristiche e organizzazioni emergenti distanti e diversi da loro. Il sistema complesso non ha una gerarchia, ma si autoregolamenta. Le relazioni nel sistema complesso sono non lineari. La complessità è data soprattutto dall'elevato numero di connessioni tra gli elementi del sistema, anche se non serve un grandissimo numero di componenti. I sistemi complessi si adattano continuamente all'ambito di riferimento, cioè mantiene la sua omeostasi caratteristica.

Ai sistemi complessi non è sufficiente aggiungere la probabilità: non sono sicuramente sistemi ergodici, quindi qualsiasi stato non ha la stessa probabilità di manifestarsi; gli elementi sono a loro volta complessi e seppur in gran numero, tale numero è finito non infinito; gli esperimenti non sono ripetibili e quindi è necessaria una comprensione *in corso d'opera*.

La comprensione e la conoscenza dei sistemi complessi avviene solo attraverso i dati che possono dare un sufficiente numero di informazioni per delineare, con una buona probabilità, comportamenti e funzionamenti del sistema o almeno di alcuni suoi fenomeni. Ai dati si devono aggiungere architetture computazionali adeguate, nuovi algoritmi e nuovi modelli di analisi dei dati.

1.3. Il mondo dei dati

Il termine Big Data venne coniato da una casa editrice o meglio il primo a parlarne fu Roger Mogoulas di O'Reilly Media[1] nel 2005. Nel 2011 i Big Data salgono a celebrità con il rapporto di McKinsey intitolato *Big Data: The next frontier for innovation, competition and productivity*. Oggi è sicuramente una parola molto di moda che piace però molto di più al settore marketing di un azienda che al settore tecnico informatico.

Big Data è un termine generico e qualitativo con un grande impatto scenico che viene continuamente alimentato dall'editoria e dagli eventi dedicati. In realtà in questa parola è insita una profonda trasformazione a livello sociale, manageriale e umano. Si indica perciò non solo una vasta disponibilità di dati ma anche la capacità di acquisirli, processarli, analizzarli ed estrarre l'eventuale valore.

La società Gartner[2] nel 1995 teorizzò la curva sull'andamento delle tecnologie emergenti (Hype Cycle for emerging trends). La visibilità all'inizio cresce esponenzialmente fino al picco mediatico (hype); il calo dell'interesse scende velocemente fino ad un picco negativo della disillusione; malgrado ciò la tecnologia continua a diffondersi facendo riprendere la curva perdendo i vincoli con il marketing e iniziando la fase della produttività.

[1] O'Reilly Media è una delle case editrici più famose al mondo nel settore informatico.
[2] Gartner Inc. è una società per azioni multinazionale che si occupa di consulenza strategica, ricerca e analisi nel campo della tecnologia dell'informazione (fonte Wikipedia).

La curva di Gartner[3] tiene in considerazione il comportamento "medio" tralasciando quelli più marginali, ma che nella realtà possono avere un impatto importante. Inoltre non indica le tecnologie che non percorrono tutta la curva e si fermano prima. Dall'altra parte alcune tecnologie entrano già nel vertice della curva, quindi senza percorrerla tutta. Da tenere conto che il contesto americano è sempre in avanti rispetto all'Italia nel posizionamento nella curva di una determinata tecnologia.

Le fonti dei Big Data si possono suddividere tra i dati prodotti in maniera autonoma dalle tracce digitali di apparecchi elettronici o informatici (*machine-generated data*) e i dati prodotti dall'utilizzo degli stessi da parte degli esseri umani (*human-generated data*). Malgrado ad oggi gli human-generated data sono di molto superiori agli machine-generated data, sono quest'ultimi che nel futuro prossimo saranno probabilmente la maggiore fonte.

Alcuni fenomeni abilitanti i Big Data sono *The Always-on Consumer (AOC)*, cioè persone che utilizzano massivamente vari *device* digitali con un uso delocalizzato. Sicuramente il fenomeno dei Social Network, con la loro imponenza di utenti registrati. Gli *Open Data*, cioè dati liberamente accessibili, sono un fenomeno di grande sviluppo per l'analisi tramite i Big Data.

Le Smart City, città urbane dotate di una vasta e diffusa serie di sensori, abilitano i Big Data prodotti per analisi strategiche ai fini sociali e di servizio pubblico. Lo IoT, *l'Internet of Things*, estende la funzionalità di internet agli oggetti: il famoso frigorifero che ci dice cosa manca da mangiare! Il *Knowledge Digitalization Process*, cioè il lento ma inesorabile processo di digitalizzazione della conoscenza, è forse il fenomeno abilitante con maggiore rilevanza per i nuovi modelli di business.

I Big Data sono caratterizzati dalle, ormai famose, tre V: Volume, Velocità e Varietà. Malgrado la grande enfasi data a queste caratteristiche, la loro rilevanza appare chiara solo anche con le altre due V necessarie: Value e Veracity. Il termine Veracity si traduce nel fatto che i dati devono essere puliti, accurati e veri. Il termine Value, cioè Valore, indica invece che i dati devono avere, appunto, un valore per le finalità e per le aziende per cui sono stati estratti, quindi per la strategia di business.

[3] https://www.gartner.com/en/research/methodologies/gartner-hype-cycle.

Il valore dei Big Data è dato dagli *insight* che emergono, cioè dall'interpretazione del fenomeno, che possa indicare delle scelte e/o direzioni operative. I Big Data possono essere definiti anche tramite una convergenza di trend emergenti: l'enorme disponibilità di dati digitali, l'elevata riduzione dei costi di archiviazione, l'aumento di potenza computazionale e gli open source, cioè la disponibilità e il continuo miglioramento di soluzioni software non proprietarie.

Questo vuol dire che i pattern emergenti dai Big Data sono la base per costruire e validare i modelli ma non sono di per sé strumenti predittivi. I dati devono essere sufficientemente adeguati per tipologia, varietà e significatività per costruire modelli predittivi efficaci. I processi che portano i dati ad essere rilevanti e quindi a modelli e a decisioni comportano la creazione di altri dati, migliori.

1.4. I big data e la velocità

Diventano un must la reattività istantanea, la riduzione di ogni spreco di tempo nell'analisi dei dati, nella progettazione di nuovi prodotti, nei cicli di produzione e logistica e l'anticipazione dei trend di mercato. In realtà la velocità non è un obbligo esogeno e imprescindibile, ma deve essere frutto di una precisa scelta strategica.

Non esiste una singola risposta al problema dell'aumento della complessità dell'ambiente competitivo. Sicuramente però è necessario un cambiamento di prospettiva strategica. Il tempo non può essere solo il tramite per il raggiungimento dei tradizionali vantaggi competitivi, cioè fare le stesse cose, solo più in fretta.

Una classificazione sui diversi significati attribuiti al tempo, proposta da Frederich W. Taylor, identifica cinque categorie o concetti differenti di tempo:

• Rate: relativo a quei problemi che riguardano l'agire più velocemente, in particolare in relazione all'idea di velocità.

• Sequence: relativo a quei problemi in cui è cruciale l'ordine in se stesso o l'ordine in cui certi eventi o azioni dovrebbero avere luogo.

• Duration: riguarda la durata di un certo evento.

• Deadlines: implica l'utilizzo del tempo come un marker per determinare il limite entro cui una certa azione dovrebbe essere completata.

• Timing: si focalizza sul momento dell'azione, sul "quando" agire.

Oggi il tempo non può essere visto nella sua singola accezione, ma attraverso un mix attento di concezioni temporali diverse. La strategia ma anche la struttura dell'azienda nel suo complesso devono incorporare la nozione di tempo e fare di esso una fonte di vantaggio competitivo sostenibile. Il concetto di *Right time enterprise*, il *tempo reale*, è l'attributo strategico di una nuova tipologia di impresa: nel senso di tempo che evita gli sprechi, ma anche nel senso di tempo necessario che serve, il tempo *giusto* non il più ridotto.

Un progetto informatico su larga scala desta delle preoccupazioni giustificate: le probabilità di successo entro i limiti di tempo e di budget, raggiugendo gli obiettivi tecnici prefissati sono circa una su dieci. Il potenziale valore perso può oscillare tra il 100 e il 170% del costo totale dell'investimento (Standish Group, 1995) . Dall'altra parte l'avviamento di modelli di business digitale e di Big Data sono fondamentali per la trasformazione organizzativa. Questi progetti naturalmente richiedono importanti investimenti nell'informatica. Le probabilità di successo di un grande progetto informatico possono essere aumentate se i dirigenti sanno dove concentrare i propri sforzi.

1.5. Big Data e la correlazione

L'avvento dei Big Data ha certamente scardinato la pratica del campionamento ma soprattutto la relazione di causalità tra fenomeni a favore della correlazione.

Il campionamento è stato effettuato per anni come soluzione migliore per estrarre informazioni dalla realtà, dato che l'osservazione del tutto era impossibile. I Big Data naturalmente non sono la totalità ma sono sufficientemente ampi, contenendo anche quei dati che inizialmente sembrano poco significativi. Il peso degli errori puntuali e localizzati quindi viene annegato nella mole dei dati che permette comunque un risultato finale soddisfacente.

L'analisi dei Big Data non cerca una relazione di casualità tra i fenomeni, ma di correlazione. La relazione tra i fenomeni è, quindi, individuata ma non spiegata. Naturalmente è rilevante il grado di correlazione per individuare i percorsi all'interno dei dati.

L'assenza di un campionamento e l'individuazione di correlazioni, quindi di *pattern* nascosti, devono essere solo la base per lo studio e la progettazione di modelli predittivi. Si passa quindi dai tanti dati ai dati esaustivi in termini di significato, validità e varietà per poter costruire un modello.

La correlazione identificata dagli algoritmi di *Data Mining* può raggiungere un grado di successo quasi del 100%, ma anche essere basata su elementi non prioritari e/o esplicativi di ciò che si sta analizzando. Famoso il caso di un algoritmo che riconosce tra un lupo e un cane con successo ma che si basa sull'elemento neve presente o no nello sfondo della foto. Si capisce bene che anche se il risultato è pressoché sempre vero, può essere altrettanto estremamente fuorviante perché non è basato su fattori distintivi ed esaustivi.

L'applicazione di algoritmi nella valutazione delle persone, come ad esempio i candidati ad un nuovo impiego, crea un preoccupante problema di discriminazione. Da qui la necessità dell'apporto delle nuove figure professionali come il *Data Scientist* con competenze interdisciplinari e del miglioramento degli algoritmi attraverso una condivisione e un lavoro di team che permetta di eliminare *bias*, errori sistematici, e discriminazioni, di essere indipendenti dalle attese e dall'interpretazione soggettiva.

1.6. Big Data e la Data Science

Fondamentalmente i Big Data necessitano di un passaggio dalla Statistica alla Data Science. Partiamo dai dati da analizzare. L'approccio statistico prevede una fase di rilevazione che si concretizza nella scelta di un campione in base al fenomeno da analizzare, nella determinazione dei caratteri della popolazione, nella raccolta e classificazione dei dati.

La determinazione di un *dataset* nel caso dei Big Data prevede un'attività molto più dispendiosa in termini di tempo. I dati sono abbondanti, ma possono essere *sporchi* e ridondanti. Altrettanto difficile è l'analisi dei dati validi per il problema di business che si vuole risolvere o indagare. Inoltre i dati diventano velocemente obsoleti.

La modalità di lavoro cambia nettamente: lo statistico lavora da solo, il Data Scientist non può che lavorare in team. Tale squadra deve essere inoltre variegata per competenze e professionalità per poter analizzare dati con così alta varietà.

L'obiettivo della Statistica è descrivere un fenomeno, mentre la Data Science si focalizza su modelli predittivi e prescrittivi. Questo focus richiede anche nuovi sistemi di visualizzazione dei dati che non possono più essere esposti su *report* e *dashboard* statiche, ma su *tool di data visualization* dinamici e interattivi.

Gli strumenti della Data Science fanno comunque parte della scienza statistica, ma devono essere affiancati al Machine Learning e alla Network Science.

La *Machine Learning* prevede algoritmi che possono estrapolare automaticamente un comportamento dai dati: la classificazione dei dati tramite classi note o non note; la regressione cioè la predizione dei valori futuri tramite serie storiche; il clustering di dati non etichettati per creare gruppi per similarità o altre caratteristiche strutturali.

La *Network Science* analizza i dati strutturandoli in nodi e relazioni: la community detection, cioè nodi con elevate relazioni all'interno di un cluster; il percorso minimo è la relazione più breve tra due nodi; la misura della centralità di un nodo all'interno della rete.

Capitolo 2. ANALISI CRITICA CON I BIG DATA

2.1. L'analisi critica

Ci siamo appena immersi, nella realtà, in questa rivoluzione dettata dai Big Data, in questo diluvio di dati, in questo *data lake*. La novità del fenomeno permette che ci siano continui esperimenti, diverse opinioni od impostazioni. E' fondamentale però cominciare ad innescare discussioni, ponendoci domande critiche verso tutti gli aspetti di un fenomeno con effetti così dirompenti nella cultura, nell'economia e nella società odierna.

Lessig (1999, citato da Boyd & Crawford, 2012, p. 664) sostiene che i sistemi sociali sono regolati da quattro forze: mercato, legge, norme sociali e architettura. I Big Data mettono in contrasto queste quattro forze. Il mercato vede solo la pura opportunità, la legislazione frena la raccolta e la conservazione dei dati e le norme sociali si scontrano con nuove questioni etiche. L'architettura giuridico-sociale è quindi scardinata dai Big Data: l'organizzazione, la "natura" di un contesto sociale, non può più essere definita dagli stessi valori, dagli stessi scopi e dalle stesse norme di prima.

Nella ricerca critica con i Big Data ci sono alcuni quesiti fondamentali a cui va data una risposta: "Possiamo considerare questi dati come attendibili, rappresentativi e validi? Analizzare questo materiale è lecito da un punto di vista etico? E la loro diffusione segnerà il tramonto delle tecniche tradizionali?" (Lombi 2015, p. 217)

La rappresentatività dei Big Data sembra scontata dall'enorme quantità di dati di cui sono composti. I ricercatori, invece, sanno bene che la numerosità campionaria non è condizione sufficiente, ma spesso nemmeno necessaria, per garantire la rappresentatività della base dati. Una ricerca rigorosa ha le fondamenta in un approccio sistematico alla raccolta e all'analisi dei dati. Questa impostazione oggi però sembra troppo dispendiosa in termini di tempo e costi. Avere a disposizione tanti dati non può significare che i metodi, e in particolare l'inferenza statistica, abbiano perso la loro rilevanza.

Se ci limitiamo ad analizzare i Big Data sotto il profilo della quantità di dati non ci troveremmo certo davanti ad un nuovo tipo di ricerca. I censimenti e alcune immense ricerche sono state realizzate anche prima del novecento. Le novità dei Big Data sono invece la difformità nella tipologia, la loro relazionalità intrinseca, la flessibilità, l'estendibilità, la scalabilità e la diffusione. Queste caratteristiche sono la vera sfida per le ricerche con i Big Data. E' quindi la comprensione del "tipo" di archivio che ci si trova a utilizzare il punto di partenza fondamentale per la ricerca.

I Big Data non possono essere rappresentativi di per sé perché in realtà mancano di completezza. Esiste ancora un elevato digital divide nel mondo di oggi: molte persone sono invisibili alla digitalizzazione, come gli emarginati, i poveri, gli ammalati e le donne in molti paesi del mondo. E se anche per tutti, il web o i social network fossero accessibili, ciò non garantirebbe comunque la rappresentatività dei dati. Molti utenti infatti sono *lurker*, leggono passivamente senza commentare, non lasciando quindi nessuna traccia digitale. Twitter Inc. ha rivelato che il 40% degli utenti attivi frequenta le piattaforme social solo per visualizzare (Twitter 2011). La ricerca invece necessita di tracce digitali, di dati utili (rappresentativi) alle domande che si pone.

Dall'altra parte i Big Data permettono ricerche su popolazioni sommerse, perché abbiamo dati che difficilmente sarebbero recuperabili con i metodi tradizionali come ad esempio i sondaggi. Possiamo avere dati da persone che non vogliono rivelare la propria identità per particolari condizioni sociali (es. prostitute, tossicodipendenti, ecc.).

L'attendibilità, cioè la riproducibilità del risultato, nei Big Data deve essere attentamente valutata. Gli algoritmi, che a molti sembrano essere il fondamento proprio dell'attendibilità, vengono definiti da Lupton (2015, citato da Lombi, 2015) come "liquidi, permeabili e mobili". I Big Data cioè non sono oggettivi e nemmeno neutrali. Sono la creazione della società digitale globalizzata che influenzano a loro volta. Contengono molte informazioni non veritiere perché create ed organizzate dagli utenti per dar forma ad un modello di desiderabilità sociale. In più, tweet e profili possono essere spesso falsi perché creati da robot o per fini commerciali e sono difficilmente riconoscibili per essere eliminati dal campione di ricerca.

La problematica più rilevante per l'attendibilità resta comunque l'interpretazione umana, soggettiva: il centro del processo di comprensione passa attraverso l'identificazione di correlazioni false, distorsioni ed ambiguità semantiche. "Un modello può essere matematicamente valido, un esperimento può sembrare valido, ma non appena un ricercatore cerca di capire cosa significhi, il processo di interpretazione è iniziato. Questo non vuol dire che tutte le interpretazioni umane sono fallaci, ma piuttosto che non tutti i numeri sono neutrali." (Boyd & Crawford 2012, p, 667).

La validità dei Big Data, cioè il rapporto tra teoria e empiria, tra misura e concetto, è un aspetto controverso. Innanzitutto ci stiamo spostando sempre più da una analisi *theory (knowledge) driven*, governata dal presupposto teorico, ad un approccio data driven, cioè guidato dai dati. Quest'ultimo può essere fortemente minato da *availability bias*, cioè giustificazioni o teorizzazioni ex post del risultato dell'analisi data driven. I Big Data non sono diversi dagli Small Data in merito al fatto che è necessaria un'analisi teorica. "La dimensione dei dati dovrebbe adattarsi alla domanda di ricerca che viene posta; in qualche caso, piccolo è il migliore." (Boyd & Crawford 2012, p. 670).

Molta attenzione va data al contesto per avere dati validi. Le connessioni devono essere valutate sulla base di una serie di misure, anche se magari con variabili diverse. Oggi facilmente un contesto è definito più dai *like* che, ad esempio, dalle tradizionali variabili demografiche. I dati in rete devono essere contestualizzati perché risentono degli effetti della piattaforma che li ospita e perché sono legati allo spazio-tempo in cui sono stati inseriti. Se i dati vengono ridotti ad un modello matematico facilmente perderanno il loro contesto e quindi la loro validità.

Le diseguaglianze, in merito all'accesso sia ai Big Data in generale che alle analisi su questi dati, hanno come risultato la creazione di una cultura ristretta della ricerca (Boyd & Crawford 2012, p. 669). Da una parte esiste un gap di competenze in materia di Big Data tra maschi e femmine, tale che i Data Scientist sono quasi tutti uomini. Questo fa sì che ci sia un possibile pregiudizio su chi pone le domande e perciò su quali domande di ricerca verranno poste e in che modo saranno formulate (Harding 2010; Forsythe 2001). Dall'altra parte spesso i ricercatori lavorano su basi dati incerte per completezza ed attendibilità, data l'ampia discrezionalità di chi fornisce i dati. La proprietà di questi dati

è infatti nelle mani di poche grandi società che non hanno alcuna imposizione nel renderli disponibili, hanno perciò un controllo totale sulla visibilità e sulla possibilità di analisi.

La questione etica è grande come i Big Data, perché il dato è molto accessibile ma ciò non può essere confuso con un utilizzo non etico. Deve esistere una tutela del dato confidenziale, della privacy e perfino del diritto all'oblio.

L'etica nella ricerca, nell'utilizzo di questi dati è un tema di responsabilità. Tutte le norme e le restrizioni dettate dalla privacy non possono difenderci dall'utilizzo non etico dei nostri dati. Troshynski et al. (2008, citato da Boyd & Crawford, 2012, p. 672) parlano di un concetto di responsabilità che deve essere applicato anche quando non ci sono limiti secondo le norme e la privacy. "Gli utenti non sono necessariamente a conoscenza di tutti i molteplici usi, profitti e altri guadagni derivanti dalle informazioni che hanno pubblicato." (Boyd & Crawford 2012, p. 673). Va costruito un network per la privacy, con un controllo sulla diffusione e sull'interpretazione, perché non può più essere gestita individualmente, tramite restrizioni all'accesso.

Ripartiamo quindi dalla base di una ricerca e cioè la scelta delle fonti: per i Big Data si tratta di selezionare i dati e i loro metadati. In quanto ai dati le principali fonti dei Big Data sono:

- *transactional data*, cioè informazioni raccolte nell'ambito degli scambi tra cittadini e amministrazioni, e tra consumatori e aziende
- *digital by product data*, cioè dati creati e inseriti dagli utenti attraverso le piattaforme social o di e-commerce
- *self-tracking e self-reporting data*, cioè dati raccolti dalle App dove gli utenti registrano i vari stati fisici o emotivi o il monitoraggio degli spostamenti, attraverso i sistemi di geo-localizzazione
- *digital texts*, ossia articoli, testi e altri contenuti in formato digitale.

Savage e Burrows (2007, citato da Boyd & Crawford, 2012, p. 664) sostengono che le scienze sociali devono accettare la sfida della crescente digitalizzazione della

comunicazione utilizzando i transactional data (soprattutto se prodotti dai social media) e identificando nuovi e più appropriati metodi e tecniche d'analisi.

La centralità dei transactional data per la divulgazione della scienza e della tecnologia, la Public Communication of Science and Technology (PCST), è dovuta soprattutto all'indispensabile utilizzo dei social media nella comunicazione e alla presenza intrinseca di scienza e tecnologia nella comunicazione digitale.

I transactional data sono prodotti di continuo e, naturalmente, senza stimoli esterni e senza uno scopo specifico, non quindi per una domanda di ricerca particolare. A queste caratteristiche positive si affiancano aspetti che ne limitano la rilevanza per la ricerca sociale.

Le politiche di forte privatizzazione, e la tutela di privacy, vera o falsa che sia, limitano molto la disponibilità dei transactional data per l'attività di ricerca. Lo sviluppo recente di questo tipo di dati non permette peraltro la necessaria distanza temporale che deve essere mantenuta dalla ricerca sociale rispetto all'oggetto di analisi. Sono infatti dati molto legati, nelle loro caratteristiche e variazioni, ai cicli della moda. Le utenze possono non essere rappresentative perché ci sono diversi tipi di utilizzatori e le loro abitudini possono cambiare nel tempo.

Sono dati genuini ma non neutrali, sia per le interazioni con le piattaforme che li contengono e sia per gli stessi algoritmi che li creano. E come per tutti i dati sociali, sono influenzati dalle convinzioni e dai valori degli stessi utenti.

Oltre ai transactional data, ultimamente anche il text-mining, cioè l'analisi del testo automatica, sta diventando uno strumento fondamentale per la ricerca. Questo per l'enorme quantità ed accessibilità di testi digitalizzati di origine completamente diversa (interviste, discorsi pubblici, questionari, forum online, blog, articoli di giornali, commenti). A ciò va aggiunto lo sviluppo di nuovi modelli matematici e di algoritmi, sofisticati ma user-friendly, e la grande potenza di calcolo disponibile ad un costo ridotto che permettono la raccolta, l'estrazione, la visualizzazione, l'analisi e l'interpretazione di un'enorme mole di documenti (Castelfranchi, 2017, p.2).

L'analisi del testo automatizzata permette prima di tutto di fare "pulizia": eliminare tutte le preposizioni, gli articoli, gli avverbi e gli aggettivi che non sono rilevanti per la ricerca

o termini che sono semplicemente sinonimi, ma declinati al plurale o al singolare, al maschile o al femminile.

L'analisi computazionale può essere utilizzata in diversi modi e momenti, variando a seconda dei dati disponibili, dell'approccio teorico deduttivo o induttivo utilizzato, degli obiettivi e delle ipotesi di ricerca e dall'utilizzo di una codifica del testo manuale o automatica.

In un tipico approccio deduttivo la codifica del testo è manuale, come pure le categorie che sono definite a priori. L'attività dei ricercatori e dei codificatori è quindi fondamentale ma può comunque essere migliorata dall'attività computazionale organizzando, visualizzando e strutturando in tabelle e data-base i dati codificati.

Nel caso in cui vengano elaborate manualmente dai ricercatori solo le categorie per una codifica automatica, è necessario creare un *codebook* gerarchico per categoria di analisi. Nel codebook deve essere dettagliato precisamente cosa, come e quando codificare. Questo elenco va affinato continuamente affinché gli algoritmi possano codificare esattamente. Quando il corpus del testo è ampio questa è ancora una strada difficilmente percorribile e forse anche poco affidabile per le continue nuove codifiche.

Un classico approccio induttivo, invece, si ha quando le categorie di analisi non sono state definite a priori, ma vengono create automaticamente dagli stessi dati identificando schemi, correlazioni, temi o concetti ricorrenti. Si potrebbero quindi individuare categorie non incluse nell'ipotesi della ricerca perché la categorizzazione automatica ignora completamente il contesto. Il text mining però è proprio questo, un'analisi induttiva, statistica della semantica dei testi e della loro struttura (Wiedemann, 2013, citato da Castelfranchi, 2017, p. 5). Coadiuvato dalla competenza e dall'interpretazione dei ricercatori, questo approccio può essere indirizzato ed addestrato per estrarre effettivamente un significato di contesto.

L'utilizzo della digitalizzazione dei media tradizionali sembra quindi una strada più sicura perché si tratta comunque di dati prodotti "naturalmente", che hanno però minori barriere di accesso e una vita di medio lungo periodo (sono disponibili archivi di articoli dei principali quotidiani anche dai primi anni '90 del secolo scorso). È assodato che anche i media non producono dati neutrali ma anzi condizionano la realtà sociale e sono condizionati dal contesto in cui operano. Inoltre gli articoli sono scritti e pensati da giornalisti, quindi specialisti e non cittadini qualunque (Pitrelli, 2017).

Per questo tipo di fonte, i documenti digitalizzati, un breve accenno al progetto TIPS può dare qualche spunto di riflessione.

TIPS (*Technoscientific Issues in the Public Sphere*) è un progetto promosso da PA.S.T.I.S.[4] che vuole sviluppare e sperimentare procedure automatiche per l'acquisizione, la classificazione e l'analisi di contenuti digitali, principalmente quotidiani e social networks. In particolare la ricerca vuole analizzare la presenza e l'evoluzione mediatica della scienza e della tecnologia. Il progetto TIPS è stato avviato come un'evoluzione del progetto SMM (*Science in the Media Monitor*), realizzato da Observa[5]. L'infrastruttura di TIPS è stata progettata per raccogliere e analizzare quotidianamente articoli di giornale, lasciando ai ricercatori un'ampia possibilità di analisi sugli argomenti di interesse.

TIPS utilizza una piattaforma a struttura modulare per raccogliere e organizzare un archivio delle testate più diffuse e rappresentative di un paese, accessibile via web. Per l'Italia il Corriere della Sera, La Repubblica, la Stampa, il Sole24Ore, Avvenire, il Giornale, il Messaggero e il Mattino; per la lingua inglese NYTimes, Guardian, Mirror, Telegraph, Times of India; per la lingua francese Figaro, Lacroix, Le Monde, Les Echos, Liberation, Parisien. Utilizza anche circa 100 blog e 100 accounts di Twitter rappresentativi per la lingua italiana.

Alcuni numeri per capire l'enorme quantitativo di dati presente nell'archivio della piattaforma TIPS già all'inizio del 2017: oltre 1.300.000 articoli pubblicati dai quotidiani italiani, oltre 750.000 articoli in inglese, oltre 700.000 articoli in francese, oltre 500.000 post pubblicati sui blog e alcune migliaia di tweets. Inoltre un campione di 162.000 articoli pubblicati dal Corriere della Sera e dalla Repubblica nel periodo 1992–2012.

L'archivio deve essere accessibile, innanzitutto per formato, con dati riconoscibili, comprensibili ed utilizzabili. Deve contenere dati sufficienti per l'analisi del fenomeno ma soprattutto deve essere facilmente integrabile con altre fonti nel momento in cui ci si accorge che la base dati non è esaustiva per la ricerca.

TIPS è un progetto multidisciplinare con ricercatori di sociologia, ICT, statistica, psicologia sociale e linguistica perché vuole anche utilizzare le potenzialità derivanti

[4] Padova Science Tecnology and Innovation Studies è un centro di ricerca per lo studio sociale dei media, della scienza, della tecnologia e dell'innovazione dell'Università di Padova

[5] Observa è un centro di ricerca indipendente, senza fini di lucro, che promuove la riflessione e il dibattito sui rapporti tra scienza e società, favorendo il dialogo tra i ricercatori, policy makers e cittadini. www.observa.it

dalla digitalizzazione dei quotidiani per indagare *the social life of data*, il rapporto fra tecnoscienza e società senza usare i transactional data (per i limiti visti sopra). Le conoscenze, le teorie e le metodologie di ricerca sono anche esse interdisciplinari e in particolare in TIPS si utilizza la *science and technology studies (STS)*, la *content analysis*, la *social representations theory* e la *computer science*[6].

TIPS utilizza una classificazione automatica tramite categorie ed indici, sviluppati dai ricercatori. Le classificazioni vengono continuamente riciclate per adattarle ai nuovi insigth e sviluppare nuove metodologie di elaborazione e di analisi automatiche. Ad esempio, utilizza un algoritmo che, attraverso la presenza e la frequenza di termini chiave nel testo dell'articolo, attribuisce loro un punteggio che, se superiore ad una determinata soglia, categorizza l'articolo come significativo per contenuto tecnoscientifico. Gli indici già disponibili in TIPS sono i seguenti:

- *salience* cioè il rapporto fra il numero di articoli inerenti a un argomento di ricerca e il totale degli articoli pubblicati nello stesso periodo e dalla stessa fonte;
- *prominence* che indica lo stesso rapporto dell'indice di salience, ma riferito alla pubblicazione sulla home-page;
- *general framing* cioè la distribuzione degli articoli riferiti alla tematica di ricerca nelle varie sezioni dei quotidiani, come ad esempio nella pagina di cronaca piuttosto che in quella di economia
- *risk* che monitora la presenza nel testo di un articolo, identificato come pertinente alla tematica di ricerca, di un insieme di parole associate alla valutazione e alla percezione del rischio.

Nell'indice di salienza degli articoli con un significativo contenuto di tecnoscienza, pubblicati da Corriere della Sera, Repubblica, Stampa e Sole24Ore dal 2010 al 2016, è visibile una certa stabilità secondo i parametri di normale oscillazione definiti da TIPS[7]. Questo malgrado il notevole aumento degli articoli sulla tecnoscienza nell'ultimo

[6] STS sono studi che ricercano le relazioni tra ricerca scientifica ed innovazione tecnologica con la società, la politica e la cultura. Content analysis cioè l'analisi del contenuto attraverso tecniche manuali o automatiche per contestualizzare i dati. Social representations theory rappresenta la società tramite una serie di valori, idee, metafore, credenze e pratiche condivisi tra i membri di gruppi e comunità.
[7] La zona di normale oscillazione definita dal TIPS è compresa tra la media più la deviazione standard e la media meno la deviazione standard.

periodo. L'oscillazione si può definire fisiologica e possiamo quindi dire che la tecnoscienza ha una presenza consolidata nella nostra società (Neresini, 2017, p, 6).

Alcuni picchi presenti sono spiegati da eventi precisi: il picco di marzo 2011 è dovuto soprattutto all'incidente della centrale nucleare di Fukushima e alla presentazione del nuovo iPad. Altri picchi invece si collegano ad eventi significativi ma non così rilevanti: sono quindi la somma di tante singole notizie, come ad esempio il rientro dalla stazione spaziale dell'astronauta Parmitano, la controversia relativa al caso Stamina, il conflitto sui brevetti fra Samsung e Apple, pubblicate nel Novembre del 2013. Il picco di Novembre 2015 scaturisce da eventi come la conferenza sul clima di Parigi e il dibattito sulla destinazione come cittadella della scienza dell'area Expo a Milano.

L'andamento del risk indicator su articoli selezionati sulla base di differenti criteri mostra, invece, un trend discendente nel tempo, ma che non deve essere interpretato come specifico della tecnoscienza, bensì generale ed intrinseco dell'indice di rischio. Comparando scienza e tecnologia si vede chiaramente che gli andamenti sono molto simili, anzi quasi sovrapponibili: la percezione pubblica del rischio non è molto distante quindi tra scienza e tecnologia (Neresini, 2017, p, 8).

Oltre alla definizione ed alla analisi degli indici anche la determinazione del *corpus*, cioè quello che si vuole analizzare, implica alcune altre problematiche. Il reperimento degli articoli incontra barriere all'accesso e problemi di formati (ad esempio i vecchi quotidiani sono archiviati come immagini, difficilmente utilizzabili per l'analisi automatica del contenuto).

La selezione di un oggetto d'analisi pertinente rispetto alla domanda di ricerca è altresì complicato, in particolare quando non è su un tema specifico come nel caso di TIPS. L'analisi di tipo comparativo aiuta molto a distinguere se una caratteristica è rilevante, distintiva per l'analisi. Ad esempio l'indicatore di salienza del TIPS ci ha rivelato che l'aumento di articoli sulla tecnoscienza non equivale ad una maggiore rilevanza pubblica della stessa.

Anche la stessa procedura operativa, per la costruzione del campione di analisi, deve essere attentamente valutata. La soluzione identificata da TIPS, ad esempio, è la ricerca dei risvolti pragmatici della tecnoscienza all'interno di un articolo: l'attività di text-mining per selezionare le fonti in TIPS è basata sulla definizione stessa di scienza. Le mie sottolineature aggiunte alla seguente definizione di scienza individuano i criteri di

selezione delle fonti in TIPS: “la scienza è un’attività sociale, dunque fatta da qualcuno, all’interno di un’organizzazione, mediante l’utilizzo di alcuni strumenti, strutturata in campi disciplinari con una specifica nomenclatura; è inoltre un’attività che produce contenuti trasmessi mediante riviste o in particolari occasioni” (Neresini, 2017, p. 9). C’è da domandarsi sempre e da subito se i criteri utilizzati per la selezione forniranno un corpus valido e rappresentativo per le domande di ricerca.

Identificare la granularità ottimale, cioè l’unità temporale giusta per strutturare, analizzare e visualizzare i dati, non è facile. Sicuramente bisogna tener conto che la scelta incide non poco su quello che si può osservare e quindi sull’efficacia dell’analisi.

La periodizzazione normalmente è in linea con l’arco temporale definito dalla granularità. Nel caso dei media è più opportuna però una periodizzazione collegata alla dinamica della comunicazione mediatica, perché questa può essere ciclica o puntuale. TIPS ha adottato la zona attorno al valore medio dell’indice di salienza come range di normale oscillazione. Come la periodizzazione non può essere fatta per periodi fissi, anche la costruzione dell’oggetto di analisi tramite campionamento potrebbe portare effetti ampiamente distorsivi (come ad esempio un campione prelevato nei giorni di picco dell’indice di salienza).

Validità: “Come possiamo infatti essere ragionevolmente sicuri che stiamo osservando proprio il fenomeno che ci interessa?” (Neresini, 2017, p.10).

La verifica della validità non si può effettuare automaticamente su grandi quantità di dati. E’ necessaria una analisi da parte dei ricercatori, manualmente con test che si concentrano su un numero piccolo di dati.

Perciò “nessuna tecnica di trattamento automatico è in grado di risolvere il problema della validità se non passando attraverso il giudizio dei ricercatori, il quale, inevitabilmente, introduce una serie di scelte, assunti e opacità di carattere epistemologico. Non è possibile fare diversamente, e la consapevolezza di tale ineludibilità non ci dovrebbe mai abbandonare, nonostante l’apparente oggettività che viene facile attribuire agli strumenti di analisi automatica” (Neresini, 2017, p.11).

Il confine tra qualità e quantità si fa più sfocato con i Big Data, costringendo i ricercatori ad utilizzare metodi misti e a lavorare in gruppi interdisciplinari.

Questi nuovi strumenti non sono nuove metodologie. Trovare nuove correlazioni, descrivere e visualizzare un fenomeno grazie ai Big Data non vuol dire aver trovato anche le variabili causali che spiegano un fenomeno sociale o il suo funzionamento. Senza modelli, siano essi concettuali o matematici, i dati, secondo Massimo Pigliucci (2009, p. 534, citato da Castelfranchi, 2017, p. 8), un filosofo della scienza, non sono altro che rumore, la scienza avanza solo quando può fornire spiegazioni. Non abbiamo inoltre adeguate metodologie per gestire gli errori o i pregiudizi nell'automazione. Dall'altra parte corriamo il rischio di vedere solo ciò che stiamo già cercando sapendo esattamente cosa il software è programmato per rilevare (Wiedemann, 2013 citato da Castelfranchi, 2017, p. 5).

Questi dati, queste fonti, sono altresì corredati dai metadati cioè dai dati che contestualizzano i dati. La ridondanza è necessaria perché i Big Data sono anche i metadati, necessari per organizzare, catalogare e consultare le informazioni raccolte.

Contestualizzare la ricerca qualitativa è una attività doverosa da sempre. E' sempre stato improbabile che gli analisti, i ricercatori secondari, potessero comprendere l'esperienza primaria del ricercatore senza l'ausilio dei metadati. Il rigore di una ricerca nel definire i propri metadati è anche un atto di responsabilità verso le persone coinvolte e le possibili conclusioni ottenute (Mills, 2018). Di chi sarà la responsabilità, ad esempio, in un progetto di Data Analytics con i Big Data condotto secondo tutte le norme di privacy, ma che con lo sviluppo e il ri-utilizzo dei dati porta all'identificazione di un soggetto?

Ancora più rilevante è ormai la certezza che senza metadati pensati e strutturati sarà impossibile la trasmissione e condivisione del sapere in futuro o almeno nel medio lungo periodo. I dati digitali sono oggi espressioni culturali e sociali, non più solo strumenti di semplificazione amministrativa o di conservazione alternativa di documenti cartacei. Un aspetto questo che da poco è stato preso in considerazione, ma che adesso risulta più importante che mai.

La Digital Cultural Heritage (DCH) non è facilmente identificabile nei Big Data e tantomeno è semplice strutturare i metadati necessari per la conservazione storica. Quanti e quali metadati tenere dell'oggetto digitale? Secondo Barbuti (2019, pag. 125) è indispensabile la giusta proporzione nella struttura dei metadati tra configurazione quantitativa e qualitativa: cioè quanti e quali dati sono sufficienti per un livello

equilibrato di informazione sulla risorsa e il suo ciclo di vita. La combinazione elementi/descrizioni, di dati/metadati, deve far conoscere i contenuti progettuali dell'archiviazione, la struttura dei metadati, il processo operativo utilizzato e le scelte effettuate. Inoltre è necessario, con lo stesso rigore, dare informazioni sulle modifiche e variazioni nel corso della "vita della risorsa" archiviata. Questa è la base fondamentale per un corretto ed agevole utilizzo e riutilizzo, adesso e in futuro, per la ricerca e anche per la consultazione in genere.

Le componenti descrittive sono "elementi fondanti per validare e certificare il dato, garantendone qualità, autenticità e leggibilità per le future generazioni." (Barbuti, 2019, pag. 135). Aggiungiamo che sono allo stesso modo fondamentali per definire il contesto del dato e quindi una ricerca critica attraverso i dati digitali, riconoscendo gli elementi di soggettività e di interpretazione comunque insiti anche nei Big Data.

E' necessario sviluppare un nuovo quadro teorico per comprendere e aumentare la percezione dei dati sottostanti ai Big Data. Lugmayr A. et al. (2017) parlano di Big Data cognitivi. Si tratta di introdurre nuove discussioni sui Big Data, affrontandoli come un nuovo sistema socio-economico-tecnico, in base al livello di analisi, senza dimenticare che anche la visualizzazione, la presentazione e l'interpretazione dei risultati sono attività cognitive.

Adottare questo approccio significa abbandonare l'attenzione sull'aumento della produttività, evitando di concentrarsi solo sulla scalabilità dei dati con il solo obiettivo di aumentare il rendimento, la quantità dei dati e i sistemi tecnici per la lavorazione dei dati.

Raccogliere il più possibile dei dati correlati, in molte forme e tipologie, non esclude che ci possano essere dubbi sulla rappresentabilità dei dati raccolti, specialmente in relazione al suo contesto applicativo, ma anche rispetto alla consapevolezza dei dati realmente disponibili. Nei campioni statistici, pur essendo molto più limitati in quantità e varietà, la rappresentatività è resa più certa dai metodi sviluppati negli anni per identificare e correggere la distorsione nei campioni selezionati. L'osservazione completa di un sistema è di solito possibile solo con una domanda di ricerca ristretta e utilizzando dati interni.

La completezza dei dati è quindi soggetta a errori a causa di giudizio umano durante il processo di definizione di un modello di sistema, dei suoi parametri e delle sue limitazioni per un particolare contesto di applicazione.

I Big Data cognitivi sono analizzati come sistema socio-tecnico, nel giusto contesto applicativo, con la giusta granularità: i dati devono poter essere riclassificati in dati che possono essere ignorati, dati non conosciuti e dati che possono essere pienamente utilizzati in processo di analisi. I *dati chiari* sono quindi i dati disponibili ma che comunque devono essere selezionati in base alla domanda di ricerca. I *dati grigi* invece sono i dati non disponibili ma che possono essere qualificati e/o quantificati tramite ipotesi perché sono conosciuti anche se non utilizzabili. I *dati oscuri* sono i dati non disponibili, sconosciuti e non quantificabili o qualificabili in nessun modo.

Le ripercussioni delle caratteristiche dei dati sul modello sono incertezza, in caso di dati oscuri, una riduzione della precisione, se vi sono dati grigi, e un possibile errato campionamento nella selezione di dati chiari. E' fondamentale, quindi, avere chiarezza della domanda a cui si deve rispondere e dell'obiettivo dell'analisi.

Un modello può avere tutti i dati che sono necessari per misurare e definire il sistema di controllo, ma non significa che sono stati presi in considerazione tutti i possibili dati relativi a ciò che si sta analizzando. Il sistema è completo perché si sono selezionate determinate variabili, escludendo però alcuni dati con diverse caratteristiche perché avrebbero aggiunto dati grigi al sistema.

Spesso l'analisi tramite i Big Data lavora su un corpo incompleto di dati. I dati selezionati posso essere estrapolati solo da una, o più di una, delle tante fonti che potrebbe avere la domanda di ricerca. Anche se l'affidabilità del rilevamento è elevata, potrebbero non essere incusi dati rilevati nel lungo termine o potrebbero esserci dati mancanti. Quest'ultimi potrebbero inoltre essere dati oscuri, oppure dati grigi, se c'è la possibilità di valutarne il grado di incompletezza e di stimarne le caratteristiche che potrebbero inficiare il risultato della ricerca. La complessità dei modelli spesso implica che non è possibile qualificarli nella loro interezza e che quindi probabilmente i dati oscuri ci sono, ma la definizione dell'obiettivo e la definizione del sistema possono circoscrivere sufficientemente il dominio così da poter raggiungere comunque l'obiettivo della ricerca.

Uno scenario di Big Data può funzionare quindi anche su dati incompleti, ma in alcuni modelli l'incompletezza non può essere descritta, perciò i dati sono oscuri. Succede

quando un sistema mette in relazione diverse origini dati specifiche senza un modello causale completo sottostante. Le variabili quindi non sono state selezionate secondo un sistema predefinito e un modello causale, ma per il solo fatto che avevano una sufficiente correlazione in passato. Un modello causale può essere introdotto a posteriori ma può non influire sull'utilità dell'analisi.

L'analisi dei social network, che tipicamente creano Big Data, è uno scenario molto generico, che non può avere dati completi su un individuo, per la complessità della *persona* e per i diversi comportamenti degli individui nei vari social. Dall'altra parte i dati presenti nei social sono talmente tanti da non poter essere analizzati manualmente. Questo tipo di scenario deve utilizzare un'analisi automatizzata su set di dati incompleti. Il rischio perciò è chiudere gli individui in un modello generato dalla piattaforma social creata per loro, bloccati in un ciclo ricorsivo, in una "*bolla filtrante*" come dice Pariser (2011).

E' importante quindi identificare e decidere quali dati sono rilevanti e quali possono essere trascurati. Oltre ai dati l'alta probabilità della correlazione rispetto alla causalità, tipica dell'analisi con i Big Data, pone il problema di distinguere le correlazioni spurie da quelle autentiche. Le correlazioni però vanno comunque valutate attentamente perché, anche se non hanno i fondamenti causali, possono fare emergere nuovi interessanti percorsi. Fondamentale è capire quando una correlazione è falsa. In realtà non esiste una correlazione falsa o autentica in senso assoluto nei Big Data. Una correlazione può essere spuria o autentica a seconda del contesto definito da uno scopo preciso. Un po' come il dilemma "è nato prima l'uovo o la gallina" non è facile decidere se e come impostare a priori un quadro teorico o trovare a posteriori un nesso causale della correlazione rilevata. Se si interviene a priori c'è il rischio di perdersi alcune correlazioni, magari rilevanti. Dall'altra parte agendo a posteriori si può incappare nel pregiudizio, nel desiderio di adattarsi alla correlazione trovata.

I Big Data cognitivi, secondo Lugmayr A. et al. (2017), sono supportati da un doppio modello: il Modello di correlazione (Machine Learning) e il Modello di causalità (comprensione umana). È necessario un interscambio continuo tra sistemi informatici che supportano le persone nel comprendere i dati e gli umani che aiutano le macchine ad apprendere, comprendere e percepire le intenzioni umane.

Si suppone quindi che i Big Data cognitivi nascano dalla sinergia tra gli umani, preoccupati principalmente di costruire modelli causali, e le macchine che applicano metodi computazionali senza il desiderio umano intrinseco (o il pregiudizio?) di avere sempre una spiegazione causale di un fenomeno.

Un modello di correlazione Machine Learning è il risultato dell'induzione dai dati, mentre la costruzione di un modello causale umano deriva dalla deduzione dai dati degli schemi esperienziali, espliciti e impliciti, ma già esistenti nello spazio esperienziale della persona. La visualizzazione dei dati, e delle loro correlazioni, attraverso il Machine Learning, fornisce nuovi stimoli per riconfigurare il modello causale umano. A sua volta questo schema concettuale può essere implementato nei sistemi computazionali attraverso nuove codifiche, regole o programmi. Questa interazione semplifica l'impatto cognitivo richiesto e rende i dati più accessibili anche per utenti non specificamente formati come gli analisti di dati.

La percezione umana delle strutture della correlazione equivale a una validazione del modello, diventa quindi uno schema causale all'interno dell'esperienza umana. È nella trasformazione di un modello di correlazione in un modello causale che si genera la comprensione umana dei dati. Il Machine Learning fa affidamento su un vasto spazio dati, mentre gli umani su uno spazio esperienziale causale. La sincronizzazione dei due modelli permette una co-conoscenza utilizzabile sia dall'uomo che dalla macchina, e in tal modo la comprensione umana è congruente con la comprensione della macchina.

La visualizzazione dei dati diventa così un'attività fondamentale. Le normali dashboard per l'esposizione dei dati non mostrano ciò che non è disponibile o non noto, dando quindi una falsa impressione di completezza. E' necessario invece che vengano incluse indicazioni su ciò che non sappiamo, i *dati oscuri* e i *dati grigi,* e i *dati chiari* che non sono stati considerati. Va abbandonata la magia dei Big Data, dell'output stupefacente creato da fantastici algoritmi di Machine Learning. L'approccio ai Big Data deve diventare invece cognitivo e percettivo, la comprensione dei dati è fondamentale per creare conoscenza e il circolo virtuoso tra uomo e macchina.

2.2. Survey contro Big Data

Gli economisti ma anche i ricercatori sociali sembrano oggi prediligere la ricerca attraverso i Big Data rispetto ai dati strutturati derivanti da surveys o dati amministrativi. C'è sicuramente molta discussione sui vantaggi e svantaggi di questo passaggio. È possibile che una sinergia, una complementarietà tra le due tipologie di dati possa essere la scelta migliore. È sicuramente necessario aumentare la comprensione delle fonti dei dati e dei metodi oltre a quelli utilizzati per i sondaggi o in combinazione con questi. I ricercatori hanno bisogno degli strumenti giusti e i sondaggi possono essere utilizzati insieme con i Big data, per identificare nuovi metodi di ricerca e anche per massimizzare il valore di ambedue. I Big Data possono definire le correlazioni, misurando i comportamenti, mentre i sondaggi possono contribuire a spiegare le cause, valutando gli atteggiamenti e le opinioni.

I sondaggi e le ricerche di mercato stanno inoltre cambiando i metodi di raccolta dei dati verso una modalità automatizzata e digitale. Si sta osservando quindi uno spostamento dalla raccolta dei dati offline ai sondaggi web o mobile, ma soprattutto una tendenza verso ricerche di mercato o sondaggi interni all'azienda anziché in outsourcing. Il rapporto ESOMAR Global Market Research[8] del 2015 evidenzia questa tendenza attraverso l'aumento della percentuale del budget speso sull'online, ma in particolare sul maggiore utilizzo di piattaforme di rilevamento presenti nel web e sondaggi online interni. Tuttavia in generale la spesa per le ricerche di mercato e i sondaggi dal 2013 al 2014 è diminuita del 6% (percentuale combinata di online, l'indagine al telefono, faccia a faccia e tramite posta/Mail).

Le informazioni comportamentali raccolte tramite social media e la Data Analytics, cioè l'analisi dei *click* dei clienti, sembrano far sparire il bisogno di parlare con qualcuno, di intervistarlo. Inoltre la grande forza del digitale è che costa poco ed è veloce. Questo implica però concentrare tutto nel breve termine, senza minimamente preoccuparsi delle conseguenze nel medio termine. Non sentire la necessità di esplorare le motivazioni delle scelte del cliente può essere una strada che porta lontano un'azienda? Oppure, come spesso succede, il cliente viene solo bombardato da suggerimenti

[8] ESOMAR è una organizzazione no-profit che promuove la ricerca di mercato, la ricerca sociale e la Data Analytics.

pubblicitari “pressapochisti”? Ad esempio il risultato della pubblicità mirata è spesso riproporre di continuo gli acquisti già fatti (quanti ne dobbiamo comprare?) o consigliare prodotti che poco hanno a che fare con noi e le nostre esigenze come clienti.

La ricerca di mercato esterna, ma anche interna, la Data Analytics ma anche l’econometria, i sondaggi con dati online ma anche offline, possono valutare e recepire sia conversazioni che osservazioni, sia rivelazioni che spiegazioni, arrivando alla motivazione del comportamento di un cliente che davvero può portare ad input fondamentali per la strategia di marchio e la comunicazione pubblicitaria. Poter misurare, conteggiare, insomma avere tanti dati sulle persone, non significa infatti poter mettere in atto una politica di *customer centricity* effettiva ed efficiente.

Dall’altra parte i modelli teorici e i metodi vanno rivisti ed integrati in seguito all’avvento dei Big Data. La progettazione di un questionario per un sondaggio, ad esempio, non era facile nemmeno in passato. Le domande sul comportamento richiedono all’intervistato di fare affidamento sui propri ricordi. Non è quindi già scontato avere una risposta veritiera, dato che in generale la precisione della memoria durante un’intervista è scarsa. Inoltre la formulazione delle domande si sta ampliando tanto che è diventata essa stessa un limite per i sondaggi (Whitaker, 2014).

Naturalmente non tutte le caratteristiche comportamentali potranno essere evidenziate tramite i Big Data, ma sembra che comunque questa sia la direzione giusta. I social media non rispondono alle domande più specifiche dei sondaggi che servono per l’analisi demografica o per l’identificazione dei driver di un modello. I ricercatori sociali si chiedono se i dati si adattano al modello, il Data Scientist si domanda quale modello è giusto per i dati. Resta comunque valida l’affermazione di Couper (2013, citato da Callegaro & Yang, 2018, p.182) per cui ormai i sondaggi non sono più l'unico strumento disponibile per rispondere a domande di ricerca, ma solo uno dei molti modi per fare ricerca.

Oggi con i Big Data e il web i dati sono abbondanti ed economici, al contrario dei dati scarsi e costosi dei sondaggi tradizionali. La ricerca si può preoccupare meno della raccolta di dati, può concentrarsi sulle domande e sull'osservazione. La psicologia cognitiva come la teoria del doppio processo (Kahneman, 2011, citato da Baker, 2017, p. 6) promuove il comportamento come un affidabile indicatore delle scelte delle persone, migliore delle analisi attraverso i sondaggi delle attitudini e intenzioni. Si prediligono

quindi gli strumenti di monitoraggio dei social media che individuano i contenuti interessanti e possono fare text mining, con analisi automatizzata del contenuto raccolto.

Come tutti gli strumenti, anche quelli afferenti al mondo digitale hanno profondità, importanza e portata storica diversificata. Ci sono quindi più metodi e più fonti di dati che possono essere usati per una ricerca. E avendo più fonti di dati anche i ricercatori passeranno da un singolo set di dati ad una sintesi, una convergenza di più fonti di dati, anche di natura eterogenea. Da ciò deriva la necessità di più strumenti di analisi, tra cui anche il sondaggio, ma non più come metodo esclusivo di ricerca. Tutto ciò sta già avvenendo. Ne è prova il passaggio dei dati dai censimenti ai dati amministrativi anche per gli istituti nazionali di statistica.

I Big data possono fornire dati in frequenza e misura nettamente superiori rispetto a quelli derivati dai sondaggi o ai dati amministrativi. I ricercatori si trovano quindi a lavorare per la prima volta con dimensioni enormi di dati e una notevole complessità strutturale. I Big Data permettono domande di ricerca impensabili con i sondaggi perché la registrazione manuale delle tante variabili su un campione così ampio sarebbe imprecisa se non impossibile. Normalmente, sia i dati amministrativi che i Big Data non sono raccolti per una specifica ricerca o almeno non per scopi accademici. Indagini di mercato, monitoraggio dell'uso dei dispositivi e altre finalità economiche o organizzative sono ciò che spinge alla raccolta ed archiviazione dei Big Data. Ciò vuol dire che questi dati sono a disposizione dei ricercatori a costi molto bassi. Questo però non vuol dire che non vada posta molta cautela nell'utilizzo dei Big Data in sostituzione dei dati strutturati e/o dei dati diretti di un sondaggio. In primo luogo, i dati digitali, pur essendo "big", non sono necessariamente completi. Le imprese potrebbero non avere interesse in alcuni dati e quindi non raccoglierli ed archiviarli. Chi progetta un sondaggio può invece porre tutte le domande che ritiene inerenti alla ricerca, anche se apparentemente non correlate. Inoltre con i dati secondari dei Big Data ci potrebbero essere problemi di privacy nel collegare informazioni da fonti diverse per ottenere le informazioni equivalenti catturate con un sondaggio.

La privacy siglata nel caso di Big Data non copre completamente le conseguenze, anche di reputazione, in caso di uso improprio dei dati. Infatti l'enorme disponibilità di dati sugli individui e la potenza di elaborazione hanno reso obsoleti gli approcci tradizionali

all'anonimizzazione. Il consenso e la riservatezza in un sondaggio, rispettivamente richiesto e garantita ai partecipanti alla ricerca, sono invece ben definiti attraverso la spiegazione dello scopo della ricerca e di come verranno utilizzati i dati.

I Big Data sono comunque *dati secondari* rispetto ai *dati primari* dei sondaggi. È quindi lecito chiedersi se si tratti di dati davvero rilevanti. Serve capire come estrapolare valore dai Big Data, con che metodi e con quali fonti. Il giudizio critico è fondamentale: "Le previsioni basate sui dati possono avere successo e possono fallire. È quando neghiamo il nostro ruolo nel processo che aumentano le probabilità di fallimento" (Nate Silver, 2012, citato da Baker, 2017).

Le discussioni sulla correlazione nei Big Data contro la causalità nella *ricerca tradizionale* devono essere seriamente intraprese, perché la ricerca deve essere sempre e comunque rigosa. Un divertente esempio di correlazioni dai Big Data senza il minimo di causalità si possono trovare nel sito web http://www.tylervigen.com/spurious-correlations.

I Big Data arrivano spesso accompagnati da un "Big Noise" (Waldherr et al., 2016, citato da Callegaro & Yang, 2018, p. 178). L'accuratezza, la completezza e la coerenza nel tempo dei dati deve essere rispettata, perché possano essere utilizzati ma soprattutto riutilizzati.

Il modello *Total Survey Error* (TSE) descritto da Groves (1989, citato da Baker, 2017, p. 3) identifica tutti gli errori che possono essere presenti nella fase di progettazione, raccolta, elaborazione e analisi in un sondaggio. Mette in risalto cioè la possibilità che i fenomeni che si vogliono misurare in un sondaggio possano essere invece misurati in modo diverso nella realtà. L'approccio di un modello TSE è che il ricercatore deve convivere con gli errori, ma deve cercare di individuarli per mitigare, o anche solo definire, la loro influenza nel progetto di ricerca. Ci possono essere errori di misurazione degli intervistatori, degli intervistati e della progettazione del questionario stesso, nonché nei metodi di raccolta dei dati.

Gli errori di *frame* sono errori relativi alla qualità del campionamento: ci possono essere elementi mancanti, duplicazioni, unità che non dovevano entrare nel campione. I dati raccolti possono contenere errori o essere obsoleti: ad esempio la mancata risposta al questionario, in tutto o in parte, può determinare un errore di campionamento. Lo

stesso processo di raccolta ed elaborazione dei dati può produrre errori nella tabulazione, codifica, immissione di dati e produzione di pesi di sondaggio.

I Big Data sono soggetti probabilmente a tutti gli stessi possibili errori di un sondaggio. La differenza è che i ricercatori del sondaggio sono ben consapevoli delle disfunzionalità: progettano e controllano il processo di ricerca cercando di evitare il più possibile questi errori. La capacità dei ricercatori, attraverso un progetto accurato, o di un modello, di circoscrivere una popolazione specifica, di specificare i dati di interesse e di verificare il processo di raccolta diminuisce di molto l'errore possibile. L'attenzione costante alla completezza, al significato e all' accuratezza del campione permette se non di eliminare, almeno di individuare gli errori.

Il futuro del sondaggio è sicuramente con i Big Data ed è perciò legato a quanto riusciamo a rendere rigorosa la ricerca con queste nuove fonti, senza errori o con errori identificati.

L'applicazione del modello della TSE ai Big Data produce la BDTE, cioè il Big Data Total Error (Japec et al. 2015). Gli errori avvengono in particolare nei tre passaggi necessari per creare un set di dati dai Big Data: inizialmente nella generazione dei dati, poi nell'attività di estrazione, trasformazione e caricamento dei dati, ed infine nell'analisi e visualizzazione (Callegaro & Yang, 2018).

La generazione dei dati attraverso i Big Data è spesso oscura e non ben documentata. Gli algoritmi di estrazione sono spesso delle "black box" (Kreuter & Peng 2014, citato da Callegaro & Yang, 2018). Questo porta a non avere il pieno controllo degli eventuali dati mancanti, di auto-selezione non verificata, di dati non di interesse e quindi di eventuali errori di completezza dei dati, di non rappresentatività.

L'estrazione, la trasformazione e il caricamento dei dati possono dare problemi di accesso ai dati, di analisi e memorizzazione da più fonti. La trasformazione e il caricamento possono portare ad errori di codifica, di ricodifica e di modifica dei dati non controllata. In questa fase gli errori possono produrre set di dati non corrispondenti per codifica, modifica e/o pulizia dei dati alle domande di ricerca.

Unire più set di dati può facilmente introdurre degli errori. Il processo di fusione chiamato ETL (Extract, Transform, Load) identifica quali dati vengono estratti dai database originari, le modifiche e trasformazioni (ricodifiche per coerenza). Il

datawarehouse prodotto quindi può contenere la stessa variabile con codifiche diverse; lo stesso nome di variabile può essere utilizzato per misurare cose diverse; regole diverse possono essere usate per determinare quando un articolo è legittimamente mancante e quando non lo è; possono esserci entità diverse (clienti, prodotti, negozi, coordinate GPS, tweet). La complessità nella gestione degli errori introdotta dai Big Data è quindi molto rilevante e non è facile identificare un *modello di errore*.

L'analisi può avere quindi errori di campionamento, selezione e di modello. La visualizzazione, infine, può amplificare gli errori o portarne di nuovi.

Il ricercatore quindi dovrebbe avere competenze non tradizionali nella ricerca delle scienze sociali: tecniche analitiche per l'analisi dei database, capacità di programmazione per l'elaborazione di datawarehouse, competenza nella visualizzazione dei dati. È necessario quindi strutturare team composti da persone con competenze nei nuovi strumenti e ricercatori sociali.

Inoltre è fondamentale, di nuovo, considerare che con i Big Data parliamo di dati secondari, cioè dati raccolti per un altro scopo e successivamente utilizzati nella ricerca. L'importanza della qualità dei dati secondari è quindi fondamentale. La valutazione deve riguardare le regole di raccolta dei dati per garantire l'integrità, la sicurezza e la disponibilità. In realtà si dovrebbero applicare le regole di processo adottate dai ricercatori per la raccolta e l'elaborazione dei dati primari, o almeno regole equivalenti.

L'archiviazione dei dati secondari per la loro conservazione nel tempo e quindi il loro riutilizzo non è un compito facile. I dati secondari, non essendo raccolti specificatamente per la ricerca, facilmente possono non essere completamente documentati o non essere conservati in un formato che ne faciliti il riutilizzo. Documentare e risalire alla provenienza dei dati è quindi un altro fattore abilitante: nei Big Data diventa particolarmente difficile, perché spesso i dati sono estrapolati da fonti che sono già state oggetto di trasformazione e/o aggregazione.

La ricerca che utilizza dati secondari, di enorme numerosità, ricavati da fonti dinamiche ed eterogenee, deve quindi utilizzare metodi per la valutazione della qualità del dato. Ciò non significa che tutti i dati debbano essere della massima qualità, ma è essenziale che la qualità sia valutata e documentata (Baker, 2017).

Il modello BDTE è ancora agli esordi: pochi sforzi sono stati dedicati all'enumerazione delle fonti di errore e dei processi di generazione dell'errore nei Big Data (Japec et al. 2015).

Un modello BDTE è necessario perché errori importanti e di grandi dimensioni sono inevitabili per i Big Data, è nella loro natura essere enormi in quantità, in varietà, in velocità e quindi intrinsecamente "rumorosi". È fondamentale studiare le fonti dei Big Data perché pensare di avere “tutti i dati” è solo una illusione e comunque non garantisce errore zero. La consapevolezza degli errori è alla base per affrontare le loro cause e la riduzione dei loro effetti come inferenze, previsioni, conclusioni e decisioni errate.

Un framework di errore totale identifica tutte le principali fonti di errore che contribuiscono ai dati errati e/o imprecisioni nelle stime, descrivendo la natura delle fonti di errore e come gli errori potrebbero influire sull'inferenza. Mappando gli errori ed approfondendo l’influenza delle componenti di incertezza, è possibile definire nuovi metodi che ne riducano gli effetti.

Un Total Error Framework per un tradizionale set di dati (righe), variabili o caratteristiche (colonne), identifica l’errore totale come la somma tra l’errore di riga, l’errore di colonna e l’errore di cella. Le variabili non specificate determinano un errore di imprecisione, i valori di variabili in errore determinano un errore di contenuto, i valori di variabili mancanti equivalgono a dati mancanti. I record mancanti invece identificano una copertura insufficiente del campione, quelli non di popolazione determinano una copertura eccessiva, i record duplicati portano ad un errore di duplicazione. Il TSE quindi evidenzia errori di rappresentazione dei dati ed errori di misura.

Nei survey tradizionali, esistono due tipi di errori fondamentali: il *bias*, che è un errore sistematico, e la varianza, che è un errore casuale. Il *bias* è la differenza tra la media delle stime sul campione replicate e il valore reale. La varianza è la variabilità di quelle stime. In altre parole, un framework di errore totale dell'indagine mostra che, quando si valutano le procedure di ricerca dei sondaggi, è necessario considerare sia la distorsione sia la varianza.

Le fonti degli errori sono dunque relative sia a problemi di rappresentazione che a problemi di misurazione. Un campionamento perfetto con domande di sondaggio errate,

oppure un cattivo campionamento con domande di sondaggio perfette produrranno ambedue errori di stima.

Una volta definita una popolazione target, devono essere identificate le persone che possono essere utilizzate per il campionamento. Malgrado questa possa sembrare una attività asettica, in realtà può produrre errori di rappresentazione: la popolazione target e la popolazione identificata spesso non coincidono. Al possibile pregiudizio di copertura si può così aggiungere il vero e proprio errore di campionamento. Più subdolo è invece il pregiudizio per mancata risposta dell'intervistato, perché i dati non pervenuti possono falsare la copertura e allontanare il campione dalla popolazione target.

Gli errori di misurazione dipendono dalle risposte degli intervistati, ma soprattutto da cosa viene chiesto: le deduzioni che facciamo possono dipendere ed essere vincolate da come poniamo le domande. Come viene formulata la domanda e che termini sono utilizzati sono variabili che possono direzionare la risposta dell'intervistato. Le domande vanno quindi costruite attentamente e le risposte vanno valutate in modo critico.

Un Framework di errori per Big Data è molto più complesso rispetto a quanto appena descritto, perché i file di Big Data spesso non sono rettangolari, non hanno una struttura gerarchica e organizzata, i dati possono essere distribuiti su più basi di dati con origini spesso eterogene (testi, sensori, transazioni e immagini).

I Big Data possono essere soggetti a *bias* di selettività, mancanza di dati ed errori di contenuto e quindi gli errori possibili nei sondaggi possono applicarsi anche ai Big Data, compreso l'errore di campionamento. Gli approcci tradizionali per la descrizione degli errori nei dati possono però essere troppo semplicistici e non adeguati alla tipologia di dati e le tecnologie utilizzate.

L'utilizzo dei Big Data non implica quindi che non dobbiamo preoccuparci dei "soliti" problemi che si affrontano con un sondaggio.

I Big Data sono qualitativamente diversi dai dati tradizionali, dai piccoli dati. D'altra parte alcuni dati, pur di grandi dimensioni, non possiedono le altre caratteristiche di Big Data, che hanno invece un set molto diverso di tipologie di dati, una gamma di attributi che si estendono oltre alle qualità essenziali dei dati, fino ai metodi, al campionamento, al riutilizzo e alla gestione. I dati tradizionali di un sondaggio hanno una portata, una

temporalità e delle dimensioni limitate per problemi di costi e di tecniche di campionamento necessarie. Sono quindi dati limitati nella generazione e nella gestione.

Murthy et al. (2014, citato da Kitchin & McArdle, 2016, p. 2) classificano i Big Data secondo sei caratteristiche che non riguardano solo i dati, ma anche la loro estrazione, analisi e visualizzazione. Riguardo ai dati l'attenzione è sulla loro latenza temporale per l'analisi: cioè se è in tempo reale, oppure quasi in tempo reale, o ancora a lotti. L'altra caratteristica rilevante per i dati è la struttura: dati strutturati, semi-strutturati o completamente strutturati. Importante per definire i Big Data è anche l'infrastruttura di calcolo utilizzata, così come l'infrastruttura di archiviazione.

Il tipo di analisi condotta, come già visto, è fondamentale: analisi supervisionata, semi-supervisionata, oppure con apprendimento automatico senza supervisione o re-applicazione. Ma è importante anche come si estraggono i dati e le tecniche statistiche utilizzate. La visualizzazione dei dati, infine, non è solo il risultato, ma determina invece il grado di comprensione e quindi di conoscenza: visualizzazione attraverso mappe, immagini astratte, visualizzazione interattiva o in tempo reale. La privacy e la sicurezza dei dati è un'altra caratteristica che definisce i Big Data.

La differenza fondamentale, però, tra i Big Data e i dati tradizionali di un sondaggio, è data da velocità ed esaustività. Fondamentale è la frequenza di generazione e la frequenza di gestione, registrazione e pubblicazione: i dati tradizionali sono lenti e campionati, i Big Data sono rapidi e (teoricamente) *tutti*. I dati possono contenere tutte le caratteristiche come volume, relazionalità, estensione e flessibilità ma non essere Big Data.

È la qualità della velocità e dell'esaustività che definisce i dati come Big Data. Ad esempio, i dati amministrativi sono prodotti in tempo reale, sono esaustivi, ma la loro pubblicazione non è sicuramente in tempo reale. Normalmente la pubblicazione dei dati amministrativi è settimanale o mensile e in forma aggregata. Pur contenendo quindi la maggior parte delle caratteristiche dei Big Data, i dati amministrativi non sono tali. Forse lo potrebbero diventare, come lo sono diventati oggi, ad esempio, i dati di borsa.

I dati possono essere generati in tempo reale ed essere voluminosi, indicizzati, relazionali e con una copertura spaziale esauriente, ma non essere Big Data perché raccolti una sola volta, o più volte ma non di continuo. Oppure sono Big Data ma diversi

da altri tipi di Big Data: non tutti i Big Data condividono le stesse caratteristiche e quindi ne esistono diverse tipologie.

Di conseguenza, le prime due V dei Big Data, volume e varietà, non sono in realtà caratteristiche chiave per definire i Big Data. Almeno non lo sono se non associate a velocità ed esaustività. Anche i "piccoli dati" possono avere la stessa varietà dei Big Data. Sicuramente non possono avere lo stesso volume, ma se questo non è associato alla velocità ed esaustività probabilmente non si parla comunque di Big Data, solo di tanti dati. Quindi il volume è solo un sottoprodotto della velocità ed esaustività. Inoltre, anche in assenza di volume e varietà, i dati possono essere comunque considerati Big Data. Kitchin R. e McArdle G. (2016) nel loro articolo "What makes Big Data, Big Data?" affermano infatti che il metro delle 3V per definire i Big Data è in realtà falso, fuorviante, responsabile della confusione sulla definizione dei confini dei Big Data.

Questo *hype* che circonda i Big Data sembra renderli di gran lunga superiori, solo per le loro caratteristiche intrinseche, ai tradizionali studi su *piccoli dati*. Il pericolo più grande è che finanziatori, imprenditori e politici rimangano abbagliati, spostando le risorse per la ricerca dai sondaggi e dai "piccoli" studi di dati ai Big Data.

Emarginando il valore della ricerca tradizionale sui dati primari, spostando tutto verso i dati secondari, si fraintende sia la natura dei big data sia il valore dei piccoli dati. Tutti i dati che vengono acquisiti sono modellati dalla tecnologia utilizzata, dal contesto in cui vengono generati e da quali dati vengono poi impiegati. Così anche i Big Data, che possono sembrare esaustivi, sono anche essi una selezione, un campione, che tuttavia non è stato definito scientificamente dal ricercatore.

Il mondo è enormemente complesso, ed è impossibile catturare un intero dominio e tutte le sue sfumature, contraddizioni e paradossi, anche con i Big Data (Kitchin, 2013). I dati digitali sono strisciate, scansioni, azioni delle persone e comportamenti espliciti, descritti. La complessità delle emozioni, dei valori, delle credenze e delle opinioni può essere dedotta dai Big Data, ma attraverso strumenti di ricerca che esplorino in maniera più completa e critica i dati. Spesso tali strumenti si definiscono partendo da piccoli studi di dati costruiti per una domanda di ricerca specifica.

2.3. I Bias (Pregiudizi) negli algoritmi

Oggi moltissime notizie che riguardano i Big Data ci raccontano di pericoli apocalittici sull'uso degli algoritmi o di affascinanti miracoli possibili tramite l'intelligenza artificiale. È necessario invece riuscire a discernere le potenzialità e i limiti di questi nuovi strumenti.

I Big Data, come ogni altra tecnologia, hanno al loro interno delle rischiose complessità: individuare questi limiti o queste distorsioni è importante per poterle prevedere e governare.

Una delle problematiche più rilevanti nei Big Data sono sicuramente i *bias* (pregiudizi) insiti nella creazione, nella gestione e nella classificazione degli algoritmi. L'utilizzo dei Big Data, in particolare nelle decisioni, dipenderà molto dall'individuazione, dal controllo e se possibile dalla risoluzione del problema dei *bias*. I pregiudizi algoritmici, con le loro distorsioni, rischiano di rendere questa tecnologia inaffidabile, parziale e potenzialmente pericolosa, soprattutto rispetto alle discriminazioni sociali.

Il Bias in statistica è la varianza di un campione di valori rispetto al risultato atteso, cioè i valori di scostamento rispetto al valore medio. I dati analizzati possono essere ad alta o bassa varianza, quindi più o meno dispersi. Ciò che è significativo però è la distorsione che crea questa varianza. Naturalmente dall'altra parte valori a bassa varianza riducono il rischio di *bias*.

Negli algoritmi dei Big Data, identificare i valori distorti o che si scostano dal valore atteso non è facile, soprattutto per la mole di dati analizzati. E anche se si è riusciti ad identificare alcuni degli errori commessi nell'istruzione degli algoritmi, correggerli è ancora più difficile.

Il *bias* nei Big Data, vista l'enorme quantità di dati, può crescere in maniera esponenziale, automatizzando e standardizzando l'errore, rendendolo sempre più nascosto all'interno dell'algoritmo. Una volta che l'errore è totalmente ignorato, sia dal ricercatore che dalla macchina, verrà ereditato da altri algoritmi. Si innesca cioè un circolo che copiando i pregiudizi automaticamente, comprometterà altre basi dati, facendo crescere in maniera esponenziale i *bias*. Questo rende ancora più necessario

utilizzare gli *small data* a supporto dei big data per identificare chiaramente i pregiudizi e i loro effetti nell'analisi effettuata.

I *bias* sono sempre presenti nelle ricerche, nei sondaggi e in generale nell'analisi dei dati. I pregiudizi possono esserci perché presenti nel ricercatore che raccoglie i dati oppure perché i dati raccolti sono distorti. Tutti abbiamo pregiudizi, è una tendenza naturale che tutti possediamo. Identificarli e ridurli è necessario per prendere decisioni migliori. Come sempre la distorsione dei dati può essere notevolmente ridotta ponendo le domande giuste, che permettono agli intervistati, ma anche ai dati secondari, di rispondere con meno influenze esterne.

I *bias* però sicuramente risiedono già nella costruzione degli algoritmi perché questi sono progettati da esseri umani. Le nostre convinzioni, cioè la nostra griglia di pensiero, intervengono nell'analisi interpretando i dati attraverso i nostri schemi cognitivi, che seppur errati, difficilmente vengono messi in discussione. Questi sono i cosiddetti *bias cognitivi*.

Gli algoritmi, che sono alla base dell'apprendimento automatico e del *Deep Learning*, permettono di progettare, sviluppare ma anche istruire delle reti neurali. Le reti neurali sono dei modelli matematici complessi ispirati dal funzionamento del cervello umano e, come tale, ancora difficilmente analizzabili in profondità.

L'apprendimento automatico prevede l'intervento, la correzione o la modifica dell'algoritmo da parte di un essere umano, in un processo di revisione ed analisi dei risultati ottenuti.

Nel Deep Learning, invece, l'intervento umano può essere omesso, quindi l'intelligenza artificiale impara dalla propria esperienza. Sicuramente la mole fornita dai Big Data permette alla macchina di avere una esperienza quasi "completa" e quindi di implementare in automatico un algoritmo che produce risultati senza errori. Il problema è che non conosciamo l'algoritmo che ha portato a tali risultati, che seppur giusti, potrebbero essere ottenuti con algoritmi discrezionali se non addirittura discriminatori. Un po' come dire che il fine non giustifica i mezzi, ma soprattutto che non conoscendo i mezzi (algoritmi) non è possibile utilizzarli per altri fini o all'interno di altri algoritmi. Sappiamo solo che, a fronte di un determinato input, otteniamo un risultato, ma cosa avvenga realmente nelle fasi di lavorazione dei dati rimane totalmente sconosciuto.

Gli algoritmi devono fondarsi su un'ottima base dati di partenza. È necessaria una iniziale attività di *training* dell'algoritmo, per determinare i pesi giusti da applicare ai vari passaggi della rete neurale. Un algoritmo costruito su un dataset non corretto, o comunque distorto già alla base, fallirà nella predizione corretta. Questo tipo di pregiudizio nei dati (*bias* dei dataset) non è diverso da quello che si può avere anche in una ricerca tradizionale.

Molto più specifico dell'apprendimento automatico e del Deep Learning è il Bias di associazione. Le associazioni compiute dall'algoritmo possono essere discriminatorie: un termine declinato sempre al femminile o al maschile assocerà sempre i dati a una donna o ad un uomo. Bisogna tener conto che la stessa lingua italiana è ritenuta oggi discriminatoria, almeno verso le donne: molti titoli professionali o ruoli istituzionali non sono declinati al femminile, non esiste ad esempio l'ingegnera o la ministra. Dall'altra parte, invece, il genere grammaticale maschile viene usato sia per donne che per uomini. Il ricercatore può ovviare a questo pregiudizio della lingua mentre l'algoritmo, che non ha pregiudizi, associa l'ingegnere o il ministro sempre ad un uomo, mentre distingue correttamente tra cameriera e cameriere.

La distorsione si può acutizzare anche se nei dati lavorati dall'algoritmo è presente solo il termine cameriera e mai cameriere: si crea un ulteriore pregiudizio di associazione di genere. Quest'ultimo è anche detto *bias* di automazione. Se il campione su cui lavora l'algoritmo è errato o non rappresentativo continuerà a replicarlo per automazione, compromettendo altri algoritmi eventualmente collegati.

Il *bias* di interazione può portare ad un auto-apprendimento discriminatorio, razzista o inopportuno, soprattutto quando l'interazione è spinta come nel caso dei tweet.

Altro *bias* è quello di conferma: una volta ottenuto sulla base di presupposti errati un risultato prossimo a quello atteso, l'algoritmo continuerà a perpetuare l'errore all'infinito. Lo si può notare spesso negli acquisti online con feedback positivo. Gli articoli vengono riproposti senza nessuna valutazione in merito alla non ripetibilità dell'azione/soddisfazione, almeno nel breve periodo. Una volta acquistato un bene durevole, come un frigorifero, difficilmente si ripeterà l'acquisto a breve, malgrado lo si riconosca come un "buon acquisto".

Le distorsioni più rilevanti rimangono legate però ai pregiudizi, di conferma o di interpretazione dei dati. Naturalmente questi pregiudizi non sono nuovi nella ricerca.

Spesso è successo che i ricercatori accettassero prove a sostegno della propria ipotesi di ricerca e rifiutassero o minimizzassero le altre che invece confutavano o non sostenevano l'ipotesi di partenza. I ricercatori devono essere pronti a riesaminare e riconsiderare le risposte degli intervistati e devono anche cavarsela con nozioni e opinioni preconcette. Come per un sondaggio, quindi, esiste il *bias* in merito all'interpretazione dei dati e alla formulazione delle domande di ricerca. In un algoritmo, in più, c'è il bisogno di una valutazione di un essere umano che sia in grado di fare un'analisi critica a cui le macchine non sono ancora arrivate. Ma anche qui può insinuarsi una distorsione nell'interpretazione dei dati, che come sempre può essere ridotta ponendo le "giuste" domande.

Un algoritmo si può definire come un procedimento per risolvere un problema o realizzare un risultato. Gli algoritmi possono essere realizzati da persone, dalla natura o dalle macchine. Il processo decisionale autonomo è il punto cruciale del potere algoritmico. Gli algoritmi di apprendimento automatico consentono ad altri algoritmi di diventare più intelligenti con decisioni basate su schemi appresi nei dati. Potrebbe essere che il risultato prodotto sia ancora troppo "grezzo", disordinato o incerto, tale per cui l'intervento umano è necessario per prendere la decisione finale in un processo. L'algoritmo però ha già orientato, indirizzato la ricerca verso un sottoinsieme di informazioni sulla base delle quali il ricercatore prenderà la decisione. Priorità, classificazione, ordinamento, associazione e filtro attirano l'attenzione su alcune informazioni a scapito di altre: tutte queste decisioni intermedie possono essere significative per ottenere un risultato corretto. A volte queste decisioni sono concatenate per formare decisioni di più alto livello e per trasformare le informazioni. La sintetizzazione, ad esempio, combina priorità e filtri per consolidare informazioni mantenendo l'interpretazione di tali informazioni.

Spesso questi criteri di priorità, classificazione, ordinamento, associazione e filtro non sono esplicitati o pubblici negli algoritmi. Difficile, quindi, comprendere il peso dei diversi fattori che contribuiscono alla classifica. In più questi criteri potrebbero essere intenzionalmente distorti. Le decisioni di classificazione comportano l'inserimento di un dato all'interno di una determinata classe attraverso l'osservazione di un numero qualsiasi delle caratteristiche di tale dato. Le classificazioni possono essere costruite da una fase di definizione delle priorità impostando una soglia o attraverso procedure

informatiche come il Machine Learning o il *clustering*. Gli algoritmi di classificazione possono trovarsi nell'incertezza su dove posizionare il dato e quindi creare distorsioni e commettere errori. A seconda di come viene implementato l'algoritmo di classificazione potrebbero esserci diverse fonti di errore. L'algoritmo impara come classificare in base alle definizioni e ai criteri utilizzati e questo può introdurre una potenziale distorsione umana nel classificatore. I falsi positivi (è falso che sia compreso nella classe) e i falsi negativi (è falso che non sia compreso nella classe) sono gli errori possibili di un algoritmo di classificazione. Le conseguenze o i rischi possono variare per le diverse parti interessate a seconda della scelta di come bilanciare falsi positivi e falsi negativi.

Le decisioni sull'associazione riguardano la definizione delle relazioni tra vari dati o classi di dati. La decisione algoritmica correlata è il raggruppamento dei dati in vari *cluster*. I criteri che definiscono l'associazione negli algoritmi spesso si basano su una funzione di similarità, che definisce, attraverso una soglia di valore sulla somiglianza, come esattamente due cose devono corrispondere secondo l'associazione data. Come per la classificazione, anche per l'associazione ci possono essere falsi positivi e falsi negativi.

Le decisioni sul tipo di filtro, invece, portano ad enfatizzare eccessivamente o a escludere determinate informazioni. La tesi di "The Filter Bubble" di Eli Pariser (2011) si basa sull'idea che, enfatizzando eccessivamente il filtro sui desideri e caratteristiche delle persone, si mettono a loro disposizione solo le informazioni "volute", amplificando i pregiudizi e negando prospettive diverse. Mentre questa "bolla" non è reale per le persone, perché dotate di un pensiero critico, può effettivamente esistere all'interno di un algoritmo.

È chiaro quindi che ci sono un numero notevole di influenze umane incorporate negli algoritmi, come le scelte dei criteri, i dati per l'autoapprendimento degli algoritmi, l'interpretazione dei risultati intermedi e finali. E bisogna tenerne conto come si farebbe in una ricerca tradizionale.

La trasparenza degli algoritmi e dei criteri è fondamentale. A volte intervengono le leggi che impongono la divulgazione delle informazioni che hanno portato a una determinata decisione. Spesso tali norme sono state introdotte perché la mancanza di informazioni potrebbe influire sulla sicurezza pubblica, sulla qualità dei servizi forniti o condurre a scelte discriminatorie. Le aziende, però, spesso devono limitare la propria trasparenza

perché esponendo molti dettagli dei loro sistemi proprietari potrebbero divenire oggetto di manipolazione, riduzione del vantaggio competitivo o della reputazione. Un altro problema nella trasparenza degli algoritmi è la loro complessità e quindi la necessaria capacità cognitiva per comprenderli.

Un approccio alternativo o complementare alla trasparenza potrebbe essere il *reverse engineering*: è il processo di individuazione delle specifiche dell'algoritmo attraverso una rigorosa analisi del dominio, l'osservazione e la deduzione di input e output al fine di definire un modello di funzionamento. Le decisioni di progettazione, gli obiettivi, i vincoli e le regole aziendali incorporati nell'algoritmo possono aprire la *black box*, analizzando ingressi e uscite per decodificare cosa sta succedendo dentro. Interviste e analisi dei documenti di progettazione dell'algoritmo possono aiutare ad identificare dati, parametri e modalità di funzionamento. Insomma, abbiamo bisogno di una teoria o almeno di conoscere lo schema teorico su cui si basa l'algoritmo per capire se il risultato, seppur corretto, è stato anche ottenuto con i mezzi "giusti". È necessario cioè capire se l'algoritmo funziona in linea con la sua progettazione o se diventa incoerente e perché.

Il *reverse engineering* diventa così un metodo scientifico per studiare la complessità dei sistemi di calcolo artificiale. Alcuni studiosi lo chiamano "il quarto grande dominio scientifico", dopo il dominio fisico, biologico e sociale (Diakopoulos, 2013, p. 23).

Come in ogni ricerca scientifica, i risultati non sono immediatamente generalizzabili, o almeno non lo sono solo perché l'analisi è effettuata sui Big Data. Il numero di osservazioni non è infatti l'unico fattore discriminante. I dati secondari estrapolati tramite i Big Data sono riferiti a particolari siti web o social, che quindi danno un campione non casuale. Gli utenti di un sito potrebbero non essere rilevanti o solo in parte rilevanti per l'analisi che si sta effettuando. Dai Big Data sono escluse tutte le altre interazioni non effettuate tramite il web, come rapporti personali in presenza o telefonici. Perciò le risposte date dall'analisi tramite i Big Data possono fornire molte informazioni ma, come in altre tipologie di analisi, queste possono essere distorte o parziali. Anche i Big Data, quindi, hanno dei limiti che vanno affrontati, individuati e se possibile mitigati.

Tutte le ricerche sociali vanno attentamente progettate, a prescindere dalla metodologia utilizzata per recuperare le risposte alle domande della ricerca. Le carenze di dati vanno affrontate attraverso la triangolazione, cercando fonti di dati aggiuntive per ampliare e

spiegare le informazioni derivate esclusivamente dai Big Data. Le persone non selezionano casualmente i siti o i social network, perciò i dati sono distorti verso determinate popolazioni, in termini di dati demografici, di retroterra socioeconomico o di competenze su Internet. Questo è un problema che affligge una particolare tipologia di Big Data, quelli che riguardano le azioni umane. Non si applica ad esempio ai Big Data generati da sistemi M2M. Le macchine raccontano ciò che i loro sensori sono stati progettati per raccogliere senza preferire un dato ad un altro. Nella progettazione ci può comunque essere un limite, ma non nel senso che può alterare la qualità del dato parziale. Così se i Big Data vengono usati per estrarre risultati relativi ad un particolare mercato, il fatto che ci sia una popolazione che non accede a quel mercato, per limiti economici o per preferenze, non altera il risultato. Lo fa invece la generalizzazione del risultato ad una popolazione di cui il campione non è rappresentativo. Questo limite è conosciuto, riconosciuto, mitigato o eliminato scientificamente negli *small data*. Le domande di ricerca, quindi, anche per i Big Data devono riflettere i pregiudizi integrati nel campione e devono essere esplicitati i limiti nel generalizzare i risultati (Hargittai, 2015).

Se poi questi dati, senza nessuna "precauzione", sono elaborati da algoritmi "opachi", è facile che si possa venire a creare una classe di persone che si ritroveranno sempre più ed inspiegabilmente escluse dalla vita normale (O'Neil, 2016).

Dopo il recente crollo finanziario, la crisi immobiliare e i fallimenti di imponenti istituzioni finanziarie, è diventata abbastanza chiara l'influenza, la spinta che i Big Data e gli algoritmi possono dare a fenomeni catastrofici. Ma le promesse di guadagni astronomici non hanno permesso di fermare questa corsa.

Effettivamente la Data Analytics permette efficienza e imparzialità. Visionare migliaia di documenti, catalogandoli e confrontandoli in pochi minuti, non è certo una possibilità umana. Sostituire giudizi soggettivi con misurazioni oggettive per innumerevoli caratteristiche sembra di certo una buona strada da percorrere. Applicare questi algoritmi a problemi reali può davvero fare la differenza. Ma pochi degli algoritmi e dei sistemi di punteggio sono stati controllati con rigore scientifico e ci sono buone ragioni per sospettare che non supererebbero tali test (O'Neil, 2016). Quindi, applicare queste analisi alla vita reale, senza nessuna o quasi provata validità, può portare conseguenze catastrofiche sulle persone.

Il successo degli algoritmi deriva principalmente dalla loro caratteristica di obiettività, di non soggettività. Ma gli algoritmi che alimentano i dati si basano su scelte fatte da esseri umani. Quindi gli algoritmi codificano il pregiudizio e l'incomprensione umana in sistemi automatici che gestiscono sempre più la nostra vita. E come in una magia, tramite i loro meccanismi invisibili, determinano un risultato, emettono un verdetto senza possibilità di appello.

In realtà molto spesso i propositi che spingono all'utilizzo degli algoritmi derivano da buone intenzioni. Ne è un esempio l'utilizzo degli algoritmi nella selezione del personale. Sia dal lato dell'azienda che da quello della risorsa umana il mercato del lavoro è distorto: trovare il candidato ideale era molto più difficile prima degli algoritmi. Un programma che può analizzare migliaia di curriculum in pochi secondi, ordinandoli in elenchi accurati, identificando i candidati più promettenti, sicuramente migliora l'efficienza nel reclutamento. In più l'introduzione della Data Analytics nelle risorse umane aziendali rende sicuramente il processo più equo rispetto al passato. L'opportunità di lavoro ottenuta solo dal "se conosci qualcuno all'interno" è un limite nella ricerca di occupazione che viene spazzato via dagli algoritmi. Allo stesso tempo questi stessi algoritmi portano nuovi e diversi limiti e pregiudizi.

La selezione del personale automatizzata serve, almeno inizialmente, ad escludere quante più persone possibile ottimizzando tempi e costi. Ma chi stanno escludendo? È questa la domanda fondamentale per essere sicuri che un algoritmo non pregiudichi in base alla razza o a un gruppo etnico, all'orientamento politico o religioso, all'orientamento sessuale. Dall'altra parte anche le aziende che ricercano il candidato "ideale" devono assicurarsi che questi automatismi invece non discrimino risorse fondamentali per professionalità, esperienza, motivazione.

Oggigiorno purtroppo si usano sempre più test sulla personalità piuttosto che test cognitivi o test sulle competenze professionali. Purtroppo, perché si ha "l'ambizione algoritmica" di poter non solo identificare il candidato perfetto per l'attività lavorativa richiesta ma anche il più produttivo e il più fedele all'azienda.

Questi test presentano domande a dir poco impertinenti, obbligano i candidati a risposte esclusive, imponendo scelte difficili tra diverse opzioni che invece spesso nella realtà sono tutte vere ("Maledetto se lo faccio, dannato se non lo faccio", O'Neil, 2016). Non si

sa, poi, questi test cosa cercano, come l'algoritmo interpreterà la risposta data, quale è lo schema concettuale che sta alla base del questionario.

Tutto questo non vuol dire che questa non sia una strada da percorrere, ma che gli algoritmi vanno resi meno opachi, continuamente analizzati e costantemente aggiornati utilizzando i feedback sui risultati ottenuti. Altrimenti, in caso contrario, diventeranno obsoleti e pregiudizievoli.

Un buon esempio lo ha dato la Xerox. Il modello dati utilizzato per il reclutamento escludeva, a buona ragione, i lavoratori che abitavano distanti dall'azienda. Effettivamente se il tragitto casa-lavoro è impegnativo è facile che il dipendente appena può opti per una azienda più vicina alla propria abitazione. Quindi la correlazione fatta dall'algoritmo era esatta. Analizzando più approfonditamente, però, si è scoperta un'ulteriore correlazione che collegava la distanza alla povertà. Insomma il nesso causale era che chi veniva da più lontano era anche il più povero. Per correttezza, quindi, più che per efficienza, Xerox ha deciso di rivedere gli stretti legami con tale correlazione eliminando così un *bias* all'interno dei propri algoritmi.

Ciò vuol dire che non sono gli algoritmi ad essere giusti o sbagliati, ma che le correlazioni che creano analizzando vastità di dati devono essere analizzate e testate, modificate e legate ad altre correlazioni. Insomma, serve una analisi critica degli esperti del settore e anche di professionalità esterne all'ambito di analisi, utile a porre le giuste domande anche agli algoritmi.

Se poi gli algoritmi vengono utilizzati non solo per ridurre il numero dei candidati, insomma per risparmiare tempo e denaro, ma veramente per scegliere una risorsa "ideale", diventa ancora più importante per l'azienda fare una attenta valutazione. I candidati di alto livello sono costosi, soprattutto in termini di un eventuale turnover (il costo si aggira circa sul 20% della retribuzione annuale, secondo il Center for American Progress). Le risorse strategiche per l'azienda devono essere valutate al di là delle esperienze e professionalità inserite in un Curriculum Vitae. È necessario avere dati sull'inventiva, la creatività, l'intelligenza, la propensione a lavorare e guidare un gruppo. Quindi la sfida dei modellisti è individuare, nel vasto mondo dei Big Data, i frammenti di informazioni correlati con originalità e abilità sociali. Un pioniere in questo campo è stata Gild, una startup con sede a San Francisco. Il suo fondatore, Luca Bonmassar, 32 anni di Massa Carrara, dice in un intervista al Fatto Quotidiano del 6 Giugno 2013:

“Vediamo che un candidato ha sviluppato certe applicazioni, analizziamo i suoi post su Facebook, i suoi tweet, ne capiamo lo stato emotivo e ne misuriamo la capacità di lavorare in un team”. È sicuramente una sfida difficile ed insidiosa: per i *bias* insiti negli algoritmi, per tutti i “dati off line” che mancano e per tutti i candidati che non frequentano attivamente i siti social. È vero anche che i modelli predittivi di aziende come Gild vogliono proporre più che escludere candidati. Difficilmente però un candidato non inserito in graduatoria dal programma verrà esaminato dalle risorse umane di un'azienda. Il risultato è che chi non adotta questa versione *on line* potrebbe non essere mai intervistato direttamente per una selezione, non si avranno mai i suoi dati primari.

Allargando lo sguardo possiamo vedere che l’intelligenza artificiale inizia ad essere presente in molti altri settori economici, trasporti, vendite, pubblicità, energia, credito, solo per citarne alcuni. In più, e soprattutto, i modelli predittivi sono oggi presenti nelle funzioni di governo, nella polizia, nella sanità, creando impatti nella democrazia e nel benessere sociale.

Gli algoritmi non sono più semplici sequenze di istruzioni, ma sono diventati strumenti per un processo decisionale automatizzato. La disponibilità di Big Data e la possibilità di processarli ha reso facile ottenere nuove informazioni attraverso i computer, velocemente e in quantità elevatissima. Tuttavia, poiché le macchine possono trattare in modo diverso persone e oggetti situati in posizioni simili, la ricerca sta iniziando a rivelare alcuni esempi preoccupanti in cui la realtà del processo decisionale algoritmico è inferiore alle nostre aspettative (Turner Lee, Resnick & Barton, 2019).

Gli algoritmi ormai sono utilizzati in una tale varietà di applicazioni da rendere necessario un comune impegno di tutti gli *stakeholder* interessati nell'affrontare proattivamente i fattori che contribuiscono ai *bias*. Si deve essere consapevoli che non esiste un software per misurare l'equità o il pregiudizio nella progettazione degli algoritmi o per affrontare gli inevitabili compromessi tra i due. L'equità è determinata da un essere umano, secondo relazioni sociali e convinzioni etiche condivise, non può essere tradotta in una formula.

Nel complesso mondo dei Big Data nemmeno la legge può preservare da possibili ineguaglianze o pregiudizi. Le linee guida etiche pubblicate dall’Unione Europea delineano sette principi di governance.

In primo luogo, enfatizzano la necessaria presenza dell'azione e della supervisione umana. Gli ingegneri che predispongono i modelli predittivi devono determinare quali decisioni automatizzate richiedono una supervisione umana. Bisogna chiedersi se alcuni algoritmi possono dare un risultato negativo non intenzionale, per chi avrà un effetto pregiudizievole e di che gravità.

Poi si parla di robustezza e sicurezza tecnica, di privacy e governance dei dati, temi importantissimi ma già abbastanza tutelati dal nuovo General Data Protection Regulation(GDPR)[9].

Concetti più confusi e difficilmente attuabili sono invece la trasparenza degli algoritmi, la diversità, non discriminazione ed equità e il benessere ambientale e sociale. Naturalmente la parità di accesso, processi di progettazione inclusivi e la parità di trattamento sono fondamentali per perseguire l'equità.

Molto rilevante è la settima linea guida che parla di responsabilità. L'assunzione di responsabilità di tutti gli *stakeholder* sembra essere l'unica garanzia di una analisi critica di tutti gli algoritmi nel processo decisionale automatizzato. Tanta più cautela e responsabilità va data nel progettare e testare qualsiasi algoritmo che venga utilizzato per prevedere risultati o per dare punteggi in merito all'ammissibilità dell'accesso ad un determinato vantaggio.

È necessario mettere costantemente in discussione i potenziali effetti legali, sociali ed economici e le relative responsabilità quando si determina quali decisioni automatizzare e come automatizzarle con rischi minimi.

Fondamentale per attenuare i *bias* iniziali è creare gruppi di lavoro inter-funzionali, dove ci sia diversità di formazione e di livello di sensibilità culturale. Esperti di vari dipartimenti, discipline e settori possono contribuire a migliorare gli standard di responsabilità e le strategie per mitigare i pregiudizi. Inoltre, collaborazioni tra settore privato, accademici e organizzazioni della società civile possono indurre una maggiore trasparenza nell'applicazione degli algoritmi a una varietà di scenari, in particolare quelli che hanno un impatto sulle classi protette o sull'interesse pubblico.

[9] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati).

Il controllo formale e regolare degli algoritmi deve diventare una attività continua perché l'identificazione e la correzione dei risultati distorti è necessaria anche molto tempo dopo che un algoritmo è stato sviluppato, testato e lanciato. Poiché i valori di chi progetta gli algoritmi e di chi li usa cambiano nel tempo, l'*audit* deve continuamente trovare il giusto equilibrio tra i risultati ottenuti e gli obiettivi dichiarati.

L'alfabetizzazione algoritmica diffusa è fondamentale per mitigare i pregiudizi (Turner Lee et al., 2019). I soggetti su cui ricadono le decisioni automatizzate hanno il diritto di capire come sono stati lavorati dagli algoritmi i propri dati. I *feedback* che possono dare gli utenti sui tanti algoritmi ormai presenti nella vita quotidiana possono essere il migliore controllo sui *bias*. Inoltre è doveroso rendere chiaro ai soggetti destinatari quando il pregiudizio algoritmico influisce negativamente sulla loro vita e come rispondere quando si verifica.

I progettisti di algoritmi che cercano di ridurre il rischio di *bias* e le complicazioni legate a esiti negativi per i consumatori attraverso la promozione e l'uso delle proposte di mitigazione possono creare un percorso verso l'equità algoritmica, anche se l'equità non è mai pienamente realizzata (Turner Lee et al., 2019).

2.4. I dati mancanti nei Big Data

Le cause e le conseguenze delle disuguaglianze sono da sempre un obiettivo di analisi per la ricerca sociale. Alle diseguaglianze "tradizionali", come razza, classe e genere, si aggiunge e si relaziona oggi la diseguaglianza digitale. L'analisi della relazione tra le disuguaglianze digitali e altre forme di disuguaglianza deve ancora essere largamente esplorata.

La disuguaglianza digitale continua ad espandersi in molte direzioni, innescando nuove forme di discriminazione. Per questo la disuguaglianza digitale merita di essere trattata, analizzata e capita così come è stato fatto per le forme più tradizionali di disuguaglianza.

Chi può accedere, chi ha più dimestichezza, chi può partecipare pienamente al mondo digitale ha un vantaggio "digitale" che gli altri non hanno. Malgrado sembri quasi impossibile ai più, molte persone non accedono ancora ad internet, anche nei paesi sviluppati. E comunque anche chi accede si differenzia per livello di capacità, di partecipazione ed efficacia nell'uso degli strumenti digitali. Queste carenze influiscono

su un numero ancora maggiore di persone, spesso le stesse già economicamente svantaggiate o tradizionalmente sottorappresentate nella popolazione.

Queste forme di diseguaglianza e di esclusione presenti online vanno attentamente valutate insieme agli svantaggi offline e viceversa. Esistono molte relazioni sociali, economiche e culturali che legano il mondo online e quello offline. In primo luogo, le disuguaglianze digitali si combinano con razza, classe, genere e altri assi tradizionali di disuguaglianza offline.

Vanno esaminati i modi in cui le tecnologie digitali sono incorporate nel tessuto sociale ed economico e come generano vari tipi di disuguaglianze. Le disuguaglianze digitali possono rafforzarsi facendo leva sulle disuguaglianze sociali esistenti o aggravare le differenze preesistenti, trasferendole in contesti online (DiMaggio & Garip, 2012, citato da Robinson et al., 2015, p. 573). Disparità distintamente digitali influenzano ambienti offline preesistenti, come la partecipazione o il tipo di coinvolgimento. Pertanto, l'esistenza di soggetti che non interagiscono, per nulla o poco, nel mondo digitale, deve far considerare la distribuzione ineguale del digitale anche negli studi che tentano di estrapolare dei risultati da campioni di persone altamente connesse.

Le disuguaglianze digitali influenzano e distorcono un'ampia gamma di aree sostanziali: il corso della vita, le differenze di genere, l'identità etnica, la stratificazione economica, la salute e l'assistenza sanitaria.

Il comportamento degli utenti online è un'estensione di quei ruoli sociali, interessi e aspettative che organizzano la vita sociale nel mondo offline (Colley & Maltby, 2008, citato da Robinson et al., 2015, p. 572).

La scienza sociale ha il compito importante di approfondire le tipologie e le caratteristiche di queste nuove disuguaglianze, così da definire futuri interventi di mitigazione delle stesse.

Si è visto chiaramente con il Covid-19 e il relativo *lockdown* quanta diseguaglianza c'è nel mondo digitale e quanto questa si interseca con le diseguaglianze socio-economiche preesistenti nel modo offline.

Nella scuola a distanza, tra i bambini e gli adolescenti, si è chiaramente capito quanto è significativa la differenza in termini di accesso, utilizzo e abilità. In un tale contesto, in cui le istituzioni richiedono ai giovani di studiare con le risorse *online*, quelli che non

hanno un accesso adeguato sono costretti a razionare il loro tempo sullo schermo, privandosi di opportunità per sviluppare preziose competenze, anche web, di cui godono i loro coetanei meno vincolati dallo svantaggio economico.

Negli adulti, si è capito quanta differenza c'è di accesso tra chi lavora già con supporti tecnologici e chi fa un lavoro manuale, diverso, non tecnologico. Differenti abilità nell'utilizzo del digitale hanno creato diseguaglianze anche tra gli stessi lavoratori "tecnologici". Una nuova disuguaglianza digitale si è creata tra individui che possono padroneggiare molteplici flussi continui di informazioni digitali e i loro pari che lottano per gestire a fatica questi flussi informativi. Lavoratori e imprenditori digitalmente svantaggiati affrontano barriere alla piena partecipazione all'economia che i loro colleghi e concorrenti più avvantaggiati dal punto di vista digitale non hanno.

La Digital Divide tipicamente divide i più anziani, i meno istruiti e gli individui economicamente svantaggiati da quelli più esperti di tecnologia, in genere più giovani, più istruiti e con più risorse economiche. La possibilità di utilizzo di canali *online* da parte di persone in difficoltà per tenersi in contatto con i propri *caregiver*, per i primi va oltre le loro capacità, per i secondi è fonte di molti vantaggi. Le misure atte a diminuire il divario di accesso non implicano comunque una immediata parità di condizioni perché permangono elevate differenze di abilità rispetto alla gamma di attività che è necessario svolgere *online*, come è divenuto evidente durante il *lockdown*.

Le carenze nelle competenze *online* sono allarmanti perché possono avere conseguenze reali nella vita delle persone. Spesso, come si è visto, tali carenze derivano anche da disuguaglianze di genere.

Le disuguaglianze digitali si intersecano con il genere in due modi principali: attraverso le differenze legate al genere nelle competenze tecnologiche e attraverso quelle legate al mercato del lavoro associato alla tecnologia.

Ad esempio, anche se le donne adottano e usano le tecnologie con stessa velocità e capacità degli uomini, gli uomini che lavorano come sviluppatori e progettisti IT sono ancora di gran lunga più numerosi. Questa assenza di donne potrebbe persistere ancora a lungo, dato che questo *gap* è presente già a livello universitario. I divari di genere nell'utilizzo dell'IT variano da paese a paese, anche all'interno del mondo sviluppato.

Ono e Zavodny (2007, citato da Robison et al., 2015) illustrano come la disuguaglianza digitale all'interno di paesi diversi rispecchi la disuguaglianza di genere esistente in quei paesi. Paesi come il Giappone e la Corea del Sud hanno divario digitale maggiore di paesi più egualitari da un punto di vista di genere come la Svezia e gli Stati Uniti. Nei paesi dove le donne non entrano nemmeno nel mondo del lavoro, queste disparità digitali possono essere particolarmente grandi.

Questo indica il ruolo che hanno le macrostrutture sociali ed economiche nell'aumentare e sostenere la disuguaglianza digitale, in termini di genere e altri attributi sociodemografici. Lo studio della disuguaglianza digitale è quindi importante per determinare come i diversi gruppi sociali accedono e utilizzano queste tecnologie e come i loro diversi impegni digitali portino alla riduzione o all'amplificazione degli svantaggi sociali. Si parla anche di riduzione perché, aumentando la disponibilità di accesso, in particolare per le minoranze razziali e piccoli gruppi etnici, il digitale può aprire a nuove fonti di informazione e di opportunità, cambiando realmente la struttura socio-economica.

In Africa, le disuguaglianze digitali hanno rafforzato le attuali disparità razziali ed economiche, escludendo un'enorme parte del continente dall'accesso a Internet. Pertanto, l'esclusione dal digitale può coincidere con altre forme di emarginazione. La già esistente stratificazione economica subisce trasformazioni significative derivanti dalle disparità digitali esistenti ed emergenti.

Si dovrebbe indagare sul ruolo dei valori specifici del "gruppo", in quanto influenzano l'adozione e l'utilizzo della tecnologia digitale in tutto il mondo. Cruciale, da questo punto di vista, è l'esplorazione delle dimensioni della disuguaglianza digitale che si relazionano con le culture nazionali e regionali. In generale, la cultura legata alla tradizione del paese o della regione può rallentare o accelerare il processo di adozione tecnologica.

Ormai anche in Italia, la condizione lavorativa e il reddito da lavoro, ma anche da pensione, prevedono un'elevata intensità di utilizzo del computer, nonché di impronte di attività online. Il passaggio allo *smartworking* ha elevato enormemente questa intensità di uso e presenza digitale. Le competenze digitali stanno giocando un ruolo sempre più critico anche nella ricerca di lavoro. Acquisire competenze digitali vuol dire quindi ottenere un vantaggio nel mercato del lavoro, con più probabilità di ottenere un posto di

lavoro, di avere salari più alti, e una più alta propensione all'attività imprenditoriale. Le competenze digitali sono preziose per interagire con le parti interessate e i clienti, la raccolta di capitali finanziari, lo sviluppo di piani aziendali, l'ideazione di modelli di business e l'aumento di capitale sociale (Chen, 2006, citato da Robinson et al., 2015, p. 575). Le donne imprenditrici sono meno efficaci nel convertire il digitale e la rete internet in vantaggi aziendali (Jennings & Brush, 2013, citato da Robinson et al., 2015, p. 575). Nel mondo del lavoro di oggi è ormai probabile che i lavoratori qualificati digitalmente manterranno o troveranno lavoro, mentre i lavoratori digitalmente non qualificati saranno ulteriormente penalizzati.

Il digitale oggi viene utilizzato anche per fornire le cure sanitarie, migliorare i risultati di salute e creare un più equo sistema sanitario. Queste tecnologie, denominate eHealth o telemedicina, vengono utilizzate per migliorare l'accesso alle cure cliniche, consentire ai pazienti di monitorare e autogestire le proprie condizioni mediche e controllare i costi (Hale, 2014, citato da Robinson et al., 2015, p. 576). Ma i gruppi sociali svantaggiati, che probabilmente sono i più soggetti a problemi di salute, sono anche quelli a cui spesso manca l'accesso, le capacità e i comportamenti associati per rendere efficace l'uso di sistemi di sanità elettronica.

Una buona parte della ricerca preliminare dimostra l'importanza della disuguaglianza digitale di primo livello nell'uso della sanità elettronica. Davison e Cotten (2003, citato da Robinson et al., 2015, p. 576) hanno riscontrato che la velocità di connessione a Internet, misurata come la quota di popolazione che usa la banda larga rispetto al modem dial-up, è uno dei fattori più importanti che spiega le differenze nelle attività online rispetto ad altri fattori di disuguaglianza digitale.

Il sito web americano HealthCare.gov, lanciato nel 2013, doveva essere la risorsa primaria per le informazioni sul piano assicurativo sanitario. Una banalità, o tale sembrava, ne ha tuttavia reso difficile l'utilizzo: il sito è stato progettato per la visualizzazione su monitor desktop grandi. Questo ha nascosto parte del sito, ha creato difficoltà nella navigazione a persone con schermi più piccoli o dispositivi portatili. Esiste quindi un potenziale di esclusione delle persone digitalmente svantaggiate dai dataset su cui si baseranno i metodi emergenti di ricerca sanitaria.

È necessario, quindi, sviluppare soprattutto analisi sulle interrelazioni delle disuguaglianze digitali con istituzioni come sanità, mercato del lavoro, scuole,

organizzazioni e stato. Tali diseguaglianze vanno definite in termini di disparità sociali esistenti come razza, classe e genere, e disuguaglianze emergenti dal regno digitale, individuando i collegamenti causali che connettono forme specifiche di vantaggio (svantaggio) offline e di coinvolgimento digitale (esclusione digitale) da parte di soggetti distinti da attributi come sesso, classe e razza o risorse come tempo e denaro.

I ricercatori sociali valutano che la raccolta dati *online* deve trarre vantaggio dalla ricerca multidisciplinare che è già stata svolta sulla disuguaglianza digitale, ma anche dal lavoro fatto sulla progettazione della ricerca e sulla metodologia dell'indagine. Dobbiamo andare avanti basandoci simultaneamente su ciò che abbiamo imparato sulla raccolta dei dati nell'era precedente a Internet e sfruttando le possibilità offerte dalle tecnologie digitali. Insomma è necessario adottare un approccio integrativo e lungimirante, ma che conserva le lezioni del passato (Robinson et al., 2015).

Joy Buolamwini è una ricercatrice del MIT di Boston e tramite uno studio dal titolo *Gender Shades* ha verificato l'accuratezza di alcuni prodotti di riconoscimento facciale come IBM Watson, Microsoft Cognitive Services e Face ++, arrivando alla conclusione che questi sistemi trattano alcune etnie in modo più impreciso rispetto alle altre. Nel caso specifico, la ricerca ha dimostrato che questi applicativi avevano una precisione del 99% per gli uomini bianchi, ma di contro un'attendibilità del 34% per le donne dalla carnagione scura. La spiegazione sta nel fatto che gli algoritmi, usati da questi sistemi, si basano su soggetti prevalentemente di tipo maschile e di carnagione chiara. IBM, in risposta ai risultati della ricerca del MIT, ha rilasciato un miglioramento al proprio sistema, che consente di ridurre notevolmente l'errore. Ha dovuto però realizzare un data set chiamato DiF (Diversity in Face) composto da 1 milione di immagini e da 10 schemi di codifica. Durante il 2018 è diventato famoso lo studio fatto da ACLU (American Civil Liberties Union), un'associazione americana a difesa dei diritti civili, che usando Rekogniction ha mescolato le foto dei parlamentari americani in un database di circa 25 mila immagini, dimostrando che nel 5% dei casi emergeva un'inesistente corrispondenza tra criminali e gli stessi parlamentari. Il dato più allarmante è che in mezzo a questi falsi positivi il 39% riguardava deputati dalla pelle scura.

La mancanza di *set* di dati etichettati in base all'etnia limita la generalizzabilità della ricerca. Manca cioè la variabile che reindirizza l'algoritmo se ci sono dati mancanti. Anche il rapporto sul genere (Ngan & Grother, 2015) del National Institute of Standards

and Technology (NIST) che ha esplorato l'impatto dell'etnia sulla classe di genere tramite l'uso di un *proxy* etnico, cioè tramite il paese di origine, non ha inserito in nessuna delle 10 località utilizzate paesi come l'Africa o i Caraibi dove si trovano popolazioni nere significative (Buolamwini & Gebru, 2018). Uno studio di Farinella e Dugelay (citato da Buolamwini & Gebru, 2018) rivela però che l'etnia sembra non avere alcun effetto sulla classificazione di genere, anche se nella loro analisi hanno usato comunque una categorizzazione etnica, pur se binaria: caucasico e non caucasico.

Difficile è capire quali sono veramente i dati mancanti, cioè quelli che hanno un effetto sugli algoritmi, sui risultati e la loro generalizzabilità. Soprattutto se si pensa, ad esempio, che nell'analisi facciale automatizzata la posa, l'illuminazione e l'espressione nella fotografia possono influire sulla precisione dei risultati. L'illuminazione è di particolare importanza quando si esegue una analisi in base al tipo di pelle. Nelle fotocamere le impostazioni predefinite sono spesso ottimizzate per esporre meglio la pelle più chiara rispetto alla pelle più scura (citato da Buolamwini & Gebru, 2018). Immagini che presentano una significativa perdita di informazioni, possono, quindi, discriminare allo stesso modo dei dati mancanti sulle etnie.

Un esempio quotidiano lo dà il popup "io non sono un robot" che spesso appare agli utenti che navigano nel web. Si richiede di riconoscere un elemento all'interno di alcune foto. Normalmente c'è sempre una foto in cui è difficile anche per una persona reale riconoscere l'elemento. Questa verifica si basa appunto sulla consapevolezza che gli algoritmi sono addestrati per riconoscere attraverso classificazioni "standard" e che l'eccezione, come il diverso posizionamento o il diverso contesto, sono catturabili solo dall'acutezza umana.

Dati e set di dati non sono obiettivi. Più precisamente la selezione per costruire dataset, le inferenze che vengono estratte, il significato dei dati sono creazioni di design umano. E i pregiudizi che inevitabilmente vengono nascosti nelle interpretazioni umane sono rilevanti quanto i Big Data. Dobbiamo collegare i dati al sistema complesso da cui provengono, altrimenti si rischia di allontanarsi fuori misura dagli obiettivi dei Big Data. Questo è successo con l'uragano Selly: il maggior numero di tweet proveniva da Manhattan e senza nessun'altra informazione di contesto, è plausibile inferire che Manhattan sia stata il fulcro del disastro. In realtà il *blackout* prolungato nelle aree più colpite dall'uragano ha creato un ridottissimo accesso ai cellulari e quindi pochissimi

tweet. Si presume che i Big Data riflettano accuratamente il mondo sociale, ma ci sono lacune significative, intere comunità non connesse. Questi enormi volumi di dati sono intrinsecamente collegati al luogo fisico e alla cultura umana. E la geografia, come le culture, hanno le loro peculiarità. Le persone che vivono in un determinato contesto socio-economico o demografico o geografico possono avere meno probabilità di possedere uno *smartphone*, di saperlo usare, di disporre di una connessione. E se questo vale per una ampia fetta di una città, di un paese o di una regione, significa che nei set di dati mancano input da parti significative della popolazione.

Presa coscienza di questa possibile assenza di dati, è possibile mitigarla arricchendo il dataset con altre tipologie di dati o riducendo la generalizzazione sapendo quale rappresentatività ha il campione di dati che si sta analizzando. Pertanto, con ogni set di Big Data dobbiamo chiederci quali persone sono escluse. Si tenga conto che, ad esempio, i problemi di segnale dei big data non scompaiono con l'aumentare dell'uso di smartphone o altre tecnologie digitali. Anzi, potrebbero esserci nuovi dati mancanti per l'introduzione di nuovi dispositivi, software e pratiche culturali. Oggi che i dispositivi personali sono utilizzati come *proxy* per le esigenze pubbliche, c'è il rischio che le disuguaglianze già esistenti diventino ancora più radicate.

I Data Scientist devono guardare alla ricerca sociale, che ha una lunga storia di domande per sapere da dove provengono i dati con cui stanno lavorando, quali metodi sono stati usati per raccoglierli ed analizzarli e quali pregiudizi cognitivi può portare l'interpretazione dei dati.

Deve esistere un controllo sulla qualità dei dati, un *audit* che verifichi, attraverso l'analisi delle eccezioni e degli scostamenti, la validità del processo di raccolta dei dati. Nel più semplice caso ad esempio dei contatori, il controllo di qualità dovrebbe concentrarsi sulla rilevazione dei dati mancanti, perché il dispositivo è fuori servizio, o dei dati discostanti, per anomalie o problemi della macchina.

Il dato statistico è ottenuto tramite la definizione a priori di una griglia concettuale e operativa e di un definito processo. La qualità del dato è intrinseca alla qualità della definizione della griglia concettuale più adatta a rispondere alle domande di ricerca, al controllo e allo studio sulla partecipazione dei rispondenti, ad una copertura della popolazione pressoché completa o almeno sotto controllo, ad un dato fornito volontariamente e scientemente. Ma soprattutto ci sono delle fasi del processo di

produzione del dato che sono rese trasparenti all'utilizzatore e anche a chi fa parte della popolazione oggetto di analisi. Anche il dato statistico ha i suoi limiti in termini di qualità e completezza del dato, ma questi sono evidenziati, conosciuti e se possibile mitigati o eliminati. Il dato statistico comprende quindi tutta una serie di informazioni aggiuntive (i metadati) che lo rendono fruibile ed interpretabile da sociologi, economisti, filosofi o statistici.

La trasparenza nella Data Analytics, negli algoritmi che utilizza, è fondamentale per valutare la qualità del dato, del risultato, anche nei Big Data. È il punto di partenza della conoscenza che deve essere noto sia ai produttori che agli utilizzatori di dati.

Integrare le fonti di Big Data con domande derivate dalla metodologia delle scienze sociali può risultare più complesso ma potrebbe far ottenere una più rigorosa ricerca qualitativa. Sapere il perché e il come, da chi e dove abbiamo ottenuto i dati arricchisce anche i Big Data di un contesto necessario per la validità del dataset su cui si effettuerà l'analisi. È necessario spostare l'attenzione dal volume dei Big Data, alla qualità, al dato tridimensionale, con la sua profondità conoscitiva (Crawford, 2013).

2.5. GDPR e Privacy

Lo sviluppo di un'economia fondata sui dati mette in discussione molte delle tutele consolidate del diritto alla privacy. La tutela dei dati è una sfida imponente con i Big Data: la raccolta continua e massiva, la trasmissione istantanea e l'incessante riutilizzo dei dati rendono difficile creare un sistema di tutela della privacy. In più, la correlata espansione delle nuove applicazioni, della robotica e della realtà aumentata, interagisce direttamente con la vita reale delle persone e rende tale tutela ancor più difficile, per il nostro desiderio di condivisione e di connessione. I rischi per la nostra persona sono nuovi e complessi, soprattutto nella sua proiezione digitale. Con i Big Data le forme di controllo sono nascoste e pervasive.

Le imprese tecnologiche hanno potuto aumentare la raccolta e la disponibilità dei nostri dati. Dall'altro lato i Governi utilizzano questi dati per il controllo delle attività svolte in rete con la finalità di combattere la criminalità e il terrorismo. Tale controllo o sorveglianza, giustificata o no, ha portato a un'intrusione nella vita di tutti noi con effetti importanti sui comportamenti individuali e collettivi e addirittura sulle nostre democrazie.

È importante preservare la fiducia degli utenti nello spazio digitale e nelle sue potenzialità. Le riforme giuridiche a livello europeo rappresentano una necessità per definire dei limiti uniformi e per anticipare le esigenze future. Il potenziale discriminatorio dei Big Data è imponente: malgrado i dati non identificativi o aggregati, i profili sono sempre più precisi ed analitici e quindi passibili di discriminazione.

Questi fenomeni sono regolamentabili solo attraverso un più rigoroso approccio etico e di generale responsabilità, perché probabilmente in un futuro molto prossimo sarà difficile mantenere un effettivo controllo sui dati. Diventa fondamentale, quindi, promuovere garanzie di trasparenza dei processi e riequilibrare i rapporti asimmetrici tra chi fornisce e chi sfrutta i dati.

Attraverso il monitoraggio continuo della rete, si individuano i temi di maggiore interesse, analizzando puntualmente la geografia dei bisogni e delle relazioni sociali per elaborare contenuti personalizzati anche nell'offerta elettorale.

La tecnologia in sé non è evidentemente né buona né cattiva, però i suoi effetti possono essere diversi e quindi emerge l'esigenza del controllo sull'enorme massa di dati disponibili *online*. Li utilizza anche la politica, in tutte le sue espressioni, dalle censure preventive, tipiche dei regimi dittatoriali, allo studio dei comportamenti per centrare meglio le campagne elettorali. Ma li utilizza anche la sicurezza pubblica per il bene del Paese, con la conseguenza paradossale che in questo caso il pubblico utilizza gli archivi dei privati per fare la prevenzione.

Vanno comunque e sempre imposti dei limiti ad un'espansione incontrollata dello sfruttamento dei dati, che sia economico o pubblico. La normativa sulla privacy è uno di questi limiti, spesso vissuta dalle imprese come un ostacolo all'utilizzo delle informazioni che sono considerate oggi un fattore di produzione e di competitività ed un attrattore di investimenti.

La generazione di dati da parte delle reti IoT sarà immensa. Gli oggetti connessi non saranno miliardi, come spesso si ripete, ma decine di miliardi. Il secondo sviluppo importante riguarderà i modelli prescrittivi e predittivi basati su piattaforme di intelligenza artificiale che renderanno ancora più sensibile il tema dell'accesso ai dati personali. Questa tecnologia importante dà grandi opportunità a costi ridotti, ma genera anche enormi rischi sulla qualità e la responsabilità dei soggetti che accedono ai dati e li utilizzano.

Ci sono tre casi di generazione e utilizzo dei dati che hanno particolare rilevanza sotto il profilo dei rischi e delle opportunità.

Il primo riguarda la generazione di dati relativi a comunicazione tra oggetti: dati relativamente semplici da raccogliere gestire e analizzare, perché rispondono a leggi fisiche. Questo tipo di raccolta pone problemi regolatori non sotto il profilo della privacy, ma sotto il profilo della sicurezza per la possibilità di accedere in modo abusivo a macchinari e impianti critici.

Un secondo caso è quello della generazione di dati relativi alla comunicazione tra persone e oggetti: I dati generati consentono di analizzare le modalità di utilizzo degli oggetti e quindi ne aumentano l'efficacia, oppure consentono di sviluppare modalità diverse per assolvere agli stessi bisogni in modo più efficiente. Le problematiche di privacy sono più rilevanti in questo caso.

Infine, vi è il caso della generazione di dati relativi ai rapporti tra persone: questo è molto più complesso come fenomeno, anche per la possibilità di combinare i dati generati da interazioni sociali con le tracce fisiche lasciate dalle persone.

Oggi la capacità di utilizzo di questi dati è concentrata nelle mani dell'*intelligence* americana, probabilmente in parte di quella cinese e soprattutto di poche gigantesche imprese private americane.

L'Europa inizialmente non ha capito la gravità delle implicazioni di questa concentrazione di potere nelle mani di poche imprese americane. L'accordo di Safe Harbour[10] tra Europa e USA ha garantito un libero accesso da parte degli Stati Uniti ai dati dei cittadini europei senza vincoli. Solo recentemente, grazie all'accordo Privacy Shield[11] siglato agli inizi del 2016, si è provveduto a porre dei limiti all'uso indiscriminato di dati personali di cittadini europei da parte di imprese americane, e sono state introdotte sanzioni nel caso di comportamenti non in linea con le normative. La Corte di giustizia dell'Unione europea (CGUE) si è pronunciata il 16 luglio 2020 (c.d. *Sentenza Schrems II*) in merito al regime di trasferimento dei dati tra l'Unione europea e

[10] Nell'anno 2000 l'Unione europea e gli Stati Uniti d'America hanno concluso un accordo, il Safe Harbor, sul libero trasferimento dei dati di cittadini europei verso gli Stati Uniti. Il 6 ottobre 2015 la Corte di giustizia dell'Unione Europea, tramite la sentenza Schrems, ha annullato l'accordo.
[11] L´accordo protegge i diritti fondamentali delle persone nell´UE i cui dati personali vengano trasferiti negli Stati Uniti, e stabilisce regole certe per le imprese che effettuano trasferimenti di dati al di là dell´Atlantico.

gli Stati Uniti invalidando la decisione di adeguatezza del Privacy Shield, adottata nel 2016 dalla Commissione europea in seguito alla decadenza dell'accordo Safe Harbor. La motivazione della mancata validità dell'accordo Privacy Shield è stata giustificata dalla Corte ritenendo "che i requisiti del diritto interno degli Stati Uniti, e in particolare determinati programmi che consentono alle autorità pubbliche degli Stati Uniti di accedere ai dati personali trasferiti dall'UE agli Stati Uniti ai fini della sicurezza nazionale, comportino limitazioni alla protezione dei dati personali che non sono configurate in modo da soddisfare requisiti sostanzialmente equivalenti a quelli previsti dal diritto dell'UE e che tale legislazione non accordi ai soggetti interessati diritti azionabili in sede giudiziaria nei confronti delle autorità statunitensi".

Il principio che deve guidare tutti gli accordi o disposizioni in merito alla tutela dei dati personali dei cittadini della Comunità Europea è dettato dall'articolo 44 del GDPR[12], in base al quale "tutte le disposizioni di detto capo devono essere applicate al fine di garantire che non sia compromesso il livello di protezione delle persone fisiche garantito da tale regolamento". Per le deroghe permesse dall'articolo 49 del Regolamento (UE) 2016/679, è opportuno notare che comunque i trasferimenti verso Paesi con una non adeguata tutela devono essere esplicitamente acconsentiti e solo con una adeguata informativa dei possibili rischi di tali trasferimenti per l'interessato, dovuti alla mancanza di una adeguatezza e garanzia normativa.

Tuttavia, anche con il nuovo Regolamento GDPR è ancora troppo presto per dire se si siano ottenuti dei miglioramenti effettivi in termini di tutela dei cittadini europei. Il tema dei Big Data pone soprattutto problemi di regolazione nell'accesso, nella elaborazione e nell'utilizzo dei dati. La regolazione non deve essere finalizzata a limitare in qualche modo le potenzialità che ne derivano, ma deve invece garantirne la trasparenza.

L'Europa è stata la prima ad avere attenzione alla tutela della privacy. Oggi molti Paesi hanno adottato regolamenti e leggi in linea con quelli europei. Ci sarebbe bisogno di raccordare le politiche di gestione delle informazioni di tutto il mondo, perché ormai è in atto una globalizzazione dei flussi di dati e i regolamenti nazionali non possono più tutelare i cittadini. Il nuovo regolamento europeo ha dato, per così dire, il via: non ha più

[12] REGOLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati).

rilevanza il luogo in cui sono stabilite le imprese, ma solo il fatto che vi siano cittadini europei interessati dalla loro attività.

Con i Big Data il concetto di privacy va oltre i dati demografici, come età e residenza, perché i dati raccolti, elaborati e scambiati riguardano, ad esempio, abitudini di consumo, spostamenti nel territorio, comportamenti online. Il GDPR, infatti, per identificare i dati soggetti a tutela fa riferimento a qualunque informazione relativa ad un individuo. In più, nel prossimo futuro con la velocità di scambio delle informazioni delle nuove tecnologie, il dato personale scomparirà a favore della nozione di dato riferibile a uno ma anche a più individui, a *cluster* che identificheranno interessi e diritti comuni.

L'art. 25 del GDPR sancisce che "Il titolare del trattamento mette in atto misure tecniche e organizzative adeguate per garantire che siano trattati, per impostazione predefinita, solo i dati personali necessari per ogni specifica finalità del trattamento. Tale obbligo vale per la quantità dei dati personali raccolti, la portata del trattamento, il periodo di conservazione e l'accessibilità. In particolare, dette misure garantiscono che, per impostazione predefinita, non siano resi accessibili dati personali a un numero indefinito di persone fisiche senza l'intervento della persona fisica.". Praticamente dice che il titolare del trattamento dei dati deve acquisire meno dati possibili e per minor tempo possibile. Assegna una responsabilità (accountability) a chi tratta i dati sensibili, e lo fa già fin dalla fase di progettazione del trattamento (quanti, quali, per quanto tempo, privacy by design) nonché per impostazione predefinita (necessità dei dati raccolti, privacy by default).

Il nuovo Regolamento contempla, inoltre, il diritto alla portabilità (*data portability*): il passaggio dei dati tra titolari deve essere validato nel contenuto, trasparente, non duplicato. Insomma è rilevato un diritto di sapere chi ha i dati, ma anche che i dati si mantengano corretti, aggiornati nel passaggio da un titolare ad un altro. Una soluzione che si sta valutando è la *one box only*: i dati sono presenti in un solo contesto tecnico, sotto l'esclusiva autonomia dell'interessato, che ne gestisce le autorizzazioni e le revoche alla legittimazione.

L'articolo 22 del GDPR sancisce: "L'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo

significativamente sulla sua persona. […] Nei casi di cui al paragrafo 2, lettere a) e c), il titolare del trattamento attua misure appropriate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato, almeno il diritto di ottenere l'intervento umano da parte del titolare del trattamento, di esprimere la propria opinione e di contestare la decisione.". Questa disposizione ha un intento antidiscriminatorio, come esprime chiaramente il considerando n. 71 del regolamento: il titolare deve garantire "[…] che siano rettificati i fattori che comportano inesattezze dei dati e sia minimizzato il rischio di errori […] e che impedisca tra l'altro effetti discriminatori nei confronti di persone fisiche sulla base della razza o dell'origine etnica, delle opinioni politiche, della religione o delle convinzioni personali, dell'appartenenza sindacale, dello status genetico, dello stato di salute o dell'orientamento sessuale.".

Le decisioni completamente automatizzate, con impatto sui diritti della persona, sono vietate ed è sempre permesso che l'individuo richieda l'intervento del fattore umano nelle decisioni, come ad esempio nell'accesso al credito. Proprio a proposito di questa attività bancaria, sempre più automatizzata, si sta riflettendo sull'equità degli algoritmi. Le centrali rischi che raccolgono i dati per la valutazione di credito hanno definito limiti ben precisi per le registrazioni dei soggetti sia con mancati o ritardati pagamenti, sia per quelli con pagamenti puntuali. Poiché i dati restano legati al passato, si assiste al perpetuarsi di una situazione immutabile: chi è stato moroso nei pagamenti non avrà più l'accesso al credito e chi in passato ha sempre pagato puntualmente avrà invece più probabilità di avere rate non pagate. Si attua cioè una discriminazione sociale perpetuando una situazione passata.

L'eccesso di perfezione degli algoritmi può bloccare l'innovazione, poiché gli errori, la fiducia, l'investimento azzardato servono per poter fare nuove scoperte.

È evidente come questa normativa vada ad incidere sulle attività di business e manageriali connesse ad attività di *analytics*, anche nell'ambito del lavoro.

Il Regolamento prevede in modo esplicito alcune prescrizioni e responsabilità di trattamento di dati sensibili connesse al rapporto di lavoro (ad esempio medicina del lavoro, valutazione della capacità lavorativa del dipendente). L'articolo 88 autorizza gli Stati Membri ad introdurre delle discipline specifiche, anche tramite contrattazione collettiva, in materia di trattamento dati nei contesti di lavoro.

L'attenzione del GDPR al mondo del lavoro è dovuta alla certezza che nel futuro anche la gestione delle risorse umane farà sempre più affidamento ai dati raccolti e alla loro analisi attraverso tecnologie di Data Science. Infatti il lavoratore deve essere sì tutelato nel momento di raccolta e trattamento dei propri dati, ma, soprattutto, quando su tali dati si basano decisioni automatizzate, in tutto o in parte, nei suoi confronti.

Ci sono tre rischi fondamentali con i Big Data e la Data Analytics. In primo luogo, l'enorme quantità di dati personali o sensibili che vengono raccolti e correlati ad altre fonti mette in grado gli algoritmi di poter fornire altre informazioni o dati ancora più personali e sensibili. In secondo luogo, emerge la possibilità che i dati della persona o sottostanti al modello contengano errori che influenzano poi la decisione. Infine, *bias* con possibili effetti discriminatori possono essere contenuti negli algoritmi che selezionano o elaborano i dati.

Secondo Kelly Trindel dell'EEOC (citato da Dagnino, 2017), nei contesti lavorativi, e non solo, il rischio delle decisioni automatizzate è che le variabili e i risultati possono risultare correlati, ma solo perché ambedue sono anche correlati con altre variabili causali. La conseguenza distorsiva che ci può essere è fondare le decisioni sulla relazione di correlazione anziché sulla relazione causale. La decisione avrà comunque un impatto reale sulla risorsa umana.

È quindi necessaria una nuova etica dei dati che coinvolga le imprese nella loro intera organizzazione. L'azienda deve cercare di identificare, conoscere, capire ed applicare i limiti e le tutele per una gestione critica ed equa dei dati raccolti, elaborati e trasformati. Una tale presa di responsabilità riferita alla propria attività rinnoverà la fiducia dei propri *stakeholder*, che non si sentiranno più minacciati dalla Data Science, attivando una collaborazione volta all'innovazione rispettosa dei diritti e dei sentimenti delle persone.

Dall'altra parte, cominciare a riflettere su regole comuni è un fatto che non riguarda esclusivamente ciò che può fare il mercato, ciò che può fare la legge, ma ciò che può fare una nuova responsabilità diffusa dei dati. È necessario un nuovo contratto sociale per la condivisione dei dati che responsabilizzi tutti gli attori di questo mondo digitale, sia per avere una tutela dei dati sensibili sia per avere una vera e aperta disponibilità di questi stessi dati.

F-Secure, una azienda specializzata in cyber security, ha fatto un esperimento nel 2014 aprendo un hotspot gratuito a Londra, con la possibilità di collegarsi al Wi-Fi. Nei termini e nelle condizioni d'uso del servizio, ha inserito una clausola *Your first-born child*, il tuo primo figlio (battezzata poi *clausola Erode*). Utilizzando il servizio si acconsente a dare il proprio primo figlio a F-Secure, quando richiesto dall'azienda. Nel caso non esista un primo figlio, la F-Secure si può rivalere sulla proprietà del proprio animale domestico. I termini di questo servizio sono validi per l'eternità. Nessuno ha letto la clausola e tutti hanno accettato per avere il Wi-Fi gratuito. "Tanto allarme da una parte, zero coscienza critica dall'altra." (Russo, 2015).

A definire il concetto di privacy ci hanno provato tutti, filosofi, psicologi, sociologi e legislatori, ma in realtà non esiste ancora un significato di privacy condiviso tra le varie discipline (Solove, 2006, citato da Miltgen & Peyrat-Guillard, 2014, p. 4), come non esiste ancora una normativa comune a tutti i paesi del mondo.

La ricerca del 2014 di Lancelot Miltgen e Peyrat-Guillard effettua una valutazione qualitativa degli atteggiamenti verso la privacy, la divulgazione e la protezione dei dati personali. Tale ricerca è stata effettuata utilizzando 14 *focus group*, in 7 Paesi dell'Unione Europea, composti da 139 partecipanti diversi per sesso, età e status professionale. La ricerca è stata effettuata attraverso l'analisi automatizzata del testo e in base a criteri di classificazione del lessico. La triangolazione con intervistati con diversi background, due moderatori, due *focus group* per paese, due metodi di analisi statistica e due pacchetti software, danno affidabilità e validità a questa ricerca. I risultati rivelano la complementarità dei metodi di ricerca e del software. La convergenza dei risultati conferma l'importanza della triangolazione per garantire la qualità dei risultati in una analisi su dati qualitativi (Lancelot Miltgen & Peyrat-Guillard, 2014).

Lo studio tende a definire il concetto di privacy analizzando le preoccupazioni della popolazione in merito al rischio avvertito di abuso nella raccolta, nell'utilizzo e nella divulgazione dei dati sensibili. Si cerca di determinare, cioè, su quali aspetti gli utenti si concentrano quando la loro privacy può essere a rischio, nonché i criteri con cui decidono se divulgare o proteggere i propri dati personali, e quindi i relativi comportamenti rispetto alla percezione del rischio. Inoltre lo studio analizza come variano questi aspetti e criteri in base alle variabili dell'età e della cultura. Queste

preoccupazioni costituiscono i limiti maggiori allo sviluppo del digitale nel commercio, nella pubblica amministrazione, nella sanità. I dati di queste percezioni del rischio possono portare a progettare strumenti, sistemi e norme utili ad affrontare davvero i problemi di privacy.

La ricerca si incentra su quattro punti fondamentali per determinare la relazione tra rischio di privacy e comportamento: capacità di controllo, livello di protezione e regolamentazione, grado di fiducia e di responsabilità percepita.

Le persone si preoccupano della possibile perdita di controllo o dell'uso improprio dei propri dati sensibili. La maggior parte delle persone che hanno partecipato allo studio considerano invasiva la raccolta dei dati *online*. I comportamenti però si distinguono tra chi la percepisce come un inganno e cerca di non farsi riconoscere, chi pensa di non avere scelta o che sia solo un vincolo per l'utilizzo, chi invece riconosce un vantaggio e chi ancora lo considera un inevitabile compromesso tra costi e benefici. La maggior parte degli intervistati ha una forte preoccupazione di una perdita di controllo, temono intrusioni alla loro privacy, perché considerano che ci sia un alto rischio e soprattutto una minaccia futura, quindi non prevedibile, di un uso improprio dei dati personali.

È evidente nei risultati della ricerca che c'è un sentito bisogno di maggiore protezione, soprattutto attraverso la normativa, in particolare perché c'è una alta preoccupazione rispetto allo squilibrio di potere tra gli utenti e chi gestisce i dati. A questo si aggiunge la generale poca fiducia nella possibilità di poter fare giustizia e di rimediare ai danni subiti. Molte volte questo si traduce in un comportamento di autoprotezione, rifiutando la condivisione di dati.

La fiducia ha un ruolo fondamentale per le persone nel decidere se rivelare i propri dati sensibili. La fiducia è declinata principalmente come mancata percezione di rischio di violazioni alla privacy che aumenta con l'esperienza di buone relazioni *online*, ad esempio con una azienda, ma anche con una buona reputazione etica di chi gestisce i dati.

La maggior parte le persone intervistate crede che la prima responsabilità sia dell'individuo che condivide i dati. Secondariamente, ma abbastanza rilevante, è la responsabilità delle aziende che detengono i dati. A seguire, una parte di responsabilità è sentita anche come condivisa con la società, le autorità di regolamentazione, i genitori.

Nonostante l'importanza di questi quattro punti focali di preoccupazione in tutti i focus group, l'enfasi su ciascuno di essi differisce, a seconda del paese natale e dell'età dei partecipanti (LancelotMiltgen & Peyrat-Guillard, 2014).

La ricerca, ad esempio, mette in luce una chiara opposizione tra la Francia, che sottolinea l'importanza della responsabilità e la Grecia che punta sulla fiducia. I partecipanti greci hanno fiducia negli istituti pubblici come importante mezzo di protezione, mentre i francesi accolgono con favore l'intervento pubblico di tutela ma sono anche scettici sul fatto che le normative vengano applicate concretamente.

Gli intervistati provenienti dalla Grecia e dalla Spagna sono convinti di poter scegliere se rivelare o meno i propri dati personali. Le persone che abitano in Polonia ed Estonia, invece, pensano di essere costretti a comunicare i propri dati alle istituzioni fidate, ma tendono a non condividere i propri dati con organizzazioni non fidate.

Circa la metà delle persone anziane e di mezza età teme e percepisce un rischio alla tutela della privacy elevato. I giovani adulti invece hanno poca fiducia ma non sono così diffidenti rispetto al possibile controllo dei dati sensibili. I giovani invece sono più sicuri e fiduciosi, probabilmente perché hanno maggiori abilità tecnologiche, ma anche più responsabili della propria protezione e hanno maggior fiducia nelle tutele normative. Questi risultati sui giovani invertono la relazione causale tra preoccupazione e ricerca di strumenti di protezione negli adulti: nei ragazzi cioè c'è minore preoccupazione, ma anche una maggiore ricerca di strategie di tutela.

Questa ricerca, quindi, mette in risalto due importanti fattori e preoccupazioni per la tutela della privacy. In primo luogo emerge la necessità di una responsabilità condivisa, tra i vari attori coinvolti nella raccolta, nell'utilizzo e nella protezione dei dati, per poter controllare la divulgazione e l'uso improprio dei dati personali. La fiducia appare necessaria per un uso effettivo e pieno del mondo digitale. Le aziende, se vogliono usufruire delle opportunità offerte dai dati digitali, dai Big Data, devono capire come creare e sostenere la fiducia, nel trattamento dei dati personali, di tutti i propri stakeholder. E così i governi e la pubblica amministrazione devono perseguire la fiducia dei cittadini, regolamentando ed attuando tali normative per un diritto al controllo, alla trasparenza nella gestione dei dati e ad una efficace rivalsa in caso di uso improprio dei dati sensibili.

Capitolo 3. I BIG DATA IN AZIENDA

3.1. La Big Data Business Intelligence

I Big Data possono creare nuove fonti di valore aziendale tramite la combinazione delle loro caratteristiche di velocità, varietà, veridicità e naturalmente volume, applicate agli oggetti del business, come clienti, prodotti, concorrenti, e agli eventi di business, come ordini, frodi, sinistri, pagamenti.

Nel mondo del management aziendale il fenomeno dei Big Data è ancora sotto osservazione per capire la corretta adozione ed utilizzo, i limiti e le opportunità di questo immenso mondo di dati. Il management aziendale se ne deve comunque interessare perché verrà sicuramente coinvolto o sconvolto. È importante sviluppare, organizzare e gestire gli appropriati fattori abilitanti di tipo organizzativo dei processi e delle risorse, tecnologico e normativo per creare conoscenza e quindi valore aziendale dai Big Data.

L'investimento necessario con i Big Data è in una cultura aziendale dinamica e interdisciplinare, più basata sui fatti (quindi sui dati), eliminando pregiudizi, ma cercando nell'esperienza e nell'intuito la conoscenza per leggere i dati e quindi generare decisioni più efficaci.

Gli elementi qualificanti i Big Data sono il VALORE AZIENDALE, il più rilevante, seguito dalla VELOCITA' e dalla QUALITA' dei dati. La VARIETA' è importante perché obbliga ad una necessaria multidisciplinarità dei progetti, dato il dominio notevolmente più ampio dei dati utilizzabili e dei campi su cui effettuare le analisi. I VOLUMI rappresentano più le componenti tecnologiche e quindi di minore rilevanza diretta, ma non di minore importanza, per il Management.

È necessario fare una corretta analisi di business della strategia dei Big Data perché il successo deriva da tanti fattori abilitanti e da capacità aziendali soddisfacenti:

- Il giusto *committment* direzionale: ricercare un forte legame tra strategia aziendale ed iniziative Big Data; creare un Comitato Direttivo Inter-funzionale che sponsorizzi gli interessi e i fabbisogni di tutte le aree aziendali; monitorare il ritorno dell'investimento con adeguati *business case* e studi di fattibilità.

- Iniziative circoscritte, mirate ma all'interno di un disegno più grande (*think big but start small and quick!*): evitare iniziative "a silos"; scegliere soluzioni IT innovative e che permettano la crescita dei volumi di dati; sviluppare la gestione dei dati statistici, in movimento e in *streaming*; integrare gli attuali *datawarehouse* per sfruttarli al meglio ed abbattere i costi; sviluppare la gestione della sicurezza e della privacy; mirare a risultati in tempi rapidi.

- Scegliere le opportune fonti informative: valorizzare ed arricchire il patrimonio informativo disponibile, sia interno che esterno, sia strutturato che destrutturato, ma che presenta un elevato grado di qualità ed affidabilità del dato; se si hanno ipotesi da verificare arricchirle di set di dati nuovi, utilizzando un approccio top-down; se invece non si hanno ipotesi da verificare utilizzare i nuovi dati raccolti cercando nuovi modelli interpretativi con un approccio bottom-up.

- Progettare l'organizzazione e sviluppare le competenze: indentificare le nuove figure professionali necessarie; organizzare nuovi modelli di Competence Center e dar loro una collocazione organizzativa; selezionare una *partnership* esterna che riduca i timori per le nuove tecnologie e velocizzi la curva di esperienza e di maturità sui Big Data; identificare il giusto equilibrio di "*make or buy*" nelle diverse fasi della Data Value Chain.

I due approcci classici della Business Intelligence Analytics sono ancora attuali anche con i Big Data. L'approccio top-down classico implica che le richieste del business devono essere verificate ed implementate con i nuovi dati a disposizione. L'approccio bottom-up deve invece partire dai nuovi dati per ricercare nuovi modelli interpretativi e nuova conoscenza.

Uno studio del 2013 è stato condotto su oltre 300 senior IT manager e *decision maker* rappresentanti di aziende di diversi settori, sia nell'ambito privato che pubblico. Tale studio è stato redatto dalla società indipendente Gcl Direct nel Regno Unito e voluta da Software Ag. La ricerca analizza come le organizzazioni stiano adottando nuove tecnologie e processi per indirizzare le sfide di business.

In particolare, la ricerca fotografa il crescente impatto dei Big data e analizza il ruolo del Data Management (Dm) e del Business Process Management (Bpm) nel guidare l'innovazione del business. Obiettivo primario delle aziende di oggi, si legge nel *report* dell'indagine, è catturare, gestire e processare tutti i dati provenienti da varie fonti,

soprattutto quelle nuove come i social media. In tal modo si possono generare in real-time conoscenze utili ai *decision maker* per costruire le basi del vantaggio competitivo di un'organizzazione. Il vantaggio competitivo, infatti, deriverebbe, secondo quanto dichiarano le aziende interpellate, proprio dalla possibilità di modellare e ridefinire i processi di business agilmente usando in modo ottimale tali informazioni.

Negli anni '70 la raccolta dei dati avveniva su supporti magnetici, nastri e dischi, e i dati erano aggregati. Le analisi, quindi, erano statiche e limitate ma soprattutto si basavano soltanto su dati a consuntivo.

Negli anni '80 con l'avvento dei database relazionali e del linguaggio SQL[13], le analisi diventano più dinamiche e le estrazioni si possono fare anche a livello di dettaglio o con diversi tipi di aggregazione. Le basi dati sono direttamente quelle operazionali. L'utilizzo delle basi dati operazionali, però, produce alcune criticità nell'attività di analisi. In presenza di numerosi applicativi, l'uniformità e la coerenza dei dati operazionali non sono garantite, data la differenza di formati, di completezza o di aggiornamento. Il disegno delle basi dati sottostanti agli applicativi operazionali è di tipo OLTP[14], quindi fortemente normalizzato per le attività transazionali come inserimenti, cancellazioni e modifiche dei dati. La normalizzazione aumenta di gran lunga il numero delle tabelle necessarie. È quindi necessaria una complessa attività utente di collegamento tra le varie tabelle per ottenere l'analisi richiesta. Inoltre solitamente i sistemi operazionali hanno una limitata profondità storica dei dati e anche qualora questi fossero presenti sono di difficile estrazione.

Negli anni '90 si introducono dei database disegnati appositamente per le analisi ma alimentati dai sistemi operazionali. Il *datawarehouse*[15] contiene dati integrati, consistenti e certificati di tutti i processi di business dell'azienda. Nasce così la Business Intelligence (BI): un sistema di modelli, metodi, processi, persone e strumenti che rendono possibile la raccolta strutturata, l'elaborazione, l'archiviazione, la trasmissione e la presentazione dei dati aziendali. Il sistema BI permette di trasformare i dati in un database di informazioni, tale da costituire un supporto alle decisioni strategiche,

[13] SQL: Structured Query Language basati su database relazionali
[14] OLTP: On Line Transaction Processing, un sistema per la gestione dei dati e delle transazioni
[15] DATAWAREHOUSE archivio informatico per l'analisi dei dati

tattiche ed operative. Rispettivamente queste informazioni saranno utili nel processo decisionale del management, delle direzioni funzionali, del personale esecutivo.

Con l'evoluzione dei sistemi di BI si è passati a basi dati multidimensionali che fondono dati e metadati, le basi dati OLAP[16], consentendo all'analista di non avere specifiche conoscenze tecno-informatiche ma di concentrarsi solo sulle problematiche di business.

L'evoluzione di oggi è verso le tecniche di Data Mining, cioè analisi dei dati per estrarre informazioni, *pattern* e relazioni non noti a priori. L'utilizzo di tecniche di Data Mining ai fini previsionali è chiamato Predictive Analytics.

L'introduzione dei Big Data e della Cognitive Business non cambiano i propositi della BI: analizzare l'enorme quantità di dati che le Aziende producono per aumentare la profittabilità e la competitività delle stesse aziende. Le sfide di business, quindi, rimangono le stesse: aiutare i manager, gli analisti e i produttori a capire cosa è vero o falso dei dati e come utilizzarli per aumentare i risultati; garantire l'allineamento tra la strategia di business e l'uso della BI, dei Big Data, della Cognitive Business per migliorare i processi aziendali strategici; gestire i fattori complessi dell'organizzazione, come ruoli e responsabilità, per determinare l'effettivo sviluppo delle applicazioni della BI e quindi l'effettivo uso delle stesse nei processi di business.

La Business Intelligence è utilizzata per descrivere l'area delle tecniche di analisi di business, dai report standard alle più sofisticate ed avanzate analisi statistiche. È l'insieme di modelli, dei reports, delle analisi statistiche e dei forecasts pensati per consentire al management di prendere decisioni finalizzate a migliorare i guadagni o ridurre i costi, o entrambi. La BI si basa su un DataWarehouse cioè il database specializzato per archiviare importanti informazioni di business (transazioni, prodotti, clienti, canali, risultati finanziari, misuratori di performance).

La visione di Business dei Big Data analizza le peculiarità di volume, velocità e varietà in merito alla possibile riduzione di costi e/o dell'aumento dei ricavi.

In merito al Volume sicuramente oggi possiamo immagazzinare elevati volumi di dati ad un basso costo. L'importante per il business però è determinare l'utilità dei dati per

[16] OLAP: On Line Analytical Processing, un sistema per l'analisi dei dati multidimensionale

creare valore, quindi non basta accumulare, è necessaria un'attenta e critica selezione dei dati.

Per quel che riguarda la Velocità, con l'esplosione dei social media e di Internet vengono prodotti milioni di dati ogni secondo. Determinare quando e come la velocità è rilevante ed utile per creare valore, è fondamentale per un progetto Big Data.

Dal punto di vista della Varietà, i Big Data catturano una moltitudine di dati, molti non strutturati e variabili nel contenuto e nella forma: stante la difficoltà di creare un dataset per questi dati così difformi dai dati amministrativi della BI, è fondamentale capire se queste diverse forme di contenuto possano aiutare la strategia di business specifica dell'azienda e dei fattori interessati.

Il Cognitive Business è semplicemente l'uso delle tecniche della scienza cognitiva per indirizzare la complessità, la dinamicità e le situazioni di business ambigue. Ci sono differenze e similitudini tra BI e Cognitive Business. Le similitudini sono che ambedue usano gli standard matematici e statistici per analizzare i sistemi dinamici del business. D'altra parte, la BI usa dati strutturati, i tipici dati di business utilizzati da decenni, mentre il Cognitive Business usa sia dati strutturati che non strutturati (dati digitali come foto, video clips, messaggi di testo, immagini di documenti e web log).

Da una prospettiva di business, quindi, è importante ampliare la BI ai dati non strutturati per migliorare i processi, le decisioni aziendali, i prodotti. Quindi il Cognitive Business va visto come un'estensione della tradizionale analisi della BI. Per fare ciò è fondamentale utilizzare il nuovo approccio di analisi insito nei Big Data: anche i dati strutturati, i tipici dati della BI, vanno rivalutati perché i nuovi sistemi di analisi permettono di incrociare, confrontare dati prima chiusi in "silos" a causa delle loro eterogeneità (Williams, 2016). La Business Analytics, infatti, è l'insieme di tutte le applicazioni dell'analisi quantitativa su database di business. L'Analytics di per sé stessa non è una novità. Sono invece un'innovazione importante le piattaforme di analisi (ad esempio SAS e SPSS[17]) che permettono confronti e interrelazioni mai pensate prima, che analizzano i Big Data e le svariate tipologie di dati di cui sono fatti.

Molto rilevante, oltre all'analisi dei dati, è la loro presentazione: anche la visualizzazione fornisce molte informazioni dai dati e quindi proposizioni di valore per le decisioni. Le

[17] Sistemi di analisi statistica dei dati

tipiche applicazioni della Business Intelligence sono sempre stati i Reports: informazioni strutturate dei dati passati dell'andamento, degli eventi e delle performance di business. I *report* possono essere più o meno dinamici, cioè modificabili in base a determinate variabili, e possono essere costruiti *ad hoc*.

Possono essere affiancati o integrati dalle Multidimensional Analyses, ovvero da analisi che sottolineano i *driver* degli eventi, degli andamenti e delle performance di business. Un'evoluzione della visualizzazione riguarda le Scorecards e Dashboards e i KPI (Key Performance Indicator): è una rappresentazione dell'analisi, convenzionale e multidimensionale, che permette una rapida valutazione degli eventi, degli andamenti e delle performance di business tramite strumenti visivi come grafici e diagrammi, semafori e indicatori a tacche.

Questi contengono informazioni principalmente indicative, riassuntive che possono essere il punto di partenza di eventuali approfondimenti. Sono rivolti al management e quindi a supporto delle decisioni strategiche. Da qui nasce la Balance Scorecard (BSC) che analizza la performance aziendale rispetto a quattro prospettive (Kaplan, 1992). La prospettiva finanziaria è valutata tramite gli indicatori finanziari classici (ROI, ROE). La prospettiva del cliente (esterna) rappresentata dai tassi di variazione del numero dei clienti o del fatturato per cliente. La valutazione dei processi (interna) può essere valutata in base alle variabili peculiari del processo aziendale analizzato. La prospettiva di innovazione ed apprendimento non è di facile misurazione, ma può ad esempio essere valutata in base al tasso di turnover, alle ore di assenza o alle ore di formazione.

La Business Intelligence Analytics utilizza Advanced Analytics cioè applicazioni automatiche che dai dati passati di andamenti, eventi e risultati di business riassumono e analizzano trends. L'evoluzione è la Predictive Analytics: applicazioni automatiche che dai dati storici di business cercano di predire o simulare i futuri risultati e gli impatti di business.

Tutte queste forme di BI analizzano i dati storici di *performance* e le loro cause per poter determinare modelli, predire i futuri risultati ed impatti e quindi aiutare ad intraprendere migliori strategie. Quando si definiscono i requisiti della BI si deve identificare anche che tipo di applicazione di BI utilizzare. Il valore dei dati raccolti risiede anche nelle informazioni che riescono a rappresentare al management le peculiarità e le variabili critiche del settore in cui l'azienda opera.

Le tecniche di *data mining* su moli di dati molto rilevanti, i Big Data, possono portare a individuare *pattern* nascosti ed informazioni con molto valore aggiunto. È necessario però ricordare che i dati vanno trasformati, puliti per poter essere utilizzati in un modello predittivo. Una volta, però, consolidato il modello predittivo e le sue fonti di dati, il processo di acquisizione, pulizia e trasformazione dei dati può essere automatizzato ed integrato al sistema di BI.

Gli strumenti di Predictive Analytics possono essere utilizzati per risolvere molteplici problemi di business:

- Ricerca di anomalie (comportamenti fraudolenti)
- Churn Analysis (ricerca dei clienti con un'alta probabilità di passaggio alla concorrenza e quindi evitarne l'uscita)
- Segmentazione della clientela (definire il profilo comportamentale del cliente e quindi attivare delle strategie differenziate)
- Previsioni su serie temporali
- Campagne pubblicitarie mirate (identificare a priori i clienti con un'alta probabilità di acquisto e quindi impiegare in modo mirato le risorse di marketing)
- *Market basket analysis* (identificare ulteriori prodotti acquistabili dal cliente in base ai suoi comportamenti abituali)

Predictive Analytics ha delle fasi predefinite e consecutive. Le prime tre fasi sono già comprese nell'implementazione di un datawarehouse di BI. I dati devono comunque essere trasformati per essere efficientemente impiegati negli algoritmi di *data mining*. La prima e più cruciale fase è la comprensione del business. La comprensione dei dati è immediatamente successiva con la conseguente fase di preparazione e pulizia dei dati. Il fulcro è la creazione di un modello predittivo. Naturalmente tale modello va testato e valutato nei risultati. L'ultima fase, tutt'altro che banale, è l'utilizzo, la messa in opera del modello.

L'idea dello sforzo congiunto e profondo che l'azienda deve compiere per attuare un progetto di BI è ben descritta dal BI Maturity Model. Le 14 variabili di cui è composto indicano tutti i fattori, aree e risorse che devono essere coinvolte. I valori delle variabili descrivono il passaggio da una fase all'altra.

L'obiettivo per ogni azienda è di spostare il proprio posizionamento verso fasi più mature, ma anche di armonizzare il più possibile le variabili all'interno della fase in cui si trova, nel modo più coerente con le scelte e i principi della BI Governance.

Di seguito l'elenco delle variabili del BI Maturity Model.

La strategia aziendale di BI. Nella fase 1 e 2 non esiste una reale strategia di BI in quanto questa è presente solo in isolate funzioni aziendali. Con la fase 3 viene ampliata a tutte le funzioni aziendali e quindi formalizzata in un capitolo dedicato alla BI nel piano strategico IT. Con la fase 4 c'è una vera e propria strategia di BI che punta a creare valore con le informazioni e a far sì che questo vantaggio competitivo sia distribuibile lungo tutta la filiera di business.

Il Budget dedicato alla BI. Nella fase 1 e 2 la BI è solo un costo, una percentuale del budget IT. Nella fase 3 la BI è valutata come fattore critico per il business e quindi all'interno del Budget IT ha una sua definizione precisa in termini di OpEX e CaPex IT[18]. Nella fase 4 il valore della BI diventa strategico e quindi il budget BI si decentra in maniera importante nelle funzioni di business.

La diffusione/penetrazione dei sistemi BI. Questa indica la quota di utenti abilitati alla BI. Nelle fasi iniziali sono praticamente solo utenti della Vendite e Controllo di Gestione (10-15%). Nella fase 3 le percentuali oscillano tra il 20-25%, mentre nelle fasi più evolute si passa ad un 50-70%.

Il grado di copertura dei fabbisogni informativi. Nella fase 1 sono "coperte" dalla BI le funzioni di amministrazione e Controllo e delle Vendite. Nella fase 2 si implementano i KPI e i cruscotti per la Direzione Generale e Divisionale. Nella fase 3 la copertura della BI arriva alle aree di Marketing, Post Vendita ed Operations fino ai sistemi di Business Performance Management aziendali. Nella fase 4 in poi la copertura si estende anche a funzioni normalmente meno coinvolte (IT, Acquisti, HR) e aumenta il grado di penetrazione nelle funzioni già coperte.

Il grado di esperienza nella BI. Nelle fasi iniziali le competenze sono soprattutto di Data Management per gli specialisti e di report dinamici e/o di analisi OLAP per gli utenti

[18] OpEx (dal termine inglese OPerating EXpense, ovvero spesa operativa) è il costo necessario per gestire un prodotto, un business o un sistema altrimenti detti costi di O&M (Operation and Maintenance) ovvero costi operativi e di gestione. CaPex (da CAPital EXpenditure, cioè le spese in conto capitale) indica l'ammontare di flusso di cassa che una società impiega per acquistare, mantenere o implementare le proprie immobilizzazioni operative. (fonte Wikipedia)

all'interno di progetti BI circoscritti. Nella fase 3 si sviluppano competenze applicative: per gli specialisti sulle parametrizzazioni e per gli utenti sulle analisi e sulla interpretazione delle informazioni. Nella fase 4 e 5 gli specialisti possiedono ormai tutte le competenze tecniche necessarie per l'autosufficienza dall'esterno della funzione IT; gli utenti sono indipendenti nelle analisi sofisticate e sviluppano una diffusa capacità analitica in azienda.

L'architettura BI. È una variabile tra le più critiche per evidenziare il grado di maturità della BI. Nelle fasi 1 e 2 la BI è applicata a problemi isolati e con strumenti prevalentemente di *office automation* o comunque specializzati e legati univocamente ad un singolo problema. I dati sono frammentati con *datamart* indipendenti, isolati. Nella fase 3 inizia l'integrazione e la razionalizzazione della BI: nasce un datawarehouse integrato, dove i metadati cominciano a dare una visione complessiva dei dati, delle misure e delle dimensioni aziendali. Cresce, se pur ancora non strutturata, la componente di Analytic Application. Nelle fasi 4 e 5 l'integrazione coinvolge anche l'Analytic Application, lo strato dei dati si articola in più livelli integrati supportando tutte le funzioni aziendali con ampiezza e profondità dei dati differenziate. La BI viene applicata sui processi di management per cui si sviluppa un vero e proprio portafoglio applicativo di BI e BPM. Inizia lo sviluppo di applicazioni BI Near Real Time e di analisi di convergenza tra dati strutturati e non strutturati.

Gli standard tecnologici. L'efficienza tecnica è data dalla costruzione di standard tecnici per la conduzione del progetti di BI (requirements), per la progettazione e lo sviluppo (modellazione dei dati, hardware e software, progettazione di interfacce), per l'avviamento e l'utilizzo (documentazione, formazione/addestramento). Nella fase 1 non ci sono standard. Nella fase 2 si definiscono standard progettuali (costi ed efficienza). Nella fase 3 si avvia un processo di razionalizzazione che riguarda i costi di progetto e di infrastruttura, di addestramento. Nella fase 4 gli standard tecnici sono a livello aziendale e soddisfano tutte le esigenze della BI dal reporting al data/text mining, il modelling di business o l'analisi dei dati georeferenziati o spaziali. I costi sono migliorati sia a livello di progetto che di infrastruttura, oltre a quelli di selezione di sistemi, ridotti i rischi di obsolescenza del software e di disallineamento delle release. Nella fase 5 si un tale consolidamento da poter rispondere alla flessibilità e il cambiamento dell'azienda per il controllo del proprio patrimonio informativo e delle decisioni.

Il Data Quality Management. Nelle fasi 1 e 2 la qualità dei dati è garantita con interventi ad hoc per la soluzione di problemi manifesti. Nella fase 3 viene definita una policy di Data Quality. Nella fase 4 vengono inseriti strumenti specializzati di Data Quality anche per garantire tempestività nell'aggiornamento dei dati. Viene istituita una policy più ampia che coinvolge i dati extracontabili gestiti direttamente dalle funzioni utente. In queste fasi alte è chiaro che il valore patrimoniale dei dati è dato dalla loro capacità informativa rilevante, ampia, affidabile e tempestiva.

La Ownership e Accountability della BI. Tale variabile organizzativa indica la proprietà del budget di BI e alla responsabilità aziendale dei progetti e dei sistemi di BI. Fino alla fase 3 il budget BI è in capo all'IT, mentre dalla fase 4 la responsabilità si decentra anche alle funzioni utente. L'accountability nella fase 1 è dell'IT, nella fase 2 è mista, nella fase 3 è più spostata verso le funzioni utente, di cui diventa esclusiva dalla fase 4 in poi. Le percezioni di ispezione e sanzione nella valutazione degli obiettivi nelle prime fasi si trasformano, a partire dalla fase 4, in reali misurazioni delle performance per scopi di miglioramento continuo.

Le unità organizzative dedicate alla BI. Nella fase 1 il team è ristretto e di specialisti IT. Nella fase 2 il team si struttura e si amplia generando un'unità formale dedicata alla BI, generalmente all'interno della dell'unità di Sviluppo Applicativo dell'IT. Nella fase 3 si distacca e diventa un'unità di line sempre all'interno dell'IT o un'unità di staff al responsabile IT. Dalla fase 4 in poi diventa un'unità di Competence Center su cui converge la governance della BI e le competenze IT e dei Keyuser di business.

Le relazioni specialisti-utenti e livelli di servizio. Due aspetti: la relazione tra IT e BI e i livelli di servizio della BI. Nella fase 1 sono inesistenti. Nella fase 2 inizia una minima relazione tra IT e BI tramite la definizione dei Business Requirements, ancora però solo con i keyuser della BI. Nella fase 3 la relazione si delinea e anche nella funzione IT e vengono identificati team dedicati. La BI inizia a delinearsi come servizio a se stante nell'IT anche se ancora valutata con i medesimi criteri degli altri servizi. Nella fase 4 la struttura organizzativa diventa più strutturata e complessa con l'introduzione di specifici SLA basati su indicatori specifici della BI. Nella fase 5 arriviamo ai Competence Center con team dedicati alla BI tramite la fusione delle competenze tecniche, progettuali e di business. Lo sviluppo della BI è attentamente monitorato tramite KPI di

servizio, processo, di costo e di soddisfazione: il così detto sistema di BI Performance Management.

L'analisi costi-benefici. Nella fase 1 non ci sono valutazioni. Nella fase 2 e 3 la valutazione è solo ex-ante e solo sugli aspetti qualitativi del business, senza nessuna misurazione in termini di incremento di ricavi e/o di riduzione di costi. Nella fase 4 e 5 oltre all'identificazione dei driver di business, si valuta anche l'impatto sul conto economico e quindi poter avere degli indicatori economico-finanziari (ROI, VAN, Payback period).

La misurazione dei risultati. Intesa come misurazione diretta degli impatti della BI sul business. Nella fase 1 non ci sono misurazioni. Nella fase 2 si inizia, in modo informale, a misurare il grado di soddisfazione degli utenti. In fase 3 le misurazioni vengono strutturate e si inizia anche a misurare l'utilizzo reale dei sistemi di BI. Nella fase 4 si aggiungono le misure quantitative degli impatti di business dei sistemi di BI. Nella fase 5 si struttura e si automatizza un sistema integrato di misurazione dei risultati della BI (business, servizio, processo, utilizzo).

La BI sourcing. Comprende sia le modalità di valutazione, la scelta delle tecnologie, i ruoli interessati che le politiche di approvvigionamento della BI. Nella fase 1 nel migliore dei casi si utilizzano gli stessi criteri per altre applicazioni aziendali. Nella fase 2 si identifica un sistema di Software selection dedicato alla BI, dove la funzione IT ha il ruolo principale. Nella fase 3 inizia una vera e propria politica di sourcing della BI, in termini di mix tra skill interni ed esterni, politiche di BI vendor management e di processo di selezione. I ruoli decisionali coinvolgono il BI manager, il CIO e le figure di business (CFO). Nella fase 4 e 5 si sposta il baricentro decisionale dalla funzione IT verso il Top Management.

3.2. Il Data Ring

Anche per i Big Data serve, per estrarre valore dai dati, un metodo che consenta di procedere in modo strutturato e replicabile. Il metodo non sostituisce la definizione di modelli teorici che servono invece per definire il come e il perché di ciò che si analizza.

Il Data Ring ha come precursore e "maestro" il Business Model Canvas di Alex Osterwalder. Questo strumento strategico non può e non vuole risolvere deficit strutturali o ampliare limiti oggettivi. È un metodo per progettare un'analisi puntuale,

reale e dinamica per definire il business model necessario agli obiettivi definiti a priori. La proposta di valore deve essere definita con i suoi annessi bisogni, risorse, priorità, canali di *partnership* ed *asset*.

Il Business Model Canvas è quindi un modello che cerca di dare una rappresentazione il più possibile oggettiva, anche verso terze parti. Fornisce una analisi della proposta di valore definita e dei relativi *partner*, risorse, canali e attività chiave, valutando anche le relazioni con i clienti e i segmenti di clientela interessati. Si tratta di un *tool* per definire infine la struttura dei costi e i flussi dei ricavi, che traccia l'evoluzione delle variabili nel tempo, nei cambiamenti e nelle loro peculiarità.

Il Data Ring nasce come strumento di analisi, flessibile ai cambiamenti ed incentrato nella proposta di valore definita da un progetto Big Data. Ha "un approccio iterativo, volto a raffinare, attraverso ripetuti cicli di operazioni concatenate, la comprensione delle dinamiche fenomenologiche, a consentire la descrizione sempre più puntuale di cause ed effetti e a permettere l'individuazione di comportamenti emergenti non banali e di non facile osservazione" (Camiciotti & Racca, 2017).

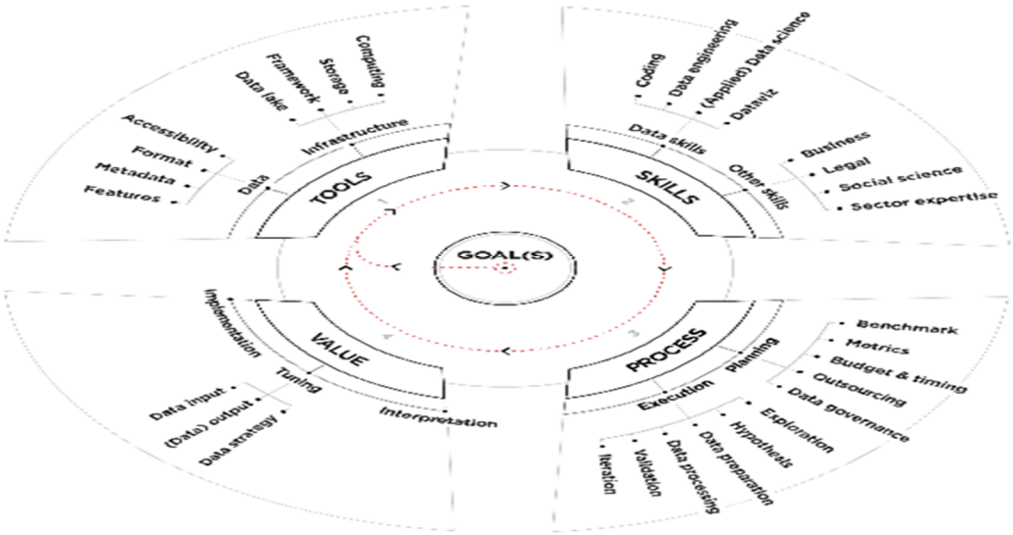

Fig. 1 - Il Data Ring

Come si può vedere nella rappresentazione visiva del Data Ring, al centro, come fase indispensabile e preliminare, troviamo i *goal*. Gli obiettivi devono essere misurabili, come quantificabili devono poter essere le cause e gli effetti di questi *goal*.

Un approccio *data-driven* non prescinde dall'identificare chiaramente l'esigenza del business e dei fattori chiave per la sua risoluzione. La valorizzazione del patrimonio

informativo è da sempre ricercata; con i Big Data si aggiungono solo caratteristiche diverse dei dati come volume, varietà, velocità e digitalizzazione. Quindi si tratta di integrare all'interno di processi questi dati perché possono incrementare di molto la capacità di comprensione umana.

Si presuppongono quindi due livelli nell'identificazione degli obiettivi: partendo da *goal* specifici, spesso ristretti ad un unico dataset, si arriva agli obiettivi di livello superiore, cioè all'integrazione dei risultati ottenuti all'interno dei processi di business.

Gli obiettivi di primo livello devono essere definiti da risorse con competenze elevate sul processo individuato e che possano, quindi, individuare le aree di miglioramento e i *benchmark* di riferimento per valutare il progetto.

Gli obiettivi di secondo livello comportano invece modifiche strutturali a procedure, competenze ed elevati investimenti. Questi *goal* perciò non possono che essere definiti da manager di alto livello con potere decisionale.

Malgrado i livelli diversi, queste due tipologie di obiettivi vanno definite a priori e soprattutto sono concatenate fra loro. Fissare solo gli obiettivi di secondo livello vuol dire non coinvolgere nell'analisi le risorse esperte del processo "pilota". Definendo invece solo gli obiettivi di primo livello, potremo trovarci nella situazione in cui i risultati non sono applicabili in modo strutturato.

Malgrado con i Big Data ci sia effettivamente il piacevole rischio di intuizioni e scoperte non pianificate, non definire gli obiettivi può portare ad elevati investimenti senza nessun risultato o con delle migliorie non applicabili al proprio business.

Un progetto Big Data presuppone sempre una riorganizzazione in termini di responsabilità, un adeguamento delle competenze, nuovi strumenti e ridefinizione del processo aziendale.

Infatti il Data Ring si compone di quattro blocchi principali collegati ai *goal*: strumenti, competenze, processo e valorizzazione.

Gli strumenti principali con i Big Data sono appunto i dati e le infrastrutture tecnologiche.

I dati raccolti devono essere adeguati all'obiettivo prefissato e più la conoscenza del dataset è elevata meno dispendioso sarà valutare l'attendibilità e l'accuratezza dei dati

raccolti. Dall'altra parte tutti i dati selezionati dovrebbero essere corredati dei propri metadati per renderli riconoscibili, comprensibili ed utilizzabili, sia dalle macchine che dalle risorse non coinvolte nel processo di selezione del dato. Questo determina il primo vincolo all'accessibilità ai dati, ma non il solo. Ci sono barriere tecnologiche all'accessibilità perché si utilizza una diversa infrastruttura IT per raccogliere ed elaborare i dati o perché i formati sono diversi. C'è sempre una barriera normativa, più o meno alta, che incide sulla piena disponibilità dei dati. Infine, le barriere strategiche non possono mancare vista l'enorme potenzialità assegnata ai Big Data e possono portare a negare l'accesso come a richiedere un compenso per l'utilizzo dei dataset.

Gli obiettivi vanno allineati anche alle infrastrutture IT in merito alle loro capacità di *storage*, accesso ed elaborazione. Esistono infrastrutture in Cloud che aiutano non poco, permettendo l'aumento di risorse computazionali o di *storage* quasi immediato. A volte però il ricorso a queste soluzioni esterne può essere critico per una azienda se i dati rappresentano un *asset* strategico. In aziende di grandi dimensioni le infrastrutture IT sono molto consolidate e gestire migrazioni verso nuovi applicativi *on line* o *near on line* può essere molto impegnativo e creare disservizi verso i clienti. Allo stesso tempo transizioni di questo tipo necessitano di competenze che potrebbero non esserci in azienda e che quindi necessitano di tempo per la formazione.

Una infrastruttura *data driven* ha quindi delle caratteristiche ben definite. 1) La scalabilità, cioè la capacità di aumentare proporzionalmente alle necessità che siano di *storage*, di risorse computazionali o di infrastrutture di *networking*. 2) La flessibilità, perché un progetto Big Data è sempre innanzitutto esplorativo e quindi le ipotesi iniziali vanno validate o modificate. 3) La bassa latenza, cioè la velocità nell'elaborazione dei dati che si deve allineare alla velocità dei Big Data. 4) La ridondanza e l'affidabilità, la *data optimization*, il *parallel processing*, cioè cercare di evitare perdite di dati per inaffidabilità del sistema, per ridondanza informativa o per sovraccarico del sistema.

Il Data Ring serve anche per determinare le competenze esistenti e necessarie al progetto. Il coinvolgimento delle risorse che hanno in carico attualmente la gestione dei dati è necessario. Come precedentemente detto, il dato deve essere integrato con il suo contesto e con i suoi metadati, ma nella realtà è facile che queste informazioni risiedano nell'esperienza delle risorse umane. Passare da una gestione umana del dato a quella della macchina vuol dire estrapolare tutte le correzioni, i dati mancanti e tutti i

parametri derivati a cui l'esperienza umana sopperisce. Questi devono diventare istruzioni, classificazioni e parametri all'interno degli algoritmi.

Le competenze del Data Scientist devono essere di *software developement* e *system administration*, nonché scientifiche, di *data visualization*, di *business development*, di privacy e *data protection*. Tutte queste competenze sono indispensabili per valutare gli obiettivi, per operare sui dati e gestire e possibilmente introdurre i risultati nel processo di business. Malgrado esistano risorse che hanno tutte queste competenze interdisciplinari e una buona esperienza sul campo, non è facile reperirle nel mercato del lavoro o formarle internamente.

Probabilmente un gruppo di lavoro con esperienze e studi differenziati può produrre un sano confronto, diverse visioni. Oltre a questo, permette al Data Scientist di dover solo organizzare e gestire molte delle competenze necessarie perché il suo *team* ha già al suo interno dei professionisti nei vari settori. Naturalmente i diversi linguaggi e i metodi di lavoro dei componenti del *team* sono allo stesso tempo una risorsa e una sfida all'integrazione.

Gli obiettivi, con i necessari strumenti e competenze, devono definire un processo per l'esecuzione del progetto. Con il passaggio dalla definizione degli obiettivi alla loro esecuzione si capisce immediatamente se il problema è stato identificato chiaramente e se è misurabile. Infatti il processo va indirizzato tramite l'esplicitazione di metriche di valutazione.

Come in ogni progetto, per definire e valutare un processo ci deve essere una pianificazione e un controllo di gestione, un *benchmark* di riferimento, un *budget*, una *timeline* e una gestione delle attività in *outsourcing* o tramite *partnership* strategiche. In un progetto Big Data questi aspetti, se pur definiti a priori, devono anche e continuamente essere adattati all'esplorazione dei dataset e alla validazione delle ipotesi. La definizione di così tanti aspetti prima di iniziare l'attività operativa serve a rendere efficace e concreto un progetto Big Data che di per sé ha una natura incerta.

Le ipotesi iniziali vanno continuamente raffinate percorrendo tutto il processo di esplorazione, pulizia e arricchimento dei dati, di implementazione degli algoritmi e dei modelli predittivi, di validazione delle ipotesi con i risultati ottenuti. Quello che si ottiene può essere già noto, o non rilevante, oppure all'opposto portare a risultati inattesi e che comportano l'allargamento del perimetro del progetto.

La fase determinante per l'applicazione dei risultati per risolvere i problemi e migliorare il processo di business obiettivo del progetto è l'interpretazione. Distinta dalla semplice spiegazione, l'interpretazione è frutto della conoscenza approfondita del contesto e delle leggi che regolano i risultati. È una fase molto complessa e critica che può portare sia alla creazione di valore che alla sua distruzione. Il passaggio dalla correlazione alla causazione senza la conoscenza dei fenomeni analizzati può condurre ad errori di alto impatto.

L'impatto deve essere misurabile, identificando indici e misure che possano effettivamente rappresentare gli aspetti qualitativi e quantitativi dei risultati. Il fenomeno va misurato in relazione al sistema di cui fa parte, per determinare l'impatto anche degli eventuali effetti collaterali che possono portare a risultati incontrollati.

Calibrare i risultati ottenuti con le ipotesi iniziali è la vera parte attuativa che si declina, appunto, nell'analisi dei dati input, output e della struttura del sistema analizzato. La scelta dei dati input ha una elevata discrezionalità umana. Malgrado si possa procedere con campionamenti cercando di misurare con piccoli dati alcune caratteristiche del processo, non è facile e a volte non è corretto generalizzare i risultati. I dati che vengono messi a disposizione dell'algoritmo e la loro qualità sono fondamentali per l'attività di addestramento dell'algoritmo. Dati errati o dati che deviano dall'oggetto di analisi possono portare a risultati sbagliati. Con una attività di riciclo continuo i dati vanno allineati al comportamento desiderato del sistema.

Attraverso la misurazione dei dati output si ricalibrano anche i dati input. È un *loop* che crea valore ed è necessario malgrado a volte questa attività sia onerosa in termini di risorse e di tempo impiegato. Nell'analisi dei dati output è determinante la competenza del Data Scientist, soprattutto nella visualizzazione dei risultati, perché anche questa attività è una interpretazione dei dati. Inoltre, se il Data Scientist ha saputo definire a priori obiettivi precisi, questi possono guidarlo nell'evidenziare risultati inaspettati o che si discostano da quanto ci si aspettava dal sistema. Ridurre gli errori permette alla delicata fase di interpretazione di essere meno incerta.

Per attuare un progetto Data Driven l'azienda deve innanzitutto dotarsi dell'organizzazione, delle competenze, della tecnologia e del *budget* appropriato per creare valore dai dati. L'approccio deve essere sistematico ed organizzato,

predisponendo a priori le tecnologie, la formazione, i ruoli e le responsabilità del processo.

Il Data Ring può essere utilizzato per questo scopo in modo analitico, cioè effettuando una *checklist* della presenza o meno dei requisiti fondamentali per un progetto Big Data. Lo si può utilizzare, invece, per descrivere in maniera esaustiva l'intero processo, nelle comunicazioni alla stampa o per pubblicazione scientifica. Per l'attuazione del progetto il Data Ring va utilizzato generando *feedback* continui, con cicli di iterazione dove il punto di partenza e il punto di arrivo sono spesso sovrapposti, dove nuovi obiettivi potranno nascere proprio dall'analisi dei dati. Non c'è uno specifico punto di ingresso o di uscita dal Data Ring perché la definizione cambia in base al progetto e all'azienda che lo sta utilizzando. Da questo deriva la forma circolare del Data Ring e l'equidistanza dal centro, e quindi dagli obiettivi, delle componenti necessarie al progetto.

I settori che ruotano all'interno del Data Ring, competenze, strumenti, processi e valorizzazione, rappresentano tutta l'azienda. Ciò vuol dire che un progetto *data-driven* è guidato dai dati, ma porta con sé tutti i settori fondamentali di una azienda. Non può essere quindi un progetto limitato all'area informatica o al team del Data Scientist o ad una funzione aziendale in particolare.

L'obiettivo individuato dai manager dell'azienda può essere molto difficile da raggiungere, ma va raggiunto a piccoli passi. L'incertezza e la complessità del digitale possono facilmente portare ad un fallimento. Per meglio dire, il progetto può e spesso deve partire come un prototipo, ma deve comunque interessare tutti i fattori.

Un grande progetto richiede risorse elevate, in termini di costi economici e sociali, di tempo e di reputazione, che potrebbero non essere ripristinabili. Persino le elevate tecnologie e i Big Data potrebbero non essere così disponibili come sembra, per dati non accurati o per i tempi necessari per definire un *tool* adeguato rispetto agli obiettivi. Un progetto che parte troppo in grande potrebbe riscontrare anche barriere interne, riluttanza al cambiamento o avversione al rischio per mancanza di coinvolgimento nelle innovazioni.

Per aziende consolidate può essere utile un approccio al digitale "stile start-up". Facendo piccoli investimenti in ridotti *use case*, cioè facendo degli esperimenti, è possibile valutare le ipotesi, i fattori abilitanti e la risposta di clienti ed utenti all'innovazione apportata. Inoltre è più facile stabile una metrica e degli indicatori di performance per

ottenere *feedback* immediati. La strategia, i processi, le competenze e gli strumenti possono velocemente assorbire i cambiamenti necessari per allinearsi agli obiettivi. Si può avere così una stima degli investimenti e dei rendimenti più reale, che, anche se strettamente collegata ai miglioramenti specifici dell'esperimento, può definire a grandi linee costi e ricavi dell'obiettivo.

In questo modo si crea una sinergia tra le doti possedute da una azienda consolidata e quelle di una start-up. Le imprese storiche, con anni o anche secoli di attività, hanno a disposizione clienti paganti, risorse finanziarie, un'elevata mole di dati sui clienti e risorse professionali in azienda. Le start-up hanno agilità ed innovazione organizzativa, di comunicazione e strategica.

Un'azienda per diventare *data-driven* deve prima di tutto investire sui propri processi di business, eliminando le distorsioni. I processi aziendali devono diventare fluidi e semplici. È necessario individuare tutti i punti che creano interruzioni al processo o attività di supporto massivo da parte dei dipendenti o dei manager. Cominciare ad automatizzare proprio questi parziali processi potrebbe essere una buona strategia per iniziare ad aggiungere valore. Diventeranno processi più veloci, più economici, più convenienti, incorporando le competenze prima affidate alle risorse umane dell'azienda. E questo probabilmente diventerà un valore anche per una trasformazione organizzativa. Piccole fasi di processo automatizzate aiuteranno i dipendenti ad acquisire, senza ritrosia, più dimestichezza con le tecnologie digitali. Ma non solo: adotteranno facilmente anche un nuovo modo di lavorare. Il tempo e l'impegno liberato dalla eliminazione delle distorsioni del processo potrà essere sfruttato dai dipendenti per acquisire nuove competenze e professionalità.

Tutti i settori del Data Ring, competenze, strumenti, processo e valorizzazione, devono essere presi in considerazione anche negli *use case*. Per quanto ridotti negli impatti, questi influenzeranno e saranno influenzati comunque da tutti i settori, e attraverso ripetizioni di cicli e ricicli porteranno agli obiettivi definiti.

La forma circolare del Data Ring rispetto alla tipica forma a piramide della Business Intelligence non cambia i fattori abilitanti al progetto digitale.

Gli strumenti, cioè i dati e l'infrastruttura per governarli, sono la Data Quality Management, il grado di copertura dei fabbisogni informativi e l'infrastruttura tecnologica del BI Maturity Model.

Le competenze, la Data Science, le conoscenze di business e legali, sono racchiuse nella variabile che determina il grado di esperienza nella BI.

Il processo nel Data Ring prevede l'individuazione di *benchmark*, metriche, *budget*, *timing*. Nel BI Maturity Model sono identificate variabili che misurano le relazioni specialisti-utenti e i livelli di servizio, l'analisi costi e benefici, la misurazione dei risultati e l'*ownership* nonché l'*accountability* della BI.

Infine, la valorizzazione, cioè la valutazione d'impatto e il tuning, possono corrispondere alla strategia e al *budget* proprio della BI.

La forma circolare del Data Ring rispetto alla tipica forma a piramide della Business Intelligence però cambia l'operatività, l'esecuzione del progetto.

La creazione del valore dai dati nel Data Ring è quasi un *loop*. Rispetto al *goal* determinato si può partire dal settore ritenuto più rilevante per l'azienda e per lo specifico progetto, ma tutti i settori vanno ingaggiati. Attraverso cicli ripetuti, ogni nuovo ciclo si modifica in base ai *feedback* del ciclo precedente. Così tramite il *tuning*, per approssimazioni successive, in una spirale di miglioramento continuo, il progetto si avvicina sempre più al centro, ai *goal*.

Diversamente, la Business Intelligence mette alla base della piramide gli strumenti, i dati e le infrastrutture per gestirli e l'architettura della BI. Più in alto inserisce le competenze, il Data Science e l'organizzazione di unità e ruoli specifici della BI. Nella parte più alta della piramide ci sono le competenze manageriali che servono per estrarre valore dai dati. Naturalmente la Business Intelligence è un sistema per produrre conoscenza ed informazione dai dati, mentre il Data Ring è un modello per qualsiasi obiettivo sia perseguito dalla strategia dell'azienda. Se la Business Intelligence è un obiettivo dell'azienda può stare al centro del Data Ring, se l'obiettivo strategico è un altro, la Business Intelligence diventa parte di tutte le fasi del Data Ring.

3.3. Big Data e velocità aziendale

Big Data e i nuovi fenomeni aziendali nel campo del Business Management collegati all'impiego delle ICT hanno come comune denominatore la VELOCITA': Real Time Enterprise, Data Explosion, Big Data (dati strutturati e non, *social web data*, *location data*, dati IOT, M2M), Next Generation Business Intelligent e Analytics (Real Time BI, BI

embedded nei processi operativi aziendali), In Memory Computing, Open Innovation e Co-Innovation, Global, Networked e Liquid Enterprise.

La velocità è condizione sempre più determinante per svolgere processi ed attività aziendali. Grazie alle economie di velocità, i Big Data, i nuovi dati, non avranno più distinzione tra dati operativi e dati direzionali. Lo stesso dato operativo sarà utilizzato per l'attività come per l'analisi, non sarà più necessario manipolare i dati operativi contabili e spostarli dai database transazionali ai database dedicati all'analisi. I dati operativi saranno disponibili in strutture dati libere con altissima velocità di accesso.

La velocità senza controllo non genera tuttavia valore e oggi il controllo è esercitato da un mix di persone e sistemi informatici. Quindi altrettanto importante è gestire le relazioni digitali tra le persone, con l'ubiquità dei dati è necessaria anche quella degli individui.

La velocità in azienda è sicuramente riduzione del tempo necessario per l'esecuzione di una attività. Oggi però la rivoluzionaria novità per la gestione aziendale è la velocità applicata alle strategie e ai processi aziendali. Deve essere possibile il cambiamento repentino, in corsa e in tempo reale degli obiettivi o delle relazioni con altri processi, senza lunghe attese dovute alla ricerca, all'elaborazione e alla presentazione dei dati, con un margine di errore ridotto ed accettabile. Come un navigatore, il sistema informativo deve essere sempre disponibile, fornendo un supporto continuo e in tempo reale, elaborando sia i dati storici disponibili, ma anche i dati in tempo reale, soprattutto quelli relativi agli eventi imprevisti, e i dati delle decisioni.

La complessità aziendale è di fatto generata da due fattori prevalenti: la crescente velocità dei cambiamenti necessari e la crescente incertezza degli scenari e quindi dei risultati delle scelte strategiche possibili.

I cambiamenti tecnologici vanno controllati attraverso le leve della cultura aziendale, e quindi delle persone, per poter creare valore. Il Management deve approntare strategie per trasformare gli atteggiamenti delle persone verso le ICT, passando dal fenomeno di moda, dalla paura, dal rifiuto al reale utilizzo delle tecnologie nella gestione aziendale. Deve velocizzare i processi di Innovation Management, sia di prodotto o di servizio che di processo.

Il management deve riuscire a far funzionare insieme elevate transazioni ed attività aziendali con l'analisi dei dati, strutturati e non, online e offline, generati da tali processi operativi.

Da parte sua il management deve velocizzare le decisioni aziendali attraverso l'utilizzo delle informazioni rilevanti, quindi opportunamente selezionate, e con sistemi di comunicazione e di collaborazione ad alta velocità.

Aumentare i cicli e i ritmi di misurazione, di monitoraggio e di analisi permette di ridefinire velocemente i percorsi, aumentando le capacità di relazionare i dati, reagendo in modo esperto, deduttivo tramite fatti documentati in continuo.

Le aziende devono essere coscienti che il sogno del *Real Time* o del *OnLine* è possibile solo per alcuni business, mentre il *Near Real Time* o il *Near OnLine* è un obiettivo fattibile e offre la velocità necessaria per nuove opportunità nelle analisi e nelle decisioni. Questa velocità però si deve concretizzare nell'intero ciclo di *Analyse, Plan, Act and Control*: gli *insight* devono concretizzarsi in azioni innovative, fattibili, veloci ed economicamente convenienti.

La velocità aziendale assume diversi significati. Riferita al *Time to innovation* implica la riduzione del tempo tra idea, concetto, *testing* e sviluppo del prodotto o servizio, ma anche una maggiore frequenza di nuovi prodotti o servizi. Nel *Time to market* una maggiore velocità riguarda la riduzione del tempo tra la disponibilità del nuovo prodotto e il suo lancio nel mercato. Nel *Time to profit* si tratta di identificare velocemente la nuova domanda e valutarne la redditività. Lo sviluppo della Co-Innovation con i clienti attuali o potenziali permette di ridurre questi rischi, migliorando la puntualità, nel senso di velocità nei processi aziendali, e la tempestività di business, riuscendo a primeggiare nell'azione prima che perda l'efficacia commerciale.

L'innovazione può essere radicale, casuale, saltuaria oppure può essere incrementale. L'innovatività è strutturare processi e competenze tali da mantenere costantemente allineata la posizione tecnologica dell'impresa, quindi il vantaggio competitivo. È significativa la *cadenza* con cui si rilasciano le innovazioni sul mercato.

La velocità dell'innovazione deve avere i seguenti importanti aspetti (Baglieri, 2012, citato da Pasini & Perego, a cura di, 2012). La velocità è relativa, dipende dal ritmo di innovazione richiesto nel mercato in cui opera l'azienda, quindi dalla velocità che è

fattore di successo competitivo. L'ascolto del cliente è molto più efficace della tecnologia nell'anticipazione delle esigenze del cliente, ma la tecnologia aumenta la velocità nel catturare e reinterpretare i *feedback* del cliente in modo originale. La semplicità è fondamentale, perché una risposta veloce ma complessa non soddisfa il cliente.

La velocità dei processi aziendali ha l'obiettivo di generare efficienza, produttività, efficienza e standardizzazione. Diventa cruciale se la produzione è trainata dalla domanda, dalle esigenze del cliente, con frequenza di consegna elevate e quantità di prodotto sempre più piccole e variabili nel tempo.

La velocità dei processi aziendali passa per l'implementazione dell'ICT, che però può essere visto come facilitatore tramite automatizzazione, dematerializzazione e informazione veloce o come ostacolo per la scarsa integrazione a causa dei diversi sistemi informativi.

La velocità di reazione ad eventi inattesi corrisponde alla resilienza, ovvero l'abilità a reagire e ri-adattarsi e ri-organizzarsi, facilmente e velocemente, ad eventi inattesi, spesso negativi. Effettuare una mappatura delle vulnerabilità individuando le azioni che possono ridurre o annullare gli impatti sull'operatività dell'azienda è il primo passo verso un'azienda *resiliente*. Poche aziende hanno sistemi decisionali basati su tecnologie di BI che possano sviluppare scenari alternativi per anticipare eventi inattesi come attacchi dalla concorrenza, aumenti dei prezzi delle materie prime o uscita di risorse chiave. Gli indici e le misure significative di *performance*, di tipo finanziario o meno, sono utili non tanto per determinare il fuori norma statico, quanto nel definire limiti di attenzione o intervalli di criticità nel loro continuo monitoraggio, a cui devono seguire contromisure tempestive. L'esternalizzazione di attività e processi spesso può vincolare questa resilienza perché si scontra con la rigidità contrattuale e i modelli di economicità dei fornitori.

L' ICT ricopre ormai un ruolo fondamentale nelle analisi, ma le soluzioni tecnologiche devono essere flessibili per lasciare gli utenti liberi di fare percorsi cognitivi e associativi che spesso non possono essere determinati a priori. Inoltre deve essere possibile l'accesso all'intero patrimonio di dati aziendali senza elaborazioni preventive che ne potrebbero modificare il contesto.

Le tecnologie In-Memory combinano innovazioni *hardware*, come riduzione del tempo di accesso ai dati, velocità di elaborazione ed elevata capacità di *storage*, ma anche

innovazioni *software*, permettendo l'elaborazione dei dati a livello del database, quindi in memoria centrale, e non a livello di applicativo.

Le tecnologie In-Memory nella loro applicazione possono seguire approcci differenti:

- Modello Associativo. I dati vengono caricati in un modello associativo, in base alle relazioni esistenti tra i diversi dati, nella memoria centrale (non c'è uno schema pre-ordinato).
- In-Memory OLAP. I dati sono caricati nella memoria centrale e le query vengono elaborate on demand.
- Excel In-Memory Add-In. Grosse moli di dati sono caricati in Excel, ma dove le relazioni tra i dati sono automaticamente definite tra i diversi dataset.
- In-Memory Accelerator. I dati sono caricati a livello di memoria centrale, ma fa leva su un sistema di indici precostituiti per velocizzare i tempi di risposta.
- In-Memory Visual Analytics. Combina un In-Memory database con i tool di *visual data exploration*, quindi veloce accesso ai dati sfruttando un ambiente grafico ed interattivo.

Le tecnologie In-Memory permettono quindi una elevata velocità di elaborazione, una elevata scalabilità ed una elevata compressione: sono performance che permettono di rivoluzionare i processi aziendali operativi e di produzione delle informazioni.

I benefici della velocità aziendale sono, in primis, i benefici in termini di velocità dell'innovazione, di esecuzione dei processi aziendali, di reazione ad eventi inattesi. L'informazione di ogni tipo deve essere disponibile in ogni luogo e in ogni tempo. Necessariamente però deve riguardare dati attendibili e completi.

Il superamento della distinzione tra sistemi informativi transazionali e sistemi di BI porta quindi alla riduzione del tempo di spostamento dei dati ma anche, e soprattutto, alla possibilità di avere i dati con la stessa granularità dei database transazionali.

Il miglioramento della catena di *performance* tecniche migliora anche le *performance* aziendali. I dati sono accessibili alla "velocità del pensiero" mantenendo così la concentrazione degli utenti, che altrimenti nel tempo di attesa sono distratti da nuove attività. La così detta Real Time Information (o Near Real Time) può essere oggi ottenuta

con la tecnologia In-Memory, portando tutti i dati e le elaborazioni nella memoria centrale.

Questa tecnologia permette una libertà di analisi che con i sistemi tradizionali di analisi non c'era, perché gli ambienti erano pensati, strutturati ed organizzati a priori. I sistemi In-Memory mettono a disposizione tutti i dati senza aggregazioni, vincoli e dataset predefiniti, lasciando così liberi gli utenti di navigare ed utilizzare il proprio percorso logico ed associativo.

Il minor Time-to-Delivery si ottiene se la maggiore velocità di implementazione dei *tool* e delle applicazioni di BI si traduce in velocità di risposta ai fabbisogni informativi espressi dal business, quindi in velocità aziendale.

La Piramide di esperienza della Business Intelligence viene scalata più velocemente tramite l'incremento dell'effetto esperienza derivante dalla possibilità di una navigazione veloce e libera con le tecnologie In-Memory. La base della piramide "Capire e dare un senso al passato e ai risultati aziendali" passa a "Anticipare i problemi e guidare il business" poi a "migliorare i processi aziendali chiave dell'azienda, soprattutto dei processi che interfacciano con i clienti e i fornitori", per raggiungere la maturità della BI con "nuovi prodotti e servizi" e "cambiamento del Business Model dell'azienda". Percorrere più velocemente la "scala dell'intelligence" significa ridurre la complessità e la stratificazione dell'architettura riducendo così le competenze necessarie e quindi ampliando il numero di potenziali utenti che possono sfruttare la BI (Pasini & Perego, a cura di, 2012).

Le tecnologie In-Memory possono incrementare i costi iniziali, ma sono sicuramente più semplici da implementare e gestire dopo. Questo riduce i costi di avvio e gestione dei progetti di BI e Analytics. Inoltre i motori di analisi, le Analytical Piattform sono molto più semplici e quindi meno costosi.

L'estrazione dei dati, con il volume elevato di oggi, sta diventando sempre più una attività *time-consuming* e complessa. La capacità di accedere rapidamente a grandi volumi di dati per scopi di analisi senza dover costruire complesse architetture di *datawarehousing*, quindi la minore complessità di gestione IT, è uno dei principali vantaggi delle tecnologie In-Memory.

La velocità senza controllo, senza finalità, non genera utilità. La velocità oggi a disposizione delle aziende può ridurre il tempo di esecuzione delle attività, ma soprattutto la finalizzazione dell'attività, modificando "in corsa" obiettivi e relazioni con altre attività. Le persone, il management aziendale devono saper convivere con la velocità, per poterla sfruttare al meglio e generare performance incrementali. Le tecnologie In-Memory possono innescare un circolo virtuoso tra rapidità dei processi aziendali, analisi dei dati e decisioni strategiche tempestive.

Fondamentale diventa quindi la conoscenza accurata dei processi e dell'esecuzione delle parti interessate all'automazione. Molti processi che sulla carta sono ben definiti, nella realtà sono soggetti ad interventi di risorse umane, dotate di elevata esperienza e professionalità, per effettuare correzioni. Spesso esistono processi reali sottostanti che nulla hanno a che fare con la sequenza di attività dettagliata nelle specifiche. Questi "processi ombra" spesso sono necessari per eliminare barriere non superabili con i processi standard, o per mancanza di tempo e risorse finanziarie. A volte non esiste nemmeno una documentazione specifica sul processo o non è aggiornata con le modifiche effettuate nel tempo. Sicuramente è una situazione che è più facile trovare in aziende di piccole dimensioni, ma anche nelle imprese di grandi dimensioni spesso la conoscenza dei processi è solo nella testa di poche risorse umane.

Una veloce implementazione di una automazione di qualsiasi perimetro del processo parte dal sistemare queste "manualità" e definire chiaramente le sequenze del processo e le interrelazioni con altre funzioni dell'organizzazione aziendale.

Inoltre, la documentazione deve essere disponibile anche dopo l'automazione e deve essere sempre aggiornata con le modifiche al processo. Le modifiche future e le nuove aree di automazione così possono essere facilmente, efficacemente e velocemente implementate.

Cambiare le modalità di lavoro, specialmente quando sono costruite in anni di esperienza, può non essere facile. È necessario prima di tutto velocizzare le attività lavorative gestite da persone fisiche. L'eliminazione della burocrazia, la semplificazione dei processi, l'eliminazione dei silos di attività, rende già la gestione dell'organizzazione più fluida, veloce. Ma soprattutto è la premessa per una azienda che vuole diventare *data-driven*. Queste transizioni, che dovranno essere effettuate, non si possono limitare

ad una singola funzione aziendale, perché l'automazione nel lungo periodo ricadrà sull'intera organizzazione dell'azienda.

I vantaggi dell'automazione spesso consistono in un ridimensionamento dell'attività lavorativa. Se tale riduzione di risorse fisiche non può velocemente essere reindirizzata in attività più qualificate e più critiche per l'azienda, il valore dell'automazione si andrà a perdere. La formazione delle risorse umane al cambiamento e all'automazione è fondamentale.

L'aggiornamento delle competenze e delle abilità dei dipendenti deve strutturarsi per procedere alla stessa velocità del processo di automazione. In modello di business *data-driven* è fondamentale la velocità, nell'agire, nel verificare il successo o l'insuccesso, nel cambiare direzione o modificare i percorsi. Si tratta di un ciclo che si svolge all'interno di un contesto, quello di oggi, complesso ed incerto. Le risorse umane, se non inserite, o meglio non coinvolte, in questo ciclo dell'automazione creeranno colli di bottiglia, difficilmente eliminabili e con ripercussioni sull'organizzazione aziendale ma anche sociale.

La cultura e la gestione del cambiamento devono diventare la cultura aziendale e la modalità di lavoro per tutta l'organizzazione dell'impresa. La gestione del cambiamento deve analizzare l'attività lavorativa del gruppo e non del singolo lavoratore, principalmente perché il cambiamento introdotto per migliorare il processo potrebbe essere, dal singolo, interpretato come punizione anziché vantaggio. Dall'altra parte l'analisi del gruppo di lavoro permette di identificare dei comportamenti, delle lacune comuni che quindi possono definire aree di intervento migliorativo senza che nessun dipendente affronti il cambiamento come un attacco a sé stesso.

I manager devono far sì che tutta l'organizzazione aziendale veda un miglioramento continuo nella realtà, identificare le strutture che possono al momento supportare il miglioramento continuo e su quali funzioni aziendali invece questo atteggiamento non è sentito. Tutta l'organizzazione deve utilizzare strumenti efficaci per promuovere la cultura del cambiamento, non solo il Data Scientist. Per una nuova cultura aziendale *data-driven* è necessario aumentare velocemente il numero di persone istruite sugli strumenti per dell'analisi dei dati e l'automazione. Le persone escluse facilmente si sentiranno inferiori in abilità e competenze, con un possibile rifiuto o comunque una percezione di svantaggio più che di vantaggio nell'utilizzo delle nuove tecnologie. I

manager devono identificare all'interno del progetto di automazione anche le competenze necessarie per ideare, eseguire e implementare l'analisi dei dati, valutando dove fosse necessario intervenire per formare le risorse umane.

L'elevata velocità di passaggio dagli algoritmi ai reali cambiamenti nell'attività lavorativa impone che i dipendenti vengano coinvolti da subito in una nuova cultura *data-driven*, al contrario dell'agire comune dove la massima priorità viene data all'esecuzione del progetto di Data Analytics. Una azienda che vuole adottare un modello di business guidato dai dati e dalle nuove tecnologie deve anche strutturare ed organizzare in modo nuovo la comunicazione, la formazione e la *governance* dell'organizzazione.

La fase di definizione dell'ambito del progetto diventa così il presupposto per il coinvolgimento delle risorse umane sulla cui attività lavorativa andrà a ricadere l'automazione. Il coinvolgimento è necessario sia per assicurarsi l'effettiva futura comprensione, adozione ed utilizzo delle tecnologie implementate, sia per utilizzare la conoscenza del business maturata dagli stessi dipendenti. In particolar modo, va tenbuto presente che le risorse che si sono qualificate professionalmente all'interno dell'azienda conoscono i precisi Key Performance Indicator (KPI) della propria attività lavorativa. Ciascuna unità aziendale ha sia dati che problemi da risolvere. In più hanno la capacità di fornire precise indicazioni sui reali miglioramenti possibili, nonché dove sono e quali sono le attività che con l'automazione possono condurre veramente a maggiori o minori aumenti di valore.

Per le aziende con modelli organizzativi e di business tradizionali o strettamente vincolati da normative, non è facile creare un'organizzazione del lavoro agile che si leghi strettamente ma flessibilmente alla strategia dei dati. Invece i processi dovrebbero essere veloci nel definire le priorità, le aree da automatizzare e il supporto organizzativo necessario. Molto più avvantaggiate sono le aziende già nate con il digitale, essendo partite da subito con gli strumenti, le competenze e la flessibilità necessaria.

I problemi che si trovano ad affrontare le aziende, anche diverse, possono essere molto simili. Non lo è invece la struttura organizzativa. La capacità organizzativa di una azienda, specialmente in un progetto *data-driven*, si distingue per le *performance* differenziali in termini di capacità di analisi, di progettazione e di implementazione.

Si tratta di conoscere il funzionamento della propria azienda, individuare quindi le aree con cui intrecciare i modelli di ricerca, sapendo portare avanti un'analisi di contesto che possa individuare nuove opportunità di creazione di valore. Non è facile implementare tale progetto attraverso l'individuazione di risorse e competenze, riuscendo a mantenerli allineati al processo di innovazione tecnologica, scegliendo di favorire l'efficienza statica o quella dinamica, la flessibilità o la rigidità e riconoscendo inoltre le peculiarità sociali della propria struttura organizzativa e dei processi decisionali.

Il passaggio a un modello *data-driven* consente di elaborare una "competenza distintiva complessa" (Venier, 2017, p.24). Per i *competitors* replicare questa competenza, nella propria organizzazione aziendale, è quasi impossibile. Una struttura organizzativa funziona in una azienda per le peculiarità delle risorse e delle competenze interne. L'analisi nel complesso e dinamico mondo di oggi deve essere fatta partendo dalla capacità organizzativa interna. Le persone, le competenze, i legami di relazione causale e la governance, consentono di portare a termine progetti con vantaggi competitivi interdetti per altre aziende. E questi fattori determineranno anche la velocità, il rischio e l'ampiezza del progetto *data-driven*. Quindi le variabili economiche e ambientali passeranno in secondo piano. La strategia aziendale sarà in primo luogo legata, in modo dinamico, ai fattori della capacità organizzativa, e viceversa.

Per "Digital Transformation dell'organizzazione intendiamo il processo di allineamento di tecnologia digitale, competenze, processi organizzativi e modelli di business, finalizzato a creare nuovo valore per gli *stakeholder* e mantenere la sostenibilità dell'organizzazione in un ecosistema di business in costante cambiamento." (Venier, 2017, p. 28).

3.4. Il team del Data Scientist

Le competenze fondamentali di un Data Scientist sono *development* (sviluppo software, costruzione di database, conoscenze delle infrastrutture IT), *science* (Statistica, Machine Learning, Network Science) e *data visualization* (teoria e *tool*). Insomma un Data Scientist è un ingegnere del dato, un analista quantitativo? Anche ma non solo. Un Data Scientist è in grado di creare *dataset* di dati non strutturati in modo tale da poterli analizzare, un analista quantitativo analizza i dati. Un ingegnere del dato struttura dati

strutturati, un Data Scientist trasforma dati non strutturati in dati strutturati analizzabili. Eppure, un Data Scientist deve avere ampie competenze matematiche, statistiche, informatiche. Ma anche curiosità per andare in profondità dei problemi, deve sapere porre le giuste domande ed ipotesi verificabili, deve avere sensibilità per i problemi aziendali ed empatia verso i clienti. E ancora abilità sociali e competenze trasversali di business, capacità di *problem solving* o meglio di *hacking*: un Data Scientist deve cioè poter risolvere problemi complessi, in modo originale, attraverso elevate competenze informatiche e di business.

Thomas H. Davenport dice: «La figura classica del Data Scientist presenta cinque volti: *hacker*, scienziato, analista quantitativo, consulente di fiducia, esperto di business». E ancora, Dino Pedreschi, professore ordinario di Informatica all'Università di Pisa, descrive lo scienziato dei dati come: «Una figura che deve avere più competenze. La prima è sapere gestire, acquisire, organizzare ed elaborare dati. La seconda competenza è di tipo statistico, ovvero il sapere come e quali dati estrarre, la terza capacità è una forma di storytelling, il sapere comunicare a tutti, con diverse forme di rappresentazione, cosa suggeriscono i dati» (Il Sole 24 Ore, 2014). Insomma si tratta di una intersezione tra competenze matematico-statistiche e di settore, mescolate sapientemente alla capacità di scoprire ed estrarre nuova conoscenza dai dati, e alla capacità di saperla raccontare.

Le Data Humanities sono la competenza più rilevante del Data Scientist: la capacità di capire la complessità dei fenomeni naturali, sociali e di business attraverso l'analisi dei dati. Attraverso il metodo scientifico, il Data Scientist fa congetture ed ipotesi ma anche confutazioni e critiche, pronto a cambiare sia strategia che opinione di fronte agli scostamenti dei risultati dagli obiettivi, mantenendo quindi una mentalità sempre aperta ai cambiamenti. Questo è il valore aggiunto che solo l'uomo può dare alla grande capacità delle tecnologie di oggi nell'analizzare una vastità, una diversità di dati, i Big Data.

L'approccio multidisciplinare necessario con la Data Science impone anche elevata capacità comunicativa. Sia perché il Data Scientist deve interfacciarsi con le funzioni di riferimento, sia perché deve riuscire a rappresentare le ipotesi o i risultati ai manager dell'azienda. Per questa parte la sua figura professionale è quasi assimilabile a quella del

commerciale, che ascolta le necessità dei clienti interni e promuove la realizzazione del servizio o del prodotto, valorizzando le caratteristiche richieste ai manager dell'azienda.

Il Data Scientist è un supereroe? Forse un po' sì ma nella realtà trovare, avere e non lasciarsi scappare una figura professionale così può essere molto difficile e rischioso per un'azienda. Concentrare tutto il proprio vantaggio competitivo in un unico individuo, tra l'altro per sua natura anche molto "volubile" per la sua curiosità intellettuale, fame di conoscenza e propensione alle grandi sfide, potrebbe diventare un problema per l'azienda.

Creare un team di Data Scientist o ancora meglio un Analytics Solution Center sembra la scelta meno rischiosa in termini di ricambio delle risorse umane e di perdita della conoscenza e dell'esperienza acquisita nel tempo.

Un Analytics Solution Center è composto da tre figure professionali: il Business Analyst, il Computer Scientist e il Data Scientist. Sono tre diverse competenze che si devono intersecare.

Il Business Analyst, cioè l'esperto di business conosce il dominio su cui si sta affrontando un progetto di Data Analytics. Può definire meglio gli obiettivi e i criteri del progetto per le esigenze aziendali e quindi guida il relativo coinvolgimento del Data Scientist. Conosce in modo approfondito il funzionamento dei processi aziendali e sa interfacciarsi con gli *stakeholder* interni ed esterni.

Il Computer Scientist ha conoscenze matematico-scientifiche ed informatiche. Esperto di programmazione avanzata, di creazione, integrazione ed ingegnerizzazione di *tool*. Conosce l'infrastruttura IT dell'azienda, i sistemi sorgente di dati e la Data Governance.

Il Data Scientist è il *team leader*, cioè la sintesi tra approccio scientifico e interdisciplinarità di business, con una elevata attitudine al *problem solving* e allo *storytelling*.

Ci sono però diversi approcci che una azienda può utilizzare per diventare *data-driven*. Aziende che affrontano per la prima volta la sfida dei dati o aziende di piccole dimensioni, possono utilizzare un modello di consulenza. Ci si affida a dei professionisti esterni della scienza dei dati e dell'analisi che lavorano a stretto contatto con le unità aziendali per definire le sfide, progettare, sviluppare e fornire l'analisi. Normalmente sono progetti a breve termine e specifici. I progetti complessi e a lungo termine sono

difficilmente accessibili perché a volte ci vogliono anni di lavoro ed analisi sulla stessa serie di problemi per ottenere grandi risultati. Nella consulenza è facile che i Data Scientist siano poco motivati e che i tempi e le priorità non siano ben definite, data la poca inclusione nei processi produttivi e decisionali di questi professionisti.

Anche quando l'automazione ha appena iniziato a prendere il via, c'è la possibilità che l'azienda si doti da subito di figure professionali al suo interno. Inizialmente le implementazioni riguarderanno una singola unità di business o saranno guidate da questa, con standard diversi e un approccio frammentato. Oppure si cercherà da subito di adottare un modello centralizzato. In questo caso, un *team* centrale guida il processo *data driven*, con tutte le competenze di Data Science detenute nel centro, fornendo alle funzioni aziendali un servizio. Questo approccio raggiunge un livello massimo di standardizzazione e fa leva sugli apprendimenti in tutta l'organizzazione. Quindi un *team* di *data science* serve l'intera organizzazione in una varietà di progetti, disponendo di finanziamenti a lungo termine e di una migliore gestione delle risorse. C'è un'alta probabilità però di creare una disconnessione con le linee di business e di avere poca familiarità con le esigenze e le difficoltà di queste ultime. Ciò può portare ad una scarsa rilevanza delle soluzioni proposte dal team di Data Science, che possono quindi non essere utilizzate dalle *business unit* o addirittura creare dei conflitti con le stesse.

Un passo in più come modello accentrato lo fa il Center of Excellence Model (CoE). Si mantiene comunque un unico centro di coordinamento, ma i Data Scientist verranno assegnati alle diverse unità dell'organizzazione: le attività di analisi sono altamente coordinate, ma gli esperti non verranno rimossi dalle rispettive unità aziendali. Ogni gruppo però risolverà solo i problemi all'interno delle proprie unità, senza una strategia aziendale generale.

Il modello Hub&Spoke ha un approccio federato: è un sistema di gestione e sviluppo della Data Science che richiama le reti, con connessioni dallo Spoke (raggio) verso l'Hub (perno) e viceversa. Questo modello si basa su un Hub centrale, quindi mantiene un centro di eccellenza, un Analytics Solution Center, costituito da un intero team di Data Scientist, Data Engineer e Business Analyst. Il raggio (Spoke) si riferisce a piccoli *team* inviati dal perno (Hub – ASC) per progettare, sviluppare e fornire analisi all'interno delle strutture organizzative.

L'Hub ha l'obiettivo di sviluppare modelli complessi e trasversali alle linee di business, soluzioni analitiche innovative e collabora nello sviluppo di programmi di formazione sui temi di riferimento. Inoltre fornisce la guida metodologica per lo sviluppo e l'evoluzione dei modelli e facilita la diffusione di *best practice* sviluppate dai diversi Spoke.

Lo Spoke ha l'obiettivo di incrementare l'utilizzo integrato dei dati e l'adozione di metodologie analitiche avanzate, attraverso lo sviluppo di modelli specifici per l'area aziendale, sia per nuove applicazioni sia per il miglioramento dell'efficacia del business esistente, e attraverso la guida e la creazione di programmi di esplorazione dati (*data discovery*), nonché analisi avanzate a supporto delle decisioni di business.

Tale modello Hub & Spoke prevede il riporto diretto degli Spoke al responsabile della struttura organizzativa in cui è inserito (Business Owner}, che fornisce indicazioni strategiche, definisce gli obiettivi e il piano di lavoro. Il riporto funzionale degli Spoke rimane legato all'Hub. Quindi l'Analytics Solution Center continuerà a fungere da centro di competenza poiché ha tutti i dati, l'infrastruttura e il personale, in un unico posto. Questo approccio può quindi servire sia ad obiettivi su scala aziendale, sia ad analisi su misura per le funzioni con diversi tipi di modellazione.

Tutti i modelli portano alcuni vantaggi e svantaggi che devono essere valutati su base individuale per scegliere e modellare la giusta soluzione per la propria organizzazione. Alcune aziende si organizzano anche con un mix o dei sottoinsiemi dei vari modelli.

Sicuramente i modelli federati, come quello *hub and spoke*, hanno due grandi qualità: una visione strategica aziendale e di lungo periodo; sviluppano esperienza e conoscenza di Data Science anche localmente nelle singole Business Unit.

Un approccio *data driven* non può prescindere dal creare, aumentare e promuovere una diffusa Data Humanties nell'organizzazione: le competenze tecniche e le *soft skill* necessarie devono essere sviluppate in tutte le risorse.

Da questo punto di vista, divertente ed esplicativo è il dialogo del film Skyfall di James Bond: in un museo, davanti a un quadro, Q, giovanissimo inventore delle tecnologie più avanzate, e il maturo James Bond (Torani, 2014).

Q: Dà sempre una certa malinconia. Una grandiosa nave da guerra trainata ingloriosamente alla demolizione. L'ineluttabilità del tempo, ti pare? Tu cosa vedi?

007: Tanta acqua e una barca. Mi scusi [e si alza per andarsene].

Q: 007 sono il nuovo addetto all'approvvigionamento.

007: Stai scherzando, spero.

Q: Perché non ho camici da laboratorio?

007: No, perché hai ancora i brufoli.

Q: La mia epidermide non è affatto rilevante.

007: Ma la tua competenza sì.

Q: L'età non è una garanzia di efficienza.

007: E la giovinezza non è una garanzia di innovazione.

Q: Oso dire che faccio molti più danni io con il mio portatile in pigiama seduto davanti alla prima tazza di Earl Grey di quanti ne fai tu in un anno sul campo.

007: Oh. E a che vi servo, allora?

Q: Ogni tanto un grilletto va premuto.

007: O non premuto. È difficile scegliere se sei in pigiama.

Lasciando un attimo da parte il supereroe della Data Science, il Data Scientist, e i professionisti sicuramente necessari come il Data Engineer e il Business Analyst, bisogna capire chi altro deve fare parte del *team*. Progettare efficaci ed efficienti *team* è una competenza manageriale tutt'altro che facile da trovare. Perciò prima di tutto c'è bisogno di un ottimo manager che riesca a mettere insieme le competenze e le abilità necessarie.

Le abilità sociali come l'empatia e la comunicazione sono fondamentali in un *team*. Tali abilità sociali non vanno intese nel senso comune del termine per cui il collaboratore deve andare d'accordo con tutti. In un *team* di Data Science è fondamentale che empatia e comunicazione permettano ai componenti del gruppo di accettare la presenza di diversi punti vista, di comprenderli e di accendere un confronto reciproco. Questo perché è un *team* interdisciplinare, che dovrebbe include risorse diverse per studi accademici, esperienza professionale e vissuta, e quindi per diversa prospettiva.

Queste diversità implicano che i componenti del *team* devono anche essere diversi per età. Le risorse in età matura hanno una esperienza professionale e un vissuto che le

rende sicuramente più competenti. Dall'altra parte le risorse giovani hanno studi accademici più recenti ed aggiornati e hanno l'innocenza di chi non ha ancora un vissuto professionale, che elimina i presupposti dell'esperienza e permette domande "anticonformiste", magari scoprendo nuovi modi di fare.

I componenti di un *team* di Data Science devono completarsi a vicenda, in una democrazia di punti di forza, ognuno il suo o i suoi, tutti importanti, ma più o meno rilevanti nella specificità di un progetto. Nessun componente eccelle, ma tutti lavorano insieme per eccellere nel progetto. In un *team* di Data Science, quindi, formato da figure professionali eccezionali, non è facile creare questo clima collaborativo. Ma deve essere altrettanto chiaro che solo un lavoro di *team* permette di affrontare le sfide con i Big Data.

Un *team* di Data Science facilmente all'inizio avrà un'organizzazione stile startup perseguendo varie iniziative, con molta attività di *brainstorming*, e l'attività e i risultati acquisiti saranno conosciuti solo all'interno del *team*.

Successivamente, quando gli obiettivi e i requisiti per ottenerli sono chiari, i componenti del *team* si possono specializzare di più nei propri ruoli. Probabilmente poi alcuni progetti si sposteranno su altri *team*, magari all'interno delle funzioni aziendali, come nel modello distributivo Hub&Spoke.

I modelli di distribuzione federati o decentrati permettono di formare risorse già presenti in azienda, completando il lavoro del *team* del Data Scientist con competenze maturate in anni di esperienza. Naturalmente la formazione interna ed esterna continuativa è indispensabile per una azienda *data-driven*.

Il *team* del Data Scientist e i talenti già presenti in diverse funzioni dell'azienda, se ben coordinati e formati, offrono quindi un supporto considerevole al Data Scientist in termini di competenze di business, informatiche e di *problem solving*.

Eppure il fattore fondamentale per il successo di un progetto sui Big Data resta in capo al Data Scientist, le persone "[…] who understand how to fish out answers to important business questions from today's tsunami of unstructured information" (Davenport, 2012).

Sicuramente queste capacità fanno parte delle caratteristiche e dell'intelligenza proprie di un individuo ma non solo, possono essere insegnate, affinate ed elevate.

Come il Data Scientist non è solo uno statistico, la Data Science non si limita alle discipline istituzionali come la matematica o l'informatica. Malgrado non sia una scienza nuova, l'etnostatistica è "differente" come lo è la Data Science.

Le discipline istituzionali volgono l'attenzione agli aspetti tecnici e scientifici. L'etnostatistica si occupa delle pratiche mondane della vita quotidiana e delle conoscenze laiche e professionali necessarie per implementare e utilizzare le statistiche (Gephart, 1988). Cioè cerca di capire la conoscenza tacita, i comportamenti sociali, anche dei ricercatori, per capire l'interpretazione data alla statistica.

Un modello di dati include il più vasto processo di ricerca, la selezione di particolari variabili e classi, l'uso di particolari tecniche statistiche, lo sviluppo dei dati. E molte delle ipotesi e delle decisioni metodologiche prese non vengono descritte negli studi, perché implicite nelle metodologie istituzionali o per interpretazioni inconsapevoli degli statistici. Si studiano quindi le caratteristiche culturali dello statistico o del ricercatore o del gruppo scientifico al quale appartiene, valutando il loro coinvolgimento nella produzione di variabili e statistiche. Mettere in discussione le metodologie istituzionali può non incontrare i favori dei ricercatori o degli statistici, perché hanno interesse a mantenere le tecniche consolidate, ma spesso perché non intravedono distorsioni per la consuetudine con tali metodologie.

Un metodo per scardinare tali presupposti invisibili e consueti è la triangolazione. Questa modalità di analisi, utilizzando diverse fonti, permette di avere diversi dettagli con diverse interpretazioni fisiche e temporali, dello stesso evento, da parte di diversi attori. I diversi punti di vista possono poi essere confrontati con i risultati degli algoritmi e dei dati. Cercare cioè di massimizzare il numero di modi in cui l'attività può essere esaminata. Un altro metodo è inserire nell'analisi alcuni dati che possano creare problemi, come in statistica lo fa il valore anomalo. L'interpretazione che ne verrà data è l'oggetto dello studio.

Anche i report o le presentazioni delle statistiche sono analizzati dall'etnostatistica perché comunque frutto di una interpretazione soggettiva. Una buona critica letteraria può svelare alcuni processi mentali sottostanti che contestualizzano i dati.

La capacità dell'etnostatistica di misurare come e cosa ha portato all'utilizzo di determinate tecniche e metodologie, fornisce informazioni fondamentali al ricercatore sociale nell'utilizzo dei dati prodotti e nella costruzione di un modello causale.

La stessa capacità deve essere, ancor di più oggi con i Big Data, in possesso del Data Scientist. Gli algoritmi di analisi dei Big Data sono frutto di elevate competenze tecniche e scientifiche. I risultati che si ottengono però sono correlazioni tra una moltitudine di dati, attraverso un auto-apprendimento guidato da decisioni umane, seppur tecnologiche.

Il Data Scientist deve avere la sagacia e la flessibilità di pensiero per individuare le correlazioni che possono portare a nuovi sviluppi ma anche la capacità di analisi e una visione di insieme per costruire un modello causale che possa essere utilizzato, ripetibile e generalizzabile. Inoltre questo modello deve aderire al business e alla sua organizzazione.

Conclusioni

La rivoluzione dei Big Data è quindi iniziata, ma non oggi, già da quasi dieci anni. Il tempo già trascorso ci fa capire l'imponenza di questo fenomeno e l'invadenza che ha nel mondo scientifico, sociale ed imprenditoriale.

Oggi viviamo in una complessità che ci regala un patrimonio di forme e strutture diverse, ma che ci mette davanti ad una nuova, sconosciuta ed importante sfida: la gestione e la conoscenza di questo mondo complesso. I Big Data sono composti di un gran numero di elementi, in continua interazione tra di loro e con sistemi diversi. La dinamicità e la dinamica di queste interazioni genera nuovi ed emergenti comportamenti dei sistemi, che non sono più riconducibili ai singoli elementi che li compongono. La linearità delle relazioni sparisce, la casualità fa posto alla correlazione. L'organizzazione di un sistema complesso creato dai Big Data non ha più una forma gerarchica, ma si autoregolamenta. E si adatta all'ambiente di riferimento.

Forse per questo in molti sostengono che i Big Data fanno tutto da soli, e non necessitano di nient'altro che di dati, di algoritmi e di analisi computazionale. Non è certo una idea nuova che "i dati parlino da soli". Una volta questo ruolo di onniscienza era svolto della Statistica. Come hanno insegnato anni di studi, invece, i dati sono sicuramente un patrimonio importantissimo, ma solo se li sappiamo interpretare correttamente.

I problemi e gli errori, che gli statistici hanno imparato a identificare, mitigare o eliminare, non sono spariti con l'arrivo della Data Analitics e con i Big Data. Semmai con i Big Data ce ne sono di più complessi ed invisibili. La ricerca con i "piccoli dati" permette di gestire meglio la complessità del fenomeno che si sta analizzando. Certo è che oggi questo tipo di ricerca non è più possibile, sia per poter sfruttare l'enorme quantità di dati a disposizione, sia per l'alto costo che implica la ricerca tradizionale. Dall'altra parte la validità dei dati e del modello di dati rimane comunque la prerogativa fondamentale in un progetto di ricerca. Probabilmente la strada migliore sarà studiare ed inventare nuove metodologie di ricerca, senza perdere ciò che è già noto grazie alle teorie e i modelli di ricerca tradizionale.

I sondaggi, ad esempio, con i Big Data, hanno una capacità di raccolta di dati prima impossibile. Eppure spesso mancano le informazioni sul contesto, e sulle caratteristiche demografiche della popolazione considerata. Soprattutto, i Big Data sono quasi sempre

dati secondari, cioè non raccolti specificatamente per una data ricerca, come appunto i dati primari. Il ricercatore sociale sa bene invece quanto rilevante è il ruolo della domanda nel sondaggio. Se non vengono poste le giuste domande agli intervistati, si può ricadere in una errata interpretazione, che a sua volta può falsare l'intero modello di analisi.

Può sembrare un paradosso, ma è necessario essere coscienti che anche ai Big Data mancano dei dati. Il *digital divide* ha un impatto molto rilevante nella completezza dei Big Data. Il ricercatore, inoltre, non sempre ha effettivamente la disponibilità dei dati, anche quando sono Big Data. E spesso non sa nemmeno che dati mancano, perché magari sono concessi da società private, che spesso ne detengono il monopolio, senza il corredo di metadati che normalmente accompagna i dati statistici.

Un'altra sfida importante con i Big Data è la conservazione nel lungo periodo di questi dati e dei relativi metadati. Come tutte le informazioni che portano ad una conoscenza, infatti, anche i Big Data devono essere corredati di un contesto, che in questo caso sono appunto i metadati. L'archiviazione digitale deve portare con sé lo storico di tutti i metadati che si sono creati nell'estrazione, classificazione e trasformazione di ogni singolo dato dei Big Data. Si tratta di un impegno notevole, ma necessario per poter tramandare al futuro la conoscenza acquisita con i Big Data.

La conservazione dei Big Data, dall'altra parte, crea non pochi problemi legati alla tutela della privacy. All'interno di questi dati ci sono numerose informazioni sensibili sulle persone, che le stesse, più o meno consapevolmente, hanno fornito. Il timore generalizzato di abusi nell'utilizzo o trasmissione di questi dati personali è uno dei più grandi limiti all'espansione del digitale nelle attività commerciali e anche di amministrazione pubblica. Molto spesso questi dati sono inoltre utilizzati per alimentare algoritmi che influiscono, attraverso decisioni automatiche, sui diritti delle persone. Queste *black box* che trasformano, integrano e classificano i dati in input possono portare a risultati distorti e discriminatori. I *bias*, i pregiudizi, presenti nei dati in input o creati dall'algoritmo, vanno identificati, mitigati e se possibile eliminati.

Per sfruttare appieno le opportunità dei Big Data serve una responsabilità e una trasparenza condivisa tra chi fornisce i dati, chi crea i modelli di dati o gli algoritmi e chi li utilizza. Questo è ancora più vero in un'azienda che cerchi di diventare *data-driven*. La condivisione della responsabilità, ma anche di conoscenze, e l'interdisciplinarietà sono

fondamentali per avere successo in una strategia aziendale con Big Data. Avere un ottimo Data Scientist in azienda è sicuramente fondamentale, ma è anche importante che sia supportato da un team eterogeneo per competenze, abilità ed età.

Tutta l'azienda deve essere coinvolta nel formare una cultura aziendale digitale, perché le decisioni, i nuovi prodotti, i nuovi processi possibili con i Big Data, devono essere compresi ed accettati da tutti i lavoratori su cui impattano. Va creata una "cultura dell'errore e del *feedback*" perché la complessità dei Big Data richiede di rivedere continuamente, in base ai risultati ottenuti, le strategie, le ipotesi di partenza e i modelli di dati progettati. L'aumento della velocità intrinseco nelle nuove tecnologie e nei Big Data deve, inoltre, necessariamente passare per una nuova organizzazione aziendale, dove il ruolo, la responsabilità e il contributo di tutti è fondamentale. I processi, sia di produzione che di informazione aziendale, vanno resi più snelli e semplici.

Tutta questa complessità dovrebbe renderci più semplice fare impresa, ma anche semplificare la vita delle persone, dei cittadini o degli anziani. E lo farà sicuramente, ma solo smettendo di banalizzare queste nuove tecnologie. I Big Data devono essere affrontati nella loro enorme complessità per poter creare valore, con le giuste competenze, nuovi modelli e un'analisi critica dei dati e degli algoritmi.

Bibliografia

Abreu Lopes, C., Bailur, S. (2018). GENDER EQUALITY AND BIG DATA. UN Women. Disponibile in: https://www.unwomen.org/en/digital-library/publications/2018/1/gender-equality-and-big-data. [26 settembre 2019]

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

Barker, A., Stuart Ward, J. (2013) Undefined By Data: A Survey of Big Data Definitions. Disponibile in: https://www.adambarker.org/papers/bigdata_definition.pdf [12 settembre 2019]

Baker, R. P. (2017). "Big Data: A Survey Research Perspective." In Total Survey Error: Improving Quality in the Era of Big Data, edited by Paul P. Biemer, Edith De Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 47-70. Hoboken, NJ: Wiley.

Barbuti, N. (2019). Ripensare i formati, ripensare i metadati: prove "tecniche" di conservazione digitale. *Umanistica Digitale*, 5 DOI: https://doi.org/10.6092/issn.2532-8816/9055

Bassa, A., (2018). Managing a Data Science Team. *Harvard Business Review*. Disponibile in: https://hbr.org/2018/10/managing-a-data-science-team [28 settembre 2020]

Biemer, P. (2010). Total Survey Error: Design, Implementation, and Evaluation, *Public Opinion Quarterly*, 74(5), 817–848. https://doi.org/10.1093/poq/nfq058

Boyd, D., Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Buolamwini, J. A., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. 2018 Conference on Fairness, Accountability, and Transparency

Caligiuri, M. (2016). Cyber intelligence, la sfida dei data scientist. Disponibile in: https://www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2016/06/cyber-intelligence-sfida-data-scientist-Caligiuri.pdf [25 settembre 2019]

Callegaro, M., Yang Y. (2018). The Role of Surveys in the Era of "Big Data". In: Vannette D., Krosnick J. (eds.) *The Palgrave Handbook of Survey Research*. Palgrave Macmillan, Cham https://doi.org/10.1007/978-3-319-54395-6_23

Camfield, L. (2019). Rigor and Ethics in the World of Big-team Qualitative Data: Experiences From Research in International Development. SAGE Publications.

Camiciotti, L., Racca, C. (2017). Creare valore con i Big data. Edizioni LSWR.

Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review*. Disponibile in: https://hbr.org/2013/04/the-hidden-biases-in-big-data [01 aprile 2013]

Castelfranchi, Y. (2017). Computer-aided text analysis: an open-aired laboratory for social sciences. *Journal of Science Communication* 16 (02), C04_en.

Casi, F. (2018). Big data ed etica dei dati, Segreteria della Consulta. Disponibile in: http://www.consultadibioetica.org/big-data-ed-etica-dei-dati-di-fiorello-casi/ [28 Dicembre 2018]

Casi, F. (2019). Cambiamento di paradigma. Segreteria della Consulta, 11 Febbraio 2019. Disponibile in: http://www.consultadibioetica.org/cambiamento-di-paradigma-di-fiorello-casi/ [20 settembre 2018]

Coffetti, E., Pasini, P., (2014). Data Scientist focus and trends. *Numbers*. Disponibile in: https://images.wired.it/wp-content/uploads/2016/02/1456304227_rivista-Numbers.pdf [28 settembre 2020]

COMITATO EUROPEO PER LA PROTEZIONE DEI DATI (2020). EDPB Domande frequenti sulla sentenza della Corte di giustizia dell'Unione europea nella causa C-311/18 — Data Protection Commissioner/Facebook Ireland Ltd e Maximillian Schrems, Adottate il 23 luglio 2020.

Crosby, P., (2019). How to Implement Data-Driven Decision Making in Your Organization. Disponibile in: https://theuncommonleague.com/blog/data-driven-decision-making [25 settembre 2019]

Dagnino, E. (2017). People Analytics: lavoro e tutele al tempo del management tramite big data. *Labor & Law*, 3(1), LLI.

Diakopoulos, N. (2013). Algorithmic Accountability Reporting: On the Investigation of Black Boxes. Report, Tow Center for Digital Journalism, Columbia University.

Davenport T.H., Patil D.J. (2012). Data Scientist: The sexiest job of the 21st century. *Harvard Business Review*, October 2012

Davenport, T.H. (2015) Big Data @l lavoro. Sfatare i miti, scoprire le opportunità, Franco Angeli Edizioni, p. 86

ESOMAR (2018). USE OF SECONDARY DATA IN MARKET, OPINION, AND SOCIAL RESEARCH AND DATA ANALYTICS. Discussion Paper.

Faelli, T. (2018). Big data e GDPR. Generali Italia TAM TAM TALKS. Disponibile in: https://www.youtube.com/watch?v=0qLJtLx0yZU [30 luglio 2018]

Feliciati, P. (2009). Gestione e conservazione di dati e metadati per gli archivi: quali standard? Disponibile in:https://www.researchgate.net/publication/37811666_Gestione_e_conservazione_di_dati_e_metadati_per_gli_archivi_quali_standard [30 giugno 2019]

Garante per la protezione dei dati personali, (2017). Big Data e Privacy La nuova geografia dei poteri. Convegno del 30 gennaio 2017 organizzato in occasione della "Giornata europea della protezione dei dati personali" 2017.

Generali (2017). Intervista a Giovanni Buttarelli: privacy e Big Data. Disponibile in: https://www.generali.com/it/info/discovering-generali/all/2017/Privacy-and-Big-Data

Gephart, Jr R. P. (1988). Ethnostatistics: Qualitative Foundations for Quantitative Research. Sage Publications.

Giribaldi, D. (2019). Discriminazione algoritmica. Intelligenza artificiale, tutti i pregiudizi (bias) che la rendono pericolosa. Disponibile in: https://www.agendadigitale.eu/cultura-digitale/intelligenza-artificiale-tutti-i-pregiudizi-bias-che-la-rendono-pericolosa/ [26 Feb 2019]

Glenna, L., Hesse, A., Hinrichs, C., Chiles, R., Sachs, C. (2019). Qualitative Research Ethics in the Big-Data Era, *American Behavioral Scientist*, 63(5), 555–559. Disponibile in: https://journals.sagepub.com/home/abs

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

Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *THE ANNALS OF THE AMERICAN ACADEMY*, 659, May 2015 https://doi.org/10.1177%2F0002716215570866

Hossain, N., Scott-Villiers, P. (2019). Ethical and Methodological Issues in Large Qualitative Participatory Studies. SAGE Publications.

Il sole 24ore, (2014). Professione scienziato del dato. Disponibile in: https://st.ilsole24ore.com/art/tecnologie/2014-10-26/professione-scienziato-dato-081257.shtml?uuid=ABHDEu6B&refresh_ce=1 [9 settembre 2018]

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. (2015) Big Data in Survey Research: AAPOR Task Force Report. *Public Opinion Quarterly*, 79(4), 839–880, https://doi.org/10.1093/poq/nfv039

Joshi, N. (2017). Avoiding bias in data analytics. Disponibile in: https://www.linkedin.com/pulse/avoiding-bias-data-analytics-naveen-joshi [22 ago 2019]

Kaplan, R., Norton, D. (1992). The Balanced Scorecard - Measures that Drive Performance, *Harvard Business Review*.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), 262–267. https://doi.org/10.1177/2043820613513388

Kitchin, R., McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, January–June 2016, 1–10.

Kitchin, R. (2017). Thinking critically about and researching algorithms, *Information, Communication & Society*, 20(1), 14-29 https://doi.org/10.1080/1369118X.2016.1154087

Lombi, L. (2015). La ricerca sociale al tempo dei Big Data: sfide e prospettive. *STUDI DI SOCIOLOGIA*, 2, 215-227 [http://hdl.handle.net/10807/67605]

Lugmayr, A., Stockleben, B., Scheib, C. and Mailaparampil, M. (2017), "Cognitive big data: survey and review on big data research and its implications. What is really "new" in big data?". *Journal of Knowledge Management*, 21(1), 197-212. https://doi.org/10.1108/JKM-07-2016

Ngan, M., Grother, P. Face recognition vendor test (FRVT) performance of automated gender classification algorithms. US Department of Commerce, National Institute of Standards and Technology, 2015.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? *Qualitative Research*, 18, 591-603.

Miltgen, C. L., Peyrat-Guillard, D. (2014). Cultural and generational influences on privacy concerns: a qualitative study in seven European countries. *European Journal of Information Systems*, pp. 103-125.

Neresini, F. (2017). Old media and new opportunities for a computational social science on PCST. Journal of Science Comunication 16 (02), C03_it.

O'Neil, K., (2016). How algorithms rule our working lives. Disponibile in: https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives [15 agosto 2020].

Palange, S. (2019). BIAS. Disponibile in: https://www.linkedin.com/in/palange/ [25 agosto 2019]

Pariser E., (2011). The Filter Bubble: What the Internet Is Hiding from You. Penguin Books.

Pasini, P., Perego, A. (2009). L'assesment della Business Intelligence in azienda: il BI Maturity Model. SDA Bocconi School of Management, Osservatorio BI e BPM.

Pasini, P., Perego, A. (a cura di). (2012). Big Data: nuove fonti di conoscenza aziendale e nuovi modelli di management. Rapporto di Ricerca per IBM.

Pasini, P., Perego, A. (a cura di). (2012). Velocità aziendale e tecnologie In-Memory: una prospettiva manageriale. Rapporto di Ricerca per SAP Italia. SDA Bocconi.

Pasini, P., Perego, A. (a cura di). (2013) Big Data Live: Casi di eccellenza. Rapporto di Ricerca per IBM. SDA Bocconi.

Pearson, T., Wegener, R. (2013). Big Data: The organizational challenge. White paper Bain & Company. Disponibile in: https://www.bain.com/contentassets/25c167a5149c42168994338f9dc99ffe/bain_brief_big_data_the_organizational_challenge.pdf [01 giugno 2017]

Pedreschi, D. (2018). Human artificial intelligence: open the black box. Generali Italia TAM TAM TALKS. Disponibile in: https://youtu.be/4qPkgASlWaM?list=PLQzVjkkNIHOOKfDhTJqmA0lWsemr75ozF [30 luglio 2018]

Pitrelli, N. (2017). 'Big data e metodi digitali per la ricerca in comunicazione della scienza: opportunità, sfide e limiti'. *Journal of Science Communication* 16 (02), C01_it.

Pratesi, M. (2017). I Big data: il punto di vista di uno statistico. Il menabò - Associazione Etica ed Economia, Menabò n. 62 14 aprile 2017 https://www.eticaeconomia.it

PWC (2017). Artificial Intelligence in HR: a No-brainer. Disponibile in: https://www.pwc.nl/nl/assets/documents/artificial-intelligence-in-hr-a-no-brainer.pdf [20 agosto 2020]

Pugliesi, R., (2018). Big data e scienze sociali. Una nuova sfida alla teoria. Dossier Economia Digitale I Copernicani, 10/2018, 52-59. Disponibile in: http://www.fondazionecomunica.org/wp-content/uploads/2018/10/Copernicani-Dossier-1-Economia-digitale.pdf

Rainie, L., Anderson, J. (2017). Code-Dependent: Pros and Cons of the Algorithm Age. Pew Research Center, February 2017. Available at: http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age.

Ramirez, E., Brill, J., Ohlhausen, M.K., McSweeny, T. (2016). Tool for Inclusion or Exclusion? Understanding the Issues, FTC Report, January 2016, Federal Trade Commission.

Rezzani, A. (2015). Dalla business intelligence ai sistemi di predictive analytics. Disponibile in: https://www.dataskills.it/dalla-business-intelligence-ai-sistemi-di-predictive-analytics/#gref [6 novembre 2017]

Rigutto, C. (2017). 'Il panorama della comunicazione visiva della scienza sul web'. *Journal of Science Communication* 16 (02), C06_it.

Robinson, L., Cotten, S.R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., M. Hale, T., Stern, M.J. (2015). Digital inequalities and why they matter, Information, Communication & Society, 18:5, 569-582, DOI: 10.1080/1369118X.2015.1012532. Disponibile in: http://dx.doi.org/10.1080/1369118X.2015.1012532

Russo, M. (2015). Privacy, Internet delle cose, big data e intelligenza artificiale: ecco perché serve un nuovo contratto sociale. Convegno "Il pianeta connesso. La nuova dimensione della privacy".

Standish Group (1995). CHAOS Report. Disponibile in: https://www.projectsmart.co.uk/white-papers/chaos-report.pdf [20 giugno 2020]

Sterett, S. M. (2019). Data Access as Regulation. SAGE Publications

Software AG (2013). Big Data Meets Big Process: Opportunities for Business Innovation. Research Report.

Torani, S. (2014). Quando il cielo crolla. Trauma e immaginario post mortem in Skyfall, Tesi di laurea in storia del cinema, a.a. 2013/2014. Disponibile in: http://www.academia.edu/15443296/Quando_il_cielo_crolla_trauma_e_immaginario_postmortem_in_Skyfall.

Turner Lee, N., Resnick, P., Barton, G. (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Wednesday, May 22, 2019. Disponibile in: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Utility Analytics Istitute, (2018). Organizational Models and Trends. UAI Research 2017-2018.

Venier, F. (2017). Trasformazione digitale e capacità organizzativa. Edizioni Università di Trieste, Trieste.

White, N., Grueger, D., (2017). Managing the digital workforce. Delotte Whitepaper. Disponibile in: https://www2.deloitte.com/content/dam/Deloitte/au/Documents/human-capital/deloitte-au-hc-managing-digital-workforce-131017.pdf [25 settembre 2020]

Williams, S. (2016). Business Intelligence Strategy and Big Data Analytics. Elsevier, Chapter 2 Business Intelligence in the Era of Big Data and Cognitive Business. Disponibile in: http://dx.doi.org/10.1016/B978-0-12-809198-2.00002-6

Whitaker, S. (2014). Big Data versus a Survey. Federal Reserve Bank of Cleveland, working paper no. 14-40.