deriamo fenomeni naturali o sociali, ma come un universo composto essenzialmente da informazioni. I fisici lo vanno dicendo da oltre un secolo: alla base di tutto non ci sono gli atomi, ma le informazioni.[26] Ammettiamolo: può sembrare esoterico. Ma oggi, attraverso la datizzazione, in molti casi possiamo cogliere e calcolare con estrema precisione gli aspetti fisici e intangibili dell'esistenza, e sfruttarli a nostro vantaggio.

Il fatto di vedere il mondo come una massa d'informazioni, un oceano di dati che si possono esplorare sempre più estesamente e sempre più approfonditamente, ci offre una prospettiva sulla realtà che prima non avevamo. È un approccio mentale che potrebbe estendersi a tutti gli ambiti della vita. Oggi abbiamo una società che privilegia i numeri perché assumiamo che il mondo si possa comprendere attraverso la matematica. E diamo per scontato che le conoscenze si possano trasmettere nel tempo e nello spazio, perché il concetto della parola scritta è profondamente radicato nel nostro modo di pensare. Domani, le nuove generazioni potrebbero essere ispirate da una «consapevolezza dei big data» – la presunzione che ci sia una componente quantitativa in tutto quello che facciamo, e che i dati siano indispensabili per l'apprendimento della società. Oggi, l'idea di trasformare le infinite dimensioni della realtà in dati appare quasi certamente innovativa alla maggior parte della gente. Ma in futuro la considereremo un fatto acquisito (come vuole l'etimologia della parola «dati»).

Con il tempo, l'impatto della datizzazione farà impallidire quello degli acquedotti e dei giornali, e la porterà a rivaleggiare con la macchina da stampa e con Internet mettendoci a disposizione i mezzi per mappare il mondo in maniera quantificabile e analizzabile. Per il momento, tuttavia, gli utilizzatori più avanzati della datizzazione stanno nelle imprese, dove i big data si usano per creare nuove forme di valore – come vedremo nel prossimo capitolo.

6. VALORE

Alla fine degli anni Novanta del Novecento, il web si stava trasformando rapidamente in un ambiente caotico, inospitale e ostile. I cosiddetti «spambot» [programmi per la raccolta di indirizzi cui inviare in automatico posta elettronica indesiderata, detta anche «spam», *n.d.t.*] riempivano le caselle di posta e inondavano i forum online. Nel 2000 Luis von Ahn, un ventiduenne che si era appena laureato, ha avuto un'idea che poteva risolvere il problema: obbligare coloro che si registravano a dimostrare la propria condizione di esseri umani. Perciò si è messo alla ricerca di qualcosa che fosse facile da fare per gli umani, ma difficile per le macchine.

Ha pensato così di presentare delle lettere a svolazzi, difficilmente leggibili, nella fase di registrazione. Le persone sarebbero state in grado di decifrarle e di digitare il testo corretto in pochi secondi, mentre i computer sarebbero rimasti disorientati. Yahoo ha implementato il suo metodo e ha ridotto quasi da un giorno all'altro il flagello dello spam. Von Ahn ha chiamato la sua creatura Captcha (acronimo di Completely Automated Public Turing Test to Tell Computers and Human Apart). Cinque anni dopo, si digitavano milioni di Captcha ogni giorno.

Captcha ha procurato a von Ahn una discreta fama e una cattedra di computer science alla Carnegie Mellon University dopo il conseguimento del PhD. Ha contribuito anche ad assicurargli, a soli ventisette anni, uno dei prestigiosi premi da mezzo milione di dollari che vengono conferiti ai «geni» dalla MacArthur Foundation. Ma

quando si è reso conto di aver costretto milioni di persone a perdere ogni giorno un sacco di tempo con quelle irritanti lettere a svolazzi – una gran massa d'informazioni che poi venivano semplicemente buttate via – non si è sentito poi così geniale.

Nel tentativo di mettere maggiormente a frutto tutta quella capacità di decifrazione, ha sviluppato un software evolutivo, opportunamente denominato reCaptcha. Invece di digitare delle lettere messe lì a caso, le persone digitano due parole estratte da progetti di scannerizzazione del testo che un programma di riconoscimento dei caratteri ottici di un computer non potrebbe identificare. Una serve a confermare ciò che hanno digitato altri utenti, e quindi segnala la natura umana del rispondente; l'altra è una nuova parola da sottoporre a disambiguazione. Per garantire l'accuratezza, il sistema presenta la stessa parola da decifrare mediamente a cinque persone diverse che la dovranno digitare correttamente, prima di certificarne la validità. Quei dati avevano uno scopo primario – dimostrare che l'utilizzatore era un essere umano –, ma anche uno secondario: decifrare parole ambigue nei testi digitalizzati.

Il valore che genera questa soluzione è immenso, se pensiamo cosa costerebbe ingaggiare delle persone per fare lo stesso lavoro. A una media di dieci secondi alla volta, per 200 milioni di reCaptcha al giorno – il volume attuale –, fanno mezzo milione di ore al giorno. Nel 2012, il salario minimo negli Stati Uniti era 7,25 dollari al giorno. Se ci si dovesse rivolgere al mercato per disambiguare le parole a cui il computer non riesce a dare un significato, si spenderebbero circa 4 milioni di dollari al giorno, ossia oltre un miliardo di dollari all'anno. Invece von Ahn ha progettato un sistema in grado di svolgere quel lavoro praticamente gratis. Era una soluzione così preziosa che nel 2009 Google ha acquistato la tecnologia dal suo inventore, mettendola gratuitamente a disposizione di qua-

lunque sito; oggi è presente in circa 20.000 siti, inclusi Facebook, Twitter e Craigslist.[1]

La vicenda di reCaptcha mette in evidenza quanto sia importante il riutilizzo dei dati. Con l'avvento dei big data, il loro valore si sta modificando. Nell'era digitale, hanno perso il ruolo di supporto alle transazioni e sono diventati essi stessi oggetto delle transazioni. In un mondo dominato dai big data, la situazione è destinata a modificarsi ulteriormente. Il valore dei dati si sposta dall'utilizzo primario ai possibili utilizzi futuri. Ciò ha conseguenze importantissime. Incide sul valore che attribuiscono le imprese ai dati in loro possesso e sul profilo dei soggetti a cui permettono di accedervi. Aiuta le aziende, o le costringe addirittura a cambiare il proprio modello di business. Ne modifica il pensiero sui dati e su come li utilizzano.

Le informazioni sono sempre state essenziali per le transazioni di mercato. I dati consentono per esempio l'individuazione del prezzo, che segnala quanto produrre. Questa dimensione dei dati è ben nota. Certe informazioni si negoziano da sempre sul mercato. I contenuti che si trovano nei libri, negli articoli, nella musica e nei film sono un esempio, al pari di informazioni finanziarie come le quotazioni azionarie. Negli ultimi decenni si sono aggiunti anche i dati personali. Negli Stati Uniti, intermediari specializzati nella fornitura di fatti personali come Acxiom, Experian ed Equifax si fanno pagare profumatamente ricchi dossier di informazioni personali su centinaia di milioni di consumatori. Con Facebook, Twitter, LinkedIn e altri social media, le nostre relazioni interpersonali, le nostre opinioni, le nostre preferenze e le nostre abitudini quotidiane sono venute ad ampliare la quantità di informazioni già disponibili su di noi.

In poche parole, pur essendo sempre stati preziosi, i dati venivano considerati funzionali alla gestione dell'impresa, o limitati a categorie relativamente ristrette come la proprietà intellettuale o le informazioni personali. Per

contro, nell'era dei big data, tutti i dati verranno considerati preziosi, in sé e per sé.

Quando diciamo «tutti i dati», ci riferiamo anche alle informazioni grezze e apparentemente più banali. Pensate alle letture fornite da un sensore termico applicato a una macchina industriale. O al flusso in tempo reale delle coordinate registrate da un rilevatore GPS, ai dati acquisiti da un accelerometro e ai livelli di carburante segnalati da un furgone adibito alle consegne – o da una flotta di 60.000 furgoni. Oppure a miliardi di vecchie parole-chiave digitate sui motori di ricerca, o al prezzo di quasi tutti i posti sui voli commerciali degli Stati Uniti, recuperato andando indietro di anni.

Fino a qualche tempo fa, non c'era modo di raccogliere, archiviare e analizzare facilmente questi dati, il che limitava sensibilmente la possibilità di estrarne del valore. Nel celebre esempio della fabbrica di spilli utilizzato da Adam Smith, con cui illustrava la divisione del lavoro nel XVIII secolo, i ricercatori avrebbero dovuto osservare l'attività di tutti gli operai non solo ai fini di quello studio, ma tutto il giorno per tutti i giorni, effettuando delle misurazioni analitiche e riportandole sulla pergamena con la penna d'oca.[2] Quando gli economisti classici studiarono i fattori di produzione (terra, manodopera e capitale), l'idea di sfruttare i dati era praticamente assente. Anche se negli ultimi due secoli il costo di raccolta e utilizzo dei dati è diminuito, fino a non molto tempo fa è rimasto relativamente elevato.

Ciò che differenzia la nostra epoca è che molte limitazioni intrinseche alla raccolta dei dati non esistono più. La tecnologia è arrivata a un punto in cui si possono raccogliere e immagazzinare spesso grandi masse d'informazioni a basso costo. Non di rado, i dati si possono raccogliere passivamente, senza grossi sforzi da parte di chi li fornisce, e talora anche a sua insaputa. E siccome il costo di archiviazione è diminuito enormemente, è più facile

giustificare la conservazione che l'eliminazione dei dati. Negli ultimi cinquant'anni, il costo dell'archiviazione digitale si è più o meno dimezzato ogni due anni, mentre la densità dei dati in memoria è aumentata di 50 volte. Grazie ad aziende come Farecast o Google – nella cui catena di montaggio digitale entrano fatti grezzi e ne escono informazioni processate – i dati cominciano ad apparire una nuova risorsa o un nuovo fattore di produzione.[3]

Il valore immediato di quasi tutti i dati è evidente per coloro che li raccolgono. In realtà, probabilmente li rilevano con un fine specifico in mente. I grandi magazzini raccolgono i dati di vendita per la contabilità finanziaria. La fabbriche tengono sotto controllo il proprio output per essere sicure che sia in linea con gli standard di qualità. I siti web registrano tutti i click dei visitatori – a volte anche la direzione del puntatore – per analizzare e ottimizzare i propri contenuti. Questi utilizzi primari dei dati ne giustificano la raccolta e la processazione. Quando Amazon registra non solo i libri che acquistano i clienti, ma anche le pagine web che si limitano a scorrere, sa che utilizzerà quei dati per offrire raccomandazioni personalizzate. Analogamente, Facebook registra gli «aggiornamenti di status» e i «Mi piace» degli utilizzatori per identificare i messaggi pubblicitari più idonei a generare ricavi.

Diversamente da quanto accade per le cose materiali – il cibo che mangiamo, una candela che brucia – il valore dei dati non diminuisce quando vengono utilizzati; si possono riprocessare all'infinito. Le informazioni sono perciò, come dicono gli economisti, un bene «non competitivo»: il fatto che una persona le usi non impedisce a un'altra di utilizzarle. E le informazioni non si logorano come fanno i beni materiali. Perciò Amazon può utilizzare i dati estrapolati da transazioni pregresse quando raccomanda dei titoli ai suoi visitatori – e li usa ripetutamente, non solo per il cliente che ha generato i dati, ma anche per molti altri.

I dati si possono sfruttare molte volte per lo stesso scopo, ma anche e soprattutto per più scopi insieme. È un punto importante se vogliamo capire quanto ci saranno utili le informazioni nell'era dei big data. Abbiamo già visto realizzarsi una parte di questo potenziale, come quando Walmart ha passato al setaccio il proprio database di vecchi scontrini e ha scoperto la lucrosa correlazione tra uragani e vendite di merendine dolci.

Tutto questo fa pensare che il valore effettivo dei dati sia molto superiore a quello ricavato dal suo primo utilizzo. Significa inoltre che le aziende possono sfruttare efficacemente i dati anche se il primo utilizzo, o ciascuno dei successivi, apporta un valore marginale; a condizione di sfruttarli innumerevoli volte.

Il «valore opzionale» dei dati

Per farvi un'idea di quello che potrebbe essere il valore effettivo generato dal riuso dei dati, considerate le auto elettriche. La loro definitiva affermazione come mezzo di trasporto dipende da tutta una serie di fattori logistici che hanno qualcosa a che fare con la durata delle batterie. Gli automobilisti devono essere in grado di ricaricarle rapidamente e comodamente, e le aziende elettriche devono fare in modo che l'energia assorbita da questi veicoli non destabilizzi la rete. Oggi abbiamo una distribuzione sostanzialmente efficace delle stazioni di servizio, ma non sappiamo ancora quale sarà il reale fabbisogno di ricarica delle batterie e quale dislocazione dei centri di rifornimento sarà necessaria.

Curiosamente, non è tanto un problema infrastrutturale quanto un problema informativo. E i big data hanno un peso decisivo nella soluzione. In un esperimento effettuato nel 2012, l'IBM ha collaborato con la californiana Pacific Gas and Electric Company e con la casa automobilistica

Honda nella raccolta di enormi masse d'informazioni per rispondere a interrogativi fondamentali su dove e quando le auto elettriche dovranno fare rifornimento, e con quale impatto per la rete. L'IBM ha sviluppato un modello previsionale sofisticato, che si basava su numerosi input: il livello di carica della batteria, l'ubicazione della vettura, l'ora del giorno e i punti di accesso disponibili nelle stazioni di ricarica del circondario. Ha poi abbinato i dati al prelievo corrente di energia dalla rete e agli andamenti storici di utilizzo dell'elettricità. L'analisi di quei flussi giganteschi di dati in tempo reale oppure storici, provenienti da svariate fonti, ha consentito a Big Blue di determinare i tempi e i luoghi ottimali per la ricarica delle batterie. Alla fine, il sistema dovrà tener conto delle differenze di prezzo delle stazioni di ricarica della zona. Bisognerà prendere in considerazione anche le previsioni del tempo: per esempio, se c'è il sole e una stazione di ricarica è strapiena di elettricità ma è prevista una settimana di pioggia in cui i pannelli solari resteranno inattivi.

Il sistema prende le informazioni generate per uno scopo e le riutilizza per un altro – vale a dire che i dati passano da usi primari a usi secondari. Ciò li rende molto più preziosi nel tempo. L'indicatore di livello della batteria dice agli automobilisti quando rimetterla in carica. I dati sul grado di utilizzo della rete elettrica vengono raccolti dall'azienda fornitrice in modo da poterne assicurare la stabilità. Questi gli usi primari. Entrambe le categorie di dati trovano degli utilizzi secondari – e offrono nuovo valore – quando vengono destinate a uno scopo completamente diverso: stabilire quando e dove ricaricare le batterie, oppure dove costruire stazioni di rifornimento per vetture elettriche. In tutto ciò sono incorporate informazioni secondarie, come il posizionamento della macchina e i consumi storici a carico della rete elettrica. E l'IBM processa i dati non una volta per tutte, ma in continuazione, così come aggiorna costantemente l'assorbimento di e-

nergia da parte delle auto elettriche e l'impatto sulla rete distributiva.[4]

Il vero valore dei dati si può paragonare a un iceberg che si sposta lentamente nell'oceano. Emerge solo la punta, mentre tutto il resto si nasconde sotto la superficie. Le aziende innovative che lo capiscono possono estrarre quel valore occulto e conseguire dei benefici potenzialmente enormi. In poche parole, il valore dei dati va calcolato in base a tutti i possibili modi con cui si potrebbero impiegare in futuro, e non semplicemente in base all'uso che se ne fa attualmente. Lo abbiamo visto in molti degli esempi citati in precedenza. Farecast sfruttava i dati relativi ai biglietti venduti per stimare le tariffe praticate in futuro dalle compagnie aeree. Google riutilizzava le parole-chiave per individuare le zone di diffusione dell'influenza. Maury riprendeva in mano i giornali di bordo di vecchi comandanti per identificare le correnti marine.

Eppure, l'importanza del riuso dei dati non viene pienamente apprezzata dalle imprese e dalla società. Pochi executive di Con Edison avrebbero mai immaginato che i record secolari di operatività e manutenzione dei cavi si potessero usare per prevenire futuri incidenti. Ci sono volute una nuova generazione di statistici e una nuova ondata di metodi e di strumenti per fare emergere il valore dei dati. Fino a poco tempo fa, anche molte Internet companies e aziende dell'alta tecnologia ignoravano quanto possa essere prezioso il riutilizzo dei dati.

Bisognerebbe vedere i dati nella stessa prospettiva in cui i fisici vedono l'energia. Parlano di energia «immagazzinata» o «potenziale», insita in un oggetto ma non utilizzata. Pensate a una molla compressa o a una palla ferma in cima a una collina. L'energia presente in questi oggetti rimane latente – o potenziale – finché non viene liberata, cioè quando la molla viene rilasciata o la palla viene spinta leggermente in avanti, facendola rotolare lungo il pendio. L'energia contenuta in questi oggetti diventa «cinetica»,

perché sono in movimento ed esercitano una forza su altri oggetti. Dopo il primo utilizzo, i dati mantengono il proprio valore, che rimane inespresso come l'energia potenziale immagazzinata nella molla o nella palla, finché non si applicano a un uso secondario e liberano ex novo la forza che hanno in sé. Nell'era dei big data, abbiamo finalmente la mentalità, la creatività e gli strumenti che occorrono per sfruttare il valore occulto dei dati.

Alla fine, il loro valore è ciò che si può ricavare da tutti i modi in cui si possono impiegare. Questi utilizzi potenziali apparentemente infiniti sono paragonabili alle opzioni – non nel senso degli strumenti finanziari ma nel senso pratico delle scelte. Il valore dei dati è la somma di queste scelte: il «valore opzionale» dei dati, per così dire. In passato, una volta realizzato lo scopo principale dei dati, pensavamo quasi sempre che avessero esaurito la propria funzione; ed eravamo pronti a cancellarli, a lasciarli scappar via. Dopotutto, eravamo convinti di averne già estratto il valore essenziale. Nell'era dei big data, i dati si possono paragonare a una magica miniera di diamanti che continua a produrre ricchezza anche dopo lo sfruttamento del valore principale. Ci sono tre modi efficaci per liberare il valore opzionale dei dati: riutilizzo, fusione dei dataset, e identificazione delle possibilità di estensione.

Riutilizzo dei dati

Un tipico esempio di riutilizzo innovativo dei dati concerne le parole-chiave digitate sui motori di ricerca. A prima vista, quelle informazioni appaiono prive di valore dopo il raggiungimento del loro scopo primario. La breve interazione tra consumatore e motore di ricerca ha prodotto un elenco di siti web e di banner pubblicitari che assolvevano una funzione specifica, legata a quel momento. Ma le vecchie queries possono essere straordina-

riamente preziose. Hitwise, un'azienda specializzata nella misurazione del traffico web di proprietà del data broker Experian, consente ai clienti di analizzare le ricerche effettuate per capire le preferenze dei consumatori. Gli operatori della moda possono usare Hitwise per stabilire se la prossima primavera andrà di moda il rosa o se tornerà in auge il nero. Google mette gratuitamente a disposizione una versione del suo programma di analisi dei termini di ricerca. Ha lanciato un servizio di previsioni economico-finanziarie insieme alla seconda banca di Spagna, il BBVA, per valutare le prospettive del settore turistico e vendere indicatori economici in tempo reale che si basano sui dati di ricerca. La Banca d'Inghilterra usa queries di ricerca relative ai beni immobili per capire se i prezzi delle abitazioni sono in aumento o in diminuzione.

Le aziende che non hanno saputo apprezzare l'importanza del riutilizzo dei dati hanno imparato la lezione a proprie spese. Per esempio, quando muoveva i primi passi, Amazon ha sottoscritto un accordo con AOL per gestire la tecnologia del suo sito di commercio elettronico. Per quasi tutti, era un banalissimo contratto di outsourcing. Ma ciò che interessava veramente ad Amazon, come spiega Andreas Weigend, l'ex chief scientist, era impossessarsi dei dati su ciò che cercavano e acquistavano gli utilizzatori di AOL, il che avrebbe migliorato la performance del proprio motore di raccomandazioni.[5] L'ingenua AOL non se n'è mai resa conto. Vedeva il valore dei dati solo in funzione del loro scopo primario, ovvero favorire le vendite. L'astuta Amazon sapeva di poter trarre dei benefici dall'impiego dei dati per uno scopo secondario.

Prendete il caso dell'ingresso di Google nel business del riconoscimento vocale con GOOG-411, una guida telefonica ad azionamento vocale che ha funzionato dal 2007 al 2010. Il colosso delle ricerche online non aveva una sua tecnologia in questo campo, perciò doveva procurarsela in licenza. Ha raggiunto un accordo con Nuan-

ce, il leader di settore, felicissimo di acquisire un cliente così prestigioso. All'epoca Nuance non aveva ancora compreso il valore potenziale dei big data: il contratto non specificava chi avrebbe dovuto conservare i record delle chiamate, così Google se li è tenuti. L'analisi dei dati permette di valutare la probabilità che un determinato frammento vocale corrisponda a una parola specifica. È un fattore essenziale per migliorare la tecnologia di riconoscimento vocale o creare un servizio completamente nuovo. All'epoca Nuance era convinta di operare nel campo del software licensing, e non nell'elaborazione dei dati Nel momento in cui ha preso coscienza del proprio errore, ha cominciato a stringere accordi con operatori della telefonia mobile e produttori di telefonini per la concessione in uso del suo servizio di riconoscimento vocale, proprio allo scopo di raccogliere i dati.[6]

Il valore di riutilizzo dei dati è una buona notizia per le organizzazioni che raccolgono o controllano vasti dataset ma ne fanno un uso limitato, come le aziende tradizionali che operano prevalentemente offline. Forse siedono su un vero e proprio geyser informativo non ancora sfruttato. Alcune aziende avranno raccolto dei dati, li avranno utilizzati una volta sola (ammesso che l'abbiamo fatto) e li avranno conservati – solo perché l'archiviazione costa poco – nelle «data tombs», come si chiamano in gergo i luoghi in cui risiedono delle informazioni molto vecchie.

Le Internet companies e le aziende dell'alta tecnologia sono in prima linea nello sfruttamento economico del «diluvio di dati», in quanto raccolgono un'infinità d'informazioni per il solo fatto di operare sulla Rete e di essere più avanti dei concorrenti nell'analizzarle. Ma tutte le imprese hanno da guadagnarci. I consulenti di McKinsey & Company citano un'azienda logistica, di cui non fanno il nome, che nel processo di consegna delle merci ha raccolto casualmente una massa enorme d'informazioni sulle spedizioni dei prodotti intorno al mondo. Fiutando

un'opportunità, ha costituito un'apposita divisione per vendere i dati aggregati sotto forma di previsioni economico-finanziarie. In altre parole, ha creato una versione offline del business di Google dedicato all'analisi delle queries pregresse.[7] O considerate SWIFT, il sistema interbancario globale per il trasferimento elettronico dei fondi: ha scoperto che i pagamenti sono correlati con l'attività economica mondiale. Di conseguenza, SWIFT offre previsioni del PIL basate sui dati relativi al trasferimento di fondi che viaggiano sul suo network.

Alcune imprese, grazie alla posizione che occupano nella catena del valore, potrebbero essere in grado di raccogliere ingenti quantità di dati, anche se non ne hanno un bisogno immediato o non sono esperte nel riutilizzarli. Per esempio, gli operatori di telefonia mobile raccolgono dati sull'ubicazione dei propri abbonati per istradare le chiamate. Per loro, questi dati hanno un utilizzo strettamente tecnico. Ma diventano più preziosi quando vengono riusati da aziende che distribuiscono messaggi pubblicitari e promozionali personalizzati e localizzati. A volte il valore non deriva dai singoli data point, ma da ciò che rivelano nel loro insieme. Ecco perché aziende di geolocalizzazione come AirSage e Sense Networks, che abbiamo incontrato nell'ultimo capitolo, possono vendere informazioni su dove si incontra le gente il venerdì sera o sulla lentezza con cui procedono le auto nel traffico. Queste preziose informazioni si possono usare per determinare il valore effettivo degli immobili o i prezzi dei cartelloni pubblicitari.

Anche le informazioni più banali potrebbero avere un valore particolare, se impiegate nel modo giusto. Pensate ancora agli operatori di telefonia mobile: sanno esattamente dove e quando i telefoni si collegano alle stazioni radio base, inclusa la forza del segnale. Hanno sempre usato quei dati per mettere a punto la performance dei propri network, decidendo dove potenziare o migliorare

l'infrastruttura. Ma essi hanno tanti altri utilizzi potenziali. I produttori di telefonini potrebbero usarli per capire cosa influenza la forza del segnale, per esempio, allo scopo di migliorare la capacità di ricezione dei loro apparecchi. Gli operatori di telefonia mobile si sono sempre rifiutati di monetizzare quelle informazioni per paura di violare le norme sulla privacy. Ma con il peggioramento dei risultati finanziari stanno cominciando a diventare più flessibili e a vedere nei dati una fonte potenziale di reddito. Nel 2012 il grande operatore spagnolo e internazionale Telefonica è arrivato al punto di creare un'azienda separata, denominata Telefonica Digital Insights, per vendere dati anonimi e aggregati sull'ubicazione degli abbonati a catene di supermercati e altre aziende.[8]

Dati ricombinanti

A volte il valore latente si può liberare solo combinando un dataset con un altro, anche del tutto diverso. Possiamo fare cose innovative mescolando i dati con modalità nuove. Un esempio in proposito ce lo fornisce uno studio pubblicato nel 2011 sulla possibilità che l'uso dei cellulari faccia aumentare l'incidenza dei tumori. Visto che nel mondo ci sono quasi sei miliardi di telefonini, poco meno di uno per ogni abitante della Terra, l'interrogativo è cruciale. Molti studi hanno cercato un legame tra i due fenomeni, ma avevano tutti quanti dei limiti: i campioni erano troppo ristretti, i periodi analizzati troppo brevi, o si basavano su dati auto riferiti zeppi di errori. Ma un team di ricercatori della Danish Cancer Society ha ideato un approccio suggestivo che si fondava su dati raccolti in precedenza.

Dall'introduzione dei telefoni cellulari in Danimarca, gli operatori hanno sempre raccolto dati sugli abbonati. Lo studio prendeva in esame i possessori di telefonini tra

il 1987 e il 1995, con l'esclusione delle aziende e di altri soggetti di cui non erano disponibili i dati socioeconomici, e ha coinvolto 358.403 persone. Le autorità sanitarie del paese tenevano anche un registro nazionale di tutti i pazienti oncologici, in cui figuravano 10.729 persone che avevano avuto tumori del sistema nervoso centrale tra il 1990 e il 2007, il periodo di indagine. Infine, lo studio ha utilizzato un registro nazionale con i dati sul titolo di studio e il reddito disponibile di ogni cittadino danese. Dopo aver combinato i tre dataset, i ricercatori sono andati a vedere se gli utilizzatori di telefoni cellulari facevano registrare una maggiore incidenza di tumori rispetto ai non-utilizzatori. Volevano capire inoltre se tra gli utilizzatori, quelli che avevano posseduto un cellulare per un periodo più prolungato avevano maggiori probabilità di ammalarsi di cancro.

Nonostante le dimensioni dello studio, i dati non erano affatto caotici o imprecisi: i dataset comportavano rigorosi standard di qualità per finalità mediche, commerciali o demografiche. Le informazioni non erano state raccolte con modalità potenzialmente in grado d'introdurre dei pregiudizi in relazione al tema dello studio. In effetti, i dati erano stati generati anni prima, e per ragioni che non avevano nulla a che fare con questa ricerca. E soprattutto, lo studio non si basava su un campione, ma su un universo vicino al fatidico $N = tutti$, ovvero aveva registrato quasi tutti i casi di tumore, quasi tutti gli utilizzatori di dispositivi mobili, per un totale di 3,8 milioni anni-persona di possesso di un telefono cellulare. Il fatto che includesse quasi tutti i casi consentiva ai ricercatori di effettuare delle verifiche a livello di sottosegmenti della popolazione, come per esempio quello di persone dal reddito elevato.

Alla fine, il team non ha riscontrato alcun incremento del rischio oncologico associato all'utilizzo dei telefoni cellulari. Perciò, quando sono state pubblicate sulla rivista medica britannica «BMJ» nell'ottobre 2011, le sue scoperte non hanno suscitato più di tanto l'interesse dei media. Ma se si fosse trovato un collegamento, lo studio sarebbe finito sulle prime pagine dei giornali di tutto il mondo, e la metodologia dei «dati ricombinanti» sarebbe stata esaltata.[9]

Nella gestione dei big data, la somma è superiore al valore delle singole parti, e quando ricombiniamo più dataset, anche quel totale vale più della somma dei suoi addendi. Oggi, gli utilizzatori di Internet hanno familiarità con i «mashup», che combinano due o più fonti di dati in maniera innovativa. Per esempio, il sito di consulenza immobiliare Zillow sovrappone informazioni sugli immobili e sui prezzi a una mappa dei quartieri e dei distretti suburbani delle città americane. Utilizza inoltre un'infinità di dati, come le ultime transazioni concluse nel distretto e le specifiche degli immobili, per indicare il prezzo di determinate abitazioni. La presentazione visuale rende più accessibili i dati. Ma con i big data possiamo andare ben oltre. Lo studio danese sulla relazione tra cancro e uso del telefono ci dà un'idea di quali siano le possibilità.

Dati estensibili

Una soluzione che facilita il riuso dei dati è quella di renderli strutturalmente flessibili fin dall'inizio, in modo che si prestino a varie applicazioni. Pur non essendo sempre attuabile – possibili utilizzi si potrebbero scoprire solo molto tempo dopo la raccolta dei dati –, ci sono vari modi per incoraggiare utilizzi multipli dello stesso dataset. Per esempio, alcune catene della grande distribuzione stanno posizionando delle telecamere all'interno dei punti vendita, non solo per identificare eventuali taccheggiatori ma anche per rilevare il flusso dei clienti e individuare i punti in cui si fermano a esaminare la merce.

Le informazioni sugli spostamenti della clientela verranno poi utilizzate per ottimizzare il layout del negozio e per valutare l'impatto delle campagne di marketing. In precedenza, le telecamere venivano impiegate esclusivamente per la sicurezza. Oggi si considerano un investimento che potrebbe fare aumentare i ricavi.

Una delle aziende più efficaci nella raccolta dei dati in una prospettiva di estensibilità è, manco a dirlo, Google. Le controverse automobili che impiegava per Street View andavano in giro scattando fotografie di case e strade, ma anche «succhiando» dati GPS, controllando le informazioni riportate sulle mappe e persino attaccandosi a reti wi-fi non protette (e recependone, forse illegalmente, i contenuti). Ognuna di quelle spedizioni produceva una massa gigantesca di dati. L'estensibilità entra in gioco perché Google impiegava i dati non solo per uno scopo primario, ma anche per tanti scopi secondari. Per esempio, i dati GPS che raccoglieva miglioravano il servizio di mappatura offerto dall'azienda ed erano indispensabili per il funzionamento della vettura autoguidata.[10]

Il costo aggiuntivo della raccolta di più flussi di dati, o di molti più data point all'interno di ciascun flusso, è quasi sempre basso. Perciò conviene raccogliere il maggior numero possibile di dati, e renderli estensibili prendendo in considerazione futuri utilizzi secondari fin dall'inizio. Questo approccio fa aumentare il valore opzionale dei dati. Si tratta di cercare delle possibilità di estensione – dove un singolo dataset si può prestare a vari scopi se si raccoglie con certe modalità. Così i dati possono offrire un duplice servizio.

Ammortizzare il valore dei dati

Con il calo dei costi che si associano all'archiviazione digitale dei dati, le imprese hanno un forte incentivo economico a conservarli, in modo da poterli riutilizzare per lo stesso scopo o per scopi analoghi. Ma c'è un limite alla loro utilità. Per esempio, sfruttando gli acquisti, l'iter di navigazione del sito e le recensioni dei clienti per raccomandare altri prodotti, aziende come Netflix e Amazon potrebbero essere indotte a utilizzare ripetutamente i record per molti anni. In quell'ottica, si potrebbe sostenere che se un'azienda non supera i limiti imposti da normative specifiche come le leggi sulla privacy, dovrebbe usare i