"Thinking Big Data in Geography"
Traduzione del capitolo 1

"Verso studi critici dei dati: tracciamento e analisi di insiemi di dati e il loro funzionamento"
Rob Kitchin e Tracey P. Lauriault

Un approccio critico ai dati

Le società hanno raccolto, archiviato e analizzato dati per un paio di millenni come mezzo per registrare e gestire le proprie attività. Ad esempio, gli antichi egizi tenevano registri amministrativi di atti fondiari, dimensioni dei campi e bestiame a fini fiscali, il Domesday Book del 1086 conteneva dati demografici, la contabilità a partita doppia è stata utilizzata da banchieri e assicuratori sin dal XIV secolo e il primo registro nazionale fu realizzato in Svezia nel XVII secolo.[1] Non è stato fino al XVII secolo, tuttavia, che il termine "dati" venisse utilizzato per la prima volta in lingua inglese, grazie alla crescita della scienza, allo sviluppo della statistica e il passaggio dalla conoscenza costruita dalla teologia, dall'esortazione e dal sentimento ai fatti, e alle prove empiriche di teorie attraverso esperimenti.[2] Nel tempo l'importanza dei dati è cresciuta, diventando fondamentale per la produzione di nuova conoscenza, lo svoglimento di attività e l'adozione di governance. I dati forniscono gli input chiave ai sistemi che individui, istituzioni, imprese e le scienze impiegano per comprendere, spiegare, gestire, regolare e prevedere il mondo in cui viviamo e siamo abituati a fare innovazioni, creare prodotti e concepire politiche.

Il volume, la varietà e l'uso dei dati sono cresciuti enormemente dal XVII secolo, e da tempo continuano la creazione e il mantenimento di insiemi di dati molto grandi, come i censimenti, o database governativi di carattere amministrativo e su risorse naturali. Tali database, tuttavia, in passato venivano creati a distanza di anni oppure per mezzo di campionamenti delle popolazioni di riferimento. Al contrario, negli ultimi cinquant'anni siamo entrati nell'era dei big data, con le seguenti caratteristiche:

- enormi nel volume, costituito da terabyte o petabyte di dati;
- ad alta velocità, creati in tempo reale o quasi;
- diversificati nella tipologia, essendo strutturati o non strutturati;
- esaustivi nel contesto, con lo scopo di rappresentare intere popolazioni o sistemi (n = totale);
- a grana fine nella risoluzione e in grado di permettere identificazioni univoche;
- di natura relazionale, con campi comuni che consentono la congiunzione di differenti set di dati;
- flessibili, con tratti di estendibilità (nuovi campi possono essere facilmente aggiunti) e scalabilità (i set di dati possono crescere rapidamente di numero).[3]

Anche se a seconda della metodologia utilizzata ci sono stime diverse per quanto riguarda la crescita della produzione di dati causata principalmente dai big data, oltre che da una forte crescita dei cosiddetti small data come video personali, foto e file audio (che presi assieme occupano enormi quantità di spazio di archiviazione), è fuor di dubbio che di recente c'è stato un significativo cambiamento nel volume di dati generati, soprattutto dall'inizio del nuovo millennio.[4] Gantz e Reinsel hanno stimato che il volume di dati è cresciuto di un fattore nove negli ultimi 5 anni e Manyika et al. hanno previsto un aumento del 40% all'anno dei dati generati a livello globale.[5] Nel 2013 il commissario europeo per l'agenda digitale Neelie Kroes ha riferito che 1,7 milioni di miliardi di byte di dati al minuto venivano generati a livello globale.[6] Tali aumenti e proiezioni per ulteriori aumenti sono dovuti alla produzione continua ed esaustiva, e non campionata, di dati nativamente digitali, in combinazione con la natura di alcuni di tali dati (ad es. file di immagini e video) e la maggiore capacità di archiviare e condividere i dati a costi marginali. Ad esempio, nel 2012 Facebook ha riferito che stava

elaborando 2,5 miliardi di contenuti (link, commenti, ecc.), 2,7 miliardi di azioni "Mi piace" e 300 milioni di caricamenti di foto al giorno, e Walmart generava oltre 2,5 petabyte (2^{50} byte) di dati relativi a oltre 1 milione di transazioni dei suoi clienti ogni ora.[7]

Questi enormi volumi di dati vengono prodotti da una serie diversificata di tecnologie per l'informazione e la comunicazione (ICT) che mediano e aumentano (nel senso di *augmentare*) sempre di più la nostra vita quotidiana come, per esempio, CCTV (telecamere a circuito chiuso) digitali, postazioni checkout nei negozi al dettaglio, smartphone, transazioni e interazioni online, sensori e scanner, social media e tecnologie con sistemi di localizzazione. Oltre ad essere prodotti da agenzie governative, enormi quantità di dati dettagliati vengono ora generati dagli operatori di telefoni cellulari, dagli sviluppatori di app, dalle società Internet, dagli istituti finanziari, da catene di negozi, dalle società di sorveglianza e sicurezza, etc. e i dati vengono regolarmente scambiati con e tra i *data broker* (mediatori di scambio dati) come merce sempre più importante. Sempre più dati analogici conservati negli archivi e i *repository* vengono digitalizzati e collegati tra loro e resi disponibili attraverso nuove infrastrutture dati e vaste aree di dati prodotti e detenuti dal governo vengono resi apertamente accessibili man mano che il movimento *open data* acquisisce trazione.[8]

Questo cambiamento nella produzione di dati ha portato a una riflessione critica sulla natura dei dati e come siano impiegati. Il concetto di dati si è evoluto, e si è iniziato a pensare ai dati come pre-analitici e pre-fattuali, cioè che esistono prima di ogni interpretazione o argomentazione, e che sono la materia prima da cui sono costruite informazioni e conoscenze. Da questa prospettiva i dati sono considerati essere entità rappresentative, che catturano il mondo sotto forma di numeri, caratteri, simboli, immagini, suoni, onde elettromagnetiche, bit e così via, e mantenendo la regola di essere astratti, discreti, aggregativi (possono essere sommati), invarianti e significativi indipendentemente dal loro formato, mezzo, lingua, produttore e contesto (vale a dire, i dati mantengono il loro significato archiviati come analogici o digitali, visualizzati su carta o schermo o espressi in diverse lingue).[9] I dati sono considerati benigni, neutrali, oggettivi e non-ideologici nella loro essenza, in altre parole uno specchio del mondo in quanto soggetto a vincoli tecnici; non contengono un significato intrinseco e possono essere presi al valore nominale.[10] In effetti i termini comunemente usati per descrivere come vengono gestiti i dati suggeriscono processi tecnici benigni: "raccolti", "inseriti", "compilati", "memorizzati", "elaborati" e "estratti". [11] In altre parole, sono solo gli usi dei dati a essere politici, non i dati stessi.

Questo modo di intendere i dati è stato messo in discussione negli ultimi anni. Contrario all'idea che i dati siano pre-analitici e prefattuali è l'argomento secondo cui i dati sono costitutivi delle idee, delle tecniche, delle tecnologie, delle persone, dei sistemi e dei contesti che concepiscono, producono, elaborano, gestiscono e analizzano i dati.[12] In altre parole, come vengono concepiti, misurati e impiegati i dati forma attivamente la loro natura. I dati non preesistono alla loro generazione; non nascono dal nulla, e la loro generazione non è inevitabile: protocolli, processi organizzativi, scale di misurazione, categorie e standard sono progettati, negoziati e discussi, e esiste un certo disordine nella creazione di dati. Come affermano Gitelman e Jackson, "«dati grezzi» è un ossimoro"; "i dati sono sempre già «cotti»".[13] I dati quindi sono situati, contingenti, relazionali, inquadrati e usati contestualmente per cercare di raggiungere determinati scopi.

Anche i database e i repository non sono semplicemente un mezzo neutro e tecnico di assemblaggio e condivisione di dati, ma sono insiemi di processi contingenti e relazionali che funzionano nel mondo.[14] Sono sistemi *socio-tecnici* complessi che sono inseriti in un più ampio panorama istituzionale di ricercatori, istituzioni e società e sono soggetti a regimi socio-tecnici "radicati in (…) pratiche ingegneristiche e industriali, manufatti tecnologici, programmi politici e ideologie istituzionali che agiscono insieme per governare lo sviluppo tecnologico."[15] Database e repository sono espressioni di conoscenza/potere, modellando quali domande possono essere poste, come vengono poste, come viene data risposta, come sono distribuite le risposte e chi le può avere.[16]

Oltre a questo ripensamento filosofico dei dati, gli studiosi hanno iniziato a pensare ai dati da un punto di vista etico, politico e economico, spaziale e temporale, e infine tecnico.[17] I dati possono riguardare tutti gli aspetti della vita quotidiana, comprese questioni delicate, e possono essere utilizzati in tutti i tipi di modi, tra cui sfruttare, discriminare e perseguitare le persone. Ci sono poi una serie di questioni vive di carattere etico e morale su come i dati vengono prodotti, condivisi, scambiati e protetti; come i dati dovrebbero essere regolati da regole, principi, politiche, licenze e leggi; e in quali circostanze e a quali fini possano essere utilizzati. Non ci sono semplici risposte a tali domande, ma l'ascesa di una generazione di dati più diffusa e invasiva e di mezzi più sofisticati di analisi dei dati creano un imperativo per un dibattito e un'azione pubblici. Inoltre, i dati sono inquadrati da preoccupazioni politiche su come sono concepiti e discussi come beni pubblici e privata. I movimenti per gli open data e i governi trasparenti, per esempio, vedono i dati come beni pubblici che dovrebbero essere liberamente accessibili. Il mondo degli affari, al contrario, considera i dati come un bene prezioso che, da un lato, deve essere protetto da regimi di proprietà intellettuale (copyright, brevetti, diritti di proprietà) e, dall'altro, deve essere sfruttabile a fini di guadagno. Infatti i dati spesso costituiscono una risorsa economica: sono spesso venduti dai governi in regime di recupero dei costi e per le società private sono *commodities* negoziabili alle quali può essere aggiunto valore o dalle quali valore può essere estratto (ad es. dati derivati, analisi, conoscenza).
Nell'era attuale i dati sono una componente chiave dell'emergente economia della conoscenza, che migliora produttività, competitività, efficienza, sostenibilità e accumulo di capitale. L'etica, la politica e l'economia dei dati si sviluppano e mutano nello spazio e nel tempo con il cambiamento dei regimi, delle tecnologie e delle priorità. Da un punto di vista tecnico, ci si è concentrati su come gestire, archiviare e analizzare enormi torrenti di dati, con lo sviluppo del *data mining* e di tecniche di analisi dei dati dipendenti dall'apprendimento automatico (*machine learning*), e ci sono state preoccupazioni rispetto a qualità, validità, affidabilità, autenticità, usabilità e origine dei dati.
In breve, stiamo assistendo allo sviluppo di ciò che Dalton e Thatcher definiscono *studi critici dei dati*: ricerca e pensiero che applicano la teoria sociale critica ai dati per esplorare i modi in cui non sono mai semplicemente neutre, oggettive, indipendente, e grezze rappresentazioni del mondo ma sono situati, contingenti, relazionali, contestuali e svolgono un ruolo attivo nel mondo.[18] Nella loro analisi Dalton e Thatcher hanno esposto sette provocazioni necessarie per fornire una critica integrale dei nuovi regimi di dati:

- situare i regimi di dati nel tempo e nello spazio;
- esporre i dati come intrinsecamente politici e identificare gli interessi di chi servono;
- scompattare la complessa relazione non-deterministica tra dati e società;
- illustrare i modi in cui i dati non sono mai grezzi;
- sfatare il mito secondo cui i dati parlano da sè e che i big data sostituiranno gli small data;
- esplorare come i nuovi regimi di dati possono essere utilizzati in modi socialmente progressivi;
- esaminare come il mondo accademico si impegna con i nuovi regimi di dati e le opportunità derivanti da tale impegno.

Siamo d'accordo con la necessità di tutte queste provocazioni. In una breve presentazione in una riunione dell'associazione dei geografi americani uno di noi ha definito una visione di come potrebbero essere gli studi critici dei dati: scompattare i complessi assemblaggi che producono, circolano, condividono / vendono, e utilizzano i dati in diversi modi; tracciare il variegato lavoro che tali assemblaggi svolgono e le sue conseguenze su come il mondo è conosciuto, governato e vissuto; sondare il più ampio panorama degli assemblaggi di dati e come interagiscono per formare prodotti, servizi e mercati che si intersecano tra loro e modellare policy e regolamentazioni. È a questa impresa che adesso noi ci rivolgiamo.

Tracciamento e spacchettamento di assemblaggi di dati

Kitchin definisce un assemblaggio di dati come un complesso sistema socio-tecnico composto da numerosi apparati e elementi tra loro intrecciati e finalizzati alla produzione, gestione, analisi e traduzione di dati e prodotti di informazione derivati per scopi commerciali, governativi, amministrativi, burocratici o di altro tipo (vedi tabella 1-1).[19] Un assemblaggio di dati consiste in più del sistema di dati o dell'infrastruttura stessa, come un sistema di big data, un repository di open data o un archivio: un assemblaggio include tutti gli aspetti tecnologici, politici, e gli apparati sociali e economici che formano la loro natura e il loro funzionamento. Gli apparati e gli elementi dettagliati nella tabella 1-1 interagiscono tra loro e si modellano attraverso una contingente e complessa rete di relazioni poliedriche. E proprio come i dati sono un prodotto dell'assemblaggio, l'assemblaggio è strutturato e gestito per produrre quei dati.[20] I dati e il loro assemblaggio sono così costituiti reciprocamente, legati in un insieme di pratiche e relazioni sia discorsive sia materiali che sono contingenti, relazionali e contestuali. Ad esempio, l'assemblaggio di dati di un censimento consiste in una grande amalgama di apparati ed elementi che modellano come il censimento viene formulato, amministrato, elaborato e comunicato e come vengono impiegati i suoi risultati. Un censimento è sostenuto da un sistema di pensiero realista; ha una serie diversificata di forme di documentazione d'accompagnamento; le sue domande sono negoziate da molti *stakeholder*; i suoi costi sono materia di discussione; la sua amministrazione e rendicondazione sono modellate da quadri giuridici e regolamenti; viene consegnato attraverso una serie diversificata di pratiche, intraprese da molti lavoratori, utilizzando una gamma di materiali e infrastrutture; i suoi dati alimentano tutti i tipi d'uso e mercati secondari. Gli insiemi di dati si evolvono e si trasformano quando nuove idee e conoscenze emergono, quando nuove tecnologie vengono inventate, quando le organizzazioni cambiano, quando vengono creati modelli di business, quando ci sono cambiamenti nell'economia politica, quando regolamenti e leggi vengono introdotti o abrogati, quando si sviluppano nuovi set di competenze, quando dibattiti hanno luogo e i mercati crescono o si restringono. E mentre i set di dati una volta generati all'interno di un assemblaggio possono apparire fissi e immutabili (ad es. un censimento compilato), essi sono in realtà aperti alla correzione e revisione, alla rielaborazione attraverso disaggregazione e riaggregazione in nuove classi o geografie statistiche, all'analisi di altri sistemi di dati, alla creazione di dati derivati da essi, a interpretazioni e intuizioni alternative. Gli assemblaggi di dati e i dati in essi contenuti sono quindi sempre in uno stato di divenire.

Questa nozione di raccolta di dati è simile al concetto di *dispositif* di Foucault, che si riferisce a un "insieme completamente eterogeneo composto da discorsi, istituzioni, forme architettoniche, decisioni normative, leggi, misure amministrative, dichiarazioni scientifiche, proposizioni filosofiche, morali [,] e filantropiche" che migliorano e mantengono l'esercizio di potere all'interno della società.[21] Il dispositif di un'infrastruttura di dati produce ciò che Foucault definisce "potere/conoscenza", cioè conoscenza che svolge una funzione strategica: "l'apparato è così sempre inscritto in un gioco di potere, ma è anche sempre collegato a determinate coordinate della conoscenza che ne deriva ma che, in egual misura, la condiziona. Questo è ciò che forma l'apparato: strategie di relazioni di forze che sostengono e sono supportate da diversi tipi di conoscenza."[22] In altre parole, le infrastrutture di dati non sono mai neutrali, essenziali, obiettive; i loro dati non sono mai grezzi ma vengono sempre cotti in qualche ricetta dagli chef incorporati all'interno delle istituzioni che hanno determinate aspirazioni e obiettivi e operano in contesti più ampi.

Tabella 1-1 Apparati e elementi di un assemblaggio di dati

Apparato	**Elementi**
Sistemi di pensiero	Modalità di pensiero, filosofie, teorie, modelli, ideologie, razionalità, ecc.
Forme di conoscenza	Testi di ricerca, manuali, riviste, siti Web, esperienza, passaparola, forum di chat, ecc.
Finanza	Modelli di business, investimenti, capitale di rischio, sovvenzioni, filantropia, profitto, ecc.
Economia politica	Policy, regimi fiscali, strumenti di incentivazione, opinione pubblica e politica, ecc.
Governamentalità e legalità	Standard di dati, formati di file, requisiti di sistema, protocolli, regolamenti, leggi, licenze, proprietà intellettuale regimi, considerazioni etiche, ecc.
Materialità e infrastrutture	Carta / penne, computer, dispositivi digitali, sensori, scanner, database, reti, server, edifici, ecc.
Pratiche	Tecniche, modi di fare, comportamenti appresi, convenzioni scientifiche, ecc.
Organizzazioni e istituzioni	Archivi, società, consulenti, produttori, rivenditori, agenzie governative, università, conferenze, club e società, comitati e consigli di amministrazione, comunità di pratica, ecc.
Soggettività e comunità	Produttori di dati, esperti, curatori, manager, analisti, scienziati, politici, utenti, cittadini, ecc.
Luoghi	Laboratori, uffici, siti sul campo, data center, server farm, business park, ecc. e relativi agglomerati
Mercato	Per i dati, i loro derivati (ad es. testo, tabelle, grafici, mappe), analisti, software analitici, interpretazioni, ecc.

Questa elaborazione di dati viene rivelata attraverso il lavoro di Ian Hacking, che ha tratto ispirazione dal pensiero di Foucault sulla produzione di conoscenza.[23] Hacking teorizza che all'interno di un insieme di dati ci siano due processi correlati che producono e legittimano i suoi dati e gli apparati e gli elementi associati, dando forma al modo in cui i suoi dati funzionano nel mondo, i quali a loro volta influenzano future iterazioni di dati e la costituzione di nuovi insiemi. In entrambi i casi egli postula che è in atto un nominalismo dinamico, in cui esiste un'interazione tra dati e ciò che rappresentano, il che porta a influenze e cambiamenti reciproci.

Il primo di questi processi è ciò che Hacking definisce "effetto loop". [24] L'effetto loop riguarda come i dati sono classificati e organizzati, come emerge un'ontologia dei dati e come può rimodellare ciò che è stato già classificato. Il loop (fig. 1-1) ha cinque fasi:

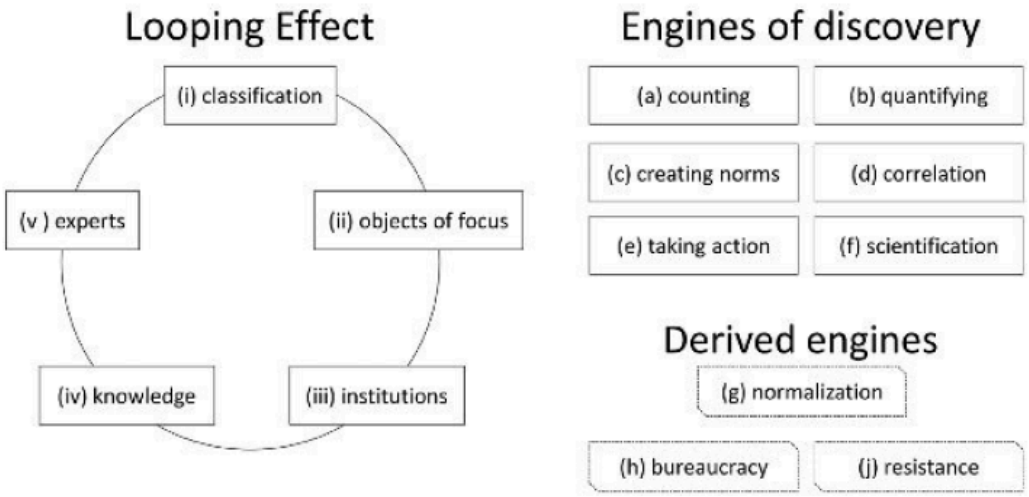

Figura 1-1
Il lavoro di un assemblaggio di dati, a seguito di Hacking, "Philosophie et histoire des concepts scientifiques", e Laney, "Gestione dei dati 3D". Creato da R. Kitchin e T. Lauriault.

1. *classificazione*, in cui elementi che sono considerati avere caratteristiche condivise vengono raggruppati o, in caso di devianza, vengono forzati a raggrupparsi;
2. *oggetti di focus* (ad es. persone, spazi, mode, malattie, ecc.) in cui, nel caso di persone, le persone alla fine iniziano a identificarsi con la classe a cui sono state assegnate o, nel caso di oggetti, le persone arrivano a capire e ad agire sugli oggetti secondo la loro classificazione;
3. *istituzioni*, che istituzionalizzano le classificazioni e gestiscono le infrastrutture dei dati;
4. *conoscenza*, che viene utilizzata per formulare, riprodurre e modificare le classificazioni;
5. *esperti* che, in quanto membri istituzionali che producono ed esercitano conoscenze, attuano la classificazione.

Attraverso questo effetto a ciclo continuo, Hacking sostiene che si verifica un processo di "immaginazione di persone", in cui sistemi di dati come il censimento o la valutazione della salute mentale delle popolazione attuano un lavoro di classificazione che finisce per rimodellare la società a immagine di un'ontologia dei dati. Esempi sono persone che si definiscono o sono definite da sintomi di salute mentale, o strutture per la salute mentale costruite e gestite da professionisti specializzati.
Il secondo dei processi consiste in ciò che Hacking definisce "motori di scoperta" che vanno oltre semplici questioni metodologiche. Egli discute di questi metodi usando una prospettiva medica, che Lauriault ha modificato per includere la composizione degli spazi e delle persone.[25] Hacking postula che ci siano un certo numero di tali motori, gli ultimi tre dei quali sono motori derivati, che sono:

a. *contare* i volumi dei diversi fenomeni;
b. *quantificazione*: trasformazione dei conteggi in misure, tassi e classificazioni;
c. *creare norme*: stabilire cosa potrebbe o dovrebbe essere previsto;
d. *correlazione*: determinare le relazioni tra le misure;
e. *azione*: utilizzare le conoscenze per affrontare e trattare i problemi;
f. *scientificazione*: creazione e adozione di conoscenze scientifiche;
g. *normalizzazione*: cercare di modellare il mondo in base alle norme (ad es. incoraggiare le diete per raggiungere gli indici di massa corporea previsti);

h. *burocratizzazione*: stabilire istituzioni e procedure per amministrare la produzione di aspettative e intraprendere azioni;
i. *resistenza* a forme di conoscenza, norme e burocrazie da parte di coloro che ne sono colpiti in maniera negativa (ad es. resistenza degli omosessuali e delle persone con disabilità a modelli medicali che li classificano, li posizionano e li trattano in modi particolari) o da parte di coloro che promuovono sistemi, interpretazioni e visioni alternative.[26]

Insieme, questi motori svolgono contemporaneamente un lavoro di raccolta dei dati e un'attività di legittimazione del lavoro di raccolta, dei dati, e del loro assemblaggio in un insieme. Ad esempio, un censimento conta una popolazione e gli aspetti della vita delle persone, trasforma queste informazioni in misure, stabilisce tassi di base, valuta le relazioni tra i vari fattori e si trasforma in conoscenza, che porta a pratiche di normalizzazione che vengono attuate da una burocrazia dedicata e correlata. Ogni fase rafforza la precedente e collettivamente tutte le fasi giustificano il lavoro svolto. Ci può essere resistenza contro la conoscenza prodotta o addirittura contro l'intero assemblaggio, come nel censimento boicottato in Germania negli anni '80 o nelle campagne per assicurarsi che l'etnia irlandese non venisse sottostimata nel Regno Unito, o il movimento per far riconoscere "neozelandese" come etnia in Nuova Zelanda (invece di "New Zealand European"), o le pressioni per il riconoscimento del lavoro non retribuito delle casalinghe. In certi casi ci possono essere trasgressioni della conoscenza prodotta, come nel caso di coloro che dichiarano la propria religione come Jedi.[27] Altre volte, il lavoro di raccolta dati può addirittura essere cancellato, come nel censimento del Canada del 2011.
Gli insiemi di dati fanno parte di un più ampio panorama composto da numerosi e correlati assemblaggi e sistemi di dati interagenti tra loro. Nel settore pubblico, ad esempio, ci sono migliaia di sistemi di dati (ognuno circondato da un più ampio assemblaggio) che interagiscono e funzionano di concerto per produrre servizi statali e forme di controllo governativo a livello locale, regionale e nazionale. Spesso questo panorama di dati si estende alla scala pan-nazionale e globale, attraverso set di dati interregionali e mondiali, accordi e infrastrutture per la condivisione di dati, e la formulazione di protocolli, standard e quadri giuridici (ad es. Global Spatial Data Infrastructures, INSPIRE). Anche le imprese creano e occupano complessi panorama di dati, vendendo, acquistando e condividendo dati da milioni di sistemi di dati, tutti parte di più ampi assemblaggi socio-tecnici. Ad esempio, il panorama dei big data è costituito da centinaia di aziende, che vanno da piccole e locali a grandi e globali, che offrono una gamma di servizi complementari e concorrenti, come elaborazione di dati, compilatori e aggregatori di specialità, analisi di dati, strumenti di segmentazione, gestione delle liste, interpretazione e consulenza, marketing, editoria, ricerca e sviluppo. Abbiamo appena iniziato a mappare vari scenari di dati, le loro spazialità e temporalità, la loro complessa economia politica, e il lavoro che svolgono nel catturare, analizzare e rimodellare il mondo. È a quest'ultimo che ora ci rivolgiamo.

Scoprire il lavoro degli assemblaggi di dati
Come notato nella sezione precedente, le raccolte di dati lavorano su scala mondiale. I dati sono sfruttati per supportare i compiti di governo delle persone e dei territori, di gestione delle organizzazioni, di produzione del capitale, di creazione di luoghi migliori, di miglioramento dell'assistenza sanitaria, di progresso della scienza e così via.
Questo sfruttamento assume molte forme, ma il principio centrale è che i dati, se analizzati e sfruttati appropriatamente, producono informazioni e conoscenze che possono essere utilizzate per rimodellare procedure operative e strutture organizzative, per identificare nuovi prodotti, per segmentare mercati, per ridurre l'incertezza e il rischio e per aumentare efficienza, produttività, competitività e sostenibilità.[28] Sebbene gran parte del lavoro a cui vengono sottoposti i dati sia vantaggioso per la società in generale (con i dati ad esempio utilizzati per migliorare la qualità della vita e per affrontare le questioni umanitarie e i problemi ambientali) c'è anche un lato oscuro di tanto lavoro sui dati. Qui

vogliamo considerare quest'ultimo, evidenziando quattro modi in cui i dati vengono utilizzati per produrre relazioni sociali e economiche pericolose: *dataveillance* (la pratica di monitorare dati digitali che riguardano dettagli personali o attività online di una persona) e erosione della privacy, profilazione e classificazione sociale, governance preventiva, e usi secondari e *control creep* (accrescimento del controllo sociale). Queste pratiche sono attualmente oggetto di molti dibattiti e vi è un urgente bisogno di studi critici in grado di informare le discussioni in corso.

Come hanno dimostrato le rivelazioni di WikiLeaks, Edward Snowden e altri informatori, il caso Maher Arar e altre sfide legali relative alla gestione errata di registri di dati e al maltrattamento di individui, dall'11 settembre in poi c'è stato un cambiamento radicale nella misura e nella natura della sorveglianza e della sicurezza governative in molti stati. Vaste quantità di comunicazioni quotidiane (telefonate, messaggi di testo, e-mail, social media), nonché l'uso generale di Internet, vengono sistematicamente raccolte da organizzazioni come la National Security Agency degli Stati Uniti e analizzate per *intelligence* strategica.[29]

Allo stesso modo, tutti gli stati raccolgono grandi banche dati di informazioni sui cittadini in relazione a tutti gli aspetti della loro vita: reddito, tasse, welfare, salute, istruzione e così via. Analogamente, le aziende generano regolarmente dati relativi a tutti gli aspetti della loro attività, compresi i loro clienti e i loro modelli di consumo. In effetti, dato il ruolo di mediazione del software in attività come lavoro, viaggi, consumi, comunicazione e giochi, è sempre più difficile vivere la propria quoidianità senza lasciare una traccia digitale.[30] Ad esempio, l'autorità olandese per la protezione dei dati stima che il cittadino olandese medio sia registrato in 250-500 database, con alcuni in un massimo di 1000 database, una cifra in crescita.[31] Questi database non comprendono solo le *digital footprints* (orme digitali) degli individui (ossia dati che gli individui stessi producono) ma anche le loro *digital shadows* (ombre digitali, ovvero informazioni sugli individui generate da altri). Coloro ai quali i dati si riferiscono hanno spesso scarso controllo su tali dati, sulla loro forma, estensione o come vengono utilizzati.[32] Individualmente, questi database offrono una visione limitata delle persone, ma quando combinati diventano molto più potenti, rivelando *pattern* dettagliati e consentendo ciò che è stato chiamato *dataveillance*: la classificazione e il filtraggio dei set di dati al fine di identificare, monitorare, tracciare, regolare, prevedere e prescrivere.[33] La diffusa generazione di dati e le pratiche di sorveglianza dei dati sollevano molte questioni relative alla privacy e ai diritti all'anonimato e alla riservatezza che stanno solo ora iniziando a essere considerate.[34]

I dati sono stati a lungo utilizzati per profilare, segmentare e gestire le popolazioni, ma questi processi sono diventati molto più sofisticati, perfezionati, diffusi e di routine con l'applicazione dell'analisi dei dati utilizzando tecniche di apprendimento automatico.[35] Mentre lo stato potrebbe profilare i propri cittadini ai fini di sicurezza e di polizia, le imprese commerciali stanno cercando di ridurre il rischio e massimizzare il rendimento attraverso un targeting più efficace dei prodotti. Mentre generazioni precedenti di profilazione hanno cercato di creare profili aggregati di popolazione o area, i quali hanno poi modellato il processo decisionale in materia di marketing e posizionamento dei prodotti (ad es. profilazione geodemografica), l'analisi di nuova generazione può funzionare a livello di individuo, combinando dati provenienti da varie fonti come le transazioni con carte di credito e carte a punti di negozi, flussi di clic, post sui social media e altri tipi di dati personali, per produrre un profilo cliente dettagliato.[36] Questi profili vengono utilizzati per classificare socialmente i clienti, identificandone alcuni per un trattamento preferenziale ed escludendone altri, e per prevedere la probabilità che i clienti siano in grado di saldare i propri pagamenti o per giudicare il loro valore previsto nel caso rimangano fedeli e la probabilità che essi vadano altrove a fare acquisti.[37] I profili sono anche utilizzati per sostenere nuove forme di prezzatura dinamica e personalizzata, su misura per il profilo del consumatore e la cronologia dei suoi acquisti, con lo scopo di raggiungere la spesa ottimale.[38] I consumatori vengono quindi regolarmente misurati e classificati e ricevono servizi differenziati in base ai dati associati e al luogo in cui vivono.

Una forma particolarmente perniciosa di profilazione predittiva è la governance preventiva. Si tratta di analisi predittive che vengono utilizzate per valutare probabili comportamenti o eventi futuri e elaborare un piano d'azione adeguato. Tale governance preventiva è stata una caratteristica dei viaggi aerei per un certo numero di anni, con i passeggeri profilati per il rischio e i livelli di controlli di sicurezza prima di iniziare il loro viaggio.[39] Più recentemente questo meccanismo è stato esteso alla polizia, con un numero significativo di forze di polizia statunitensi che lo hanno utilizzato per identificare potenziali futuri criminali e per dirigere il pattugliamento di aree sulla base di un'analisi dei dati storici sulla criminalità, delle registrazioni degli arresti e delle note reti sociali dei criminali.[40]
In tali casi, le ombre dei dati degli individui fanno più che seguirle: le ombre li precedono, con lo scopo di sorvegliare comportamenti che potrebbero non verificarsi mai.[41] Di conseguenza, le persone vengono trattate in modo diverso in previsione di qualcosa che potrebbero fare o meno. Dati i loro effetti sulle vite degli individui e la loro natura di *black box* (sistemi che danno risultati che però non mostrano come tali risultati siano stati ottenuti), le pratiche di profilazione predittiva, classificazione sociale e governance preventiva richiedono molta più attenzione, così come le aziende che sviluppano e svolgono tali compiti.
Il lavoro svolto dai sistemi di dati in tutti questi casi si basa sulla generazione di un eccesso di dati. In effetti, i big data si basano sulla generazione, l'archiviazione e il collegamento di quanti più dati possibili nella speranza di ricavare da essi valore e conoscenza. Anziché essere generati e utilizzati per eseguire un'attività specifica, i dati possono essere riconfezionati, venduti e riutilizzati per tutti i tipi di usi secondari. Tale strategia è in contrasto con la politica di minimizzazione dei dati, una delle basi della privacy e della protezione dei dati nell'Unione Europea e nel Nord America. Questa policy prevede che i dati debbano essere generati e utilizzati solo per eseguire una determinata attività e che debbano essere conservati solo per il tempo necessario a svolgere tale attività.[42] Un esempio di dove si stia violando la premessa della minimizzazione dei dati si ha con il control creep, in cui i dati generati per una forma di governance sono appropriati e usati per un'altra.[43] Chiaramente il control creep si è verificato principalmente nel contesto della sicurezza, con l'industria aerea e i dati amministrativi del governo riutilizzati per la profilazione e la valutazione del rischio passeggeri; con le telecamere installate per controllare gli accessi auto in centro città e invece utilizzate anche per le attività di polizia; con i dati dei social media riutilizzati per condurre indagini criminali e intraprendere profilazione predittiva.[44] Il control creep è evidente anche in una serie di altri domini, ad esempio, nell'uso della posizione personale, dei consumi e dei dati dei social media per valutare il rischio di credito o l'idoneità per l'occupazione di una persona.[45] Date le implicazioni per le libertà civili da uso secondario dei dati, è necessario esaminarne le conseguenze e progettare nuovi approcci alla protezione dei dati, come la *privacy by design*.[46]

Conclusione
Dalton e Thatcher concludono la loro richiesta di studi critici sui dati ponendo cinque domande che ritengono necessitino di ulteriori studi, tutte relative ai big data:

1. Quali condizioni storiche portano alla realizzazione dei big data come sono oggi?
2. Chi controlla i big data, la loro produzione e la loro analisi? Quali motivi e imperativi guidano il suo lavoro?
3. Chi sono i soggetti dei big data e quali conoscenze stanno producendo?
4. Come vengono effettivamente applicati i big data nella produzione di spazi, luoghi e paesaggi?
5. Che cosa si deve fare con i big data e quali altri tipi di conoscenza potrebbero aiutare a produrre? [47]

Ci sono molte altre domande che possono essere aggiunte a questo elenco, non solo allargando l'obiettivo fino a includere gli open data, nonché archivi e repository di dati, ma anche considerando i

più ampi paesaggi, assemblaggi e mercati dei dati. Piuttosto che produrre un ampio elenco di domande, vogliamo concludere chiedendo un maggiore lavoro concettuale e una ricerca empirica per sostenere e approfondire studi critici dei dati.

I modi in cui i dati vengono generati, le analisi utilizzate per elaborare e estrarre informazioni da essi, le industrie che crescono intorno a loro, la loro più ampia definizione politica economica e come sono impiegati richiedono tutti un impegno critico. Mentre esiste una ricca e diversificata tradizione di teoria sociale critica che può essere diretta verso insiemi di dati e un più ampio panorama di dati, tale teoria deve essere perfezionata e messa a punto per dare un senso ai dati e al loro lavoro nel mondo, con lo sviluppo di una nuova teoria se necessario. Tuttavia abbiamo solo appena iniziato a concettualizzare criticamente i dati, i loro apparati e i loro elementi. Tale pensiero deve essere integrato con una riflessione più normativa sull'etica e sulla politica dei big data, degli open data e i diversi sistemi di dati oggi esistenti.

Tali valutazioni concettuali e normative devono essere accompagnate da una serie diversificata di *case studies* empirici che esaminano tutti gli aspetti della governance, degli affari e della scienza basati sui dati, che scompattano i gruppi di dati e mappano un più ampio panorama dei dati. Il nostro approccio suggerito è quello di utilizzare metodi come etnografie, interviste, focus group e osservazioni dei partecipanti per approfondire il funzionamento degli assemblaggi, per tracciare le genealogie di come il panorama dei dati è cambiato nel tempo e nello spazio, per mappare le materialità e le infrastrutture che costituire infrastrutture di dati e decostruire il regime discorsivo che accompagna le iniziative basate sui dati.[48]

Intraprendere questo lavoro concettuale e empirico è ciò su cui la nostra ricerca si focalizzerà nei prossimi anni come parte del progetto Programmable City, basato sui nostri studi iniziali su largs scala.[49] Questo vasto progetto sta esaminando le intersezioni di big data e open data, *ubiquitous* computing, software e algoritmi, sviluppi di *smart cities* a Dublino e Boston, scompattando una serie di assemblaggi di dati e tracciando il panorama dei dati di ogni città. Non abbiamo dubbi sul fatto che molti altri saranno impegnati in studi simili, vista la crescita di forme di scienza, affari e governo basate su dati. Speriamo che ciò che questa ricerca produrrà sia un insieme variegato di brillanti studi critici sui dati.

Note

1. Dupaquier and Dupaquier, Histoire de la démographie; Bard and Shubert, Encyclopedia of the Archaeology; Poovey, History of the Modern Fact; Porter, Rise of Statistical Thinking.
2. Poovey, History of the Modern Fact; Garvey, "facts and FACTS"; Rosenberg, "Data before the Fact."
3. Kitchin, "Big Data and Human Geography," 262; boyd and Crawford, "Critical Questions for Big Data"; Dodge and Kitchin, "Codes of Life"; Laney, 3D Data Management; Marz and Warre, Big Data: Principles; Mayer-Schönberger and Cukier, Big Data: Revolution; Zikopoulos et al., Understanding Big Data.
4. See, for example, Hilbert and López, "World's Technological Capacity"; Gantz and Reinsel, Extracting Value from Chaos; and Short et al., How Much Information?
5. Gantz and Reinsel, Extracting Value from Chaos; Manyika et al., Big Data: Next Frontier.
6. Rial, "Power of Big Data."
7. Constine, "How Big Is Facebook's Data?"; Open Data Center Alliance, Big Data Consumer Guide.
8. Lauriault et al., "Today's Data"; Kitchin, Data Revolution.
9. Floridi, "Data"; Rosenberg, "Data before the Fact."
10. Pérez-Montoro and Díaz Nafría, "Data."
11. Gitelman and Jackson, "Introduction."

12. Bowker and Star, Sorting Things Out; Lauriault, "Data, Infrastructures"; Ribes and Jackson, "Data Bite Man"; Kitchin, Data Revolution.
13. Gitelman and Jackson, "Introduction," 2, citing Bowker, Memory Practices.
14. Star and Ruhleder, "Steps toward an Ecology"; Kitchin and Dodge, Code/Space.
15. Ruppert, "Governmental Topologies"; Hecht, "Technology, Politics, and National Identity," 257.
16. Lauriault, "Data, Infrastructures"; Ruppert, "Governmental Topologies."
17. Kitchin, Data Revolution.
18. Dalton and Thatcher, "Critical Data Studies."
19. Kitchin, Data Revolution, 24.
20. Ribes and Jackson, "Data Bite Man."
21. Foucault, "Confession of the Flesh," 194.
22. Foucault, "Confession of the Flesh," 196.
23. Hacking, "Biopower"; Hacking, "Making Up People"; Hacking, "Tradition of Natural Kinds"; Hacking, "Philosophie et histoire des concepts scientifiques"; Hacking, "Kinds of People."
24. Hacking, "Tradition of Natural Kinds"; Hacking, "Philosophie et histoire des concepts scientifiques"; Hacking, "Kinds of People."
25. Lauriault, "Data, Infrastructures."
26. Hacking, "Philosophie et histoire des concepts scientifiques."
27. Hannah, Dark Territory; UK Census, Irish in Britain; Middleton, "Email Urges 'New Zealander'"; Waring, If Women Counted; Singler, "SEE MOM IT IS REAL."
28. Kitchin, Data Revolution.
29. Amoore, "Biometric Borders"; Bamford, Shadow Factory.
30. Kitchin and Dodge, Code/Space.
31. Koops, "Forgetting Footprints."
32. CIPPIC, On the Data Trail.
33. Clarke, "Information Technology"; Raley, "Dataveillance and Countervailance."
34. Solove, "Taxonomy of Privacy"; Elwood and Leszczynski, "Privacy, Reconsidered."
35. Weiss, Clustering of America; Goss, "'We Know Who You Are'"; Parker, Uprichard, and Burrow, "Class Places and Place Classes"; Singleton and Spielman, "Past, Present and Future."
36. Siegel, Predictive Analytics.
37. Graham, "Software-Sorted Geographies"; Minelli et al., Big Data, Big Analytics.
38. Tene and Polonetsky, "Big Data for All."
39. Dodge and Kitchin, "Flying through Code/Space"; Amoore, "Biometric Borders."
40. Siegel, Predictive Analytics; Stroud, "Minority Report."
41. Stalder, "Privacy Is Not the Antidote"; Harcourt, Against Prediction.
42. Tene and Polonetsky, "Big Data for All"; CIPPIC, Submissions to the House of Commons.
43. Innes, "Control Creep."
44. Lyon, Surveillance Studies; Pither, Dark Days; Gallagher, "Staking Out Twitter."
45. O'Reilly, "Creep Factor."
46. Information and Privacy Commissioner/Ontario, Seven Principles of Privacy.
47. Dalton and Thatcher, "Critical Data Studies."
48. Kitchin, Data Revolution.
49. Programmable City, http://www.nuim.ie/progcity/; Lauriault, "Data, Infrastructures"; and Kitchin, Data Revolution.