



*Dipartimento di Scienze Politiche    Cattedra Metodologia delle scienze sociali*

## Big Data: quale potere predittivo per le scienze sociali

RELATORE  
Prof.ssa Albertina Oliverio

CANDIDATO  
Gian Paolo Garelo  
080122

ANNO ACCADEMICO 2016/2017

# INDICE

<b>1. Introduzione .....</b>	<b>2</b>
<b>2. Capitolo 1: Un metodo, tanti metodi .....</b>	<b>3</b>
<b>1.1 Previsione scientifica</b>	
<b>1.2 Un problema di filosofia</b>	
<b>1.3 Una soluzione</b>	
<b>3. Capitolo 2: Big Data .....</b>	<b>9</b>
<b>2.1 Definizione</b>	
<b>2.2 Nowcasting</b>	
<b>2.3 Razionalità</b>	
<b>4. Capitolo 3: Limiti dei Big Data .....</b>	<b>18</b>
<b>5. Conclusioni .....</b>	<b>25</b>
<b>6. Abstract.....</b>	<b>26</b>

## INTRODUZIONE

Fin dall'alba dei tempi l'uomo ha sempre cercato di prevedere il futuro. Inserito in un contesto ambientale, la conoscenza delle variabili che influivano sulla sua vita gli permetteva di trarre vantaggi quotidiani e in ultima istanza garantiva la sua sopravvivenza. Dall'uomo delle caverne fino ai giorni nostri ne è passato di tempo, ma una cosa che non è cambiata è quel desiderio innato di sapere cosa avverrà, di poter dominare il futuro. Se inizialmente la sfida era quella di prevedere il moto di oggetti fisici, ora l'oggetto di questa ricerca è più ambizioso e sicuramente più complesso: l'uomo.

La fisica classica esaurisce in pieno il suo compito per i moti degli oggetti sulla Terra, ma quando si passa ad investigare l'essere umano, con le sue convinzioni, credenze, abitudini si rende necessario l'uso di nuove tecniche e strumenti. I modelli fisici e i concetti correlati di Teoria, Dati sperimentali, Previsione e Legge costituiranno la base sulla quale costruire una teoria con la quale cercare di prevedere il comportamento umano con una sostanziale novità tecnologica: i "Big Data".

Nel primo capitolo si procederà con la comparazione delle scienze fisiche e sociali mettendo in luce lo sviluppo delle teorie epistemologiche nel corso della storia fino ad arrivare alle conquiste recenti come il modello nomologico deduttivo.

Se per corroborare le teorie i fisici hanno per secoli condotto esperimenti e raccolto dati, ora sono gli stessi oggetti osservati, le persone, a disseminare su Internet moli gigantesche di informazioni che possono essere studiate e analizzate per comprendere i loro comportamenti. Ma le azioni umane da che cosa dipendono? O ancora, se risulta inutile ai fini della nostra ricerca conoscere le cause di una azione, con che tipo di razionalità agisce l'uomo? Questi temi saranno trattati nel secondo capitolo; partendo da una definizione del concetto di Big Data si mostreranno le applicazioni empiriche, e verrà approfondito il concetto di razionalità umana.

Infine nel terzo capitolo si evidenzieranno i limiti tecnologici e concettuali di una teoria che si appoggia sui Big Data, e di come le problematiche circa la rappresentatività del campione e le conseguenze inintenzionali, per citarne alcuni influenzino i risultati.

Considerando che già poco più di metà della popolazione mondiale ha accesso a Internet, le prospettive per questo ambito appaiono ampie e da investigare.

## CAPITOLO 1 “Un metodo, tanti metodi”

### 1.1 Previsione

La previsione svolge un ruolo di controllo fondamentale per tutte le teorie scientifiche. Ogni ricercatore, che sia di una scienza naturale o umanistica, elabora delle leggi o delle teorie scientifiche al fine di spiegare fatti già noti e prevedere altri che non lo sono<sup>1</sup>.

Se la legge o la teoria effettivamente hanno predetto il fenomeno, ciò costituirà una conferma sperimentale che le renderà valide. La previsione però costituisce l'ultima fase di un processo scientifico basato su una teoria formulata dal ricercatore, che permette di enunciare una legge.

Quest'ultima sarà corroborata con delle prove sperimentali che decreteranno la validità della teoria.

Questo algoritmo di passaggi scientifici è chiamato metodo.

### 1.2 Un problema di filosofia

Per arrivare a spiegare un fenomeno possono essere usati molti metodi: si può ricorrere all'intuizione, la forma privilegiata di conoscenza per Platone, ci si può basare sulle proprie esperienze in materia, ma una spiegazione per poter essere considerata “scientifica” deve basarsi su un metodo altrettanto scientifico.

Nel corso dei secoli gli uomini di scienza si sono interrogati su quale fosse il metodo per raggiungere una conoscenza *certa*, fino ad arrivare alla creazione di una branca della Filosofia della scienza che affronta in pieno questo interrogativo: l’*“Epistemologia”*.

Una delle personalità scientifiche più importanti del secolo scorso, in questo campo, fu “Karl Popper”, epistemologo e filosofo austriaco naturalizzato britannico propositore della falsificabilità come criterio di demarcazione tra scienza e non scienza.

Per Popper una teoria per essere considerata scientifica deve essere falsificabile ovvero nella sua formulazione deve vietare alcuni accadimenti. La ricerca scientifica si deve quindi muovere

---

1. Dario Antiseri *Epistemologia e didattica delle scienze*, Armando editore, Roma, 2000

nell'ottica dell'individuazione dell'errore nella teoria, non nella sua conferma, perché mentre la conferma sperimentale sul piano empirico è sempre e solo temporanea, la sua smentita sul piano logico è definitiva.

Celebre in tal senso è il cosiddetto esempio del cigno nero. Prendiamo un enunciato universale "tutti i cigni sono bianchi" e sottoponiamo la nostra teoria a controllo sperimentale. Per essere sicuri che tutti i cigni siano bianchi, in teoria, si dovrebbero controllare tutti i cigni presenti e futuri, ma esaminando un campione limitato di cigni non potremmo giungere alla conclusione che effettivamente tutti i cigni siano bianchi.

Supponiamo che un giorno arrivi da un paese straniero un cigno nero, ecco quell'unico cigno rappresenta la falsificazione della teoria, l'enunciato di partenza che si proponeva come universale è stato smentito da quell'evento. Questo risultato rappresenta per Popper la fine di quella teoria che deve essere scartata immediatamente. Una teoria scientifica più prove sperimentali riesce a sopportare più sarà valida, fino ad arrivare al momento in cui verrà smentita.

In questo senso il falsificazionismo popperiano non accetta teorie secondarie alla principale, che giustificano fenomeni che la principale non riesce a spiegare, sono da rifiutare quindi le ipotesi ad hoc e le ipotesi ausiliarie.

Il Falsificazionismo popperiano lancia inoltre un pesante attacco al metodo induttivo. Come dimostra l'esempio del cigno nero, la validazione è solo temporanea e dall'osservazione particolare non si può arrivare ad una conclusione generale. Questo tipo di ragionamento, chiamato dai logici "inferenza induttiva" muove da premesse particolari e arriva a conclusioni generali. È una inferenza non valida in quanto dalla premessa iniziale non discende la conclusione generale. I limiti del metodo induttivo erano già stati sottolineati nei secoli passati dal filosofo scozzese David Hume, mettendo in luce la radice psicologica che porta ad affidarsi a questa metodica.

Nella vita quotidiana si dà per scontato il "principio dell'uniformità della natura" che porta a credere che in virtù del fatto che i fenomeni siano sempre accaduti in un determinato modo continuo a farlo anche in futuro, un chiaro esempio di inferenza induttiva.

L'essere umano per definizione non può essere considerato alla pari di un oggetto, il modello causa-effetto forse può essere applicato ma con alcune distinzioni se si rifiuta un'ottica riduzionistica che per secoli è stata presente nella filosofia della scienza occidentale.

Diversi metodi e approcci sono stati cambiati nel corso degli anni con lo scopo di dare scientificità alle ricerche in campo sociale mettendo in luce le concezioni culturali e ideologiche dei periodi storici nelle quali sono nati.

In pieno Positivismo, la società, avendo beneficiato delle scoperte e innovazioni tecnologiche, riponeva una fiducia nella scienza quasi illimitata fino ad arrivare a formulare visioni deterministiche del sapere scientifico. A questo proposito famoso è il cosiddetto "demone di Laplace": un'intelligenza

sovrumana, che secondo lo scienziato francese, se fosse un osservatore onnisciente e disponesse di tutte le informazioni e le leggi che regolano la natura si accorgerebbe della struttura causale dell'universo e sarebbe in grado di spiegare il passato e prevedere il futuro.

Ecco le parole testuali:

“Possiamo considerare lo stato attuale dell'universo come l'effetto del suo passato e la causa del suo futuro. Un intelletto che ad un determinato istante dovesse conoscere tutte le forze che mettono in moto la natura, e tutte le posizioni di tutti gli oggetti di cui la natura è composta, se questo intelletto fosse inoltre sufficientemente ampio da sottoporre questi dati ad analisi, esso racchiuderebbe in un'unica formula i movimenti dei corpi più grandi dell'universo e quelli degli atomi più piccoli; per un tale intelletto nulla sarebbe incerto ed il futuro proprio come il passato sarebbe evidente davanti ai suoi occhi”<sup>2</sup>.

Il determinismo fisico poggiava le sue basi sulla solida visione newtoniana dell'universo visto come un gigantesco meccanismo “preciso come un orologio” similitudine usata per indicare la massima attendibilità e perfezione meccanica. Un sistema, quello newtoniano, quindi dove ad ogni movimento corrisponde un'equazione in grado di descrivere la sua traiettoria. Con queste premesse per lo scienziato britannico era possibile arrivare a formulare un “sistema del mondo” o di come diremmo oggi della “Teoria del tutto”; ambizione che rivelerà i suoi limiti solamente secoli dopo la sua enunciazione, con lo studio della realtà microscopica dell'atomo e delle vaste estensioni dello spazio interstellare<sup>3</sup>.

Questa impostazione concettuale porterà a trasporre questi stessi metodi matematici allo studio del comportamento umano. Emblematica è l'opera “Fisica sociale” dello statistico Adolphe Quetelet incentrata sullo studio del cosiddetto “uomo medio”, un uomo che poteva essere considerato perfettamente nella media statistica della popolazione per le sue caratteristiche generali e quindi che poteva essere preso a modello per una efficace pianificazione sociale.

Nella fattispecie lo strumento statistico centrale nel ragionamento dello scienziato belga è la curva di distribuzione normale o gaussiana, che, dopo essere stata applicata con successo a innumerevoli fenomeni naturali, come la distribuzione dell'altezza di una popolazione o la distribuzione degli errori nelle osservazioni astronomiche, venne adoperata sull'uomo; l'uomo medio per Quetelet non era solo un'astrazione matematica, bensì un ideale morale<sup>4</sup>.

- 
2. Pierre Simon de Laplace *Saggio filosofico delle probabilità* 1812
  3. Ian Stewart *Dio gioca a dadi?* Bollati Boringhieri Torino, 2010, p. 7
  4. Ian Stewart *Dio gioca a dadi?* Bollati Boringhieri Torino, 2010, p. 50

Questa impostazione è stata “istituzionalizzata” nelle scienze sociali da colui che si definì il padre della sociologia Auguste Comte; infatti il termine sociologia è usato per indicare “lo studio positivo delle leggi fondamentali proprie dei fenomeni sociali”. Lo scienziato francese immaginava lo sviluppo della storia dell’uomo come un percorso orientato verso il progresso grazie al ruolo giocato dalla scienza, che nella gerarchia comteiana occupa il primo posto e rappresenta il modello a cui tutte le altre si devono ispirare. In questa ottica è chiara la definizione di Comte della sociologia come una “fisica sociale” ovvero un sistema governato da leggi alle quali gli uomini obbedivano al pari degli oggetti nel mondo naturale.

Le scoperte in campo fisico dei secoli successivi segnarono il definitivo tramonto di qualsiasi tipo di determinismo. L’introduzione in campo fisico di concetti come quello della probabilità ad opera di Maxwell nello studio dei gas diede una svolta significativa alla concezione del sapere scientifico, che lungi dalla certezza matematica, ora si basa su stime e probabilità.

Le teorie di Einstein misero in luce come l’applicazione delle leggi di Newton sia corretta solamente per casi specifici, ovvero per oggetti a “basse velocità”, e che nello studio di fenomeni astronomici non sia più valida.

Le conseguenze in ambito sociale vengono recepite dall’epistemologo Karl Popper il quale fa della fallibilità e dei limiti della conoscenza i criteri chiave che lo porteranno alla formulazione del criterio di fallibilità.

Nonostante le scienze siano state depurate in parte dalla pretesa di determinismo, figlia com’è evidente di quel contesto storico, il ruolo della spiegazione scientifica è sempre presente come garante di oggettività.

Guardando al modello della spiegazione causale proprio della fisica, le scienze sociali hanno cercato di spiegare i fenomeni sociali nel modo più oggettivo possibile: un esempio è costituito dal “*comportamentismo*”. Questo approccio studia i comportamenti dei soggetti esaminati senza investigare le motivazioni, i pensieri tutte cause che non possono essere misurate sperimentalmente.

Il comportamento umano viene quindi considerato la “risposta” ad uno stimolo, meccanismo stimolo-risposta che riprende il modello causa-effetto.

Celeberrime sono le ricerche di Ivan Pavlov sugli stimoli sugli animali, nella fattispecie sui cani. Lo scienziato riuscì a collegare e influenzare il comportamento dell’animale partendo dallo schema stimolo risposta. Quando al cane viene somministrato il cibo esso reagisce allo stimolo con una risposta fisica, ovvero la salivazione. Associando per un numero di volte la ciotola di cibo con un rumore esterno prodotto artificialmente, l’animale inizierà a salivare al solo udire il suono, senza che sia presente realmente il pasto. Questo dimostra come il comportamento possa essere influenzato con uno stimolo artificiale.

Un altro tipo di apprendimento è quello del tipo “prove ed errori” teorizzato dallo psicologo statunitense Edward Lee Thorndike. Con i suoi esperimenti riuscì a sottolineare come il comportamento del soggetto studiato (sempre un animale) possa essere influenzato dalla ricompensa che avrà nel caso di successo della prova a cui è sottoposto. La ricompensa è determinante in quanto permette al soggetto l’associazione fra comportamento corretto da eseguire e premio. Ripetendo gli esperimenti è evidente come il numero di errori compiuti nelle prove diminuisca sensibilmente, mettendo in luce la funzione “didattica” della ricompensa.

Su un bambino invece, risultati simili sono presenti negli esperimenti condotti dallo psicologo John Watson. L’associazione in questo caso consisteva nel mostrare un topo ad un bambino di 11 mesi associandolo ad un rumore di forte intensità. Il bambino, a cui in un primo momento la sola visione del topo non suscitava paura, con il susseguirsi degli esperimenti provò quello stato d’animo anche in assenza del rumore, dimostrando quindi come la risposta di paura potesse diventare una risposta condizionata al rumore.

Le tecniche di apprendimento vennero approfondite negli studi di Burrhus Skinner che pose l’accento sul concetto di rinforzo. Le ricompense o premi (i rinforzi) all’interno di un processo di apprendimento svolgono un fondamentale ruolo di “modellamento attivo” sul soggetto portandolo a trovare la soluzione e a diminuire nel corso delle prove a cui sono sottoposti gli errori compiuti. Tecniche che risultano efficaci anche nell’educazione primaria.

Anche se influenzato dalle premesse positiviste comtiane, lo “studio del suicidio” di Emile Durkheim rappresenta un caposaldo per la ricerca sociologica. Il merito di Durkheim è quello di investigare con la sua opera il “fatto sociale” del suicidio basandosi su statistiche e dati, quindi utilizzando metodi empirico-scientifici. Tenendo sempre presente una struttura di tipo causale nelle scienze sociali, individua le cause non negli stati di coscienza individuale bensì in altri fatti sociali quali la religione, la famiglia, politica che sono riconducibili in ottica generale al concetto di “integrazione” dell’individuo nella società. Con queste premesse egli fu in grado di spiegare una serie di dati empirici rilevanti: come il motivo per il quale il tasso di suicidio fosse più alto fra le popolazioni di religione protestante rispetto a quelle cattoliche, tra i celibi che tra i coniugati.

Storicamente il dibattito sul metodo da applicare alle scienze sociali ha avuto la sua massima espressione sul finire dell’800 in Germania con quello che viene chiamato “Methodenstreit”. A livello epistemologico questo scontro tra visioni differenti rappresenta il cuore della questione sul metodo, ovvero se si possa parlare di spiegazione scientifica anche per le scienze sociali, e quindi dare dignità a queste discipline al pari delle scienze fisico-naturali. Le posizioni in gioco sono due, la prima è sostenitrice del “monismo metodologico” ovvero che sia possibile applicare la spiegazione causale al mondo delle scienze sociali ottenendo risultati validi la seconda sottolineando come l’oggetto di studio delle scienze sociali sia sostanzialmente differente da quello delle scienze naturali propone un

“dualismo metodologico” che presuppone l’uso di concetti come il “senso”, il “significato” e la “comprensione”. Questi ultimi vengono osteggiati aspramente dal primo gruppo in quanto allontanano la ricerca sociale dall’oggettività.

Una soluzione eclettica che riesce a superare l’impasse del *Methodenstreit* è quella della “sociologia comprendente” di Max Weber. Lo studioso tedesco compie un’opera di sintesi tra l’unicità dei fenomeni sociali che alla luce dello storicismo tedesco poteva essere compresa solo tramite l’empatia e la volontà di trovare una spiegazione oggettiva all’evento in esame.

Lo strumento metodologico che permette di compiere questa operazione è quello dell’idealtipo. L’idealtipo è un quadro teorico dell’individuo dove vengono studiate dei tratti chiave rilevanti per lo scienziato ai fini della spiegazione di un fenomeno. Essendo un modello astratto vengono trascurati innumerevoli dettagli e condizioni psicologiche e soggettive, permettendo una depurazione da tutte quelle caratteristiche che da parte dello storicismo tedesco rappresentavano un ostacolo a qualsiasi generalizzazione e universalizzazione. Una presa di distanza dal metodo empatico e dallo psicologismo. Lo strumento dell’idealtipo permette quindi generalizzazioni empiriche che collegano delle cause a degli effetti, celebre è il lavoro “L’etica protestante e lo spirito del capitalismo”. Partendo dall’idealtipo del protestante, mosso per motivi religiosi alla valorizzazione del lavoro, all’accumulazione della ricchezza, egli spiega come l’effetto sia stato lo sviluppo del sistema capitalista. Parlare di cause ed effetti però sembra azzardato, alla luce degli studi weberiani. A livello epistemologico è più corretto parlare di correlazioni in quanto nelle scienze sociali non sono presenti “leggi” come propriamente intese nel campo delle scienze naturali; esse tendono a indicare cosa generalmente gli individui fanno in determinate situazioni, ma non cosa *necessariamente* faranno<sup>1</sup>.

Come esito del “*Methodenstreit*” dalle scienze sociali e fisiche è accettato il modello causale proposto dai filosofi Carl Gustav Hempel, Paul Oppenheim e Karl Popper conosciuto come modello “nomologico deduttivo”. Il modello è costituito da due parti: l’explanandum, ovvero il fenomeno che deve essere spiegato e l’explanans, costituito dalle condizioni iniziali più la legge generale. Per spiegare un evento E (explanandum), si parte dalle premesse iniziali e dalle leggi generali (explanans), e si deduce logicamente il fenomeno E. Sono presenti nel modello la parte nomologica, costituita dalla legge generale e quella deduttiva, che permette di arrivare logicamente in modo valido alla spiegazione causale del fenomeno.

---

1. Albertina Oliverio *Individuo, natura, società* Mondadori Milano 2015 p. 31

## CAPITOLO 2 BIG DATA

Quanti più dati il ricercatore riesce raccogliere per controllare la sua teoria tanto più questa sarà accurata. Grazie ai progressi degli hardware nel campo digitale vengono raccolte sempre più informazioni scritte sotto forma di dati, questo permette di avere riscontri empirici precisi. In particolare nell'ultimo decennio è diventato popolare il termine "Big Data" il quale indica una grande quantità di dati e informazioni che, con i tradizionali strumenti di analisi, non possono essere processati o analizzati.

Ecco la definizione di IBM: "Il 90% tutti i dati nel mondo è stato creato solo negli ultimi due anni e ogni giorno ne creiamo 2,5 quintilioni bytes. Questi dati vengono dappertutto: dai sensori usati per raccogliere informazioni climatiche, i post sui social media, le fotografie digitali e i video, i registri delle transazioni degli acquisti fino ai segnali GPS dei cellulari solo per nominarne alcuni. I Big Data sono questo<sup>1</sup>.

Le loro caratteristiche sono: il volume, la varietà e la velocità. Per quanto riguarda il volume dei dati immagazzinati nel mondo stiamo assistendo ad un incremento significativo di questo fenomeno: basti pensare che nell'anno 2000, 800.000 petabytes (PB) di dati erano immagazzinati nel mondo, secondo recenti studi il volume previsto nel 2020 sarà di 35 zettabytes (ZB)<sup>2</sup>.

Un altro aspetto problematico è la varietà dei formati. Nell'universo digitale i dati sono codificati in formati differenti a seconda della loro provenienza che può essere un sensore, uno smart device, una e-mail, un documento facendo sì che i dati da analizzare siano dati "grezzi" "semistrutturati o "non strutturati" rendendo impossibile una lavorazione di questi con le tradizionali piattaforme di analisi.

Infine, una sfida importante per le piattaforme di calcolo è quella di analizzare i Big Data in tempo reale.

Questa capacità permette per esempio ad un'azienda di collocare più efficacemente un prodotto sul mercato, consente di cambiare il portafoglio di azioni di un cliente in tempo reale a seconda degli indici finanziari.

---

5. IBM:<http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>

IBM: *Big data, bigger outcomes* (PDF)

6. Per avere un'idea del volume il lettore consideri che 800 PB=0,0008 ZB *Understanding Big Data* [www.ibm.com](http://www.ibm.com)

Parallelamente all'aumento della capacità di memoria hardware negli ultimi anni si è assistito ad un incremento significativo dell'utilizzo di Internet.

Partendo dal numero di 16 milioni di utenti attivi nel 1995, nel dicembre 2005 si è arrivati al 1.018 miliardi di persone connesse. Numeri che continuano a crescere fino ad arrivare a 3.885 miliardi di utenti nel giugno di questo anno, ovvero il 51,7% della popolazione mondiale.

Secondo le stime dell'“International Telecommunication Union”, l'agenzia delle Nazioni Unite specializzata per la tecnologia dell'informazione e della comunicazione, nell'anno 2016 gli utenti di Internet sono stati 3,384,58 miliardi a fronte di una popolazione mondiale di 7,432,66 miliardi di persone, ovvero più della metà<sup>3</sup>. Per “utente” si intende un individuo che può accedere a Internet da casa, indipendentemente dalla frequenza con la quale si connette. I requisiti per l'accesso sono un apparecchio hardware funzionante, un abbonamento ad Internet attivo e la possibilità per l'individuo di accedere ad Internet in qualsiasi momento, ovvero senza impedimenti<sup>4</sup>.

I Big Data rappresentano quindi le tracce digitali che le attività quotidiane di miliardi di persone lasciano con l'utilizzo di sistemi ICT<sup>5</sup>.

Il largo utilizzo di questo strumento rappresenta per lo scienziato sociale una fonte immensa di informazioni per le sue ricerche. Si parla a livello accademico di “Social media mining” per indicare che il web, e le tracce lasciate dagli utenti sia considerabile al pari di una miniera da cui estrarre elementi preziosi.

Applicazioni di questo fenomeno sono già presenti nel mondo delle previsioni a breve termine.

In opposizione al concetto di “previsione” (forecast) già affrontato nel primo capitolo si parla oggi di “nowcasting” ovvero la previsione a brevissimo termine, quasi in tempo reale del fenomeno studiato.

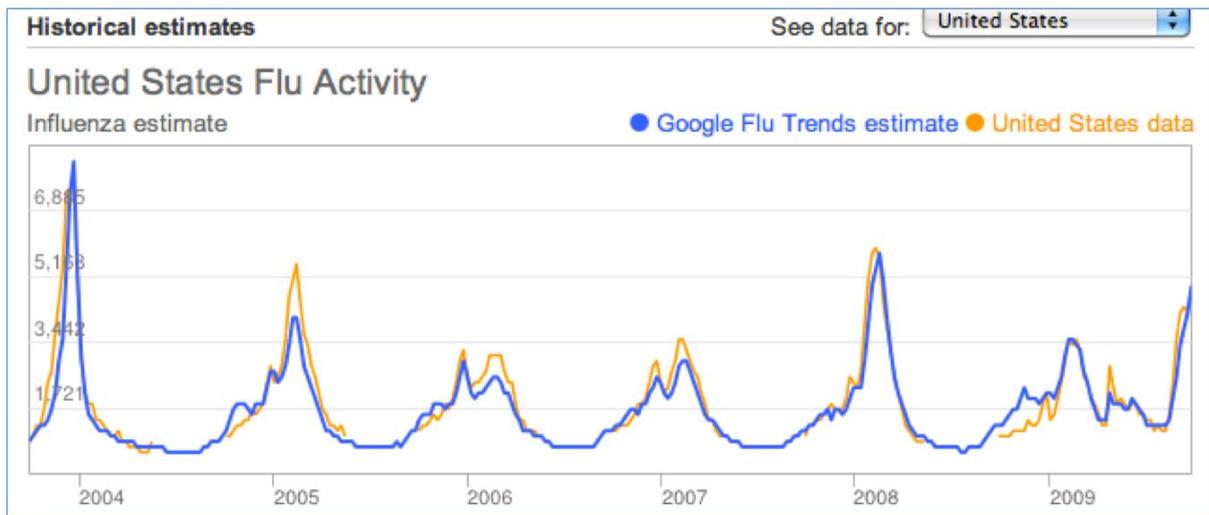
Questo termine nasce in ambito meteorologico indicando appunto una tecnica per una previsione del tempo a breve termine, usando una stima delle variabili in gioco e assumendo una sostanziale continuità del comportamento osservato grazie proprio al ridotto intervallo di tempo considerato<sup>6</sup>.

- 
7. [http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx\\_end-2017\\_estimates\\_for\\_key\\_ICT\\_indicators](http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx_end-2017_estimates_for_key_ICT_indicators) (excel)
  8. <http://www.internetlivestats.com/internet-users/#definitions>
  9. Information Communication Technology. Indica l'insieme delle tecnologie che consentono il trattamento e lo scambio delle informazioni in formato digitale.
  10. <https://www.metoffice.gov.uk/learning/science/hours-ahead/nowcasting>

Passando al campo economico, il “nowcasting” viene definito dal “working paper della Banca Centrale Europea” come una predizione del presente, del futuro immediato e del passato appena trascorso<sup>7</sup>. A differenza del campo meteorologico dove partendo dalla conoscenza certa del meteo giornaliero si possono fare previsioni sul futuro, in economia si rende necessario una conoscenza sempre più accurata del passato appena trascorso, del presente e nei limiti dell’immediato futuro in quanto si parla di valori stimati e quanto più precise sono le stime quanto più efficaci saranno le policies adottate.

Stime accurate che permettono di essere determinanti in altri settori chiave come quello epidemiologico. Un esempio popolare è lo studio<sup>8</sup> che ha dimostrato come i query logs<sup>9</sup> possano predire l’incidenza delle malattie influenzali (Giannotti 2014).

- 
11. <https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1275.pdf?ba24acc83cc23e6e268f6be4069ceb46> NOTA: questo Working Paper non può essere considerato rappresentativo della posizione della B.C.E. in merito all’argomento. Le posizioni espresse sono quelle degli autori e non rappresentano necessariamente quelle della B.C.E.
  12. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. *Detecting influenza epidemics using search engine query data*. Nature 457:1012–10155. (2008)
  13. Per “query logs” si intendono le ricerche su Internet, nella fattispecie su Google.



**In colore blu la stima di Google Trend**

**In colore arancione i dati ufficiali forniti da U.S. Centers for Disease Control**

Come mostra il grafico la correlazione fra le ricerche della popolazione statunitense sull'influenza e i dati ufficiali forniti dal governo americano è evidente, le curve appaiono sovrapponibili lungo quasi tutto il grafico con scarti poco significativi.

La differenza sostanziale fra i dati forniti dalle ricerche su Google e quelli ufficiali è la tempestività con la quale sono disponibili. Gli ultimi necessitano più tempo per essere elaborati e fornire dei risultati accurati; nella fattispecie se il governo fosse intervenuto basandosi sui dati delle ricerche la risposta sarebbe stata tempestiva, in quanto i dati sarebbero pervenuti in tempo reale e verosimilmente efficace data la corrispondenza fra i due grafici.

Dimostrato come il grado di precisione può essere raggiunto con ricerche di Nowcasting, un'applicazione che sta diventando sempre più frequente in ambito sociale è il "Twitter mood", che ha riscontrato un notevole successo in ambito economico.

Partendo dall'assunto della sociologia comportamentale secondo cui le emozioni dell'individuo incidono sul suo processo di "decision making", in teoria estendendo il ragionamento ad un'ottica macro, sarebbe possibile comprendere come l'umore della popolazione influenzi l'insieme delle decisioni prese, per esempio quelle in ambito economico.

È questa la domanda posta nella ricerca "Twitter mood predicts the stock market" dove si studia il grado di correlazione fra l'umore delle persone, misurato attraverso Twitter e le decisioni economiche nello Stock market usando come indice da analizzare il Dow Jones Industrial Average (DJIA). Il periodo analizzato va da marzo a dicembre 2008, nel quale sono stati analizzati 9,7 milioni di tweet postati da 2,7 milioni di individui. L'analisi delle emozioni dei tweet condivisi è stata fatta con l'utilizzo di due algoritmi: il primo, chiamato Google-Profile of Mood States (GPOMS), registra il livello di sei stati d'animo: vitalità, calma, sicurezza, allerta, gentilezza e felicità, il secondo invece

più generico “Opinion Finder” suddivide il “sentiment” fra positivo e negativo. Cercando una relazione fra la variabile umana e quella economica effettivamente si è trovata una correlazione tra l’indicatore della calma GPOMS e il comportamento del Dow Jones Industrial Average<sup>10</sup>.

Risultato straordinario che sembra confermare la potenza e la precisione di queste tecniche predittive. Sempre attraverso lo studio dei commenti su Twitter è stato possibile prevedere con un certo grado di precisione gli incassi dei film al botteghino<sup>11</sup>. Il metodo di previsione è molto semplice, consiste nel conteggio sulla piattaforma Twitter del nome del film ma solo dopo la sua uscita.

Anche in questo campo, la predizione grazie ai social media si è rivelata essere più precisa dei precedenti metodi di previsioni. Uno strumento quindi, quello della Social Media Analysis che appare indispensabile in sempre più campi, per avere previsioni accurate.

Le applicazioni e le potenzialità di questi strumenti risultano enormi per le scienze sociali.

Usando le parole di Giannotti (2014) i Big data “sono le briciole che registrano la nuda verità sui comportamenti individuali e collettivi con una precisione senza precedenti, in modo che le diverse dimensioni della nostra vita sociale trovano un’immagine riflessa nello specchio digitale: desideri, opinioni, stili di vita, movimenti, relazioni”<sup>12</sup>.

Questi dati rappresentano quindi ciò che la gente pensa e desidera con una sostanziale differenza qualitativa. Precedentemente, nella fase di raccolta di informazioni soprattutto con lo strumento statistico del sondaggio, le risposte del soggetto intervistato erano influenzate dal cosiddetto “social desirability bias” ovvero la tendenza del soggetto a fornire risposte all’intervistatore che siano socialmente accettate e desiderate anziché dichiarare ciò che egli pensa in realtà<sup>13</sup>.

Questo comportava quindi dei risultati falsati e poco veritieri, pericolo che viene superato grazie all’analisi delle ricerche effettive del soggetto su Internet, che di fronte alla tastiera agisce in modo più spontaneo come sostiene Rudder in *Dataclysm: Who We Are When We Think No One’s Looking*. Inoltre il sistema di raccolta dati sul Web è meno dispendioso in termini di tempo e più discreto rispetto al sondaggio tradizionale perché è il soggetto studiato a fornire direttamente le informazioni, senza essere disturbato.

---

14. [https://www.researchgate.net/publication/47405468\\_Twitter\\_Mood\\_Predicts\\_the\\_Stock\\_Market](https://www.researchgate.net/publication/47405468_Twitter_Mood_Predicts_the_Stock_Market) <https://www.technologyreview.com/s/421251/twitter-mood-predicts-the-stock-market/>

15. Bernardo A. Huberman *Predicting the Future With Social Media*

16. Giannotti *Big data e social mining: i dati a saperli ascoltare raccontano storie* 2014

17. <https://www.alleydog.com/glossary/definition.php?term=Social%20Desirability%20Bias>

L'evidenza di come le risposte del sondaggio possano essere influenzate da un "bias" sociale mostra come non si possa partire dall'assunto della perfetta razionalità dell'individuo quando si studiano i fenomeni sociali. Essendo l'individuo l'oggetto di questo studio si rende necessario un approfondimento del modo in cui prende decisioni e secondo quali motivi. Innanzitutto si deve distinguere se sta agendo singolarmente o se sia inserito in un gruppo. Si consideri la prima casistica in quanto più vicina alla realtà delle ricerche su Internet.

Nelle scienze sociali per capire il senso delle scelte individuali e quindi in ultima istanza per spiegarle, gli approcci individualistici si sono basati sulla teoria della razionalità dell'attore. Tale approccio partendo da un livello di analisi "micro" dei fenomeni sociali studia il comportamento dell'attore e le sue interazioni con gli altri. Interazioni e scambi che vengono trattati in un'ottica economica in quanto costituiscono un trade off quotidiano al quale tutti gli attori sono sottoposti: in regime di scarsità la relazione con l'Altro si rende necessaria. Da queste premesse ne scaturisce un modello nato nel campo economico e poi prestato alle scienze sociali più in generale, il modello dell'"homo oeconomicus".

L'individuo, considerato come un agente economico, agisce razionalmente; si assume che disponga di una informazione perfetta circa le possibili alternative d'azione e le sue conseguenze, nonché di una capacità cognitiva adeguata sia a valutare le diverse opzioni mettendole a confronto tra loro sia a individuare una relazione che le leghi alle rispettive conseguenze, sia a ordinare tali conseguenze dalla migliore alla peggiore, sia, infine, a selezionare l'azione (i mezzi) migliore sulla base di un calcolo razionale teso a massimizzare i "vantaggi" o "benefici" e a minimizzare gli "svantaggi" o "costi" della propria azione (Oliverio 2015). Questo modello si rifà a quello che era diventato un proprio e vero "paradigma della razionalità" nelle scienze sociali. Partendo da queste premesse allora fu possibile sviluppare una teoria matematica in grado di individuare quale scelta avrebbe compiuto l'individuo agendo razionalmente. La "teoria dell'utilità attesa" formulata da John von Neumann e Oskar Morgenstern fornisce un quadro normativo nel quale l'individuo, partendo dalle sue preferenze, fra le varie alternative sceglie quella che massimizza la sua utilità attesa ottenuta moltiplicando l'utilità della stessa per la sua probabilità di occorrenza. La teoria assume il principio della coerenza del comportamento dell'individuo e della transitività delle preferenze.

Questo modello di razionalità che costituì il paradigma dell'economia neoclassica, venne applicato alle scienze sociali a opera di economisti come Kenneth Arrow, Mancur Olson, James Buchanan, Gary Becker per descrivere ed in una certa misura prevedere le scelte individuali e collettive. Cambiando le variabili in gioco ed utilizzando strumenti matematici come la teoria dei giochi, gli alberi decisionali, il calcolo probabilistico si sono potuti analizzare numerosi campi dell'agire umano: dalla decisione di sposarsi a quella di interrompere una gravidanza.

Nonostante la vasta applicazione della teoria normativa dell'agire umano in campo sociale e della sua appropriatezza nell'analisi dei comportamenti, a partire dal XX secolo il paradigma della razionalità inizia ad entrare in crisi. Uno dei primi limiti che vengono evidenziati è quello circa l'informazione; l'individuo empiricamente non gode mai di una informazione completa, ovvero di tutta la rosa di alternative fra cui scegliere, ma l'informazione è sempre parziale. L'ottica meramente economica della massimizzazione dell'utilità, secondo calcoli di costi e benefici viene criticata sottolineando il ruolo giocato dalle emozioni, dal contesto e dal comportamento degli altri. Di fronte a questi limiti la critica si è divisa in due: una parte contro qualsiasi forma di modello razionale considerando la razionalità un'astrazione e l'altra sostenitrice della razionalità di fondo dell'agire umano se pur depurata dalle influenze emotive e sociali.

La psicologia cognitiva ha fornito risultati empirici importanti nel mostrare come la razionalità effettivamente non sia alla base del comportamento umano. Esempio è l'esperimento condotto nel 1953 da Maurice Allais. Ad un gruppo di persone vengono sottoposte due alternative, nella prima si può scegliere tra due opzioni A e B, la seconda tra C e D. Scegliendo A si ottiene una vincita sicura di 1000€ scegliendo B si ha una probabilità del 10% di vincere 5000€ una probabilità del 89% di vincerne 1000 e l'1% di probabilità di non vincere niente. Nel secondo caso scegliendo C si ha una probabilità dell'11% di vincere 1000€ e l'89% di vincere 0€, con l'opzione D si ha il 10% di probabilità di vincere 5000€ e il 90% di non vincere nulla. Per la teoria dell'utilità attesa se A è preferito a B allora C è preferito a D. Nel primo caso si ha la certezza di una vincita sicura di 1000€ mentre con l'opzione B si rischierebbe (1% di probabilità) di non vincere niente; nel secondo caso la probabilità di guadagnare 1000€ a fronte di 5000€ è più probabile dell'1%. I risultati empirici di questo esperimento hanno dimostrato che nel primo caso veniva data più rilevanza alla probabilità, quindi A è stata l'opzione più scelta nel secondo invece è stata scelta l'opzione D grazie al valore molto più alto della ricompensa a discapito della minore probabilità di accadimento. Questa anomalia di comportamento è conosciuta come "paradosso di Allais" e dimostra come venga violato l'assioma dell'indipendenza delle decisioni concetto fondamentale della teoria dell'utilità attesa: gli individui agiscono a seconda della situazione e non perseguendo il principio della massimizzazione dell'utilità.

Una serrata critica economica deriva dall'economista austriaco Friedrich von Hayek insistendo sempre sul concetto di informazione perfetta. Secondo il principio della dispersione della conoscenza, non esiste una mente centrale pianificatrice in grado di padroneggiare tutte le informazioni e dati economici, bensì la conoscenza è dispersa, ovvero distribuita tra tutti gli individui in un contesto sociale. Le informazioni a disposizione degli individui sono tuttavia parziali e portano come diretta conseguenza alla fallibilità dell'agente stesso nelle sue decisioni; punto cardine nella sua visione sociale fondata sulla libertà dell'individuo che darà poi ispirazione alla concezione della "Società aperta" popperiana.

Interessante è la proposta originale da parte dell'economista e psicologo Herbert Simon sul concetto di razionalità normativa. Considerati i limiti della teoria in termini di variabili quali abitudini, norme, valori Simon ne propone una che spieghi l'agire dell'individuo fondata sul concetto di "razionalità limitata". Essendo impossibile individuare la scelta ottima per l'individuo, a causa della limitatezza delle informazioni a disposizione e per mancanza di potenza di calcolo nell'elaborare tutte le alternative possibili in termini di vantaggi o svantaggi, egli propende per soluzioni che ritiene "soddisfacenti". Questa soluzione permette di integrare al "paradigma della razionalità" dei limiti che nonostante la arginino lasciano valido il ragionamento razionale a fronte di soluzioni che si possono definire "ragionevoli" e non ottimali.

Dal punto di vista della psicologia cognitiva, forte dei riscontri empirici ottenuti attraverso un gran numero di esperimenti, il modello dell'homo oeconomicus e del paradigma della razionalità vengono respinti in toto. A causa delle capacità cognitive limitate e dell'alto grado di influenzabilità dell'essere umano nel processo di "decision making" gli assunti della razionalità appaiono svuotati di significato in quanto sistematicamente violati sperimentalmente. In opposizione alla corrente comportamentista, la psicologia cognitiva nasce con la sostanziale differenza di voler studiare i meccanismi mentali, i processi cognitivi che guidano le decisioni umane. Il punto di partenza è quindi lo studio di ciò che avviene dentro la mente umana e non come il comportamentismo, limitato allo studio delle azioni esteriori degli individui.

Uno dei limiti al quale l'individuo è soggetto nella fase decisionale è rappresentato dal contesto, ossia di come viene formulata la domanda all'individuo. Gli psicologi cognitivi Tversky e Kahneman dimostrarono questo condizionamento in un esperimento chiamato "problema della malattia asiatica" (Kahneman, Tversky, 1981) strutturato come segue: al primo gruppo di persone viene chiesto di scegliere tra due piani di azione per contrastare l'arrivo di una epidemia straniera, l'opzione A prevede che 200 persone si salvino, con l'opzione B c'è un terzo di probabilità di salvare tutte le 600 persone e due terzi di non salvarne alcuna; al secondo gruppo viene presentato il programma C con il quale muoiono 400 persone e il programma D dove c'è un terzo di probabilità che nessuno muoia e due terzi che muoiano tutti. Dal punto di vista numerico le risposte A e C, e B e D sono identiche ma le risposte dei gruppi presentano differenze significative: il programma A è stato scelto dal 72% del primo gruppo mentre nel secondo gruppo sempre il 72% ha scelto il programma D. Alla luce dell'uguaglianza contenutistica dei programmi è impossibile spiegare con il modello razionale un "rovesciamento delle preferenze" se non con gli effetti del "frame" ovvero della presentazione del problema decisionale. L'esperimento dimostra come il processo decisionale sia viziato da "euristiche" ovvero delle scorciatoie mentali che vengono percorse permettendo una trattazione dell'informazione più veloce. Nonostante siano spesso usati nella quotidianità, possono portare alla commissione di errori sistematici chiamati "bias".

Come appare evidente dai risultati delle ricerche sulla razionalità non si può investigare la realtà sociale partendo dall'assunto della perfetta razionalità dell'attore.

### CAPITOLO 3 Limiti dei Big data

“Le rivoluzioni in scienza sono state spesso precedute da rivoluzioni nella misurazione”

Sinan Aral (citato da Cukier, 2010)

Dopo aver visto con che tipo di razionalità agiscono le persone, quindi chiedendoci che modello di comportamento adottare per prevedere le loro decisioni passiamo ad analizzare più approfonditamente la tecnologia dei big data.

I Big Data rappresentano non solo un nuovo e potente strumento, ma nell’ottica del metodo scientifico, anche un cambio di paradigma<sup>1</sup>. Vengono considerati una nuova forma di empirismo che segnerebbe “la fine della teoria”. Non sarebbe più necessario partire da delle ipotesi, verificarle sperimentalmente, raccogliere dati e formulare una teoria, ma partire proprio dai dati stessi: sui dataset a disposizione, l’intelligenza artificiale applica una serie di algoritmi, fino a trovare quello che meglio descrive il comportamento del fenomeno. Questo processo rappresenta una rivoluzione copernicana nel metodo scientifico. Sono i dati stessi elaborati con algoritmi a mostrare la loro struttura, senza l’utilizzo della teoria. Come sottolinea Anderson 2008 nel suo provocatorio articolo “the end of the theory”, gli strumenti concettuali di ipotesi teorie e leggi diventano obsoleti e rappresentano solo un’ulteriore complicazione mentale di fronte a dati, che se ben analizzati matematicamente mostrano le relazioni presenti. Usando le parole di Prensky (2009): “Gli scienziati non devono più fare ipotesi plausibili, costruire modelli, e testarli con esperimenti basati su dati e simulazioni. Invece, possono estrarre il set completo di dati dalle ricorrenze che rivelano effetti, producendo conclusioni scientifiche senza una ulteriore sperimentazione”.

In termini scientifici si sta affermando che è sufficiente la “correlazione” tra dati, senza la ricerca di una “causazione”.

Se nel mondo scientifico i Big data porteranno ad un cambiamento di paradigma in senso Kuhniano è da vedere, per quanto riguarda le scienze sociali però si possono trarre alcune considerazioni.

L’applicazione di questo nuovo empirismo rappresenta per una parte degli scienziati il trionfo positivista, quel progetto di controllo e predizione sociale grazie alla quantità e soprattutto alla qualità di informazioni a disposizione che non ha precedenti nella storia umana. Tuttavia bisogna evidenziare alcuni limiti come quello sulla rappresentatività del campione.

---

1. Rob Kitchin *Big Data, new epistemologies and paradigm shifts* 2014

<http://bds.sagepub.com/content/1/1/2053951714528481>

Nonostante sempre più persone usino il web e internet si deve tener conto di quali siano queste persone e soprattutto del fatto che rappresentano solo una parte di tutta la società. Analizzando i dati del campione appare evidente una sovrarappresentazione di fasce di popolazione come i giovani e una della popolazione maschile<sup>2</sup>.

Riguardando le previsioni possibili grazie ai big data ci accorgiamo che queste sono realmente affidabili solo per le rilevazioni “fisiche” ovvero quelle che prevedono gli spostamenti delle persone in termini di spazio e di tempo (con la tecnologia GPS) mentre non sono affidabili per quanto riguarda ciò che la gente pensa e come questo possa influenzare le variabili sociali, nella fattispecie del Twitter mood, quella economica del Dow Jones Industrial Average. Nonostante la scoperta della correlazione tra lo stato d’animo di calma e il movimento dell’indice finanziario, questo fatto non ci dice niente sulla “causazione” dell’uno sull’altro. In questo senso sono molte le correlazioni tra variabili completamente svincolate da una relazione causale vedi le cosiddette correlazioni spurie<sup>3</sup>.

Anche considerando una moltitudine di informazioni e dati, piccole variazioni del fenomeno non misurabili in fase osservativa possono portare a grandi conseguenze a livello macroscopico. Nel passaggio tratto dal libro “Caos” di James Gleick viene descritto il momento in cui si scopre nel campo meteorologico il concetto di “effetto farfalla”.

“La ragione di questo stato di cose era l’effetto farfalla. Per piccoli fenomeni meteorologici, e per chi faccia previsioni su scala mondiale, piccolo può significare tempeste e tormento, ogni previsione si deteriora rapidamente. Errori e incertezze si moltiplicano, diffondendosi a cascata attraverso una catena di elementi di turbolenza, da turbini di polvere e temporali improvvisi fino a vortici continentali, osservabili solo da satelliti. I moderni modelli meteorologici lavorano con una griglia di punti distanziati fra loro di un centinaio di chilometri, e anche a questa scala, alcuni dati di partenza devono essere congetturati. Per comprendere le difficoltà di previsione nelle scienze sociali è utile un paragone con la meteorologia. Ma supponiamo che la Terra potesse essere coperta da sensori separati di una trentina centimetri uno dall’altro, che salissero a intervalli di trenta centimetri per l’intera altezza dell’atmosfera. Supponiamo che ogni sensore fornisse letture perfettamente esatte di temperatura, pressione, umidità e di ogni altra quantità desiderabile per un meteorologo.

A mezzogiorno in punto un computer infinitamente potente prenderebbe tutti i dati e calcolerebbe gli sviluppi in un ciascun punto. Per le 12.01, per le 12.03, e via dicendo.

---

2. <http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx>

3. <http://www.tylervigen.com/spurious-correlations>

Il computer sarebbe non di meno incapace di predire se un giorno, a un mese di distanza, Princeton, nel New Jersey, avrà sole o pioggia.

A mezzogiorno gli spazi fra i sensori nasconderanno fluttuazioni di cui il computer sarà all'oscuro, piccole deviazioni dalla media.

Alle 12.01 quelle fluttuazioni avranno già creato piccoli errori a qualche decina di centimetri di distanza.

Ben presto gli errori si saranno moltiplicati alla scala di qualche metro, e così via sino a giungere alla scala dell'intero pianeta”.

Questa efficace similitudine metereologica, mostra che addirittura in una scienza fisica naturale, piccole quantità trascurate possono generare a livello macroscopico notevoli cambiamenti. Considerando la misurazione sociale ancora più complessa, in quanto ci si basa su indicatori poco precisi per definizione appare evidente l'impossibilità della predizione sociale, il sogno positivista.

Conseguenze sociologiche sono state individuate e analizzate da Friedrich von Hayek. Rimanendo con la similitudine metereologica il sociologo austriaco combatteva aspramente contro qualsiasi tipo di potere centralizzato che si arroga il diritto di pianificare la vita sociale. Dal punto di vista economico ha già dimostrato come il progetto totalitario di questa intelligenza calcolatrice era destinato a fallire per il teorema della dispersione della conoscenza, al pari di un super computer dotato di una straordinaria potenza di calcolo e avente a disposizione tutte le informazioni. È impossibile prevedere il corso delle azioni in un contesto di scambio continuo nella società a causa del numero infinito di conseguenze derivanti da quella azione. Oltrepassando la volontà e quindi il calcolo razionale dell'azione dell'individuo nella società, a causa dei continui scambi e interazione (il mercato), si propagano le cosiddette conseguenze inintenzionali. Esse portano alla creazione di istituzioni regolanti il vivere comune, grazie alla loro nascita dalla “pratica” sociale con un processo continuo di prove ed errori e non frutto di un'intelligenza centrale pianificatrice<sup>1</sup>.

Proprio come le piccole fluttuazioni non rilevabili tra un sensore e l'altro nell'atmosfera provocano infiniti cambiamenti, così le infinite (in termini matematici) interazioni sociali producono cambiamenti nella società impossibili da prevedere.

Concludendo l'uomo risulta essere qualcosa di più della semplice somma delle sue azioni, la sua razionalità è limitata e le conseguenze delle sue azioni difficilmente prevedibili.

---

1. Friedrich von Hayek *La società libera* trad. it. Vallecchi, Firenze 1969

Inoltre, parallelismi efficaci per comprendere come nessuna macchina sarà mai in grado di comprendere concetti come il significato e il senso vengono proposti i seguenti esempi.

Marche (2012) sostiene che gli artefatti culturali come per esempio la letteratura non possono essere trattati meramente come dati. Uno scritto non è semplicemente un ordine di lettere e parole; è un significato contestuale e convenzionale gli algoritmi non sono in grado di catturare e decifrare il significato o il contesto e sostiene Marche trattano: tutta la letteratura come se fosse la stessa...l'analisi algoritmica dei romanzi e degli articoli dei giornali è necessariamente al limite del riduzionismo. Il processo di trasformare la letteratura in dati rimuove la specificità. Rimuove il gusto. Rimuove tutti i raffinamenti della critica. Rimuove la storia dell'accoglienza del lavoro.

Jenkins (2013) quindi conclude: il valore dell'arte, la qualità di una pièce teatrale o di un dipinto, non è misurabile. Potresti mettere ogni tipo di dati nella macchina: dati, colori, immagini, ricezione al box office, e niente di tutto ciò potrebbe spiegare di che opera d'arte si tratta, cosa significa e perché è così potente. Tutto ciò richiede L'Uomo, non la macchina.

## CONCLUSIONI

I Big Data stanno rivoluzionando le scienze sociali ed epistemologiche. Le prospettive sono ampie tanto da portare a parlare di un cambio di paradigma nel metodo scientifico, dalla causazione alla correlazione. Se sono evidenti i successi del nowcasting come riportato negli esempi del secondo capitolo, la loro applicazione rimane limitata: per un motivo tecnologico, l'elaborazione di dati complessi e un motivo strutturale, l'imprevedibilità del comportamento umano e delle sue conseguenze. Nonostante il numero crescente di informazioni non sarà mai possibile prevedere quali decisioni prenderà una persona, le variabili sono troppe e l'essere umano non è un decisore coerente: è pieno di contraddizioni, la strada per la decisione è costellata di errori. L'utilità dei Big data è innegabile ma tutte queste montagne di dati analizzati quantitativamente dovranno essere sempre interpretati dall'uomo: un compito che ancora non può essere sostituito dalla macchina<sup>1</sup>. Un cambiamento che porta a ripensare il ruolo dello scienziato sociale, ovvero quello di diventare l'interprete dei dati e trovare una relazione causale tra variabili oltre la semplice correlazione mostrata dalla macchina.

---

1. Bauman, Zygmunt *“Non c'è spiegazione senza teoria il fattore umano sarà sempre necessario”*  
Staglianò Riccardo *La Repubblica* 8/09/2013 <https://blog.wired.it/arabafenice/2013/09/09/big-data-e-scienze-sociali-interviene-zygmunt-bauman.html>

## Bibliografia

- Antiseri D. *Epistemologia e didattica delle scienze*, Armando editore, Roma, 2000
- Cukier K. *Data, data everywhere* The Economist, 25/02/2010
- Develin K. Lorden G. *The numbers behind numbers. Solving crime with mathematics* Plume 2007
- Durkheim E. *Le regole del metodo sociologico* Ed. di Comunità, Milano. 1979
- Durkheim E. *Il suicidio*, UTET, Torino, 1969
- Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. *Detecting influenza epidemics using search engine query data*. Nature 457:1012–10155. 2009
- Giorello G. *Popper e la filosofia della scienza* La biblioteca di Repubblica Roma 2011
- Gleick J. *Caos. La nascita di una nuova scienza* trad. Libero Sosio 2012
- Hayek F. *La società libera* trad. it. Vallecchi, Firenze 1969
- Hayek F. *La via della schiavitù* Trad. D.Antiseri R. De Mucci Rubbettino 2011
- Hayek F. *The fatal conceit* www Bartley III 1991
- Hayek F. *Legge legislazione e libertà. Critica dell'economia pianificata* Trad. P.G. Monateri il saggiatore 2010
- Infantino L. *Ordine senza piano. Le ragioni dell'individualismo metodologico* armando editore 2008
- Kahneman D., Tversky A. *The framing of decisions and the psychology of choice* Science New series Volume 211 Issue 4481 30/01/1981 pp. 452-458
- Laplace P. *Simon de Saggio filosofico delle probabilità* 1812
- Oliverio A. *Individuo, natura, società* Mondadori Milano 2015
- Peruzzi A. *Modelli della spiegazione scientifica* Firenze university press 2009
- Popper Karl R. *Logica della scoperta scientifica* trad. Mario Trincherò Piccola biblioteca Einaudi 2010
- Popper Karl R. *La società aperta e i suoi nemici* trad R.Pavetto Armando Editore 2002
- Popper Karl R. *Miseria dello storicismo* trad C. Montaleone Feltrinelli 2013
- Rudder C. *Dataclism: Who We Are (When We Think No One's Looking)* Crown Pub 2014
- Stewart Ian *Dio gioca a dadi?* Bollati Boringhieri Torino, 2010

- Weber Max *L'etica protestante e lo spirito del capitalismo* trad. A.M. Marietti BUR biblioteca univ. Rizzoli 1991
- Weber Max *Economia e società. Comunità* Donzelli 2005

## Sitografia

- Allais Maurice *Le Comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine* 1953  
<https://wolfweb.unr.edu/homepage/pingle/Teaching/BADM%20791/Week%206%20Decision%20Making%20Under%20Uncertainty/Allais%20Paradox%201953.pdf>
- Bańbura M. Giannone D. Reichlin L. *Nowcasting* Working paper series n.1275 2010  
<https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1275.pdf?ba24acc83cc23e6e268f6be4069ceb46>
- Bauman, Z. “Non c'è spiegazione senza teoria il fattore umano sarà sempre necessario” Staglianò Riccardo *La Repubblica* 8/09/2013  
<https://blog.wired.it/arabafenice/2013/09/09/big-data-e-scienze-sociali-interviene-zygmunt-bauman.html>
- Bollen J, Mao H, Zeng X, *Twitter Mood predicts the Stock Market* 2015  
[https://www.researchgate.net/publication/47405468\\_Twitter\\_Mood\\_Predicts\\_the\\_Stock\\_Market](https://www.researchgate.net/publication/47405468_Twitter_Mood_Predicts_the_Stock_Market)
- Giannotti F. *Big data e social mining: i dati a saperli ascoltare raccontano storie* 2014  
[http://www.sobigdata.it/sites/default/files/BigData%26SocialMining\\_Giannotti.pdf](http://www.sobigdata.it/sites/default/files/BigData%26SocialMining_Giannotti.pdf)
- Huberman B. *Predicting the Future With Social Media* 2014  
[https://www.researchgate.net/publication/45909086\\_Predicting\\_the\\_Future\\_with\\_Social\\_Media](https://www.researchgate.net/publication/45909086_Predicting_the_Future_with_Social_Media)
- IBM: *Big data, bigger outcomes* (PDF) <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html> [www.ibm.com](http://www.ibm.com)
- Jenkins T. *Don't count on big data for answers*. The Scotsman, 2013.  
<http://www.scotsman.com/the-scotsman/opinion/comment/tiffany-jenkins-don-t-count-on-big-data-for-answers-1-2785890>
- Kitchin R. *Big Data, new epistemologies and paradigm shifts* 2014  
<http://bds.sagepub.com/content/1/1/2053951714528481>
- Marche S. *Literature is not data: Against digital humanities*. Los Angeles Review of Books 2012 <http://lareviewofbooks.org/article.php?id¼1040&fulltxt>
- Prensky M. *H. sapiens digital: From digital immigrants and digital natives to digital wisdom* 2009 <http://www.innovateonline.info/index.php?view=article&id=705>

## ABSTRACT

Social sciences aim to explain and predict social phenomena. In epistemological terms social sciences use natural science as a model differing on the studied object: the man. Positivist ambition of predicting the future seems to be realized thanks to Big Data, a huge quantity of data and information available on the Internet shared by millions of users around the world. Data constitute a crucial part in the scientific discovery. In order to be scientifically valid, a theory must follow a scientific method, in which empirical data collected through experiments validate or invalidate the starting hypothesis.

Over the centuries the meaning of the word “scientific” has changed. Starting with the Galilean method, constituted of hypothesis, experiment and theory, confirmations of the starting hypothesis were searched through experiments. A radical change of philosophical thought was introduced by Karl Popper who criticized the confirmation process by supporting the “falsification approach”. Whereas the inductive thought looks for confirmations which are only temporary and generalize the results, the deductive method is logically correct and it allows to generate valid conclusions. Over the course of history the concept of scientific knowledge changed according to the historic and social context. Starting from the 18th century the influence of the physical-natural science on the society structure began to be evident. Newton physical discoveries revolutionized the scientific knowledge, by summarizing in short and simple equations which gave the illusion to predict the movement of every object in the universe. The predictive euphoria was soon transferred to the human behavior and the society. The first applications reflects the positivist thought in which there is the belief that is possible to predict the human behavior as it was a physical object. However the scientific discoveries of the last century marked the end of the deterministic vision introducing probabilistic and statistical concepts.

In the social science field it is always difficult to find a clear cause-effect relation between the studied phenomena, so correlation appears a valid instrument. For instance in his original study (*The protestant ethic and the spirit of capitalism*) Max Weber finds a correlation between the protestant values of the population analyzed and how they brought to a primitive form of capitalist society. The German scholar realized an important synthesis between two different epistemological schools of thought which clashed in Germany at the end of the 19th century in the debate known as “Methodenstreit”. The two positions differed each other on the way the social science should use natural science as a model: the historicist school sustained the uniqueness and specificity of the social phenomenon and so arguing that it is impossible to find regularities and laws, on the other hand the positivist school proposed a transposition of the scientific method to the social field.

In the modern philosophy of science a widely accepted model is the “deductive-nomological”. It consists in a reasoning which is based on a nomological part, with the presence of a law and starting conditions (explanans), and a deductive one because the phenomenon to be explained (explanandum) is logically deduced from the starting conditions.

This model can be successfully used for natural and social sciences both.

Once the vital part played by the data has been stressed, the recent technological revolutions in terms of huge quantities of data have led to wonder how this will affect social science. In terms of hardware great progresses have been achieved in the last years: regardless the size a bigger amount of digital data can be stored. Simultaneously the production of data and the sharing of information have hugely increased in fact millions of data are shared everyday on the Internet. However these data are represented in different format and it makes difficult for modern computer to elaborate them effectively. These data are called “Big Data” and they are mainly characterized by three factors: volume, variety and velocity.

The applications of the big data studies on the social phenomena begin to show how it is possible to predict very short term events. These predictions are called “Nowcasting”, literally “predict the present”. For instance the study “Twitter mood predicts the stock market” shows the correlation between the mood of people and a specific “Dow Jones Average” trend. Given the amount of personal information that people freely share on the Internet the perspectives to predict their behaviour seem to be wide.

To understand human behavior it is necessary analysing their decision making method. Starting from the “homo oeconomicus” model the paradigm of rationality has established in the social sciences. It is based in three principles: the optimal choice of the rational actor, an order of preference and a perfect information available. In the last century the cognitive psychology stressed the limits of this theory showing how in the empirical reality the principles of the theory are always violated as the Allais’ paradox the Asiatic disease problem show. Considering the critics to the paradigm of rationality two different school of thought appeared: the first one promoted to abandon totally the rational view because of the clear limits, the second accepting the limits continued believing in the basically rational structure. A radical shift occurred in the rationality field, starting from a normative conception the evolution of the research brought to focus on a more descriptive approach.

If it is impossible to take as assumption the rational behavior of the human beings it seems difficult for a Big Data technology based computer to predict the human decisions. Nevertheless the quantity of data available can benefit the social scientist work by redefining it: the computers show the correlation and the scientist personally interpretes the results and find cause-effect relations where it is possible.