FILOSOFIA DELLE SCIENZE E DINTORNI / 1

Collana diretta da Giovanni Boniolo

Non solo una scienza, ma tutte. Anche tecnica e tecnologia (se si sapesse in modo condiviso dove termina la prima e iniziano le altre). Non solo riflessione epistemologica, ma pure etica, sociologica, antropologica, economica ecc. Per capire criticamente certi aspetti del mondo e della società in cui si vive, dove si è e dove si sta andando. Insomma, *filosofia delle scienze e dintorni*, sia della filosofia sia delle scienze.

Sabina Leonelli

La ricerca scientifica nell'era dei Big Data

Cinque modi in cui i Big Data danneggiano la scienza, e come salvarla

MELTEMI

Meltemi editore
www.meltemieditore.it
redazione@meltemieditore.it

Collana: *Filosofia delle scienze e dintorni*, n. 1
Isbn: 9788883539015

Sede legale: via Ruggero Boscovich, 31 – 20124 Milano
Sede operativa: via Monfalcone, 17/19
20099 Sesto San Giovanni (MI)
Phone: +39 02 22471892 / 22472232

Indice

A Fany Sferidu e Luca Leonelli,
che mi hanno insegnato a essere
aperta al mondo

Introduzione

Viviamo nell'*era della post-verità*. Nel mondo politico come in quello sociale, la facilità con cui le informazioni si diffondono tramite tecnologie digitali e social media rende più difficile che mai capire quali fonti di informazioni siano affidabili, e su che base. Per ogni tipo di asserzione – dalla veridicità del cambiamento climatico a che tipo di dieta sia più indicata per i diabetici – è sufficiente fare una breve ricerca su Google per trovare sia voci di assenso sia di dissenso. Su internet ci sono tantissimi dati, generati e messi *online* da ricercatori universitari ma anche da mille altre fonti di informazione – dai servizi sociali a quelli commerciali, dalla sanità locale alle istituzioni pubbliche, dai supermercati ai social media. Questo oceano di dati si trasforma inevitabilmente in una cacofonia di interpretazioni dissonanti. Troviamo dati che "provano" che bere vino regolarmente fa male alla salute, ma anche dati che "provano" che i malati di cuore farebbero bene a bere un bicchiere di vino al

giorno. Dati che confermano l'effetto negativo della plastica sull'ecosistema marino e dati che lo smentiscono. Dati che indicano l'effetto negativo dell'inquinamento sulla salute del pianeta e dati che indicano il contrario. E – cosa forse ancora più sconcertante – troviamo persone che usano esattamente gli stessi dati per trarre conclusioni opposte, e lo fanno in modi spesso difficili da valutare per chi non ha competenze specifiche. In questi casi, ci troviamo spesso a indagare credenziali e competenze di coloro che si offrono come interpreti dei dati, e il nostro giudizio su *cosa* credere si riduce a un giudizio su *a chi* credere. E così, nel nostro mondo iperconnesso e multimediale, finiamo per allontanarci sempre più da decisioni basate su dati di fatto, basandoci invece sulle opinioni di chi riteniamo degno di fiducia. Lo *status* della ricerca scientifica come fonte affidabile di verità si è indebolito di conseguenza, fino a essere visto da politici, giornalisti o imprenditori senza scrupoli come equivalente a qualsiasi altra opinione e quindi privo di legittimità.

Vista tale situazione, può sembrare paradossale che questa decade sia anche considerata come l'*era dei dati*: un momento rivoluzionario per l'innovazione tecnologica e i meccanismi di ricerca, e un trionfo della base empirica della conoscenza sulla pura speculazione. Grazie alle tecnologie digitali e a sistemi di ricerca e comunicazione sempre più globalizzati, abbiamo a disposizione enormi quantità di dati – un mare di fatti che aspettano di essere studiati e interpretati, e la

cui analisi tramite algoritmi per l'apprendimento automatico è un fattore fondamentale nello sviluppo dell'intelligenza artificiale. I cosiddetti Big Data contengono la promessa di cambiare radicalmente il modo in cui si fa ricerca e crea conoscenza, sia dentro il mondo accademico sia fuori. L'analisi dei Big Data permette di pianificare, condurre e disseminare ricerca in modi innovativi e – come viene spesso ripetuto – più efficienti che nel passato. Soprattutto, cambia il modo in cui possiamo assemblare e integrare dati che vengono da fonti molto differenti, e diventa molto più facile creare modi per analizzare vaste quantità di dati di tipo diverso in modo veloce e relativamente poco costoso. La disponibilità di dati in grandi quantità incentiva la creazione di sistemi computazionali sempre più potenti per potere analizzarli, e a sua volta la creazione di questi sistemi incentiva investimenti nell'accumulo dei dati. L'accesso e l'analisi dei dati diventa dunque il motore della ricerca: un modello di innovazione che chiameremo *Datocentrico*. La disponibilità dei Big Data, e la facilità con cui vengono prodotti, è una fantastica opportunità di estrapolare nuove scoperte e perfezionare metodi di computazione sempre più autonomi e sofisticati, basandosi sul più vasto archivio di fatti mai registrato nella storia dell'uomo.

Com'è possibile, in questo mondo di Big Data così facilmente accessibili, ritrovarsi tanto persi su ciò che costituisce conoscenza affidabile? Come siamo arrivati a dubitare di ogni verità, in una realtà piena zeppa di fatti? E che conseguen-

ze ha questa situazione per lo sviluppo di tecnologie come l'intelligenza artificiale, che continueranno a trasformare la nostra società in maniera radicale negli anni a venire? Per rispondere a queste domande è necessario capire due cose fondamentali. La prima è il legame tra la produzione dei dati e la produzione di conoscenza, e il modo in cui i dati devono essere gestiti così da poter confermare o smentire un'asserzione. La seconda è costituita dall'enorme difficoltà e dalle immense risorse necessarie per processare e analizzare i dati tanto da usarli per creare interpretazioni affidabili e soggette a valutazione critica. L'obiettivo di questo libro è chiarificare questi due aspetti, in modo da illustrare come i Big Data devono essere preparati e manipolati allo scopo di facilitare analisi e interpretazioni, e riflettere sul profondo legame tra presupposti teorici, metodi e tecnologie usati per analizzare i dati e l'affidabilità della conoscenza che ne viene ricavata. Il libro propone di dimostrare come l'adozione affrettata dei Big Data, e di modi automatizzati per interpretarli, può avere conseguenze disastrose per la credibilità e la qualità del sapere che viene prodotto – e come questa prospettiva può e deve essere evitata per il bene dell'umanità e del pianeta intero.

Le mie osservazioni si basano su quindici anni passati a seguire e analizzare i processi con cui i ricercatori producono, gestiscono e interpretano dati come fonti di conoscenza. Nel mio lavoro in filosofia della scienza mi interesso dei sistemi creati nel corso della storia per concepire descri-

zioni e spiegazioni di come funziona il mondo. Questo si lega in particolar modo all'epistemologia, ossia al ramo della filosofia che studia la maniera in cui si ottiene conoscenza. Mi affascina soprattutto la capacità umana di superare i nostri limiti intellettuali, fisici e sociali per sviluppare idee ingegnose e tecnologie sofisticatissime di considerevole impatto sociale e ambientale. Per questo motivo, ho concentrato le mie ricerche sullo studio delle pratiche e delle strategie usate dagli scienziati per generare, disseminare e analizzare i dati. Ho indagato i modi in cui i dati vengono fatti viaggiare attraverso contesti diversi, e quelli in cui i ricercatori – specialmente coloro che lavorano in istituzioni pubbliche nell'ambito biologico e biomedico – gestiscono e analizzano i loro dati: come li maneggiano e li archiviano, come ne parlano, come giustificano le proprie azioni e soprattutto come li trasformino in fonte di conoscenza. Ho intervistato centinaia di scienziati in tutto il mondo, incluse molte nazioni europee, gli Stati Uniti, il Sudafrica, la Nigeria, la Cina e l'India; e ho partecipato in prima persona alla creazione e allo sviluppo di grosse infrastrutture per la gestione dei dati, comprese le regolamentazioni e le istituzioni introdotte recentemente dalla Commissione Europea per facilitare l'uso dei dati di ricerca per stimolare innovazione e perseguire il bene pubblico[1]. E

[1] Dal 2016 a oggi ho lavorato come esperta in Scienza Aperta e rappresentante della Global Young Academy per la Commissione Europea.

chiaramente, come ricercatrice a capo di svariati progetti, produco e gestisco dati di molti tipi (fotografie, video, documenti storici e migliaia di pagine di trascrizioni delle mie interviste con ricercatori); anche questa un'esperienza che ha contribuito alla visione dei dati che propongo in questo libro.

Fin da subito, questo approccio dimostra come la vita dei dati sia immensamente complessa, specialmente quando questi vengono custoditi, standardizzati, distribuiti e aggregati tramite banche dati e altre piattaforme digitali. L'attenzione al ruolo dei dati nella ricerca scientifica sottolinea come l'utilizzo dei Big Data pone tanti rischi quante opportunità, sia per i ricercatori sia per la società in generale. Molti dei problemi che questo libro mostra essere parte integrante della gestione dei dati vanno ben al di là del mondo della ricerca, e si manifestano tutte le volte che cerchiamo di giudicare le basi empiriche della conoscenza a cui ci affidiamo – che sia quando leggiamo una notizia sorprendente, decidiamo se usare una certa medicina o cerchiamo informazioni su Google. La ricerca scientifica offre un microcosmo nel quale questioni metodologiche ed epistemologiche su come i dati generano conoscenza possono essere affrontate ed estese a situazioni di ricerca altrove (per esempio nel giornalismo, nella politica, nel settore privato e nei servizi pubblici).

Lo studio di come i dati viaggiano attraverso contesti diversi insegna che non c'è modo di separare in maniera nitida i dati scientificamente

rilevanti da quelli che non lo sono. Tutto dipende dalla situazione in cui i dati vengono usati. Dati personali come altezza, peso e situazione familiare, per esempio, possono essere di interesse medico se usati da un dottore per scopi diagnostici; di interesse scientifico se usati da un epidemiologo per studiare la salute della popolazione; di interesse affettivo se raccolti per tracciare un albero genealogico; o di interesse commerciale se usati da una catena di supermercati per identificare preferenze d'acquisto. I dati acquisiscono un valore diverso a seconda delle mani in cui cadono, e la loro rilevanza non è mai riferita a un ambito solo. Inoltre, i dati possono essere valorizzati in tanti modi e per motivi diversi allo stesso tempo. È proprio questa molteplicità che li rende interessanti come oggetti di analisi: da una parte, i dati promettono di documentare aspetti della realtà in maniera fedele e accurata, in modo da facilitarne lo studio; dall'altra, il valore attribuito ai dati ogniqualvolta vengono utilizzati ha un effetto determinante sulla maniera in cui vengono gestiti e sulla loro interpretazione.

Un fattore comune e cruciale alla comprensione del ruolo dei dati nella società contemporanea è il riconoscimento che *tutti* i tipi di dati (che siano o no prodotti e usati da ricercatori, e che siano o no riconosciuti come fonti legittime di conoscenza) hanno un potenziale valore commerciale, specialmente quando vengono aggregati per analizzare e prevedere comportamenti di massa. Basta dare un'occhiata alla lista delle industrie di più grande e veloce successo

a livello nazionale e mondiale per realizzare che le compagnie e le *start-up* che si occupano di analisi dei dati sono cresciute in maniera esponenziale nell'ultima decade, e che i loro servizi sono ormai accettati come una parte indispensabile di qualsiasi settore – dall'organizzazione di una campagna elettorale al lancio di un nuovo prodotto. Collezionare, mobilitare e analizzare dati non è un'occupazione ristretta al mondo della ricerca, ma piuttosto un'espressione fondamentale dello sviluppo economico di stampo capitalistico che caratterizza il libero mercato globale. Non a caso Google, Apple, Facebook e Amazon sono cresciute a velocità vertiginosa fino a essere noverate tra le più ricche e potenti corporazioni del mondo, e *start-up* italiane come EnergyWay, Instal e Cloud4Wi sono riconosciute tra le più promettenti per il nostro sviluppo economico sia a livello nazionale sia all'estero. La crescita economica si basa sempre di più sulla creazione di servizi personalizzati e ottimizzati per rispecchiare le esigenze di clienti specifici, come per esempio il calcolo del consumo energetico di un'unità familiare o le varie apps sui nostri telefonini che promettono di misurare la nostra attività fisica e le condizioni di salute ogni giorno. Questi servizi sono resi possibili dallo sviluppo di sofisticati algoritmi e da strategie per analizzare i comportamenti dei consumatori, che a loro volta funzionano solo se capaci di attingere a vaste fonti di dati sulle persone e sulle loro condizioni di vita (dimora, ambiente, trasporti e così via).

Queste considerazioni illustrano come l'epistemologia, l'etica e l'economia politica siano aspetti complementari e integranti per la comprensione del funzionamento dei Big Data. Solo tramite una visione complessiva del ruolo sociale, culturale, economico e politico dei dati possiamo comprendere l'impatto dei Big Data sul mondo scientifico, e cosa questo voglia dire per la società. Una premessa cruciale per la mia analisi in questo libro è che la scienza dei Big Data non è facilmente discernibile o separabile dal mondo al di fuori della ricerca: come vedremo nei prossimi capitoli, interessi e valore commerciale, politico, affettivo ed economico sono inevitabilmente congiunti con l'eventuale valore scientifico dei dati come fonte di conoscenza.

Capitolo primo
Cosa sono i Big Data?

Big Data tra quantità e qualità

Ci sono tanti modi di caratterizzare i Big Data[1]. Un punto di partenza accettato da molti è la *quantità*. Le tecnologie digitali sviluppate negli ultimi trent'anni consentono enorme capacità di produrre, conservare e analizzare un numero crescente di dati. Non a caso le due caratteristiche più spesso associate ai Big Data sono il volume e la velocità. Il *volume* si riferisce alla dimensione dei *file* usati per archiviare e disseminare i dati, che grazie al potere crescente dei processori elettronici sta aumentando vertiginosamente e in maniera impossibile da percepire chiaramente per il sistema cognitivo umano (chi di noi comprende veramente la differenza tra un triliardo e un quadriliardo, cifre che per gente che lavora

[1] Kitchin & McArdle (2016) identificano ventisei modi di descrivere i Big Data che sono utilizzati nella letteratura scientifica. Il seguente sito ne contiene ancora di più: https://datascience.berkeley.edu/what-is-big-data/

con Big Data sono relativamente normali?). La *velocità* si riferisce al ritmo incalzante e sempre più serrato con cui i dati vengono generati da tecnologie come, per esempio, il sequenziamento del genoma.

Nell'enfatizzare il numero dei dati e il formato digitale, questa definizione non tiene però conto di quattro fattori che riguardano la *qualità* dei Big Data e sono fondamentali per il loro utilizzo:

1) La varietà dei tipi di dati in uso, che comprende dati in formati non-digitali (come per esempio dati stampati su carta) e dati che pur essendo in formato digitale non sono facilmente analizzabili tramite algoritmi (come per esempio le fotografie);

2) il fatto che quel che viene percepito come grande quantità o velocità di dati dipende completamente dalle tecnologie usate per produrli, archiviarli e analizzarli, e quindi cambia continuamente da un anno all'altro. Per esempio, mentre all'inizio del millennio i Big Data erano quelli troppo numerosi per essere annotati con una normale *spreadsheet* di Microsoft Excel, adesso si pensa ai trilioni di dati ottenuti tramite l'uso di *social media* come Facebook; mentre tre secoli fa si pensava alle collezioni di migliaia di osservazioni fatte da metrologi, cartografi e astronomi in giro per il mondo, difficilissime da analizzare e integrare in mappe geografiche senza accesso a computer[2];

[2] Come illustrato da vari studiosi, queste sfide non sono nuove alla storia della scienza. Ricercatori in astronomia, me-

3) la dipendenza dell'analisi dei dati dal contesto in cui essi vengono valutati e usati, che può variare immensamente a seconda della situazione e delle domande poste dagli analisti – un fattore fondamentale per la mia analisi, su cui tornerò nel capitolo tre;

4) il fatto che non è possibile analizzare i Big Data senza avere accesso ai cosiddetti *metadati*, ossia a informazioni sulla loro provenienza (come sono stati generati, rispetto a cosa e in quali circostanze) che permettono agli analisti di valutare se i dati sono affidabili e quali interpretazioni sono plausibili.

Per tenere conto di questi aspetti, altre caratteristiche sono state associate ai Big Data negli ultimi anni (figura 1)[3]. Oltre a varietà dei formati, si parla anche di *varietà* dei fenomeni a cui i dati possono riferirsi, e di approcci usati per analizzarli; di *veridicità* nell'interpretazione dei dati e nel modo in cui rappresentano la realtà[4]; di *validità* dei dati rispetto ai modi in cui ven-

teorologia e tassonomia si occupano di come gestire enormi quantità di dati da centinaia di anni. Per studi storici di questo fenomeno, si vedano le edizioni speciali *Data-Driven Research in Biology and Biomedicine* nella rivista "Studies in the History and the Philosophy of the Biological and Biomedical Sciences", specialmente il contributo di Müller-Wille & Charmantier (2012); e *Historicizing Big Data* nella rivista "Osiris" (Aronova *et al*. 2018).

[3] Per una discussione delle "V" usate per caratterizzare i Big Data (che variano da tre a dieci a seconda di chi li legge), si può consultare per esempio Nordmandeau 2013; Ward & Backer 2013; Mayer-Schönberger & Cukier 2013; Marr 2015; Kitchin 2014; Borgman 2015; Sætnan *et al.* 2018.

[4] Cai & Zhu 2015; Floridi & Illari 2014.

gono analizzati; di *volatilità* nel tempo, ossia la capacità dei dati di rimanere affidabili e leggibili nonostante l'evoluzione di nuove tecnologie di archiviazione; e di *valore* loro assegnato da settori diversi della società, anche questo infinitamente variabile a seconda del periodo storico o della località.

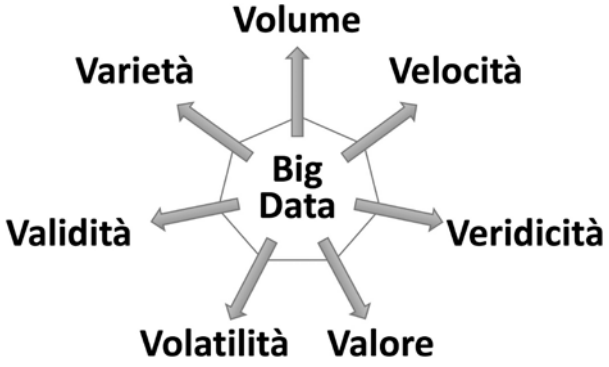

Figura 1. Le sette "V" dei Big Data (realizzazione Michel Durinx).

I Big Data non sono dunque solo "tanti dati". Quello che davvero li caratterizza sono i vari modi in cui vengono prodotti e veicolati tra diversi settori sociali. In questo consistono il potere e la vera promessa dei Big Data: *permettere di instaurare connessioni tra settori e approcci con cui nel passato* è *risultato difficile – sia per barriere sociali sia per motivi tecnici – dialogare direttamente*. Invece che cercare di definire cosa

siano i Big Data in termini di caratteristiche fisiche e quantità, propongo quindi di caratterizzarli in virtù di come vengono usati. I Big Data sono dati di tipi e provenienze diversi che vengono messi in relazione l'uno con gli altri, spesso in forma digitale e in modi che si prestano all'apprendimento automatico, così da produrre nuove forme di analisi e conoscenza. Come discusso da due eminenti sociologi dei dati, boyd e Crawford, l'espressione Big Data segnala "la capacità di esplorare, aggregare e relazionare vasti insiemi di dati"[5]. Per capire come funzionano i Big Data, dobbiamo quindi volgere lo sguardo verso le strutture, le istituzioni e le abilità/professioni che rendono possibile questa capacità.

Rivoluzione o esagerazione? I Dati Aperti e l'approccio Datocentrico

I Big Data sono stati accolti e descritti come una rivoluzione nel modo di acquisire conoscenza – nelle parole di Mayer-Schönberger e Cukier, "una rivoluzione nel modo in cui lavoriamo, viviamo e pensiamo"[6]. Sulla scia di questo tipo di dichiarazioni trionfali, le opportunità insite nell'analisi di Big Data tramite algoritmi sempre più intelligenti ha generato alte aspettative da

[5] Boyd & Crawford 2012, p. 663.

[6] Mayer-Schönberger & Cukier (2017) propongono un approccio trionfalistico al potere dei Big Data che ho fortemente criticato in passato (Leonelli 2014).

parte di governi, industria e ricercatori in tutto il mondo. Da una parte, la disponibilità di Big Data promette di rendere più accurati i modelli usati per calcolare e predirre scenari futuri – che questo riguardi le previsioni del tempo o la probabilità di un determinato individuo di ammalarsi di cancro. Dall'altra, la possibilità di relazionare tra di loro dati di origine diversa promette di aprire nuove direzioni di ricerca, rivelando correlazioni e collegamenti finora invisibili ai ricercatori ma facilmente identificati dai computer. L'uso dei Big Data viene quindi strettamente associato a un'accelerazione non solo nella produzione di nuovo sapere scientifico, ma anche nella traslazione di questo sapere in innovazioni e prodotti per l'uso di tutti i giorni – alcuni dei quali, come per esempio il monitoraggio della diffusione delle malattie infettive e la capacità di prevenire disastri ambientali, possono servire nella risoluzione delle grandi sfide sociali del nostro tempo come il cambiamento climatico, le pandemie e l'inquinamento.

I Big Data sono anche legati a una rivoluzione tuttora in corso nella comunicazione dei risultati della ricerca, che passa sotto il nome di Scienza Aperta. Questa è l'idea – sicuramente non nuova per i ricercatori, ma sempre più problematica da realizzare in pratica vista la competitività del mondo scientifico e la crescente privatizzazione delle scoperte tramite patenti e diritti d'autore – che sia i testi, sia il software, sia i dati prodotti nel corso della ricerca debbano essere resi accessibili in maniera facile e priva di costo a chiunque

voglia usarli. La spinta verso la Scienza Aperta è particolarmente forte nei confronti di progetti sponsorizzati da enti pubblici. Vari governi, europei e non, hanno recentemente argomentato che i loro cittadini, in quanto sostenitori della ricerca pubblica tramite il pagamento delle tasse, hanno il diritto di accedere a tutti i risultati prodotti dagli scienziati grazie a finanziamenti statali, inclusi i loro dati[7]. Questa nozione dei Dati Aperti, o Open Data, è particolarmente attraente per coloro che lavorano su Big Data, con l'idea che più dati sono disponibili da analizzare e mettere in relazione liberamente tra loro, più le opportunità di fare nuove scoperte crescono[8]. Per questo motivo, molte grosse corporazioni, come per esempio l'industria farmaceutica GlaxoSmithKline e il gigante biotecnologico Monsanto (presto Bayern), stanno "aprendo" alcuni dei loro dati, nella speranza che questo faciliti collaborazioni con il settore pubblico e apra la strada ad analisi più fruttuose e meglio informate di quelle che queste compagnie possono sviluppare *in-house*.

[7] Per un'analisi dettagliata del concetto di Scienza Aperta si veda il rapporto della Royal Society (Boulton *et al.* 2012). Per un'analisi delle sue implicazioni per il modo in cui la ricerca viene valutata e valorizzata, si veda il rapporto che ho scritto a questo proposito per la Commissione Europea (European Commission 2018).

[8] Per un'analisi del ruolo degli Open Data nel mondo dei Big Data, si veda l'ottimo rapporto di Science International (2015). I principi associati all'uso di Open Data sono discussi da Mauthner & Parry (2013).

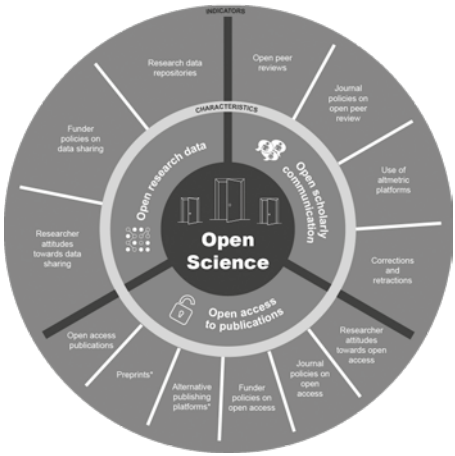

Figura 2. La Scienza Aperta secondo la Commissione Europea (fonte: Open Science Monitor, URL https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor_en).

Il legame tra i Big Data e gli Open Data è *il vero motivo per cui l'avvento dei Big Data ha una portata rivoluzionaria.* Di per sé, ricavare vantaggio grazie a nuove forme di aggregazione dei dati tramite nuove tecnologie non è cosa nuova per il mondo della ricerca. La diffusione della stampa nel Diciassettesimo secolo ha trasformato i modi in cui i biologi comunicano e verificano la scoperta di nuove specie, e ha determinato lo sviluppo di sistemi moderni per quantificare la biodiversità (ossia il numero e il tipo di specie che si trova in ogni ecosistema, che è fondamentale per identificare casi di estinzione

o l'evoluzione di nuove specie). L'invenzione di sistemi di archiviazione come le *punch cards* ha cambiato completamente il lavoro dei demografi e degli epidemiologi nel diciannovesimo secolo, aiutandoli a discernere i comportamenti di intere popolazioni a livello nazionale. Tecniche statistiche e lo sviluppo di calcolatori sempre più potenti nei primi anni del Ventesimo secolo hanno portato un simile livello di innovazione. Ed è evidente come l'intera storia della medicina sia fatta di tentativi di mettere insieme tipi di dati completamente diversi (dalla dieta all'anatomia, dalla storia familiare all'esposizione al clima) per trovare combinazioni di fattori che si associano in maniera regolare all'insorgere di una malattia. Questi sforzi hanno a loro volta generato nuovi approcci all'archiviazione e alla visualizzazione dei dati, come per esempio i tesauri usati per standardizzare il vocabolario medico in modo che i medici si comprendano tra di loro e possano scambiarsi osservazioni; le tecniche usate per assicurarsi che dati acquisiti in un lungo lasso di tempo rimangano comparabili; e i registri che custodiscono dati personali rilevanti per la ricerca epidemiologica in maniera affidabile e confidenziale, che permette agli scienziati di analizzarli senza necessariamente tradire la *privacy* dei pazienti[9].

[9] Si veda per esempio l'analisi delle tecniche di inferenza epidemiologiche di Broadbent (2013) e la discussione della standardizzazione del ragionamento medico proposta da Ankeny (2014).

Non è dunque semplicemente l'opportunità di analizzare vaste quantità di dati con nuove tecnologie che distingue questo momento storico da epoche precedenti nella storia della scienza. Quello che lo rende straordinario è piuttosto *lo status acquisito dai dati stessi* come componente e risultato fondamentale della ricerca scientifica, la cui produzione implica un obbligo alla distribuzione in modo che i dati vengano riutilizzati da più persone possibile (Open Data) e quindi fungano da ponte tramite cui diverse parti della società possono comunicare e lavorare insieme (Big Data). Fin dall'avvento di quel che è generalmente riconosciuto come il primo giornale scientifico in Europa, le "Philosophical Transactions" della Royal Society a Londra (fondato nel 1665), i dati sono stati concettualizzati e manipolati come oggetti privati, proprietà degli scienziati che li producono e che soli hanno la capacità di interpretarli correttamente. Ancora adesso, la stragrande maggioranza delle pubblicazioni scientifiche pubblica solo una piccola parte dei dati prodotti all'interno di un progetto, selezionata apposta in quanto prova più convincente della veridicità dell'interpretazione proposta dagli autori. Tutti i dati che non funzionano bene come prova in questo senso, o che addirittura potrebbero essere interpretati in modo molto diverso, vengono eliminati e tipicamente non sono esposti a ulteriore analisi da parte di altri scienziati. In questo approccio, che chiameremo *teoria-centrico*, l'utilità dei dati sta nel funzionare come prova della plausibilità di una ipotesi già

data, confermando così asserzioni che i ricercatori interpretano come nuove scoperte. La conoscenza teorica è vista come guida e riferimento fondamentale non solo per come condurre ricerche scientifiche, ma anche per come concepirne l'obbiettivo finale. Nella visione teoria-centrica, la cosa più importante che gli scienziati producono sono nuove teorie, tipicamente espresse tramite formule o testi e pubblicate tramite articoli e libri che non a caso riempiono gli scaffali di ogni biblioteca universitaria. Tutti gli altri componenti della ricerca, dai dati ai modelli, dalle tecniche alla strumentazione, sono visti come secondari alla creazione delle teorie.

Questo approccio sta ora lasciando il passo a un modo molto diverso di pensare alla ricerca. Nella visione datocentrica della scienza, l'obbiettivo della ricerca comprende la produzione di diversi tipi di risultati, che includono sia le teorie che i modelli, le tecniche investigative, i metodi sperimentali e i dati stessi. I dati quindi non sono più semplicemente un passo verso la creazione di nuove teorie. Nell'approccio datocentrico, i dati sono visti come entità pubbliche che hanno valore scientifico indipendentemente dal loro ruolo di prova per una determinata ipotesi e che possono essere interpretati in modi diversi a seconda delle abilità e degli interessi dei ricercatori che li analizzano. Stiamo dunque assistendo a u*na rivalutazione radicale del potenziale dei dati nel generare conoscenza* – una rivoluzione in cui gli sforzi e le tecnologie rivolte alla distribuzione, mobilitazione, visualizzazione e integrazione dei

dati sono viste non solo come strumenti di scoperta, ma come importanti modalità di scoperta di per sé stessi[10]. Questa rivoluzione nella comunicazione dei dati ha ricadute anche sul settore privato, in cui il potenziale valore economico dei dati non sempre ne facilita il rilascio in maniera "aperta", ma sicuramente incoraggia la compravendita dei dati e il loro uso sia come transazione sia come valuta per scambi commerciali. Ci sono numerosi modi di "aprire" i dati, molti dei quali non prevedono un'apertura totale e senza costo a chiunque li richieda, ma piuttosto la possibilità di acquisirli a determinate condizioni[11].

Big Data in movimento: il potere delle infrastrutture

L'ascesa del datocentrismo e la combinazione di Big e Open Data ha enormi implicazioni per

[10] Hey *et al.* (2009), un volume prodotto da scienziati della Microsoft e liberamente accessibile su internet, offre vari esempi di questa tendenza. Una difesa dettagliata di questa interpretazione del datocentrismo si trova nel volume *Data-Centric Biology* (Leonelli 2016a).

[11] Gli entusiasti degli Open Data nel campo scientifico vorrebbero vedere tutti i dati di rilevanza per la ricerca come beni pubblici e disponibili senza condizioni a chiunque li richieda. Questa posizione non è però sostenibile, visti i costi sostenuti dal settore privato nel creare dati e la possibilità di abuso legata alla circolazione di alcuni tipi di dati (come per esempio quelli personali). Il panorama di modalità di aperture dei dati è quindi molto vario, con diverse opzioni e possibilità di disseminazione. I principi FAIR, per esempio, sono stati proposti come alternativa a modelli di apertura radicale (WIlkinson et al., 2016).

come la ricerca scientifica – e in particolare l'organizzazione e l'utilizzo dei dati – viene condotta, organizzata, governata e valutata. La capacità dei dati di servire come fonte di conoscenza scaturisce dalla loro *mobilità*: ossia, dalla loro capacità di viaggiare attraverso diverse situazioni di analisi e riuso e di essere relazionati con quanti più tipi di dati diversi possibile. Senza mobilità non ci sarebbero Big Data, perché i dati non riuscirebbero mai a uscire dalle situazioni specifiche in cui sono generati, e sarebbe impossibile aggregare e confrontare tra di loro dati raccolti in circostanze diverse.

L'importanza della mobilità spiega come mai dati prodotti in formati standard e da tecnologie altamente diffuse e convenzionali, come per esempio i dati geografici GPS basati su misurazioni satellitari, sono di altissimo valore scientifico e commerciale: grazie al livello di standardizzazione e digitalizzazione di questi dati, è facile distribuirli e metterli in relazione con altri tipi di dati (come per esempio la posizione di ristoranti o l'intensità del traffico), il che a sua volta permette di generare mappe e indicazioni stradali in tempo reale come quelle di GoogleMaps. La maggior parte dei dati rilevanti per l'analisi scientifica, e specialmente in campo biologico e biomedico, non sono però standardizzati in questo modo. Tutt'altro: questi dati sono tipicamente diversissimi sia nelle tecniche usate per produrli, sia nel formato, sia nel tipo di oggetto a cui si riferiscono. Osservazioni raccolte sulla forma delle foglie di una particolare

specie di pianta, per esempio, possono variare nel modo in cui i ricercatori misurano la superficie della foglia, nei criteri usati per scegliere le foglie da misurare, nel tipo di strumenti usati per rilevare la ruvidità e il colore della foglia, nella frequenza con cui le misure sono prese e perfino nei nomi usati per indicare la pianta in questione e le diverse parti della foglia. Queste variazioni dipendono dalla tradizione scientifica del gruppo e dal tipo di uso che i ricercatori fanno dei dati così raccolti: ricercatori che sperano di investigare correlazioni tra la forma e il profilo genetico della pianta tendono a usare misure, nomi e criteri diversi dai ricercatori interessati alla relazione tra la grandezza delle foglie e la loro velocità di crescita[12].

È importante notare come questa pluralità di approcci non si verifichi per caso o per mancanza di coordinazione tra i ricercatori. Ha invece una funzione epistemica molto precisa. Questi campi scientifici, come pure la ricerca sull'ambiente, il clima e la geologia, sono dedicati allo studio di fenomeni altamente variabili nel tempo e nello spazio. Spiegare le caratteristiche e i comportamenti di particolari specie, individuo, organo o ecosistema inevitabilmente comporta investigare la specificità di questi sistemi in modi che permettano di interagire con essi in maniera efficace (inventando così strategie per aumentare il ritmo

[12] Ho discusso questo esempio nel dettaglio in Boumans & Leonelli (2019). Per una discussione particolareggiata di casi equivalenti in campo biomedico, si veda Leonelli (2017a).

di crescita di piante economicamente importanti come il grano o interventi utili alla cura di particolari malattie, come la terribile piaga causata dal fungo *Fusarium* che minaccia di far sparire per sempre la banana dal mercato occidentale). Nel corso dei secoli, gli scienziati hanno elaborato delle metodologie altamente specializzate per poter studiare e comprendere le proprietà uniche di questi fenomeni. La varietà e la potenziale incompatibilità dei dati che risultano da questi studi non devono quindi sorprendere: sono piuttosto una conseguenza inevitabile della necessità di produrre sistemi di conoscenza che si adattino il più possibile alla natura degli svariati oggetti e processi di interesse scientifico – una situazione che i filosofi chiamano "pluralismo"[13].

Il pluralismo scientifico genera enormi problemi nel mobilitare i dati. Prima di tutto, bisogna trovare modi per archiviarli in maniera sicura e accessibile a chiunque voglia cercare di analizzarli. Gli archivi non possono però essere semplici contenitori ("*data dumps*") in cui sbattere i dati appena vengono prodotti. L'archivio funziona solo se accompagnato da sistemi per cercare e trovare dati in maniera sistematica e rilevante per il loro riuso. Come già segnalato da tanti filosofi del Ventesimo secolo, tra cui Michel Foucault e Jacques Derrida, la chiave per la gestione di un archivio è il modo in cui lo si ordi-

[13] Referenze classiche per un'introduzione al pluralismo sono i volumi Dupré (1983), Kellert *et al.* (2006) e Chang (2012).

na[14]. La struttura dell'archivio, le parole chiave usate per categorizzare e richiamare i dati, i modelli e gli algoritmi usati per visualizzarli – tutti questi elementi hanno un effetto decisivo su come i dati vengono interpretati e riutilizzati, specialmente nel casi in cui l'interpretazione è generata da persone che non hanno avuto nulla a che fare con la creazione dei dati stessi e quindi non ne conoscono le circostanze di produzione. Allo stesso tempo, la necessità di dare ordine ai dati trasforma la loro gestione in un complesso problema epistemologico e manageriale, specialmente visto che non è per nulla chiaro che tipo di ordine sia più o meno utile all'interpretazione dei dati e alla loro trasformazione in nuove conoscenze – e nemmeno se un tale ordine "ideale" esista, visto che l'organizzazione dei dati è condizionata dal contesto e dallo scopo per cui i dati vengono analizzati.

Come abbiamo già visto, particolarmente importante nel contesto dei Big Data è la capacità di relazionare dati ottenuti da fonti diverse e quindi spesso non immediatamente comparabili. Questo ha conseguenze non solo per l'ordine, ma anche per la formattazione dei dati – la forma che gli viene data per farli circolare (per esempio il tipo di *file* usato per digitalizzarli). Implica fare in modo che il formato dei dati sia tale da permettere di visualizzarli tutti insieme, trasformando così gruppi diversi di dati in un'unica fonte di conoscenza. Il formato è dunque parte inte-

[14] Per esempio Foucault (1967) e Derrida (1995).

grante delle condizioni che determinano come i dati vengono interpretati, e la scelta di quale formato usare è lungi dall'essere un problema risolvibile in maniera parzialmente o interamente automatizzata, o basata su soluzioni universalmente applicabili. Tipicamente, la formattazione dei dati richiede giudizi bene informati sia da conoscenza dei formati tecnicamente possibili che da familiarità con il campo e i fenomeni in questione. Questi giudizi sono spesso compito di coloro che mantengono le banche dati (curatori, informatici, archivisti, scienziati dei dati) e sono mirati a valutare le conseguenze scientifiche dei formati possibili e verificare l'impatto di eventuali cambiamenti sui modi in cui i dati possono essere aggregati e analizzati.

La valutazione e la scelta di ordine e formati in cui mobilitare i dati sono operazioni fondamentali sia al funzionamento interno di ogni banca dati sia al modo in cui ognuna di loro si relaziona alle miriadi di altre infrastrutture cariche di dati potenzialmente utili da comparare e relazionare con i propri. Le infrastrutture usate per gestire e mobilitare i dati devono essere il più possibile facili da coordinare e legare tra loro (un'elusiva proprietà che i ricercatori chiamano *interoperabilità*)[15], così da permettere ai dati di entrare a far parte del *network* globale che è l'universo dei Big Data. Tale ambizione si traduce in un'enorme sfida logistica e scientifica, vista la complessa ecologia di queste infrastrutture che

[15] Si veda per esempio Sansone *et al.* (2012).

variano enormemente per scopi, funzionamento, finanziamenti, prospettive future e posizione geografica.

L'archivio – che nel caso di Big e Open Data prende spesso la forma di una banca dati digitale resa disponibile alla consultazione di un vasto pubblico tramite pubblicazione su internet – è un sito di grande potere nell'epistemologia dei Big Data. Il modo in cui le banche dati vengono strutturate determina chi può usarle, dove e per quali fini; quali dati sono accessibili e quali no; e che tipo di interpretazione è possibile darne. Questo potere, che i sociologi della scienza descrivono come il *potere delle infrastrutture*[16], viene esercitato dagli esperti coinvolti nello sviluppo e nella manutenzione delle banche dati, che a loro volta si dividono in categorie diverse a seconda dei tipi di dati e del campo in questione. Dal punto di vista computazionale, troviamo immancabilmente esperti in tecnologie dell'informazione, informatica e ingegneria elettronica, che sviluppano il *software* e l'*hardware* necessario al funzionamento delle banche dati. Ci sono poi esperti in tecniche di archiviazione e catalogazione; esperti nel modo in cui i dati gestiti da una banca dati devono essere formattati così da essere comparabili con dati gestiti da altre banche dati a livello locale o internazionale; ed esperti nel dominio di applicazione dei dati stessi, capaci di gestire, categorizzare e visualizzare

[16] Per esempio Bowker (1994), Hine (2006), Wouters *et al.* (2013), Mongili & Pellegrino (2014), Ribes & Polk (2015).

i dati in modo che siano intelligibili e usabili da utenti che operano in ambiti specifici e diversi tra di loro (da quello medico a quello ambientale, per esempio)[17].

Seguendo i viaggi dei dati

Il modo in cui i dati vengono trasportati e resi riutilizzabili è dunque altamente distribuito. Già all'interno di ogni banca dati difficilmente troviamo situazioni in cui un unico individuo comprende tutti gli aspetti e i tipi di *expertise* coinvolti nella mobilitazione dei dati. Molto più spesso, la comprensione delle scelte e delle tecniche tramite cui i dati sono formattati e mobilitati è distribuita tra svariati individui, ognuno con mansioni e prospettive diverse, e non sempre necessariamente in contatto gli uni con gli altri o comunque capaci di comprendersi a vicenda. Le cose si complicano ulteriormente una volta che i dati viaggiano da una banca dati a un'altra, formando così legami tra infrastrutture che hanno origini e intenti diversi, e usano presupposti e criteri molti differenti l'una dall'altra per selezionare, classificare, formattare e visualizzare i dati. Nel viaggiare da un sito a un altro, da una situazione di ricerca a un'altra, e da un archivio a un altro, i dati stessi si trasformano sia nella loro forma sia nel contenuto – un fatto ine-

[17] Sulle tensioni che caratterizzano la comunicazione tra questi gruppi, si veda Edwards *et al.* (2011).

vitabile e necessario al trasporto dei dati da un contesto all'altro e al loro riutilizzo come parte di Big Data, ma spesso dimenticato da chi pensa ai dati come a rappresentazioni immutabili e fedeli della realtà (tornerò su questo punto, importantissimo dal punto di vista epistemologico, nel capitolo tre).

La complessità e l'importanza epistemica e scientifica di questa situazione mi ha portato a spendere vari anni letteralmente "seguendo i dati": cercando cioè di investigare come esattamente i dati viaggiano, in che condizioni e con quali risultati e implicazioni per i ricercatori coinvolti, la conoscenza prodotta e i settori sociali che da quella conoscenza sono influenzati (figura 3). Questa investigazione è resa necessaria dalla mancanza di tracce del passaggio dei dati, dovuta all'abitudine dei ricercatori di non citare adeguatamente le banche dati che usano nel loro lavoro. Questo rende molto difficile tracciare la maniera in cui dati prodotti in un sito specifico vengano poi assorbiti da una o più banche dati e altre forme di mobilitazione, e da lì trovati da ricercatori e analisti interessati alla loro interpretazione.

I miei studi tipicamente iniziano da un'analisi della storia, della struttura e dell'organizzazione di grosse banche dati usate per custodire e mobilitare dati nel mondo della ricerca. La ricostruzione dei motivi alla base delle scelte fatte in ogni banca dati, e dei modi in cui i dati vengono gestiti, è sempre molto difficile, visto che queste scelte sono fatte in momenti diversi da persone

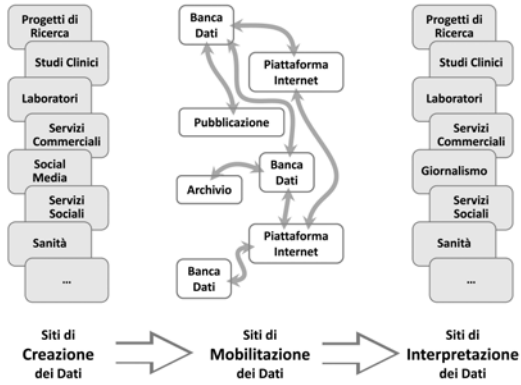

Figura 3. Rappresentazione grafica dei viaggi dei dati volta a illustrare due tipi di movimenti: da siti di produzione a siti di mobilitazione, e da lì a siti di interpretazione. È importante notare come i siti coinvolti in tutti e tre gli stadi di viaggio (di cui nella figura sono menzionati solo alcuni esempi) possano variare enormemente sia nel tempo sia nello spazio: i dati scientifici viaggiano regolarmente da un continente all'altro, in modi segnati da interruzioni e fermate previste o impreviste (dati archiviati da lungo tempo possono per esempio essere improvvisamente riscoperti o rimanere sepolti in un magazzino senza mai rivedere la luce del giorno). (copyright Sabina Leonelli, realizzazione Michel Durinx).

diverse e spesso rispondono a situazione politiche, scientifiche ed economiche differenti che non vengono documentate in maniera sistematica. Le migliori banche dati sono quelle strutturate con una chiara visione dell'impatto che avranno sulla futura interpretazione dei dati e del modo in cui la percezione di questo impatto dipende da presupposti e interessi specifici.

Non tutte le banche dati si basano su una visione chiara dei propri obbiettivi e strategie di gestione dei dati. Molte sono state messe in piedi semplicemente per avere un posto dove custodire i dati prodotti da progetti o servizi, senza una precisa motivazione o razionalizzazione delle scelte fatte nell'organizzare i dati. Questa situazione genera enormi problemi quando il volume dei dati cresce e diventa sempre più difficile organizzarli in modo che continuino a essere facilmente accessibili.

Neanche partire da un'idea ben definita è però sufficiente a creare una banca dati efficiente e rispettata dai suoi utenti. I tipi di pubblico a cui la banca dati si rivolge spesso cambiano negli anni, come anche i loro bisogni e il modo in cui si aspettano di poter consultare i dati – senza parlare dei continui cambiamenti nelle tecnologie e nei *software* usati per mantenere archivi digitali. Anche il personale delle banche dati cambia, spesso senza grande continuità e consapevolezza di come le scelte fatte a monte condizionano le decisioni prese dieci o venti anni dopo. Di conseguenza, i modi in cui i dati vengono estratti e analizzati dalla stessa banca dati tendono a diversificarsi sempre di più man mano che passa il tempo – una situazione che ho documentato in vari casi in cui ho seguito i dati sia attraverso la loro gestione nelle banche dati sia nel modo in cui vengono usati dai ricercatori che utilizzano queste infrastrutture. Tracciare il legame tra la struttura di un archivio e il modo in cui i dati vengono utilizzati non è un'operazione

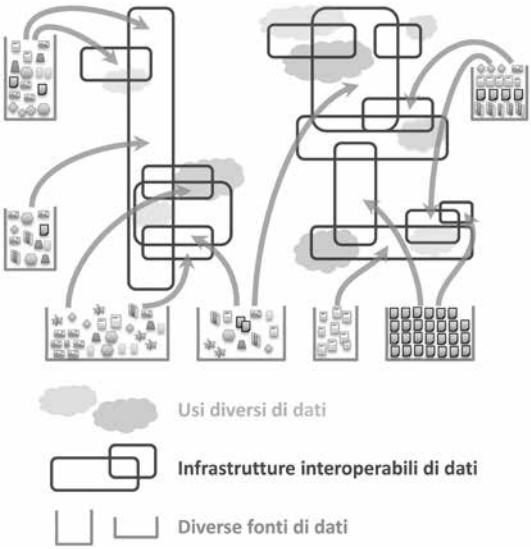

Figura 4. Rappresentazione grafica dei viaggi dei dati, volta a illustrare la complessità delle interdipendenze tra varie fonti di dati, le infrastructure coinvolte nel loro trasporto e le porzioni di dati limitate che vengono messe in uso per facilitare determinate scoperte. Per quanto complicata, questa rappresentazione statica e astratta non tiene conto dell'ulteriore complessità di aggiornare e correlare queste risorse nel corso del tempo (copyright Sabina Leonelli, realizzazione Michel Durinx).

che si può risolvere una volta per tutte: questo legame va invece problematizzato e regolarmente aggiornato a seconda di come il mondo intorno all'archivio continua a evolversi.

Ricostruire i viaggi dei dati attraverso sistemi di gestione e analisi così complessi, caratterizzati da tante relazioni di dipendenza da realtà socia-

li, biologiche e tecnologiche in continuo cambiamento, è un modo per sottolineare la natura tecnica e poco trasparente di questi processi e l'impatto sostanziale e drammatico che hanno sulla ricerca datocentrica e i suoi risultati. Le difficoltà insite nel ricostruire i viaggi dei dati spiegano almeno in parte il fatto che il ruolo fondamentale degli archivi è spesso sottovalutato da coloro che pensano ai Big Data come a un'opportunità di accelerare il progresso scientifico e allo stesso tempo renderlo meno costoso – inclusi molti governi europei, che vedono l'avvento dei Big Data come un modo per tagliare la spesa dedicata alla ricerca senza però danneggiare l'innovazione. La mia ricerca ha documentato come questa sia in realtà un'illusione pericolosa, dalle conseguenze potenzialmente drammatiche sia per il mondo della ricerca sia per la società in generale. La manutenzione di banche dati che permettono la mobilitazione e l'interpretazione dei dati richiede investimenti considerevoli, l'implementazione di strategie di partecipazione *sociale* e controlli e regolamentazioni adeguati, nonché l'assunzione di nuovi tipi di esperti specializzati nei vari aspetti della gestione dei dati. In mancanza di questi investimenti, l'utilizzo di Big e Open Data rischia di trasformarsi in un disastro con conseguenze severe per l'affidabilità e l'impatto sociale della conoscenza così prodotta. Il prossimo capitolo esplora cinque tipi di rischio concretamente associati alla mala gestione dei Big Data.

Capitolo secondo

Segnali di allarme: cinque modi in cui i Big Data danneggiano la ricerca

1. *Conservatismo: il problema dei dati vecchi*

Nonostante l'uso dei Big Data venga comunemente visto come fonte di innovazione, ci sono buoni motivi per sospettare che affidarsi ai Big Data rafforzi invece il conservatismo nei processi e nei risultati della ricerca. Questo avviene, prima di tutto, come conseguenza della crescente complessità delle banche dati e delle rispettive scelte di *standard* e algoritmi usati per selezionare, classificare e visualizzare i dati. Questa complessità rende sempre più difficile e costoso mantenere le banche dati in modo che riflettano i più recenti sviluppi computazionali e scientifici, e crea un considerevole incentivo verso il mantenimento di categorie prestabilite – un atteggiamento dogmatico che poco si addice alla ricerca.

Come già notato nel primo capitolo, il modo in cui i Big Data vengono ordinati e formattati ha un effetto determinante per quello in cui poi

vengono analizzati e interpretati. Una delle ragioni per questo effetto è che le categorie usate per ordinare i dati hanno inevitabilmente valore semantico. In altre parole, queste categorie riflettono le assunzioni concettuali e metodologiche di creatori, gestori e/o utenti dei dati stessi, che a loro volta derivano dal loro modo di concepire il mondo. Nei miei lavori filosofici ho analizzato il valore di questa impalcatura concettuale, argomentando che deve essere vista come una forma di teoria – che chiamo *teoria classificatrice*. Uno dei migliori esempi di questo tipo di teoria sono le cosiddette "bio-ontologie": un sistema di organizzazione dei dati estremamente popolare nelle scienze della vita, che si basa sull'identificazione e definizione precisa di fenomeni a cui i dati possono riferirsi, e sulle relazioni tra di loro[1]. Teorie come le bio-ontologie sono utilissime – molti direbbero indispensabili – nell'organizzare i dati, ma estremamente laboriose da modificare e aggiornare.

In uno studio pubblicato nel 2011, io e alcuni degli artefici della bio-ontologia più usata nella biologia contemporanea (la Gene Ontology, inventata per organizzare dati genetici) abbiamo discusso vari casi in cui il sistema è stato adeguato a nuovi sviluppi scientifici[2]. Un esempio è la rielaborazione legata alla scoperta che il citoscheletro, tradizionalmente interpretato come un organo esterno al nucleo cellulare, in realtà

[1] Leonelli 2012; 2016a.
[2] Leonelli *et al.* 2011.

contiene frammenti di nucleo. Questa scoperta ha costretto i curatori della Gene Ontology a cambiare non solo la loro definizione di citoscheletro, ma anche i modi in cui questo termine era relazionato con altri termini che descrivono l'anatomia della cellula. Questi cambiamenti hanno avuto un impatto significativo sulle relazioni tra i dati associati a queste categorie; e il risultato è che un biologo che cercasse quali dati sono associati al citoscheletro, nel 2011 avrebbe trovato risposte diverse da quelle del 2013. Il caso più eclatante è la stessa categoria del "gene", che è notoriamente ambigua e indica cose diverse a seconda della tradizione intellettuale e della prospettiva teorica adottata da chi la usa. C'è chi pensa ai geni come al codice della vita, parte fondamentale della riproduzione; e chi invece li vede come una delle tante componenti (che includono gli stimoli ambientali e la struttura dell'intero tessuto cellulare) che interagiscono nello sviluppo dell'organismo[3]. A seconda del punto di vista, l'importanza che viene data a dati genetici rispetto a dati cellulari cambia notevolmente, e la teoria classificatrice usata per ordinare questi dati riflette fedelmente queste preferenze concettuali.

Se c'è una cosa che il datocentrismo proprio non è, quindi, è la "morte della teoria"[4]. Non

[3] Barnes & Dupré 2008; Müller-Wille & Rheinberger 2012.

[4] L'idea che l'analisi dei Big Data implichi la morte della teoria, ossia l'idea che i presupposti teorici usati per produrre e gestire i dati non sono rilevanti alla loro interpretazione, ha

solo i dati vengono sempre prodotti in relazione a precise aspettative e presunzioni concettuali, ma le stesse strategie scelte per mobilitare i dati hanno sempre un significato preciso dal punto di vista teorico, le cui basi empiriche e implicazioni è importante rendere esplicite e rivalutare regolarmente per controllare che siano ancora adeguate – un compito reso ancora più urgente a causa dalla continua accelerazione nella produzione di sapere scientifico legata al numero crescente di investimenti nella ricerca a livello mondiale. Nonostante questo, all'interno di ogni banca dati – specialmente quelle che funzionano da più di dieci anni – è comune trovare confusione e ignoranza su motivi e criteri usati nel corso degli anni per scegliere, adattare e aggiornare le modalità di gestione dei dati. Da una parte, c'è spesso troppa gente in ballo per riuscire a tener dietro in maniera sistematica a: chi ha fatto quali cambiamenti, perché e sulla base di quali criteri. Dall'altra, il ritmo di lavoro e gli incentivi che caratterizzano la gestione dei dati non incoraggiano chi lavora in questo campo a prendersi il tempo di documentare le proprie scelte in maniera dettagliata. La maggioranza delle infrastrutture usate per mobilitare i dati si nutre di finanziamenti a corto termine e risponde all'obbligo di dimostrare la propria utilità il prima possibile ai finanziatori, in modo da avere l'opportunità di

fatto scalpore soprattutto in seguito a un editoriale di Chris Anderson pubblicato nella rivista "Wired" nel 2008, chiamato appunto *The end of theory* (Anderson 2008).

ricevere ulteriori fondi. Chi gestisce i dati è dunque sotto la grande pressione di sviluppare una risorsa che diventi quanto prima funzionale. C'è poco tempo per valutare con calma i vantaggi e gli svantaggi di diversi modi di ordinare dati, inclusa la semantica delle categorie che vengono adottate a questo scopo, e ancora meno per tenere nota delle svariate azioni intraprese per creare e mantenere l'archivio. Difficilmente, quindi, il personale che decide di cambiare lavoro dopo aver passato vari anni a impostare una banca dati si lascia dietro una guida scritta sul modo e sulle ragioni usate fino ad allora per ordinare e formattare i dati.

Il risultato è quello che Bruno Latour ha chiamato una "scatola nera": una tecnologia costruita sulla base di svariate assunzioni – per esempio, riguardo alle caratteristiche degli oggetti sui cui i dati sono stati prodotti, o su quale sia il migliore formato per mobilitare i dati in questione – che diventano componenti invisibili e quindi incontestabili dell'infrastruttura[5]. I filosofi William Wimsatt e James Griesemer, lavorando in collaborazione con la psicologa Linda Caporeal, hanno ulteriormente affinato l'idea di scatola nera nel loro lavoro sulle "impalcature" usate nell'evoluzione culturale, che è facilmente applicabile al caso dei Big Data. Le impalcature sono le presupposizioni concettuali, sociali e materiali necessarie alla costruzione di teorie, tecnologie o infrastrutture. Una volta che il processo di co-

[5] Latour 1987.

struzione è concluso, le impalcature solitamente finiscono per essere rimosse, lasciando però – proprio come le impalcature usate per costruire edifici – un'impronta importante su quello che è stato costruito. Per chi è interessato a intervenire sulla costruzione o a verificarne la solidità, è indispensabile capire quali impalcature siano state usate nel processo di costruzione. Ma questo vuol dire ripercorrere passo per passo i processi e le strategie usati per impostare la costruzione, il che non è affatto semplice una volta che il prodotto in questione è finito e la memoria di quel che è stato fatto viene progressivamente persa[6].

In una situazione in cui le scelte fatte nell'impostazione dei dati diventano sempre più difficili da ricostruire man mano che passa il tempo, diventa più difficile anche individuare parti dell'infrastruttura che necessitano di aggiornamento o che sono state nel frattempo contestate da nuovi sviluppi scientifici. La situazione si aggrava notevolmente una volta che teniamo conto della moltitudine di banche dati che popolano ogni branca della ricerca scientifica, ognuna delle quali contiene assunzioni che influenzano la circolazione e l'interoperabilità dei dati, e che però spesso non vengono aggiornate in maniera regolare e affidabile. Giusto per dare un'idea concreta dei numeri, la prestigiosa rivista scientifica "Nucleic Acids Research" pubblica ogni anno un'edizione speciale sulle nuove banche dati di rilevanza per la biologia molecolare: le nuove infrastrutture

[6] Caporael *et al.* 2014.

incluse nell'edizione speciale erano 56 nel 2015, 62 nel 2016, 54 nel 2017 e 82 nel 2018. Queste non sono che una piccola parte delle centinaia di banche dati create ogni anno in relazione alle scienze della vita, che a loro volta sono legate alla quantità ancora maggiore di banche dati usate in medicina, scienze ambientali e agronomia. Il fatto che molte di queste infrastrutture abbiano fondi a corto termine genera una percentuale crescente di risorse che rimangono consultabili *online*, ma sono da tempo defunte, nel senso che nessuno è più coinvolto nel prendersene cura – una situazione non sempre visibile agli utenti, che spesso si affidano al contenuto delle banche dati senza verificare se queste siano attivamente mantenute. A che punto queste infrastrutture diventano obsolete? E quali sono i rischi insiti nel tessere un tappeto sempre più esteso di infrastrutture che dipendono l'una dall'altra, vista la disparità tra i modi in cui vengono gestite e le difficoltà nell'individuare e comparare i loro prerequisiti e le teorie e impalcature usate per costruirle?

Uno di questi rischi è proprio il conservatismo dilagante che viene dall'insistere sul riciclo di dati vecchi, le cui caratteristiche e modalità di gestione diventano sempre più opache col passare del tempo, invece che incoraggiare la produzione di dati nuovi le cui caratteristiche rispondano in maniera specifica alle esigenze e le circostanze di chi li usa. In discipline come la biologia e la medicina, che studiano organismi viventi e quindi per definizione in continuo svi-

luppo e evoluzione, la crescente fiducia in dati vecchi di cui si nutre l'analisi dei Big Data è particolarmente allarmante. Non è affatto detto, per esempio, che dati raccolti su batteri o funghi prelevati da una foresta e portati in laboratorio dieci, venti, cento anni fa siano una fonte affidabile per spiegare il comportamento della stessa specie di batteri adesso o nel futuro – e infatti i biologi che studiano l'evoluzione di patogeni dannosi all'uomo hanno grossi problemi nell'usare le varie fonti di dati disponibili attualmente in modo da ricostruire la diffusione geografica di infezioni e patologie nel corso degli ultimi decenni[7]. Lo stesso vale per dati raccolti su culture cellulari in vitro, che sono soggette a stretti controlli proprio per assicurarsi che rimangano le stesse da una generazione all'altra, e nonostante questo tendono comunque – come la vita tende a fare, secondo quanto abbiamo tutti imparato da Jeff Goldblum in *Jurassic Park* – a evadere i controlli e mutare in modi imprevedibili[8].

Incoraggiare i ricercatori a tornare continuamente ad analizzare e rianalizzare dati vecchi ha sicuramente il vantaggio di ridurre gli investimenti necessari a creare nuovi dati, un fattore che fa gola a molte delle agenzie che sponsorizzano l'uso di Big e Open Data – specialmente visti i costi astronomici spesso legati all'acquisto delle tecnologie e dei materiali usati per produrre dati, per cui anche un sempli-

[7] Leonelli 2018.
[8] Landecker 2007.

ce esperimento biologico può venire a costare centinaia di migliaia di euro. Questa scelta può però generare un altro tipo di costo, molto più difficile da quantificare e di impatto potenzialmente disastroso per la scienza: lo sviluppo di ricerca conservatrice e poco creativa, volta a massimizzare quello che già sappiamo sul mondo naturale invece che esplorarne la continua evoluzione in modi sempre nuovi e di crescente sofisticazione.

2. *Inaffidabilità: il problema dei dati inattendibili*

Un altro enorme rischio che diventa evidente appena considerate le condizioni di mobilitazione dei dati è la difficoltà di valutarne la qualità. Uno dei problemi principali dell'affidarsi a una banca dati abbandonata è la mancanza di garanzie che i dati che vi si trovano siano ancora credibili. Ancora più preoccupante è il fatto che – abbandonate o no – le banche dati attualmente in circolazione variano enormemente nel loro approccio al controllo della qualità dei dati. Alcune infrastrutture non hanno nessun meccanismo per filtrare i dati affidabili da quelli che lo non lo sono, e giustificano questa scelta con l'idea che i loro utenti hanno interessi e criteri di valutazioni diversi, e non è quindi compito della banca dati decidere quali dati sono accettabili per quale scopo. Altre infrastrutture vedono questo controllo come una responsabilità

fondamentale delle banche dati nei confronti dei loro utenti. Questa posizione è frequente specialmente in aree in cui gli utenti non hanno né il tempo né le conoscenze necessarie per verificare i dati trovati nelle banche dati, e quindi finiscono per dare per scontato che i dati messi su internet siano automaticamente affidabili.

La diversità di approcci tra banche dati già pone ostacoli considerevoli alla circolazione dei dati, specialmente visto che gli utenti non sono sempre informati e consapevoli delle differenze tra i criteri di inclusione usati da banche dati diverse. Ancora più problematico è il fatto che, anche quando i gestori delle banche dati si preoccupano di garantirne l'affidabilità, non esistono criteri universali per giudicare quali dati siano "buoni" e quali no. È vero infatti che i giudizi sulla qualità e sull'affidabilità dei dati come fonte di conoscenza variano fortemente a seconda degli scopi della ricerca e della maniera in cui i dati vengono usati[9]. Per esempio, molti ricercatori clinici vedono i dati sull'espressione genetica (i cosiddetti *microarrays*) come intrinsecamente affidabili, perché prodotti tramite strumentazione standardizzata e quindi più facilmente comparabili tra loro rispetto ai dati fisiologici acquisiti sui pazienti. Ho intervistato molti biologi che invece pensano a questi dati come inaffidabili,

[9] Il volume sulla qualità dell'informazione edito da Phyllis Illari e Luciano Floridi (Floridi & Illari 2014) contiene numerosi esempi e riflessioni sulle implicazioni di questo fenomeno; si veda anche Edwards (2010) e Leonelli (2017b).

perché i campioni e la strumentazione usata per produrli sono altamente sensibili a cambiamenti di temperatura e illuminazione nei laboratori in cui vengono prodotti[10]. Il risultato della diatriba è che i *microarrays* vengono spesso incorporati in banche dati biomediche ma non in quelle biologiche – una situazione paradossale visto che molti ricercatori che lavorano in oncologia, per esempio, usano entrambi questi tipi di banche dati come fonte per i loro studi.

In generale, l'adozione di formati e strategie di classificazione standardizzate possono essere di grande aiuto nel rendere le scelte di gestione dei dati chiare e facilmente rivisitabili, facilitando così anche i controlli sulla qualità e sull'affidabilità dei dati stessi. Il pluralismo che caratterizza il lavoro scientifico, e quindi la diversità dei metodi usati dai ricercatori per studiare la natura, rendono però impossibile l'adozione di *standard* universali – e anche quando gli *standard* esistono, grandi sforzi sono dedicati all'adattarli alle condizioni specifiche della ricerca in questione[11]. L'applicabilità degli *standard* dipende dall'abilità di coloro che li usano di mettere i dati in relazione con determinate situazioni di ricerca. Esistono certo metodi statistici per controllare vari aspetti dei dati, dal modo in cui vengono aggregati alla loro omogeneità. La statistica non aiuta però a verificare se i metodi usati per produrre i dati hanno senso, viste le proprietà degli

[10] Leonelli (2012).
[11] Bowker & Star 1999; Timmermans & Epstein 2010.

oggetti dello studio in questione. Per esempio, un progetto sul ritmo circadico delle piante – i diversi modi in cui il loro metabolismo funziona durante le 24 ore di una giornata – può produrre dati perfetti dal punto di vista tecnico, ma che possono essere giudicati problematici rispetto ai geni che vengono analizzati, alla frequenza con cui i dati vengono generati e alle condizioni in cui le piante stesse sono cresciute.

Oltre che dipendere dall'uso, la valutazione della qualità dei dati dipende infatti dalla familiarità che i ricercatori hanno con le condizioni in cui i dati vengono prodotti. Chi ha esperienza degli strumenti e dei materiali usati per generare i dati dispone di una conoscenza intima dei controlli e dei metodi usati, che a sua volta gioca un ruolo significativo nel giudicarne l'affidabilità. Una volta che i dati viaggiano in contesti diversi da quelli in cui sono stati prodotti, questa conoscenza intima viene spesso a mancare, creando ulteriori ostacoli nella valutazione della qualità dei dati e della credibilità di particolari interpretazioni. Inoltre, come tutte le attività umane, anche la produzione dei dati può essere fatta male o con motivazioni che hanno poco a che fare con la ricerca della verità. Senza criteri chiari con cui valutare i dati, è impossibile individuare dati inattendibili o prodotti apposta per creare una visione distorta della realtà. Basti pensare agli sforzi fatti dall'industria del tabacco negli ultimi cinquant'anni, ben documentati dai colleghi Naomi Oreskes e Brian Conway, per produrre dati che provassero che il fumo fa

bene alla salute[12]. Il rischio nel collezionare Big Data è quello di costruire un enorme castello di carte, mettendo insieme vaste quantità di dati senza riguardo per le potenziali differenze nella loro attendibilità come fonte di conoscenza.

Questo problema è chiaramente riconosciuto e ampiamente discusso dai gestori delle banche dati e dai ricercatori che ne fanno uso. Tutti sono d'accordo sulla soluzione. Da una parte, le banche dati devono fare il possibile per verificare le motivazioni e il rigore dei metodi usati per produrre i dati, e acquisire quante più informazioni possibile sulle circostanze di produzione dei dati, in modo da poter fornire ai propri utenti un quadro il più completo possibile della storia dei dati. Le banche dati sono quindi responsabili della *decontestualizzazione* dei dati: la scelta di quali dati includere e soprattutto di quali informazioni dare agli utenti sulla loro provenienza (meta-dati), due fattori fondamentali alla mobilitazione dei dati al di fuori dei loro siti di provenienza. Dall'altra parte, visto che i criteri di qualità variano a seconda della situazione di uso, ogni utente che si appresta a creare nuove interpretazioni dei dati deve prendersi la responsabilità di indagarne l'affidabilità nei confronti del proprio ambito di ricerca. Questa è la *ricontestualizzazione* dei dati: gli utenti esercitano il proprio giudizio, basato sulle loro conoscenze e abilità, per formulare una loro interpretazione del significato

[12] Oreskes & Conway 2010.

dei dati e della loro credibilità[13]. Le banche dati forniscono supporto indispensabile al processo di ricontestualizzazione tramite la selezione di metadati che descrivano in modo appropriato le scelte fatte nella produzione e organizzazione dei dati, dando quindi agli utenti la possibilità di valutare per conto proprio il significato e l'impatto di esse[14].

La differenza tra le conoscenze e gli interessi di chi produce i dati e di chi li ri-interpreta per nuovi scopi è alla base del problema della qualità dei dati, visto che questi due gruppi di ricercatori spesso divergono nei criteri che usano nella loro valutazione. Allo stesso tempo, questa differenza è anche il motivo per cui l'analisi dei Big Data ha la capacità di portare a nuove scoperte: è proprio il fatto che chi riusa i dati tende a valutarli con occhi nuovi – e nuove competenze e interessi – che genera la possibilità di interpreta-

[13] Un fattore che va menzionato, anche se non c'è spazio per discuterlo in maniera approfondita in questo libro, è l'importanza che l'accesso a materiali e strumenti originariamente usati per produrre dati può avere nella loro ricontestualizzazione. Spesso l'unico modo per un biologo di comprendere il potenziale significato di dati trovati su internet è di valutare a sua volta gli stessi campioni (che siano colture cellulari o particolari varianti genetiche della stessa specie) usati da chi ha creato i dati. Alcune banche dati, come quelle legate agli organismi modello, includono informazioni sui campioni originali all'interno delle loro funzionalità, ma questo richiede considerevoli risorse inaccessibili alla maggior parte delle infrastrutture (Ossorio 2011, Leonelli 2016a).

[14] Per una discussione dettagliata sui processi di decontestualizzazione e ricontestualizzazione, e sul loro significato scientifico e filosofico si veda Leonelli (2016a).

zioni innovative e diverse da quelle date in passato. Questa osservazione è fondamentale per capire l'indissolubile legame tra l'analisi dei Big Data e il rischio dell'affidarsi a dati inattendibili. Il rischio di inaffidabilità è parte integrante del modo in cui i Big Data vengono mobilitati e analizzati, e la scienza datocentrica non potrebbe esistere in sua assenza.

3. *Mistificazione: il problema dei dati parziali*

Il rischio di affidarsi a dati vecchi o inattendibili può già sembrare un prezzo alto da pagare per la consultazione dei Big Data, ma non è secondo me quello più alto. Il rischio più significativo consiste piuttosto nel problema dei dati parziali: ossia il fatto che le banche dati forniscono informazioni molto selettive, rappresentando dunque solo una piccola parte della realtà senza però necessariamente fornire gli strumenti necessari per analizzare le conseguenze di questo limite.

La natura selettiva delle banche dati non è un problema di per sé. Ogni studio scientifico necessariamente semplifica la realtà in modo da poterne investigare un aspetto specifico, e questa capacità di focalizzare e modellare il mondo pezzo per pezzo è alla base del successo della ricerca scientifica. Quello che mi preoccupa non è quindi il fatto stesso di ridurre la realtà, nella sua infinita complessità, a un campione di dati limitati nella loro portata e rappresentatività. Questa

riduzione è una componente essenziale di ogni processo di analisi. Il "rischio di mistificazione" è posto invece dalla tendenza degli utenti dei Big Data a dimenticarsi che quello che stanno manipolando non è un campione comprensivo né particolarmente ben bilanciato della realtà, ma piuttosto una selezione fatta in parte per limiti pratici e in parte per ragioni concettuali. Ignorare la natura selettiva dei dati messi *online*, e le ragioni per cui quei dati e non altri sono accessibili e analizzabili tramite algoritmi, aiuta sicuramente ad accelerare l'interpretazione dei dati – ma allo stesso tempo facilita la produzione di interpretazioni che mistificano i fatti invece che aiutare a comprenderli.

Quali sono le maggiori fonti di distorsione nella selezione dei dati inclusi negli archivi digitali? Uno sguardo veloce alle banche dati più usate al mondo per scopi di ricerca rivela immediatamente che la maggior parte delle iniziative di successo si occupa di dati facilmente trattabili dal punto di vista computazionale come per esempio le sequenze genetiche, che sono espresse tramite lettere (A, C, G, T) facilmente analizzabili per mezzo di algoritmi[15]. Costruire una banca dati che collezioni dati difficili da digitalizzare e analizzare, come per esempio fotografie o disegni fatti a mano, richiede investimenti mol-

[15] Per uno studio dettagliato di come si è arrivati a questa notazione e del suo impatto sulla digitalizzazione della biologia molecolare, si vedano Rheinberger (2011), November (2012) e Stevens (2013).

to più consistenti che un'infrastruttura dedicata a dati in forma numerica o simbolica. Esistono certo iniziative molto interessanti che supportano la mobilitazione e l'analisi computazionale di immagini complesse, come per esempio nel caso delle risonanze magnetiche usate per rivelare strutture anatomiche. Ma questi rimangono sforzi più contenuti rispetto all'enorme quantità di risorse devolute alla circolazione di dati più facilmente trattabili – un fattore che influenza e riduce in maniera significativa la diversità dei dati accessibili *online*.

Un'altra osservazione immediatamente evidente quando si considera il panorama dei Big Data è che le infrastrutture che godono di migliore fama sono quelle gestite da istituti ben finanziati, ricchi di risorse umane, situati in località di potere nel mondo scientifico (Boston, Singapore, Cambridge, Oxford, Pechino, San Francisco) e rigorosamente mantenuti e discussi in lingua inglese. Non solo: questi istituti e le rispettive banche dati tendono a lavorare in tradizioni di ricerca particolarmente in voga e ben riconosciute dal mondo politico e industriale – motivo per cui beneficiano di fondi considerevoli e di notevole visibilità nel mondo scientifico, il che le rende capaci di pubblicizzare i loro sforzi a livello globale e incoraggiare ricercatori di tutto il mondo ad accettare i loro presupposti, piuttosto che perdere tempo a contestarli[16]. Questo

[16] Con Rachel Ankeny abbiamo analizzato nel dettaglio queste configurazioni di modalità di ricerca e strutture isti-

genera un'enorme distorsione nella provenienza dei dati resi accessibili tramite queste infrastrutture. Non solo, chi crea le banche dati tende ad adattarle alle proprie preferenze metodologiche e concettuali, generando così archivi che contengono soprattutto i dati propri e di colleghi con preferenze simili. È anche chiaro che chi è esposto in prima persona alla creazione delle banche dati ne capisce meglio il funzionamento, ed è quindi più capace di usare queste infrastrutture a suo vantaggio – sia per mobilitare i propri dati sia per analizzare quelli degli altri.

Questi privilegi hanno implicazioni pesanti per il livello e la qualità della partecipazione attratta dalle banche dati in questione. Uno studio che il mio gruppo di ricerca ha condotto in vari laboratori africani tra il 2014 e il 2016 dimostra che chi lavora in questi ambienti è spesso in soggezione rispetto a chi fa ricerca in ambiti meglio riconosciuti a livello internazionale[17]. Questo vuol dire che, da una parte, molti ricercatori africani non osano contestare l'operato di banche dati americane o europee, indipendentemente da quanto questo sia corretto e utile dal loro punto di vista e, dall'altra, che raramente questi ricercatori riescono a farne uso in maniera otti-

tuzionali e amministrative, che chiamiamo repertori di ricerca (Ankeny & Leonelli 2016). Importanti per comprendere l'impatto delle istituzioni sulla ricerca sono il lavoro di Pestre (2003) e quello di Cambrosio *et al.* (2014) sui "regimi di conoscenza" nonché quelli di Daston (1995) e Strasser (2011) sull'economia morale delle comunità scientifiche.

[17] Bezuidenhout *et al.* (2016; 2017).

male. Una delle ragioni è il presupposto, spesso usato da queste banche dati, che gli utenti abbiano accesso alle versioni più nuove del *software* necessario all'analisi, mentre molti ricercatori che non lavorano a Stanford o al MIT dispongono di programmi meno aggiornati – e non hanno sempre accesso a una connessione internet abbastanza potente da permettergli di scaricare e lavorare con Big Data[18]. Un'altra ragione è il timore di condividere i propri dati tramite infrastrutture internazionali espresso da molti dei ricercatori che abbiamo intervistato, causata dalla paura che ricercatori con più visibilità e mezzi più potenti di loro se ne approfittino (per esempio, per produrre analisi in maniera molto più veloce di quella disponibile a coloro che hanno strumenti e infrastrutture meno sofisticate)[19].

Questa situazione non risulterà sorprendente ai lettori, vista la complessità dei processi di mobilitazione dei dati descritti finora e la varietà di competenze, strumenti e investimenti necessari per riuscire a realizzarli. È importante però riflettere sul risultato di questo forte legame tra il potere (economico e culturale) dei siti di mobilitazione dei dati e il tipo di dati che diventano disponibili come Big Data. I dati resi disponibili tramite infrastrutture digitali, che costituiscono

[18] Vermeir *et al.* (2018).

[19] Questo fattore è particolarmente evidente nei laboratori meno attrezzati (Bezuidenhout *et al.* 2017), ma paradossalmente viene citato anche da ricercatori che lavorano nei migliori laboratori al mondo (Fecher *et al.* 2015, Levin *et al.* 2016, Levin & Leonelli 2016).

la fonte della maggior parte degli studi sui Big Data, sono estremamente selettivi e privilegiano i risultati di gruppi di ricerca di successo, che lavorano in ambito anglofono e in situazioni di agio economico (relativamente ad altri). Difficilmente i dati prodotti da gruppi di ricerca in siti meno visibili e ben attrezzati vengono inclusi, e le condizioni di inclusione, nei pochi casi in cui questa avviene, sono dettate dalle élite universitarie. Questo genera un'enorme disparità sia (1) nelle fonti e nei tipi di dati che possono essere analizzati come Big Data (che possiamo chiamare *disparità di rappresentazione*) sia (2) nella possibilità da parte di ricercatori in tutti il mondo – senza parlare dei cittadini la cui vita è profondamente influenzata da come i loro comportamenti sono datificati e quindi resi trattabili per analisi scientifica – di criticare gli strumenti, gli *standards* e le infrastrutture usati per mobilitare i dati (che chiameremo *disparità di partecipazione*). Queste due forme di disparità sottolineano e peggiorano notevolmente la diseguaglianza che già esiste tra tradizioni di ricerca e la gestione del rapporto tra scienza e società a livello globale, in modi che generano una preoccupante distorsione e potenziale mistificazione della realtà rappresentata dai Big Data.

Un esempio ovvio della disparità di rappresentazione è il fatto che il gruppo meglio documentato dalla maggior parte dei dati biomedici in circolazione è quello delle classi medio-alte della popolazione di nazioni altamente sviluppate, e soprattutto coloro di origine caucasica e

genere maschile. Non è certo una novità nella storia della scienza. Quello che diventa particolarmente preoccupante nell'era dei Big Data è la facilità con cui questo problema di campionamento e rappresentazione viene messo da parte. Discipline come la medicina, la sociologia e l'epidemiologia hanno passato gli ultimi due secoli a sviluppare metodi molto sofisticati per identificare le fonti di distorsione e discriminazione nei loro dati, nonché i modi in cui la ricerca viene condotta e interpretata così da tenerne conto. La scelta di che settore del mondo investigare, e quindi di quale tipo di dati ottenere, è una parte fondamentale del lavoro scientifico e del modo in cui i ricercatori giustificano le scoperte che ricavano dai loro studi. Questa riflessività e questa capacità di aprirsi alla critica, fondamentale per la credibilità della scienza stessa, non sono sempre rispettate nell'ambito dei Big Data. C'è addirittura chi pensa che l'accesso ai Big Data renda inutile ogni riflessione sul campionamento e la rappresentatività dei dati che vengono analizzati: se si accetta l'idea che i Big Data danno informazioni su *tutto*, si accetta anche l'idea che basti metterli insieme per ottenere una piattaforma empirica affidabile e bene equilibrata – e in questo senso incontestabile – per la ricerca futura[20].

Questo libro invece sostiene che i Big Data diano informazioni su *molto poco* e in maniera che tende a precludere, o comunque a rendere

[20] Per esempio in Mayer-Schönberger & Cukier (2013).

difficoltoso, ogni tipo di opposizione costruttiva. L'idea che i Big Data racchiudano una rappresentazione completa della realtà è un'illusione che sta distruggendo lo spirito critico con cui i ricercatori affrontano l'analisi e l'interpretazione dei dati empirici. Il rischio consiste della distorsione o nell'occlusione dei motivi per cui i ricercatori selezionano dati rilevanti per la loro ricerca. A causa della parzialità insita nelle fonti e nei tipi di dati che vengono mobilitati *online*, i ricercatori che lavorano su Big Data si trovano spesso a lavorare su campioni scelti non per motivi scientifici ma per motivi di pura convenienza – economica, politica o culturale – che perlopiù non vengono chiaramente segnalati nelle banche dati, e il cui potenziale di distorsione non viene quindi tenuto in conto nell'analisi dei dati stessi. Questa situazione riflette un fenomeno sociale molto più ampio, ossia il monopolio crescente di compagnie con grandi risorse finanziarie e tecnologiche – Google in particolare – sullo sviluppo di strumenti di gestione e analisi dei dati. L'immediata conseguenza è il ruolo sempre più passivo giocato dal resto della società nel determinare quali dati contino, per cosa, e come vengono utilizzati.

Non intendo qui negare che l'avvento dei Big Data abbia portato reale cambiamento nel modo in cui si fa ricerca, ma piuttosto contestare l'ottimismo con cui alcuni analisti hanno descritto l'impatto di queste tecnologie sulla geografia e l'inclusività della scienza. Gli storici Bruno Strasser e Paul Edwards, per esempio,

sottolineano come i Big Data abbiano aperto opportunità per posti come Singapore e la Cina di sfidare l'egemonia scientifica occidentale[21]. Io sono più pessimista: vedo gli sviluppi a Oriente come il risultato di enormi investimenti piuttosto che dell'adozione della tecnologia di per sé, e l'uso dei Big Data come un'opportunità di rafforzare ulteriormente il ruolo della ricchezza economica nel determinare che tipo di ricerca si ritenga significativa e affidabile. Contrariamente alla visione di Big e Open Data come portatori di democrazia e istigatori di partecipazione sociale nella ricerca, il modo in cui la scienza è governata e finanziata non sembra essere sfidato dai Big Data, ma anzi la diseguaglianza di potere e visibilità tra diverse nazioni e comunità scientifiche continua a crescere. Il divario digitale tra chi non solo ha accesso ai dati ma ha anche la capacità di usarli e chi non ce l'ha si sta allargando – portandoci quindi da una situazione di *digital divide* a una situazione di "*data divide*". Il che ci conduce a esaminare il prossimo problema.

4. *Corruzione: il problema dei dati disonesti*

Il trionfo di criteri finanziari e opportunistici su quelli scientifici nella scelta di quali dati vengono messi *online* segnala una profonda tensione nel mondo della scienza datocentrica. Specialmente nel settore della ricerca pubblica

[21] Strasser & Edwards 2018.

esistono sforzi di grande sofisticazione volti a stabilire modi per ricontestualizzare i dati, capirne la provenienza e le modalità di gestione e, quindi, interpretarli in maniera che rifletta le circostanze e i limiti sotto cui sono stati prodotti e mobilitati. Proprio grazie al livello di apertura e trasparenza di queste iniziative, che contano per esempio i gruppi di lavoro della Research Data Alliance e il lavoro associato alla costruzione in corso della European Open Science Cloud, è possibile documentare i mille modi ingegnosi con cui chi gestisce banche dati nella sfera pubblica cerca di controbattere i rischi delineati finora e, se non proprio evitarli, perlomeno, mitigarli. Rimane però il fatto che la stragrande maggioranza dei dati prodotti a scopo di ricerca (senza contare i dati prodotti in altri ambiti e poi assorbiti da banche dati in modo da supportare nuove scoperte) è generata in ambito di interesse commerciale e spesso privatizzato.

Questa tendenza ha due conseguenze di grossa rilevanza per il ruolo dei dati nella ricerca scientifica. Per prima cosa, la produzione e il possesso di dati in ogni ambito sono stati mercificati nel senso marxista del temine. Non solo i dati sono trattati come beni commerciali e come tali soggetti alle leggi del libero mercato, ma la compravendita dei dati e la crescente consapevolezza del loro fondamentale ruolo nell'economia capitalista ne hanno stimolato enormemente la mobilitazione attraverso piattaforme digitali[22].

[22] Per un'analisi di questo fenomeno e di come si è evoluto

Una porzione crescente delle risorse necessarie per collezionare, custodire e analizzare i Big Data è quindi sotto il controllo di enti con interessi primariamente commerciali, sia in ambito pubblico (governo) sia in ambito privato (corporazioni attive nell'ambito della ricerca), con sempre meno opportunità date a chi ha meno potere economico e sociale nel partecipare alla costruzione di strumenti e strategie di analisi e interpretazione. In altre parole, stiamo assistendo alla costruzione di un'oligarchia (per non dire di un monopolio) sull'informazione e la produzione della conoscenza, in cui la logica dell'esclusione gioca un ruolo più forte della spinta verso l'inclusività associata all'idea di Open Data.

La seconda conseguenza della privatizzazione dei Big Data è che questa rende più difficile l'apertura non solo dei dati stessi ma, soprattutto, delle informazioni su come sono stati prodotti e gestiti per facilitarne l'interpretazione. La mercificazione dei dati si accompagna infatti a un'ambiguità sul loro *status* di bene pubblico o privato. Prendiamo il caso dei dati personali, ossia dati che identificano caratteristiche di un particolare individuo (come per esempio nome, indirizzo, numero di conto corrente). Non è raro che corporazioni come Facebook, Google e le centinaia di compagnie nate nell'ultimo decennio per agevolare l'acquisizione e la vendita di dati per scopi commerciali difendano allo stesso

negli ultimi anni si vedano, per esempio, Thrift (2005), Beer (2016) e Srnicek (2017).

tempo due idee apparentemente opposte: quella che i dati personali sono in gran parte beni pubblici in quanto facilmente accessibili (come per esempio il nostro nome, cognome e indirizzo) – e che quindi sono riutilizzabili per qualsiasi scopo una volta che le persone in questione hanno dato il loro assenso – e quella che i dati personali sono, se non necessariamente privati, almeno *privatizzabili*, e quindi soggetti a essere venduti e comprati come ogni altro prodotto. Nonostante la contraddizione, l'intero modello finanziario di queste compagnie si basa sull'acquisizione e sul riuso dei dati personali di milioni di persone in tutto il mondo in modi che eludono il più possibile le regolamentazioni vigenti nelle nazioni in cui questi utenti risiedono. La confusione regnante su cosa voglia dire "possedere" i dati è ampiamente e regolarmente sfruttata per incoraggiare i cittadini a concedere i diritti sull'uso dei propri dati a una quantità crescente di compagnie, spesso sulla base di accesso a servizi che agevolano la loro vita di tutti i giorni (come informazioni sul traffico, il cinema più vicino e le previsioni del tempo per il fine settimana), ma senza consapevolezza dei modi in cui questi dati possono essere riutilizzati e venduti come merce sul libero mercato.

Il fatto che questa compravendita dei dati – e specialmente dei dati personali – può avere conseguenze pesanti per individui e comunità sta fortunatamente diventando sempre più chiaro anche a chi non è direttamente coinvolto in questo tipo di lavoro. Uno dei modi più eclatanti

di diffondere il messaggio è l'introduzione di iniziative volte a regolamentare il traffico di dati e gli usi a cui possono essere sottoposti. L'esempio più progressista di questo tipo di regolamentazione è la legislazione per la protezione dei dati personali (General Data Protection Regulation, o GDPR) varata nel 2018 dalla Commissione Europea, il cui intento è precisamente quello di proteggere i cittadini dall'abuso dei loro dati, problematizzarne la compravendita e il riuso e istigare forme più sofisticate di archiviazione e mobilitazione dei dati stessi[23]. La GDPR segue la scia di numerosi rapporti preparati da istituzioni come l'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE) e le Nazioni Unite, che da tempo consigliano l'introduzione di regolamentazioni di questo genere, nonché di misure per favorire il dibatto e l'educazione pubblica sulle conseguenze dell'uso di tecnologie e misure di sorveglianza sempre più sofisticate da parte di enti pubblici e privati.

Quello che è forse meno evidente e discusso nella sfera pubblica è quanto sia la privatizzazione dei dati sia la regolamentazione dei loro viaggi abbiano implicazioni molto serie per il mondo della ricerca e per la conoscenza che viene prodotta. Prima di tutto, queste forme di controllo e mobilitazione si traducono in un ulteriore modo di selezionare quali dati vengono mobilitati in maniera aperta e trasparente. Le

[23] Per dettagli sulla GDPR e su come funziona si veda il recente libro di Curioni (2017).

corporazioni di solito rilasciano dati che ritengono di minor valore commerciale e per la cui interpretazione ritengono di aver bisogno di aiuto dal settore pubblico. Questo introduce un'ulteriore distorsione delle fonti e dei tipi di dati accessibili *online*, con dati più dispendiosi e complessi – ma anche potenzialmente più interessanti e fruttuosi come base empirica per la conoscenza – tenuti all'interno degli archivi di chi li produce, al riparo da sguardi indiscreti e dalla possibilità di riuso per altri scopi. Anche i modi in cui i cittadini – inclusi i ricercatori – vengono incoraggiati a interagire con le banche dati e i siti di interpretazione dei dati vengono radicalmente ristretti a forme di partecipazione che generano ulteriore valore commerciale, come per esempio la valutazione di applicazioni digitali ("*rate your app*"!) che vengono usati per migliorare i prodotti – e quindi il rendimento economico – dei prodotti sviluppati sulla base dei dati. Vari sociologi hanno recentemente descritto questo tipo di partecipazione sociale come una forma di sfruttamento o comunque di lavoro non retribuito; un fenomeno che gli economisti descrivono come parte essenziale della *sharing economy*[24].

Queste modalità di sfruttamento dei dati – e di chi li produce e/o fornisce – rappresentano una spinta verso il potenziamento del valore economico dei dati *a scapito di quello scientifico*. Come discusso nell'introduzione di questo libro,

[24] Sullo sfruttamento insito nell'uso dei Big Data, si veda Prainsack (2017), Prainsack & Buyx (2017) e Srnicek (2017).

i dati hanno sempre tanti tipi di valore che vanno da quello scientifico a quello affettivo, commerciale, politico e culturale. Per esempio, Niccolò Tempini ha mostrato come i dati personali estratti da *social media* come Twitter e usati per produrre conoscenza medica siano necessariamente valutati sia come elementi scientifici sia come prodotti commerciali, informazioni costitutive del senso di identità degli individui ed elementi la cui condivisione è alla base della formazione di gruppi sociali[25]. Questi modi di valorizzare i dati non sono necessariamente in conflitto l'uno con l'altro, ma tensioni e divergenze emergono spesso e hanno un impatto decisivo su come i dati viaggiano e vengono interpretati. Nel caso della compravendita di dati personali tra compagnie che si occupano di analisi, il valore dei dati in quanto prodotti commerciali – che include la valutazione della velocità e l'efficienza con cui l'accesso a certi tipi di dati può aiutare a generare nuovi prodotti – ha spesso priorità su questioni scientifiche come, per esempio, la rappresentatività, l'affidabilità e il conservatismo dei dati e dei metodi usati per analizzarli. In molti casi, questo può sfociare in decisioni scientificamente problematiche o anche solo disinteressate a investigare le conseguenze delle assunzioni e delle procedure utilizzate – una mancanza di interesse che si traduce facilmente in ignoranza sulle discriminazioni, sulle ineguaglianze e sui potenzia-

[25] Tempini (2017); si vedano anche Harris *et al.* (2016) e Leonelli (2016a).

li errori nei dati che vengono presi in considerazione. Questo tipo di ignoranza è altamente strategica ed economicamente produttiva, visto che permette l'uso dei dati senza dover farsi scrupoli sulle potenziali implicazioni scientifiche e sociali. In questo scenario, il giudizio sulla qualità dei dati si riduce a un giudizio sulla loro utilità per produrre l'analisi o la previsione richiesta dal cliente nel breve termine. Non ci sono incentivi, in questo sistema, che incoraggino la considerazione delle implicazioni di questo tipo di analisi sul lungo termine.

Il rischio è dunque che la mercificazione dei dati si accompagni a una crescente separazione dei dati stessi dal proprio contesto, senza alcuna possibilità di ricontestualizzazione. L'interesse nella storia dei viaggi dei dati, la pluralità del loro valore affettivo o scientifico e il riconsiderarne la provenienza nel lungo termine scompare, e viene sostituito dalla fossilizzazione crescente del valore economico dei dati – in un processo parallelo a quello che Marx ha notoriamente descritto come "alienazione". È chiaro come questo tipo di valorizzazione dei dati apra la strada alla produzione, alla gestione e all'analisi di dati per scopi disonesti e tendenziosi. E qui torniamo alla questione della *post-verità* menzionata nell'introduzione di questo libro. In situazioni in cui il valore commerciale attribuito ai dati supera di gran lunga l'interesse circa il loro valore scientifico, è perfettamente possibile abbandonare completamente la ricerca di dati veritieri, corretti e la cui manipolazione preve-

da una rappresentazione affidabile della realtà. Proliferano così procedure di fabbricazione di dati con il solo scopo di fornire credibilità a posizioni e ipotesi prestabilite e convenienti dal punto di vista politico, commerciale o sociale. In questi casi, la produzione dei dati non può avere l'esito di modificare quello in cui già si crede, perché gli unici dati che contano sono quelli che possono essere usati per sostenere e rafforzare opinioni già presenti, o per migliorare prodotti già pianificati indipendentemente dal loro valore scientifico e sociale.

Lo scandalo scoppiato nel 2018 intorno all'acquisizione e all'uso di dati personali da Facebook per scopi politici è un'ottima illustrazione di questo meccanismo all'opera. In questo caso, ricercatori impiegati dall'istituto privato Cambridge Analytica sono stati pagati da varie entità politiche, inclusa la campagna a favore dell'uscita del Regno Unito dall'Unione Europea nel referendum del 2016, per analizzare i dati personali presenti su Facebook con lo scopo di generare metodi efficaci di persuasione grazie ai quali cittadini individuati come "vulnerabili" potessero venire bombardati da messaggi che li incitavano a votare per un determinato esito. La veridicità dei dati stessi viene così giudicata puramente in funzione dell'efficacia dell'intervento sociale e politico che la compagnia è pagata per fare. Dati giudicati irrilevanti o che non si accordano a queste preferenze vengono eliminati dall'analisi, e una volta individuati i segmenti della popolazione su cui intervenire rimane poco

interesse ad esplorare potenziali contraddizioni o problemi nei dati stessi (o a confrontarli con altre fonti di informazione) – generando così una conoscenza parziale, inaffidabile e corrotta. In casi come questo, la preferenza nel valutare il valore dei dati su base commerciale è in netto contrasto con la funzione principale dei dati, che è epistemica. Se questo non viene tenuto in conto quando si producono, disseminano e analizzano dati, si rischia non solo la completa commercializzazione della ricerca ma soprattutto la produzione di interpretazioni plasmate dagli interessi e dagli scopi di specifiche entità – che queste siano grosse corporazioni o specifici gruppi di individui – in maniera impossibile da contestare[26].

5. *Danno sociale: il problema dei dati sensibili*

Il caso di Cambridge Analytica esemplifica anche un altro problema, lasciato per ultimo perché pervade non solo tutti gli altri problemi analizzati finora, ma anche il ruolo sociale della ricerca, in maniera più generale. Questo è il problema dei dati sensibili, ossia dati che possono essere usati per rappresentare caratteristiche di individui o gruppi di persone. L'analisi sempre più sofisticata di questi dati, e l'opportunità di relazionarli tra di loro offerta dai Big Data, aprono la porta a una comprensione sempre mi-

[26] Leonelli (2016a); Ebeling (2016); Sunder Rajan (2017); Murphy (2017).

gliore delle esigenze reali dei cittadini, e quindi a decisioni politiche, sociali e ambientali meglio informate e più efficienti. Allo stesso tempo, una gestione sbagliata di questi dati, o l'adozione di metodi o scopi di ricerca problematici dal punto di vista etico e sociale, possono facilmente generare enormi danni alle persone coinvolte – rendendole per esempio vulnerabili a sorveglianza e manipolazione da parte di malintenzionati, oppure generando conoscenza inaffidabile o parziale su di loro, che viene poi usata da servizi sociali, commerciali, medici o assicurativi per stabilire che tipo di assistenza dare e a quali condizioni.

Viste le promesse di innovazione sociale e tecnologica fatte in relazione all'uso dei Big Data, che variano dalle auto senza pilota a modi per ottimizzare il consumo di energia, è facile sottovalutare la gravità e la varietà dei problemi causati da una gestione dei dati immorale e poco attenta alle implicazioni sociali del loro uso. Leggi come la GDPR e varie altre regolamentazioni rivolte alla protezione dei dati tendono a concentrarsi sui diritti degli individui, segnalando come l'efficienza con cui tipi di dati diversi vengono relazionati tra di loro può generare rischi considerevoli per singoli cittadini. Questo è sicuramente vero, specialmente vista la diseguaglianza, la corruzione e l'inaffidabilità insite in alcuni dei sistemi usati al momento per produrre, mobilitare e interpretare i Big Data. Ma quello che è forse ancora più preoccupante, e meno discusso in ambito pubblico, è il rischio

posto dall'uso di Big Data per *gruppi* di cittadini. La considerazione di potenziali danni alla collettività invece che all'individuo è importantissima, perché consente di estendere notevolmente l'insieme dei dati definiti come *sensibili*. Questioni etiche e sociali emergono non solo in relazione a dati personali, ma anche a dati che documentano le caratteristiche di una particolare località e forniscono quindi indicazioni che possono essere usate per giustificare interventi di vario tipo, che a loro volta possono avere effetti positivi o negativi sui residenti. Dati sulla salute, la distribuzione demografica e l'uso dei trasporti fatto dagli abitanti di un determinato quartiere, per esempio, possono essere usati per giustificare la costruzione di un parco o l'approvazione di un nuovo blocco di edifici. Visti in quest'ottica, dati sul clima, l'ambiente e la biodiversità presente in una certa area geografica, pur senza identificare un determinato individuo, possono avere conseguenze significative per le comunità umane che in quell'area risiedono – il che ne fa dati sensibili. E proprio questa caratteristica – la rilevanza di questi tipi di dati nella produzione di conoscenza sulle abitudini e le preferenze delle persone – ne sottolinea sia il potenziale che i rischi.

Uno dei progetti in cui sono stata coinvolta recentemente è un tentativo (tipico nell'ambito della ricerca sui Big Data) di relazionare dati medici con dati climatici e dati estratti da *social media* come Twitter, per capire quanto l'incidenza di asma e altre malattie respiratorie stagionali, segnalata da persone che si lamentano dei loro

sintomi su Twitter, sia associata a particolari condizioni climatiche. L'uso di dati estratti da Twitter (come per esempio "oggi non riesco proprio a respirare" e "sarebbe stato un bel giro in bici se non fosse stato per il maledetto polline!") è particolarmente produttivo per questo tipo di ricerca a causa della mancanza di dati che documentino tipi meno violenti di asma – a sua volta dovuta all'alto numero di situazioni in cui i pazienti si sentono male ma non vanno necessariamente dal dottore, e quindi non lasciano traccia del loro malessere negli archivi medici. Questi dati sono usati per generare proiezioni su quando e dove la gente comincia a soffrire di asma, che vengono poi testate confrontandole con dati medici su quando e dove occorrono casi più gravi, sul livello di spesa pubblica in ogni ospedale legata a malattie respiratorie e sul tipo di vegetazione a cui i cittadini di quelle regioni sono esposti (per verificare che tipo di polline o di erba è presente e associato ai vari focolai). Questo insieme di Big Data permette così di generare spiegazioni della frequenza e delle caratteristiche di epidemie di asma nonché della loro associazione al microclima e alla flora locale con l'obbiettivo di aiutare la sanità pubblica a gestire gli ospedali e a decidere quante risorse investire nella cura dell'asma in regioni diverse, e quando e quanto rafforzare o diminuire queste risorse nel corso dell'anno.

Il progetto è un ottimo esempio delle grandi opportunità offerte dai Big e Open Data: una conoscenza dettagliata delle condizioni in cui l'asma emerge, che possono sicuramente

aiutare studi medici dediti alla prevenzione e alla cura delle malattie respiratorie; e una base fattuale per organizzare e motivare interventi statali e scelte su come equipaggiare gli ospedali nel futuro. Questi sono universalmente visti come vantaggi volti al bene pubblico. Eppure, anche in questo tipo di ricerca troviamo rischi considerevoli associati alla selezione, gestione e potenziale interpretazione dei dati. La scelta di usare Twitter, per esempio, è condizionata fortemente dal fatto che questo è uno dei pochi social media che permette (in maniera limitata) il riutilizzo dei suoi dati per scopi di ricerca – mentre Facebook e Instagram chiedono un sacco di soldi per rilasciare dati sui loro utenti, e sono quindi più difficili da usare per un progetto di ricerca finanziato da modesti fondi pubblici. Purtroppo Twitter è anche una piattaforma con utenti di tipo abbastanza specifico: la maggior parte è tra i venticinque e i quarantacinque anni, residente in città piuttosto che in campagna, di classe medio-alta e con un bagaglio culturale più alto rispetto alla media della popolazione. I dati estratti da Twitter tendono quindi a rappresentare malamente gruppi residenti in zone rurali e con meno accesso alla sanità pubblica, che però sono quelli più esposti ai cambiamenti stagionali associati all'asma (come la stagione del polline e del taglio dell'erba nei campi).

Il mio ruolo nel progetto è proprio quello di analizzare le potenziali conseguenze di questa parzialità e individuare modi in cui i ricercatori possano tenerne conto nel corso della loro ricer-

ca e nelle forme di conoscenza che finiscono per produrre. In questo caso, l'esigenza di tenere conto delle implicazioni sociali della natura dei dati sensibili usati nel progetto ha creato l'esigenza di fare ricerca supplementare sugli utenti di Twitter e sulla loro distribuzione sul territorio rispetto al resto della popolazione, il che aiuta a quantificare i limiti di questi dati nel rappresentare la popolazione nella sua interezza; e interrogarsi a ogni passo sui modi in cui lo specifico campione fornito dai dati Twitter si relaziona a parti della popolazione che ne sono escluse. Queste considerazioni prendono tempo e possono essere interpretate come un rallentamento indebito dei processi di ricerca. Possono anche essere interpretate come un indebolimento dei risultati acquisiti, specialmente da parte del governo che preferisce ricevere risposte chiare ai propri quesiti invece che risposte temperate da ammonimenti sulle limitazioni della conoscenza ottenuta e sulla potenziale discriminazione insita nei dati. I ricercatori del progetto avrebbero certo più successo mediatico e finanziario se proponessero una soluzione facile, veloce e non ambigua per prevedere l'insorgere di un'epidemia tramite l'analisi automatica di Big Data, senza stare a preoccuparsi delle possibili eccezioni o del potere discriminatorio di un tale strumento per coloro che non vengono rappresentati nei dati che lo alimentano. L'onestà con cui i ricercatori del progetto segnalano i limiti delle loro proiezioni – e i modi in cui queste devono essere temperate e corrette a seconda della situazione –

è invece altamente preferibile dal punto di vista scientifico e sociale, proprio perché la conoscenza così prodotta segnala in maniera esplicita le condizioni in cui può essere ritenuta affidabile.

Quello di come utilizzare i dati sensibili è dunque un problema sia epistemico sia etico, in cui non è possibile distinguere i criteri usati per produrre conoscenza affidabile da quelli usati per assicurarsi che i metodi usati non rinforzino discriminazioni sociali ingiuste e arbitrarie. La mancanza di una netta separazione tra condotta scientificamente corretta e condotta eticamente corretta è particolarmente rilevante nel caso dei Big Data. La riflessione sulle conseguenze sociali dell'utilizzo di dati vecchi, parziali, inaffidabili e corrotti è sempre inesorabilmente legata a una valutazione del valore etico delle scelte fatte nella selezione, gestione e interpretazione dei dati. In questo senso, il valore scientifico e quello etico dei dati non solo non sono necessariamente in conflitto ma sono tipicamente associati l'uno all'altro. Questo non è sempre riconosciuto dagli scienziati stessi, la cui fretta di raggiungere risultati – dovuta all'immensa pressione esercitata da università e sponsor – li rende restii a prendersi il tempo di valutare le eventuali conseguenze sociali dell'uso di certi dati. Anche nel caso del progetto sull'asma appena descritto, l'attitudine degli scienziati coinvolti potrebbe essere descritta come schizofrenica: da una parte sono genuinamente contenti di lavorare con un filosofo e molto interessati a produrre previsioni accurate e affidabili dal punto di vista sia scien-

tifico sia sociale, dall'altra sono frustrati dal fatto che gestire i dati in maniera etica – per esempio indagando esattamente che fonti di dati eravamo autorizzati a usare, che tipo di discriminazione potevano contenere e come potevano essere resi visibili ad altri interessati a replicare alla nostra ricerca – richiede tempo e sforzo, e quindi inevitabilmente prolunga il processo di ricerca e ne rende i risultati meno sensazionali.

Questa schizofrenia è purtroppo familiare e ben comprensibile a qualsiasi ricercatore, visto il modo in cui la ricerca accademica è finanziata e valutata in molti paesi europei, primo tra tutti il Regno Unito ma anche, in maniera crescente, l'Italia. Specialmente considerate le aspettative generate dai profeti del "riuso facile" dei dati, i finanziatori di ricerca fatta sui Big Data cercano risultati velocissimi e di grande impatto economico, spesso senza tenere conto del tempo e della fatica necessari a curare e a indagare i dati in modo da verificarne l'affidabilità, la rappresentatività e l'impatto sociale[27]. I meccanismi di attendibilità tipicamente riconosciuti nel mondo scientifico non aiutano questa situazione, in quanto la credibilità e la visibilità dei ricercatori è spesso associata alla quantità di pubblicazioni prodotte e alla natura sensazionalistica delle loro dichiarazioni, che attrae l'attenzione dei giornali

[27] Per un esame approfondito degli incentivi all'opera nel mondo accademico di fare ricerca intorno a Big e Open Data si veda il rapporto che ho scritto per l'Unione Europea nel 2017 (European Commission 2017).

e dei finanziatori. Il potenziale di corruzione e l'incentivo a lasciarsi indietro ogni scrupolo morale è dunque continuamente presente, anche se in forme diverse, sia per la ricerca condotta in ambito privato sia per quella condotta in ambito pubblico – ed è facile vedere come questo si scontri con l'ambizione di produrre basi solide, affidabili e socialmente accettabili per la conoscenza.

L'etica come parte integrante della scienza

Questa è la situazione in cui studiosi di quella che il filosofo Luciano Floridi chiama l'infosfera – ossia il modo in cui l'introduzione di tecnologie digitali sta cambiando il mondo – avvertono il potenziale distruttivo dell'uso dei Big Data e l'urgente bisogno di focalizzare sforzi di gestione e uso dei dati in maniera attiva e riflessiva in direzione del miglioramento della condizione umana[28]. Nelle parole di Floridi, "le ICT dischiudono grandi opportunità, le quali, però, implicano l'enorme responsabilità intellettuale di comprendere tali tecnologie e sfruttarle nel modo più appropriato"[29]. A questi avvertimenti si deve aggiungere un elemento importante specialmente in ambito di uso dei dati per scopi di

[28] Floridi (2017). Si veda anche il rapporto su Data Governance preparato dalla Royal Society & British Academy nel 2017.

[29] Floridi (2017).

ricerca: è essenziale che *questioni etiche e sociali siano viste come parte integrante delle esigenze tecniche e scientifiche associate alla gestione e all'analisi dei dati*. La gestione etica dei dati non si ottiene solamente tramite la regolamentazione della compravendita della ricerca e della gestione dei dati privati, che pure è un enorme passo avanti come dimostrato dalla GDPR, né con l'introduzione di controlli sulla ricerca che viene finanziata, per quanto anche questi siano importanti. Per garantire che l'uso dei Big Data sia il più scientificamente e socialmente avanzato possibile, è necessario andare oltre la concezione dell'etica come qualcosa di esterno ed estraneo alla ricerca, che si occupa dei presupposti e delle conseguenze della scienza ma non dei contenuti. Questioni e implicazioni etiche devono invece essere sollevate a ogni passo della ricerca fatta sui Big Data, diventando quindi una componente fondamentale dell'educazione e dell'operato di coloro che si curano dei dati e dei metodi usati per visualizzarli e analizzarli[30]. Giudizi e scelte etiche si celano in ogni aspetto della gestione dei dati, incluse quelle scelte che a prima vista sembrano puramente tecniche e quindi neutrali dal punto di vista sociale.

Questo diventa particolarmente evidente in contesti di ricerca lontani da obbiettivi di valore pubblico e collettivo, che invece seguono scopi determinati dai governi per guadagno politico a breve termine, da corporazioni con grandi inte-

[30] Leonelli (2016b).

ressi finanziari o da metodi di valutazione scientifica incentrati sulla quantità invece che sulla qualità dei risultati. Ma non solo: anche la ricerca fatta nel nome del bene pubblico può essere problematica quando non si ferma a valutare cosa esattamente "bene pubblico" significhi e per chi, e quali siano le conseguenze più ampie dell'elaborazione di certi tipi di analisi di Big Data. Uno dei problemi più grossi in questo ambito, nonché più difficili da controllare e gestire, è quello del cosiddetto "duplice uso": ossia il fatto che una tecnologia sviluppata con buone intenzioni possa sempre anche essere sfruttata in modi eticamente problematici (come per esempio nel caso dell'estrazione di dati personali da Twitter, che possono essere usati per migliorare la vita di gruppi e di individui, ma anche per sorvegliarne il comportamento e maturare sistemi di sorveglianza e assicurazione sempre più predatori e strumentali)[31]. Questo non è un fenomeno limitato ai Big Data: tutte le tecnologie sono soggette allo stesso problema, visto che non è mai possibile controllare come un certo strumento viene usato una volta che è stato prodotto e distribuito. Il fatto che tecnologie come algoritmi e banche dati abbiamo sempre un potenziale duplice uso è quindi ovvio, una situazione che rende le aspettative trionfalmente positive associate all'uso di Big e Open Data ancora più assurde.

Certo, Big e Open Data hanno il potenziale di rafforzare la partecipazione nella ricerca, l'a-

[31] Rappert & Selgelid (2013).

vanzamento della conoscenza e l'efficienza dei processi investigativi: ma per le stesse ragioni hanno anche il potenziale di compromettere – o addirittura di sabotare – la qualità e l'affidabilità della conoscenza prodotta con metodi scientifici, danneggiando così irreparabilmente la percezione sociale del valore della scienza. In questo contesto è vitale trovare modi di gestire i dati così da incoraggiare il rispetto dei diritti e della dignità umana sia a livello individuale sia a livello collettivo[32]. L'integrazione dei Big Data esemplifica il legame indissolubile tra problemi tecnici, come quelli di custodire i dati in maniera sicura e di verificarne la validità, e problemi etici, come quello di stabilire il potenziale impatto dell'uso di Big Data su individui e comunità[33]. La gestione della confidenzialità e della sicurezza dei dati è critica per lo studio e il trattamento dei soggetti della ricerca e per il modo in cui i dati sono relazionati tra di loro[34].

[32] Vayena & Tasioulas 2016.

[33] Dove *et al.* 2016; Mittelstadt & Floridi 2016; Leonelli 2016b.

[34] Tempini & Leonelli 2018.

Capitolo terzo

Come evitare il peggio: l'approccio relazionale all'epistemologia dei Big Data

Alla base di questo libro è il tentativo di capire cosa voglia dire al giorno d'oggi produrre conoscenza che ha basi empiriche e in che modo la scienza continui a rapportarsi e a distinguersi da altre forme di conoscenza che non si basano sullo studio empirico della natura. Immancabilmente, questo comporta l'acquisizione di strumenti per capire il ruolo che i dati giocano nell'ispirare, correggere, confermare o smentire le nostre intuizioni, e cosa questo comporti per il processo di estrazione e di conoscenza dai dati. A sua volta, una chiara epistemologia dei dati – Big, Open o meno – è un elemento fondamentale per lo sviluppo di strategie per ottimizzare i processi di ricerca datocentrici e renderli il più robusti possibile nei confronti dei cinque problemi fondamentali che abbiamo discusso nel capitolo precedente. Questo capitolo offre quindi una visione dell'epistemologia dei dati – un quadro filosofico su cosa essi siano, come forniscono informazioni e come sono usati per creare conoscenza – che aiuta a individuare e

comprendere meglio le fonti di precarietà insite nel trattamento contemporaneo dei Big Data. Il prossimo capitolo mostrerà poi come questa base filosofica serva quale punto di partenza per proporre modalità di intervento sulla produzione, gestione e analisi dei Big Data volte a mitigare i rischi posti da conservatismo, inaffidabilità, mistificazione, corruzione e danno sociale descritti nel capitolo precedente.

Visioni contrastanti del ruolo dei dati nei processi di ricerca

Finora ho affrontato la questione di cosa sono i Big Data, ma non ho discusso nel dettaglio cosa siano i dati in senso più generale, preferendo partire da una discussione dei problemi legati alla gestione e all'utilizzo dei Big Data come fonte di conoscenza. Ora voglio dimostrare come molti dei problemi osservati nel trattamento dei Big Data siano dovuti almeno in parte a una concettualizzazione ben radicata ma profondamente sbagliata dei dati e del modo in cui essi contribuiscono alla produzione di conoscenza – una posizione filosofica che è spesso presa in maniera inconsapevole e implicita, e che chiamo la "visione rappresentativa" dei dati.

Secondo questo approccio, i dati consistono in rappresentazioni affidabili della realtà che vengono prodotte attraverso l'interazione tra l'uomo e il mondo. Le interazioni tramite cui i dati sono generati possono avvenire in qualsiasi

situazione sociale, che questa avvenga o no per scopi di ricerca. Esempi variano dalla biologa che misura la circonferenza di una cellula in laboratorio, annotando poi il risultato in un *file* Excel, al maestro che conta il numero di alunni nella sua classe e lo scrive sul registro. Quello che conta come dato in queste interazioni sono gli oggetti creati nel processo di descrizione e misurazione del mondo: questi oggetti, che possono essere tanto digitali (il *file* Excel) quanto concreti (il registro di classe), costituiscono una traccia della realtà che fornisce il punto di partenza necessario per studiarla e trarne nuove forme di conoscenza. Questo è il motivo per cui i dati formano una legittima base per il nostro sapere empirico: la produzione dei dati è equivalente alla "cattura" di caratteristiche del mondo in modi che le rendono soggette allo studio sistematico. Secondo la visione rappresentativa i dati sono dunque oggetti dal contenuto fisso e immutabile, il cui significato come rappresentazioni della realtà va investigato e progressivamente rivelato tramite l'utilizzo corretto di metodi scientifici. I dati generati dallo studio della forma delle cellule sono esaminati per migliorare la conoscenza della fisiologia e della struttura dell'organismo. I dati creati nel contare gli alunni di una classe possono essere aggregati con altri dati simili raccolti in altre classi e altre scuole, generando così una base empirica per valutare la densità degli alunni relativa al territorio e la frequenza con cui vanno a scuola – e così produrre conoscenza delle caratteristiche e

dei limiti delle strutture scolastiche attualmente in uso.

Nella visione rappresentativa, i dati sono quindi il ponte tramite cui possiamo accedere al mondo in maniera sistematica, contestabile e riproducibile da altri. Questa funzione è spesso presentata in opposizione alla conoscenza che deriviamo dalla percezione dei nostri sensi che ci è fondamentale nella vita di tutti i giorni, ma che può rivelarsi illusoria e traditrice – come quando apro gli occhi in una stanza buia e non so più distinguere se sia la mia vista a non funzionare o se sia andata via l'elettricità. Come discusso nelle opere di innumerevoli filosofi sia nella tradizione analitica (Locke) sia in quella continentale (Kant), la conoscenza acquisita tramite la nostra percezione non è verificabile in maniera oggettiva, perché dipende dal nostro punto di vista e dalle nostre capacità cognitive e sensoriali. I dati sono concettualizzati come un'alternativa a questo potenziale solipsismo: ossia come oggetti pubblici che possono essere scambiati, discussi e criticati da chiunque e il cui significato non dipende dalla percezione di un unico individuo. In questa interpretazione, i dati formano una base oggettiva per l'acquisizione di conoscenza, ed è questa oggettività – la possibilità di derivare il sapere dall'esperienza umana uscendo però dai confini della percezione soggettiva – che rende la conoscenza propriamente *empirica*.

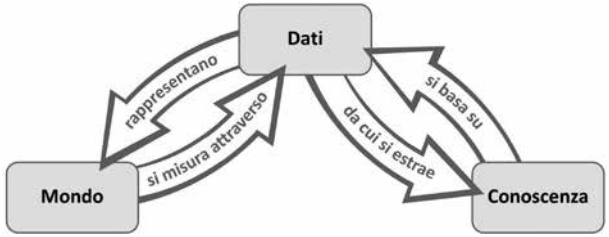

Figura 5. La produzione di conoscenza secondo la visione rappresentativa dei dati (copyright Sabina Leonelli, realizzazione Michel Durinx).

Di primo acchito, questa posizione può sembrare inoppugnabile. Pensare ai dati come a rappresentazioni oggettive (di parti) della realtà sembra essere indispensabile alla credibilità della scienza e all'esistenza stessa dell'empirismo come alternativa alla conoscenza puramente teorica e razionalista. Secondo la visione rappresentativa dei dati, i metodi e i risultati scientifici sono superiori ad altre forme di studio della realtà in quanto capaci di investigare il mondo in maniera fedele e accurata. Se gli esperti riescono a creare rappresentazioni credibili del reale nei loro studi, per esempio pianificando esperimenti intelligenti, la conoscenza che se ne deriva diventa automaticamente affidabile. Seguendo la stessa logica, la retorica tipicamente legata ai Big Data è quella di accumulazione del sapere in maniera induttiva: l'accumulo di dati ottenuti tramite metodi affidabili genera una montagna di fatti pronti all'analisi, e più fatti si producono e legano tra di loro, più conoscenza se ne può derivare.

Eppure, l'analisi delle condizioni in cui i Big Data vengono concretamente usati rivela parecchi problemi nella visione rappresentativa del ruolo dei dati nella ricerca[1]. Prima di tutto, è chiaro che non tutti (ricercatori o meno) hanno gli strumenti e le conoscenze necessari per trarre significato dai dati, specialmente quelli prodotti da laboratori e metodi specializzati come nel caso di molti dati scientifici. I dati non parlano mai da soli, e tipi di dati diversi necessitano di preparazione e strumenti di analisi differenti per poter essere interpretati. Il modo in cui i dati vengono interpretati dipende quindi almeno in parte dal tipo di abilità e dale conoscenze in possesso di chi li studia[2]. Di per sé questo non è necessariamente un problema per la visione rappresentativa dei dati, i cui sostenitori possono semplicemente rispondere che ci sono modi corretti e modi scorretti di interpretare la maniera in cui i dati rappresentano la realtà e che chi si prende la responsabilità di analizzarli deve possedere capacità adeguate a farlo in maniera corretta. Ma cosa conta come un'interpretazione "corretta" nel contesto di Big e Open Data, in cui i dati

[1] Anche Floridi ha una visione critica della concettualizzazione dei dati come oggetti dal contenuto semantico fisso e indipendente dal contesto. La sua alternativa alla visione rappresentativa, che è compatibile e parallela al mio approccio ma non focalizzata in maniera specifica sul mondo della ricerca, è incentrata sullo studio del concetto di informazione (Floridi 2017).

[2] de Regt 2017; de Regt *et al.* 2009.

vengono mobilitati e riutilizzati in mille maniere e per scopi diversi?

Il fatto che non ci sia solo una interpretazione possibile di un insieme di dati è estremamente difficile da giustificare all'interno della visione rappresentativa. Come già notato nel capitolo precedente, è proprio la possibilità di usare metodi, capacità e criteri interpretativi diversi che rende la mobilitazione e l'aggregazione di Big e Open Data così produttiva e potenzialmente rivoluzionaria. L'insistenza sull'utilità di ricontestualizzare i dati in ambiti diversi e darne quindi interpretazioni divergenti ma egualmente valide, dimostra che l'analisi dei dati non è un processo puramente oggettivo in cui il significato dei dati viene progressivamente rivelato. I dati possono essere usati per rappresentare molti aspetti diversi della realtà e la validità di ognuna di queste interpretazioni dipende dalle circostanze specifiche del processo di analisi, incluse l'abilità di manipolazione e le presupposizioni teoriche che permettono a persone e/o algoritmi di organizzare e di visualizzare i dati in modo da corroborare una certa concettualizzazione del reale. In altre parole, l'interpretazione dei dati è continuamente *mediata* dal punto di vista e dalle capacità di chi li usa. La concettualizzazione dei dati come oggetti dal significato fisso e indipendente dal contesto non si accorda con questa osservazione fondamentale, e anzi genera aspettative sbagliate sul modo in cui i dati forniscono informazioni sul mondo. A causa dell'approccio

rappresentativo, molti vedono i dati come fatti incontestabili e privi di aspetti teorici e soggettivi – un presupposto che non tiene conto della storia dei dati e della loro provenienza, né delle circostanze concettuali, materiali e sociali in cui possono essere interpretati.

Un secondo problema della visione rappresentativa dei dati consiste nello spiegare come mai la gestione dei dati ha un tale peso sul modo in cui vengono interpretati. Come abbiamo visto, la presentazione dei dati, il modo in cui vengono identificati, selezionati e inclusi (o esclusi) nelle banche dati, e le informazioni fornite agli utenti per aiutarne la ricontestualizzazione sono fondamentali alla produzione di conoscenza e ne condizionano fortemente i contenuti. Di più: l'analisi di come i dati viaggiano da un contesto all'altro rivela che le aspettative e le capacità di chi manipola e mobilita i dati determina non solo il modo in cui vengono trattati ma anche quel che viene visto come "dato" stesso e il formato in cui è disponibile – il che a sua volta ne influenza la trattabilità e la potenziale interpretazione.

Prendiamo per esempio il caso di osservazioni botaniche fatte per divertimento e in seguito usate come dati. Fare fotografie e annotare osservazioni sulle piante incontrate lungo il cammino è un passatempo preferito per tante persone, che spesso scelgono di custodire gli oggetti così ottenuti per puro piacere personale – per esempio in quaderni e annotazioni tenute a casa e mostrate ad amici e parenti in occasioni spe-

ciali. Difficilmente chi ha questo hobby segue metodi standardizzati per generare le proprie foto e le note scritte: come ampiamente documentato nella storia delle scienze delle piante, e nonostante il successo di classificazioni tassonomiche come quella di Linneo, ognuno tende a costruire il proprio sistema di osservazione a seconda delle personali preferenze estetiche, concettuali e affettive e dei propri limiti fisici e pratici (una donna di ottant'anni non scala alberi in cerca di una particolare varietà di rampicante; un genitore con tre figli piccoli non ha due ore al giorno da dedicare a riordinare e annotare le foto fatte nel parco). Spesso capita che questi amanti del mondo naturale decidano di mettere foto e osservazioni, magari quelle più belle dal punto di vista tecnico o riguardanti piante o località a loro care, su Instagram o su qualche altro sito web. Capita anche che scienziati interessati a documentare la morfologia delle piante di una determinata località scoprano l'esistenza di queste foto e decidano di usarne alcune per scopi di ricerca[3]. Le foto giudicate rilevanti per lo studio morfologico degli scienziati in questione (un cri-

[3] Considerare oggetti prodotti nel tempo libero da cittadini qualunque come dati scientifici è una componente fondamentale della *citizen science* – la scienza dei cittadini – che a sua volta fa parte del movimento verso la Scienza Aperta. La creazione e la condivisione di dati sono uno dei metodi più in voga per coinvolgere non-professionisti nella creazione di conoscenza scientifica, com'è anche dimostrato dalla storia dell'astronomia e delle scienze naturali – in cui l'acquisizione di dati è da lungo tempo organizzata alla stregua di una collaborazione tra dilettanti e ricercatori.

terio ben diverso da quelli usati dai creatori delle foto per scegliere quali foto pubblicare) sono quindi estratte dai siti web usati dai dilettanti, formattate in modo da essere interoperabili con foto e osservazioni prese da altri in altri luoghi e inserite in una banca dati botanica con lo scopo di permetterne l'analisi. Viene così generato un insieme di oggetti diverso nel formato e nei contenuti da quello originariamente creato dai dilettanti, ma anche più facile da comparare e aggregare con foto di altre piante già presenti nel sistema. Una volta che le foto sono state inserite in questo contesto, diventando in questo modo dati potenzialmente utilizzabili per la produzione di conoscenza, è perfettamente possibile anche per scienziati con interessi e specialità diversi da quell'intento morfologico selezionare alcune delle foto e modificarle ulteriormente per i loro scopi. Per esempio, alcune delle immagini potrebbero essere ingrandite e studiate da patologi interessati a investigare la velocità di diffusione di un'infezione che ha l'effetto di annerire le foglie delle piante.

Assistiamo così a una situazione in cui una particolare combinazione di interessi, capacità, accessibilità e caratteristiche degli oggetti in questione determinano cosa viene identificato come dato – ossia, quali delle foto vengono considerate come base empirica per l'analisi e la risoluzione di un particolare quesito. In questo come in mille altri casi in cui i dati viaggiano attraverso contesti diversi, quello che viene valorizzato come dato – e quello che viene scartato come

elemento irrilevante all'analisi –[4] non è sempre lo stesso oggetto o lo stesso insieme di oggetti, ma continua a cambiare (nel formato, nella parte dell'oggetto che diventa rilevante, o nel modo in cui l'oggetto si lega ad altri)[5]. E come cambia l'oggetto così cambia il modo in cui le sue caratteristiche possono essere concettualizzare e manipolate, generando interpretazioni diverse su quello che questi oggetti – questi dati – possono rappresentare.

La visione rappresentativa non può che leggere questa manipolazione come una distorsione del significato originale dei dati, e infatti i suoi esponenti spesso insistono su una netta distinzione tra i dati "grezzi" ("*raw data*") e quelli che sono stati processati o comunque ulteriormente elaborati per renderne possibile l'analisi. L'elaborazione dei dati grezzi è riconosciuta come una tappa necessaria per poterli utilizzare come fonte di conoscenza, ma è tipicamente vista con sospetto da chi interpreta il dato come una rappresentazione della realtà: secondo questa prospettiva, la difficoltà dell'interpretare i dati consiste nell'assicurarsi che i metodi usati per formattarli non tradiscano il loro significato originario. Nella visione relazionale, invece, la distinzione tra dati "grezzi" e dati "elaborati" diventa irrilevante dal punto di vista epistemi-

[4] McAllister 2007; Loettgers 2009; Woodward 2010; Boumans & Leonelli 2019.

[5] Si vedano anche le analisi di Niccolò Tempini, Mary Morgan e James Griesemer contenute del volume di prossima uscita *Varieties of Data Journeys* (Leonelli & Tempini 2019).

co e comunque difficile da mantenere nel caso di Big e Open Data, in cui: (1) la differenza tra dati "originali" e dati "processati" non è né evidente, né nitida, né particolarmente significativa per l'analisi scientifica in cui i dati grezzi possono essere tanto problematici quanto quelli elaborati, a seconda dei metodi e degli strumenti usati per generarli[6]; (2) usare i dati nel formato e nell'ordine in cui sono originariamente creati è spesso impossibile, come attestano i tanti metodi usati dai ricercatori per "pulire" i dati e renderli processabili (incluse tecniche statistiche di riduzione e normalizzazione)[7]; e (3) gli stessi oggetti che vengono selezionati e interpretati come dati cambiano a seconda del contesto di ricerca. Secondo gli esponenti della visione rappresentativa dei dati, queste tre situazioni sono letteralmente incomprensibili: se i dati costituiscono rappresentazioni fedeli della realtà, come possono essere continuamente modificati e nonostante questo continuare a essere utili come fonte di conoscenza? Questa riluttanza nel riconoscere l'importanza epistemica dei processi di elaborazione dei dati si traduce in riluttanza nel devolvere attenzione a tali processi e documentarli così da renderli visibili e contestabili. A sua volta, questa tendenza a occultare o dimenticarsi delle trasformazioni subite dai dati nel corso nei loro viaggi genera la concezione della banca dati come "scatola nera" che abbiamo già visto ave-

[6] Si veda anche Gitelman (2013).
[7] Mayo (1996).

re conseguenze altamente problematiche per la cura e l'utilizzo dei Big Data.

Un terzo problema della visione rappresentativa dei dati è legato alle difficoltà, discusse nel capitolo precedente, di trovare criteri universali e assoluti per valutarne la qualità e l'attendibilità[8]. Tali criteri non esistono proprio perché la qualità dei dati deve essere valutata in relazione agli scopi precisi della ricerca in cui vengono usati e al tipo di dati in questione (il cui formato varia enormemente a seconda dei metodi usati per generarli). Il pluralismo negli scopi e nei metodi in uso nel mondo della ricerca riflette la varietà e la diversità della realtà che ci circonda e che cerchiamo di comprendere e controllare, ed è a sua volta rispecchiato dalla varietà di tipi di dati prodotti dalle nostre interazioni con il mondo. La scelta di cosa conti come dato attendibile è quindi inevitabilmente legata alle circostanze specifiche del loro uso. Tornando per un momento al nostro esempio botanico, le fotografie di una pianta possono essere interpretate come dati utili allo studio di tanti aspetti diversi, tra i quali lo sviluppo morfologico di quella specie, i sintomi di un'infezione, l'effetto di determinate condizioni metereologiche sulla colorazione delle foglie e la presenza di parassiti in una determinata località. Ognuna di queste interpretazioni è in parte condizionata dalle caratteristiche fisiche delle fotografie (ci sono insetti sulle foglie o no? Di che colore sono le foglie?) e in parte dal

[8] Leonelli (2012), Canali (2016).

modo in cui chi usa questi oggetti ne accentua la trattabilità come dati (magnificando le parti delle foto in cui compaiono insetti, estraendo le misure delle dimensioni delle foglie tramite analisi computazionale e così via). Non è quindi ovvio che i dati contengano, in maniera oggettiva e indipendente dal contesto di uso, determinate informazioni sul mondo. Le caratteristiche degli oggetti che vengono considerate come dati certamente delimitano il tipo di uso e di interpretazione che se ne può trarre, ma allo stesso tempo è possibile ottenere informazioni diverse dagli stessi oggetti a seconda di come questi vengono gestiti e interpretati – il che di nuovo smentisce l'idea che i dati stessi siano rappresentazioni fedeli, oggettive e immutabili della realtà.

Se l'affidabilità dei dati non proviene dalla loro capacità di rappresentare il mondo in maniera oggettiva e immutabile, allora da dove viene? L'alternativa qui proposta è di concettualizzare i dati in maniera *relazionale* piuttosto che rappresentativa. Invece che come una rappresentazione del reale, il dato va concepito come un oggetto messo in relazione a un quesito irrisolto in modi e per motivi che dipendono dalla situazione in cui il quesito viene posto. Nella visione relazionale, *qualsiasi oggetto può svolgere il ruolo di 'dato' a patto che (1) venga trattato come potenziale fonte di conoscenza empirica e (2) sia possibile mobilitarlo in modo da renderlo accessibile a più persone*. In altre parole, le intenzioni e le aspettative dei ricercatori hanno un forte impatto su quali oggetti vengono selezionati come

potenziali fonti di conoscenza, ma questi oggetti devono essere concretamente visibili e ispezionabili da più di una persona, in modo da fungere come prova delle asserzioni a cui vengono legati. I dati non sono tali se esistono solo nella testa di chi li usa: almeno in teoria, essi devono essere accessibili ad altri che possano valutarne il valore scientifico e verificarne l'affidabilità come basi empiriche di conoscenza. Secondo l'approccio relazionale, il significato assegnato ai dati non dipende quindi solo dalle loro caratteristiche fisiche e da quello che rappresentano, ma anche dalle motivazioni e dagli strumenti usati per analizzarli e per difendere particolari interpretazioni; e l'affidabilità dei dati dipende soprattutto dalla credibilità e dal rigore dei processi usati per produrli e analizzarli.

Questa visione riconosce che qualsiasi oggetto può funzionare come dato, oppure smettere di farlo, a seconda delle circostanze – un'osservazione ben nota a chi si occupa di dati storici, spesso custoditi in archivi dimenticati da tutti e quindi ridotti a oggetti senza significato, o di dati ricavati da attività che non hanno nulla a che fare con la ricerca ma che possono comunque essere usati per generare nuova conoscenza, come per esempio la quantità di prodotti organici venduti in Italia o il numero di compagnie fallite dopo la crisi economica del 2008. Inoltre, l'approccio relazionale riconosce che gli oggetti che vengono considerati come dati sono spesso modificati nel corso dei loro viaggi attraverso siti di produzione, disseminazione e riuso. Non solo i dati pos-

sono cambiare formato, ma questi cambiamenti possono avere un forte impatto su come, dove e da chi vengono usati come fonte di conoscenza. Di conseguenza, diventa importantissimo documentare i processi di gestione e trasformazione dei dati, particolarmente nel caso dei Big Data in cui questi viaggiano in lungo e in largo per canali digitali, e vengono raggruppati, analizzati e interpretati in forme e modi diversi. La visione relazionale incoraggia così la cura dei dati e l'attenzione alla loro storia, sottolineando la loro qualità di oggetti in continua evoluzione e sottoposti a modificazioni talvolta radicali, e le implicazioni di questa qualità per il loro potere di confermare o smentire ipotesi.

Un'ovvia obiezione a questo modo di pensare ai dati come soggetti a continui cambiamenti consiste nell'osservare che queste trasformazioni rendono molto difficile seguire i dati mentre viaggiano da un posto all'altro. Come possiamo dire che i dati usati in un determinato contesto siano gli stessi che sono stati prodotti in un contesto diverso, se il loro utilizzo e la loro stessa forma continuano a cambiare e ad adattarsi a nuove situazioni? E come fanno i dati a mantenere una loro integrità se, ogni volta che vengono spostati e rivalutati, le loro caratteristiche fisiche possono cambiare? La risposta a questa obiezione è di stampo pragmatico[9]. È perfettamente possibile pensare ai dati come a entità storiche che evol-

[9] La mia teorizzazione è ispirata dal filosofo pragmatista americano John Dewey, come ho spiegato in Leonelli (2016a).

vono e cambiano attraverso la riproduzione e l'accumulo di esperienze diverse, ma la cui provenienza può e deve sempre essere ricostruita, almeno in teoria, per poterne valutare la validità. Il modo migliore di teorizzare i movimenti e le trasformazioni dei dati è quello che usiamo per pensare alla riproduzione degli organismi viventi e al modo in cui le caratteristiche di ogni generazione vengono passate, anche se in modo imperfetto e imprevedibile, a quella che la segue. I dati possono quindi essere concettualizzati come *stirpi*: dinastie di oggetti che si trasformano nel passaggio da una forma all'altra ma il cui studio dipende almeno in parte dalla nostra capacità di valutarne l'origine e la provenienza.

L'idea del dato come stirpe porta a un'altra obiezione importante per la visione relativistica del dato, ossia quella concernente il relativismo insito nel considerare i dati sempre e solo in rapporto a contesti specifici invece che come fonte di conoscenza oggettiva. Non si rischia di creare un concetto di dato completamente relativistico, in cui qualsiasi cosa venga usata come dato lo diventa automaticamente, e qualsiasi oggetto può essere legittimamente interpretato come base per affermazioni di qualsiasi tipo? Questa preoccupazione è particolarmente acuta vista la plateale mancanza di rispetto per fatti assolutamente evidenti a tutti, come per esempio che il nostro pianeta sia rotondo e che non sia sotto il controllo di rettili provenienti da un altro pianeta, da parte di chi insiste nel creare verità soggettive e completamente distaccate dal mondo reale

sulla base di interesse finanziario o convinzione personale[10].

La visione rappresentativa dei dati ha un modo apparentemente semplice e convincente per rispondere a questa obiezione, il che spiega perché tanti si ostinino a pensare ai dati in questo modo: se i dati sono una rappresentazione oggettiva di un determinato aspetto del mondo, allora basta comparare i dati al mondo per capire se sono corretti e se ha senso credere o no alla loro affidabilità. Purtroppo però, e sicuramente per la maggior parte dei dati usati in ambito scientifico, una tale valutazione non è così facile. In che modo i dati genetici, come per esempio la sequenza GTTACCTGAAA, rappresentano in maniera chiara e ineccepibile le informazioni ereditarie all'interno di una cellula? Cosa vuol dire per i numeri 34, 72 e 91 rappresentare le dimensioni di un organismo? In che senso un'ecografia rappresenta un feto nell'utero materno e fornisce informazioni sul suo stato di salute? Molto spesso gli oggetti che i ricercatori usano come dati non assomigliano in maniera ovvia alle parti del mondo che servono a documentare, e la loro interpretazione richiede abilità e conoscenze precise (come quelle usate per interpretare un'ecografia come rappresentazione di una cardiopatia nel feto). Altrettanto spesso i criteri e i metodi usati per attribuire un significato a questi oggetti cambiano con il tempo e il con-

[10] L'esempio dei terrapiattisti e dei seguaci di David Icke è solo uno dei tanti scenari estremi del momento.

testo. Questo non vuol dire che questi criteri e metodi siano arbitrari, ma piuttosto che la loro sofisticazione e la precisione con cui vengono applicati in situazioni diverse continua a crescere passo a passo con l'avanzare delle tecnologie e delle conoscenze scientifiche. Esistono criteri e metodologie sempre più sofisticate per stabilire e verificare se un oggetto può essere usato in maniera affidabile come fonte di conoscenza su un determinato fenomeno, e a che condizioni. E infatti non esiste metodo o criterio scientifico che consenta di interpretare le fotografie satellitari del nostro pianeta come prova che la terra è piatta, mentre esistono mille modi di identificare errori nel ragionamento usato da chi promuove queste idee, dimostrando così che la conoscenza prodotta in questo modo non sta in piedi. Accettare una concezione relazionale dei dati non è quindi equivalente ad accettare un totale relativismo. Anzi: questa visione incoraggia alla continua vigilanza sui motivi, sui metodi e sugli scopi per cui certi oggetti vengono identificati, manipolati e proposti come dati in supporto di determinate asserzioni, sia al di fuori che all'interno del mondo scientifico.

Dai dati alla conoscenza: una questione di ordine

Accettare la visione relazionale dei dati ha ripercussioni su come si concettualizza l'intero processo di ricerca empirica, riassunte nel grafico in figura 6.

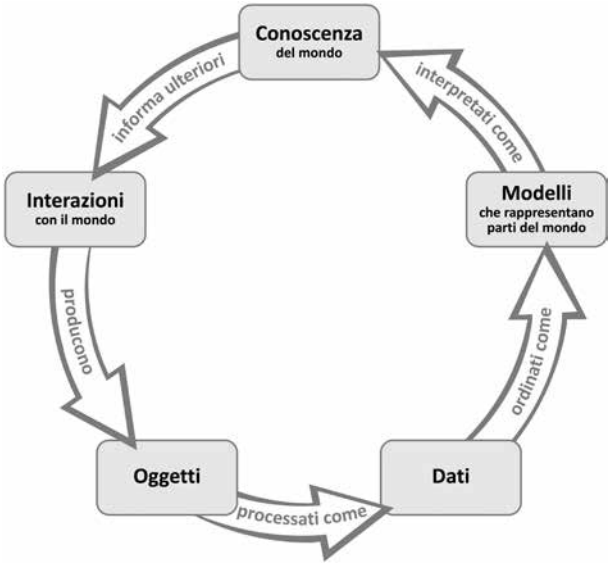

Figura 6. Il processo di ricerca associato alla visione relazionale del dato (copyright Sabina Leonelli, realizzazione Michel Durinx).

La ricerca empirica inizia, come nel caso della visione rappresentativa, dall'interazione tra l'uomo e il mondo. Queste interazioni producono artefatti di vario genere, come per esempio numeri, misure, simboli, fotografie, descrizioni e grafici. Alcuni di questi oggetti sono poi selezionati e processati con lo scopo di fungere, almeno potenzialmente, come fonte di conoscenza. Gli oggetti così manipolati sono quelli che chiamiamo dati. Come ho già spiegato, quello che questi dati sono usati per rappresentare non è deter-

minato solo dalle caratteristiche fisiche dei dati stessi, ma anche dai presupposti e dal contesto di coloro che ne valutano il potenziale significato. Interpretare i dati come fonte di conoscenza include quindi due ulteriori passaggi: (1) la creazione di modi di ordinare i dati – spesso chiamati *modelli* nel mondo scientifico – che ne manifestano una certa funzione rappresentativa[11] e (2) l'uso di questi modelli come base empirica per la produzione di conoscenza. In questa visione della ricerca, quindi, la funzione rappresentativa dei dati continua a essere presente, ma non sono i dati come oggetti in sé che rappresentano – in modo più o meno affidabile e accurato – una parte della realtà. Quello che svolge la funzione rappresentativa è il modello dei dati che chiunque li interpreti compone quando sceglie come organizzarli. In altre parole, è un determinato *ordine* di dati – la maniera in cui vengono visualizzati e resi così rilevanti per un certo tipo di analisi – che rappresenta un aspetto del mondo e lo rende accessibile allo studio scientifico[12]. Mettere ordine ai dati è il modo in cui questi diventano trattabili come prove di determinati fatti e come fonti di conoscenza nuova. I dati non

[11] Leonelli (in corso di stampa), Bokulich (2018), Suppes (1962).

[12] Come ho già segnalato nel capitolo 1, questa visione dei dati ha radici nel lavoro di molti filosofi, e forse in maniera più evidente di Michel Foucault e Jacques Derrida. Non è il mio obbiettivo in questo libro tracciarne le origini quanto specificarne le caratteristiche e l'importanza nel contesto contemporaneo.

sono quindi di per sé stessi una base oggettiva per il sapere: è il modo in cui li organizziamo e visualizziamo che determina il significato loro assegnato.

Si consideri di nuovo l'esempio dei dati botanici per illustrare questo processo in maniera più concreta. In quel caso, l'amatore che fa fotografie camminando nel bosco produce oggetti tramite la sua interazione col mondo – le foto – che vengono poi processati da ricercatori con l'aspettativa che possano fungere da dati (per esempio, quando le foto vengono formattate per inserimento nell'archivio digitale). I ricercatori organizzano e ordinano i dati così ottenuti in modi che li aiutano a rappresentare fenomeni diversi: nel caso dei morfologi, la forma di una particolare specie in una certa località; nel caso dei patologi, i potenziali sintomi di un'infezione delle foglie. Questi modelli sono poi testati per verificarne l'affidabilità e la rilevanza ai fini dei fenomeni che documentano – per esempio, i ricercatori controllano che il modello dei sintomi dell'infezione ricavato dall'analisi delle immagini trovate *online* rispecchi le caratteristiche di modelli di dati provenienti da altre fonti, e quando possibile tornano nella località in questione per verificare la veridicità del modello[13]. Se i modelli sono giudicati come adeguati, vengono usati come fonte di conoscenza sul modo in cui l'infezione si manifesta nelle piante in questione.

[13] Per esempio in Shavit & Griesemer (2009) e Leonelli & Tempini (2018).

Se non vengono giudicati come adeguati, i ricercatori tornano ad analizzare i dati e provano a ordinarli in modi diversi – il che a volte comporta cambiare radicalmente il tipo di oggetto che viene considerato come dato e/oppure l'aspetto della realtà che si sta studiando[14].

Abbiamo già notato come una delle caratteristiche principali dei Big Data, che li rende particolarmente interessanti come fonti di conoscenza, sia la possibilità di esaminare e comparare tanti tipi di dati diversi, ottenuti in modi spesso molti differenti ma di potenziale rilevanza per l'analisi dello stesso fenomeno. L'analisi comparativa di cosa si può imparare dalla giustapposizione di tipi diversi di dati viene spesso chiamata *triangolazione*, ed è fortemente associata all'idea che quanti più tipi di dati confermano una certa interpretazione, tanto più quella interpretazione si può giudicare affidabile ed empiricamente corretta. È però essenziale specificare che la triangolazione funziona solo se i dati che vengono giustapposti hanno origini diverse tra di loro (provenendo dunque da stirpi differenti). Se i dati sono creati dallo stesso gruppo di ricercatori, con gli stessi strumenti e sulla base degli stessi presupposti teorici è infatti probabile che vengano interpretati in maniera simile, ma

[14] Come discuto nel dettaglio in Leonelli (in corso di stampa), l'identificazione dei fenomeni studiati dai ricercatori è fortemente condizionata dalla scelta e dalla gestione dei dati – un punto epistemologicamente significativo che è stato discusso anche da Bogen & Woodward (1988), McAllister (2007), Teller (2010), Massimi (2011) e Feest (2011).

non è chiaro se l'interpretazione dipenda dalle similarità tra i dati stessi o dalla loro affidabilità come fonte di conoscenza. Se dati che non condividono la stessa storia e non sono parte della stessa dinastia puntano nella medsima direzione, abbiamo invece una conferma più credibile del fatto che l'interpretazione sia corretta. La filosofa Alison Wylie ha discusso nel dettaglio i modi in cui dati di tipi diversi sono usati per triangolare interpretazioni e rinforzare così la base empirica del ragionamento scientifico, concludendo che conoscere l'origine e la storia di come i dati sono gestiti permette di verificare in maniera adeguata se e come certi modi di ordinare (modellare) i dati siano attendibili quali fonti di conoscenza[15].

Nella visione rappresentativa dei dati, è difficile riconoscere il ruolo fondamentale giocato dalla storia dei dati in quanto oggetti nella loro interpretazione come fonti di conoscenza. Per chi pensa ai dati come rappresentazioni oggettive e immutabili del mondo, le condizioni in cui essi vengono formattati e ordinati importano poco: quel che importa è che si riesca a svelare il loro reale significato. Questo approccio si accompagna facilmente all'idea che mettere insieme tanti dati costituisca di per sé un incremento della base empirica della conoscenza. L'accumulo dei dati equivale all'accumulo di tanti fatti, un vero e proprio tesoro da cui estrarre nuove scoperte tramite tecniche induttive e statistiche. È semplice capire come chi adotta questa visione dei dati

[15] Wylie (2002, 2017); Chapman & Wylie (2015).

sia facile preda delle false promesse legate all'uso dei Big Data, come per esempio l'idea che questi siano universalmente attendibili, imparziali e usabili per qualsiasi tipo di analisi[16].

La visione relazionale dei dati si affianca invece a una visione meno utopica delle condizioni in cui i Big Data possono essere usati come fonti attendibili ed efficaci di conoscenza. Nella visione relazionale, la derivazione di conoscenza comporta il *posizionamento* di oggetti scelti per svolgere la funzione di dati (e quindi delle loro caratteristiche fisiche) *in relazione* ad altri elementi cruciali per l'interpretazione, come per esempio lo scopo della ricerca, i presupposti concettuali su cui è basata e il tipo di conoscenza – teorica o pratica – che si vuole ricavare[17]. Questo posizionamento comporta quindi presupposti e scelte ben più ampie di quelle coinvolte nell'applicazione di metodi statistici. Le procedure con cui i dati sono processati e ordinati sono fondamentali al loro uso come fonte di conoscenza, e alla scelta e all'uso di criteri per giudicarne le potenzialità rappresentative nei confronti della

[16] Non voglio dire che la visione rappresentativa dei dati sia necessariamente incompatibile con una visione dei Big Data molto più sofisticata, ma piuttosto che la visione relazionale si accorda molto meglio a una gestione e a un uso dei Big Data che siano sempre attenti a potenziali problemi metodologici – il che la rende preferibile ai miei occhi.

[17] La mia concezione di conoscenza si accorda con quella proposta da Chang (2017) su ispirazione egualmente pragmatista: conoscenza come abilità di agire, di cui la conoscenza espressa in forma linguistica è solo una delle possibili componenti.

realtà. La visione relazionale dei dati riconosce quindi l'enorme lavoro necessario per documentare i viaggi dei dati e renderne possibile lo scrutinio nel corso di processi di interpretazione.

Capitolo quarto

Come incoraggiare il meglio: verso una scienza partecipativa e responsabile

Torniamo ora alle questioni fondamentali con cui abbiamo iniziato la nostra discussione dei Big Data. Come possono essere usati per difendersi dalla *post-verità* e dalle tante forze che cercano di manipolare i fatti per guadagno personale? Come facciamo ad accertarci che la conoscenza scientifica a cui ci affidiamo ogni giorno abbia fondamenta empiriche solide e attendibili? La risposta a queste domande non consiste nell'agitare una bacchetta magica e invocare un particolare metodo o figura professionale come la perfetta e unica soluzione alla crisi epistemica dei nostri tempi. Non esiste magia che possa risolvere in un colpo solo le tensioni e l'incertezza insite nella molteplicità di voci, settori e interessi coinvolti nei viaggi dei dati. Come illustrato nel capitolo precedente, partire da una visione relazionale dei dati vuol dire accettare che non esistono riferimenti fissi né tecniche infallibili per giudicarne la qualità e il valore scientifico ed epistemico, ma quello che conta è il modo in cui

vengono ordinati e visualizzati in relazione alla situazione e allo scopo dell'analisi.

Questa osservazione non deve scoraggiare chi si affida alle scienze – inclusa la medicina e la tecnologia – come punto di riferimento fondamentale per comprendere sé stessi e il mondo. La visione relazionale dei dati ha infatti una conseguenza importantissima per la gestione della ricerca scientifica e dell'intelligenza artificiale: quella di sottolineare lo stretto legame tra le decisioni prese nell'organizzare i dati e il modo in cui essi vengono interpretati. Tecniche come l'apprendimento automatico rendono sempre più facile l'automatizzazione di alcune di queste decisioni, che possono essere determinate e implementate da algoritmi capaci di evolvere e migliorare a seconda delle esperienze fatte. Ma il giudizio umano continua a permeare il modo in cui gli algoritmi privilegiano alcune fonti e alcuni tipi di dati rispetto ad altri; la scelta di tecniche di visualizzazione e analisi statistica; gli scopi, le presupposizioni e le preferenze incorporate nel processo di analisi.

Queste decisioni sono tanto scientifiche quanto etiche, e dimostrano l'importanza dell'etica e della partecipazione sociale nello sviluppo dei sistemi che permettono la mobilitazione e il riuso dei dati. Scelte che possono sembrare puramente tecniche – quale forma di calcolo probabilistico usare, a che tipo di classificazione affidarsi – sono in realtà cariche di implicazioni per come la conoscenza che se ne deriva può trasformare la società. E per quanto gli scienziati coinvolti

nell'analisi dei Big Data siano primariamente responsabili per le decisioni che prendono, le loro scelte non possono essere assunte al di fuori del contesto sociale in cui la conoscenza viene prodotta e utilizzata. All'interno di una società democratica, questo implica un impegno continuo verso la ricerca del dialogo e del confronto tra ricercatori e altri gruppi sociali la cui esperienza di determinate situazioni li mette in grado di contribuire in maniera decisiva alla valutazione dei presupposti e delle scelte fatte nella produzione, selezione, disseminazione e interpretazione dei dati. Genitori, imprenditori, pazienti, insegnanti, ristoratori hanno una conoscenza unica e preziosa di cosa comporti curare i figli, gestire un'azienda, convivere con la malattia, educare nuove generazioni e soddisfare le preferenze del pubblico tutelandone allo stesso tempo la salute. Nella visione relazionale dei dati, questo tipo di conoscenza deve essere messo a confronto con i metodi scientifici usati per analizzare i Big Data, ed essere incorporato nei viaggi dei dati per quanto possibile. La capacità dei dati di stravolgere separazioni rigide tra tipi diversi di *expertise* – che abbiamo visto all'opera ogni volta che i Big Data cominciano a viaggiare – può e deve essere sfruttata per stravolgere e migliorare la comunicazione tra settori sociali che sono a loro volta diventati troppo specializzati e incapaci di confrontarsi tra di loro in maniera costruttiva e necessaria per lo sviluppo di una società funzionale. I dati inevitabilmente viaggiano attraverso tanti mondi sociali diversi, ed è attraverso questi

viaggi che acquisiscono valore epistemico come fonte di conoscenza: riconoscere questa realtà è un passo importante verso una concezione di produzione scientifica che non esclude il mondo esterno alla ricerca, ma invece lo abbraccia e lo incorpora nella formulazione di giudizi su cosa costituisce conoscenza – sia affidabile, sia equa – dal punto di vista scientifico ed etico.

Alcuni degli addetti ai lavori a questo punto ribadiscono che, proprio a causa della natura fortemente distribuita e tecnica del lavoro informatico, non è mai possibile per gli esperti coinvolti prevedere quali possano essere le conseguenze etiche e sociali delle loro scelte, o tantomeno coinvolgere settori sociali diversi in questa valutazione – e quindi il legame tra etica e scienza, pur convincente dal punto di vista teorico, si spezza appena confrontato con le limitazioni pratiche dell'uso dei dati nella vita di tutti i giorni. Nel mondo concreto dell'analisi dei Big Data, mi dicono, non c'è modo di valutare le implicazioni delle proprie azioni, perché queste implicazioni emergono solo una volta che un determinato programma, strumento o tipo di analisi è messo in uso. Ma a quel punto, continuano i critici, non c'è più modo di rivisitare i presupposti su cui le tecniche di analisi sono state costruite, un po' perché sono tipicamente incorporate in un apparato tecnologico e computazionale estremamente complesso, e un po' perché quasi nessuno di coloro che applicano questo apparato a situazioni di ricerca ha una buona comprensione della "scatola nera" creata da questa tecnologia.

In un esempio concreto: da una parte, chi costruisce algoritmi di analisi di testi per Google non può preoccuparsi di tutti i modi possibili in cui questi algoritmi possono essere abusati, perché il potenziale abuso tipicamente emerge solo quando gli algoritmi sono incorporati nell'enorme apparato di Google e resi accessibili al pubblico; e dall'altra parte, una volta che gli algoritmi sono assimilati e pubblicizzati in questo modo, è molto difficile modificarli e individuare esattamente quali delle loro caratteristiche risultano problematiche dal punto di vista sociale. La conclusione di questa linea di ragionamento è che nessuno degli addetti ai lavori – né chi produce i dati e le relative infrastrutture, né chi ne fa uso per produrre conoscenza – sembra potere assumersi responsabilità per le conseguenze delle scelte fatte nella gestione dei dati. Ci ritroviamo così in preda al determinismo tecnologico: nonostante la consapevolezza che alcune delle implicazioni dell'analisi dei Big Data sono fortemente negative, non sembra esserci modo per frenare l'abuso. In quest'ottica tutte le tecnologie coinvolte nell'analisi dei Big Data – che siano ufficialmente riconosciute come intelligenza artificiale o no – possono essere interpretate come superiori all'uomo, per il semplice fatto che ne soverchiano oramai la capacità di giudizio e deliberazione.

Io credo che sia non solo possibile ma assolutamente necessario evitare questa rassegnazione al determinismo tecnologico, che è deleteria sia per la qualità e la credibilità della conoscenza scientifica sia per il suo impatto sociale – e che

la visione relazionale racchiuda intuizioni importanti su come farlo. Le prossime due sezioni esaminano brevemente due strategie che chi lavora con Big Data può utilizzare per tenere conto delle possibili implicazioni etiche e sociali del proprio operato, riasserendo una forma limitata ma cruciale di controllo sulla conoscenza prodotta e così mitigando – se non evitando completamente – i rischi epistemici discussi finora.

L'integrazione dell'etica nella ricerca scientifica

La prima strategia è l'adozione di procedure per l'integrazione dell'etica nelle scelte tecniche di gestione e analisi dei dati. Questo richiede prima di tutto di abbandonare l'idea che per valutare il valore etico di un'innovazione sia necessario disporre di un pronostico preciso del suo potenziale impatto sociale. Un conto è spendere tempo e risorse nell'immaginare e verificare come una certa innovazione potrebbe essere incorporata in vari contesti – uno sforzo fondamentale per la produzione di conoscenza e tecnologie reattive alle esigenze sociali e ai presupposti culturali di chi le usa. Un altro conto è cercare di elaborare modi di prevedere, quantificare e controllare completamente le esatte implicazioni; un requisito impossibile per qualsiasi novità. Questo vale perfino per innovazioni in campo medico, dove l'approvazione di nuovi tipi di trattamento è soggetta a anni di test e verifiche severe, ma non ci sono modi garantiti per evitare effetti collate-

rali inaspettati, specialmente nel lungo temine. È quindi assurdo pensare che l'analisi di potenziali implicazioni etiche e sociali abbia senso solo quando se ne ha un quadro preciso.

In secondo lungo, si deve abbandonare l'idea che la ricerca debba focalizzarsi solo su innovazioni con effetti socialmente positivi. Questo non è un presupposto realistico per nessun tipo di innovazione: tutte possono infatti essere usate per scopi in qualche modo dannosi ad alcune parti della società, e tutte sono legate a situazioni di rischio e incertezza. Si pensi solo all'impatto che tecnologie come l'intelligenza artificiale stanno avendo sul mercato del lavoro, dove milioni di persone – dai tassisti agli insegnanti, dagli operai agli avvocati – rischiano di essere rimpiazzati da una serie di algoritmi, come è già successo a tanti prima di loro grazie all'avvento del *personal computer*. In questo caso non c'è modo di evitare l'impatto sociale senza arrestare completamente lo sviluppo tecnologico, un'opzione estrema che priva l'umanità dei vantaggi e delle opportunità offerte da queste stesse tecnologie. L'acquisizione di conoscenza porta sempre vantaggi e svantaggi, e il problema etico consiste nel valutare come questi si relazionino gli uni agli altri nonché il peso che possono avere su realtà sociali con caratteristiche differenti. Nel caso dell'impatto dell'intelligenza artificiale sul mercato del lavoro, la valutazione etica concerne quale tipo di caratteristiche e funzionalità privilegiare nello sviluppo di nuovi algoritmi, quali settori possono usufruire meglio di questi cambiamenti e che

tipo di riorganizzazione culturale, pedagogica e sociale sia meglio incentivare per favorire un impatto positivo sulla popolazione (per esempio, riconoscendo che la gestione dei Big Data ha un enorme potenziale di creare nuove forme di impiego, che però richiedono una formazione diversa da quella che la stragrande maggioranza della popolazione sopra i trent'anni ha ricevuto nel sistema educativo italiano).

In altre parole, la valutazione etica consiste nel tenere conto a ogni passo, per quanto possibile, delle circostanze in cui un certo risultato potrebbe essere utilizzato nonché della situazione e delle aspettative degli utenti – e usare questa conoscenza come base per decisioni tecniche riguardanti le fonti, i formati, la classificazione e l'analisi dei Big Data. Questo comporta l'esplorazione di chi potrebbe essere interessato alla conoscenza che sta venendo prodotta; per quale motivo; chi viene incluso o escluso da questo potenziale uso; come si potrebbe modificare il processo di ricerca in modo da rendere i risultati meno discriminatori, più sostenibili e più o meno inclusivi a seconda delle esigenze.

È importante notare come porsi queste domande non garantisca l'esistenza di risposte adeguate, e spesso generi compromessi e dubbi invece che soluzioni ottimali. La valutazione etica non fornisce nessuna certezza sul domani, ma porsi queste domande rimane un passo fondamentale verso l'acquisizione di maggiore consapevolezza sul potenziale impatto etico e sociale della gestione dei dati, che di per sé stessa au-

menta il senso di responsabilità di chi produce, mobilita e analizza i dati nei confronti dell'impatto di queste procedure. In questo modo, la produzione di conoscenza viene posizionata rispetto a chi ne può fruire, e l'etica diventa parte integrante del processo di ricerca.

Ci sono molti esempi di innovazioni sviluppate grazie all'analisi di Big Data il cui effetto nefasto si sarebbe potuto moderare se solo chi le ha sviluppate si fosse interrogato in maniera seria e sistematica sull'impatto delle scelte fatte. Prendiamo di nuovo il caso di Facebook, che specialmente nei primi anni di vita ha notoriamente fagocitato e rivenduto i dati personali dei suoi utenti senza nessun riguardo per le potenziali conseguenze, diventando così un vero e proprio Grande Fratello – uno strumento di sorveglianza usato da molte compagnie e istituzioni come fonte di informazioni sui cittadini, trascurando il fatto che questo tipo di *social media* non rispecchia necessariamente la vita reale delle persone e può quindi generare conoscenza completamente inattendibile. Questo tipo di abuso, per cui Facebook fatica ancora ad assumersi responsabilità, ha inizialmente aiutato la compagnia a crescere, ma nel lungo termine le si è ritorto contro, danneggiando enormemente la sua immagine e il rapporto di fiducia con gli utenti. Tanti altri esempi sono legati al tentativo di usare Big Data per scopi medici. Un altro caso ben noto è quello di Google Flu Trends, un programma lanciato nel 2008 con lo scopo di usare i Big Data generati dalle ricerche fatte su Google per prevedere

l'apparizione di epidemie di influenza. L'idea era di sfruttare il fatto che molti degli utenti che ricercano parole come "influenza", "sintomi" e "febbre" lo fanno molto prima di chiamare un dottore, e a volte addirittura come alternativa al sistema sanitario. Google sperava quindi di analizzare questi dati per ricavarne pronostici molto più affidabili di quelli derivati dall'analisi di dati medici ufficiali, e infatti dichiarò nel 2012 che il programma permetteva di identificare focolai cinque giorni prima che l'infezione diventasse visibile ai servizi di sanità. Il programma non teneva però conto della varietà di terminologie usate dagli utenti per descrivere i sintomi, nonché della quantità di ricerche, formalmente simili a quelle di chi si sente male, fatte in realtà per motivi totalmente diversi – in altre parole, troppo poca ricerca era stata fatta sui potenziali utenti, sugli usi di Google e sul modo in cui questi potessero pervertire l'efficacia dell'analisi dei dati. E infatti questa incarnazione di apprendimento automatico è diventata un emblema di conoscenza inattendibile: mentre nel 2013 il programma non riuscì a prevedere un'epidemia particolarmente grande, nel 2015 un'analisi indipendente delle conclusioni di Google mostrò che il numero di casi diagnosticato da Google Flu Trends era il doppio di quelli effettivamente verificati[1].

Questo tipo di errore nell'identificazione di dati rilevanti, e nella loro classificazione e ana-

[1] Lindstrom (2016).

lisi, dimostra l'impossibilità di separare l'etica dalla valutazione della robustezza, sofisticazione tecnica e credibilità dei metodi scientifici[2]. Nel caso di Google Flu Trends, la conoscenza ottenuta è tanto inaffidabile quanto discriminatoria per gente che l'influenza ce l'ha davvero ma viene esclusa da questo tipo di ricerca. La sfida più grossa per l'uso dei Big Data è proprio quella di generare meccanismi di riflessione e responsabilità, a ogni stadio della gestione dei dati, che aiutino a identificare il prima possibile le potenziali fonti di errore e discriminazione, e permettano di correggere e, se necessario, sanzionare decisioni che risultano essere problematiche dal punto di vista sociale. Siamo sicuramente ancora molto lontani – se mai questo sarà possibile – dal creare tecnologie che possano rimpiazzare la capacità umana di valutare il contesto e le implicazioni delle proprie azioni. Al momento, il modo in cui i Big Data vengono gestiti sta creando una crescente discrepanza tra la crescita esponenziale di banche dati e algoritmi per l'analisi di Big Data e la mancanza di procedure e principi che permettano una valutazione seria del loro impatto. Negli ultimi cinque anni centinaia di programmi simili a Google Flu Trends sono stati creati, spesso senza nessun tipo di controllo e senza la capacità e la volontà di fermarsi a riflettere sul loro effetto sociale e sull'affidabilità scientifica.

[2] Sull'idea del ruolo fondamentale dei valori etici nella scienza, si veda anche Douglas (2009); e Elliott *et al.* (2016) nel caso della ricerca datocentrica.

La produzione, compravendita e analisi dei dati è spesso fatta "perché si può" e non in base a criteri solidi dal punto di vista tecnico ed etico.

La partecipazione sociale e l'importanza di rallentare i tempi di ricerca

L'alternativa è l'introduzione di procedure esplicitamente finalizzate all'identificazione di tali criteri e alla riflessione sulla loro rilevanza in ogni fase dei viaggi dei dati. Si entra qui nel merito della seconda strategia per evitare il determinismo tecnologico, che consiste nell'adozione di processi deliberativi basati sulla consultazione sociale come base per le decisioni tecniche prese da chi analizza i Big Data. Questo tipo di consultazione può apparire utopico agli addetti ai lavori, che spesso operano con poche risorse e sotto grosse pressioni finanziarie. Eppure, l'istituzione di procedure che consentano un dialogo sociale esteso sul trattamento dei dati è un modo immediato e costruttivo di esplorare le implicazioni etiche e sociali della ricerca, invocando l'aiuto di chi queste implicazioni le vive in maniera diretta.

Questa lezione è stata assorbita da tempo dal mondo dei servizi digitali e dei *social media*, in cui il parere degli utenti è richiesto regolarmente e usato per migliorare la qualità e l'utilità delle tecnologie in questione. Ottenere riscontri su elementi tecnici coinvolti nell'analisi dei Big Data pone però due ulteriori difficoltà: una ri-

guarda la *mancanza di incentivi* che possano aiutare a implicare nello sviluppo di sistemi digitali persone che già sono oberate di altre responsabilità; l'altro consiste nel trovare modi intelligenti di *coinvolgere* persone con una preparazione non-scientifica in decisioni che possono apparire incomprensibili al di fuori di una cerchia ristretta di informatici. Ancora una volta, la visione relazionale dei dati può aiutarci nell'affrontare questi ostacoli.

Prima di tutto, la costruzione di procedure, regolamentazioni e strumenti il cui scopo esplicito sia aumentare il dialogo sociale sui sistemi di produzione, gestione e interpretazione dei dati è una base fondamentale per prendere decisioni su cosa è etico, rispetto a quali situazioni, e per chi. La GDPR, ossia la legislazione europea varata nel 2018 per proteggere i cittadini dall'abuso dei loro dati personali, è un passo avanti in questo senso, in quanto richiede a chi riusa dati di documentare accuratamente il modo in cui essi vengono gestiti e di impostare e mantenere un dialogo tra chi analizza i dati e chi ne è l'oggetto. In preparazione all'entrata in vigore di questa legislazione, molti gestori di dati in campo pubblico e privato sono stati forzati a riesaminare i presupposti e le modalità con cui li organizzano e analizzano, e a trovare modi per migliorare la comunicazione e il dialogo con i propri utenti. Questo è sicuramente un esercizio dispendioso che ha l'effetto di rallentare e limitare la produzione di conoscenza nel corto termine, ma che ha anche grande potenzialità

di migliorarne la qualità e l'impatto sociale nel lungo termine – e dimostra come l'implementazione di dialogo sociale sulla gestione dei dati sia molto più semplice da implementare durante il processo di costruzione di banche dati, piuttosto che in maniera retroattiva.

In secondo luogo, abbiamo già visto come anche all'interno del mondo della ricerca non esiste chi ha una comprensione perfetta e totale dei sistemi usati per gestire i dati. Ci troviamo dunque già in una situazione in cui persone con punti di vista e capacità diversi devono collaborare per creare un sistema che funzioni nel suo insieme. La partecipazione non richiede quindi una totale rieducazione della popolazione e la trasformazione di tutti i cittadini in esperti informatici, ma piuttosto la creazione di canali di comunicazione in cui analisti si confrontino con gruppi di potenziali utenti in modo da esaminare e discutere il loro operato. Questa comunicazione deve essere il più libera possibile, in modo da facilitare uno scambio equo tra tecnici e pubblico, incoraggiando così i tecnici a modificare i loro sistemi digitali tenendo conto delle esigenze e delle obiezioni emerse nel dialogo. Allo stesso tempo, tutti i partecipanti lavorano sotto limiti tecnici e metodologici precisi, che gli esperti in ingegneria elettronica e programmazione devono fare il possibile per comunicare agli altri, e che anche chi di queste cose non capisce nulla deve essere aperto a considerare. Non c'è simmetria in scambi di questo genere, né esistono garanzie che la comunicazione funzioni bene e

risulti in apprendimento reciproco – e in diversi momenti e situazioni dello sviluppo di Big Data richiedono sicuramente sforzi differenti da parte di chi è coinvolto. Ma l'apertura al dialogo e al confronto sociale più ampio possibile rimane fondamentale per la gestione e l'analisi dei Big Data per scopi di ricerca.

Un ottimo esempio di questo tipo di scambio, e del modo in cui possa contribuire alla qualità della ricerca, è il sistema con cui lo stato inglese gestisce la sperimentazione sugli animali – un altro settore in cui l'applicazione di principi etici nel processo di ricerca dipende enormemente dal caso specifico ed è estremamente controverso dal punto di vista sociale (ricordiamoci che in Inghilterra esistono molti gruppi dedicati alla salvaguardia della vita animale, alcuni dei quali hanno adottato tattiche di intimidazione e violenza nei confronti di biologi che usano animali per le loro ricerche). Sia lo stato sia le stesse comunità scientifiche si sono impegnati nel creare spazi di dialogo in cui i ricercatori possano discutere il loro motivi per utilizzare animali e ricevere spunti per ridurne il numero e migliorarne il trattamento in modi compatibili con gli obbiettivi dei loro progetti. Inoltre, tutti i progetti che utilizzano animali vengono esaminati regolarmente da ispettori governativi con competenze in biologia e in questioni etiche e legali. Il continuo confronto tra ricercatori e ispettori non è volto semplicemente all'implementazione di regole prestabilite, ma soprattutto all'incoraggiamento di riflessione su come la ricerca si sta

sviluppando e sulle possibili implicazioni sugli animali usati, e l'elaborazione di giustificazioni esplicite delle scelte fatte e dei compromessi così raggiunti. Il risultato è la creazione di momenti in cui i ricercatori possono sospendere temporaneamente la loro frenetica attività di ricerca e prendere tempo per chiedersi come migliorarne i metodi e l'impatto, per esempio modificando il trattamento degli animali a seconda dei risultati ottenuti fino ad allora[3].

Questo tipo di incoraggiamento a prendersi momenti di riflessione e valutazione del proprio operato può sembrare banale, ma è in realtà rivoluzionario rispetto agli incentivi e alle strutture istituzionali ed economiche in cui la ricerca sui Big Data viene tipicamente condotta. Troppo spesso i Big Data sono visti come un mezzo per accelerare enormemente la produzione di conoscenza a scapito di ogni "scrupolo". La costruzione di procedure che incoraggiano la gestione etica dei Big Data può invece aiutare a migliorare l'attendibilità e la qualità della conoscenza prodotta, il suo valore metodologico, la responsabilizzazione dei ricercatori coinvolti, la sostenibilità delle banche dati coinvolte e l'attenzione all'uso dei dati per scopi realmente innovativi. Proprio come nel caso dello *slow food,* la *slow science* costituisce una valida alternativa al modello di utilizzo dei Big Data che prevede una crescente alienazione tra le procedure di ricerca

[3] Ho discusso questo esempio e la sua rilevanza per la scienza dei dati in Leonelli (2016b).

e le preferenze, esigenze e sfide che ne caratterizzano il contesto sociale.

Molti degli esperti che lavorano con Big Data sono i primi a sottolineare la contraddizione tra la complessità della gestione dei dati e l'aspettativa che forniscano conoscenza attendibile in maniera semplice, veloce e socialmente accettabile. Proprio per questo, ci sono ingegneri, informatici e archivisti che insistono sull'adozione di un *codice di comportamento* per la scienza dei dati che, in maniera simile al giuramento ippocratico fatto dai medici, incoraggi chi analizza i Big Data a prendersi responsabilità per le eventuali conseguenze delle loro scelte – inclusa quella di confrontarsi con altri per quanto possibile allo scopo di identificare in maniera più efficace e partecipativa quali possano essere tali conseguenze[4].

Esistono anche istituzioni create apposta per la salvaguardia dei diritti dei cittadini i cui comportamenti vengono studiati tramite Big Data. La banca dati *Secure Anonymized Information Linkage* (SAIL) in Galles, per esempio, è stata messa in piedi quindici anni fa per custodire e rendere anonimi dati sensibili usati per la ricerca medica, ma si è progressivamente evoluta in un centro capace di organizzare consultazioni tra tanti tipi di esperti, mediare tra le esigenze di pazienti, dottori e ricercatori, e consigliare gli

[4] Vari esempi di cosa questo codice possa contenere sono dati da Boyd (2012), Dove *et al.* (2016), e Zook *et al.* (2017); si vedano anche i miei suggerimenti alla fine di questo capitolo.

scienziati su che tipo di ricerca fare sui questi dati, e in quale modo. Questa funzione di mediazione sociale ha permesso a SAIL di trasformarsi da risorsa puramente strumentale a parte integrante del processo di ricerca[5]. Questa esperienza è condivisa da molte delle banche dati con cui ho lavorato nel corso degli anni, la cui sopravvivenza e il cui progressivo riconoscimento come punto di riferimento cruciale per intere comunità di ricercatori è in gran parte dovuto alla loro capacità di facilitare comunicazione tra i vari settori coinvolti nei viaggi dei dati e fornire opportunità per confronti e per l'elaborazione di soluzioni comuni. Anche l'organizzazione della Scienza Aperta riconosce e valorizza il ruolo della mediazione e della comunicazione sociale; e l'istituzione di consorzi volti proprio a favorire questi scambi è stata identificata come una componente cruciale per l'ulteriore sviluppo di forme di Scienza Aperta[6].

Inoltre, la crescente attenzione a questioni etiche sta fungendo da stimolo per lo sviluppo di nuovi tipi di algoritmi e tecnologie di analisi, il cui scopo è proprio consentire l'indagine di dati potenzialmente utili per la ricerca senza però allo stesso tempo mobilitare grandi quantità di dati considerati come sensibili o comunque privati. Un esempio è la creazione

[5] Jones *et al.* (2014); abbiamo analizzato la storia e le caratteristiche di questa infrastruttura in Tempini & Leonelli (2018).

[6] Si veda Vallance *et al.* (2016) e Leonelli (in corso di stampa).

di algoritmi e accordi tra banche dati che consentono di collegare e analizzare alcune caratteristiche dei dati custoditi in località diverse senza dover accedere alle banche dati nella loro interezza. Questo consente l'analisi dei dati senza necessariamente doverli condividere e spostarli da un posto all'altro, ritenendo così il controllo su chi può averne accesso e sugli scopi per cui possono essere utilizzati[7].

Nessuna di queste soluzioni è ideale o universale nel suo impatto, e ognuna di esse ha più o meno senso a seconda del tipo di dati e di situazione sociale e culturale in cui questi ultimi vengono analizzati. Tutte queste soluzioni sono però modi per incoraggiare la partecipazione sociale nella ricerca, e la concettualizzazione della ricerca stessa come in continuo dialogo con valori e problemi sociali. È chiaro che ci sono momenti nei viaggi dei dati – per esempio quelli in cui nuove tecniche di programmazione e di custodia di essi vengono elaborate – in cui i ricercatori operano in maniera separata e indipendente da immediate questioni sociali, e questa relativa indipendenza gioca una funzione importante nello sviluppo di tecnologie e conoscenze che trascendano il momento storico in cui sono state concepite. Allo stesso tempo, però, i ricercatori hanno una forte responsabilità nei confronti delle applicazioni immediate dei loro risultati, che include l'uso dei metodi e dei

[7] Si veda anche Richards *et al.* (2015) e l'esempio di DataSchield (Burton *et al.* 2015).

concetti scientifici sviluppati nel corso di secoli proprio per migliorare la qualità e l'integrità dell'analisi dei dati; e tutti gli altri settori sociali hanno la responsabilità di interessarsi e aprirsi al confronto sui presupposti e sulle scelte tramite le quali vengono realizzati gli strumenti di produzione di conoscenza. Ci sono già tanti casi in cui gente che non ha né formazione né ruoli professionali nel mondo della ricerca contribuisce in modi decisivi all'analisi di Big Data: si pensi solo alla realizzazione di "*health apps*" volte alla quantificazione della nostra forma fisica; dati sull'ambiente e il clima; dati prodotti da servizi sociali e demografici; dati forniti da pazienti per la ricerca biomedica. È fondamentale riconoscere che gli esperti in Big Data non sono solo quelli che vengono pagati per fare analisi di dati e che conoscono metodi statistici e computazionali. La *data expertise* comprende anche la conoscenza delle condizioni in cui i dati vengono raccolti, il modo in cui vengono raccolti, e le implicazioni del loro uso – e questa *expertise* svolge un ruolo fondamentale nell'indirizzare le scelte tecniche.

Principi guida per facilitare la trasformazione dei Big Data in conoscenza affidabile

Come si può tradurre questa analisi in pratica? Questa sezione, che conclude il capitolo, identifica otto "principi guida" per la gestione pratica dei Big Data, ognuno dei quali ha una o più conseguenze pratiche per vari settori sociali

(sicuramente non esaustive, ma scelte per esemplificare almeno alcune delle possibili ramificazioni). Questi principi derivano dall'analisi sviluppata sia in questo libro sia in vari altri studi, citati nel testo, sull'impatto sociale dei Big Data. Questa lista, seppure altamente semplificata e incompleta, vuole essere un punto di riferimento iniziale per chi è coinvolto nella creazione, disseminazione o riutilizzo dei dati nel mondo della ricerca o altrove.

Principio 1: il "dato" è una categoria relazionale.

Non c'è dato senza relazione. I dati, che siano Big o meno, sono interpretabili solo in base a una rete di relazioni concettuali, materiali e sociali che deve essere resa esplicita in modo da giustificare i risultati dell'analisi.

– *Conseguenze pratiche*: una gestione dei dati attenta sia alle circostanze in cui sono stati prodotti sia a quelle in cui sono mobilitati è indispensabile al loro riuso. La storia dei dati – gli stadi dei loro viaggi, i materiali da cui sono stati estratti e le trasformazioni che hanno subito – deve essere documentata in maniera esplicita e facilmente accessibile da chi cerca di analizzarli.

Principio 2: La manutenzione regolare e a lungo termine delle infrastrutture è necessaria per giustificare la fiducia nei Big Data.

L'accumulo e l'interoperabilità dei Big Data richiedono un enorme apparato concettuale, ma-

teriale e istituzionale nella forma di infrastrutture, banche dati, regolamentazione e programmi di formazione adeguati. Ciò a sua volta richiede finanziamenti specifici e sostanziosi volti al mantenimento e all'aggiornamento periodico di questo apparato nel lungo termine.

– *Conseguenza pratica*: le istituzioni di ricerca devono lavorare col governo a livello nazionale e con consorzi e associazioni internazionali per sviluppare e mantenere sistemi efficienti per la cura e la manutenzione dei dati (come per esempio nel caso della *European Open Science Cloud*, con cui molti governi e associazioni di ricerca stanno collaborando in direzione della creazione di un sistema federale per la custodia e la cura di dati prodotti da finanziamenti europei)[8].

Principio 3: Infrastrutture e abilità di gestione dei dati sono essenziali all'estrazione di conoscenza dai Big Data.

È cruciale che i ricercatori impegnati nell'analisi di Big Data si interessino al funzionamento delle banche dati e degli algoritmi usati per mobilitare e analizzare i dati, in modo da poter valutare in maniera critica l'impatto di questi strumenti, delle relative metodologie e dei sistemi di classificazione sulla conoscenza estratta dai dati. Allo stesso tempo, le istituzioni di ricerca

[8] Si veda il sito della Commissione Europea dedicato all'EOSC: http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

devono riconoscere l'importanza di assumere esperti nella gestione e analisi dei dati che possano aiutare i ricercatori a negoziare le complessità insite nel loro esercizio.

– *Conseguenza pratica*: Gli esperti sui dati (i cosiddetti *data scientists*) devono essere valorizzati dalle università e dal mondo industriale come un nuovo tipo di figura professionale indispensabile all'utilizzo corretto dei Big Data. I ricercatori in ogni campo devono ricevere un'educazione di base per l'utilizzo di tecnologie, infrastrutture e metodi relativi all'analisi dei Big Data.

Principio 4: Lo spazio per la ricerca esplorativa deve essere preservato.

È importante evitare di incanalarsi nell'utilizzo di certi tipi di dati solo perché sono più facili da produrre o mobilitare.

– *Conseguenza pratica*: L'analisi dei Big Data deve essere rivolta anche all'identificazione di aree di conoscenza su cui non ci sono dati disponibili, e quindi stimolare tanto il riciclo di dati vecchi quanto la produzione di nuovi dati.

Principio 5: La ricerca scientifica deve avvantaggiarsi di quante più fonti di dati possibile, tenendo conto dei rischi di discriminazione e ineguaglianza legati all'uso dei Big Data.

La triangolazione dei Big Data aiuta a produrre conoscenza più affidabile solo nei casi in

cui i dati usati hanno storie differenti e separate (ossia, come discusso nel capitolo 3, provengono da stirpi diverse). Qualora questo non sia possibile, i ricercatori devono evidenziare i modi in cui diversi tipi di dati possono o meno essere comparati, rifacendosi a metodi già ben sviluppati nei campi scientifici di riferimento.

– *Conseguenza*: La scelta di quali dati inserire in una banca dati deve essere documentata in maniera esplicita e informata da metodi di campionamento già da tempo usati nelle scienze naturali e sociali.

Principio 6: L'etica, la sicurezza e la responsabilità sociale sono parti integranti della ricerca datocentrica.

Non c'è modo di separare il valore etico dei dati da quello epistemologico e scientifico, e per quanto l'analisi dei dati possa essere regolamentata e standardizzata, chi si incarica di organizzare e analizzare Big Data rimane responsabile per il modo in cui le regole vengono applicate in ogni caso specifico. Scelte fondamentali all'estrazione di conoscenza dai Big Data hanno implicazioni decisive per il livello di conservatismo, attendibilità, parzialità, corruzione e sensibilità sociale incorporato nei processi e nei risultati della ricerca. Per quanto distribuite tra tante persone con competenze differenti e altamente specializzate, tutte le fasi del lavoro associato ai viaggi dei dati – inclusa la loro pianificazione, creazione, mobilitazione e analisi – comportano

quindi una dose di responsabilità verso le conseguenze sociali degli strumenti e della conoscenza che sono così ottenuti.

– *Conseguenza pratica*: Ogni elemento coinvolto nella gestione dei Big Data deve essere valutato sia per il suo valore tecnico sia per le sue implicazioni etiche e sociali, come indicato dai vari codici di comportamento per *data scientists* emersi recentemente per guidarne l'operato.

Principio 7: L'uso dei Big Data per scopi di ricerca è legato al dialogo sociale sui presupposti usati per analizzarli nei vari contesti di applicazione.

La partecipazione civica è un requisito essenziale alla disseminazione e all'uso dei Big Data per produrre conoscenza scientifica, in quanto aiuta i ricercatori a identificare potenziali problemi in relazione a specifiche situazioni di riuso, e a trovare soluzioni adeguate.

– *Conseguenza pratica*: una o più forme di consultazione e dialogo sociale devono essere incorporate all'interno delle procedure per la costruzione e il mantenimento di infrastrutture e di tecniche di analisi dedicate ai Big Data. Istituzioni rivolte alla mediazione tra diversi utenti di Big Data possono essere estremamente utili sia alla gestione dei dati stessi sia alla produzione di conoscenza affidabile e appropriata al contesto sociale.

Principio 8: È fondamentale che ogni settore sociale coinvolto nell'utilizzo di conoscenze e tecno-

logie derivate dall'analisi di Big Data si interessi della base empirica di queste conoscenze e abbia gli strumenti necessari per interagire con le scelte fatte a livello tecnico.

– *Conseguenze pratiche*: La creazione di forme di dialogo e partecipazione sociale nei processi scientifici deve essere valutata e promossa come parte importante della ricerca da governi, università e finanziatori, migliorando così gli incentivi per coinvolgere gli scienziati in questi sforzi. Il curriculum scolastico nazionale a ogni livello (da scuole elementari a superiori) deve includere discussioni ed educazione pratica sulla gestione e sull'interpretazione dei dati, mentre lo stato e la società civile devono impegnarsi a organizzare corsi di aggiornamento e occasioni di dialogo per adulti.

Conclusione

Questo libro ha cercato di dimostrare come quello che conta come dato, e il modo in cui i Big Data generano conoscenza, dipende dalle tecnologie coinvolte nel produrre, mobilitare e analizzare i dati – e quindi dalle decisioni e dalle scelte fatte dalle persone responsabili per la gestione dei dati, dalle risorse economiche e sociali a loro disposizione, nonché dagli incentivi e dagli scopi per cui i dati vengono collezionati e analizzati. Ho difeso due messaggi fondamentali. Uno è che il pluralismo e la variabilità nel tipo di conoscenza e nei metodi usati nel mondo della ricerca sono preziosi, e devono essere riconosciuti e sfruttati invece che eliminati come fonte di complicazioni. Il secondo è che l'impatto dei Big Data sulla ricerca e sul futuro della conoscenza umana dipende dal modo in cui tutti noi, e specialmente chi si occupa di dati nel proprio lavoro, ci confrontiamo con tre ambiti strettamente correlati tra di loro:

1. La gestione delle banche dati: come e da chi vanno gestite le infrastrutture responsabili

per custodire e ordinare i dati, che modalità di accesso ai dati sono preferibili, e quali tipi di dati (e relative informazioni) dovrebbero essere collezionati e riutilizzati?

2. L'attendibilità dei dati: chi giudica la qualità dei dati e delle tecnologie usate per analizzarli, come, e a che punto del processo di ricerca e interpretazione? Come possiamo assicurarci che i dati disponibili in rete siano gestiti correttamente e ci siano soluzioni sostenibili per aggiornare le collezioni e gli algoritmi attualmente in essere?

3. La partecipazione e l'interazione con i dati: come si può organizzare un dialogo tra tutti coloro che hanno interesse nella gestione e nell'uso dei dati o nella conoscenza che ne viene tratta? Che ruolo hanno i diversi tipi di abilità ed *expertise* coinvolti nell'interpretazione dei dati, e come possono interagire in maniera costruttiva? Chi è escluso, sfruttato o vittimizzato dall'uso dei Big Data, e cosa si può fare per mitigare l'impatto negativo nei loro confronti?

Spendere tempo nel confrontarsi con queste domande può essere percepito dai ricercatori come una perdita di tempo prezioso e un ostacolo ulteriore a studi già molto complessi e di difficile risoluzione. Alcuni ricercatori sono anche – e comprensibilmente – preoccupati che un eccessivo peso conferito a considerazioni etiche costituisca un ostacolo alla produzione di alcune forme di conoscenza, sia perché aggiunge ulteriori barriere amministrative a processi di ricerca già oberati dalla burocrazia (per esempio, quando i ricercatori devono chiedere permesso per ri-

usare dati personali in nuovi contesti di ricerca) sia perché alcuni dei modi in cui i dati vengono gestiti ne riducono la capacità di essere usati come fonte di conoscenza (per esempio quando procedure di "anonimizzazione" riducono fortemente la granularità dei dati generati da ricerca su soggetti umani). La protezione di dati personali può essere tanto problematica, per esempio in casi di ricerca fatta in più località o culture diverse, da scoraggiare i ricercatori dal porre certi tipi di domande, rendendo così impossibili alcuni tipi di ricerca e, quindi, di conoscenza.

Questi sono svantaggi considerevoli dell'approccio che ho proposto, ma non c'è modo di evitarli senza incorrere nel pericolo ben più grande di produrre conoscenza inaffidabile e dannosa, secondo quanto descritto nel secondo capitolo. In più, abbiamo visto come alcuni di questi svantaggi siano legati a opportunità di migliorare sia le procedure sia i risultati della ricerca, favorendo un rallentamento del ritmo di produzione scientifica in favore della sua qualità etica e scientifica. Prestare maggiore attenzione e sforzo nel determinare quali dati siano più appropriati a un certo tipo di progetto, e come debbano essere curati in modo da essere riutilizzati da altri con interessi diversi, stimola la creazione di meccanismi che rendano le banche dati più sostenibili; migliora la qualità e l'attendibilità dei dati stessi; e riconfigura le relazioni tra tutti coloro che sono coinvolti nella ricerca, come per esempio quelle tra pazienti e ricercatori clinici, in modo da rendere il processo investigativo più

aperto all'integrazione con fonti di conoscenza esterne a quelle tradizionalmente riconosciute come "scientifiche", e quindi potenzialmente più riflessivo e informato sulle possibili implicazioni future dell'uso dei Big Data.

Bibliografia

Anderson C., *The end of theory. The data deluge makes the scientific method obsolete*, in "Wired", Giugno 2008. URL: https://www.wired.com/2008/06/pb-theory/.

Ankeny, R.A. *The overlooked role of cases in causal attribution in medicine* in "Philosophy of Science" *81*(5), 2014, pp. 999-1011.

Ankeny R.A., Leonelli S., *Repertoires: A Post-Kuhnian Perspective on Scientific Change and Collaborative Research*, in "Studies in the History and the Philosophy of Science: Part A", 60, 2016, pp. 18-28.

Aronova E., van Hoerzen C., Sepkoski D., *Introduction: Historicizing Big Data*, "Osiris", 32 (1), 2018, pp. 1-17.

Barnes B., Dupré J., *Genomes and What to Make of Them*, University of Chicago Press 2008.

Beer D., *Metric Power*, Palgrave Macmillan, Basingstoke 2016.

Bezuidenhout L., Leonelli S., Kelly A., Rappert B., *Beyond the Digital Divide. Towards a Situated Approach to Open Data*, in "Science and Public Policy", 44, n. 4, 2017, pp. 464-475.

Bezuidenhout L., Kelly A., Leonelli S., Rappert B., *"$100 Is Not Much To You": Open Science and Neglected Accessibilities for Scientific Research in Africa*, in "Critical Public Health", 2016, pp. 1-11.

Bogen J., Woodward J., *Saving the Phenomena*, in "*The Philosophical Review*" 97 (3), 1988, pp. 303-352.

Borgman C.L., *Big Data, Little Data, No Data*, MIT Press, Cambridge 2015.

Boulton G., Campbell P., Collins B. *et al.*, *Science as an open enterprise. The Royal Society Science Policy Centre Report 02/12*, The Royal Society Publishing, London 2012.

Boumans M., Leonelli S., *From Dirty Data to Tidy Facts: Practices of Clustering in Plant Phenomics and Business Cycles*, in Leonelli S., Tempini N. (a cura di), *Varieties of Data Journeys*, 2019.

Bowker G.C., *Science on the run: information management and industrial science at*

Schlumberger, 1920-1940, MIT Press, Cambridge 1994.

Bowker G.C., Star S.L., *Sorting Things Out*, MIT Press, Cambridge 1999.

Boyd D., *Critical Questions for Big Data*, in "Information, Communication & Society", 4462, June 2012, pp. 37-41.

Boyd D., Crawford K., *Six Provocations for Big Data*, in "Data & Society Paper" 2011.

Broadbent A., *Philosophy of Epidemiology*, Palgrave Macmillan, Basingstoke 2013.

Bokulich A., *Using Models to Correct Data: Paleodiversity and the Fossil Record*, in "Synthese", 2018.

Burton P.R., Murtagh M.J., Boyd A. *et al.*, *Data Safe Havens in Health Research and Healthcare*, in "Bioinformatics", 31, n. 20, 2015, pp. 3241-3248.

Cai L., Zhu Y., *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*, in "Data Science Journal", 14, p. 2, 2015.

Cambrosio A., Keating P., Nelson N., *Régimes Thérapeutiques et Dispositifs de Preuve en Oncologie. L'organisation des Essais Cliniques, des Groupes Coopérateurs aux Consortiums de Recherche*, in "Sciences Sociales et Santé", 32, 2014, pp. 13-42.

Canali S., *Big Data, Epistemology and Causality. Knowledge in and Knowledge out in EXPOsOMICS*, in "Big Data & Society", 3, n. 2, 2016, pp. 1-11.

Caporael L.R., Griesemer J.R., Wimsatt W.C., *Developing Scaffolds in Evolution, Culture and Cognition*, MIT Press, Cambridge 2014.

Chang H., *Is Water H_2O?*, Springer Netherlands, Dordrecht 2012.

Chang H., *Pragmatist Coherence as the Source of Truth and Reality*, in "Proceedings of the Aristotelian Society", CXVII, n. 2, 2017.

Chapman R., Wylie A. (a cura di), *Material Evidence. Learning from Archaeological Practice*, Routledge, Oxon and New York 2015.

Curioni A., *La Protezione dei Dati. Guida Pratica al Regolamento Europeo*, Mimesis Editore, Milano 2017.

Daston L., *The Moral Economy of Science,* in "Osiris", 10, 1995, pp. 2-24.

Derrida J., *Mal D'Archive: Une Impression Freudienne*, Éditions Galilée, Parigi 1995.

Directorate-General for Research and Innovation (European Commission), *Incentives and Rewards to Engage in Open Science Activities. Thematic Report No. 3 for the Mutual Learning Exercise Open Science: Altmetrics and Rewards*, Publications Office of the European Union, Luxembourg 2017.

Directorate-General for Research and Innovation (European Commission), *Implementing Open Science. Strategies, Experiences and Models. Thematic Report No. 4 for the Mutual Learning Exercise on Open Science: Altmetrics and Rewards,* Publications Office of the European Union, Luxembourg 2018.

Douglas H., *Science, Policy and the Value-Free Ideal*, University of Pittsburgh Press, Pittsburgh 2009.

Dove E.S., David T., Meslin E. *et al.*, *Ethics Review for International Data-Intensive Research*, in "Science", 351, n. 6280, 2016, pp. 1399-1400.

Dupré J., *The Disorder of Things. Metaphysical Foundations of the Disunity of Science*, Harvard University Press, Cambridge and London 1983.

Ebeling M.F.E., *Healthcare and Big Data. Digital Specters and Phantom Objects*, Palgrave Macmillan, New York 2016.

Edwards P.N., *A vast machine: Computer models, climate data, and the politics of global warming*, MIT Press, Cambridge 2010.

Edwards P.N., Mayernik M.S., Batcheller A.L. *et al.*, *Science Friction. Data, metadata, and collaboration*, in "Social Studies of Science", 41, n. 5, 2011, pp. 667-690.

Elliott K.C., Cheruvelil K.S., Montgomery G.M., Soranno P.A., *Conceptions of Good Science in Our Data-Rich World*, in "BioScience", 66, n. 10, 2016, pp. 880-889.

Fecher B., Friesike S., Hebing M., *What Drives Academic Data Sharing?*, in "PLoS ONE", 10, n. 2, e0118053, 2015.

Feest U., *What Exactly is Stabilized When Phenomena are Stabilized?*, in "Synthese" 182 (1), 2011, pp. 57-71.

Floridi L., *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, trad. it M. Durante, Raffaello Cortina Editore, Milano 2017.

Floridi L., Illari P. (a cura di), *The Philosophy of Information Quality. Synthese Library 358*, Springer, Cham Heidelberg New York Dordrecht London 2014.

Foucault M., *Le Parole e le cose*, trad. it E. Panaitescu, Rizzoli, Milano 1967.

Gitelman L., *"Raw data" is an Oxymoron*, MIT Press, Cambridge 2013.

Harris A., Kelly S., Wyatt S., *CyberGenetics. Health Genetics and New Media*, Routledge/Taylor & Francis Group, Londra 2016.

Hey T., Tansley S., Tolle K., *The fourth paradigm. Data-intensive scientific discovery*, Microsoft Research, Redmond 2009.

Hilgartner S., *Constituting large-scale biology. Building a re-*

gime of governance in the early years of the Human Genome Project, in "BioSocieties", 8, 2013, pp. 397-416.

Hine C., *Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work*, in "Social Studies of Science", 36, n. 2, 2006, pp. 269-298.

Jones K., Ford D.V., Jones C. *et al.*, *A Case Study of the Secure Anonymous Information Linkage (SAIL) Gateway. A Privacy-Protecting Remote Access System for Health-Related Research and Evaluation*, in "Journal of Biomedical Informatics", 50 (Special Issue on Informatics Methods in Medical Privacy), August, 2014, pp. 196-204.

Kellert S.H., Logino H.E., Waters C.K. (a cura di), *Scientific Pluralism*, University of Minnesota Press, Minneapolis 2006.

Kitchin R., *The data revolution. Big data, open data, data infrastructures and their consequences*, SAGE, Londra 2014.

Kitchin R., McArdle G., *What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets*, in "Big Data & Society", 3, n. 1, 2016, pp. 1-10.

Landecker H., *Culturing Life: How Cells Became Technologies*, Harvard University Press, Cambridge 2007.

Latour B., *Science in Action: How to Follow Scientists and Engineers through Society*, Harvard University Press, Cambridge 1987.

Leonelli S., *When Humans Are the Exception. Cross-Species Databases at the Interface of Clinical and Biological Research*, in "Social Studies of Science", 42, n. 2, 2012, pp. 214-236.

Leonelli S., *What Difference Does Quantity Make? On the Epistemology of Big Data in Biology*, in "Big Data and Society", 1, 2014, pp. 1-11.

Leonelli S., *Data-Centric Biology: A Philosophical Study*, Chicago University Press, Chicago 2016a.

Leonelli S., *Locating ethics in data science: responsibility and accountability in global and distributed knowledge*

production, in "Philosophical Transactions of the Royal Society: Part A", 374, n. 2083, 2016b, 20160122.

Leonelli S., *Biomedical Knowledge Production in the Age of Big Data. Analysis conducted on behalf of the Swiss Science and Innovation Council SSIC*, Bern 2017a, URL: https://www.swir.ch/it/pubblicazioni.

Leonelli S., *Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology*, in "Data Science Journal", 16, n. 32, 2017b, pp. 1-11.

Leonelli S., *The Time of Data. Time-Scales of Data Use in the Life Sciences*, in "Philosophy of Science", 85 (5), 2018.

Leonelli S., *Scientific Agency and Social Scaffolding in Contemporary Data-Intensive Biology*, in Wimsatt W., Love A.C. (a cura di), *Beyond the Meme. Articulating Dynamic Structures in Cultural Evolution*, University of Minnesota Press, Minneapolis 2019.

Leonelli S., in corso di stampa. *What Distinguishes Data from Models?*

Leonelli S., Diehl A.D., Christie K.R., Harris M.A., Lomax J., *How the Gene Ontology Evolves*, in "BMC Bioinformatics", 12, 2011.

Leonelli S., Tempini N., *Where Health and Environment Meet. The Use of Invariant Parameters in Big Data Analysis*, in "Synthese", Special issue on Philosophy of Epidemiology, 2018.

Levin N., Leonelli S., *How Does One "Open" Science? Questions of Value in Biological Research*, in "Science, Technology and Human Values", 42, n. 2, 2016, pp. 280-305.

Levin N., Leonelli S., Weckowska D. *et al.*, *How Do Scientists Understand Openness? Exploring the Relationship between Open Science Policies and Research Practice*, in "Bulletin for Science and Technology Studies", 36, n. 2, 2016, pp. 128-141.

Lindstrom M., *Small Data: The Tiny Clues that Uncover Huge Trends*, St Martin's Press, New York 2016.

Loettgers A., *Synthetic Biology and the Emergence of a Dual Meaning of Noise*, in "Biological Theory", 4, n. 4, 2009, pp. 340-355.

Marr B., *Big Data. Using SMART big data, analytics and metrics to take better decisions and improve performance*, John Wiley & Sons, Hoboken 2015.

Massimi M., *From Data to Phenomena: A Kantian Stance*, in "Synthese" 182 (1), 2011, pp. 101-116.

Mauthner N.S., Parry O., *Open Access Digital Data Sharing: Principles, Policies and Practices*, in "Social Epistemology", 27, n. 1, 2013, pp. 47-67.

Mayer-Schönberger V., Cukier K., *Big Data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, trad. It. R. Merlini, Garzanti, Milano 2013.

Mayo D., *Error and the Growth of Experimental Knowledge,* University of Chicago Press, Chicago 1996.

McAllister J.W., *Model Selection and the Multiplicity of Patterns in Empirical Data*, in "Philosophy of Science", 74, n. 5, 2007, pp. 884-894.

Mittelstadt B.D., Floridi L. (a cura di), *The Ethics of Biomedical Big Data*, Springer Switzerland, Basel 2016.

Mongilli A., Pellegrino G. (a cura di), *Information Infrastructure(s). Boundaries, Ecologies, Multiplicity*, Cambridge Scholars Publishing, Cambridge 2014.

Müller-Wille S., Charmantier I., *Natural history and information overload. The case of Linnaeus*, in "Studies in History and Philosophy of Science Part C", 43, 2012, pp. 4-15.

Müller-Wille S., Rheinberger H., *A Cultural History of Heredity*, University of Chicago Press, Chicago 2012.

Murphy M., *The Economization of Life*, Duke University Press, Durham 2017.

Normandeau K., *Beyond volume, variety and velocity is the issue of big data veracity*, sul sito "Inside Big Data", 2013, URL: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity.

November J., *Biomedical Computing: Digitazing Life in the United States*, The John Hopkins University Press, Baltimora 2012.

Oreskes N., Conway E.M., *Merchants of Doubt*, Bloomsbury Press, Londra 2010.

Ossorio P., *Bodies of data: Genomic data and bioscience data sharing*, in "Social Research", 78, n. 3, 2011, pp. 907-932.

Pestre D., *Regimes of knowledge production in society. Towards a more political and social reading*, in "Minerva", 41, 2003, pp. 245-261.

Prainsack B., *Personalised Medicine. Empowered Patients in the 21st Century?*, New York University Press, New York 2017.

B. Prainsack, A. Buyx, *Solidarity in Biomedicine and Beyond*, Cambridge University Press, Cambridge 2017.

Rappert B., Selgelid M.J., *On the Dual Uses of Science and Ethics: Principles, Practices and Prospects*, ANU Press, Canberra 2013.

de Regt H.W., *Understanding Scientific Understanding*, Oxford University Press, Oxford 2017.

de Regt H.W., Leonelli S., Eigner K., *Scientific Understanding: Philosophical Perspectives*, University of Pittsburgh Press, Pittsburgh 2009.

Rheinberger H., *Infra-Experimentality: From Traces to Data, From Data to Patterning Facts*, in "History of Science" 49(164), 2011, pp. 337-348.

Ribes D., Polk J.B., *Organizing for Ontological Change. The Kernel of a Research Infrastructure*, in "Social Studies of Science", 45, n. 2, 2015, pp. 214-241.

Richards M., Anderson R., Hinde S. *et al.*, *The collection, linking and use of data in biomedical research and health care. Ethical issues*, Nuffield Council on Bioethics, Londra 2015.

Sætnan A.R., Schneider I., Green S. (a cura di), *The Policy and Politics of Big Data*, Routledge, Oxon 2018.

Sansone S.A., Rocca-Serra P., Field D. *et al.*, *Toward Inte-*

roperable Bioscience Data, in "Nature Genetics", 44, n. 2, 2012, pp. 121-126.

Science International, *Open data in a big data world. An international accord*, ICSU, ISCC TWAS, & IAP, Parigi 2015.

Shavit A., Griesemer J.R., *There and back again, or the problem of locality in biodiversity surveys*, in "Philosophy of Science", 76, 2011, pp. 273-294.

Srnicek N., *Platform Capitalism*, Polity Press, Cambridge and Malden 2017.

Stevens H., *Life out of Sequence. A Data-Driven History of Bioinformatics*, University of Chicago Press, Chicago (IL) 2013.

Strasser B.J., *The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine, 1979-1982*, in "Isis", 102, 2011, pp. 60-96.

Strasser B.J., Edwards P., *Big Data is the Answer… But What is the Question?*, in "Osiris"
32 (1), 2017, pp. 328-345.

Sunder Rajan K., *Pharmocracy. Value, Politics, and Knowledge in Global Medicine. Duke University Press*, Durham 2017.

Suppes P., *Models of data*, in Nagel E., Suppes P., Tarski A. (a cura di), *Logic, methodology and philosophy of science*, Stanford University Press, Stanford, 1962.

Teller P., *Saving the Phenomena Today*, in "Philosophy of Science" 77 (5), 2010, pp. 815-826.

Tempini N., *Till Data Do Us Part. Understanding Data-based Value Creation in Data-Intensive Infrastructures*, in "Information & Organization", 27, 2017, pp. 191-210.

Tempini N., Leonelli S., *Concealment and Discovery: The Role of Information Security in Biomedical Data Re-Use*, "Social Studies of Science", in corso di stampa.

Timmermans S., Epstein S., *A World full of Standards but not a Standard World: Toward a Sociology of Standardization*, in "Annual Review of Sociology", 36, 2010, pp. 69-89.

Thrift N., *Knowing capitalism*, SAGE, Londra 2005.

Vallance P., Freeman A., Stewart M., *Data Sharing as Part of the Normal Scientific Process. A View from the Pharmaceutical Industry*, in "PLoS Medicine", 13, n. 1, e1001936, 2016.

Vayena E., Tasioulas J., *The Dynamics of Big Data and Human Rights. The Case of Scientific Research*, in "Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences", 374, n. 2083, id 20160129, 2016.

Vermeir K., Leonelli S., Shams Bin Tariq A. *et al.*, *Global Access to Research Software: The Forgotten Pillar of Open Science Implementation. A Global Young Academy Report*, Global Young Academy, Halle 2018.

Ward J.S., Barker A., *Undefined by data. A survey of big data definitions*, School of Computer Science at the University of St Andrews, St. Andrews 2013.

Wilkinson M.D. *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*, in "Scientific Data" 3:160018 doi: 10.1038/sdata.2016.18, 2016.

Woodward J., *Data, Phenomena, Signal, and Noise*, in "Philosophy of Science", 77, n. 5, 2010, pp. 792-803.

Wouters P., Beaulieu A., Scharnhorst A., Wyatt S. (a cura di), *Virtual Knowledge. Experimenting in the Humanities and the Social Sciences*, MIT Press, Cambridge 2013.

Wylie A., *Thinking from Things. Essays in the Philosophy of Archaeology*, University of California Press, Berkeley 2002.

Wylie A., *How Archaeological Evidence Bites Back. Strategies for Putting Old Data to Work in New Ways*, in "Science, Technology, and Human Values", 42, n. 2, 2017, pp. 203-225.

Zook M., Barocas S., Boyd D. *et al.*, *Ten Simple Rules for Responsible Big Data Research*, in "PLoS Computational Biology", 13, n. 3, e1005399, 2017.

Ringraziamenti

La ricerca che ha portato a questo libro è stata finanziata dai seguenti enti pubblici, a cui sono enormemente grata per il loro supporto: European Research Council (DATA_SCIENCE grant award 335925, "The Epistemology of Data-Intensive Science"), Leverhulme Trust (award RPG-2013-153), Australian Research Council (award DP160102989), U.K. Medical Research Council and Natural Environment Research Council (award MR/K019341/1) e U.K. Economic and Social Research Council (award ES/P011489/1). Giovanni Boniolo mi ha dato l'idea di scrivere questo libro, che senza il suo incoraggiamento e le sue capacità editoriali non esisterebbe. Il Centro di Logica e Filosofia della Scienza dell'università di Ghent in Belgio, che mi ha ospitato nella primavera del 2018, mi ha fornito uno spazio perfetto per redigere il manoscritto – e senza il caffè del bar Emmy dubito che sarei riuscita a finirlo in tempo. Michel Durinx mi ha aiutato a impostare le figure e la bibliografia con considerevole efficienza e disponibilità, e

Stefano Canali ha fornito suggerimenti editoriali all'ultimo minuto. Purtroppo non c'è modo per ringraziare adeguatamente le centinaia di scienziati e colleghi di filosofia, sociologia, antropologia e storia della scienza con cui ho avuto il privilegio di discutere queste idee negli ultimi dieci anni. Le seguenti persone meritano una menzione individuale visto il ruolo fondamentale che hanno giocato nel plasmare la mia ricerca sui dati nel corso del progetto "DATA_SCIENCE" (in ordine alfabetico): Rachel Ankeny, Elizabeth Arnaud, Ruth Bastow, Bill Bechtel, Louise Bezuidenhout, Marcel Boumans, Alberto Cambrosio, Hasok Chang, Adrian Currie, Gail Davies, John Dupré, Lora Fleming, Luciano Floridi, Jean-Paul Gaudillière, James Griesemer, Gregor Halfmann, Mary Morgan, Rebecca Lovell, Staffan Müller-Wille, Barbara Prainsack, Hans Radder, Ed Ramsden, Brian Rappert, Hans-Jörg Rheinberger, Beckett Sterner, Kaushik Sunder Rajan, David Teira, Niccolò Tempini, Sally Wyatt e Alison Wylie. L'affetto dei miei amici di sempre continua a darmi forza e ispirazione ogni giorno – un grazie enorme a tutti, specialmente a Elena e Fabio che hanno letto e commentato il manoscritto prima della pubblicazione, e a Stefano che mi tiene ancorata al dibattito italiano. E, come sempre, devo tutto al supporto della mia famiglia e alla pazienza e all'amore di Leonardo, Luna e Michel.

Finito di stampare
nel mese di settembre 2018
a Digital Team - Fano (PU)