

Big Data. . . Perché (oh perché?) questa scienza sociale computazionale?

David O'Sullivan

Molti altri, in questo volume e altrove, hanno commentato e commenteranno le implicazioni politiche, sociali, economiche, culturali ed etiche dei big data.¹ Sono pienamente d'accordo sul fatto che tali implicazioni siano importanti. In effetti credo che siano gli aspetti più urgenti dei big data che le scienze sociali impegnate in modo critico devono affrontare. C'è una lotta da combattere nell'arena pubblica sulle molte direzioni preoccupanti verso cui si stanno muovendo gli impulsi politico-economici che guidano i big data. I modi complicati in cui il movimento dei big data (se riusciamo a chiamarlo così) è intrecciato con uno stato in espansione, autoritario e di sorveglianza, che contemporaneamente permette ed è schiavo di una retorica di "perturbazioni" del libero mercato, richiedono la nostra attenzione, se vogliamo che siano contrastati da alternative più umane. Esiste ovviamente una sostanziale letteratura che si rifiuta di chinarsi davanti al trionfalismo tecnologico (o è fatalismo?) di lavori come "Too Big to Know" e "Big Data".² Molte delle preoccupazioni sollevate dal momento attuale sono abilmente dissezionate in "Digital Disconnect", "The Filter Bubble", "You Are Not a Gadget", "Who Owns the Future?", e "To Save Everything, Click Here", tra molti altri.³ Data la centralità dei dati geografici in questo diluvio di dati, è sicuramente importante che i geografi diventino più visibili in questa conversazione pubblica.⁴ Contributi sostanziali, come "Code/Space", rimangono saldamente in stile accademico ma forniscono comunque una base per contributi futuri che affrontano aspetti più specificamente spaziali dei big data e i suoi impatti.⁵

Riconoscendo l'importanza di questi dibattiti più ampi, desidero tuttavia concentrarmi su preoccupazioni metodologiche più ristrette. Tra quanto sono riusciti o si dice siano riusciti a realizzare i big data, di sicuro c'è l'aver posto saldamente i metodi quantitativi e computazionali nell'agenda delle scienze sociali.⁶ Accolgo con favore tale sviluppo e lo vedo come un'apertura per una visione più plurale della geografia e delle altre scienze sociali. Tuttavia, è improbabile che l'apertura ci porti in qualsiasi luogo nuovo se continuiamo a interpretare i big data esclusivamente sulla base della novità dei dati stessi. I dati, per quanto "grandi", sono fortemente limitati nel modo in cui rappresentano i *processi*. Data la centralità del concetto di "processo" per sviluppare una comprensione sofisticata di come funziona il mondo, questo è più di una limitazione dei big data. Se la comprensione, la spiegazione e l'intervento efficace nel mondo sono gli obiettivi perseguiti, allora dobbiamo porci domande sullo stile delle scienze sociali computazionali a cui dovremmo puntare. Ci sono comunque molti indizi che (eccessiva) eccitazione e clamore nei confronti dei big data rischiano di farci perdere di vista tali domande. Ciò sarebbe un male sia per gli oppositori sia per i sostenitori del potenziale dei big data per le scienze sociali; attirare l'attenzione su questi temi è quindi il mio obiettivo.

Nella prossima sezione espongo la mia interpretazione dell'epistemologia dei big data e suggerisco perché i big data abbiano avuto tanto successo, cioè come tecnologia ampiamente adottata, non necessariamente come un modo efficace per comprendere il mondo. Da lì passo a considerare un dualismo persistente nel modo in cui gli strumenti computazionali sono stati impiegati nelle scienze, vale a dire una distinzione tra approcci top-down, aggregati o statistici alla spiegazione (tra cui possono essere collocati i big data) e bottom-up, approcci emergentisti spesso associati alla scienza della complessità.⁷ Sebbene queste due tradizioni condividano elementi sostanziali nel loro patrimonio intellettuale, producono prospettive fortemente divergenti su spiegazione, comprensione e previsione e suggeriscono stili intellettuali e direzioni metodologiche molto diversi per la scienza sociale computazionale. Le due tradizioni affrontano anche la questione centrale del processo in modo molto diverso. L'attuale apertura a un maggiore utilizzo di computer e dati quantitativi (secondari*) è un'opportunità per una scienza sociale migliore, più efficace che rischiamo di perdere

* raccolti da altre persone che non sono l'utente dei dati

se il paradigma dei big data rimane dominante. Tale pericolo ha aspetti sia scientifici che etici, riportando l'argomento nel contesto più ampio considerato all'inizio.

La misteriosa ascesa dei Big Data

I big data sembrano usciti dal nulla, molto rapidamente, ma è un'illusione. Le copertine oramai icona di riviste sull'argomento, come "Big Data: Science in the Petabyte Era" di Nature e "Data Deluge" dell'Economist, hanno reso popolare il termine "big data" ma testimoniano sviluppi già ben avviati.⁸ La velocità con cui una tale "tendenza" rilevata (e inevitabilmente amplificata) dai media si è trasformata in una direttiva primaria per tutta la scienza è stata sorprendente. Aver vissuto in Nuova Zelanda fino alla fine del 2013 mi ha in qualche modo protetto da questa valanga, ma anche lì all'inizio del 2013 i big data erano inevitabili, come evidenziato da una serie di discussioni sull'argomento, sponsorizzate dalla Royal Society of New Zealand e dalla National Library e trasmesse da Radio Nuova Zelanda, finanziata dallo stato.⁹ Come è accaduto altrove, le iniziative nazionali di finanziamento e infrastrutture scientifiche sono state rapidamente agganciate a quest'ultimo imperativo strategico urgente, aprendo opportunità redditizie per società private che portano la banda larga ultraveloce nella Nuova Zelanda rurale. Come è accaduto altrove, questi sviluppi sono spesso esplicitamente collegati all'evoluzione parallela delle "smart cities".¹⁰ I dettagli dell'esperienza della Nuova Zelanda nel settore dei big data e delle città intelligenti sono geograficamente, politicamente e culturalmente specifici, ma i punti in comune con ciò che sta accadendo altrove sono sorprendenti.¹¹ Il caso della Nuova Zelanda pone immediatamente in primo piano la questione spesso discussa di cosa esattamente rende i big data grandi. Certamente il volume di dati generati in Nuova Zelanda non si qualificerebbe come grande in molti altri contesti. Le ormai familiari tre V – volume, velocità e varietà – caratteristiche presumibilmente definitive dei big data sono state, direi a ragione, rubate e usate in un rapporto di business intelligence.¹² Si sono dimostrate insufficientemente descrittive per molti, portando a un'industria accademica che propone e discute attributi aggiuntivi (preferibilmente quelli che iniziano con la lettera V). Piuttosto che aggiungere a quel dibattito, ritengo che i big data siano principalmente una questione epistemologica.¹³

Tale posizione può essere grossolanamente definita come un'affermazione secondo cui, data una sufficiente quantità di dati, il mondo può essere conosciuto (se non completamente, abbastanza bene per qualsiasi scopo particolare) e che siamo attualmente sulla soglia di un'era in cui la capacità tecnologica di assemblare insieme di dati così completi è (finalmente!) arrivata. Come tale, questo momento preannuncia la realizzazione di un sogno di alcuni tipi di positivismo deterministico.¹⁵

Questo è ovviamente un terreno filosoficamente estremamente traballante, anche se sembra che le difficoltà epistemologiche non siano un fattore dissuasivo per l'adozione dell'approccio.¹⁶ Laddove i big data possono essere distribuiti, il punto in molti contesti non è, dopo tutto, capire il mondo ma semplicemente conoscerlo abbastanza bene da realizzare un profitto o, ancora più in basso, essere plausibilmente in grado di affermare che un giorno un profitto potrebbe essere realizzato usando intuizioni ricavate dai dati.¹⁷ Basti dire che, mentre io, insieme a molti geografi, sono incuriosito dalle possibilità che tali insiemi di dati possono aprire, non sono per niente persuaso dalle grandiose affermazioni fatte per i big data. Comprendere il mondo richiede ancora che sviluppiamo attentamente teorie, consideriamo le implicazioni di quelle teorie per ciò che ci aspettiamo di osservare nel mondo e sottoponiamo tali aspettative a un controllo attraverso l'osservazione empirica, usando metodi multipli, solo alcuni dei quali sono migliorati dalla rete della raccolta di big data. Nonostante le affermazioni epistemologiche non convincenti, come mai questo particolare approccio computazionale è arrivato a dominare un pensiero così recente sull'uso del calcolo per conoscere il mondo?

Mi vengono in mente almeno tre risposte. In primo luogo, giustifica (retroattivamente) la raccolta di dati che poggia su fondamenti etici e legali discutibili. Non importa come siamo arrivati a essere in possesso di questi vasti archivi di dati; pensa solo a cosa possiamo fare con loro! Un caso di fini che giustificano i mezzi su scala sociale. In secondo luogo, il regime tecno-sociale dei big data è

fattibile in un contesto in cui la raccolta di più dati è diventata necessaria nell'ambito delle pratiche commerciali quotidiane. Non è chiaro che ciò che è emerso sia stato deliberatamente pianificato da nessuno dei principali protagonisti. Ad esempio, Brin e Page, in un primo articolo sul motore di ricerca di Google, notano che “ci aspettiamo che i motori di ricerca finanziati dalla pubblicità siano intrinsecamente distorti verso gli inserzionisti e lontano dalle esigenze dei consumatori”, suggerendo che la creazione di una società pubblicitaria raccoglitrice di dati non era la loro intenzione originale.¹⁸ Tuttavia, in un ambiente aziendale in cui la costruzione di un pubblico aveva la priorità sul difficile compito di vendere servizi che spesso non erano ovviamente utili, era imperativo che tali servizi fossero gratuiti nel punto di utilizzo, lasciando la pubblicità mirata e la sorveglianza degli stessi come uno dei pochi modelli di business sostenibili e sostenibili.¹⁹ Questo percorso si basa su una speculazione finanziaria che i dati alla fine ripagheranno ma, indipendentemente dall'eventuale correttezza di tale speculazione, una volta che un'azienda avvierà questo percorso, avere più dati sugli utenti non può che essere un bene, e la tecnologia disponibile ha reso possibile l'assemblaggio di vasti set di dati.

In terzo luogo, dal punto di vista del profitto, non vi è dubbio che i big data possano funzionare. In effetti, da questo punto di vista il profitto (o, più in generale, l'efficienza) è l'unica metrica che conta: “L'imperativo di correlazione capitalista è chiaro: la correlazione spuria va bene, purché sia correlazione spuria redditizia”.²¹ Questo porta a una presa di posizione sulla conoscenza che non è preoccupata della spiegazione: proprio come non ho bisogno di capire come funziona il mio telefono per usarlo, le aziende e i governi non hanno bisogno di capire come o perché funzionano gli algoritmi per gestirli o, comunque, non si interessano finché gli algoritmi continuano a funzionare bene.²²

Big Data come metodologia

Questo per una spiegazione schematica dell'ascesa dei big data. In termini pratici in cosa consistono i big data come metodo? Dato il modo in cui il termine viene utilizzato, non è facile precisarlo, ma data la rilevanza della quantità di dati nell'approccio, l'accento è posto su metodi di analisi dei dati su larga scala, di vario tipo statistico. Va sottolineato che questo orientamento pone immediatamente i dati davanti alla teoria, poiché i dati e il mondo che si presume illuminino vengono prima di qualsiasi considerazione delle domande da affrontare. In ogni caso parlando in modo molto generale, i modelli statistici sono adatti ai dati per identificare i fattori che tengono conto delle variazioni dei dati. A dire il vero, ora è disponibile una gamma più ampia di metodi rispetto a qualsiasi altro momento, inclusi l'apprendimento automatico, il data mining e i metodi di riconoscimento dei modelli, insieme a più approcci esplorativi, in particolare la visualizzazione interattiva.²³

Nel caso ideale, un singolo ricercatore, analista o un piccolo team di analisti potrebbe utilizzare metodi di visualizzazione esplorativa per acquisire familiarità con un set di dati. Questo a sua volta potrebbe spingere alcune idee sui pattern che si possono trovare, i metodi che più probabilmente enfatizzano quei pattern e da lì i modelli statistici più adatti a far progredire la comprensione dei fenomeni rappresentati dal set di dati. Per molti scienziati che lavorano oggi con set di dati più grandi che mai, questa è una descrizione ragionevole di come procederebbero. Inoltre, non è molto nuovo o diverso da come avrebbero proceduto in passato. Ma il volume e la velocità dei dati contano. Quando i set di dati sono molto grandi e in rapida evoluzione, lo spazio per un approccio esplorativo è limitato, poiché le esigenze computazionali sono potenzialmente illimitate. I problemi devono essere resi tracciabili predeterminando ciò che costituisce un modello di interesse in termini di dati noti per essere disponibili. In ambienti aziendali o di altro tipo in cui la “intelligence utilizzabile” è premiata, molto probabilmente vincolata, prevarranno probabilmente approcci automatizzati. In questi contesti, gran parte del processo decisionale su quali schemi seguire devono essere delegati alle statistiche diagnostiche generate da qualunque metodo venga distribuito. In un pezzo visionario Stan Openshaw (con una certa eccitazione) ha anticipato il tipo di monitoraggio

continuo che questo approccio implica, insieme a un grave svantaggio: l'identificazione di molti schemi e correlazioni spurie.²⁴

Tali considerazioni richiedono che i criteri che specifichino ciò che è interessante e ciò che non sia ulteriormente ristretto, anticipando la ricerca aperta di modelli che potrebbero in primo luogo ispirare la raccolta di dati estesi. Sarebbe assurdo sostenere che non vi è alcun potenziale in dati più dettagliati e aggiornati più frequentemente per descrivere meglio come si sviluppa il cambiamento sociale nel tempo. Forse fenomeni sociali precedentemente sconosciuti possono essere osservati come risultato della migliore risoluzione temporale in tali dati. Le analogie vaghe con i progressi resi possibili in altri campi dall'invenzione del telescopio o del microscopio non sembrano del tutto errate.²⁵

Soprattutto quando i big data vengono usati insieme ad altri approcci, ci sono sicuramente motivi per (cauto) ottimismo riguardo alle sue possibilità scientifiche sociali. In un contesto specificamente geografico, Miller e Goodchild identificano alcuni degli aspetti interessanti di questo approccio e indicano interessanti continuità con precedenti lavori in geografia quantitativa.²⁶

La sfida di lunga data della geografia di creare un ponte tra il locale, il particolare, il globale e il generale è centrale nella loro argomentazione secondo cui l'approccio ai big data, se distribuito con cura e attenzione, promette una "geografia basata sui dati", in particolare all'inizio della fase di scoperta della ricerca quando l'obiettivo è quello di sviluppare buone idee e spiegazioni potenziali.

Anche in questo caso, un punto più grande qui è che i metodi associati ai big data partono dal livello aggregato e implementano un approccio statistico per identificare le relazioni tra gli attributi dei dati in modi più o meno tradizionali. La modalità di spiegazione è statistico-inferenziale, basata su un modello di congiunzione costante di causalità, piuttosto che su un resoconto realista, meccanicistico o orientato al processo.²⁷

Gli attuali set di dati di grandi dimensioni, in particolare quelli che vengono aggiornati di frequente, danno un'impressione di dinamismo e, per estensione, possono essere considerati in grado di offrirci una ricca rappresentazione del processo. In verità questa è poco più che un'impressione. I set di dati di grandi dimensioni, anche se aggiornati di frequente, non rappresentano alcun concetto di processo. A volte è implicito che la velocità di tali set di dati, la loro diffusione e la frequenza di aggiornamento in qualche modo catturino il processo. In realtà, la maggior parte di tali dati sta semplicemente aggiornando rapidamente le istantanee degli eventi. Nulla è registrato nei dati dei processi o meccanismi che guidano le modifiche che si verificano nei dati. Il processo e il cambiamento sono quindi resi come "una cosa dopo l'altra", senza alcuna nozione di processo o meccanismo nei dati stessi.

Invece, i dati impongono concetti piuttosto rigidi di identità su persone e luoghi, riducendo il processo dal "divenire" al mero cambiamento nei valori degli attributi associati a entità sociali altrimenti invariate (sia individui che istituzioni). Questa distinzione, e la necessità di un diverso approccio ai dati che prendere seriamente il processo, il tempo e il cambiamento comporta, è stata per molti anni un punto focale nella comunità delle scienze dell'informazione geografica.²⁸

È difficile concepire qualsiasi mezzo con cui il processo possa essere adattato ai big data sotto forma di dati. Piuttosto è nell'uso creativo e nella pratica, attraverso l'analisi da una prospettiva teoricamente informata, che i concetti di processo vengono aggiunti ai dati. Ironicamente, tale analisi teoricamente informata è uno degli approcci più fortemente evitati dai sostenitori più accaniti dei big data.

Calcolo biforcuto

Indipendentemente dai suoi limiti epistemologici e dalla cecità sui processi, i big data sono chiaramente in ascesa, almeno per ora. Per i miei attuali scopi è istruttivo considerare i big data come gli ultimi sviluppi nella distribuzione del calcolo nel commercio e nel governo e prestare particolare attenzione agli approcci computazionali alternativi per la comprensione del mondo che al momento non sono così fortemente diffusi. Le origini del calcolo moderno risiedono nella seconda guerra mondiale.²⁹ Quel contesto ha visto i computer e la ricerca operativa strettamente

correlata applicata alla soluzione della produzione pratica e alle sfide logistiche della crescente guerra moderna su larga scala.³⁰ Di particolare importanza per lo sviluppo dell'informatica sono stati i difficili problemi matematici sorti in questi contesti, come la decodifica di messaggi cifrati, i complessi problemi di ottimizzazione e la simulazione di reazioni nucleari.

Dalle origini dei computer possiamo identificare due ampie applicazioni. Le prime sono le applicazioni di calcolo a set di dati troppo grandi per il calcolo manuale, per produrre soluzioni in forma chiusa a problemi matematicamente ben definiti, utilizzando vari tipi di analisi numerica.

Tali calcoli si basano su algoritmi per la manipolazione di matrici di grandi dimensioni, su metodi di interpolazione e approssimazione e sulla matematica dell'algebra lineare e dell'ottimizzazione.³¹

Sebbene di routine, tale calcolo richiede più risorse di elaborazione, in particolare quando le dimensioni dei set di dati aumentano, quando i requisiti computazionali associati possono crescere con le potenze quadrate, cubiche o anche di ordine superiore delle dimensioni del problema.³²

In linea di principio, tali calcoli non sono difficili, ma sono intensivi dal punto di vista computazionale. Questo dominio di calcolo è associato alla gestione più efficiente dei sistemi logistici, all'ottimizzazione dell'allocazione delle risorse nei sistemi di produzione e al campo generalmente noto come ricerca operativa.³³ Il fenomeno dei big data si colloca esattamente in questa tradizione.

È possibile identificare una seconda ampia area di applicazioni in cui il calcolo viene applicato in modo iterativo per simulare in prospettiva sistemi reali o ipotetici. In questo caso le applicazioni stesse potrebbero non essere molto impegnative per i dati, ma eseguire ripetutamente (spesso) calcoli semplici nel tempo e / o nello spazio porta a complesse richieste computazionali nell'insieme. Tale calcolo è impiegato nella simulazione di sistemi target di interesse, in meteorologia, progettazione di prodotti, applicazioni militari (simulatori di volo e di combattimento) e, forse più familiari a un pubblico generale, nei giochi per computer.³⁴

È utile considerare questi due approcci computazionali come si sono sviluppati in una particolare disciplina (geografia) per fornire un resoconto più specifico delle differenze tra loro. La rivoluzione quantitativa della geografia ha visto la sua biforcazione in qualche modo correlata, nella divergenza tra metodi teorici orientati al modello e applicazioni più pragmatiche di statistiche inferenziali a dati primari e secondari, esemplificata da testi come "Statistical Analysis in Geography" di King.³⁵

Non passò molto tempo prima che le particolari sfide dei dati spaziali complicassero e compromettessero considerevolmente quest'ultima impresa.³⁶

Nel frattempo, al di là di alcune aree e sottocampi specializzati come la modellistica urbana, l'approccio orientato al modello ha trovato solo un'accettazione limitata prima dei drammatici sconvolgimenti epistemologici degli anni '70 e '80. Come osserva Thrift, questa "ghettizzazione della teoria della complessità in geografia è stata una tragedia, poiché c'erano le potenzialità per un'interazione molto più ampia".³⁷

L'emarginazione degli approcci orientati alla complessità all'interno della geografia quantitativa, è istruttiva, poiché sottolinea la misura in cui qualsiasi tentativo di mappare approcci metodologici su predisposizioni politiche o di altro tipo è destinato al fallimento.

Anche dal punto di vista computazionale la distinzione che sto disegnando è in qualche modo artificiale, dal momento che molti degli stessi strumenti e algoritmi computazionali sono ugualmente applicabili ai big data o alla scienza della complessità, il punto dell'informatica generale è proprio la sua natura versatile e riprogrammabile. La verità è più complicata e sfumata di quanto suggerirebbe qualsiasi semplice account binario. Da un lato, la simulazione dipende da calcoli ripetitivi, spesso di routine, non realizzabili a mano. Dall'altro, l'applicazione di soluzioni in formato chiuso a piccoli insiemi di dati può consentire l'esplorazione iterativa e interattiva di molte possibili soluzioni e portare da lì al concetto di spazio di soluzione e, in definitiva, a una comprensione più sfumata del problema originale. Quando l'analisi numerica viene applicata in questo modo, può trasformare le domande poste ai dati. Questa posizione più esplorativa nei confronti dei set di dati ha avviato la tendenza alla visualizzazione interattiva di set di dati più grandi e più complicati.

Due culture della computazione?

È allettante tentare di mappare questi due stili computazionali (soluzioni chiuse ed esplorazione a tempo indeterminato) direttamente su due manifestazioni culturali dell'informatica: autoritaria, corporativa, statista, da grande fratello, big data da un lato e liberatoria, empowering individuale, personal computing dall'altra.³⁸

Il dualismo è profondamente radicato in molti resoconti della storia dell'informatica.³⁹ La dualità è particolarmente enfatizzata dalle start-up "rivoluzionarie" o "dirompenti" autodidatte della Silicon Valley, che implementano la loro versione della politica contro-culturale degli anni '60.⁴⁰

Queste contraddizioni sono brillantemente sezionate da Turner nel suo libro "From Counterculture to Cyberculture", e molte delle contraddizioni dell'élite autocoscientemente liberale (spesso libertaria) ma conservatrice sono raccontate in modo divertente nel tuttora durevolmente rilevante "Cyberselfish" di Borsook.⁴¹

Un simile scontro di culture è evidente nel contrasto tra un olismo new age nella lingua e nell'iconografia attorno al caos e alla scienza della complessità e le agende più autoritarie, di istituzione e di affari che finanziano e consumano questa scienza.⁴²

Quindi un importante centro di scienza della complessità è stato il Santa Fe Institute (SFI), che secondo Helmreich "a volte è considerato il gemello buono di Los Alamos, interessato alla tecnologia della vita, piuttosto che alla tecnologia della morte", e stabilito in parte attraverso gli sforzi di George Cowan, ex direttore del laboratorio di Los Alamos.⁴³

O ancora: "la maggior parte degli scienziati della SFI sono diffidenti nei confronti di qualsiasi associazione con i movimenti New Age", eppure libri come "At Home in the Universe" di Kauffman si sforzano (e falliscono nella maggior parte dei casi) a rimanere sul lato scientifico di una linea sorprendentemente sottile tra bizzarra New Age e rigore scientifico distaccato quando si tratta delle più grandiose affermazioni della scienza della complessità.⁴⁴

Tali strane contraddizioni intellettuali (se non culturali) possono essere un risultato inevitabile quando metodi scientifici riduttivi, così bravi a spiegare fenomeni senza tempo, almeno dal punto di vista umano (ad esempio, il sistema solare, l'evoluzione), vengono applicati a sistemi in cui le modalità storiche di spiegazione, con le relative contingenze ed eventi casuali, sono state predominante.

Una mappatura semplicistica delle origini culturali delle due tradizioni computazionali in discussione su particolari programmi politici o economici è chiaramente insostenibile. Studi scientifici e tecnologici in numerosi settori hanno ripetutamente e in modo convincente dimostrato la natura altamente contingente delle relazioni tra le tecnologie e le politiche che tali tecnologie incorporano e producono.⁴⁵

Quindi, mentre è allettante suggerire che la modellazione orientata verso la complessità e dal basso verso l'alto è inesorabilmente associata ad approcci anti-autoritari e più aperti alla conoscenza, mentre gli approcci statistici big data, top-down, classificatori e inferenziali sono allineati con gli interessi del potere, ciò è manifestamente falso. Non c'è nulla di intrinseco in nessuno dei due approcci che determina i fini ai quali possono o devono essere distribuiti. Il calcolo in forma chiusa potrebbe essere utilizzato per ottimizzare la produzione efficiente e l'equa distribuzione di servizi medici o pubblici, mentre la simulazione può essere (e quasi sicuramente è stata) utilizzata per esplorare possibili strategie per l'invasione e l'occupazione illegale di un altro paese.

Epistemologie computazionali contrastanti

Anche se non c'è una mappatura uno a uno dell'approccio computazionale su particolari ordini del giorno politici o economici, dovrebbe tuttavia essere chiaro che questi due ampi approcci come pratiche scientifiche distintive incorporano una diversa concezione del processo. Essi inoltre tendono verso atteggiamenti contrastanti sull'uso del calcolo nel promuovere la comprensione dei sistemi socioeconomici, politici e culturali. In effetti, qui è la mia tesi che dovremmo scegliere quali

approcci all'uso del calcolo hanno maggiori probabilità di far progredire la nostra comprensione del mondo e quindi adottarli per quel motivo.

È probabile che tale pragmatismo vedrebbe la deformazione dei big data a favore di modelli di complessità e altri approcci computazionali più in sintonia con il processo e la spiegazione, come nelle discipline umanistiche digitali.⁴⁶

Questo contrasto tra scienza della complessità e big data è schematicamente elaborato nella tabella 2-1.

Tabella 2-1 Approcci diversi alla scienza della complessità e ai big data

Complexity science	Big data
Theory embedded in models	Correlation and classification
Process	Temporal snapshots
Open-ended exploration of process implications	Exploration of already-collected data
Bottom-up orientation	Top-down orientation
Multiple levels and scales	Two levels (aggregate and individual)
Many alternative histories (or futures)	"Just the facts" (or optimal solutions)

Gli approcci orientati alla complessità e basati su modelli riguardano esattamente il processo. È una rappresentazione del processo (comunque limitata) che guida le dinamiche di tali modelli e un'indagine aperta sui comportamenti dei modelli può essere considerata un'esplorazione delle condizioni di possibilità del sistema da modellare. Prima di eccitarsi troppo, si dovrebbe riconoscere che in molti casi la nozione di processo incorporata in tali modelli non è significativamente più ricca di quella implicita dalle istantanee storiche dei big data. Il cambiamento viene spesso eseguito in termini di modifica dei valori di attributi di entità altrimenti fisse e stabili.

Tuttavia, l'attenzione è rivolta al cambiamento, così come alle circostanze che producono cambiamento, una prospettiva che obbliga gli utenti di modelli a considerare direttamente processi e meccanismi.

Interessanti strutture modello che combinano sia il cambiamento di attributo che il cambiamento strutturale sistemico sono un possibile progresso in questo senso.⁴⁷

L'esplorazione a tempo indeterminato di modelli dinamici genera un atteggiamento diverso, più umile e più provvisorio nei confronti della conoscenza rispetto alla predeterminazione e quindi all'individuazione di soluzioni "ottimali" o pattern interessanti.⁴⁸

Il calcolo simulativo pone "mondi possibili" (sotto forma di risultati di simulazione in diversi scenari o configurazioni di modello) e riconosce implicitamente la natura speculativa dell'esercizio.⁴⁹ La modellistica computerizzata speculativa in questa ottica ha portato al riconoscimento, in alcune delle scienze matematiche (e anche in altre discipline), dei limiti alla conoscenza e alla previsione, sotto forma di effetti dinamici come il caos, e di proprietà come l'emergenza, la dipendenza dal percorso, il feedback positivo e l'adattabilità, tutti elementi che probabilmente precluderanno una previsione affidabile di un comportamento del sistema nel tempo.⁵⁰ Il riconoscimento di queste caratteristiche di sistema nel quadro della complessità della scienza richiede, almeno potenzialmente, una maggiore umiltà da parte degli scienziati e un riconoscimento dei limiti di conoscenza inerenti alla natura dei sistemi oggetto di studio.⁵¹

Capita in questo modo, la scienza della complessità sottoscrive un approccio pluralistico alla conoscenza che riconosce l'importanza di comprendere i sistemi a più livelli, da più prospettive, e utilizzando una varietà di metodi.⁵²

Riconoscendo che i sistemi sociali sono composti da individui complessi organizzati in famiglie che svolgono molteplici ruoli in una serie di istituzioni di varie strutture organizzative con una serie di scopi e traguardi, si dovrebbe realizzare, quando si guarda attraverso una lente scientifica di

complessità, che nessuna comprensione capillare top-down di come funzionerà la società è possibile. Inoltre, è probabile che diversi metodi siano appropriati per attendere a ciò che sta accadendo in ciascuno di questi diversi contesti.⁵³

È importante notare che i big data e la scienza della complessità non sono così lontani l'uno dall'altro come sembrano inizialmente. Entrambi riguardano l'adattamento di semplici modelli alle osservazioni: modelli statistici derivati da dati osservati da un lato e modelli di simulazione sintetica dall'altro. Allo stesso tempo, le differenze importanti nell'epistemologia delineate qui sono reali.

Un approccio modellistico orientato alla complessità della conoscenza ci consente di pensare ai dati non come prove concrete e precise della realtà ma come un insieme di modelli che limitano uno spazio di modelli plausibili e speculativi la cui struttura e meccanismi possono rendere conto di tali pattern e che possono quindi essere utile nella costruzione di spiegazioni teoriche orientate al processo dell'esistenza di tali pattern.⁵⁴

I dati in questo contesto diventano un passaggio intermedio nello sviluppo delle spiegazioni. Al contrario, secondo il modello più spesso adottato nel mondo dei big data, i dati stessi sono i fenomeni, e la spiegazione è meno sulla comprensione dei processi e dei meccanismi, cioè sulla spiegazione del mondo, e più sulla descrizione dei dati, e a quel punto si presume che i fenomeni stessi siano stati compresi.

Conclusioni

Come ho cercato di dimostrare, un orientamento al processo è assente dall'epistemologia dei big data, ma è sicuramente centrale per qualsiasi approccio coerente alla spiegazione geografica e delle scienze sociali in generale. Altri approcci computazionali offrono di più in questo senso, ma sono stati meno importanti negli ultimi anni. Tutto ciò potrebbe essere sintomatico delle mode intellettuali della scienza. Il caos e la teoria della complessità hanno avuto entrambi il loro posto al sole, e "complessità" rimane una buzzword molto usata. Forse questi approcci sono stati provati ma non sono riusciti a mantenere la promessa iniziale, come suggerisco che probabilmente anche i big data faranno.

Potrebbe esserci del vero in questa visione, anche se dipende da una prospettiva strana e attenta alla moda su come dovremmo valutare il metodo scientifico. Più seriamente, non riesce a capire quanto sia grande una sfida alle modalità convenzionali di spiegazione scientifica nel prendere sul serio le incertezze introdotte dalla complessità. L'impresa della complessità indica un mandato molto più ampio per i modi di spiegazione storici e narrativi, che vanno contro la natura dei modi dominanti di spiegazione scientifica. La tensione è aggravata dal mantra delle "regole semplici, comportamento complicato", così spesso usato per vendere la scienza della complessità, che presenta modelli semplici come punto di arrivo di approcci orientati alla complessità quando in realtà sono solo l'inizio. Così come è sciocco credere che il mining di big data possa fornire risposte a ogni domanda di scienze sociali, sarebbe sciocco sostenere che i semplici modelli scientifici di complessità possono rispondere a tutte le domande.

Riconoscere e valorizzare il pluralismo nei metodi è la chiave degli approcci computazionali orientati alla complessità che preferisco. Ciò implica due cose. Innanzitutto c'è ovviamente posto per i big data.⁵⁵ Sarebbe assurdo discutere diversamente. Senza dubbio quando i set di dati contemporanei e i metodi di data mining vengono applicati a domande di autentico interesse socio-scientifico, verranno identificati nuovi fenomeni e emergeranno nuove prospettive su vecchie domande. Ma comprendere questi nuovi fenomeni richiederà approcci diversi da quelli dei big data. Ciò porta immediatamente al secondo punto: che altri approcci alla geografia e alle scienze sociali rimangono vitali per qualsiasi strada coerente. Ciò che disturba gran parte dell'hype nei confronti dei big data è l'apparente desiderio di avanzare su tutti i fronti contemporaneamente: i big data, non contenti di essere un approccio rivoluzionario alle scienze sociali, devono diventare un intero sistema di vita, una scelta di vita sociale, una nuova modalità di governance, di affari e di scienza.

La tragedia è che questa posizione nei confronti dei big data potrebbe facilmente screditare tutti gli approcci computazionali verso il mondo sociale, non solo facendo commettere errori alla scienza

ma anche diventando un sistema pervasivo di sorveglianza sociale, la cui necessità non è chiara, al di là del desiderio di profitto di grandi interessi corporativi e ansia dei dati di uno stato di sorveglianza.⁵⁶ La psicologia dei big data fa una promessa del tutto falsa: che se solo i dati fossero più grandi, ne sapremmo ancora di più. Esistono certamente contesti in cui ciò potrebbe essere vero (ad esempio la Square Kilometer Array in astronomia), ma non sono quelli sociali.⁵⁷ Gran parte di ciò che viene rivelato dai big data sociali ci è già noto o ci possiamo arrivare in altri modi che collocano azioni e decisioni umane in contesti sociali molto più ricchi. È probabile che perderemmo ben poco genuino interesse scientifico *non* registrando e memorizzando ogni microinterazione persona-macchina. Non ci sono motivi scientifici per “collezionare tutto”, solo imperativi commerciali (e persino quelli basati su una speculazione selvaggia), il che ci riporta alle importanti questioni politico-economiche menzionate all’inizio di questo capitolo.⁵⁸

C’è un pericolo nel focalizzarsi sul metodo come ho fatto io in questo capitolo, sull’enfatizzare i mezzi anziché i fini. Sono le posizioni etiche, non le scelte metodologiche, che incidono maggiormente sull’impatto della ricerca.⁵⁹ Alla fine i due non possono essere districati e se si vuole realizzare l’indubbio potenziale dei metodi computazionali nelle scienze sociali, è importante scoprire cosa si può imparare dagli errori del passato, riconoscere i limiti di tutti i nostri dati e concentrarsi sullo sviluppo di approcci computazionali alla geografia e alle scienze sociali che si allineino meglio alla gestione di tali problemi. Una tale scienza sociale non sarà solo una migliore scienza in quanto scienza, ma sarà anche più eticamente difendibile come risultato diretto del riconoscimento dei limiti esplicativi dei dati.

Insomma, non dovremmo solo sfidare l’economia politica dei big data ma dovremmo essere profondamente (e sonoramente) diffidenti nei confronti della sua epistemologia, non solo all’interno delle tradizioni critiche scettiche della quantificazione, ma anche dal punto di vista dei metodi quantitativi e computazionali più interessanti a disposizione.

Note

1. “Big data” è una frase problematica da usare correttamente data la sua mescolanza intenzionale (e fastidiosa) di un aggettivo singolare e un sostantivo numerabile plurale - “numerosi dati” sarebbe stato un conio più accurato, seppure meno convincente. Inoltre, la frase “big data” è diventata una combinazione complessa di tecnologie, idee e pratiche, in modo tale da poter essere considerata una frase singolare. In un forse inutile tentativo di trattenere l’ondata di offese contro la grammatica, dove considero il significato di “molti dati”, considero i “dati” al plurale, mentre quando il significato è grande dato come un’idea, lo considero singolare .
2. Weinberger, *Too Big to Know*; Mayer-Schönberger and Cukier, *Big Data: A Revolution*.
3. McChesney, *Digital Disconnect*; Pariser, *Filter Bubble*; Lanier, *You Are Not a Gadget*; Lanier, *Who Owns the Future?*; Morozov, *To Save Everything*.
4. Farmer and Pozdnoukhov, “Building Streaming GIScience.”
5. Kitchin and Dodge, *Code/Space*.
6. Barnes, “Not Only . . . But Also”; Barnes and Sheppard, “Nothing Includes Everything”; Burrows and Savage, “After the Crisis”; Johnston et al., “Mutual Misunderstanding and Avoidance”; Wily, “Strategic Positivism.”
7. Coveney and Highfield, *Frontiers of Complexity*; Manson, “Simplifying Complexity”; Mitchell, *Complexity: A Guided Tour*; O’Sullivan, “Complexity Science and Human Geography”; Thrift, “Place of Complexity”; Waldrop, *Complexity: The Emerging Science*.
8. Questi articoli sono apparsi nel numero di settembre 2008 di *Nature* e nel numero di febbraio-marzo 2010 dell’*Economist*.
9. Radio NZ, <http://www.radionz.co.nz/national/programmes/bigdata>
10. Per informazioni sugli investimenti strutturali, vedi “Fast Broadband: New Zealand’s Internet Upgrade,” Ministry of Business, Innovation, and Employment, <http://www.med.govt.nz/sectors-industries/technology-communication/fast-broadband>.

Un'impressione dell'alta priorità data ai big data nella scienza e innovazione nazionali è data dai report de the New Zealand Data Futures Forum, <http://www.nzdatafutures.org.nz/>.

11. Come Kitchin, Data Revolution, ci ricorda che dovremmo aspettarci.
12. Laney, 3D Data Management.
13. Boyd and Crawford, "Critical Questions for Big Data."
14. Anderson, "End of Theory."
15. Wyly, "New Quantitative Revolution."
16. Pigliucci, "End of Theory in Science?"
17. Wyly, "Automated (Post)Positivism"; Wilson, "Location-Based Services."
18. Brin and Page, "Anatomy of a Large-Scale Hypertextual Web," 18.
19. Delle spiegazioni su come le intenzioni siano andate alla deriva sono fornite in Zuckerman, "Internet's Original Sin."
20. Wilson, "Location-Based Services."
21. Wyly, "Automated (Post)Positivism," 681 (enfasi originale).
22. Vedi Flood, Fires; e il più vecchio Wallace and Wallace, Plague on Your Houses, su cui si basa Floods.
23. Vedi Miller e Han, Geographic Data Mining, per una survey di questi sviluppi in geografia.
24. Openshaw, "Two Explanatory Space-Time-Attribute Pattern Analysers."
25. Brynjolfsson and McAfee, "Big Data Boom."
26. H. Miller and Goodchild, "Data-Driven Geography."
27. Sayer, Method in Social Science.
28. O'Sullivan, "Geographical Information Science."
29. Ceruzzi, History of Modern Computing; Dyson, Turing's Cathedral.
30. DeLanda, War in the Age of Intelligent Machines, offre una panoramica storica critica, mentre Shrader, History of Operations Research, dà un'idea dell'ampiezza del contesto in un resoconto storico ufficiale in tre volumi.
31. Aspray and Gunderloy, "Early Computing and Numerical Analysis."
32. Cormen, Algorithms Unlocked.
33. Light, From Warfare to Welfare.
34. Crogan, Gameplay Mode.
35. Chorley and Haggett, Models in Geography; King, Statistical Analysis in Geography.
36. Gould, "Statistix Inferens."
37. Thrift, "Place of Complexity," 60n2.
38. È una forte tentazione se vivete nella California del Nord.
39. Vedi, per esempio Barbrook, Imaginary Futures; and Levy, Hackers, o vedete lo spot Apple durante il Super Bowl del 1984; see Golumbia, Cultural Logic of Computation.
40. Frank, Conquest of Cool.
41. Turner, From Counterculture to Cyberculture; Borsook, Cyberselfish.
42. Thrift, "Place of Complexity."
43. Helmreich, Silicon Second Nature, 43. Los Alamos è naturalmente base di uno dei laboratori statunitensi più strettamente associati con gli sviluppi passati e attuali di armi nucleari; vedi Waldrop, Complexity, 53–69.
44. Helmreich, Silicon Second Nature, 41; Kauffman, At Home in the Universe.
45. Feenberg, Critical Theory of Technology; Latour and Woolgar, Laboratory Life; Winner, Whale and the Reactor.
46. Burdick et al., Digital Humanities.
47. Gross and Blasius, "Adaptive Coevolutionary Networks."
48. Naturalmente la scienza di costruzione dei modelli non è sempre un esercizio aperto. Una quantità considerevole di teoria economica è costruita sull'inversione della sequenza di costruzione dei modelli e esplorazione che porta al raffinamento di teorie testabili.

Invece i modelli forniscono supporto retorico per posizioni teoriche già saldamente stabilite e sono raffinati per adattarsi alla teoria, piuttosto che il contrario. Nel processo i modelli diventano più distaccati dalla realtà, non più vicini; vedi Lawson, *Economics and Reality*; e Keen, *Debunking Economics*. Mi preoccupo qui di un approccio all'uso di modelli di simulazione discusso in un mio libro recente; vedi O'Sullivan and Perry, *Spatial Simulation*. L'uso appropriato della simulazione in scienza è un'area filosofica difficile, che rimane poco esplorata dai filosofi della scienza, ma vedi Winsberg, *Science in the Age of Computer Simulation*; e Weisberg, *Simulation and Similarity*.

49. Casti, *Would-Be Worlds*.

50. Gleick, *Chaos*.

51. Coveney and Highfield, *Frontiers of Complexity*; Manson, "Simplifying Complexity"; O'Sullivan, "Complexity Science and Human Geography"; Cilliers, *Complexity and Postmodernism*; Richardson, Cilliers, and Lissack, "Complexity Science."

52. Harvey and Reed, "Social Science."

53. Manson and O'Sullivan, "Complexity Theory in the Study of Space and Place."

54. Grimm et al., "Pattern-Oriented Modeling."

55. Miller and Goodchild, "Data-Driven Geography."

56. Lazer et al., "Parable of Google Flu"; Crampton, "Collect It All"; Crawford, "Anxieties of Big Data."

57. See Taylor, "Square Kilometre Array."

58. Crampton, "Collect It All."

59. Lake, "Methods and Moral Inquiry."