[slide 1]
“Thinking Big Data in Geography”
Traduzione del capitolo 3

[slide 2]
“Small data e slow data in un’era di big data”
Renee Sieber e Matthew Tenney

I big data sono diventati un paradigma per un presunto nuovo tipo di dati e una nuova forma di scoperta della conoscenza.[1] L’iperbole che circonda i big data, anche se non necessariamente vera, ora ha un impatto su quasi tutti gli aspetti della nostra vita sociale e anche sul modo in cui ricercatori e professionisti lavorano con le informazioni digitali.

[slide 3]
Il linguaggio dei big data denota il lavoro di tali ricercatori come “nuovo” contro “vecchio” e “grande” contro “piccolo”.

[slide 4]
Tale impatto colpisce i temi primari della ricerca, promuovendo il ruolo delle persone come “sensori”, della comunicazione come “contenuto” e della vita in generale come “comportamento”.
I fattori in gioco sono numerosi, ad esempio, con mercati redditizi per l’intermediazione di dati che creano una domanda per un ambiente di business “data-now”.[2] I governi, dal livello municipale a quello nazionale, stanno passando da vecchi a nuovi modi di amministrare i servizi pubblici e risorse attraverso, ad esempio, l’analisi dei dati.[3]

[slide 5]
C’è comunque scetticismo che accompagna le iniziative del settore pubblico e privato mentre i big data scavallano “il picco dello hype, e si muovono verso la valle della disillusione”.[4] Sono anche sempre più numerose le critiche che espongono le insidie teoriche e le carenze pratiche dei big data nel produrre risultati significativi o attuabili.[5] In ambito accademico, però, la letteratura recente fa ben poco per intaccare le speranze esagerate riposte nei big data.

[slide 6]
Come osserva Wilson, gli scienziati sociali sono spesso pronti a vedere i big data come una finestra sulla vita quotidiana dei loro creatori, sostenendo che “per molti [del grande pubblico], twittare, pubblicare, ritwittare e condividere è simile alla respirazione.”[6]
Siamo attratti dai big data per un motivo pragmatico. Un autore di questo capitolo ha presentato un articolo di rivista incentrato sulle piattaforme di mappatura Web 2.0 applicate allo sviluppo della comunità locale.

[slide 7]
Secondo un revisore ben intenzionato, il numero di contributi (n <100) era troppo piccolo ed è stato raccolto troppo lentamente per garantire il rigore necessario per fare dichiarazioni informate sui bisogni o sui desideri di una comunità. Questa critica rappresenta una tendenza a inquadrare tecnologie geospaziali e metodologie nelle scienze sociali in modo tale da renderle sempre più conformi agli approcci usati per i big data.

[slide 8]
I processi sociali sono quindi articolati in termini di volume e si considerano validi solo se svolti su piattaforme basate sulla tecnologia cloud.

[slide 9]
Questo tipo di determinismo dei dati, in base al quale la tecnologia determina il modo in cui comprendiamo la società, può portare a una nuova forma di distorsione della ricerca. Le caratteristiche della tecnologia specificatamente in uso prevalgono sul valore dei dati piccoli e lenti e resuscitano dibattiti ricorrenti nel contesto della geografia sul contrasto tra metodi quantitativi e qualitativi.[7]
Questa non vuol essere una critica postpositivista sui metodi quantitativi dei big data. Di fatti, siamo agnostici riguardo al valore dei big data e alle metodologie della data science. Non siamo nemmeno completamente scoraggiati dal ruolo crescente che i dati digitali hanno nell'indagine delle scienze sociali, un fenomeno che può essere studiato e che a sua volta può anche svelare altri fenomeni da studiare.[8] La nostra preoccupazione, però, è che, facendo ricerca di dati solo in relazione ai big data, i data scientist esacerbino le problematiche già presenti all'interno di discipline come la geografia e aiutino lo sviluppo di una scienza sociale ossessionata dai dati. Pensare che i dati debbano essere sempre "big" (ad esempio, caratterizzati da grande volume e alta velocità) genera aspettative di validità e verità e crea una normatività nelle scienze sociali su come la ricerca debba essere condotta.

[slide 10]
Affrontiamo questi problemi discutendo prima le molteplici origini del concetto di big data, l'emergere degli *small data* in risposta ai fallimenti dei big data e la difficile situazione degli small e *slow* data quando sono costretti a integrarsi con l'epistemologia dei big data.

[slide 11]
Big Data, da scopi industriali a scopi industriosi
Secondo Hidalgo, pochissimo dello "hype proviene dalle persone che lavorano davvero con set di dati di grandi dimensioni. Invece, la maggior parte del clamore è scatenato da persone che vedono i "big data" come una buzzword e un'opportunità di marketing: consulenti, organizzatori di eventi e accademici opportunisti in cerca dei loro 15 minuti di fama."[9] Innanzitutto va ribadito che non c'è una definizione universale di big data. Invece, "big data" come termine ha origini diverse e soddisfa molteplici esigenze derivanti dal mondo accademico, industriale e dei media. Mashey è stato forse il primo a spiegare il concetto prima che il termine diventasse comune.[10] Ha parlato di "infrastress" (stress dell'infrastruttura): grandi quantità di dati stavano imponendo richieste eccessive alle infrastrutture di elaborazione esistenti e richiedevano nuovi approcci per l'archiviazione dei dati, nuove strutture di database, nuovi servizi di accesso immediato e analisi sei dati. Per Mashey, i big data non offrivano nuove finestre sul mondo sociale o dei clienti delle aziende.[11] Piuttosto, i big data rappresentavano una sfida tecnica per gli ingegneri dell'hardware e del software informatico. Weiss e Indurkhya hanno usato il termine "big data" nel titolo di un capitolo nel loro libro "Data mining predittivo: una guida pratica".[12] Gli autori, con esperienza nell'intelligenza artificiale, consideravano la gestione dei big data un'opportunità che non era priva di sfide intrinseche: "Raccolte di dati molto grandi (…) vengono ora compilate in data warehouse centralizzati, consentendo agli analisti di utilizzare metodi potenti per esaminare i dati in modo più completo. In teoria, i "big data" possono portare a conclusioni molto più forti per le applicazioni di data mining, ma in pratica sorgono molte difficoltà". [13] Ciò che è iniziato come un problema di gestione dei dati si è presto trasformato in un nuovo modo di concepire i dati.

[slide 12]
Tony Hey e colleghi ricercatori della Microsoft sono arrivati al punto di proclamare l'emergere di un "quarto paradigma" della scienza, in cui i big data richiedono nuovi metodi per affrontare l'attuale "diluvio di dati scientifici".[14] Elwood, Goodchild e Sui spiegano i tre precedenti paradigmi dominanti della scienza del ventunesimo secolo: "l'empirico (descrizione di fenomeni naturali), il

teorico (uso e verifica di modelli e leggi generali) e il computazionale (simulazione di fenomeni complessi usando data set piccoli fittizi, artificiali o dal mondo reale).”[15] Quel diluvio di dati, secondo Chris Anderson della rivista Wired, ha segnato la fine della teoria.[16] Gli attuali modelli basati sulla teoria sono inadeguati e inappropriati di fronte alle opportunità offerte dal volume dei big data. Non abbiamo più bisogno di fare affidamento su ipotesi per scoprire relazioni all'interno di sistemi complessi.[17] Invece ora vogliamo affidarci a una scienza guidata dai dati (data-driven) per esplorare e scoprire relazioni laddove prima non se ne conoscevano. Secondo Jim Gray, un altro dipendente di Microsoft, la scienza basata sui dati è simile a un “macroscopio”, una situazione in cui i problemi di ricerca vengono studiati come se si stesse scrutando milioni di interazioni attraverso un microscopio.[18] Un approccio induttivo e ad alta intensità di dati per la scienza “servirebbe da nuova arca su cui possiamo sopravvivere all'attuale diluvio di big data”. [19]
L'argomento è che i big data differiscono sostanzialmente da un semplice set di dati molto grande e che questa differenza richiede nuovi approcci che si distaccavano dalle tradizionali norme e pratiche scientifiche. Nel diluvio di dati, la precisione è stata offerta al posto dell'accuratezza, che non può essere valutata perché ci sono troppi dati. Fonti di big data come Wikipedia utilizzano il *crowdsourcing* per perfezionare i dati.[20] La spiegazione più di successo della differenza tra set di dati molto grandi e big data è venuta da Laney, che ha riflettuto sulle sue esperienze di prima mano con le sfide della memorizzazione e della gestione dei big data.

[slide 13]
Laney ha caratterizzato questo nuovo tipo di enorme set di dati attraverso quelle che sono diventate note come le tre V essenziali. Ha sostenuto che i set di dati erano molto più voluminosi di prima; c'erano nuove piattaforme su cui i dati venivano continuamente trasmessi in streaming o erano disponibili a velocità molto più elevate.[21] Inoltre, i dati erano ora accessibili e analizzabili come record individuali, al contrario di interi set di dati non scompattabili. Molti di questi dati ora si manifestano in forme altamente non strutturate e disordinate. Il volume, la varietà e la velocità rappresentavano sfide per l'integrazione dei dati e l'interoperabilità del sistema che li gestisce. Oggi diverse teorie sui big data sono occasionalmente accompagnate da una quarta, quinta e persino sesta V come supporto necessario per l'utilità di applicazioni specifiche per big data (ad es. veridicità, vitalità (*viability* in inglese, nel senso di fattibilità), variabilità, valore). È con le sfide (e le opportunità) di queste V che le caratteristiche dei big data hanno iniziato a formare un sostituto di una definizione unica, che illustriamo con un'equazione:

big data = *f* (volume, varietà, velocità, *e in più:* veridicità, vitalità, variabilità, valore)

I big data si sono solidificati attorno a queste prime tre caratteristiche di volume, varietà e velocità.
Restano comunque difficoltà a caratterizzare il fenomeno dei big data. Un recente sondaggio informale della School of Information presso l'Università della California a Berkeley ha chiesto a “più di 40 pensatori leader nel campo dell'editoria, della moda, del cibo, delle automobili, della medicina, del marketing e di altri settori come esattamente definirebbero l'espressione “big data.”[22]

[slide 14]
Non sorprende che i risultati abbiano rivelato quasi quaranta definizioni diverse di big data che ruotavano in gran parte attorno alle esigenze di ciascun partecipante. Ciò che diventa evidente dopo aver esaminato poche risposte è che “big data” è diventata una frase colloquiale che valorizza il potenziale di approfondimenti a grana fine, rilevanti per obiettivi specifici piuttosto che un magico accesso a un set di dati dai risultati generalizzabili e sfruttabili per numerosi e nuovi casi d'uso. Secondo Floridi, molte delle definizioni di big data si basano su ambiguità e ragionamento circolare: i dati sono “grandi” solo in relazione al nostro attuale potere computazionale.[23]

[slide 15]
Come affermano M. Graham e Shelton, il "modificatore" "big" "[è] sempre relativo e rappresenta un obiettivo mobile."[24] Quelli che sono attualmente considerati "small data" erano estremamente "big" mezzo secolo fa, e le nozioni contemporanee dei big data probabilmente saranno minuscole già tra mezzo secolo. Batty sostiene che l'ambiguità concettuale ha portato a un'attenzione comune alla creazione e all'uso di big data che va oltre il semplice focus solo sul loro volume.[25] Essere semplicemente grandi (ad esempio, più grandi di una tabella di Excel) non rende i dati preziosi, non dà loro valore. I big data rimangono quindi "un concetto astratto" che si distingue "dalle masse di dati, [per] altre caratteristiche, che ne determinano la differenza con "dati enormi" o "dati molto grandi"". [26] Inoltre, i big data diventano inestricabili dalle capacità del software e dell'hardware disponibili al momento.[27] Se un insieme di dati supera la capacità di un foglio di calcolo, allora è da considerarsi "big". Va detto però che i big data non sono interamente il prodotto di macchine calcolatrici. Possiamo far risalire le origini dei big data al periodo compreso tra il 1880 e il 1940, quando presso l'Harvard College Observatory mezzo milione di osservazioni del cielo notturno sono state accumulate interamente da esseri umani.[28] Nonostante le ambiguità o forse proprio a causa loro, "big data" è rapidamente passato da essere un termine usato per descrivere la raccolta e la gestione dei dati a uno slogan di marketing che prometteva di migliorare le pratiche commerciali e di individuare con maggiore precisione i clienti.[29] Gli slogan si sono spostati da "Big Data as Boogeyman (i big data come l'uomo nero)", segnale dei primi dubbi sul fatto che i costi potessero superare il potenziale valore dei big data, a "The Big Data Gold Rush (la corsa all'oro dei big data)", in cui ai big data è stata attribuita la creazione di un mercato dei dati del valore di $125 miliardi.[30]

[slide 16]
L'emergere dei big data come industria commerciale li ha presentati come un prodotto che potrebbe essere sfruttato per la business intelligence e come sostegno della data science come potenziale vantaggio competitivo per le aziende che la adottano. I big data pongono inoltre in primo piano il ruolo dell'information technology (IT) nelle aziende, non più una funzione che la dirigenza ha relegato in un dipartimento IT (ad es. buste paga, inventario e proiezioni), bensì una funzione chiave tramite cui un'azienda agile risponde rapidamente a dati in costante evoluzione.[31] L'agilità richiede nuovi investimenti IT per gestire i flussi di dati attraverso, ad esempio, nuove analisi, visualizzazioni e interfacce utente. In questo modo i dati non sono più uno slogan di marketing, ma tornano al loro contesto originale nell'informatica.
Le prime campagne di marketing sono state abbinate a frasi come "sovraccarico di dati" o "infobloat", che descrivevano la scienza dei big data come soluzione a contenuti eccessivi e ingombranti.[32] Al di fuori di alcuni casi aneddotici, i big data continuano a fallire nel fornire informazioni e valore. Un sondaggio condotto su oltre trecento dipartimenti IT ha rilevato che molti grandi progetti orientati ai dati non hanno mai lasciato le fasi di pianificazione, con prove di concetto e prototipi che non riescono a raccogliere valore per le loro imprese.[33]

[slide 17]
Lo stesso sondaggio ha rilevato che la mancanza di conoscenza empirica ha già provocato errori costosi. I risultati del sondaggio hanno rispecchiato una crescente disillusione nella comunità imprenditoriale. Siti di notizie come ZDNet e Forbes hanno suggerito che i big data sono stati presentati in maniera esagerata rispetto ai risultati: "I big data sono difficili (e il dominio di pochi). Usarli su vasta scala e attendere i benefici può richiedere del tempo."[34] Ciò ha ulteriormente contribuito all'ambiguità nelle definizioni dei big data. Le crescenti quantità di dati digitali sono viste come una panacea per le industrie e le attività scientifiche, pur essendo accompagnate da aspettative sempre meno convinte. Per raggiungere il valore promesso dai big data, sembra necessario dover portare qualcosa di più.

Dai big data agli small data

Nonostante le promesse di vari momenti da "eureka!" nell'analisi dei dati o nella loro visualizzazione, i big data non riuscivano a estrarre valore dalla massa di dati. Erano troppo grandi, troppo veloci e troppo eterogenei; erano incomprensibili e impersonali. "Mentre alle aziende (e ai computer!) piacciono i big data, la maggior parte delle persone ha bisogno solo di small data", ha affermato Fidelman, perché, rispetto ai big data, "è più facile analizzare e testare piccoli set di dati per differenziare il segnale utile dal rumore bianco ed estrarne un risultato significativo."[35]

[slide 18]

La segmentazione di big data è stata sempre più spesso vista come una cura per la mancanza di utilità dei big data: "una buona strategia per risolvere la "maledizione dei big data"… è la suddivisione intenzionale di grandi serie di dati in serie di dati più piccole."[36]

[slide 19]

Una definizione è presto emersa dal settore privato per formalizzare il concetto di "small data", che "collega le persone con approfondimenti (*insight*) immediati e significativi (derivati da big data e/o fonti locali), organizzati e confezionati - spesso visivamente - per essere accessibili, comprensibili e praticabili nelle attività quotidiane."[37] Gli small data hanno fatto eco a ambiguità simili riscontrate nei big data, ma almeno gli small data sono stati visti come un modo per mantenere le promesse dei big data senza imporre sforzi straordinari per l'estrazione di valore dai dati. Il concetto di small data è emerso per almeno due motivi. Innanzitutto, gli small data offrono un modo per ricavare valore dai set di dati utilizzando la stessa data science e la stessa analisi dei dati progettate per rivelare il valore nei big data. In secondo luogo, gli small data confermano il primato dei big data nell'inquadrare la totalità dei dati disponibili. Non avremmo dati di piccole dimensioni senza i big data, perché "prima del 2008, i dati erano raramente considerati in termini di "small" o "big". Tutti i dati erano, in effetti, quelli che a volte vengono definiti "small data" indipendentemente dal loro volume."[38] Ci rivolgiamo alla letteratura recente per diverse prospettive dell'emergere degli small data rispetto ai big data.

[slide 20]

L'emergere degli small data ci aiuta a costruire un'epistemologia dei big data, anche se a costo di una parte dell'integrità degli small data:

small data = big data – some data

La prima prospettiva è che gli small data sono semplicemente una porzione digeribile di big data. Gli insight immediati e significativi derivano dall'estrazione ponderata di sottoinsiemi di dati. Spesso questo set di dati più piccolo viene estratto perché risponde a una particolare organizzazione o necessità.[39] Il processo iniziale condotto su qualsiasi big data è quello di ridurre il set di dati in modo significativo.[40] Il processo rappresenta sia una riduzione sia un riconoscimento del fatto che "i dati allo stato brado" non siano mai grezzi.[41]

La riduzione dei big data riduce al minimo il costo della gestione dei dati e presumibilmente quindi massimizza le informazioni da set di dati altrimenti gonfiati e sovradimensionati. In altre parole, l'utilità si ottiene campionando e rimuovendo dati ridondanti, errati e irrilevanti. In questo modo produciamo set di dati utilizzabili per le nostre attività. Secondo Lu e Li, i data scientist conducono raramente analisi su big data; alla fine utilizzano small data: "La maggior parte delle volte, l'accesso diretto a tutti i dati non è né possibile né fattibile dal punto di vista computazionale, costringendo le persone a sondare le proprietà dei dati osservando un campione. A causa dell'enorme dimensione dei dati, spesso anche un campione di dimensioni discrete è troppo costoso da ottenere considerando il traffico di rete coinvolto e le quote giornaliere imposte dalle circostanze. Per considerazioni pratiche, ci dobbiamo spesso limitare al più piccolo campione possibile."[42]

Jacobs fornisce un esempio delle sfide computazionali legate alla gestione dei big data.[43] Il ricercatore ha generato un database sintetico composto da 6,75 miliardi di record a 16 byte che aveva lo scopo di emulare un database basato su un censimento (ad esempio età, religione, reddito e indirizzo) per ciascuna persona sul pianeta. Il valore di tale fonte di dati sarebbe innegabilmente utile per i ricercatori di geografia e altre discipline, ed è stato facile archiviare i record per l'intera popolazione mondiale su un singolo laptop di media fascia del 2009. Jacobs ha sostenuto che l'archiviazione dei dati non è il fattore limitante; la vera sfida è l'analisi di tali dati.[44] Per ricavare informazioni da enormi processi di analisi numerica, in particolare quando i dati hanno dimensioni sia temporali sia spaziali, è necessario che lo scienziato dei dati rispetti la "aggregazione [di dati] dipendente dall'ordine con cui tale aggregazione è avvenuta (per esempio, funzioni cumulative, funzioni a finestra mobile, etc.)."[45] L'accesso casuale dovuto al campionamento nella maggior parte dell'analisi dei big data distrugge i contesti temporali e spaziali dei dati. I dati di piccole dimensioni possono mantenere una topologia laddove i dati di grandi dimensioni non lo fanno. L'esempio di Jacobs mostra una contraddizione ricorrente nell'accumulo di grandi set di dati. Acquisiamo i dati anche se non riusciamo ad accumulare le risorse tecnologiche necessarie per gestire tale problematica "grandezza", e inoltre, anche dovessimo avere tali risorse potremmo non renderci conto del fatto che non le stiamo usando in maniera uniforme su tutti i dati a nostra disposizione. Floridi fa luce sulla sfida paradossale in cui il valore dei big data merita "tecniche e tecnologie maggiori in quantità e migliori in qualità, che" devono ridurre "i big data a dimensioni accettabili".[46] Pertanto questa prima formalizzazione di small data enfatizza la distillazione di dati da controparti più grandi al fine di evitare le limitazioni computazionali esistenti e il sovraccarico analitico. La "grandezza" crea valore solo quando diventa piccola. Anche se le nostre capacità di archiviazione crescono, probabilmente avremo comunque bisogno di "frammentare" i dati per poterli analizzare.

[slide 21]
Qui gli small data diventano un pezzo, un datum che proviene dai big data, che viene curato ed elaborato dalle macchine in base alle esigenze di una specifica operazione che l'individuo vuole eseguire:

small data <= cervello umano

Il fallimento di molti progetti di big data può essere attribuito a una paralisi decisionale in presenza di troppi possibili strumenti da usare, troppe fonti di dati e troppe potenziali applicazioni disponibili per i big data.[47] Tali strumenti sono però essenziali, non si può prescindere da loro poiché l'incomprensibilità è considerata una caratteristica intrinseca dei big data e dobbiamo usare tali strumenti per liberarcene. Una seconda prospettiva sugli small data si riferisce alla loro capacità di migliorare la comprensione dei big data.[48] Qui gli small data vengono espressi come dati di dimensioni sufficientemente piccole per la comprensione umana.[49] Un documento di lavoro di Markowsky utilizza questa definizione antropocentrica di small data per giustificare l'intervento umano nella creazione di un sottoinsieme di big data in modo che "possa essere facilmente compreso dalla mente umana e facilmente visualizzato dall'occhio umano".[50] Gli small data, in questa prospettiva, sono simili alla descrizione di cui sopra del tentativo di rendere i dati a modelli familiari e gestibili di dimensioni ridotte. Invece di fare affidamento esclusivamente su una soluzione tecnologica per ricavare insight sui big data (ovvero, attraverso l'analisi computazionale), questa prospettiva abbraccia un approccio tradizionale all'interpretazione dei dati. Il cervello umano diventa il "computer analitico" piuttosto che dover dipendere esclusivamente da algoritmi o correlazioni statistiche che macinano insiemi di dati altrimenti incomprensibili.[51] Lo scopo di questa caratterizzazione degli small data è di aiutare le persone a utilizzare i big data, in modo che possano ricavare informazioni e verificare le proprie intuizioni. Questo processo si rivela difficile non solo da scorporare e replicare,[52] è anche difficile da condividere, come discuteremo di seguito.

[slide 22]

small data = big data ∩ me

Un diverso restringimento del diluvio digitale identifica come small data solo quei dati che riguardano direttamente la mia persona. Troviamo questa prospettiva nel contesto del movimento del "sé quantificato" (persona digitale), che nasce dall'enorme successo delle tecnologie indossabili e mobili.[53] Dispositivi come Fitbit creano volumi e velocità di contenuto relativamente grandi basati sui singoli individui. Queste sono le tracce digitali generate da quasi tutti gli aspetti della tecnologia che utilizziamo, che possono a loro volta essere analizzate per ricavare insight sui nostri comportamenti individuali. Estrin e il Small Data Lab della Cornell University considerano questo tipo di dati come small data, che "possiamo concepire … come nuovo tipo di prove mediche, prove in cui n = me, perché integrano gli studi tradizionali sulla popolazione (N grande) con dati che riguardano solo me nel tempo."[54]
Rispetto ai big data, i big data ∩ me non sono né anonimi né aggregati. Sono intesi essere comprensibili perché riguardano un individuo specifico e rispecchiano i tentativi di valutare i big data discretizzando i dati in blocchi digeribili. Gli small data acquisiscono così una caratteristica di personalizzazione, lontana dalla collettività dei big data. Questi small data rappresentano una nuova fonte altamente personale di informazioni preziose, in cui i dati arrivano in profondità nel corpo, ad esempio con dispositivi medici integrati e connessi come i pacemaker.
Big data ∩ me rivela un dilemma morale di base per gli small data. Gli individui possono generare i dati, ma questo tipo di small data è in gran parte fuori dalla portata d'uso da parte di questi stessi individui, e i dati sono ulteriormente offuscati da (una mancanza di) diritti sulla proprietà dei dati e considerazioni sulla privacy.[55] Gli small data rappresentano un distacco dai big data in termini di distillazione e comprensibilità dei dati, ma questa stessa prospettiva evidenzia una distinzione in chi è l'utente di tali dati. Nei big data l'utente finale concide con l'analista dei dati. Negli small data, invece, l'utente finale può essere la stessa fonte dei dati (o il raccoglitore di dati, come vedremo con i censimenti nazionali) ma non necessariamente l'analista dei dati. È probabile che gli small data del sé quantificato vengano aggregati tra più individui per essere poi analizzati da un analista. L'analisi e le visualizzazioni di dati sono costruite a partire da tali aggregazioni, che vengono in un momento successivo personalizzate sulla base dei dati di un individuo. Né i dispositivi hardware né il software di un prodotto come Fitbit sarebbero stati mai sviluppati se il tutto fosse dipeso da un singolo individuo. Nonostante la fortissima personalizzazione dei dati, comunque l'analisi e l'estrazione di valore di tali dati necessitano di una aggregazione che diminuisce la personalizzazione ma aumenta il volume e rende più solide le statistiche. In questo contesto iniziamo a vedere l'interazione indissolubile tra grande e piccolo:

[slide 23]

small data = big data / dominio

Una quarta prospettiva sugli small data si riferisce alla riduzione dei big data da parte di domini specifici come ad esempio la geografia. Spesso questa riduzione è correlata a dati sul sé quantificato: "i dati sul consumo di energia della mia famiglia, i tempi degli autobus locali, le spese del governo - questi sono tutti small data".[56] È importante sottolineare che questo tipo di prospettiva rappresenta anche un dominio, ad esempio, dell'energia, dei trasporti, della pubblica amministrazione. Molti di questi flussi di dati contengono geolocalizzazioni esplicite (ad es. ubicazioni delle fermate degli autobus, posizione degli autobus) o implicite (ad es. spese governative, che sono legate alle giurisdizioni locali). Alla riunione del 2012 dell'American Association of Geographers c'è stata una sessione speciale intitolata "I limiti dei big data e il valore

degli studi sugli small data", che ha portato a un numero speciale della rivista scientifica GeoJournal.[57] In questa sessione Goodchild e Kitchin hanno inserito i dati geografici nella categoria degli small data. Il loro esempio principale è il censimento nazionale perché il suo volume assomiglia alla caratteristica più comune dei big data e per il ruolo centrale che un censimento svolge in molte indagini geografiche.[58] Goodchild rafforza la prospettiva che gli small data siano basati sul dominio invece che sul volume. Sostiene che, nel giro di soli due anni, gli small data si sono evoluti dall'agire come proxy per i big data ad essere un termine generale che colloca il "tradizionale approccio geografico" all'interno della pratica degli studi sugli small data: "I big data si distinguono da quello che propongo di definire small data per la loro mancanza di normali processi di controllo qualità, documentazione e rigoroso campionamento… Gli small data, esemplificati dai prodotti del censimento, hanno fornito tutte queste cose, con il risultato che l'analisi degli small data permette una pronta generalizzazione."[59]

[slide 24]

Qui i big data vengono ridotti a un dominio specifico o diventano più comprensibili quando resi compatibili con i metodi di un dominio specifico. In un lavoro successivo H. Miller e Goodchild asseriscono il valore che la geografia apporta ai big data[60]. Sostengono che i geografi possiedono una lunga esperienza con il volume dei dati (ad esempio, con le immagini rilevate a distanza di Landsat), nonché la velocità e la varietà dei dati (ad esempio, con informazioni geografiche volontarie (VGI: Volunteered Geographic Information, contenuti multimediali geolocalizzati da piattaforme di social media; pensate per esempio alle fotografie geolocalizzate che gli utenti pubblicano su Google Maps). I metodi tradizionali sono stati sviluppati durante la rivoluzione quantitativa sul campo, sono sopravvissuti al contraccolpo culturale contro la geografia, e sono rifioriti nel "contro-contraccolpo" fornito dalla scienza dei GIS. Ciò ha reso la disciplina probabilmente meglio preparata rispetto ad altre per impegnarsi con i big data e fondere quell'impegno con ricerche di scienze sociali minori.[61] I metodi tradizionali potrebbero essere applicati a dati geografici più recenti, come VGI, che assomigliano ai big data nella loro "confusione, non essendo strutturati, essendo raccolti senza controllo di qualità e spesso non accompagnati da documentazione o metadati."[62] Questa prospettiva sugli small data apre da un lato le fonti di dati geografici tradizionali (es. censimento nazionale) a nuove tecniche analitiche e dall'altro lato apre nuovi tipi di dati, come VGI, ai metodi geografici tradizionali.
L'implicazione è che le discipline possono affermare la loro rilevanza trasformando i big data in dati più significativi – "domando" i dati. A sua volta, uno specifico dominio della conoscenza acquisisce rilevanza dalla sua associazione con una nuova fonte di valorizzazione. Posizionando i dati di una disciplina in relazione ai big data, si dimostra che una disciplina è equipaggiata per affrontare una nuova fonte di dati ed essere sufficientemente importante per essere ascoltata da altre discipline. Il valore di questo posizionamento nei confronti dei big data all'interno della geografia è stimolato dalle affermazioni della sua capacità unica e potente di scrutare all'interno di sistemi sociali stratificati e complessi. Queste considerazioni sono sostenute da dichiarazioni come la seguente: "immagina, ad esempio, la geografia umana e la ricerca nel campo delle scienze sociali che potrebbero essere intraprese con il set di dati messo insieme dal team del presidente Obama per le sue campagne elettorali del 2008 e del 2012".[63] Hyman sottolinea che la speculazione dei media ci ha fatto sopravvalutare le tecniche di elaborazione dei dati usate durante le campagne elettorali, che, in realtà, sono state attuate con dati relativamente piccoli che potevano essere analizzati con carta e matita.[64] Ciò rispecchia lo hype nel settore privato riguardo alla promessa della scoperta di nuove conoscenze grazie alla combinazione dell'esperienza di un dominio con i big data e la data science. Vogliamo vedere un grande potenziale dei big data e delle sue analisi anche quando in realtà potrebbero non esistere.
Goodchild ha presentato gli small data come dati caratterizzati da controllo della qualità, da documentazione adeguata e rigore.[65] Questo rispecchia Kitchin e Lauriault, che offrono una

definizione formale degli small data in cui "gli small data sono (…) caratterizzati da un volume generalmente limitato, da una raccolta non continua, da una varietà ristretta e di solito sono generati per rispondere a domande specifiche."[66] In questo utilizzo degli small data c'è una generale mancanza di distinzione tra dati e informazioni, o poca chiaezza su quale ruolo abbiano i dati nelle varie tecniche di raccolta, analisi e utilizzo nell'ambito dell'attuale ricerca geografica. Nonostante questa omissione, gli autori evidenziano una differenza critica tra dati big e small: la maggior parte dei dati prima della decantata "grandezza" avevano uno scopo ed erano organizzati con un preciso intento. Gli small data sono dati goal-oriented, creati con obiettivi specifici. Discuteremo più avanti che questi obiettivi, una delle diverse caratteristiche distintive degli small data, possono disperdersi e quindi danneggiare lo status degli small data.

[slide 25]

Più significativo per noi è il fatto che questa definizione consolidi i legami intrinseci tra small data e big data, poiché i primi sono definiti con i modificatori dei secondi e, come riconosciuto dagli autori, sono compatibili con la scienza e le pratiche dei big data:

> in primo luogo, nonostante la rapida crescita dei big data e delle analisi di dati a loro associate, gli small data continueranno a prosperare perché hanno una comprovata capacità di rispondere a domande specifiche. In secondo luogo, i dati di questi studi saranno sempre più raggruppati, collegati tra loro e scalati attraverso nuove infrastrutture di dati, con un impulso verso l'armonizzazione degli small data con gli standard, i formati, i metadati e la documentazione dei big data, al fine di aumentare il loro valore attraverso la loro aggregazione e condivisione. In terzo luogo, il ridimensionamento degli small data li espone alle nuove epistemologie della data science e al loro inserimento in nuovi mercati multimiliardari sviluppati da broker di dati, con il rischio di coinvogimento in pratiche potenzialmente dannose come la data surveillance, la classificazione sociale, il control creep e la governance preventiva, ossia pratiche per le quali gli small data non stati originariamente concepiti.[67]

La definizione precedente evidenzia un'ultima prospettiva sugli small data: che gli small data non sono legati ai big data ma possono fungere da parametro di input per le tecniche di analisi dei big data:

small data ≠ big data, ma valore = *big_data_analytics*(small data)

La nostra conclusione è che queste diverse prospettive sugli small data, piuttosto che offrire visioni diverse sui dati, riaffermano invece il discorso sui big data. Gli small data sono più comprensibili, possiedono più rigore e così via, specialmente quando tali dati ci riguardano. Tuttavia, anche gli small data sono posizionati rispetto ai big data con il presumibile scopo di raccogliere tutti i vantaggi derivanti dalle tecniche dei big data. Non appena proviamo a definire big data e small data, portiamo alla luce un problema circolare: i big data trovano valore solo quando diventano small, ma gli small data, secondo alcuni, ottengono valore solo quando vengono riassemblati in qualcosa che assomiglia ai big data.

[slide 26]

Sostanzialmente giriamo avanti e indietro tra dati grandi e piccoli e di conseguenza non otteniamo chiarezza su nessuno dei due gruppi. Le prospettive degli small data attirano anche l'attenzione sulla lunga relazione tra un dominio come la geografia e le tecnologie di Internet. Small data come dati georeferenziati rappresentano (da un lato) una separazione da e (dall'altro lato) un inserimento

in un'epistemologia dei big data in cui (dal primo lato) un censimento può arricchire un set di dati con un obiettivo, ma (dal secondo lato) le V multiple dei big data iniziano a essere importanti per tutti i tipi di dati (anche gli small). In queste inclusioni di data che sono nominalmente big all'interno degli small data, i proponenti di queste definizioni finiscono per ridefinire la geografia, separando i dati e le pratiche geografiche del passato dai principi alla base dei dati futuri. Tutti i dati, in particolare i dati geografici, diventano parte dei big data.

[slide 27]

Quando gli small data non sono la risposta, a prescindere dalle loro dimensioni

Maggiore potenza e controllo, rotture col passato e nuovi insight sono fattori che accompagnano spesso i concetti di big data e small data. Haklay, Singleton e Parker identificano il modo in cui i neologismi, in particolare quelli associati a Internet, sono comuni in molti campi di ricerca che tentano di affermare la propria legittimità allineandosi alle tematiche dominanti.[68] Per noi, i neologismi servono come abbreviazione di epistemologia, un modo per raggiungere la verità che, per i big data, sta nella loro capacità di essere valorizzati (ad esempio, monetizzati). Per definizione, la maggior parte dei neologismi è benigna o passa inosservata. Occasionalmente, l'inserimento di un neologismo nel mainstream può portare a impianti concettuali poco produttivi. Sosteniamo che il neologismo dei big data può fallire perché è ambiguo, spesso deliberatamente. In gran parte tale ambiguità deriva da una mancanza di contesto e di intenti, difetto che è presumibilmente corretto dagli small data. Allo stesso modo, gli small data possono far perdere di vista il loro intento originario e trascurare la diversità di prospettiva tra i ricercatori nella loro comprensione teorica e metodologica dei dati. In uno speciale numero di GeoJournal sui big data e la geografia, i curatori Burns e Thatcher descrivono nel loro editoriale le conseguenze di un neologismo che fornisce un inquadramento tutt'altro che produttivo: "Nell'organizzazione di questo numero, è diventato chiaro che anche tra un piccolo gruppo di autori che lavorano su un unico concetto da noi suggerito, i big data, la loro influenza sulla società e il loro significato nella vita quotidiana differiranno radicalmente a seconda dei focus della ricerca considerati dagli autori come importanti, distinti o superflui. Ciò che un autore mostra come un problema fondamentale per l'epistemologia derivante dall'analisi dei big data, è per un altro autore semplicemente un prerequisito per iniziare a lavorare su un altro obiettivo."[69]

Sia i big data sia gli small data comportano perdita di informazioni

[slide 28]

Il volume è la parte più importante del neologismo dei big data. Se le dimensioni contano in questo contesto neologismo, l'adagio "più è meglio (*more is better*)" esprime l'omaggio che gli small data devono pagare ai big data. Kitchin spiega che i big data rivendicano uno spazio di osservazione esaustivo in cui vengono catturate intere popolazioni, in contrasto con le strategie di campionamento pianificate rappresentative degli small data.[70] Tuttavia, catturare intere popolazioni oggi non è più realistico nemmeno per i big data, a causa della limitazione sugli accessi, l'inevitabile selection bias nella raccolta dei dati e numerosi altri fattori (ad es. digital divide tra i potenziali contributori, ontologie diverse delle fonti di dati online). Questa ossessione per le dimensioni è forse radicata in una confusione di base tra dati e informazione. Wu è uno dei principali data scientist di un'azienda di big data ma rimane scettico. Spiega l'inganno di "più è meglio": "È vero che i dati ti danno informazioni, ma il grande errore dei big data è pensare che un maggior numero di dati ti darà proporzionalmente più informazioni. In effetti, più dati hai, meno informazioni ottieni in proporzione ai dati. Ciò significa che le informazioni che è possibile estrarre da qualsiasi insieme di big data diminuiscono in modo asintotico all'aumentare del volume di dati."[71]

[slide 29]

Paradossalmente, più dati si hanno, più informazioni si possono perdere. Il segnale può essere sommerso dal rumore e dai bias. Gli small data presumibilmente offrono una maggiore comprensione contestuale – sono dati orientati al cervello umano – e quindi potrebbero ridurre la perdita di informazioni. Tuttavia, è possibile perdere informazioni nell'applicare il neologismo dei big data agli small data. La sezione precedente menziona come Jacobs ha dettagliato la potenziale perdita di informazioni nell'analisi del censo perché l'analisi non può mantenere la topologia dei dati.[72]
Ciò vale sia che i dati siano big sia che siano small (ricordiamo che alcuni dati del censimento sono considerati small). Lo spazio di analisi potrebbe essere ancora insufficiente per l'attività preposta. Se gli small data vengono campionati o "analizzati" in modo casuale in modo simile ai big data, allora qualsiasi struttura sottostante (ad es. l'ordine di registrazione dei dati) probabilmente verrà distrutta.
I big data comportano la perdita di informazioni sul loro ciclo di vita, principalmente a causa della necessità di essere riusati. Tuttavia, la perdita di informazioni sul ciclo di vita della raccolta dei dati può avvenire anche con dati di piccole dimensioni. Armstrong e Armstrong, ad esempio, hanno criticato l'approccio del Canada alla raccolta di dati censuari nazionali.[73] Hanno raccomandato di riesaminare i dati dal punto di vista di coloro che tali dati avrebbero dovuto rappresentare e far luce sulla relazione tra i dati e le ipotesi teoriche fatte durante le varie fasi del ciclo di vita dei dati stessi. Ciò corrisponde a una patologia tipica dei big data, in cui ci si dimentica troppo spesso che il contesto della creazione di un dato è importante tanto quanto il dato stesso. Come afferma Snickars nella sua critica del data mining utilizzato da aziende come YouTube, "se il contenuto è re, allora il contesto è la sua corona".[74]

[slide 30]
Gli small data possono perdere la verifica con la stessa facilità dei big data
Small data come quelli di un censimento cercano di essere esaustivi in termini di acquisizione di dati demografici sociali su intere popolazioni in determinati periodi di tempo. I dati del censimento mancano della velocità e della varietà per essere considerati big data; tali dati sono inoltre limitati in termini di accesso poiché la loro disponibilità è limitata ai profili campionati. Per proteggere la privacy dei cittadini, Statistics Canada limita la segnalazione di alcune caratteristiche geodemografiche a un campione del 20 percento e fornisce la geodemografia in forma aggregata a meno che non ci sia un'autorizzazione speciale. Il vantaggio di questo set di dati ufficiale e autorevole non risiede necessariamente nel suo resoconto verificato sulla popolazione ma nel fatto che la raccolta di dati sia controllata e irregimentata. Ciò è coerente con le caratteristiche degli small data basati su un dominio specifico, con le proprie regole internamente coerenti in materia di controllo di qualità, documentazione e rigore. Queste regole offrono mezzi concreti per rilevare possibili distorsioni dei campioni e/o dell'intero set di dati, con la possibilità di correggere errori sistematici.
L'elevazione del censimento a quintessenza degli small data implica che questo tipo di fonte di dati sia considerato un resoconto più verificabile e autentico degli insight sulla società rispetto ai big data. Sarebbe molto difficile ottenere livelli simili di granularità geografica o demografica attraverso i social media di maggior successo, per numerose ragioni: l'accesso limitato ai dati proprietari (di cui sono in possesso le aziende private che gestiscono i social), l'intrinseca incertezza legata a informazioni di profilo non verificabili. Wilson illustra la natura non verificabile dei dati dei social media con uno scherzo trasformatosi in notizia virale sulla morte dell'attore Morgan Freeman.[75] Come Wilson ci ricorda, la verità non è un prerequisito dei big data, ma l'errata attribuzione di autenticità è possibile anche negli small data.[76] La quinta affiliazione religiosa più comune riferita al censimento nel Regno Unito è stata "Jedi Knight". Il caso di centinaia di migliaia di cavalieri Jedi che abitano nelle isole britanniche è forse più facilmente individuabile rispetto ai bias nei big data, dove gli adolescenti segnalano erroneamente la loro età per aggirare le restrizioni su determinati servizi web. Qui c'è un paradosso. Probabilmente ci sono più richieste di accuratezza

sugli small data rispetto ai big data a causa dell'architettura di controllo specifica del dominio che caratterizza i primi. Eppure H. Miller e altri sostengono che gli small data trarrebbero beneficio da un'analisi indipendente dal dominio e basata sui nuovi metodi di qualità dei dati ideati per i big data. Questo stratagemma, però, come detto sopra ostacola l'obiettivo di utilizzare il cervello umano come computer analitico piuttosto che dipendere esclusivamente dagli algoritmi.

[slide 31]
I ricercatori finiscono per diventare studiosi di big data
L'indipendenza o meno di un dominio da altre discipline implica una trasformazione della conoscenza in tale dominio. La ricerca sugli small data potrebbe trasformare gli studiosi in studiosi junior di big data. Man mano che gli small data vengono potenziati, Kitchin vede nuove opportunità per la data science e una maggiore disponibilità di finanziamenti per la ricerca.[77] Ciò rispecchia l'emergere della GIScience (o scienza dei GIS), menzionata in precedenza. GIScience, emergendo da un contraccolpo accademico sulle implicazioni culturali della geografia, oltre che da un dibattito "strumento o scienza" sui GIS, è diventata una disciplina incentrata sulla questione se il posizionamento dei GIS come scienza conferiva maggiore legittimità alla ricerca.[78]
I GIS come forma di uso di uno strumento tecnologico rischiano di essere visti come inferiori a una vera e propria GIScience. L'uso di strumenti è dominio dei professionisti praticanti, mentre un'etichetta a connotazione più scientifica potrebbe portare a una migliore posizionamento in ambito accademico, con la promessa di pubblicazioni più prestigiose, sovvenzioni maggiori e assunzioni più frequenti. Trasformare i big data in una scienza e quindi posizionare gli small data all'interno di big data potrebbe presumibilmente far ottenere benefici simili a quelli di GIScience. Vediamo già la pubblicizzazione di nuove posizioni accademiche nella scienza dei dati geospaziali. Se siamo in grado di aggregare piccoli set di dati, per eseguire analisi con il data mining, gli small data potrebbero ottenere un valore rinnovato nell'accademia.
Secondo Kitchin e Lauriault, "I dati di questi studi saranno sempre più raggruppati, collegati e ridimensionati attraverso nuove infrastrutture di dati, con un impulso verso l'armonizzazione degli small data rispetto a standard, formati, metadati e documentazione dei dati."[79] Le esortazioni all'armonizzazione dei dati non comportano automaticamente una reale armonizzazione, spesso perché queste forme digitali di aggregazione standardizzata possono entrare in conflitto con le culture istituzionali. Da un punto di vista culturale, gli ideali della torre d'avorio possono seguire le virtù democratiche propagandate dai sostenitori della condivisione dei dati. Quella stessa cultura può punire l'aggregazione di dati. Il motto di "pubblicare o perire (*publish or perish*)" rimane profondamente radicato nella cultura della ricerca. In un'università sempre più neoliberista, che inietta valori capitalisti come la concorrenza nel mondo accademico, la condivisione dei dati può significare che un ricercatore perde un'opportunità di avanzamento di carriera e un incremento della sicurezza del lavoro. In effetti, la stessa strutturazione dei dati può far parte del lavoro di ricerca: "Gli scienziati ora hanno ampia scelta quando si tratta di formati di dati. In effetti, è abbastanza comune per i ricercatori inventare nuovi formati per ogni nuova tecnica e addirittura per ogni nuovo esperimento. Ciò rende notevolmente più difficile il lavoro di integrazione di set di dati di grandi dimensioni."[80]
Trevor Garrett è il principale ricercatore del progetto nazionale olandese per la creazione di un'infrastruttura internazionale di condivisione dei dati.[81] Egli sostiene che il ridimensionamento efficace attraverso le strutture di dati assomiglia a una sorta di pensiero magico. Le infrastrutture sono fortemente desiderate ma in realtà non riescono nemmeno ad avvicinarsi ai loro obiettivi dichiarati. Il ministro delle finanze del Canada ha divulgato informazioni su un progetto da $15,7 milioni finanziato dai contribuenti per costruire un "repository digitale affidabile per i record, ma a causa di un cambiamento nell'approccio alla raccolta dei dati, tale repository non è mai stato utilizzato".[82] L'obiettivo era quello di raccogliere dati governativi risalendo fino al 1890, ma l'host del repository, Library and Archives Canada, ha attualmente un arretrato di quasi centomila scatole, alcune delle quali non sono state toccate da più di vent'anni. Le funzioni di ricerca del repository

sono segnalate come inefficienti, il che è particolarmente problematico per quanto riguarda le informazioni sul vergognoso sistema scolastico residenziale per i nativi indigeni del Canada, informazioni fondamentalmente necessarie per la Commissione per la verità e la riconciliazione del Canada. Il "pensiero magico" ha pervaso i preparativi per l'archiviazione dei documenti cartacei, ma il Canada deve ancora elaborare una strategia per gestire l'arrivo imminente di documenti esclusivamente digitali.

Si sa di più sul perché le persone si rifiutino di condividere i propri dati di quanto si sappia sul perché dovrebbero condividerli. Wallis, Rolando e Borgman hanno esaminato gli utenti e i collaboratori di una piattaforma di condivisione di una rete di sensori, in cui la maggiore preoccupazione dei partecipanti era la perdita di informazioni nell'aggregazione di dati: l'aggregazione su questo portale separava i dati dal contesto della documentazione che avrebbe consentito la corretta attribuzione ai contributori originali dei dati stessi.[83] Quando è avvenuta condivisione di dati, è successo prevalentemente nelle interazioni tra persona e persona e non attraverso infrastrutture digitali impersonali. Gli autori hanno confermato che esistono pochi incentivi istituzionali per rendere i dati interoperabili e quindi utilizzare dati condivisi. Quando c'è poco incentivo a condividere i dati, è difficile prevedere il sostegno finanziario per un'infrastruttura per rendere interoperabile la "ricchezza e varianza che è probabile che esiste in parti della lunga coda della ricerca scientifica e tecnologica."[84]

Sia che si tratti di small data o di big data, consentire l'interoperabilità può richiedere un profondo cambiamento in ciò che viene valutato nel processo di ricerca. Una spinta verso l'interoperabilità può spostare l'attenzione all'interno del paradigma linguaggio dei mezzi/fini. Invece di utilizzare i dati come mezzo per generare risultati, diventano un fine a se stesso. Abbiamo assistito a questo cambiamento nell'implementazione dei GIS, in cui i dati hanno da tempo acquisito un valore che è indipendente dai motivi della loro generazione.[85] Una lotta nel contesto dei GIS è quella per documentare i dati in modo sufficiente da conservare la memoria istituzionale della loro provenienza, classificazione e intenzione. Le difficoltà nella creazione e nella conservazione dei metadati spaziali sono note da tempo; l'automazione non ha migliorato in maniera significativa la loro raccolta.[86] C'è anche la sfida di preparare i dati in un modo che ne permetta il riuso da parte di un pubblico sconosciuto e per scopi non ancora noti. Il riuso è un presupposto cruciale per i big data, ma il perseguimento di tale scopo può spostare risorse dall'uso vero e proprio dei dati alla loro preparazione. In realtà, la maggior parte dei ricercatori e dei professionisti non sono intesi essere produttori di dati (ovvero, produrre dati per il bene dei dati) ma collezionisti di dati, in situazioni in cui i dati aiutano a raggiungere obiettivi predefiniti.

[slide 32]

La situazione critica dei dati molto piccoli e lenti

Crediamo che in un futuro non lontano vedremo un notevole impatto sui progetti locali, di dimensioni ridotte e con risultati lenti da raggiungere, a causa della pressione per passare ai big data e alla relativa data science. Queste attività rischiano di essere trasferite a ciò che chiamiamo, per mancanza di una frase migliore, "studi di dati molto piccoli e lenti (*very small and slow data*)". Sosteniamo che i dati molto piccoli e lenti non siano necessariamente sovvertiti dai big data ma sono costretti a integrare gli approcci ai big data. Questo riallineamento si verifica su numerosi fronti, compresa la creazione di aspettative di avere una "grandezza" per il proprio set di dati, che diventa rappresentante dell'importanza di uno studio, dei suoi dati e del rigore dei suoi metodi, nonché una maggiore possibilità di accesso alle risorse (ad es. finanziamenti riservati a studi di tipo big data).

Dati molto piccoli e lenti possono essere considerati parte del processo dei metodi qualitativi delle scienze sociali, come casi di studio, relazioni etnografiche o resoconti biografici. Questi set di dati potrebbero essere normativi, ad esempio, quando esplorano aspetti della giustizia sociale. I dati molto piccoli e lenti sono la scala a cui vengono raccolti molti dati nelle scienze sociali. I dati tendono ad essere altamente dettagliati e richiedono lunghi periodi di tempo per essere raccolti

perché presumibilmente offrono riflessioni sfumate e relazioni topologiche profonde, sono incorporati in contesti storici e antropologici e, presumibilmente, sono comprensibili da parte degli esseri umani anche senza l'uso di algoritmi. Secondo Ballantyne, questi tipi di studi dovrebbero descrivere il disordine di ciò che accade sul campo, per essere poi distillati in storie attraverso le quali spieghiamo i dati della nostra ricerca.[87] Alcuni sostengono che i big data ci consentono di sfuggire a un'era di scarsità di dati scarsi in modo da poter vivere in ambienti ricchi di dati.[88]
Un'altra prospettiva è la promessa che i nostri scarsi archivi di dati possano descrivere ambienti ricchi di informazione. Il punto di questa visione è che il valore si trova proprio nel rumore che viene scartato dai big data per ottenere il segnale significativo. Dati molto piccoli e lenti possono derivare dal ragionamento che si svolge nella codifica dei valori-chiave in molti approcci di analisi del contenuto, in altre parole, possono essere il prodotto dei processi di pensiero dei ricercatori nel determinare la loro "scelta di ciò che è più reale".[89] Pur sostenendo un approccio molto "small e slow" alla raccolta dei dati, anche se fatto in modo digitale, non sosteniamo automaticamente che gli unici dati validi siano i più piccoli e lenti. Numerose ragioni precludono questo tipo di studi (ad es. obiettivi di studio incompatibili, vincoli di risorse, obiezioni dei partecipanti). La nostra definizione di dati molto piccoli e lenti è imperfetta e soggetta alle stesse critiche che abbiamo discusso sopra. Scegliamo il termine come una provocazione e semplicemente mettiamo in discussione il desiderio di sottoporre tutti i dati alle ipotesi incorporate nel neologismo dei big data.[90]
Facciamo affidamento sulla prima equazione,
big data = f (volume, varietà, velocità e forse veridicità, vitalità, variabilità, valore ...)
e sulle V in essa contenute per suggerire alcuni modi in cui questo spostamento verso i big data è particolarmente nocivo per gli small e slow data.

[slide 33]
Posizionamento normativo attraverso le dimensioni
Dati molto piccoli e lenti mettono in grande rilievo le ipotesi su come il volume dei big data posizioni in maniera normativa la ricerca delle scienze sociali su piccola scala. Poiché il volume può essere misurato numericamente e ordinalmente, incorpora in maniera intrinseca il concetto di gerarchia. Più grande è meglio (*bigger is better*). Qualsiasi sistema gerarchico o accoppiamento dicotomico presuppone un'etica, in cui una scelta è strumentalmente superiore a un'altra scelta, o in cui uno "dovrebbe" fare una scelta piuttosto che un'altra (cioè, ci sono pratiche giuste e pratiche sbagliate). La piccolezza riflette anche il tipo di dati rappresentati. Facendo riferimento all'aneddoto introduttivo sul Web 2.0 per lo sviluppo di una comunità sociale, un piccolo numero di osservazioni VGI potrebbe essere considerato errato rispetto a un gran numero di osservazioni, in parte perché le osservazioni VGI sono asserite da utenti e non provenienti dallo studio di esperti. Implicitamente, set di dati molto piccoli e lenti richiederebbero un rafforzamento, sia imponendo l'accuratezza (minimizzazione della distanza tra misurazione e valore reale) sia, in un modello di crowdsourcing tipo Wikipedia, la precisione (minimizzazione della distanza tra misurazioni ripetute). Solo con questa stratificazione, questo accrescimento di asserzioni, un approccio può acquisire valore. Potrebbe non esserci alcun perfezionamento, tuttavia la quantità costruisce validità perché permette maggiore precisione. In un'epistemologia dei big data in cui i modi di conoscere sono collegati a un gran numero di contributi che si triangolano l'uno con l'altro, si presume che la precisione corregga gli errori.

Un'ipotesi più ampia riguarda il modo in cui la dimensione dei big data (o degli small data, come previsto con insiemi di dati molto grandi come il censimento di un paese), ci convince che con volume, portata e scalabilità possiamo ottenere nuovi insight. Rispetto ai dati grandi e piccoli, i dati molto piccoli e lenti, a meno che non vengano aggregati, limitano la nostra capacità di massimizzare gli insight.

[slide 34]
Hardt svela i pregiudizi dei big data nel suo articolo, "Quanto sono scorretti (*unfair*) i big data." Hardt traccia i metodi con cui i big data possono diluire il punto di vista delle minoranze, che sono statisticamente sopraffatte dal volume delle opinioni della maggioranza. Invece di raggiungere la lunga coda dell'opinione pubblica, i big data possono tradursi in una regressione a una media "bianca (in senso razziale)" che diluisce le voci della minoranza pur dando l'impressione che tali voci siano ascoltate.[91] Ciò si riferisce all'osservazione di Elwood e Leszczynski che troppo spesso confondiamo disponibilità con accesso.[92] Solo perché chiunque può partecipare a una piattaforma di social media non significa che tutti parteciperanno. Contrariamente al presupposto che i big data sono neutrali mentre i dati molto piccoli e lenti sono distorti, sia i big data sia le relative tecniche di analisi offrono uno specchio sociale dei nostri pregiudizi:

> Dato che siamo in procinto di utilizzare l'apprendimento automatico per rendere praticamente tutti i tipi di decisioni sugli esseri umani in settori come l'istruzione, l'occupazione, la pubblicità, l'assistenza sanitaria e la polizia, è importante capire perché l'apprendimento automatico non è di default né imparziale né giusto.
> Ciò è in contrasto con il diffuso equivoco che le decisioni algoritmiche tendano ad essere giuste, perché, sai, la matematica riguarda le equazioni e non il colore della pelle.[93]

Velocità piuttosto che sfumature
Gli small data, rispetto ai big data, soffrono di una mancanza di velocità "in tempo reale" sia nella loro creazione che nella loro raccolta. Dati molto piccoli e lenti possono permetterci di osservare come cose buone arrivino a coloro che aspettano con pazienza. Spesso tale attesa si rende necessaria, che i ricercatori lo vogliano o meno.

[slide 35]
Una caratteristica della raccolta di dati molto piccoli e lenti è la costruzione della fiducia tra il ricercatore e i soggetti della ricerca (da non confondere con la fiducia [*trust*] calcolata algoritmicamente utilizzata in molti progetti di social network e big data). Per raccogliere dati molto piccoli e lenti da interviste approfondite, è necessario concedere tempo per coltivare una relazione e una fiducia personali. Allo stesso modo, i numeri degli intervistati possono essere di piccole dimensioni con al massimo una dozzina di intervistati; le interviste possono avere periodicità irregolare o addirittura del tutto inesistente (interviste singole o una sequenza lungo mesi), possedere una relazionalità debole ed essere limitate nella varietà (ad esempio, solo trascrizioni di testo). Invece di supporre che i dati molto piccoli e lenti siano deboli negli insight, le informazioni raccolte da questi tipi di dati possono essere riccamente strutturate e supportate da metodi rigorosi. Allo stesso tempo, gli insight derivati dalla lentezza di alcuni metodi possono essere incompatibili con la nostra era ad accesso istantaneo, in cui i dati vengono costantemente aggiornati in un flusso continuo. Numerose situazioni potrebbero richiedere un accesso immediato. Potremmo, ad esempio, salvare vite umane grazie alla rapidità offerta dal rilevamento delle crisi da parte dei cittadini.[94] Tuttavia, nel conformarci alle ipotesi dei big data, rischiamo di abbandonare lo studio lento preferendo il rapido ma superficiale.

Armonizzare i dati più piccoli e lenti
Infine, consideriamo la varietà. Gli sforzi necessari per mantenere il valore di dati molto piccoli e lenti in un futuro di big data possono obbligare i ricercatori a garantire che i loro dati siano collegabili e scalabili, come nel caso degli small data. L'ipotesi alla base dell'armonizzazione è che i dati acquisiscono valore nella loro aggregazione. L'idea espressa in una frase sarebbe "ogni pixel non utilizzato è un pixel sprecato." Se i dati esistono solo nel silos proprietario di un rapporto di ricerca, non riescono a raggiungere il loro potenziale. Perché tali dati non dovrebbero essere

riutilizzati? Poiché la condivisione e il riutilizzo, in particolare il riutilizzo digitale, sono diventati intrinseci alla ricerca attuale, queste domande diventano un attacco *ad hominem*. Che cosa ha l'obiettore contro il riutilizzo dei dati, specialmente se quel riutilizzo genera nuove conoscenze? L'idea di base è che il ricercatore è immorale per non aver tentato di estrarre più approfondimenti dagli studi preesistenti, se gli insight possono essere accumulati in combinazione con altri set di dati. Una chiara espressione di moralità sta nell'attribuzione delle proprietà salvavita ai dati collegati: "Esempi di potere dei dati collegati sorgono quotidianamente. In Gran Bretagna, il Times ha raccolto dati grezzi ma linkati tra loro sugli incidenti in bicicletta da DirectGov e ha pubblicato una mappa che mostra dove si sono verificati gli incidenti in bicicletta, in modo che i ciclisti potessero essere consapevoli dei punti pericolosi lungo le strade della città."[95]

[slide 36]
Allora dati molto piccoli e lenti possono essere espressi così:

big data = n * (very small and slow data)

dove n è la soglia alla quale i dati diventano legittimi nell'epistemologia dei big data. Tuttavia, i dati molto piccoli e lenti ci consentono di considerare il contrario: "perché non dovremmo sprecare il pixel?" Alcuni dati non possono e non devono essere riutilizzati. I dati "sacri" esemplificano la natura conflittuale della condivisione dei dati.

[slide 37]
Rundstrom scrisse che in molte culture indigene alcune conoscenze potevano essere conosciute solo da un piccolo numero di persone (ad esempio, gli anziani).[96] Altri non avevano diritto a quella conoscenza. Alcuni gruppi accetterebbero la perdita della conoscenza indigena se non ci fossero più anziani piuttosto che consentire la registrazione di tale conoscenza. In un altro caso, un gruppo indigeno consentirebbe di distruggere il suo sito sacro piuttosto che consentirne la mappatura e l'esposizione di tale conoscenza a un pubblico più vasto. L'ipotesi che alcuni set di dati debbano essere persi e non essere riproposti o mai resi pubblici viola l'etica secondo cui tutti i dati dovrebbero essere disponibili per essere collegati e aggregati.
Quando conduciamo ricerche con o su popolazioni emarginate in studi di dati molto piccoli e lenti, spesso mettiamo la nostra ricerca in un contesto critico. Questi includono la posizionalità e la soggettività nei confronti degli individui con cui e su cui conduciamo ricerche (ad esempio, "uno degli autori è una donna bianca di classe media che co-conduce ricerche con popolazioni indigene, che in realtà sono un sottoinsieme di un più ampio gruppo indigeno"). Tecnicamente, un approccio armonizzato e collegato ai dati può allegare questi dettagli ai dati estratti a causa del loro polimorfismo (ad esempio, un file da allegare a un singolo record nel database). Un collegamento iniziale, tuttavia, non può garantire che tale collegamento venga successivamente mantenuto. Un collegamento può anche escludere una revisione etica. In effetti, considerazioni etiche potrebbero non consentire un riuso dei dati. Sia che stiamo collegando o aggregando, potremmo perdere molto nell'armonizzazione di dati molto piccoli e lenti. Forse alcuni set di dati non dovrebbero essere ridimensionati.

[slide 38]
Conclusione
In questo capitolo siamo passati dai big data agli small data ai very small and slow data e viceversa. Questa struttura ci consente di meditare sulla retorica altalenante dei big data. Vale a dire, i big data sono troppo grandi o veloci per essere compresi o gestiti a livello computazionale. Non riescono a produrre valore come pubblicizzato, quindi dobbiamo ridurre i dati a dimensioni gestibili.[97] I dati piccoli, molto piccoli e lenti offrono valore attraverso la raccolta di dati mirata, ma rischiano di essere considerati irrilevanti per le nuove tecniche di analisi, di visualizzazione e, in definitiva, per

la creazione di insights. Di conseguenza, siamo invitati a utilizzare varie aggregazioni per ridimensionare gli small data ai big data. Ma i dati risultanti potrebbero perdere il loro contesto e diventare troppo grandi per essere compresi. Quindi li dobbiamo ridurre…ripetete la retorica secondo le vostre necessità.

Non siamo contro il valore di nessuna dimensione (ovvero, le V suggerite dai big data) di un set di dati rispetto a un altro. Invece sosteniamo che l'iperbole dei big data permea tutti i dati. Indipendentemente dalla dimensione dei dati, la tentazione è quella di posizionare tutti i dati all'interno delle opportunità – gli insight e le nuove valorizzazioni – offerte dai big data. Generalizzare una dimensione di dati come rappresentativa di tutta la ricerca nelle scienze sociali travisa la natura degli small e very small data e il valore di tutte le dimensioni dei dati nel futuro di una disciplina come la geografia. Anche considerare gli small data come una rappresentazione unaria della geografia genera una riduzione filosofica mal concepita della disciplina. Le rappresentazioni unarie tradizionale e i neologismi servono più a offuscare artificialmente il nostro lavoro che a chiarire il futuro della disciplina. Speravamo di dimostrarlo con il conio ironico del nostro stesso termine: "dati molto piccoli e lenti".

Prevediamo un numero sempre maggiore di richieste per il riposizionamento di dati molto piccoli e lenti come dati compatibili con gli small data, che insieme verranno riposizionati come contribuenti ai big data. Preferiamo l'accettazione di diversi set di dati e di approcci diversi, anche quelli con alcuni dati mai archiviati, condivisi o collegati. In questo capitolo abbiamo cercato di sgonfiare la tracotanza dei big data che sono prematuramente collegati agli small data, i quali probabilmente incontreranno difficoltà nel realizzare le affermazioni esagerate sull'utilità di assomigliare ai "big". Sollecitiamo, insieme a molti altri, la moderazione nell'adottare tali epistemologie per le discipline delle scienze sociali come la geografia, perché mancano di indirizzare questioni più grandi che potrebbero far deragliare la pertinenza dei big data nel futuro della ricerca nelle scienze sociali. Adottando l'accettazione pluralistica di diverse dimensioni dei dati in geografia, dovremmo cercare di bilanciare in qualche modo critici da un lato e opportunisti da un altro, alla maniera di M. Graham e Shelton: "Riteniamo che una conversazione più ampia sul meme dei big data e sui modi con cui è in grado di reindirizzare e spostare l'attenzione, la conversazione, le risorse e le pratiche lontano da altri problemi urgenti non solo ci consentirà di evitare le implicazioni più problematiche dei big data ma anche di lavorare per un'integrazione più produttiva dei big data con i paradigmi di ricerca esistenti."[98]

Note

1. Ruppert, "Rethinking Empirical Social Sciences."
2. Simon, Too Big to Ignore.
3. Schmidt and Cohen, New Digital Age.
4. Buytendijk, Hype Cycle.
5. See Lerman, "Big Data"; Kaisler et al., "Big Data"; and Lazer et al., "Parable of Google Flu."
6. Wilson, "Morgan Freeman Is Dead," 345.
7. E. Sheppard, "Quantitative Geography."
8. Wilson, "Morgan Freeman Is Dead."
9. Hidalgo, "Saving Big Data."
10. Mashey, "Big Data."
11. Mashey, "Big Data."
12. Weiss and Indurkhya, Predictive Data Mining.
13. Weiss and Indurkhya, Predictive Data Mining, as quoted in Diebold, "On the Origin(s)."
14. Hey, Tansley, and Tolle, Fourth Paradigm, 1.
15. Elwood, Goodchild, and Sui, "Prospects for VGI Research," 371.
16. Anderson, "End of Theory."
17. Anderson, "End of Theory"; Bar-Yam, "Limits of Phenomenology."

18. Hey, Tansley, and Tolle, Fourth Paradigm, 223–24.
19. Elwood, Goodchild, and Sui, “Prospects for VGI Research,” 371.
20. Raymond, Cathedral and Bazaar.
21. Laney, 3D Data Management.
22. Dutcher, “What Is Big Data?”
23. Floridi, “Big Data.”
24. M. Graham and Shelton, “Geography and the Future of Big Data,” 256.
25. Batty, “Big Data, Smart Cities and City Planning.”
26. M. Chen, Mao, and Liu, “Big Data: A Survey.”
27. Wilson, “Morgan Freeman Is Dead.”
28. Nelson, “Big Data.”
29. Diebold, “On the Origin(s).”
30. Baldwin, “Big Data as Boogeyman”; Peters, “Big Data Gold Rush”; Press, “6 Predictions.”
31. Davenport, Barth, and Bean, “How ‘Big Data’ Is Different.”
32. Floridi, “Big Data.”
33. Infochimps, “CIOs & Big Data.”
34. Greenberg, “10 Reasons.”
35. Fidelman, “These Smart, Social Apps”; Walker, “Small Data Is Beautiful.”
36. Walker, “Small Data Is Beautiful.”
37. Bonde, “Defining Small Data.”
38. Kitchin and Lauriault, “Small Data,” 464.
39. See, e.g., Barber and Harfoush, “Synchonizing Small Data”; daCosta, Rethinking the Internet of Things; and A. Paul and Bruns, “Usability of Small Crisis Data Sets.”
40. M. Chen, Mao, and Liu, “Big Data: A Survey.”
41. Davies and Frank, “There’s No Such Thing as Raw Data.”
42. J. Lu and Li, “Bias Correction,” 2658.
43. Jacobs, “Pathologies of Big Data.”
44. Jacobs, “Pathologies of Big Data.”
45. Jacobs, “Pathologies of Big Data,” 40.
46. Floridi, “Big Data,” 436.
47. Buhl et al., “Big Data.”
48. See, e.g., Gutierrez, “Big Data vs. Small Data”; Pollock, “Forget Big Data”; and Markowsky, “In Praise of Small Data.”
49. Markowsky, “In Praise of Small Data.”
50. Markowsky, “In Praise of Small Data,” 1.
51. Couldry and Powell, “Big Data from the Bottom Up.”
52. Rast, “Context as Assumptions.”
53. Simon, Too Big to Ignore.
54. Small Data Lab at Cornell University, http://smalldata.tech.cornell.edu/; Estrin, “Small Data,” 33.
55. Elwood and Leszczynski, “Privacy, Reconsidered”; Estrin, “Small Data.”
56. Simon, Too Big to Ignore.
57. The special issue was GeoJournal 80, no. 4 (2015).
58. Goodchild, “Quality of Big (Geo)Data.” Considerando che abbiamo fornito diverse distinzioni di dati piccoli e grandi, autori e sostenitori tendono a enfatizzare solo il primo V, volume. Se i governi sono i principali produttori degli small data, allora vedremo sicuramente una notevole varietà degli small data, ad esempio i dati provenienti dalla rete idrica, dai trasporti e dalle aree del parco. Inoltre i censimenti soddisfano un requisito di velocità per i big data. I censimenti hanno una periodicità temporale, ad esempio annuale o decennale. Piccoli dati ci consentono di affinare la

caratterizzazione della velocità ad alta e crescente velocità. La componente di velocità dei big data verrebbe chiamata in modo più preciso accelerazione.
59. Kitchin, “Big Data and Human Geography,” 3.
60. H. Miller and Goodchild, “Data-Driven Geography.”
61. See Couclelis, “Construction of the Digital City”; Sui and Bednarz, “Message Is the Medium”; and E. Sheppard et al., “Geographies of the Information Society.”
62. Goodchild, “Quality of Big (Geo)Data,” 8.
63. Kitchin, “Big Data and Human Geography,” 263.
64. Hyman, “Small Data.”
65. Goodchild, “Quality of Big (Geo)Data.”
66. Kitchin and Lauriault, “Small Data,” 463.
67. Kitchin and Lauriault, “Small Data,” 464.
68. Haklay, Singleton, and Parker, “Web Mapping 2.0.”
69. Burns and Thatcher, “Guest Editorial,” 3.
70. Kitchin, “Big Data and Human Geography.”
71. Wu, “Big Data Fallacy.”
72. Jacobs, “Pathologies of Big Data.”
73. Armstrong and Armstrong, “Beyond Numbers.”
74. Snickars, “If Content Is King.”
75. Wilson, “Morgan Freeman Is Dead.”
76. Wilson, “Morgan Freeman Is Dead.”
77. Kitchin, “Big Data and Human Geography.”
78. Wright, Goodchild, and Proctor, “Demystifying the Persistent Ambiguity of GIS.”
79. Kitchin and Lauriault, “Small Data,” 464.
80. James, “Out of the Box,” 119.
81. http://www.etriks.org/author/trevor-garrett/.
82. Office of the Auditor General of Canada, “Chapter 7.”
83. Wallis, Rolando, and Borgman, “If We Share Data.”
84. Wallis, Rolando, and Borgman, “If We Share Data,” 15.
85. Onsrud and Pinto, “Diffusion of Geographic Information.”
86. Olfat et al., “Spatial Metadata Automation.”
87. Ballantyne, “Geomatics and the Law.”
88. H. Miller and Goodchild, “Data-Driven Geography.”
89. Mills, Sociological Imagination, 67.
90. “Piccolezza” e “molto piccola e lenta” sono classificazioni soggettive. Gli etnografi probabilmente non qualificheranno il loro lavoro come “molto piccolo”. Un lavoratore del censimento difficilmente prenderebbe in considerazione un piccolo censimento, anche con la definizione di Kitchin e Lauriault “Small Data”. Riconosciamo questi disaccordi anche quando utilizziamo i termini.
91. Hardt, “How Big Data Is Unfair.”
92. Elwood and Leszczynski, “New Spatial Media.”
93. Hardt, “How Big Data Is Unfair” (original emphasis).
94. Goodchild, “Citizens as Sensors.”
95. Fischetti, “Web Turns 20.”
96. Rundstrom, “Teaching American Indian Geographies.”
97. Floridi, “Big Data.”
98. Graham and Shelton, “Geography and the Future of Big Data,” 259.