Scienze economiche e statistiche
Data science
Istituto Lombardo (Rend. Lettere) 150, 109-123 (2016)

Osserviamo una nuova rivoluzione industriale guidata da dati digitali, computazione e automazione. Attività umane, processi industriali e ricerca portano alla raccolta ed analisi di dati a una scala senza precedenti che stimolano nuovi prodotti, nuovi processi produttivi e nuove metodologie scientifiche.
(Commissione Europea 2014, *Towards a thriving data-driven economy*)

LA SCIENZA DEI DATI: UNA NUOVA SFIDA MULTIDISCIPLINARE

ANTONIETTA MIRA (*)

Nota presentata dal m.e. Benito Vittorio Frosini

(Adunanza del 26 maggio 2016)

SUNTO. – Il presente intervento intende delineare e approfondire la nuova sfida multidisciplinare costituita dalla Scienza dei Dati. Al fine di rendere il lettore consapevole della portata rivoluzionaria di questa nuova disciplina, passerò in rassegna diversi esempi di successo scaturiti dall'analisi di dati eccezionali in termini di quantità e profondità, ossia i cosiddetti *Big Data*. La mia analisi non mancherà di sottolineare l'impatto economico e sociale dei *Big Data* facendo anche riferimento alle questioni etiche che naturalmente scaturiscono in presenza di dati e informazioni sensibili.

***

ABSTRACT. – The present manuscript aims to delineate and deepen the new multidisciplinary challenge of Data Science. In order to make the reader aware of the revolutionary scope of this new discipline, I will review various examples of success arising from

(*) Università della Svizzera Italiana, Switzerland. Università dell'Insubria, Varese, Italy. E-mail: antonietta.mira@usi.ch

the analysis of so called "Big Data". My analysis will not fail to underline the economic and social impact of the Big Data revolution by also referring to the ethical issues that naturally arise in the presence of sensitive data and information.

Il mio intervento intende delineare e approfondire la nuova sfida multidisciplinare costituita dalla Scienza dei Dati. Ritengo opportuno iniziare dalla definizione di *Big Data*. Una prima e approssimata definizione richiama la traduzione italiana del termine, cioè dati enormi, che, come suggerisce l'etimologia latina, ossia *ex-norma*, indica dati straordinariamente grandi, nel senso che la loro dimensione è fuori dall'ordinario. Secondo tale accezione, il termine si carica di un significato soggettivo che varia sulla base dei ricercatori che si trovano ad analizzare i dati e del momento in cui l'analisi viene condotta: ciò che è *big* per voi probabilmente non lo è per me che sono abituata ad analizzare enormi mole di dati e ciò che è *big* oggi probabilmente non lo sarà più domani grazie all'aumento continuo ed esponenziale della capacità di calcolo e di memoria dei super-computer. Il termine *Big Data*, inoltre, racchiude categorie eterogenee di dati, come ad esempio testi, immagini, video e dati relazionali generati da processi di scambio sui social media – come *Facebook* e *Twitter* – oppure estrapolati da scambi di email, telefonate e metadati tra cui coordinate geografiche e temporali associate a foto e video che collezioniamo mediante sistemi di sensori e apparecchi mobili. I *Big Data* arrivano a noi da fonti diverse a *velocità*, *volume* e *varietà* straordinari. Per avere una misura delle cosiddette *3V* che li caratterizzano si pensi che il 90% dei dati mondiali è stato generato negli ultimi due anni e che ogni minuto su *YouTube* vengono caricati circa trecento minuti di video, visti per l'equivalente di trecentoventitré giorni da un milione di utenti l'82% dei quali ha fra i quattordici ed i diciassette anni e che ogni secondo su *Google* si effettuano in media 2.3 milioni di ricerche (i dati si riferiscono al momento della stesura del presente lavoro). Stiamo vivendo quindi una crescita esponenziale nella produzione di dati e in parallelo un calo nei costi di memorizzazione ed elaborazione degli stessi: a titolo di esempio si pensi che oggi bastano seicento Euro per comprare un disco fisso che contenga tutta la musica del mondo. I *Big Data* si presentano tipicamente in modo non strutturato, distribuiti su supporti diversi e, laddove non sono grandi, sono sicuramente complessi. Soprattutto in quest'ultimo caso, ma anche in generale, piuttosto

che *big* sarebbe più opportuno definirli *smart data* nel senso che sono gli statistici e i *data scientist*, con le loro analisi, a renderli "intelligenti", cioè fruibili estraendone informazioni. Sono custoditi in banche dati e il termine "banche" sottolinea il valore delle informazioni che da questi possiamo estrarre con un processo di *data mining*. Vale tuttavia la pena di sottolineare che la catena di valore della Scienza dei Dati è tale solo se l'informazione si trasforma in conoscenza che a sua volta deve essere in grado di supportare decisioni e azioni che spesso richiedono competenze multidisciplinari.

1. La scienza dei dati dentro e fuori le università

La Scienza dei Dati, di cui i *Big Data* rappresentano l'oggetto di analisi, sta cambiando il modo in cui lavoriamo dentro e fuori dalle università: da un lato crea un nuovo paradigma per la ricerca scientifica; dall'altro, all'interno delle imprese e delle organizzazioni in generale, definisce nuove modalità di lavoro. Continuando il parallelismo, all'interno delle università la Scienza dei Dati sta creando una cultura multidisciplinare in cui ricercatori, attivi in diverse aree del sapere, uniscono le forze per estrarre conoscenza dai *Big Data*. Nelle imprese, invece, la Scienza dei Dati sta creando una cultura dove *leader* aziendali e personale dedicato all'informazione e alla tecnologia uniscono competenze per realizzare valore dai dati. Come conseguenza di simili cambiamenti, si è passati dallo scienziato eclettico di stampo leonardesco alla specializzazione del sapere scientifico in cui abbiamo assistito, tra la fine dell'800 e gli inizi del 900, alla circoscrizione degli oggetti di osservazione delle diverse discipline. I *Big Data*, con la loro mole e profondità, sono come un masso che sta frantumando lo specialismo e scardinando i modi tradizionali di fare scienza. A ricomporre la spaccatura interviene la figura del *data scientist* inteso non come singolo individuo ma come insieme di ricercatori che mescolano competenze e, forti dei loro specialismi, sono in grado di generare nuove scoperte *data-driven*.

1.1 Esempi di utilizzo dei Big Data *e di successo della scienza dei dati*

Per rendere la misura della portata rivoluzionaria della Scienza dei Dati riporto alcuni esempi di successo di questa nuova disciplina.

Il primo è relativo alla scoperta del bosone di Higgs. Il *Large*

Hadron Collider produce seicento milioni di collisioni fra particelle al secondo, ciascuna delle quali genera circa un megabyte di dati. Basta dunque un rapido calcolo per capire la mole di dati e di analisi che ha supportato questa scoperta: 10^{15} byte di informazioni al secondo – vale a dire un *petabyte* di dati, l'equivalente di duecentodiecimila DVD. Ritengo che quest'ultimo esempio metta ben in evidenza il valore intrinseco dei *Big Data* che consentono anche a discipline mature come la fisica delle particelle di progredire analizzando eventi estremi, improbabili, ma assolutamente significativi e la cui osservabilità è garantita dalle enormi quantità di esperimenti che a loro volta generano enormi quantità di dati.

Un altro esempio di applicazione dei *Big Data* sono gli esperimenti che hanno portato alla conferma sperimentale dell'esistenza delle onde gravitazionali e contestualmente dei buchi neri – esistenza prevista dalla teoria della relatività generale di Einstein (1915). La distanza tra i due specchi degli interferometri laser a causa dall'onda gravitazionale, la più debole dell'universo, cambia di un decimillesimo del diametro di un protone e questo è stato misurato! Il rapporto segnale-rumore è di circa 1 a 1000. A questa scoperta hanno partecipato quindici paesi per un totale di centrotrentatré istituzioni e oltre mille ricercatori.

Un terzo esempio di successo è il progetto *Human Genome* iniziato nel 1990 e concluso nel 2003 con la produzione della mappa di circa venticinquemila geni. Oggi le tecniche di *Next Generation Sequencing* generano in un giorno tanti dati quanti quelli prodotti in dodici anni dal progetto *Human Genome*. Quest'ultimo esempio fornisce l'idea della veloce evoluzione delle capacità di generare dati nella genomica – evoluzione comune a molti altri ambiti di ricerca e forse ancora più evidente nelle scienze sociali dove, fino a qualche anno fa, i dati erano pochi, per lo più qualitativi e raccolti attraverso interviste e questionari.

Per quanto riguarda le scienze sociali, la rivoluzione apportata dai *Big Data* ha la stessa portata che l'introduzione del telescopio e del microscopio hanno avuto nell'astronomia e nella fisica e nelle scienze biologiche e mediche, consentendo l'esplorazione dell'infinitamente grande e dell'infinitamente piccolo. Le scienze sociali oggi dispongono non soltanto di dati profondi raccolti attraverso i social media, i mezzi di comunicazione multimediale e l'internet delle cose ma anche di metodi statistici che consentono di esplorare le infinitamente complesse

relazioni sociali. La *Fig. 1*, tratta dal lavoro di un collaboratore (Onnela *et al.* 2007), rappresenta un campione di un network di comunicazioni telefoniche di sette milioni di soggetti con ventitré milioni di connessioni e illustra quanto detto poc'anzi. Per analizzare dati di tipo relazionale come questi si può ricorrere a due classi di modelli: i modelli statistici e i modelli meccanicistici.

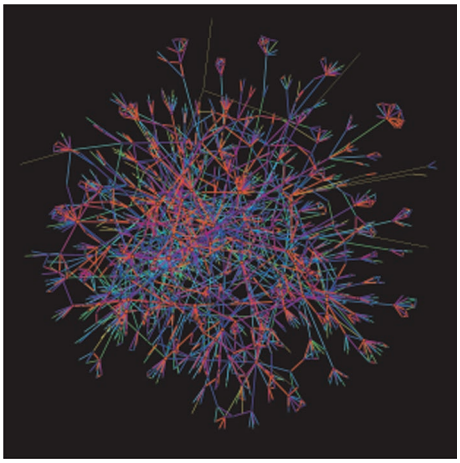

Fig. 1 – Network di ventitré milioni di comunicazioni telefoniche intercorse tra sette milioni di soggetti.

I primi permettono di fare inferenza in modo tradizionale ma presentano lo svantaggio di non essere scalabili rispetto ai grandi numeri e di incorporare con difficoltà le conoscenze relative al dominio in cui i dati sono stati generati. I modelli meccanicistici, al contrario, partono dalla conoscenza delle teorie e dei meccanismi sociali che governano la formazione delle osservazioni e scalano bene in presenza di dati abbondanti ma non possono essere analizzati con i metodi tradizionali di inferenza statistica. Al proposito, un contributo mio e del Prof. JP Onnela della *Harvard School of Public Health*, ha consentito di superare quest'ultimo limite e di rendere quindi possibile anche per modelli meccanicistici l'inferenza in tutti i suoi aspetti: dalla stima dei parametri, al test d'ipotesi, alla scelta fra modelli alternativi. Siamo riusciti nel nostro intento pensando i modelli meccanicistici come delle scatole nere in cui, fissati in entrata dei valori dei parametri, si può velocemente simulare in uscita una configurazione del network corrispondente.

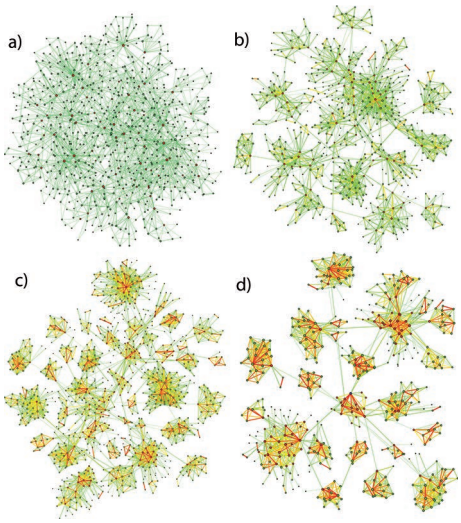

Fig. 2 – Realizzazioni della rete rappresentata in *Fig. 1* simulate considerando quattro diverse configurazioni dei parametri tratte da un modello meccanicistico.

In *Fig. 2* sono riportate quattro realizzazioni della rete di comunicazioni telefoniche simulate considerando quattro diverse configurazioni dei parametri del modello meccanicistico. Ciascuna di queste immagini è da intendersi come una collezione di pseudo-dati generati dal modello e va confrontata con i dati osservati con l'obiettivo di individuare la realizzazione che più si avvicina a questi ultimi. Nelle immagini in *Fig. 2* a livello visivo si nota con una certa facilità che gli pseudo-dati nel pannello (a) sono i più simili alle osservazioni riportate in *Fig. 1*: ne consegue che fra le quattro diverse configurazioni dei parametri è più verosimile che quella che ha generato le osservazioni reali sia la stessa che ha anche generato gli pseudo-dati nel pannello (a). Questa semplice osservazione sta alla base della metodologia statistica nota come ABC acronimo di *Approximate Bayesian Computation* (Marin *et al.* 2012).

2. Evoluzione temporale della scienza dei dati e del suo oggetto

La definizione di *Big Data* presentata all'inizio dell'elaborato assume connotazioni diverse a seconda nel periodo storico di riferimento e mi consente di fare un salto indietro nel tempo e introdurre quella

che ritengo essere la prima storia di successo dei *Big Data*. Siamo a Rothamstead, nella stazione di ricerca sperimentale in agricoltura essenziale per la competitività inglese. Fondata nel 1843, la stazione ha prodotto dati in modo ininterrotto dal 1856 e si qualifica come il primo esempio di laboratorio dedicato esclusivamente all'elaborazione dei dati. Nel 1919 è lì che viene assunto con l'obiettivo di *"investigate the possibility of analysing the vast amount of data accumulated*" Sir Ronarld Fisher (1890-1962), noto statistico e primo *data scientist ante litteram*. Analizzando i dati a propria disposizione, Fisher si accorse che *data is rich but information poor*. Nel corso di cinque anni rivoluzionerà il processo di analisi dei dati raccolti rendendolo efficiente e porrà le basi per la moderna Scienza dei Dati inventando concetti quali il disegno degli esperimenti, l'analisi discriminante e la stima di massima verosimiglianza – tutti strumenti molto utili ancora oggi. La successiva storia di successo si svolge più o meno negli stessi anni ed è relativa alla *Guinness*, produttrice di birra, fra le prime compagnie ad adottare tecniche di controllo e miglioramento di qualità. William Gosset, *alias* Student (1876-1937), abbandona le procedure consolidate basate su processi di tipo *trial and error* e introduce miglioramenti sostanziali al processo di controllo della qualità sistematizzandolo. Al proposito, lo scienziato introduce il cosiddetto *test t di Studen*t e nel 1908 pubblica la sua scoperta utilizzando proprio lo pseudonimo di Student. Motivo della sua scelta era aggirare il divieto imposto dalla *Guinness* di pubblicare i risultati delle ricerche condotte internamente al fine di proteggere il vantaggio competitivo dell'azienda.

Facciamo ora un salto in avanti. Il volo AF 447 da Rio a Parigi si è schiantato nel 2009. Alla fine del 2010, dopo quasi due anni di vane ricerche dei resti, la società *Metron* è stata incaricata di elaborare una mappa di probabilità per identificare le aree in cui più verosimilmente il volo era precipitato. La mappa è stata prodotta considerando una serie di informazioni *a priori* quali i dati di volo e le condizioni meteo locali al momento del disastro. La squadra *Metron* ha utilizzato il metodo di ricerca bayesiano che si basa sulla combinazione di tali informazioni *a priori* con i risultati delle ricerche. La stessa metodologia era stata usata in precedenza con successo durante la guerra fredda per individuare i sottomarini sovietici. Le operazioni di ricerca sono iniziate nel luogo identificato dallo studio *Metron* come il più probabile per il ritrovamento e dopo una settimana sono stati rinvenuti i resti dell'ae-

reo. Quello tra *Big Data* e utilizzo di informazioni *a priori* è stato dunque un matrimonio di successo: quando si cerca un ago in un pagliaio, come un aereo perso in una zona dell'oceano grande quanto la Svizzera, il bosone di Higgs o le onde gravitazionali, eventuali informazioni *a priori* sono fondamentali per indirizzare la ricerca.

Il presidente Obama nello *State of the Union Address* del 2015 ha annunciato di aver stanziato duecentoquindici milioni di dollari per un progetto dedicato allo sviluppo della medicina di precisione, il cui obiettivo principale era l'utilizzo di *marker* biologici, genetici e psicosociali nel processo di scelta delle cure più adeguate per ottenere la massima efficacia e i minimi effetti indesiderati. Una coorte di un milione di americani è stata dotata di dispositivi elettronici che trasmettono continuamente tutte quelle informazioni che vengono classificate come *Big Data* per la loro mole e per il loro grado di dettaglio. Per procedere ad un'analisi sistematica e ottenere un quadro più preciso in termini di prevenzione, diagnosi e trattamento per diverse patologie, il campione viene stratificato in classi omogenee di pazienti con riferimento alle caratteristiche genetiche, comportamentali ed ambientali. Tale approccio, utilizzato anche nell'ambito della farmaco-genomica, risulta valido e apprezzabile anche dal punto di vista statistico e molto più condivisibile del concetto di medicina personalizzata. Personalizzare la cura per il singolo paziente richiede risorse enormi e tempistiche lunghe: molto più sensata ed efficiente è allora l'idea di personalizzare la cura per una coorte di persone omogenee con riferimento alle caratteristiche che sappiamo, *a priori*, potrebbero influire sull'effetto della cura stessa.

Rimanendo in ambito medico, sofisticate analisi statistiche vengono condotte per curare anche le malattie rare che, per definizione, generano tipicamente pochi dati in quanto i casi sono pochi. Nel campo si stanno facendo notevoli passi avanti grazie al raggruppamento di dati multicentrici, studiati mediante meta-analisi e processati con tecniche statistiche all'avanguardia che tengono in conto le diverse condizioni in cui i sotto-campioni sono stati raccolti. Nel settore medico, anche la neonata medicina partecipativa, che prevede il coinvolgimento del paziente nella gestione della propria patologia, beneficia degli strumenti messi a disposizione dalla Scienza dei Dati. Il paziente è generatore di informazioni oltre che utente di informazioni disponibili in internet – dove è piuttosto difficile discernere la veridicità delle fonti. I dati generati dal paziente sono da lui inviati con applicazioni specifiche che, ad

esempio, misurano la glicemia, la pressione arteriosa e il ritmo cardiaco. Il coinvolgimento del paziente nelle fasi di anamnesi e trattamento aiuta il medico ad avere un quadro più preciso e fornisce un'informazione quantitativa bio-umorale in tempo reale. Un esempio è l'applicazione per la glicemia dei pazienti diabetici sviluppata dai colleghi del laboratorio di *Biomedical Informatics* dell'Università degli Studi di Pavia, intitolato Mario Stefanelli, professore di ingegneria biomedica. In un simile contesto si corre tuttavia il rischio della cosiddetta socializzazione della salute ovvero l'inserimento di terzi soggetti nel rapporto fra medico e paziente: così facendo, l'organizzazione sanitaria non si pone più solo come erogatore di servizi medici e servizi correlati ma anche come detentore di informazioni. Inoltre, l'apparato scientifico-tecnico sempre più sofisticato porta il medico a vedere il paziente come somma di organi e non più come insieme olistico e psicosomatico. Dal lato suo, il paziente vede il medico di medicina generale non più come punto di riferimento ma come tramite di smistamento verso specialisti di organo o di apparato. Con queste nuove tecnologie la medicina ha sicuramente guadagnato in termini personalizzazione e tempestività della cura ma forse sta perdendo in umanità diventando depersonalizzata.

In ambito sanitario, vale la pena di introdurre un ultimo esempio. Per gli ospedali dell'Ontario IBM ha creato un sistema di monitoraggio dei nati prematuri: attraverso l'utilizzo di sensori si rilevano parametri vitali quali pulsazione cardiaca, saturazione di ossigeno e pressione arteriosa. Ogni sensore è dotato di una spia luminosa che si accende di verde nel caso in cui il parametro vitale associato sia nella norma. L'evidenza tuttavia mostra che se le spie luminose di un neonato sono tutte verdi e a lungo, quasi sempre si produce un'infezione. Viene naturale chiedersi perché i bambini siano infettati. I medici non sono ancora stati in grado di determinare alcuna concatenazione causale e di individuare meccanismi funzionali capaci di spiegare la relazione tra i due eventi. Quel che è importante, in tal caso, è che certi antecedenti producono quasi sempre certe conseguenze e che, a discapito del senso e della logica medica, un software scientifico permetta di esercitare una migliore prevenzione e di salvare nuove vite.

Tutti gli esempi riportati ci consentono di avanzare una riflessione profonda riguardo il rapporto tra *Big Data* e conoscenza. I ricercatori in futuro avranno a disposizione una quantità sempre più grandi di dati

e stabiliranno delle correlazioni. La tentazione di basare la conoscenza sulle correlazioni diventerà sempre più forte a scapito della comprensione del fenomeno stesso ma io continuerò a ripetere ai miei studenti che *correlation is not causation* mettendoli in guardia rispetto ai pericoli di questo modo di procedere.

3. Impatto sociale dei *Big Data*

Passiamo ora all'analisi dell'impatto economico, sociale ed etico dei *Big Data*. È notevole l'impatto della Scienza dei Dati sull'economia: negli Stati Uniti, consumi e spese legati ad internet, se misurati come un settore, risultano più grandi di quelli legati all'agricoltura o all'energia. Beni e servizi distribuiti attraverso internet sono responsabili per una parte significativa e crescente del PIL mondiale. I dati sono diventati un importante fattore di produzione come ben dimostrato dalla grafica riportata in *Fig. 3* dove si può seguire l'evoluzione temporale delle cinque più grandi imprese mondiali per capitalizzazione. Nel 2016 compaiono *Apple, Alphabet, Microsoft, Amazon* e *Facebook*, tutte società che hanno fatto della generazione e analisi dei dati la loro attività principale e trent'anni fa non esistevano. Nel 2001 figura per l'ultima volta *Walmart*, tutt'ora il più grande rivenditore fisico del mondo, ma superato per capitalizzazione da *Amazon*, nonostante quest'ultima impieghi un decimo dei dipendenti. L'infografica riportata in *Fig. 3* rinforza l'idea che «i dati siano il nuovo petrolio». Infatti le varie *Exxon, Shell, Total*, in prima posizione fino al 2011, sono state sostituite da aziende che vivono di dati.

Se ci focalizziamo sulle spese per la ricerca, progetti che coinvolgono *Big Data* necessitano di di grandi investimenti che si stanno realizzando soprattutto in ambito pubblico. Ne nomino solo alcuni.

Il progetto europeo *Human Brain* partito nel 2013 è cofinanziato dalla Commissione Europea e coinvolge novanta università e scuole di alta formazione di ventidue paesi. *Human Brain* si propone di costruire in dieci anni una mappa dell'attività di cento miliardi di neuroni. È un esempio di modello politico di investimento in ricerca che punta sulla cosiddetta *Big Science* favorendo grandi consorzi con obiettivi ambiziosi e molto visionari che sortiscono però l'effetto collaterale di imporre a un determinato settore di ricerca una visione unitaria in opposizione ad un modello di ricerca basato sulla competizione, la creatività e la molteplicità dei modelli interpretativi.

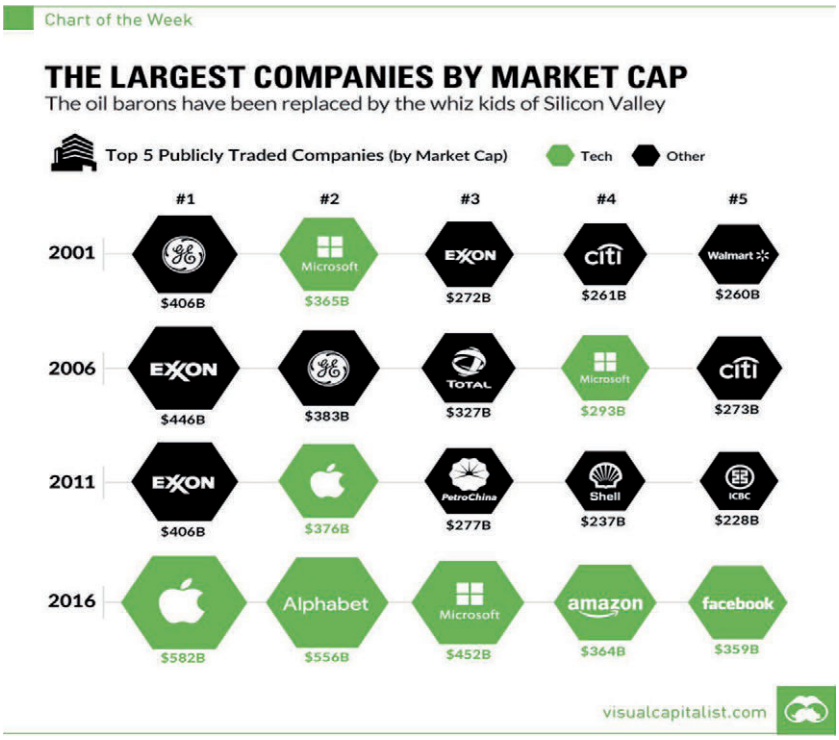

Fig. 3 – Evoluzione temporale delle cinque più grandi imprese per capitalizzazione.

Nell'annunciare il parallelo progetto *US BRAIN*, Obama, nella *State of the Union Address* del 2013 ha detto: *"every dollar we invested to map the human genome returned $140 to our economy. Now is the time to reach a level of research and development not seen since the height of the Space Race"*. La cosa interessante è che i due progetti non sono competitivi ma complementari. Sempre con riferimento alla Scienza dei Dati, il governo inglese ha stanziato seicento milioni di sterline per far partire l'*Alan Turing Institute*, istituto nazionale per la Scienza dei Dati. Nel nuovo polo di ricerca *Human Technopole* che sorgerà sull'area EXPO uno dei sette centri sarà dedicato alla Scienza dei Dati. Dal canto loro i privati non solo sostengono questi grossi progetti ma istituiscono anche con una serie di premi in denaro spesso piuttosto consistenti. Ne nomino solo due. *Netflix* nel 2006 ha istituito un premio di un milione di dollari per l'algoritmo in grado di migliorare il sistema

esistente che raccomanda agli utenti quali film vedere. L'*Heritage Health Prize* è un premo di tre milioni di dollari volto alla ricerca di algoritmi per prevedere ricoveri ospedalieri partendo da dati multi dimensionali sui pazienti.

4. *Big Data* e questioni etiche legate alla *privacy*

Veniamo ora alle questioni etiche e di *privacy*. Siamo diventati contemporaneamente sia produttori sia consumatori di dati ma, mentre la produzione è generalizzata, l'accesso ai dati è selettivo. In relazione alla prima questione si pensi, a titolo di esempio, che alcune popolazioni del centro Africa dispongono di cellulari intelligenti con applicazioni che permettono di segnalare disastri naturali, epidemie o indicare se i pozzi sono asciutti consentendo interventi tempestivi. Dall'altro lato, la ricerca scientifica, la tecnologia e le infrastrutture per estrarre informazioni dai dati e rielaborarle spesso non sono disponibili nei paesi che al meglio potrebbero beneficiare delle rivoluzioni dell'informazione e della tecnologia apportate dai *Big Data*. È bene lavorare per una democratizzazione dei dati e perché non si diffonda la preoccupazione – del tutto ragionevole – che la Scienza dei Dati aumenti la diseguaglianza a causa di un'asimmetria di conoscenze nella gestione e nell'analisi degli stessi. Per evitare questo occorre rendere consapevoli tutte le persone che il patrimonio di dati che generano è fonte di informazioni, conoscenza e quindi valore – questioni, queste, note già negli anni '80 ma che esplodono con i *Big Data* che insidiano sempre più pesantemente la *privacy*. Il problema è ancora più grave nel momento in cui investe persone minorenni. Quando un *teenager* scarica dal web un'applicazione e accetta le condizioni d'uso, sta inconsapevolmente accettando uno scambio di valori: rinuncia a parte della sua *privacy* per avere un servizio. Ritengo opportuno che ciascuno di noi venga reso consapevole di generare dati che diventano veri e propri fattori di produzione: non si tratta infatti soltanto di rinunciare alla propria riservatezza ma di capire quale processo seguano i dati dal momento in cui vengono resi disponibili al momento in cui generano valore.

Tra e sfide intellettuali poste dai *Big Data*, voglio riferirmi alla gestione delle cosiddette *smart city*. Queste ultime vengono intese come insieme di bisogni che trovano risposte in tecnologie, servizi e applicazioni riconducibili a domini diversi quali la gestione trasparente, effi-

ciente e sostenibile di temi come salute, ambiente, governo, benessere, educazione, mobilità ed economia. Tali tecnologie, servizi ed applicazioni non costituiscono di per se, ne singolarmente ne collettivamente, una *smart city*: per essere tali richiedono infatti l'integrazione in una piattaforma che assicuri inter-operabilita e coordinamento, ma soprattutto la definizione di appropriati strumenti di governo e finanziamento, elementi essenziali alla realizzazione della *smart city*. Al centro della sfida vi è la costruzione di un nuovo genere di bene comune, una grande infrastruttura tecnologica e immateriale che faccia dialogare persone e oggetti, integrando informazioni al fine di migliorare la nostra quotidianità e promuovere lo sviluppo economico. Per il successo di queste iniziative occorre che i grandi attori istituzionali come governo, accademia e industria, abbiano una scala di priorità e obiettivi misurabili condivisi. Nella nostra area di riferimento, un caso all'avanguardia che mira a rendere Milano una città ancora più intelligente è il progetto *Urbanscope* sviluppato dal Politecnico. Dal punto di vista dell'individuo una *smart city* può essere anche intesa come ambiente in cui il cittadino è immerso in uno spazio urbano che gli fornisce informazioni, soluzioni e impulsi grazie all'analisi delle sue preferenze e che, attraverso complessi algoritmi, indirizza le sue scelte e quelle di individui con un profilo simile. Quest'ultima riflessione mi porta a introdurre un altro dei pericoli della Scienza dei Dati per la società. Le applicazioni che suggeriscono cosa fare, guardare e leggere rischiano di appiattire e standardizzare il gusto degli utenti: a noi dunque l'incombenza di utilizzarle in modo intelligente, cogliendo gli stimoli che risuonano con le nostre preferenze e rafforzano le nostre opinioni rispetto al comportamento medio.

Mi accingo a concludere il mio intervento facendo riferimento alle sfide che i *Big Data* pongono agli "addetti ai lavori", statistici o ricercatori in generale. Da loro punto di vista, il pericolo più grande è che i *Big Data* possano condurre a risultati errati nel caso in cui la loro analisi non sia supportata da una corretta metodologia scientifica. Al proposito, un esempio comune è che piccole anomalie nei dati possano essere confuse con segnali. La vera difficoltà sta nel porre le giuste domande e adottare la metodologia appropriata: per questo motivo è importante che l'analisi dei *Big Data* non sia lasciata solo a colossi come *Google, Microsoft, Twitter* o *Facebook* che sono guidati da logiche commerciali. Al contrario, è necessario che le università si facciamo promotrici della Scienza dei Dati che ha la potenzialità di creare un nuovo paradigma per la ricer-

ca scientifica. In passato la ricerca ha lavorato in modalità di conferma – come nel caso della fisica delle particelle – occupandosi per lo più della verifica di singole ipotesi in situazioni in cui il fenomeno di interesse era osservato insieme ad un numero enorme di variabili di disturbo**.** Poi è subentrata una modalità esplorativa – pensiamo all'astronomia e alla genomica – in cui è necessario vagliare un gran numero di ipotesi prima di individuare quelle che meritano una ricerca approfondita. Oggi disponiamo invece di una mole consistente di dati che possono essere usati per personalizzare le analisi. Rispetto al passato, in cui i dati erano raccolti per rispondere a domande di ricerca specifiche, oggi molti dati esistono a prescindere da domande scientifiche: non si tratta dunque di dati sperimentali ma osservazionali a partire dai quali si cercano correlazioni e regolartià empiriche che possono generare nuove ipotesi e teorie ispirate dai dati e non più solo verificate sui dati.

L'Istituto Interdisciplinare di Scienza dei Dati dell'USi, alla cui fondazione ho contribuito e che ora dirigo, ha nel suo portafoglio progetti e ricerche avviate che utilizzano i *Big Data*. Io e i miei collaboratori abbiamo a disposizione, ad esempio, dati su tutti i brevetti registrati negli ultimi cinquant'anni e li analizziamo per cogliere le dinamiche dell'innovazione e i flussi di conoscenza. Altri progetti studiano i dati tratti da Wikipedia per capire i fattori critici di successo di alcune pagine, altri ancora impiegano dati longitudinali sul commercio internazionale o sugli scambi interbancari fra gli istituti di credito europei. Disponiamo anche di dati racconti attraverso i social media con la doppia finalità di studiare la diffusione di notizie online per valutare se ci sono differenze nelle dinamiche di notizie vere e false e per la costruzione di indicatori di benessere sociale che permettano di fare previsioni in tempo reale attraverso l'analisi dei sentimenti, delle opinioni e delle emozioni. L'istituto di Data Science è supportato e collabora con il Centro Svizzero di Calcolo Scientifico che, ad oggi, possiede il super computer più potente d'Europa e il quarto al mondo.

Vorrei concludere dando uno sguardo al futuro. Ritengo che l'approccio multidisciplinare tipico della Scienza dei Dati sia vincente anche per l'accesso a finanziamenti competitivi per la ricerca. Nel mio ambito specifico di ricerca, per esempio, sta nascendo l'esigenza di creare omogeneità tra il modo di pensare di tipo computazionale – che prevede astrazione, modularità, scalabilità e robustezza – con il paradigma di

ricerca di tipo inferenziale – che è probabilistico ed utilizza i concetti di campionamento e di rischio e mira ad effettuare previsioni e supportare decisioni. Fino ad ora le scienze computazionali e la statistica si sono sviluppate in modo separato. Mentre la natura multidisciplinare della Scienza dei Dati sta innestando un processo di armonizzazione tra le due discipline perché queste possano beneficiare l'una del paradigma metodologico dell'altra. In particolare, l'obiettivo è sensibilizzare i *computer scientist* al concetto di rischio statistico e render gli statistici inferenziali più familiari con aspetti più strettamente computazionali. Per darvi l'idea di come la fusione disciplinare si potrebbe perseguire, cito il motto latino del *d vide et mpera*. Se il singolo problema viene affrontato separatamente dalle due discipline secondo le proprie competenze specifiche, l'obiettivo ultimo di entrambe le comunità dovrebbe essere quello di instaurare un dialogo aprendosi ad approcci metodologici complementari. A mio parere, la Scienza dei Dati può diventare una guida per scoperte scientifiche, ispirare innovazione, e può essere uno strumento trasversale in grado di aumentare la collaborazione e la condivisione fra ambiti disciplinari diversi favorendo l'interazione positiva fra i saperi. Ritengo tuttavia che la Scienza dei Dati non si possa sostituire al giudizio umano – che deve sempre mediare le nostre relazioni con computer e algoritmi di analisi dei dati. Non possiamo inoltre prescindere da curiosità e creatività e abbiamo il dovere di educare i giovani ad un approccio multidisciplinare, armonizzando diverse visioni proprio come sta accadendo all'approccio statistico e computazionale.

5. Ringraziamenti

Desidero ringraziare Francesco Bartolucci, Federica Bianchi ed Alessandro Lomi con cui ho discusso i temi oggetto della presente nota e che hanno contribuito alla sua revisione. Eventuali refusi o errori rimangono comunque piena responsabilità dell'autore.

BIBLIOGRAFIA

JM. Marin *et al.*, *Approximate Bayesian computational methods,* Statistics and Computing 22, 6 (2012), 1167-1180.

JP. Onnela et al., *Structure and tie strengths in mobile communication networks*, Proceedings of the National Academy of Sciences 104, 18 (2007), 7332-7336.