

Il computer come macroscopio

Big Data e approccio computazionale per comprendere i cambiamenti sociali e culturali.

1. Contesto, Computer e macroscopio

1.1. Un mondo globale di artefatti digitali con conseguenze sociali

Nella società contemporanea le persone interagiscono attraverso le piattaforme di social media, cercano informazioni avvalendosi di algoritmi, mentre stanno nascendo una serie di applicazioni il cui scopo è creare una forma economica nuova in cui la tecnologia funge da strumento abilitatore fra chi offre un servizio e chi ha bisogno di un servizio.

Tutto questo accade mentre le persone si spostano nella propria città o all'interno dei confini globali resi accessibili dalla rete delle infrastrutture di trasporto, mantenendosi costantemente in contatto con la rete mediante dispositivi mobili.

Un mondo così complesso, in cui si mescola la componente sociale, dimensione geografica e opportunità permessa dalla tecnologia, ha bisogno di nuovi strumenti concettuali ma anche operativi per poter comprendere le sfide della contemporaneità.

Sociologia, economia, antropologia, psicologia sociale nascono dalla crisi del XVIII secolo, intesa come rottura con la società precedente.

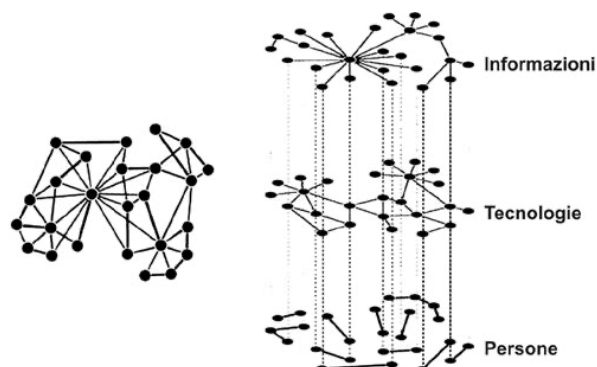
Questa società ha visto una serie di cambiamenti importanti: l'ascesa di nuove classi sociali (borghesia), l'emergere di nuove forme economiche (il capitalismo) basate su strategie di produzione diverse (l'industrialismo), che per scopi politici ed economici avevano bisogno di comprendere le popolazioni altre (il colonialismo) e i cambiamenti dell'individuo e dei suoi rapporti con il mondo.

Tutto ciò necessitava di una conoscenza diversa del mondo disposta a sviluppare tecniche di ricerca empirica e teorie sulla modernità. Per questo il XIX secolo è il periodo d'oro delle scienze sociali, perché il mondo stava diventando troppo complesso per non tentare di comprenderlo e provare a governarlo.

es. il tema della globalizzazione: questo settore di ricerca ha richiesto una serie di nuovi strumenti concettuali per comprendere una serie di processi difficilmente spiegabili con altri strumenti, come l'emergere di aziende economiche transnazionali (le multinazionali), movimenti globali di resistenza allo *status quo* economico, istituzioni economiche sovranazionali (es. WTO, World Trade Organization), le caratteristiche della cultura contemporaneamente globale e locale.

Se la globalizzazione ha messo al centro della riflessione delle scienze sociali lo spazio fisico e le sue trasformazioni, il mondo contemporaneo sta mettendo al centro della propria riflessione la tecnologia. È la tecnologia che è responsabile di una nuova forma di complessità con cui le scienze sociali sono costrette a fare i conti.

La società di oggi è composta da tre diversi *layers*, ognuno con delle proprie regole e caratteristiche:



1. Persone: le persone interagiscono con altre secondo delle dinamiche precise che ormai fanno parte del bagaglio culturale della nostra società. Processo caratterizzante: sociale.
2. Tecnologie digitali: intese come artefatti che memorizzano, elaborano, scambiano informazioni. Hanno un doppio ruolo di strumenti abilitanti (consentono alle persone di fare qualcosa che senza di esse sarebbe difficile o impossibile fare. es. comunicare a distanza) e artefatti vincolanti (hanno delle caratteristiche che alcuni comportamenti possono essere limitati a causa del loro stesso utilizzo. es. la necessità di una rete dati per poter accedere a Internet). Le tecnologie per loro natura interagiscono con altre tecnologie per poter svolgere il loro compito. Sono elementi di un'infrastruttura. Processo caratterizzante: essere oggetti fisici con caratteristiche digitali.
3. Informazioni: per certi versi l'interazione sociale può essere ricondotta a uno scambio di specifiche informazioni mediante artefatti, in specifici contesti con delle specifiche conseguenze. Le persone si scambiano informazioni e vi accedono (es. un numero nella rubrica). Le informazioni possono essere anche prodotte dagli stessi artefatti. Processo caratterizzante: frutto dell'interazione umana/scambio d'informazioni tecnologiche.

Questo schema è dinamico perché ogni *layer* impatta sugli altri, in una catena di feedback che sarebbe difficile da interpretare se non la si scomponesse nei suoi elementi.

es. un gruppo di persone decide di darsi un appuntamento in un luogo preciso su whatsapp. (lo strato delle persone).

Ogni livello ha una funzione specifica e ha delle conseguenze sugli altri, in una catena di effetti che ha delle conseguenze sociali. Il livello di azione di questo processo non è solo locale, ma può essere tranquillamente esteso a tutto il mondo.

In questo senso possiamo proporre l'ipotesi che la società contemporanea può essere concettualizzata come un mondo globale di artefatti digitali con conseguenze sociali.

E per capire questa società serve un macroscopio.

1.2 Studiare la società attraverso un macroscopio:

Il macroscopio è quello strumento che permette di vedere i processi di grandi dimensioni che accadono nella società contemporanea e questo strumento assume le forme di un computer, inteso come l'artefatto che consente di utilizzare il software. Perché il computer può essere considerato un macroscopio?

- Strumento per affrontare dimensione di scala dei fenomeni sotto analisi:
La grande quantità di dati da affrontare. C'è bisogno di tanti dati per studiare il mondo contemporaneo, sia che siano fenomeni locali, macroregionali o globali, la loro dimensione è tale che i dati che fanno riferimento a questi fenomeni devono essere trattati con strumenti informatici, perché altrimenti la loro gestione potrebbe essere problematica. Prima si lavorava a campione.
- Computer come laboratorio: per sviluppare delle analisi:
 - a. Per poter analizzare i dati è necessaria una formalizzazione che permetta di capire i fenomeni a cui fanno riferimento (l'uso della statistica, la rappresentazione dei dati in forma di grafo, l'applicazione di tecniche informatiche) è necessario l'utilizzo del computer sia perché è possibile manipolare le variabili nel nostro tentativo di comprendere il fenomeno, sia perché il computer è l'unico strumento in grado di gestire gli enormi database frutto della raccolta dei dati.
 - b. Il confronto tra Anobii e Flickr: hanno raccolto lo scambio di messaggi fra gli utenti delle due piattaforme. I messaggi raccolti sono stati divisi in tre gruppi: *status sociale* che facevano

seguito ad azioni come il like o il follow; *supporto sociale* dove come dare il benvenuto; *scambio di conoscenza* per condividere esperienze personali. Hanno sviluppato un algoritmo in grado di classificare automaticamente i messaggi secondo i tre gruppi e si sono concentrati su come nel corso dello scambio di messaggi, questi potessero cambiare da un gruppo all'altro. Su Anobii è più presente lo status sociale, mentre su Flickr è presente una gamma più ampia di tipologie di messaggi. Un fattore chiave è come evolvono i messaggi nel tempo. Le conversazioni lunghe mostrano un mix fra scambio di conoscenza e sostegno sociale. Questo studio mostra due cose: da un lato la quantità di informazioni che è possibile raccogliere con queste metodologie di ricerca, dall'altro come il computer diventa uno strumento indispensabile per analizzare i dati e ottenere risultati interessanti.

- Computer come strumento per trovare gli schemi ricorrenti: negli ultimi anni per esigenze di analisi e per necessità di comunicare i dati in modo non banale, si sono sviluppate altre forme di rappresentazione delle informazioni testuali e quantitative. *Tagcloud* e *Streamgraph* sono dalle piattaforme di social media, le prime servono per visualizzare a colpo d'occhio le etichette di testo (tag) più ricorrenti, le seconde permettono di visualizzare l'andamento di più variabili rispetto al tempo. Esistono diversi approcci alla rappresentazione delle informazioni, da un'impostazione più interna alla tradizione statistica-matematica, ad approcci più vicini al design e alla comunicazione. La tendenza a rappresentare graficamente enormi quantità di dati ha dato vita al settore → *visual analytics*: è un modo per studiare i dati che li analizza solo quando vengono rappresentati graficamente; negli ultimi tempi si è sposata con il *visual design*.

Ci sono però degli altri motivi che hanno portato a una crescita dell'uso delle tecnologie informatiche nelle scienze sociali, che sono appannaggio di un contesto sociale e culturale completamente cambiato, un contesto in cui gli elementi chiave sono: i big data, i social media e l'ascesa della big social science.

1.3 Big Data: un nuovo mercato che riguarda gli scienziati sociali:

Big Data: si fa riferimento a uno specifico settore industriale il cui ambito è lo sviluppo di tecnologie per la raccolta, organizzazione, analisi di grandi quantità di dati. Ovviamente data l'onnipervasività delle tecnologie informatiche, questo settore ha conseguenze sociali, politiche, economiche, e quindi anche un impatto su temi come privacy e sorveglianza.

Nel 2011 era usato per indicare un settore del mercato dell'informatica che mira alla gestione di database di informazioni dalle dimensioni enormi. Descritto in questo modo, appare evidente il motivo per cui quasi tutti i documenti che affrontano il tema dei big data, sono report di mercato dei principali player del mondo delle ICT, i quali sono interessati a definire questo nuovo settore tecnologico, ma vogliono dichiarare ai propri clienti che sono essi stessi fornitori di soluzioni tecnologiche relative ai big data.

Dato che non ha una definizione rigorosa, quella più utilizzata è quella delle tre V di Gartner:

1. Volume: grande quantità di dati; le tecnologie di questo settore sono legate a un'enorme quantità di dati. Inizialmente la scienza che acquisiva la maggior quantità di dati era la fisica; oggi i social producono tantissimi dati. Di solito quando si parla di big data non si danno delle dimensioni effettive dei database perché è facile che ciò che si intende per enorme oggi, possa essere la normalità domani.
2. Velocità: tempi rapidi di raccolta dati; altrettanto rapidamente devono essere archiviati e analizzati. Uno dei problemi classici della progettazione dei database è il tempo di accesso ai dati. Nel caso dei big data, poiché la quantità di informazioni archiviate è effettivamente

enorme, è necessario approcciare al problema dell'accesso in maniera completamente diversa e con delle tecnologie che siano scalabili linearmente.

3. Varietà: eterogeneità. La crescita dell'industria delle ICT e l'aumento di complessità del mercato, ha fatto sì che i database si trovino ad archiviare dati molto diversi tra loro. Questa situazione rende più complicata la gestione del database che deve essere in grado di trattare dati molto diversi tra di loro. Tre tipologie di dati:
 - a. Dati strutturati: struttura univoca, esplicano direttamente i dati che stanno conservando es. codice fiscale.
 - b. Dati semistrutturati: non hanno una struttura rigida ma seguono uno schema. es. HTML
 - c. Dati non strutturati: non hanno una forma. es. testo libero.

La teoria di database vuole due tipologie:

- a. Testuali: usati per la ricerca d'informazione di testi;
- b. Contestuali: raccolta d'informazione d'immagini, video ecc.

La loro caratteristica in Big Data è che gestiscono i dati in maniera eterogenea, gestiscono le due tipologie di dati.

Con il passare del tempo sono state associate altre due variabili ideate da Intel:

4. Valore: capacità di fornire analisi utili e interessanti in grado di prevedere eventi e processi futuri attraverso strumenti di computazione avanzata come il machine learning, i modelli statistici e gli algoritmi basati su grafi.
5. Veridicità: i dati devono anche avere una certa affidabilità quando il loro utilizzo è alla base di decisioni delicate e quindi non devono essere errati, rovinati o avere altre caratteristiche che ne potrebbero limitare l'uso.

Uno dei motivi chiave che rendono interessante l'universo dei big data anche per le scienze sociali è l'idea della *predittività*, ovvero la possibilità di prevedere eventi grazie all'enorme quantità di informazioni che vengono raccolte all'interno dei database.

I big data, grazie alla raccolta di una quantità sterminata di informazioni e mediante l'applicazione di sofisticate tecniche di analisi sarebbero in grado di estrapolare informazioni rispetto al futuro con una certa affidabilità.

L'articolo di Chris Anderson, all'epoca editor di Wired, in cui dichiarava la "fine della teoria" in quanto il diluvio di dati reso possibile dai big data avrebbe reso obsoleto il metodo scientifico tradizionalmente inteso. L'argomentazione si basa sull'idea che la possibilità di scoprire un gran numero di correlazioni fra i dati avrebbe reso possibile la scoperta di processi nascosti tra i dati.

La riflessione di Anderson sembra trovare conferma in Google Flu: un servizio che serve per prevedere l'andamento delle epidemie di influenza stagionale usando i milioni di ricerche che le persone fanno tutti i giorni.

L'idea è che quando una persona sospetta di avere l'influenza, cerca su Google informazioni sul tema. Non tutti coloro che cercano queste informazioni sono malati, ma aggregando questi dati è possibile trovare una correlazione fra l'effettivo diffondersi dell'epidemia influenzale e il numero di ricerche fatte. È stato per molto tempo un classico case study per mostrare la forza predittiva dei big data, grazie anche alla sua capacità di anticipare di 10 giorni l'epidemia influenzale rispetto alle fonti tradizionali.

In realtà a ben vedere Google Flu si basa su una teoria del comportamento umano in situazioni di incertezza informativa. Esiste tutta una letteratura che ritiene che per quanto enormi, i dati non possano parlare da soli, hanno sempre bisogno di una teoria che guidi sia il criterio di costruzione del database. Non basta avere una correlazione fra due eventi per dire che l'uno è causa dell'altro.

Tutte le tecnologie quando appaiono nel mercato vengono incorporate in un contesto sociale con una narrazione che ne giustifica l'uso. Hanno bisogno di un'ideologia che le legittimi. Solitamente un prodotto viene lanciato anche se non è completamente ottimizzato; trovando una nicchia di mercato viene perfezionato.

Una tecnologia non si sviluppa in un contesto sociale solo per motivi ingegneristici.

Con i big data hai così tanti dati da poter prevedere i processi che ti interessano. es. azienda che può prevedere i dati di mercato.

In sociologia è impossibile prevedere il comportamento umano.

Un ulteriore elemento che rende i big data interessanti per le scienze sociali è che spesso le domande conoscitive sollevate da chi lavora con queste tecnologie sono domande tipiche delle scienze sociali. Con un'iperbole potremmo dire che i big data sono un settore latente delle scienze sociali. Questa affermazione porta con sé due conseguenze.

1. Le imprese che si confrontano con i big data si pongono domande non banali sul funzionamento della società e sul proprio impatto sociale, per quanto mediato dall'esigenza di avere informazioni sul mercato. La necessità di conoscere i comportamenti dei propri clienti, le indagini di mercato, le strategie di condivisione dei contenuti ecc. sono tutte domande del nuovo marketing contemporaneo basato sull'analisi dei dati.

Queste domande portano anche questioni interessanti sulla società contemporanea: quali sono gli stili di vita emergenti, qual è il rapporto fra consumo e società, come cambiano le strategie di consumo delle informazioni.

Temi come identità, comportamenti sociali, atteggiamento culturale possono essere desunti anche dalle dinamiche di consumo; pertanto, i big data, quando utilizzati dalle imprese commerciali vengono usati per rispondere a domande che interessano anche il sociologo, il politologo, l'economista, lo psicologo sociale e perfino l'antropologo.

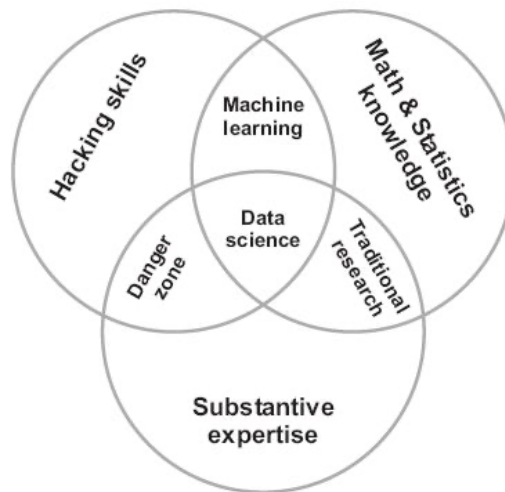
2. I professionisti di questo settore lavorano su temi che sono problemi sociali espressi in termini tecnologici. Uno dei cambiamenti più interessanti legati allo sviluppo dei big data è la nascita di una nuova figura professionale: il data scientist o lo scienziato dei dati.

Anche per il data scientist non esiste una definizione rigorosa (negli anni '90 le persone che facevano questo lavoro si chiamavano *data miner*). Il termine comincia a circolare a partire dal 2008, quando DJ Patil di LinkedIn e Jeff Hammerbacher di Facebook cominciarono a usare questo termine per descrivere il proprio lavoro.

Il data scientist un ricercatore con competenze in informatica e matematica o statistica che analizza i dataset tipici dei big data.

Drew Conway, data scientist presso il Project Florida, ha usato un diagramma di Venn per indicare l'intersezione di tre competenze chiave che sono:

1. Capacità di hacking (inteso come manipolazione dei dataset e per pensare "per algoritmi").
2. Conoscenza matematica e statistica (per la creazione e la ricerca di modelli).
3. Conoscenza del settore nel quale si vogliono applicare queste competenze (*substantive expertise*, competenza nel merito).



Se il data scientist è uno studioso che usando strumenti informatici e matematici lavora con dataset relativi a processi sociali, è facile desumere che le domande che si porrà sono da scienziato sociale. Qual è il modello di diffusione dei memi dentro Facebook? Come si diffonde una notizia su Twitter? Non sono domande da informatico, ma sono domande da studioso di fenomeni sociali che deve usare strumenti informatici e big data per poter rispondere.

Se per rispondere a una domanda bisogna rispondere a varie domande sociali, il lavoro di data scientist è un mix di lavoro sociale, statistico e matematico.

- Social informatics: domande da cripto-sociologo.

Se una persona si chiede quali sono i pattern sociali dei database che sta analizzando, non si sta facendo una domanda da informatico ma da scienziato sociale.

1.4 Social Media: le più grandi fonti di dati di comportamenti sociali

Da quando è apparso Internet come tecnologia di comunicazione fino al suo boom grazie allo sviluppo del World Wide Web, è stato chiaro che ci sarebbero stati profondi cambiamenti nello studio delle relazioni sociali mediate: Internet si presentava come un mondo nuovo, ricco di processi interessanti che poneva questioni non banali su alcune delle acquisizioni delle scienze sociali e degli studi sui media.

Con la quotidianizzazione dei social media, sono aumentate le domande teoriche, sono incrementate le ricerche empiriche ed è aumentato in modo esponenziale la quantità di persone che usano le piattaforme, nonché la disponibilità di dati per la ricerca.

Lo studio di internet può essere distinto in due paradigmi:

1. Internet come spazio esterno → (ci devi entrare):

Il paradigma è stato identificato tra la fine '80 e i primi 2000; Internet si presenta come un luogo da esplorare che si colloca al di fuori delle persone e dello studioso.

Questa percezione è anche dovuta ai limiti delle tecnologie di accesso alla rete di questo periodo. L'esperienza di navigazione in rete è paragonabile al superare una soglia, ed è per questo motivo che molti ricercatori dell'epoca hanno usato la dicotomia online/offline per studiare la rete.

Questa dicotomia è stata rafforzata dall'orizzonte antropologico con cui veniva vista la rete: l'idea è quella della frattura antropologica, l'umanità in Internet è un'umanità diversa, in cui c'è la prevalenza dell'anonimato, le interazioni sociali sono limitate e sono prevalentemente basate sul testo. Gli oggetti sociali che vengono studiati sono le community che pur assumendo diverse configurazioni tecnologiche sono basate su due principi cardine:

- a. l'interazione testuale
- b. la partecipazione attiva.

L'impostazione metodologica delle ricerche è quella di considerare la rete come territorio, un nuovo luogo sociale che ha bisogno di essere cartografato.

Quando Internet comincia a diventare fenomeno di massa, grazie anche alla diffusione del Web 2.0 nasce il paradigma:

2. Internet come spazio interno (sei già dentro)

Copre il periodo dal 2004 fino a oggi.

La rete è un luogo che deve essere compreso e si trova intorno a noi. Anche in questo caso l'esperienza d'uso della rete non fa che rafforzare questa percezione. L'uso della rete è data da tecnologie mobili, wi-fi e uso sistematico delle app; per questo motivo la dicotomia online/offline è sostituita dal concetto di *always on*, sempre connessi e sempre dentro la rete.

L'orizzonte antropologico di riferimento è quello della continuità: i processi sociali che avvengono in rete non sono necessariamente nuovi, ma sono essenzialmente diversi poiché hanno una scala molto grande e sono resi possibili dal ruolo abilitante della tecnologia.

Gli oggetti sociali che vengono studiati sono i social network sites, ovvero siti che consentono la creazione di contatti sociali attraverso una pagina Web. A guidare le interazioni nei social network, è anche l'uso performativo dei contenuti multimediali ma soprattutto l'uso indicale che i social network sites mettono a disposizione attraverso una funzione tecnica che permette di esprimere apprezzamento o interesse.

L'impostazione metodologica di questi studi è che la rete è una mappa dei fenomeni sociali. Dato che l'uso dei social media riguarda ogni aspetto della vita quotidiana, ciò che accade in rete per certi versi accade anche fuori dalla rete e viceversa. Il rapporto fra l'evento che accade nel mondo e l'evento che accade su Internet è un rapporto di mappa/territorio

Es. la pratica dei selfie:

Il ricercatore potrebbe porsi essenzialmente due domande di ricerca:

1. Se esiste una sorta di canone nello scattarsi una foto e poi diffonderla nei social media, che magari dipende dal desiderio che la foto raccolga il maggior numero di like e circoli in maniera virale all'interno di Internet. In questo caso la domanda è da paradigma di Internet come spazio esterno, poiché si stanno cercando le regole sociali e tecnologiche della pratica del selfie.
2. Se la pratica dei selfie è sintomatica di qualcos'altro, come uno specifico modello culturale come il narcisismo o semplicemente la voglia di raccontare sé stessi all'interno del flusso di immagini che invadono ogni giorno i social media. Le questioni che ci poniamo in questo modo sono da paradigma di Internet come spazio interno, in cui la pratica digitale deve essere ben compresa perché rimanda a un altro processo culturale più ampio.

Da quanto fin qui detto, i social media sono un ambito di ricerca piuttosto interessante e molto affascinante, sia perché hanno avuto una diffusione esponenziale negli ultimi anni, sia perché consentono ai ricercatori di raccogliere enormi quantità di dati.

Quello della (relativa) facilità della raccolta dati sui social media è un tema delicato, in quanto esistono diverse strategie di raccolta dati:

1. Accesso tramite API (Application Programming Interface): è un protocollo per accedere a una serie di informazioni messe a disposizione dalle stesse piattaforme di social media. Il vantaggio è la possibilità di costruire strumenti *ad hoc* per procedere alla raccolta di informazioni secondo le esigenze della propria ricerca; lo svantaggio è che si è alla mercé delle piattaforme che decidono a quali dati si possa accedere e in che modo.
2. Servizi sviluppati da terze parti: esiste un universo di applicazioni di *social media analytics*, ovvero strumenti che consentono di raccogliere dati ed elaborare semplici metriche per poter

effettuare delle analisi. Alcune volte questi servizi sono gratuiti o open source. Questi servizi, dato che incorporano nel loro codice le istruzioni delle API, hanno gli stessi limiti che queste impongono loro.

Altre volte questi servizi sono a pagamento; in questo caso non hanno limiti attribuibili alle API, perché spesso hanno accordi commerciali con le singole piattaforme. Questi servizi sono *freemium*, ovvero in parte gratuiti (*free*) e in parte a pagamento (*premium*). In questo modo il cliente potenziale può usare la versione gratuita per sperimentare lo strumento e quando avrà bisogno di performance migliori, potrà decidere di passare alla versione a pagamento. L'unico problema di utilizzare questi servizi per scopi di ricerca sociale, è che sono pensati per rispondere a domande di marketing e non di conoscenza scientifica.

3. *Data reseller*: società da cui è possibile acquistare dataset presi dai social media per poterli utilizzare per i propri scopi. Solitamente ogni società, pur offrendo servizi standard, di solito un servizio di database in tempo reale e un servizio di raccolta dati storici, è specializzata in uno specifico settore. Per esempio Gnip o Datasift.

Uno dei social media che meglio ha saputo sfruttare il proprio ruolo di fornitore di dataset da big data è Twitter: ha un ecosistema di offerta di dati completo che comprende API, servizi di terze parti, data reseller, partner certificati.

Il flusso dati costante relativo a tutti i tweet che vengono inviati nel Web ha anche un suo nome preciso: *firehose*. Il fatto che Twitter fornisca i propri dati in maniera così diversa ha fatto sì che sia il social media più studiato con approcci computazionali, ovvero con approcci che usano algoritmi o altre tecniche di analisi dei dati di tipo informatico.

La conseguenza è stata che Twitter è stato considerato come un organismo modello per i big data originati dai social media.

1.5 Big Social Science: la scienza sociale fatta in grande

Politica della ricerca:

- Il frame commerciale dei big data, il potere di riferimento è militare, politico ed economico; devi avere la possibilità di comprare i dati.
- Microsoft research labs, yahoo! Labs, google research, facebook data science
- Il rischio dell'assenza di ricerca su ambiti non commerciali
- Ambiti che mescolano domande di ricerca e domande commerciali (*healthcare*)
- Comunità spontanee di appassionati (*quantified self*)
- Attivisti dei diritti digitali (movimento open data) per dati usati in una ricerca pubblica e finanziati pubblicamente.

Fisica delle particelle, programmi di ricerca spaziali, progetto Genoma umano, sono tutti esempi di quella che viene chiamata *big science*, ovvero progetti che hanno bisogno di ingenti finanziamenti.

Big science: è stato utilizzato Derek de Solla Price per descrivere lo sforzo economico e organizzativo alla base del progetto *Manhattan*.

Nella sociologia della scienza, la riflessione sul concetto di *big science* va di pari passo con il concetto di collettivizzazione della scienza, ovvero un processo di industrializzazione del sapere scientifico in cui i progetti di ricerca sono sempre meno appannaggio di singoli scienziati e sempre più di pertinenza di grandi comunità di ricercatori che per i loro studi gravitano intorno al raggiungimento di uno specifico obiettivo di ricerca.

L'uso sistematico dei *big data* nelle scienze sociali e la relativa necessità di strumenti informatici per il loro studio proietterà tali discipline nel settore della *big science*, contribuendo a creare una *big social science*.

Esiste una letteratura scientifica sulle caratteristiche storiche della *big science* e sulle proprietà sociologiche della collettivizzazione della ricerca scientifica. Riteniamo interessante sottolineare questo aspetto perché permette di spiegare alcuni limiti, problemi e opportunità di queste discipline. Tre fattori della collettivizzazione della scienza che sono: le attrezzature, la cooperazione, la politica della ricerca:

1. Attrezzature: la ricerca contemporanea è caratterizzata dalla necessità di una specifica dotazione tecnologica che renda il ricercatore in grado di svolgere le proprie attività di ricerca come la raccolta e l'analisi dei dati. Quasi tutti i settori più all'avanguardia hanno bisogno di specifiche attrezzature che alcune volte assumono la forma di strumentazioni tecnologiche sofisticate, altre volte assumono la forma di vere e proprie cattedrali della ricerca scientifica. Per le scienze sociali: programmi software specializzati nella raccolta dei dati dai social media, se si vuole lavorare con un sistema completo di dati o bisogna avere i fondi necessari per acquistare dai *data reseller*, oppure bisogna lavorare all'interno delle grandi organizzazioni produttrici di dati. Questo scenario ha una profonda conseguenza sulla praticabilità delle ricerche delle scienze sociali e non tutte positive.

Se da un lato c'è la possibilità di collaborare con le compagnie del mondo dei social media dall'altro c'è il rischio che si crei un nuovo divario digitale fra i ricchi di *big data* e i poveri.

Un *data scientist* in Facebook ha un accesso ai dati che nessun altro ricercatore può avere: pertanto l'unico modo per aggirare questo limite è quello di procedere a una divisione del lavoro fra ricercatori sociali e professionisti dei big data delle grandi piattaforme digitali o compagnie.

Questa situazione porta alla cooperazione → da chi progetta la raccolta dati; chi si occupa concretamente della raccolta; chi prepara i dati attraverso l'organizzazione del database; a chi applica modelli di analisi o altri strumenti. I *big data* costringono gli studiosi di scienze sociali a collaborare con gli esperti di informatica e viceversa, in quanto le competenze per utilizzare tali apparecchiature sono piuttosto specifiche.

Una componente che rende necessaria la collaborazione fra scienziati sociali e informatici: l'ambiguità dell'interpretazione dei dati.

Quando vengono applicati modelli computazionali per l'analisi dei dati, è importante distinguere fra:

- Il fenomeno che si sta analizzando.
- Il modello che viene a essere applicato.

Dato che questa strategia è recente non è facile una corretta interpretazione dei risultati → necessaria una collaborazione multidisciplinare:

- Bisogna applicare modelli di analisi sostanzialmente nuovi.
- Bisogna distinguere quali sono le caratteristiche del processo sociale analizzato una volta espunte le proprietà formali.

es. → Nel 2010 chi scrive, assieme ad altri autori ha svolto una ricerca sulle caratteristiche della circolazione delle notizie in Twitter. La metodologia adottata è stata quella di concentrarsi sulle notizie che avessero la caratteristica di essere *trending topic* di Twitter e poi focalizzarsi sugli utenti con molti follower e da questi caricare il grafo completo.

La rappresentazione della quantità di tweet/retweet rispetto al tempo ci ha permesso subito di distinguere le notizie in due categorie:

- notizie esogene → si propagavano di Twitter ma circolavano nei mass media; notizie come la crisi economica in Grecia o l'eruzione del vulcano islandese Eyjafjallajökull
- notizie endogene → si diffondevano su Twitter e poi passavano nei mass media, contribuendo all'incremento della circolazione della notizia; notizie come l'ultima puntata di Lost o il fracasso

delle vuvuzela, le famigerate trombette protagoniste dei mondiali di Calcio in Sudafrica nel 2010.

Le curve di diffusione dei *trending topic* analizzati mostravano delle similitudini; il contributo dei retweet alla diffusione delle news era molto diverso per le due tipologie di notizie. Paragonando questi comportamenti a modelli sociali abbiamo ipotizzato che le notizie esogene si comportano secondo le caratteristiche dell'agenda setting; le notizie endogene sembravano rispettare le caratteristiche degli eventi televisivi → attivano le conversazioni online in maniera spontanea, attivando delle specifiche *fan culture* fino ad attirare l'attenzione dei media mainstream.

C'è stato un'interessante divisione del lavoro fra la componente e l'interpretazione sociologica in cui l'una per avere senso aveva bisogno dell'altra, e viceversa. Questa nuova forma di collaborazione fra scienziati sociali e informatici sta dando vita a una nuova cultura scientifica e accademica.

Nuovi strumenti di ricerca e un numero importante di ricercatori chiedono che ci siano dei fondi dedicati per questi studi; pertanto, è necessaria una politica della ricerca che legittimi l'investimento tale che la ricerca delle scienze sociali nell'ambito dei big data sia economicamente sostenibile.

Svantaggi: il *frame* attraverso cui si giustifica l'investimento in questo settore sia quello commerciale: i big data sono importanti per mantenere alto il livello di competitività delle multinazionali e consentono di creare nuovi prodotti o servizi semplicemente attraverso una analisi non banale dei dati prodotti internamente da un'azienda → di conseguenza gli studiosi lavorano per grandi aziende come: Microsoft Research Labs, Yahoo! Labs, Google Research, Facebook data science.

Se i big data vengono applicati solo in ambiti di business, c'è il rischio che gli ambiti non immediatamente legati al business possano essere dimenticati o non abbiano il dovuto interesse da parte dei soggetti coinvolti; esistono dei modi per aggirare questi limiti:

La nascita di specializzazioni dei big data che rispondono contemporaneamente a domande di ricerca e domande commerciali → es. settori come la salute o la gestione della città si stanno configurando come settori specializzati sui big data in cui la differenza fra domande di ricerca e domande a interesse commerciale sfumano le une rispetto alle altre.

In aggiunta, stanno emergendo comunità spontanee di appassionati dati che grazie alle opportunità consentite loro da tecnologie specializzate diventano produttori e consumatori di dati: *quantified self*, appassionati che attraverso tecnologie commerciali raccolgono dati relativi alle proprie performance sportive o semplicemente ai loro dati biometrici per motivi di fitness o per stare bene e creano una sorta di resistenza soft, per cercare di avere la massima libertà possibile nella gestione di questi dati.

Nonostante la proficua collaborazione ricerca pubblica/ricerca privata, resta necessario uno spazio in cui i big data possano essere utilizzati al di là del loro valore commerciale e del loro essere proprietà privata. In questo senso il movimento open data può essere un buon *trade off* tra offerta di dati proprietà dei dati. La necessità di una infrastruttura di ricerca pubblica che produca dati a disposizione degli scienziati sociali è un'opzione che nel prossimo futuro sarà difficilmente rinviabile.

Le piattaforme di *e-social science*, ovvero l'uso di approcci computazionali avanzati nelle scienze sociali, sono nate proprio quando è aumentata la ricerca di dati, ovvero agli inizi del 2000. Progetti come NCeSS (National Centre for e-Social Science, UK) hanno lo scopo di mettere a disposizione degli scienziati sociali, una serie di strumenti che servono per soddisfare le mutate esigenze nell'*e-social science*.

È necessaria una politica della ricerca o di collaborazione con le imprese private in questo senso, anche perché l'incremento dei dati nelle scienze sociali sta facendo nascere non solo un modo diverso di fare ricerca, ma anche settori completamente nuovi i cui primi risultati sono al momento molto affascinanti.

2. Paradigmi. I programmi di ricerca computazionali.

2.1 Dati, modelli e algoritmi: alla ricerca di una forma

Dati: servono per creare modelli; la modellistica è la possibilità di utilizzare dati per spiegare dinamiche complesse; algoritmi: quando un modello può essere fatto girare sul computer (deve avere delle caratteristiche precise). Il legame tra i tre è che tutti i dati possono creare modelli ma questi ultimi non possono sempre creare algoritmi.

Dati: cosa sono? In maniera estremamente semplificata:

- Frutto di una specifica rilevazione, frutto di una costruzione metodologica. Bisogna avere un protocollo per identificarlo e questo dà vita a un'informazione strutturata.
- Qualunque tipo di informazione strutturata inseribile in un database.
- La costruzione del dato: essendo figlio di un criterio di rilevazione e strutturazione → sono costruiti, non ha nulla di oggettivo.
- sono rilevati in base alle esigenze; (ovvero registrati e inseriti in un database, una raccolta di dati) secondo le specifiche dello studioso e della ricerca che si sta effettuando es. → età: è una convenzione, solitamente si risponde con l'ultimo compleanno (un adulto risponderà con un numero intero; un bambino anche con un decimale per una percezione diversa). Se si vuole calcolare il reddito serve solo l'anno.
- non tutti i dati numerici sono numerali: sono solo dei modi per organizzare un'idea; "etichette numeriche": es. → età: 48 e 24 anni: numeri; es. "quanto ti piace x da 1 a 10": numerale.
- i dati possono essere anche testi: la sociologia è spaccata in due approcci di ricerca →
 - a. *approccio quantitativo*: lavorano con numeri usano la statistica, rilevano le informazioni con il questionario ed estendono i risultati raggiunti alla popolazione di cui il campione è un sottoinsieme rappresentativo; molto esteso con i dati ma standardizzato (metafora del letto di Procruste);
 - b. *approccio qualitativo*: lavorano con i testi, non usano la statistica, rilevano le informazioni con le interviste e i risultati raggiunti restano limitati al gruppo di persone studiate dato che non esiste una procedura formale per considerarli rappresentativi di una popolazione più ampia; (studio di Propp sulle favole russe).

La questione oggi è molto più sfumata. La riflessione metodologica ha mostrato somiglianze di ricerca con numeri e testi; anche i testi usano quantificazioni statistiche (es. linguistica computazionale, statistica linguistica e lessicografia); la centralità delle strategie di rilevazione: l'uso del computer come strumento di organizzazione di informazioni in database, fa sì che il problema venga centrato sulle strategie di rilevazione e non sulla natura dei dati, poiché i computer riescono a gestire dati molto diversi tra loro → conseguenza della diffusione dei big data: il processo di datizzazione (*datafication*).

- I dati sono relazioni: es. soggetto A con B → I primi a considerare le relazioni come tipologie di dati, sono stati gli antropologi. I legami di parentela sono stati considerati come un ottimo grimaldello concettuale per capire le culture "altre", fino ad arrivare alle riflessioni dell'antropologia strutturale che con i diagrammi di parentela ha provato a elaborare una rappresentazione grafica delle relazioni di uno specifico gruppo sociale. Bisognerà aspettare la *social network analysis* per poter trasformare questi dati in strumenti di analisi quantitativa.
- Importanza dei metadati: dati sui dati, ovvero informazioni che servono per descrivere dei dati. Se opportunamente interrogati possono dare informazioni che possono arricchire la conoscenza nei risultati di ricerca. es. selfie → raccogliere una serie di autoscatti che girano nei social media → classificazione delle fotografie per cercare di identificare se ci sono dei

pattern. Da queste foto si possono ricavare le info: Exif (Exchangeable image file format): data, ora, marca e modello della fotocamera/dispositivo mobile, coordinate geografiche ecc.

Teoria dei giochi → si crea una matrice di decisioni; a ognuna di queste viene aggiunto un valore (positivo o negativo). Sono formalizzazioni dove i numeri sono traduzioni semiotiche di altre cose (es. scelta).

Trasformare un'informazione in dato è un passaggio semiotico. Con una semiotica (sistema di trasformazione di senso) si possono usare i dati come qualunque cosa; questi possono essere analizzati una seconda volta e ottenere dei risultati diversi).

Modello: in cosa consiste? Non è possibile sviscerare la questione di che cosa sia un modello, possiamo però descrivere il modello rispetto agli obiettivi della nostra argomentazione. Contiene l'autonomia delle variabili e vede come interagiscono tra di loro:

- Descrizione analogica di tipo schematico di un processo delle scienze umane o sociali: è una rappresentazione semplificata della realtà, pertanto per essere tale il modello deve descrivere un processo facendo riferimento a un'analogia o a una metafora. es. Modello fisico all'università di Glasgow del funzionamento dei mercati ispirato alla "Teoria dei sentimenti morali" di Smith: rappresentazione con un sistema ad ingranaggi dell'orologio. Sposti una leva e si ripercuote nel sistema. Analogico: hai bisogno di una semiotica che traduca la teoria in modello comprensibile: es. scacchi dove viene usato il linguaggio bellico.
- a. Descrizione analogica: isomorfismo: ha la stessa forma pur comportandosi in maniera diversa. Questa descrizione deve essere:
- b. Schematico: descritto un linguaggio astratto: schematico in questo senso non vuol dire semplificato; è necessario un linguaggio sufficientemente astratto che esprima l'essenza delle componenti in gioco in un modello e come si comportano le une rispetto alle altre. (matematica, linguaggio naturale, linguaggio *markup* UML): il caso del modello a invarianza di scala; fenomeni gaussiani e fenomeni *power law*; il collegamento preferenziale.

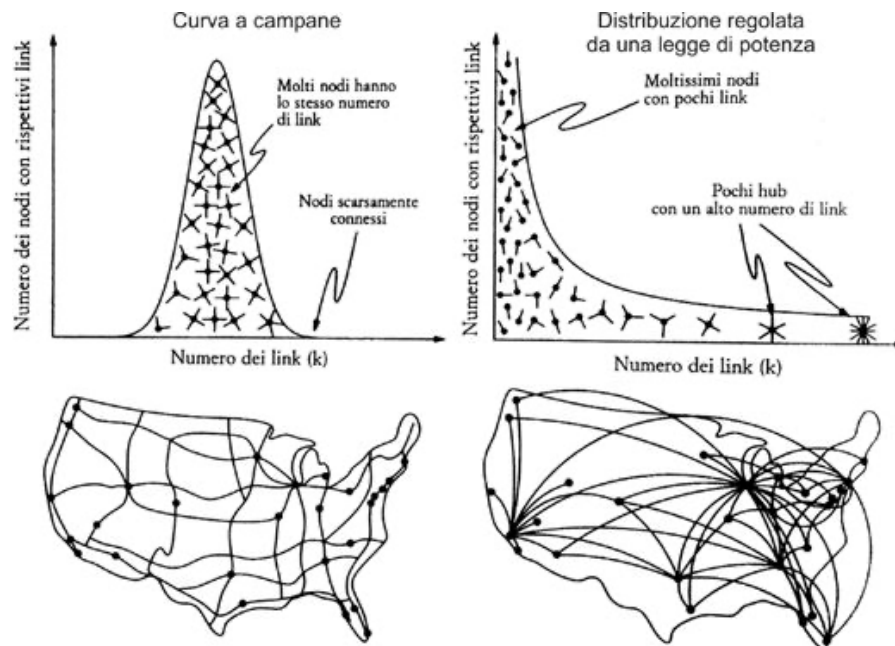
Anche il linguaggio naturale può essere usato per descrivere processi, usando elaborazioni schematiche come classificazione e organizzazione. es. → Propp e il suo studio delle fiabe russe dal quale ricava 21 modelli ricorrenti.

Esistono poi linguaggi grafici in grado di descrivere processi come l'UML (Unified Modelling Language), il cui scopo è senza dubbio legato alle esigenze della programmazione informatica, ma che si prestano anche per la descrizione di processi complessi.

Due modi per fare i modelli:

1. Inventare un modello arbitrario: più faticoso.
 - Il caso del modello a invarianza di scala: il più importante modello di teoria contemporaneo:
 - *Modello della curva di Gauss* → la curva descrive il seguente fenomeno: in una distribuzione qualsiasi, i valori della distribuzione tenderanno ad addensarsi intorno a un valore medio, mentre tutti gli altri valori tanto più si allontanano dal valore medio, tanto più saranno sempre meno probabili.
 - *Fenomeni gaussiani e fenomeni power law* → sono alcune particolari distribuzioni dove esistono pochi valori molto rappresentati e molti valori poco rappresentati → distribuzione condizionata es. *ranking* delle parole in un *corpus*.
 - *Collegamento preferenziale o "preferential attachment"*: quando arriva un nodo in una rete preferisce connettersi con i nodi con tante connessioni piuttosto che poche. Quando analizziamo un fenomeno, se questo può essere descritto in termini rete e se mostra una distribuzione da legge di potenza, siamo in presenza di una rete a invarianza di scala. → processo di attaccamento preferenziale (effetto san Matteo). es. andare in un pub affollato e non in uno vuoto. È arrivato a questo ragionamento studiando la struttura di internet; se una

pagina ha molti link le pagine nuove che nascono avranno la tendenza a collegarsi con le prime, altrimenti ci sarà una distribuzione casuale.



La curva che descrive la distribuzione è identica alla curva della legge di potenza. Questo vuol dire che sono lo stesso fenomeno? Assolutamente no: la semantica che spiega perché la curva di distribuzione è simile (ma diversa) è specifica del fenomeno studiato. Pertanto, il modello di invarianza di scala è un buon modello perché la sua schematizzazione matematica aiuta a capire altri settori.

Fenomeni diversi descritti dalla stessa distribuzione:

<u>Tipo di distribuzione</u>	<u>Criterio alla base della distribuzione</u>	<u>Tipo di fenomeno</u>
Legge di potenza	Collegamento preferenziale	Reti a invarianza di scala (internet)
Legge di Zipf	Principio del minimo sforzo	Uso delle parole in un corpus di testi
Distribuzione di Pareto	Principio di Pareto 80/20	Relazione fra cause ed effetti
(non presente)	Effetto San Matteo	Meccanismo di distribuzione delle ricompense (ricercatori)

2. Si cerca un modello precedente a cui rimandarlo.

Una volta definiti i modelli, è necessario che questi modelli siano computazionali, ovvero che possano essere resi in modo da poter essere trattati dal computer. In pratica dai modelli bisogna passare agli algoritmi.

Algoritmo: cos'è?

Solitamente la definizione è di una procedura generale, una sequenza di passi il cui scopo è la soluzione di un problema specifico o una classe di problemi. (nei libri d'informatica viene fatta la metafora con una ricetta culinaria). Questa procedura gode di alcune proprietà:

- **Effettività:** la procedura deve essere eseguibile.

- Finitezza di espressione: la lista delle operazioni da svolgere deve essere finita. Inizia con lo step 1 e finisce con n : deve concludersi altrimenti entra in un *loop* infinito.
- Finitezza di calcolo: esplicita le condizioni per cui un algoritmo deve fermarsi.
- Determinismo: ogni passo dell'esecuzione deve essere seguito da una e una sola operazione: passo 1 → passo 2 ecc. non deve esserci spazio alla casualità.

Riassumendo → un algoritmo è essenzialmente una procedura computazionale che serve per raggiungere uno specifico risultato; una procedura che può essere eseguita da un computer.

Consentono di applicare dei modelli di analisi a enormi database in cui i dati sono frutto di diverse fonti. La successione dati → modelli → algoritmi: questo passaggio può essere interpretato come:

1. Indicazione metodologica: è il modo di lavorare quando dai dati vogliamo passare a procedure computerizzate. Dati → applicare modelli → vedere se sono computerizzabili.
2. Esigenza epistemologica: i modelli altro non sono che schemi di ragionamenti organizzati in modo tale che siano trattati dai computer, ovvero algoritmi.

I modelli devono essere algoritmi?

Le scienze cognitive si sono chieste se il computer potesse essere un buon modello del funzionamento del cervello umano → Sì.

Di contro, le scienze sociali si sono via via chieste se l'algoritmo potesse essere considerato una teoria o un modello → sì, soprattutto per l'impostazione che considera il computer un ottimo strumento per simulare processi sociali reali. Questo non vuol dire che i modelli delle scienze sociali devono essere degli algoritmi: l'algoritmo può anche essere un aiuto per districarsi nella selva dei dati che si hanno a disposizione.

Sta di fatto che per questa nuova generazione di scienze sociali il legame tra dati/modelli/algoritmi è tale che il loro utilizzo congiunto non può essere considerato circostanziale ma strutturale.

2.2 Programmi di ricerca computazionali nelle scienze sociali:

Il computer nelle scienze sociali c'è da tantissimo tempo. La flessibilità che permette da quando sono nati i linguaggi di programmazione è recente.

Computer e scienze sociali: strumento o ambiente? Se il computer è un semplice strumento, allora le scienze sociali non sono cambiate di molto da quando alla fine del XIX secolo studiavano la società di massa; se il computer è una componente fondamentale delle scienze sociali contemporanee, allora ci stiamo avvicinando a una prospettiva molto più vicina ai problemi sollevati dal XXI secolo.

- Strumento per raccogliere i dati → Strumento.
- Strumento per analizzare i dati → Ambiente.
- Modello di processi altrimenti non studiabili → Ambiente.

Quando il computer è entrato a far parte in maniera sistematica di alcuni settori delle scienze sociali ha portato una serie di innovazioni che non sono solo di ordine metodologico. Negli ultimi tempi il rapporto tra "computer" (software) e le scienze sociali è passato da computer come strumento per semplificare delle procedure a un ambiente: elaborare ragionamenti in maniera computazionale.

Per questa ragione è necessaria una mappa delle scienze sociali che hanno messo al centro del loro progetto il computer. Da un lato perché molte di quelle idee sono state scoperte o riscoperte in tempi recenti, quando computer e ricerca scientifica sono diventati un binomio indissolubile. Dall'altro perché in questo modo è possibile avere un quadro ampio di ciò che sta succedendo nelle scienze dell'uomo e della società. Attualmente la disciplina con un esplicito richiamo all'uso del computer è la

scienza sociale computazionale (*computational social science*), ma la sua istituzionalizzazione è di gran lunga più recente delle idee e delle ricerche a cui fa riferimento.

Complessità socio-tecnologica e approccio computazionale:

- Schemi esplicativi di un mondo sempre più complesso: le variabili da analizzare sono sempre di più; più aumentano meno è controllabile il processo esplicativo. es. battito d'ali di una farfalla in Giappone → uragano in America.

Necessità di una mappa delle scienze sociali computer-friendly

- Molte delle idee classiche sono state riscoperte in tempi recenti: c'è sempre stato un sottoinsieme degli studiosi di scienze sociali che hanno usato il computer per i loro studi anche se erano minoritari.
- Avere un quadro ampio di come stanno cambiando le scienze sociali.

La scienza sociale computazionale (*computational social science*)

- Istituzionalizzazione recente a fronte di idee e ricerche "antiche".
- Le quattro aree di ricerca della CSS (Cioffi-Revilla 2014)
 - a. Estrazione di informazione sociale automatizzata → possibilità di fare analisi dei testi conversazionali.
 - b. Reti sociali → *network analysis*.
 - c. Complessità sociale → usare modelli computazionali per spiegare variabili legate tra di loro.
 - d. Simulazione sociale → simulo dei processi sociali all'interno del computer per intervenire sulle variabili es. tempo, rapporto popolazione nativa/immigrata durante una pandemia.

Non discipline ma programmi di ricerca.

- Imre Lakatos → filosofo che ha dato una soluzione al problema della logica della scoperta scientifica e sviluppo storico dei problemi scientifici.
La filosofia della scienza degli anni '60/'70: genesi delle teorie scientifiche; stavano diventando sempre diverse es. teoria astrofisica → non basata su esperimenti; altre si basavano su modelli esplicativi. Nascono due scuole:
 - a. Falsificazionisti → la forma scientifica è la più precisa rispetto ad altre forme di conoscenza. Popper è il maggiore esponente; è un sostenitore della logica della scoperta scientifica. Le teorie scientifiche possono essere verificate per quanti non riescono a spiegare → falsificazionismo: se sai quando NON puoi applicare una teoria scientifica: la teoria è pienamente scientifica; se si può applicare in tutti i casi: non è scientifica. Il limite è che questa posizione è astratta, non ha un rapporto con l'effettivo procedere della ricerca.
 - b. Convenzionalisti → Kuhn: concetto di paradigma → (forme flesse dentro un'espressione verbale) quando studi una teoria scientifica ne impari tutto; gli esperimenti non possono essere *in crucis*. Dove si diceva una questione logica → qui è comunitaria. Convenzionalismo: La correttezza di una teoria scientifica viene decisa dalla comunità di pratica: la comunità usa la teoria per spiegare più fenomeni in quel quadro di riferimento.
- Né Popper né Kuhn, né falsificazioni né convenzionalismo con Lakatos:
- Regole metodologiche in una comunità scientifica: euristiche negative, euristiche positive → ci sono degli assetti dentro la conoscenza scientifica non dimostrabili. I postulati danno vita a una spiegazione teorica divisa in:
 - Euristiche positive: proteggono i formulati;

- Euristiche negative: danno le spiegazioni;
Applichi partendo dalla negativa → positiva → postulato



- Le tre caratteristiche dei programmi di ricerca:
 - Obiettivo teorico ed empirico
 - Metodologia condivisa
 - Successione di teorie che definiscono l'avanzamento scientifico, mantenendo inalterato il postulato di riferimento.
- Programmi di ricerca che sono entrati a far parte della CSS (*computational social science*):
 - Sociologia analitica.
 - Scienze delle reti: come si è evoluta la network science.
 - Simulazione sociale.
 - Memetica: applicare problemi epidemiologici alla diffusione di modelli.
 - Cliometria o cliodinamica: branca della storia che usa la cliomedia per spiegare problemi storici.
 - Economia comportamentale: comportamenti economici spiegati in base psicologica.
 - *Digital humanities* e *culturomics*: applicabilità dei *big data* all'analisi dei fenomeni culturali.

Per esempio, se seguiamo una delle argomentazioni più recenti su questo settore interdisciplinare, vediamo che le principali aree di ricerca sono quattro: l'estrazione di informazione sociale automatizzata (un campo a metà fra il text mining e l'analisi informatica del contenuto), le reti sociali, la complessità sociale e la simulazione sociale (Cioffi-Revilla 2014: 12-17). Tutte queste aree di ricerca hanno solide basi storiche in diversi ambiti di

ricerca: dalla teoria generale dei sistemi (Bertalanffy 1968), alla ricerca sui media con strumenti informatici (in particolare la semantica quantitativa: Lasswell, Leites 1949), la social network analysis (Forsyth, Katz 1946), la simulazione al computer di processi politici (Sola Pool, Abelson 1961).

Una mappa delle scienze sociali alla prova degli approcci computazionali, dicevamo. Ma in che modo costruire la mappa? Per quanto possa sembrare strano, il modo meno interessante di costruire questa mappa è affidarsi alla tradizionale distinzione delle scienze sociali. Ovviamente esistono approcci computazionali in sociologia, antropologia, economia e psicologia sociale, ma spesso questi approcci si sovrappongono e si mescolano, così come un campo disciplinare che si trova all'incrocio di diverse discipline è logico si comporti.

Per questo motivo invece che seguire le divisioni disciplinari, abbiamo preferito far riferimento al concetto di programma di ricerca.

In filosofia della scienza il termine è stato utilizzato da Imre Lakatos per distinguere la propria posizione filosofica secondo cui bisognava rifuggire sia dal falsificazionismo di Popper (una teoria scientifica è tale se falsificabile, ovvero se si possono definire i criteri di quando non funziona) che dal convenzionalismo delle filosofie successive a Kuhn (i criteri di accettazione di una teoria scientifica sono storico-sociali e interni alla comunità di riferimento) (Lakatos 1970). Secondo Lakatos, il programma di ricerca – ovvero l'obiettivo che si pone una teoria scientifica – altro non è che le regole

metodologiche che la comunità scientifica si pone per definire quale linea di ricerca evitare (che Lakatos chiama euristiche negative) e quali linee di ricerca perseguire (dette euristiche positive) dando vita così a una successione di teorie (Lakatos 1970: 208). Non dettaglieremo oltre la questione della metodologia dei programmi di ricerca scientifici di Imre Lakatos, quello che ci preme sottolineare è che i programmi di ricerca hanno tre caratteristiche: un obiettivo teorico ed empirico, una metodologia condivisa (per quanto problematizzata), e una successione di teorie che delineano l'avanzamento del sapere scientifico. Nelle pagine seguenti tratteremo i diversi programmi di ricerca che si sono posti come obiettivo quello dell'avanzamento della conoscenza scientifica facendo affidamento all'approccio computazionale come strumento privilegiato di conoscenza.

2.2.1 Sociologia analitica: la spiegazione tramite meccanismi

Spiega i fenomeni sociali tramite i meccanismi sociali; tutti i fenomeni sociali sono spiegati tramite la dinamica macro-micro-macro, un grande evento può essere spiegato tramite i suoi componenti che devono essere ricondotti all'evento.

Recente orientamento teorico ed empirico delle scienze sociali, per recente: a partire dagli anni '80. Autori più significativi: Jon Elster, Raymond Doudon, James Coleman.

L'ultimo si è inventato un sistema di rappresentazione → barca di Coleman, spiega la dinamica macro-micro-macro.

- Concetti chiave → macro-micro-macro

Spiegare processi macro con dinamiche micro (azione sociale)

- Concetti chiave → spiegazione tramite meccanismi
- Meccanismi → entità e attività organizzate per generare un risultato. Sono micro-processi sociali legati tra di loro – non variabili – che spiegano un fenomeno macro.
- Il meccanismo non dev'essere necessariamente uno schema statistico → il meccanismo esplicativo non è statistico. Le componenti si uniscono tra di loro perché si spiega il fenomeno tramite micro-processi.
- Spiegazione statistica: cercare legame fra variabili con strumenti statistici.
- Spiegazione per meccanismi: cercare spiegazioni logiche e poi cercare sostegno empirico. = il meccanismo spiega e la correlazione conferma.

cercare
differenza

Legami sociologia analitica e approccio computazionale.

- I meccanismi sociali sono modelli che possono sviluppare algoritmi.
- Teorie dell'azione sociale derivate dalle scienze cognitive → teoria per cui la società si costituisce in base al comportamento delle persone; qui usano le teorie che nascono dalle scienze cognitive es. linguaggio, interazione.
- Uso di approcci teorici ed empirici pesantemente computazionali (simulazione sociale, interesse vero *large dataset*)

2.2.2 Scienza delle reti: la struttura del piccolo mondo

Studiando le reti ci si è resi conto che le reti sociali hanno delle ricorrenze topologiche (es. 50 reti sociali/ 50 neurali → le sociali hanno delle proprietà in comune) che le altre non hanno.

Le reti *matematiche* possono essere:

- Ordinate → tutte collegate.
- Disordinate → collegate secondo un grado di casualità.

} Le reti sociali stanno in mezzo

La scienza delle reti (*network analysis*) → nasce metodologica e diventa teorica.

- Programma di ricerca che mette insieme sociologia e matematica.

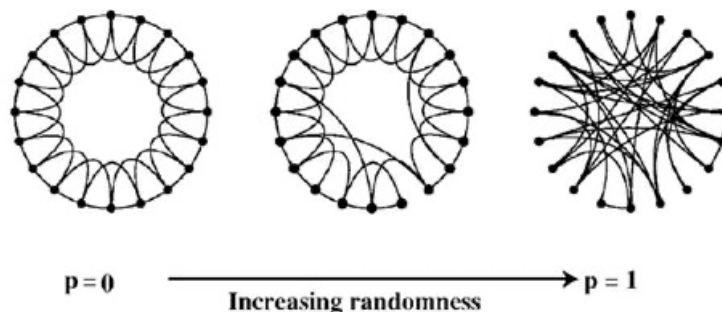
- Duncan Watts e Albert L. Barabasi → sono coloro che hanno studiato questa scienza.

La teoria del piccolo mondo (*small world theory*): la versione classica.

- Principio dei sei gradi di separazione: ogni persona al mondo è connessa da una media di sei contatti. es. gioco persona normale a Donald Trump.
- Esperimento di Stanley Milgram (1967) pioniere della *social network analysis*. → mandare sei pacchetti da un campione dal Nebraska a Kansas e deve arrivare a un agente di cambio di Boston (Massachusetts). L'aspettativa era la perdita dei pacchetti, invece sono arrivati il 29% dei pacchetti. Lui aveva ideato un meccanismo per tracciare i passaggi → per passare da un punto generico a un punto preciso: media di 5,2 passaggi. La ricorrenza statistica è stata trovata anche in altri casi.
Spiegazione sociologica → equilibrio tra omofilia (un nodo si connette con un altro se ha delle somiglianze sociali) ed eterofilia (connessione in assenza di somiglianze sociali).

La teoria del piccolo mondo (*small world theory*): la versione moderna.

- Riverniciata 2.0: Duncan Watts e Steven Strogats (1998)
- Le reti del piccolo mondo, l'esperimento viene fatto da un punto di vista computazionale → né ordinate né casuali (è una rete parzialmente ordinata) → ogni nodo è connesso a un certo numero di nodi vicini (ordine) e questi nodi ad altri nodi (casualità) secondo una frazione definita dal parametro $0 < p < 1$ (p : probabilità di connessione; 0: disordinata; 1: ordinata) cioè: Costruiamo legami con altre persone in maniera probabilistica con prevalenza di omofilia.
- Cluster e lunghezza media dei percorsi più brevi dipendono da p ; la media è 3.
- Il lavoro sui grafi casuali di Edrös e Rényi (1960) che diventano reti *small world*. → il fenomeno piccolo mondo emerge in una rete di qualunque dimensione. Come appare?
Appare superato un valore soglia (numero connessioni)



Il motivo per cui le reti del mondo reale non sono casuali: forza dei legami deboli → autore: Mark Granovetter → la gente trovava lavoro tramite un passaparola tra persone che non si conoscevano:

- Legami forti: tutti connessi tra loro (*cluster*), legami deboli alcuni sono connessi tra loro → legami deboli sono strumenti di accesso ad altri *cluster*.

L'oracolo di Kevin Bacon:

Kevin Bacon è un attore americano ma soprattutto è un hub, ha lavorato così tanto che conosceva tutti gli attori di Hollywood. Un gruppo di studenti hanno cercato di sviluppare un modello matematico sulla base di questa osservazione. Perché è interessante? Con questo giochetto si possono fare gli esperimenti, è una rete small world omofila: tutti i collegamenti, con qualsiasi tipologia di film il numero medio di passaggio è tre, massimo 4. Le reti in varianza di scala sono quelle che hanno le reti a pallotta, il mondo dei film è una rete del genere. Questa rete è una rete in varianza di scala ma anche small world. Il funzionamento deriva dalla presenza delle bridge person.

Legami network science e approccio computazionale:

- I modelli di reti sono basati su un gran numero di dati oppure sono reti del mondo reale: sono diventate uno studio ad alta matrice computazionale perché oggi studiare una rete di facebook,

linkedin ecc è diventato molto più semplice quindi la network analysis è riuscita ad unire i dati e le reti del mondo reale.

- La scienza delle reti propone modelli a metà tra modelli matematici dei fenomeni e l'interpretazione sociologica. L'interpretazione sociologica e linguistica è che possiamo applicarla a probabilmente qualsiasi cosa, spiegando l'esistenza di una determinata correlazione. L'esistenza di un legame non significa spiegare il perché esiste questo legame; con la statistica è la stessa cosa, dimostro l'esistenza di qualcosa ma non il motivo per il quale esiste.

2.2.3 Simulazione sociale: il mondo ricostruito nel computer.

Simulazione sociale vuol dire tradurre un modello di spiegazione delle scienze sociali in una forma computazione per analisi teoriche ed empiriche.

Il computer diventa un ambiente sperimentale simulato. Gli algoritmi non hanno tanto di matematico, sono spesso strutture logiche "if...then" quindi li puoi studiare logicamente e il vincolo logico lo puoi trasformare ed inserire all'interno di un computer.

Piccole variazioni iniziali nel sistema, creano grandi reazioni a lungo termine nel sistema, tutto ciò è non deterministico quindi lo stato finale del sistema non dipende dallo stato iniziale del sistema stesso. → caos deterministico, un disordine che assume una forma. È per questo che si parla di apprendimento nelle reti neurali.

Tutte le scienze sociali partono dal presupposto che le variabili che descrivono un processo sono talmente tante e legate che quando lavoriamo alcune cose non le consideriamo perché aumenta l'instabilità del fenomeno, più variabili metto più precisa sarà la spiegazione ma ogni volta che aumento le variabili diventa più complicato da spiegare.

Se vuoi spiegare processi molto complessi, utilizzi la statistica multivariata, in alcuni casi può essere più interessante però utilizzare il legame non lineare quindi utilizzando variabili probabilistiche. L'unico strumento che si ha è proprio la simulazione sociale. Fai esperimenti che esistono nel mondo reale ma che per motivi etici non puoi farlo (esempio: vedere come cambia la situazione lavorativa cancellando i sindacati dal mondo).

Le *miles stones* della *social simulation* risalgono agli anni '60 attraverso: la previsione elettorale, analisi dell'ecosistema terrestre dei referendum e della segregazione sociale.

Le tre componenti della simulazione sociale:

- 1) Componente teorica: strumento per esprimere le teorie delle scienze sociali assieme al linguaggio verbale e al linguaggio matematico. Ci sono due tipologie di modelli: informali e formali. I modelli informali sono quelli figurativi e/o verbali mentre i modelli formali sono matematici e/o simulativi.
- 2) Componente metodologica: modello ovvero versione semplificata di un fenomeno reale, a metà tra l'approccio quantitativo e qualitativo. La possibilità di fare esperimenti, avere risultati generali ed essere predittivo.
- 3) Analitica: la capacità di sollevare domande chiave per comprendere i fenomeni sociali come nello studio delle società artificiali (mondi possibili) o della complessità sociale (società come sistema sofisticato di variabili interdipendenti).

I diversi approcci alla simulazione sociale:

- System dynamics: sistemi di variabili in feedback costante.
- Automi cellulari: uno spazio in forma di griglia su cui si muovono enti dette cellule.

- Agent based models: simulazione di più enti che interagiscono fra di loro. L'ente di partenza può essere qualsiasi cosa: una persona, una famiglia, un'azienda, un sistema di aziende. Formalizzando l'ente di partenza studio come interagisce con gli altri.

2.2.4 Memetica: le idee come virus culturali

La memetica è un programma di ricerca basata sullo studiare l'evoluzione culturale come caratterizzata dagli stessi processi secondo il modello darwiniano.

- Le origini: il gene egoista (Dawkins 1976): darwinismo applicato sul codice genetico e non su organismi viventi. Gli organismi biologici non sono altro che strumenti al servizio dei geni, i quali servono a replicare il codice genetico. Ipotesi che è stata presa seriamente:
- Lo sviluppo: Daniel Dennet (1995): analogia evoluzione/algoritmo: evoluzione come algoritmo biologico. Questo vuol dire che l'evoluzione è un processo algoritmico in sé indipendente dal substrato (cioè biologia, cultura). Si può arrivare a una scienza della cultura: Memetica. Usa concetti come: replicazione, variazione, selezione.

Due versioni della memetica:

1. La versione forte: i meme come geni. Hanno le stesse caratteristiche dei geni e seguono le regole dell'evoluzione (darwiniana). Lo sviluppo è lamarckiano: i memi si adattano all'ambiente e si diffondono. Per Dawkins un meme è l'unità minima di trasmissione culturale. Il modo più semplice e intuitivo per considerare un meme è quello di pensarlo come una specie di idea ma non è chiaro cos'è un'idea. Molti di questi problemi sono risolti con:
2. La versione debole: i meme sono entità che consentono diffusione e propagazione di idee e concetti. Viene considerata come epidemiologica: la corrispondenza meme/gene è un'analogia. Questa versione ha una componente antropologica (Sperber) e la componente pop (memetica con media con Rushkoff e Brodie). Questa memetica ha contagiato la cultura di massa. Fantascienza e memetica: la spiegazione delle mode (Willis, Gibson). L'ambiente mediale dove hanno avuto maggiormente successo è Internet: meme ironici, politici.

Legami memetica e approccio computazionale:

1. Applicazioni del modello SIR/SER
2. Memetica e social simulation: usi metodologie di simulazioni sociali con modelli di memetica. Spesso hanno a che fare con concetti come l'identità.
3. Diffusione delle informazioni: information cascades (Easley, Kleinberg), cyber cascades (Sunstein).

2.2.5 Cliometria e cliodinamica: la storia come analisi quantitativa

Alla fine degli anni '70 sociologia empirica, antropologia culturale, economia politica, psicologia sociale sono solo alcune delle discipline che in quegli anni hanno ottenuto importanti risultati teorici ed empirici, risultati a cui l'uso del computer non è estraneo.

Cliometria: programma di ricerca di applicazione della teoria economica e metodi quantitativi allo studio della storia. Metodi di ricerca:

- Uso di database quantitativi e i modelli statistici ed economici.

I database quantitativi si presentano nella forma di serie storiche, ovvero dati cronologicamente organizzati relativi al fenomeno studiato oppure che vengono utilizzati come informazione indiretta per comprendere il fenomeno sotto analisi.

- Modelli di regressione per fare diverse tipologie di analisi come: regressione ecologica, analisi longitudinale ed altre analisi dove come fulcro c'è l'analisi comparata della variabile nel tempo.
- Analisi controfattuale: dove si elaborano delle ipotesi alternative di un processo o fenomeno spiegando perché non si sia verificato.

es. Fogel e ruolo economico ferrovie in America → la storiografia statunitense aveva considerato lo sviluppo della ferrovia come un fattore fondamentale per lo sviluppo economico degli Stati Uniti. Fogel elabora l'ipotesi che se la rete ferroviaria non si fosse costruita, l'economia americana sarebbe cresciuta in maniera uguale: al posto della ferrovia un sistema di canali/fiumi navigabili che avrebbe ridotto il costo del trasporto.

es. Fogel e schiavi in sud America → il sud aveva un'economia basata sul lavoro degli schiavi. Con questo commercio venivano acquistati beni di lusso e nulla reinvestito nelle stesse regioni. Attraverso una analisi della redditività delle piantagioni schiaviste, Fogel notò che gli schiavi erano un sistema efficiente compatibile con una economia di scala se venivano usati ininterrottamente tutto l'anno. Mentre erano strategici per la produzione di cotone, non lo erano per economie basate su altri prodotti. Gli schiavi avrebbero potuto essere usati in una economia industriale, ma semplicemente non erano efficienti.

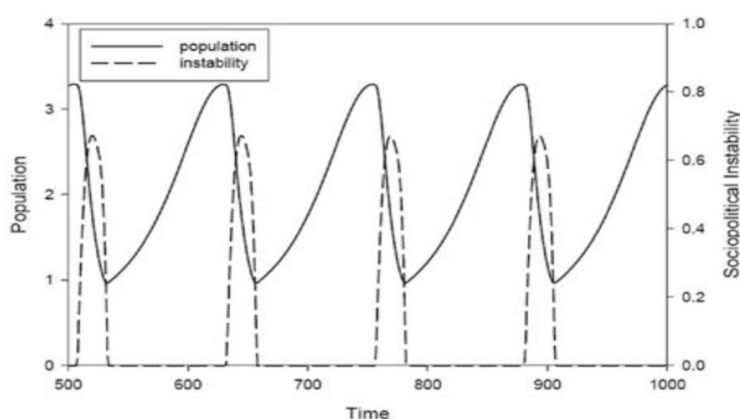
- La cliometria è interna al dibattito storiografico poiché studia i fenomeni storici nella loro specificità. È interessata ai cicli storici decennali, quindi lo sviluppo di un fenomeno in un lasso di tempo determinato.
- La cliodinamica è esterna al dibattito storico in quanto l'obiettivo è lo studio dei fenomeni storici alla ricerca di leggi e regolarità.

Uno dei punti chiave è l'enfasi nella componente predittiva. L'idea è applicare la modellistica biologica alla possibilità di trovare delle leggi nella storia, ovvero regolarità sistematiche che permettano di elaborare delle previsioni sull'andamento di fenomeni.

È interessato a cicli storici in lunga durata e il succedersi dei cicli di stabilità/violenza dell'andamento storico.

Nell'ambito della previsione un ruolo importante è attribuito alle predizioni retrospettive → esperimenti storici in cui si confrontano le affermazioni basate su dati relative allo stato di una variabile in un contesto storico a partire da due teorie concorrenti.

Modello demografico strutturale: la crisi sociale è frutto della composizione di due forze: la crescita costante della popolazione e l'impatto sulle istituzioni politiche, economiche e sociali. I cicli di stabilità/violenza vengono identificati come *secular cycles*.



Modelli matematici non-lineari, uso della *social simulation*, interesse verso la teoria della complessità, rendono la cliometria un programma di ricerca interessante il cui scopo è rendere la storia una scienza a tutti gli effetti.

2.2.6 *Economia comportamentale: attori cognitivi e mercati predittivi*

Una tendenza dell'economia contemporanea è l'economia comportamentale, detta anche cognitiva e sperimentale.

Uno dei dibattiti centrali della scienza economica si è fondato sulla netta distinzione con le scienze psicologiche è curioso notare che uno dei più importanti campi di ricerca dell'economia contemporanea sia nato grazie alla reintroduzione della psicologia nell'alveo della teoria economica.

In che cosa consiste il programma di ricerca dell'economia cognitiva e perché risulta interessante per le scienze sociali che usano un approccio computazionale? Un ruolo chiave è quello della teoria neoclassica:

es. l'epistemologia di Chicago si basava sui seguenti principi: l'agire economico è governato dalla razionalità individuale, nella ricerca economica le preferenze individuali non sono osservabili e qualunque fenomeno economico è riconducibile al comportamento degli individui.

Il punto chiave di questa impostazione è basata sul concetto di razionalità individuale, intesa come processo di calcolo mentale in cui lo scopo dell'attore economico è quello di massimizzare le conseguenze positive delle sue scelte economiche. L'idea è che gli individui sanno che a ogni propria azione corrisponde una conseguenza; pertanto, è possibile ordinare le proprie azioni sulla base delle conseguenze che esse avranno su sé.

I suoi limiti sono stati mostrati tramite gli studi di Allais → i soggetti devono scegliere fra due gruppi di decisioni, in cui un elemento della coppia rappresenta una violazione degli assiomi alla base dell'utilità attesa. I soggetti tendono a non massimizzare le proprie utilità → razionalità limitata in cui il punto chiave è la costruzione del contesto della decisione, perché grazie a questo processo è possibile operare una scelta.

Tversky e Kahneman → il contesto della decisione opera come una cornice che può veicolare le scelte degli attori: se in una decisione rischiosa le alternative sono rappresentate come fonte di perdita, gli individui si comportano come sfavorevoli al rischio; invece, se le alternative sono presentate come fonte di guadagno, gli individui saranno favorevoli al rischio. La *Prospect Theory* ritiene che il processo di decisione consista in due momenti distinti, un momento di editing delle decisioni e un momento di valutazione delle alternative.

L'economia cognitiva è interessante per l'approccio computazionale perché il corso alla psicologia cognitiva ha permesso di modellizzare i processi decisionali alla base degli attori economici. L'approccio sperimentale in economia è diventato un test per verificare il funzionamento delle teorie.

Lo sforzo di comprendere il comportamento degli attori economici ha portato all'analisi di particolari contesti economici collettivi come il comportamento dei mercati, da cui è scaturita la linea dei mercati predittivi → sono dei particolari tipi di mercati elettronici, i quali simulano una situazione di compravendita di titoli finanziari e vengono utilizzati per la previsione di eventi futuri. I trader devono registrarsi sul sito e versare una quota in denaro, poiché questi mercati usano (e pagano) con denaro vero. es. *L'Hollywood Stock Exchange* in cui è possibile contrattare titoli relativi all'industria cinematografica dalle candidature agli Oscar fino al guadagno ai botteghini.

I mercati predittivi non sono perfetti: anch'essi sono soggetti a errori dovuti a diverse motivazioni come la sopravvalutazione delle decisioni, diffusione incontrollata di notizie non verificate o manipolazione da parte di trader spregiudicati.

2.2.7 *Digital humanities e culturomics; il computer strumento dell'umanista*

Il rapporto stabile sia teorico che metodologico fra computer e scienze umane è frutto del programma di ricerca delle *digital humanities*. Qual è il contributo che il computer e l'informatica possono dare all'avanzamento delle scienze umane?

È necessario distinguere due momenti che sono *humanities computing* e le *digital humanities*.

- *Humanities computing* → programma di ricerca il cui obiettivo è applicare le possibilità di calcolo e analisi rese possibile dal computer anche nell'ambito delle scienze umane.
Fanno riferimento al primo periodo della relazione fra informatica e scienze umane. Studioso fondatore di questo approccio: padre Busa.
Questo periodo è anche quello più importante per la creazione di una comunità di studiosi che si riconoscono nello studio delle scienze umane attraverso l'uso sistematico di strumenti computazionali.
Quasi tutti gli studiosi di questo primo periodo fanno risalire l'approccio delle *humanities computing* alle prime riflessioni sul linguaggio, comunicazione e informatica facendo riferimento ad autori come Turing.
In questa fase nasce anche la metalinguistica, TEI, SGML e XML.
- *Digital humanities* → condividono con le *humanities computing* lo stesso programma di ricerca, con in più la componente legata alla diffusione di Internet e alla nascita del World Wide Web.
L'emergere della nuova denominazione deve essere considerato come un cambiamento concettuale, come l'ingresso del concetto di ipertesto. In questa fase nasce il *Memex*.

La dimensione computazionale lascia spazio alla componente digitale, allargando la concezione di ricerca umanistica che comprende aree come i *game studies*, letteratura *fandom* e culture del remix.

La contaminazione fra *digital humanities*/ studi sulla comunicazione online ha fatto sì che alcuni temi chiave delle scienze sociali contemporanee siano entrati a far parte del bagaglio degli umanisti digitali.

Ciò che rende le *digital humanities* un programma di ricerca molto interessante in un approccio computazionale delle scienze sociali e umane è l'interesse verso il processo di sviluppo e uso dei modelli nel campo delle discipline umanistiche. In pratica se si volessero descrivere le due forme principali in cui c'è la collaborazione fra computer e scienze umane, potremmo distinguere fra la computazione per le scienze umane e la computazione nelle scienze umane.

L'approccio che coniuga letteratura, strumenti digitali e utilizzo di database negli ultimi tempi si è collegato con alcuni progetti legati ai big data → nasce la *culturomics* → fa riferimento a un progetto di ricerca che mette insieme le *digital humanities* con il progetto di Google per la digitalizzazione dei libri.

La *culturomics* nelle intenzioni dei suoi ideatori è l'applicazione dei big data allo studio della cultura umana; attraverso l'interrogazione di enormi database è possibile studiare nella sua componente diacronica alla ricerca di pattern linguistici o altre strutture che potrebbero suggerire un'interpretazione storica-culturale di diversi processi.

L'approccio fortemente quantitativo ma ricco di sfumature interpretative negli studi della cultura è tipico di altri orientamenti di analisi come la *culturale analytics*.

2.3 *Scienza sociale computazionale: modelli interdisciplinari per lo studio della complessità con un obiettivo sociale*

Scienza sociale computazionale: programma di ricerca che consiste nell'indagine interdisciplinare dell'universo sociale a diversi livelli – dall'individuale al collettivo – attraverso lo strumento della computazione.

Le diverse definizioni convergono su 3 dimensioni chiave: interdisciplinarietà, uso di modelli, ruolo sociale.

1. Interdisciplinarietà → variabilità delle discipline che contribuiscono a creare un programma di ricerca il cui scopo è la risposta articolata a domande complesse sulla società e sulla cultura contemporanea.

È il risultato delle esigenze di:

- ricchezza di analisi → deve affrontare problemi che si fondano sulla stessa esigenza: l'importanza della comunicazione nei fenomeni umani e sociali e la caratteristica della complessità.

I sistemi sociali e i loro processi si fondano su uno scambio di informazioni e sull'attivazione di canali di comunicazione; in questo modo si legittima l'uso dell'impostazione computazionale come strumento di ricerca. La complessità è un fenomeno tipico delle scienze sociali.

- orientamento metodologico → se alla base della società c'è lo scambio di informazioni, si può presupporre che lo scambio possa essere studiato con l'uso sistematico dello strumento informatico.

2. Uso di modelli → centralità assegnata ai processi di modellazione e necessità di una maggiore attenzione ai dati.

È aumentata l'esigenza di maggiore integrazione fra teoria e ricerca. L'uso dei modelli consente di dare una maggiore rilevanza al ruolo dei dati; è possibile fare esperimenti nella forma di esperimenti mentali attraverso la manipolazione delle variabili alla base dei modelli.

3. Ruolo sociale → il ricorso ai modelli fa sì che non possa esimersi dal rispondere alle sfide poste dal mondo in cui viviamo. Le sfide diventano non solo concettuali ma politiche, per le conseguenze che hanno in un'ottica consulenziale.

Shneiderman chiama "Scienza 2.0" quella nuova fase nella collaborazione scienze umane e sociali con scienze informatiche che hanno il loro punto d'incontro nella quantità di dati prodotti nei *setting* tecnologici.

3. Casi. Ricerche e controversie della scienza sociale computazionale

3.1 Strategie di ricerca della scienza sociale computazionale

Le ricerche che vengono compiute nella SSC sono ricerche interessate a studiare fenomeni contemporanei.

La rete si comporta come un enorme laboratorio i cui eventi che accadono possono essere descritti come degli esperimenti sociali naturali: quando accade qualcosa i dati sono già pronti per essere analizzati e il fenomeno accaduto è pronto per essere modellizzato.

Le grandi piattaforme di social media sono protagoniste della vita sociale e relazionale contemporanea. Quando un fenomeno globale viene studiato grazie ai big data che esso produce all'interno delle piattaforme dei media sociali, la SSC affronta la questione seguendo due linee argomentative.

Un fenomeno può essere studiato nella sua specificità o generalità.

3.1.1 La struttura della rete dei contatti di Facebook rivela i partner

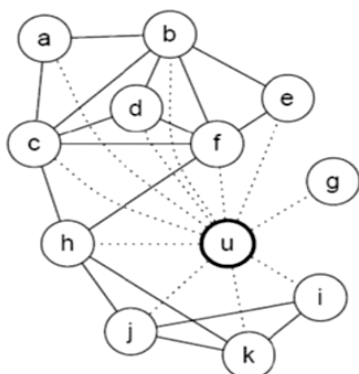
Ricerca di Backstrom e Kleinberg → domanda: esiste una struttura della rete di una persona che rivela che uno dei nodi è il proprio partner?

La topologia della rete sociale è stato un elemento importante per la *social network analysis*. Esiste una letteratura sulle caratteristiche dei legami e sul ruolo che hanno nell'ampliare il capitale sociale es. paper di Granovetter sulla possibilità di trovare un lavoro tramite legami deboli.

Un concetto usato in queste analisi è quello di *embeddedness*: il numero di amici reciproci che due persone condividono. Il concetto di *embedding* è uno strumento non in grado d'identificare le relazioni affettive → definiscono *dispersion*: considera non solo il numero di amici reciproci di due persone ma anche la struttura della rete di questi amici condivisi.

Il campione in analisi sono persone che dichiarano su FB di essere in una relazione ma il partner è anonimo. → si studiano gli amici e i loro legami reciproci per provare a identificare quale fosse il partner a partire solo da proprietà strutturali della rete.

Si può ipotizzare di selezionare le connessioni che partono da u che hanno la massima *embeddedness* e provare a identificare il partner v . Non è il massimo come tecnica perché sia u che v possono avere delle sottoreti che non condividono. I collegamenti tra u e il partner possono avere una bassa *embeddedness*, ma spesso coinvolgono contatti reciproci con un numero elevato di focus di interesse. Per questo ipotizzano che il collegamento tra u e v presenta una struttura dispersa.



L'utente ha un'alta *embeddedness* con i vicini c, b, f (pari a 5: si contano i link del nodo che ne ha di più) e con h, k, i. L'utente u e h sono gli intermediari (nodi che li connettono) di c-f e j-k. La connessione con dispersione maggiore è u-h: perciò è possibile che h sia partner di u .

L'indice mostra una buona performance che arriva fino a prevedere il partner nel 75% dei casi. I risultati sono interessanti perché dalla rete dei contatti reciproci di due persone è possibile desumere il tipo di relazione che lega le due persone, e, la regola dell'*embeddedness* non vale in tutti i casi: se due persone si dichiarano una coppia ma l'indice di dispersione è basso, c'è il 50% di probabilità che la coppia si romperà nei mesi successivi.

3.1.2 Gli archetipi delle conversazioni su Twitter

Per quanto twitter non sia rappresentativo degli utilizzatori di internet, ha un ruolo importante nell'ambito giornalistico.

Twitter è uno spazio interessante per studiare in che modo si creano conversazioni che circolano intorno a un tema specifico.

Pew Research Centre + Social Media Research Foundation → mettere a punto una tassonomia delle conversazioni che avvengono su twitter.

Per analizzare la *topic network* di twitter sono stati presi in considerazione diversi parametri: dimensione e struttura della rete e delle sue sottoreti, analisi delle parole, uso degli hashtag, circolazione di url.

Sono state costruite mappe su vari argomenti che sono state sottoposte a un'analisi osservativa per cercare le strutture ricorrenti all'interno dei diversi grafi ottenuti grazie all'uso di software per la visualizzazione delle reti.

Per creare delle mappe dettagliate sono state raccolte:

- Analisi della densità di rete
- Isolamento delle principali clique
- Identificazione dei *bridge*
- Identificazione degli *hub*
- Isolamento dei piccoli gruppi

Ha permesso di identificare sei strutture distintive

1. Il pubblico polarizzato → due gruppi con poche connessioni tra loro. I temi discussi sono caldi e portano persone a dividersi e a radicalizzare le proprie posizioni.
2. Folla concatenata → le discussioni sono caratterizzate da persone altamente interconnesse con pochissimi partecipanti isolati.
3. Gruppi relativi a brand → conversazioni relative a prodotti, servizi, personaggi popolari; molti partecipanti disconnessi.
4. Gruppi comunitari → temi popolari che si sviluppano intorno a molti piccoli gruppi, spesso organizzati intorno a nodi rilevanti
5. Rete broadcast → discussione che prende le mosse da notizie dell'ultimo minuto o sono il risultato di specifiche fonti informative ben conosciute; presenza di nodi rilevanti in cui ci sono moltissime persone che retwittano.
6. Rete di sostegno → strumento della gestione dei clienti per intervenire in caso di problemi relativi a prodotti e servizi.

La ricerca sulla tipologia di conversazioni mostra che queste danno vita a configurazioni molto diverse che dipendono dall'argomento trattato e dal coinvolgimento delle persone che partecipano alla conversazione.

3.1.3 *Cospirazioni e razionalisti alla prova di Facebook*

Un settore di studio affascinante è la diffusione delle notizie attraverso i social media.

Un caso specifico di questo processo è quella della circolazione di contenuti che contengono ipotesi complottiste o cospirazioniste.

Quello della diffusione delle teorie cospirazioniste è solo uno degli aspetti della circolazione di informazioni inaccurate/false.

Università Pavia + Lucca + Roma → obiettivo produrre un atlante delle informazioni scientifiche e alternative presenti su Facebook e studiare i modelli di consumo delle notizie da parte di cospirazionisti e razionalisti. Il dataset va dal periodo 2010-2014, dati ottenuti facendo riferimento alle API di Facebook e usando solo informazioni disponibili pubblicamente.

Lo studio si è concentrato sugli utenti polarizzati che hanno messo il 95% dei like nelle pagine di una sola categoria.

Lo studio della correlazione fra le interazioni possibili sulle pagine e sui post, mostra che i cospirazionisti sono più attivi dei razionalisti nel mettere like e condividere i post delle pagine. Questo comportamento può essere interpretato come chiusura cognitiva, ovvero come maggiore disponibilità a consumare e diffondere solo contenuti che sostengono il proprio punto di vista come interagiscono i razionalisti con i contenuti cospirazionisti e i cospirazionisti con i contenuti razionalisti? Identificati gli utenti polarizzati. Fatto 100 il numero dei like messi sulle pagine, il 76,79% dei like delle pagine razionaliste vengono da utenti polarizzati, mentre nel caso delle pagine cospirazioniste la percentuale è 91,53% → pagine cospirazioniste vengono seguite da utenti cospirazionisti.

Gli utenti razionalisti tendono a commentare i post razionalisti nel 90,29% dei casi e i post cospirazionisti nel 9,71% dei casi, di contro i cospirazionisti commentano i post cospirazionisti nel 99,08% dei casi e i post razionalisti nel 0,92%. Questo sta a indicare che i cospirazionisti sono più propensi a interagire con la propria community di riferimento rispetto ai razionalisti.

Per i post palesemente falsi, entrambe le categorie vi interagiscono ma i cospirazionisti sembrano più attivi rispetto agli utenti razionalisti.

3.1.4 *Il contagio delle emozioni di Facebook*

Cosa succede quando leggiamo il feed di Facebook? Cosa accade se questi contenuti hanno una specifica caratterizzazione emotiva? Di norma non siamo abituati a fruire delle emozioni altrui in forma testuale, se riusciamo a immedesimarci nelle emozioni altrui solo tramite l'interazione testuale alla fine come cambierà il nostro umore? È possibile modificare lo stato emotivo di una persona attraverso i messaggi circolanti in Facebook?

Esistono evidenze empiriche a sostegno dell'idea che le espressioni emotive di una persona su Facebook possono predire le emozioni espresse dai suoi amici anche qualche giorno dopo.

Lo studio ha seguito un protocollo di tipo sperimentale in cui è stato manipolato il feed per verificare se le persone sottoposte a una serie di contenuti con una precisa caratterizzazione emotiva avrebbe modificato in maniera coerente il tipo di contenuti postati da queste stesse persone.

Un gruppo di persone è stato soggetto a contenuti di carattere positivo e un altro a negativo. Questa manipolazione riguardava solo la bacheca pubblica. L'esperimento è stato svolto per una settimana con partecipanti casuali.

Le persone sottoposte al News Feed con ridotta presenza di post positivi hanno diminuito l'uso di parole che esprimevano emozioni positive, nelle persone sottoposte a post negativi hanno diminuito le parole esprimenti emozioni negative.

È stato dimostrato che siamo sensibili alle espressioni emotive degli altri anche tramite interazioni testuali. Le persone sottoposte a un numero minore di post con contenuti emotivi, sono stati meno emotivamente espressivi nei giorni successivi, rafforzando così l'idea che chi legge contenuti positivi è più propenso a esprimere emozioni positive.

3.2 Controversie e limiti della scienza sociale computazionale

Nella scienza sociale computazionale esiste un "lato oscuro" che sta facendo tornare sulla scena pubblica in modo pesante domande sulla legittimità di queste analisi.

Questi studi hanno a che fare con le persone, quindi non si può dimenticare che il materiale su cui si lavora sono esseri umani.

3.2.1 Datafrenia: Gramellini, Princeton e l'estinzione di Facebook

Vi è la paura dell'esasperata quantificazione del mondo e dei rapporti sociali.

Datafrenia → radicalizzazione della quantofrenia, neologismo coniato da Sorokin per descrivere la fiducia acritica nei confronti delle possibilità di quantificazione dei fenomeni sociali.

Questa posizione contraddice tutte le riflessioni della moderna filosofia della scienza secondo cui l'osservazione non solo ha bisogno di teoria ma l'osservazione stessa è carica di teoria.

Se mai ci fosse bisogno di rimarcare quanto sia necessario un approccio critico all'analisi dei big data, è utile descrivere la controversia relativa alla ricerca di Princeton sull'estinzione di facebook → il dipartimento ritiene che facebook tra il 2015/2017 perderà l'80% degli utenti. La base dei dati parte dalla ricerca del termine su Google Trends; su questi è stato applicato il modello SIR a cui sono state fatte delle modifiche ed è stato nominato modello irSIR.

SIR nasce come una funzione matematica per spiegare il declino di MySpace sempre usando dati da Google Trends. Trovata la funzione e applicata ai dati di facebook viene data la data di morte. Il passaggio metodologico è errato.

Si presuppone che le persone cerchino il social sempre da google, senza considerare app, la pagina direttamente salvata o gente perennemente loggata.

Dire che i social network si evolvono come malattie nega l'intenzionalità dell'agire umano. In aggiunta, usare MySpace come modello è sbagliato perché i due social hanno caratteristiche diverse.

In risposta → data scientist di facebook applica lo stesso modello all'università di Princeton mostrando che sta perdendo d'interesse.

L'applicazione di modelli a fenomeni socio-tecnici come il ciclo di vita dei social network è molto interessante ed è il campo di azione della SSC.

Il mondo dei social media, essendo all'incrocio di tecnologia, innovazione, impresa, economia, è molto difficile da prevedere con modelli statistici, per quanto sofisticati.

Questa vicenda apre una questione sul fenomeno dei *paper* non sottoposti a *peer review* che entrano nel circuito comunicativo.

3.2.2 *Dataumanesimo: dietro i numeri ci sono le persone*

L'obiettivo delle grandi data company contemporanee è quello di orientare il comportamento sociale verso specifici obiettivi commerciali.

La questione dell'etica della ricerca in ambito computazionale sta diventando sempre più pressante. Quando si manipolano i dati bisogna ricordare che a essere manipolati sono informazioni relative alle persone.

Facebook contagion → lo studio diventa controverso quando si passa ad analizzare in dettaglio la metodologia di ricerca adottata. Gli autori dello studio hanno dichiarato il loro esperimento legittimo poiché l'accettazione dei termini di utilizzo può essere considerato una sorta di consenso informato.

Le critiche allo studio sono molteplici, ma tutte gravitano intorno alla domanda se sia legittimo che un social network usi i propri utenti come delle cavia. Considerare i termini di servizio come consenso informato non è corretto poiché pochi leggono fino in fondo il contratto che Facebook obbliga a sottoscrivere con la forma del silenzio-assenso.

I social media devono essere considerati come parte di noi stessi e pertanto i protocolli sperimentali devono essere rigidi alla stregua dei protocolli utilizzati per la ricerca farmaceutica.

Se Facebook crea una cortina digitale virtualmente impenetrabile per poter mantenere il controllo dei dati prodotti dagli utenti, Twitter invece ha una politica diversa rendendo accessibile molte informazioni attraverso le API. Accessibile non vuol dire automaticamente etico.

I dati di superficie fanno riferimento a un gran numero di persone; quelli di profondità a pochi individui o piccoli gruppi.

In mezzo a questi due concetti di dati, c'è il concetto di statistica e di campionamento, ovvero la possibilità di dire come si comporta un universo sociale di riferimento analizzando una versione ridotta dei soggetti che ne fanno parte.

L'utilizzo dei social media e dei big data permette di superare la dicotomia fra dati di superficie e dati di profondità.

Un utilizzo delle piattaforme di social media costante nel tempo, permette non solo di avere informazioni dettagliate, ma anche di vedere come queste informazioni cambiano nel tempo.

3.2.3 *Dataveglanza: da House of Cards a Minority Report*

Netflix ha un sistema con cui raccoglie pareri e altre informazioni e li usa per migliorare l'offerta. La raccolta sistematica di queste informazioni ha fatto sì che il database di Netflix sappia sui propri clienti molte più informazioni di quanto gli stessi clienti siano disposti ad ammettere.

Le loro analisi avevano previsto che *House of Cards* sarebbe stato un successo e così fu.

Monitoring e profiling sono i due processi chiave che hanno portato le tecnologie dalla sorveglianza a un livello nuovo e più sofisticato.

Dataveglanza è la sorveglianza basata sui dati ed è stata delineata da Clarke. Questa tecnica non fa altro che digitalizzare e industrializzare delle pratiche di controllo sociale che esistono da tantissimo tempo.

A seconda del tipo di dati su cui gli algoritmi vengono applicati, possono esserci minacce alla libertà individuale. Il rischio è che si verifichi quello che possiamo chiamare effetto *Minority Report* se un

algoritmo predittivo sostiene oltre una certa soglia la probabilità che un soggetto possa delinquere senza che ancora abbia compiuto alcun atto criminoso, come ci si deve comportare?

Se è legittimo pensare che piattaforme come facebook siano al contempo strumenti per la connettività sociale e strumenti di scienza sociale computazionale cosa succede se vengono implementate le funzioni per migliorare la ricerca delle informazioni in questi spazi?

Esistono diversi tentativi per salvaguardare la riservatezza delle informazioni personali.

Una soluzione è definire nuovi diritti per l'era digitale come il diritto all'oblio, ovvero la possibilità di essere cancellati dai database dei motori di ricerca o in cui alcune informazioni che ci riguardano non siano di facile accesso.