Big Data e Privacy by Design

Giuseppe D'Acquisto

g.dacquisto@gpdp.it

"All'uomo sensibile e immaginoso, che viva, come io sono vissuto gran tempo, sentendo di continuo ed immaginando, il mondo e gli oggetti sono in certo modo doppi. Egli vedrà cogli occhi una torre, una campagna; udrà cogli orecchi un suono d'una campana; e nel tempo stesso coll'immaginazione vedrà un'altra torre, un'altra campagna, udrà un altro suono. In questo secondo genere di obbietti sta tutto il bello e il piacevole delle cose".

Giacomo Leopardi – Zibaldone

1. Cosa è "big" nei "Big Data"?

Cosa vi ha portato a leggere queste pagine? Ipotizziamo. Siete studenti e questo testo vi è stato segnalato per formarvi sui temi della privacy e delle tecnologie nell'ambito di un corso universitario o di un master. Oppure, siete dei professionisti – tecnologi o giuristi – e conoscete il dibattito che è attualmente (2016) in corso sul nuovo Regolamento europeo in materia di protezione dei dati personali e l'obbligo che sarà presto introdotto per chi farà uso di dati personali di trattarli secondo il principio di privacy by design di cui ci vogliamo qui occupare. O, magari, conoscete gli autori e ne avete seguito nel tempo le altre pubblicazioni e volete continuare questo dialogo a distanza con noi. E, più concretamente, cosa state leggendo? Forse un file pdf sul vostro PC, oppure la stampa di quel file, o ancora il libro che avete acquistato e che contiene questo testo come suo primo capitolo. E cosa vogliamo dirvi, quale messaggio intendiamo trasmettervi? Be', dovrete spendere una parte del vostro tempo per completare la lettura e farvi un'idea sull'argomento e sul punto di vista che ne hanno gli autori. Se riflettiamo sui vari passaggi di questo semplice processo di formazione della conoscenza, legato a questo particolare e specifico esempio, possiamo sin da subito farci un'idea di cosa significhi "big" quando si parla di Big Data.

Analizziamo le diverse fasi temporali che portano da noi autori, intenti a digitare sulla tastiera questi pensieri, a voi lettori che li leggete in una delle forme che abbiamo appena detto. Si tratta, a ben vedere, di una concatenazione di eventi (ancora) poco integrati e pieni di discontinuità. Nel momento in cui stiamo scrivendo, ad esempio, disponiamo di alcune note scritte su dei fogli di carta e su un file ausiliario, su cui abbiamo preso nota dei vari temi che affronteremo e che consulteremo di quando in quando per orientare il ragionamento e la sua scrittura. Né i fogli di carta, tuttavia, né il file ausiliario sono "consapevoli" del testo che sta per essere scritto. È rimesso interamente all'autore ogni collegamento tra le note e il testo che leggete. Quest'ultimo poi è totalmente isolato dal resto del mondo: non "vede", né "è visto" dagli altri file, che magari pure parlano dello stesso argomento, contenuti nello stesso PC o che si possono trovare in internet sul tema dei Big Data. Inoltre, se guardiamo all'oggetto "libro", il suo essere collocato in una libreria di 10 o di 10.000 libri, né per il libro, né per il lettore fa alcuna differenza. Anche qui, è interamente rimessa al lettore la costruzione di quella complessa relazione tra i "concetti" contenuti nei libri che chiamiamo conoscenza. Eppure, consolidato per quanto sia

questo processo nella storia e nelle nostre abitudini, non è impossibile pensare che si possa fare meglio e più di così. Pensiamo a quanto più spedito potrebbe essere il tempo di redazione e quanto più accurata l'esposizione dei vari punti trattati in queste pagine se, ad esempio mentre digitiamo specifiche parole chiave, o addirittura se potessimo farlo con un maggiore livello di astrazione persino quando affrontiamo particolari "concetti", ci fossero presentate tutte le fonti pertinenti, magari già organizzate per una consultazione critica e con una interfaccia di facile uso, che favoriscano la più compiuta espressione del nostro modo di intendere il tema. E se le librerie parlassero? Quanti suggerimenti potrebbero darci i libri se potessero comunicarci a prima vista i loro "concetti" e se fossero in grado di interconnetterli, e quanti interessi e spinte a conoscere sarebbero capaci di suscitare in noi?

Se poi guardiamo al modo in cui il lettore e questo testo si sono "incontrati", altre discontinuità emergono, Probabilmente tra il momento in cui si è venuti a conoscenza di questo lavoro e il momento in cui se ne stanno leggendo i contenuti è trascorso un certo tempo, vuoi perché si è passati per un cambio di mezzo (ad esempio, la conoscenza è avvenuta tramite una serie di collegamenti su internet e la lettura avviene su una stampa o su un libro acquistato in una libreria o, ancora, su internet e ricevuto successivamente a casa tramite un corriere), vuoi perché tra il momento in cui siete entrati in contatto con il documento o il file e la sua lettura avete attraversato vari "contesti" che vi hanno impedito, o non invogliato a leggere i contenuti, prima che altre incombenze in cui eravate impegnati venissero completate (ad esempio, la lettura del documento che rimandava a questo che state leggendo, oppure altre attività, disgiunte dalla lettura in cui siete impegnati adesso). In ogni caso, è praticamente certo, voi avete "voluto" scaricare questo file, o stamparlo, o comprare questo libro. Potete aver ricevuto molti aiuti nella ricerca (dal motore di ricerca, dai link presenti in altre pagine web che vi hanno portato al file, da un riferimento bibliografico che avete giudicato pertinente, dal suggerimento automatico del vostro venditore di libri online che vi ha presentato questo libro tra quelli che rientravano tra i vostri interessi), ma l'ultimo passo per arrivare alla lettura di queste pagine è una vostra scelta. Persino irrazionale: voi non sapete (se non per grandi linee) ciò che leggerete e se vi interesserà fino in fondo. Potreste, alla fine della lettura, aver perso il vostro tempo. In questa discontinuità di tempi, in genere, l'*hic et nunc* che ci ha spinti a passare dall'assistenza che abbiamo avuto nel "cercare" alla volontà che abbiamo esercitato nel "trovare" si perde, e con esso talora parte del nostro interesse a conoscere. Nel passare dal "cercare" al "trovare" avete, in altri termini, sostenuto un rischio, il cui costo è interamente vostro.

Qui è il punto. Per vedere internet (che poi è il mondo nella sua rappresentazione digitale) in una prospettiva "Big Data" e immaginare come potrebbe essere, bisogna concentrarsi sulla differenza che esiste tra cercare e trovare, sul "costo" che sussiste nel passaggio tra l'una e l'altra attività e su chi lo sostiene.

Torniamo all'esempio dei libri per formulare, allargandola, la stessa domanda che ci siamo fatti per le librerie: e se internet parlasse? Già oggi, lo vediamo, internet ci dice molte cose e non c'è, di fatto, nulla che non possa essere cercato su internet. Dall'albergo in cui trascorreremo le prossime vacanze, a questo testo. Ma tanto l'albergo, quanto questo documento sono come i libri nella libreria dell'esempio. La rete di interconnessioni che lega tra loro gli oggetti informatici presenti su internet, ciò che chiamiamo web, è infatti il frutto di decisioni locali e unilateralmente prese, senza coordinamento, da chi immette i contenuti. In altri termini, è chi pubblica un contenuto a decidere con quale altra risorsa quel contenuto è collegato, creando riferimenti o hyperlink tra quel contenuto e le altre risorse da lui prescelte come pertinenti. L'insieme di tutti gli hyperlink, che rimandano da una risorsa all'altra, creano la "ragnatela" che "copre" tutta la rete. Su questo meccanismo di rimandi lavorano i motori di ricerca, che classificano tutta l'informazione "ricercabile" ordinandola secondo un

criterio di autorevolezza delle fonti che si basa sul numero di collegamenti "entranti" e di visite ricevute: una risorsa è tanto più rilevante quanto maggiore è il numero di collegamenti entranti da altre risorse e di visite ricevute, e quanto più questa ultime sono a loro volta richiamate da altre risorse. Si parla di "saggezza della folla" (*wisdom of the crowd*) per indicare questo meccanismo di ordinamento che non ha un *dominus*, essendo distribuito e determinato dalla totalità degli utenti del web e dalla frequenza delle loro visite ai diversi siti. Google ha reso questo processo misurabile, trasformandolo in un algoritmo, *Pagerank*, che ha perfezionato nel tempo fino a fare diventare il suo motore di ricerca ciò che oggi è: la porta di accesso al web. È, obiettivamente, il modo più efficiente che sia mai stato realizzato per cercare una informazione. Cercare, però, non trovare. Se l'informazione da Google ordinata sia o meno pertinente per noi, è una scelta interamente rimessa al "ricercatore". Anche se l'ambizione di trovare è sempre stata palesemente perseguita da Google (si pensi al bottone "Oggi mi sento fortunato", che è stato introdotto per rimandare direttamente al primo risultato di una ricerca, nella *speranza* che fosse il più rilevante per quella specifica query), il web per come funziona oggi non è in grado di compiere questo ultimo passo: il motore di ricerca "cerca" e noi (ancora) troviamo ciò che ci serve. Questo passaggio, con i rischi e costi associati è ancora interamente nostro.

Cosa manca per compiere questo ultimo passo? Serve una rappresentazione dei dati idonea allo scopo: servono dati "più grandi".

Un dato è tanto più grande quanto più ampia è la sua "sfera di influenza", ossia quanto maggiore è il numero di attributi con cui il dato è descritto e il numero di fenomeni che è potenzialmente in grado di spiegare. Ogni attributo aggiunto alla descrizione di un dato diventa immediatamente una nuova dimensione da esplorare per collegamenti tra quel dato e altri dati, tra il fenomeno rappresentato da quel dato e altri fenomeni rappresentati da altri dati. Più elevato è il numero di descrittori, più verosimile potrà risultare il collegamento di un dato con altri dati, ciascuno a propria volta reso "grande" da un più ricco insieme di attributi. Ciò consentirà di trovare connessioni tra fenomeni che prima erano nascoste, o persino impossibili. Due fenomeni potranno essere messi in relazione tra loro perché i dati che li rappresenteranno mostreranno comunanze, esprimibili dalla presenza in entrambi del medesimo insieme di descrittori, che svolgeranno il ruolo di chiave di collegamento tra l'uno e l'altro.

Un esempio aiuterà a chiarire. Guardiamo queste due immagini.

Figura 1. Primo termine di una relazione Big Data

Figura 2. Secondo termine di una relazione Big Data

Nella prima, vediamo una partita di basket. Nella seconda, un uomo incappucciato, che si affaccia da un balcone. Queste due immagini hanno *qualcosa in comune*. Osservando i completi indossati dai giocatori e le loro capigliature, ci accorgiamo che la partita non è stata giocata di recente. A meno che non siamo appassionati o esperti, poco altro attrae la nostra attenzione. L'altra immagine ci lascia intuire una situazione di pericolo, ma nessun elemento che individui univocamente l'evento a cui si riferisce, anche qui a meno di non essere studiosi o esperti. Eppure, già oggi, queste due foto (questi due dati) compaiono tra i risultati di una interrogazione per immagini al motore di ricerca Google, inserendo come chiave di ricerca la query Olimpiadi+Monaco+1972. La prima immagine, infatti, si riferisce alla storica vittoria della nazionale dell'URSS su quella USA nella finale di basket di quei giochi olimpici, mentre la seconda è l'immagine-simbolo dell'attacco terroristico avvenuto nel corso di quelle stesse olimpiadi al villaggio degli atleti. Sono entrambe immagini molto celebri, tuttavia fino a pochi anni fa dovevate sapere cosa hanno in comune per metterle in relazione l'una con l'altra. Eravate, in altri termini, voi stessi a dover fare il collegamento tra i due fenomeni rappresentati dai due dati, e questo o era ovvio (se eravate esperti) o pressoché impossibile. Già oggi la situazione è obiettivamente diversa: il motore di ricerca stesso mette in relazione i due dati costituiti dalle due immagini e ci offre la possibilità di mettere in relazione i due eventi a cui si riferiscono.

Proviamo ad individuare il processo che porta il motore di ricerca a proporre questa associazione ragionando dapprima in modalità "Small Data", quindi ripetiamo lo stesso processo in modalità "Big Data" per capire come può funzionare uno "schema Big Data". In modalità "Small Data" le due immagini (i due dati) venivano caricate su due server indipendentemente l'una dall'altra, come rappresentato nella figura 3. Ciascuno degli uploader decideva a quali altri dati ciascuna immagine fosse collegata, secondo criteri unilateralmente stabiliti. Nell'esempio in figura, l'uploder 1, immettendo l'immagine della finale di basket (dato D1) sul server S1, la collegava con un hyperlink ad un altro sito di sport, mentre l'uploader 2, immettendo l'immagine dell'uomo incappucciato (dato D2) sul server S2, la collegava con un hyperlink ad un altro sito di storia moderna. Il motore di ricerca, recependo queste scelte dei due uploader e applicando la *wisdom of the crowd* faceva entrare il dato D1 nel circuito dei siti di sport e il dato D2 nel circuito dei siti di storia contemporanea, indicizzandoli separatamente. Da quel momento i due dati avevano vite separate: il primo era in grado di suscitare curiosità legate alla sfera degli eventi sportivi, il secondo collegabile ad altri fatti di terrorismo riconducibili al medesimo contesto, succedutisi nel corso degli anni '70. Solo il "ricercatore", appassionato o esperto, era in grado di riconnetterli.

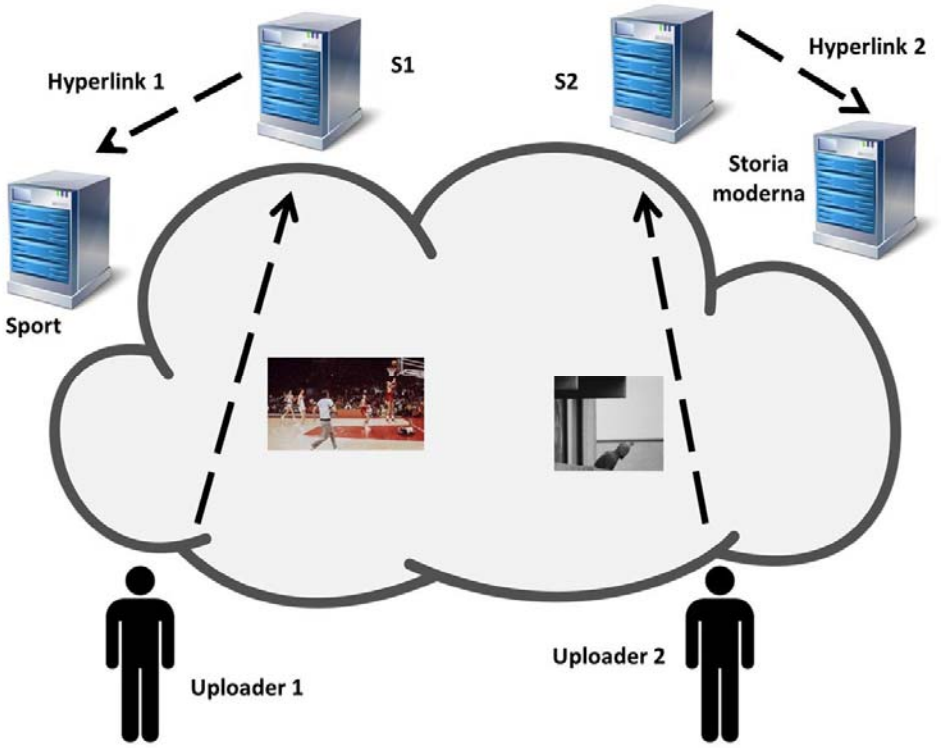


Figura 3. Relazione tra dati in una rete Small Data

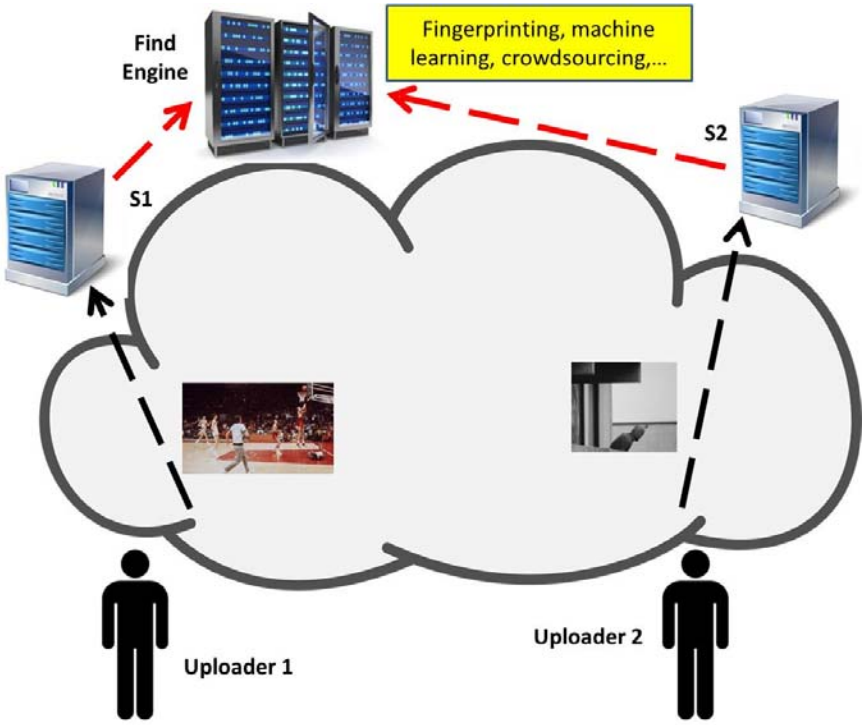


Figura 4. Relazione tra dati in una rete Big Data

In “modalità Big Data”, il motore di ricerca – sempre più “find engine” e non soltanto “search engine” – effettua un processing sui due dati immessi che ne fa emergere il tratto comune, ossia il fatto di riferirsi a due eventi connessi agli stessi giochi olimpici di Monaco del 1972. La situazione è rappresentata nella figura 4. Ma, dove è il cambiamento radicale di questo modo di procedere rispetto allo scenario “Small Data”? Sarà più chiaro se consideriamo un’ulteriore conseguenza della capacità di collegamento tra dati. Immaginiamo che il “find engine” effettui autonomamente una ulteriore associazione, che lega i due fenomeni riconducibili allo stesso evento ad un terzo dato. L’intera vicenda è infatti diventata il soggetto di molti documentari e film, e dunque, partendo dalla stessa chiave comune, il “find engine” potrà proporre *a tutti* l’associazione tra le due immagini e, per esempio, il film del 2005 di Steven Spielberg “Munich”, che ripercorre proprio quelle vicende, magari suscitando curiosità *in qualcuno* e stimolando la visione del film.

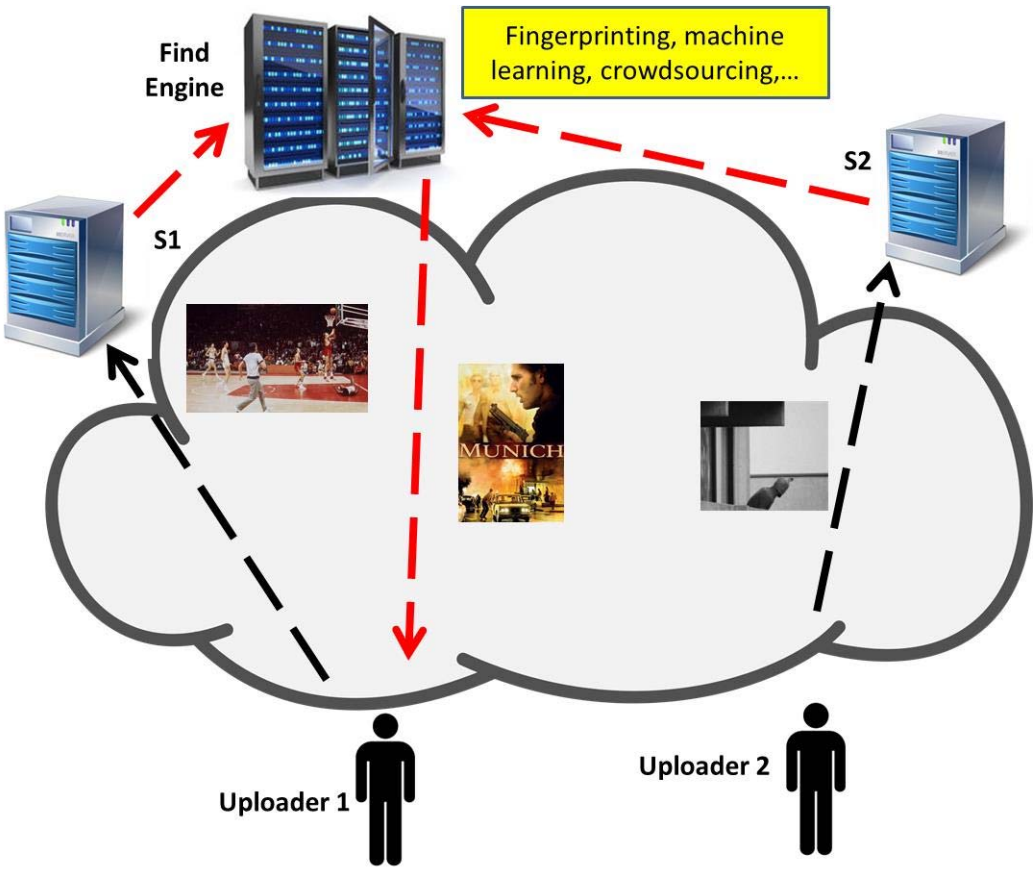


Figura 5. La generazione di nuova conoscenza in una rete Big Data

Ricapitoliamo. Partiamo dal dato D1 (l’immagine di una partita di basket d’altri tempi). Da questo il “find engine”, applicando diverse possibili tecniche di processing estrae una serie di descrittori (operazione che non poteva essere effettuata in un contesto “Small Data”) cercando ogni possibile corrispondenza tra questi nuovi descrittori e i descrittori di altri dati esistenti in internet (a loro volta oggetto delle medesime tecniche di processing che li rendono “più grandi”). Viene trovata una nuova corrispondenza con un altro dato D2 (l’immagine dell’uomo incappucciato al balcone) e una nuova prospettiva di conoscenza viene offerta a tutti (la relazione tra un evento sportivo e un evento della nostra storia più recente). Infine, una ulteriore opportunità di appro-

fondimento e un nuovo "bisogno" vengono stimolati (la possibile visione di un film connesso con le due vicende). L'apparizione del terzo collegamento è rappresentata nella figura 5. Questa è una esemplificazione di ciò che potremmo definire "schema Big Data".

Un altro modo per descrivere questo schema è mediante il ricorso a tabelle, che per semplicità e comodità suddivideremo in due sezioni: la sezione "Small Data" in azzurro a sinistra, la sezione "Big Data" in giallo a destra. Il tipo di processing realizzato dal "find engine" consiste nell'individuare quali nuovi attributi i dati D1 e D2 sono in grado di rivelare e se ve ne siano di comuni o assimilabili (il campo "Olimpiadi 1972" evidenziato in giallo per tutti i dati). Nella figura è rappresentato il caso delle due immagini e del collegamento tra loro e con il terzo dato D3 (il film di Spielberg, anch'esso evidenziato in giallo nella sezione "Big Data" della tabella).

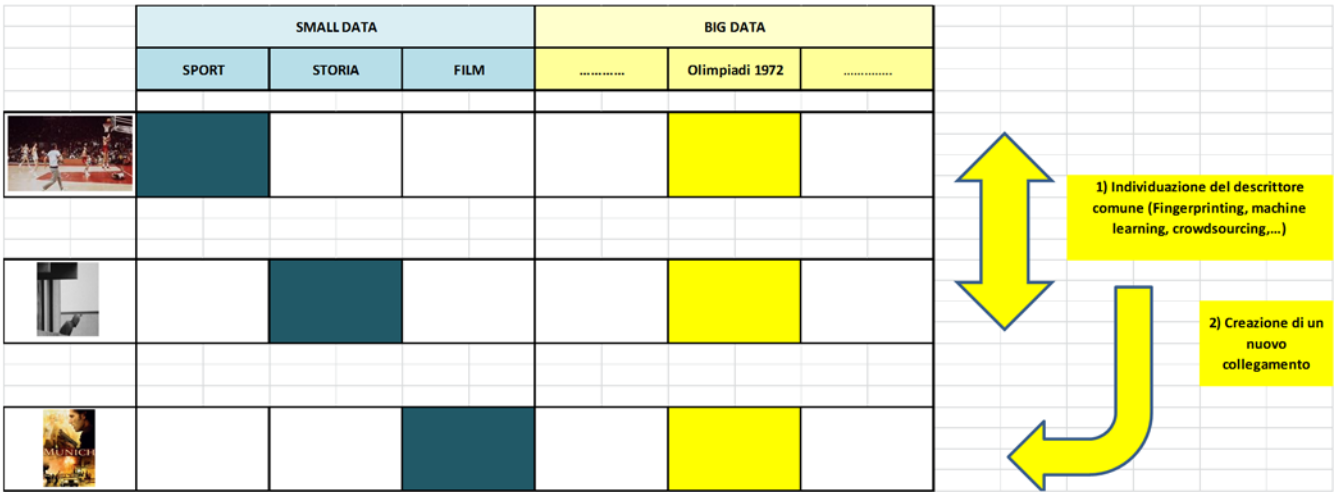


Figura 6. I trattamenti di dati in uno schema Big Data

Il motore di ricerca (nella sua veste di "search engine") effettua già oggi queste corrispondenze ancora timidamente, ossia limitandosi a proporre queste associazioni come una opzione e rimettendo a noi l'ultimo passo dell'esercizio di volontà che porta all'azione, ossia a "trovare" il collegamento. Presto (nella nuova veste di "find engine") potrà farlo più decisamente, grazie all'accresciuta conoscenza dei propri utenti acquisita con l'osservazione storica delle loro abitudini d'uso dei vari servizi, e più efficacemente, sia selezionando, a seguito di un'accurata creazione di profili, i soggetti a cui rivolgere queste proposte di associazione con più elevata probabilità di successo, sia con lo sviluppo di interfacce utente che non implicheranno discontinuità in ciò che chiamiamo "user experience" e che faranno percepire la proposta più come una opportunità che come un fastidio.

Per svolgere la fase di estrazione di descrittori da un dato, molte tecniche sono disponibili. Queste potranno essere impiegate anche in combinazione tra loro per arricchire la descrizione di un dato. Alcune sono di natura passiva, ossia possono essere svolte in modo automatico e senza intervento umano, altre invece, che coinvolgono la sfera dei "significati" da attribuire ai dati, sono evidentemente di natura attiva e non possono prescinderne. Rientrano nella categoria delle tecniche passive le seguenti metodologie di processing

- Hashing [*url1*], ossia la creazione di una o più impronte digitali univoche di un dato. Queste possono essere ottenute in modo svincolato dal significato del dato, mediante l'applicazione di tecniche crittografiche. Si presta ad essere impiegata anche a dati non testuali, come tracce audio, video, immagini
- Machine learning [*url2*], ovvero l'individuazione automatica delle categorie a cui il dato appartiene. Si distingue in "unsupervised machine learning" (o clustering), nel caso in cui le categorie non siano preidentificate e l'algoritmo individua a partire dall'analisi di tutti i dati un insieme di categorie di riferimento a cui associare ciascun dato, e "supervised machine learning", nel caso in cui siano disponibili le categorie e l'obiettivo è assegnare un nuovo dato alla categoria corrispondente. Idonea ad essere impiegata in un'ampia gamma di situazioni per dati testuali o multimediali
- Trasformata di Fourier [OS99], ovvero l'estrazione dello spettro di frequenze contenute nel dato. Particolarmente impiegato per dati multimediali quali video, audio e immagini, ma anche per dati numerici, quali serie storiche delle misurazioni raccolte su un fenomeno
- Concordanze e stilometria [*url3*], ossia la creazione di indicatori misurabili sul tipo di scrittura, applicabili a qualsiasi testo presente sul web

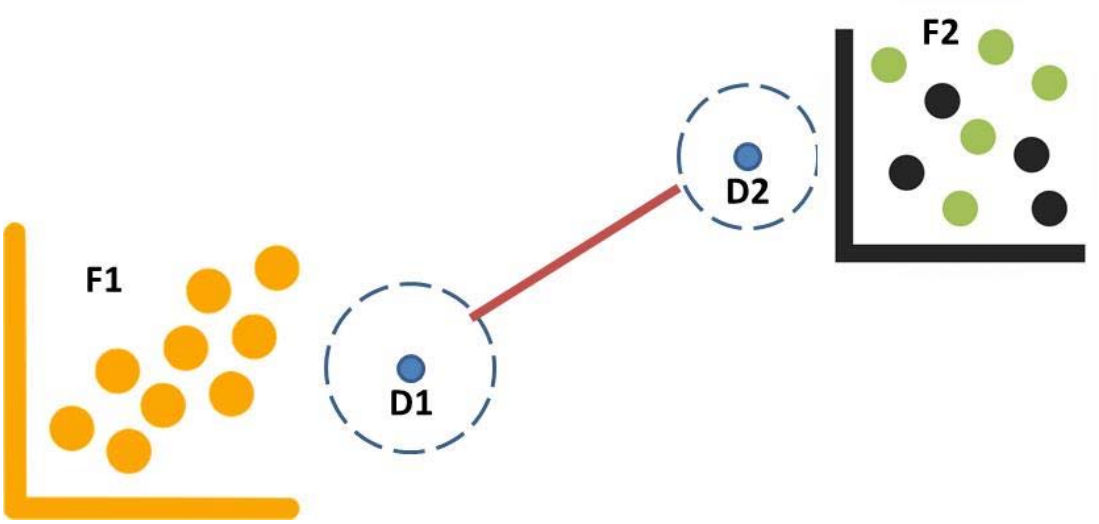
Tra le tecniche attive, invece, si possono citare

- Fingerprinting [*url4*] [*url5*] [*url6*], ossia la classificazione semantica di un contenuto per parole chiave (nell'esempio il dato D1 potrebbe avere come descrittori le parole chiave "basket", "URSS", "USA", "anni 70" e così via). Quanto più ricca è la descrizione del dato, tanto più esso sarà idoneo a essere messo in collegamento con altri dati per applicazioni "Big Data". Risulta molto efficace per classificare dati non testuali, quali brani musicali, video, immagini
- Crowdsourcing [*url7*], ovvero l'insieme di informazioni su un dato che sono fornite dagli stessi utilizzatori del dato. Ad esempio, sono molto efficaci l'estrazione delle parole chiave a partire dai commenti inseriti dagli utenti su un dato (commenti a video, notizie ecc.), ovvero l'analisi dei suggerimenti e feedback forniti sulle traduzioni automatiche, o ancora l'uso dei risultati dei test CAPTCHA [*url8*] che molti siti realizzano per verificare che l'utente che sta accedendo al servizio non sia un programma automatico

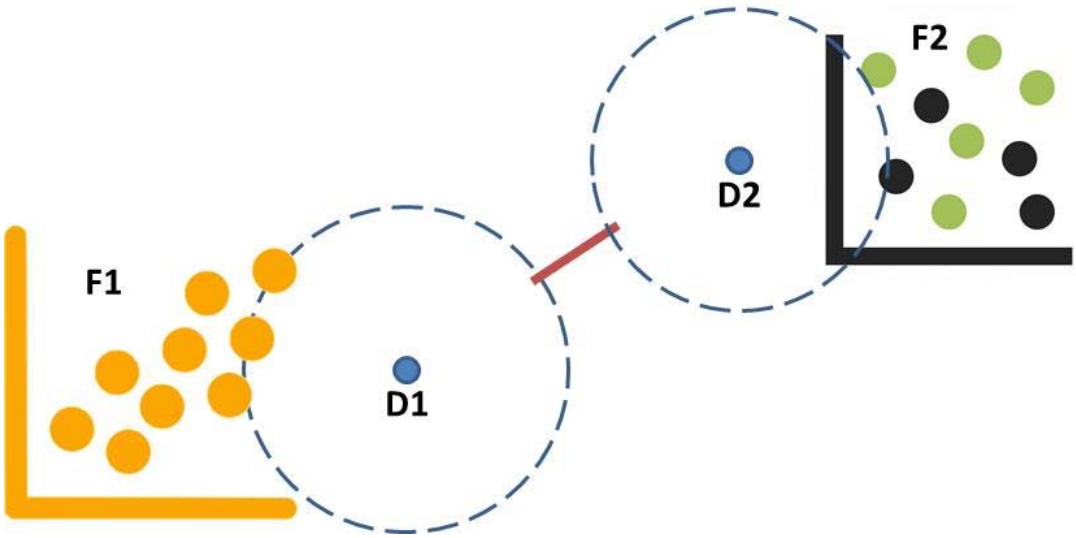
Queste le modalità disponibili per rendere i dati "più grandi". Nell'esempio di cui ci siamo serviti per introdurre lo "schema Big Data", abbiamo immaginato che esse potessero essere impiegate su dati già sufficientemente "grandi", in quanto riferibili a fatti noti (le vicende accadute nel corso dei giochi olimpici del 1972) e dunque facilmente correlabili l'uno all'altro. La prospettiva Big Data è che questo schema di emersione dei collegamenti possa essere *applicato a qualsiasi dato*, anche il più minuto e apparentemente privo di significato, in modo da creare una ragnatela incomparabilmente più fitta di come è il web attualmente, costruito, lo si ribadisce, su collegamenti tra i dati scelti unilateralmente da chi li immette in rete.

Natura non facit saltus, e dunque c'è da aspettarsi che il pieno dispiegamento delle potenzialità di questo schema di generazione di nuova conoscenza (una volta affrontate alcune questioni di cui si parlerà diffusamente più oltre) sarà frutto di una evoluzione più che di una rivoluzione. Eppure, ne vediamo già i primi esempi. Dal completamento automatico delle query di ricerca, alla disambiguazione dei risultati di una ricerca, dalla ricerca per immagini, alla possibilità di trovare "in diretta" il titolo di una canzone che ascoltiamo. Sono tutti risultati di questa accresciuta interconnessione tra dati, ottenuta in modo centralizzato mediante processing, che si "sovrappone" alla rete di relazioni tra dati decisa da chi li ha immessi per la prima volta su internet, in modo da ar-

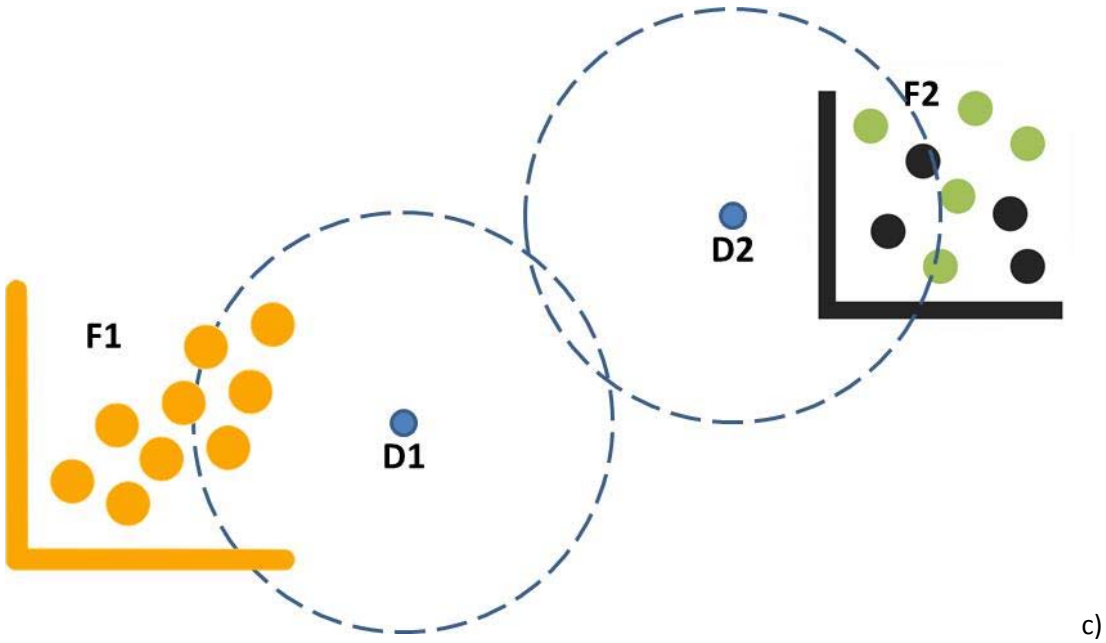
ricchirla e potenziarne le capacità di spiegare il mondo. Con uno sguardo prospettico (e, ovviamente, con qualche inevitabile problema legato alla difficolta di mettere a fuoco scenari futuri) possiamo ipotizzare che presto grazie ai Big Data, con l'aiuto della rete potremo sapere in quale luogo o in quale occasione è stata scattata una foto o chi è ritratto in un dipinto, associando questo nuovo dato a tutte le possibili relazioni con quel luogo, quell'evento o quel quadro, oppure ricevere una bibliografia-filmografia-discografia selezionata sull'argomento-film-brano musicale che ci interessa, o una guida turistica personalizzata dei luoghi che intendiamo visitare, o dei programmi di learning sugli argomenti più diversi. Inoltre, molte barriere ancora esistenti legate alle differenze linguistiche e culturali saranno superate, grazie all'impiego di strumenti di traduzione assistita sempre più sofisticati e precisi fino a raggiungere una piena corrispondenza semantica tra sorgente e traduzione (scritta o audio). Molte delle complessità che oggi affrontiamo (cambi di formato, o di mezzo) o inefficienze (tempi di ricerca) dovrebbero essere rimosse, rimettendo alla rete l'onere di cercare i collegamenti o trovare soluzioni, e l'uso di internet dovrebbe diventare molto più fluido. Tutto questo, senza voler considerare scenari che implichino l'uso di dati associati alle "cose", a cui sarà dedicato il prossimo capitolo.



a)



b)



c)

Figura 7. Big Data come fenomeno non lineare

Naturalmente, rispetto a questo scenario di impiego dei Big Data, molti casi intermedi possono presentarsi, quali quelli in cui più soggetti decidono di mettere in comune i loro dati con una specifica finalità. Ad esempio, un fornitore di servizi di mappe potrebbe, in aggiunta alla ricerca di un percorso, fornire all'utente una serie di offerte rese disponibili lungo il tragitto da diversi fornitori di beni o di servizi con cui ha stretto un accordo (per la vendita di titoli di viaggio, per soggiorni in hotel o servizi di ristorazione, per un coupon di sconti in una catena di negozi e così via). La presenza di una finalità non altera significativamente lo "schema" dei trattamenti, in quanto il fornitore di mappe nell'esempio ricercherebbe sempre corrispondenze tra dati (nell'esempio: l'ubicazione del negozio nelle vicinanze del percorso, la presenza di offerte che rientrano tra gli interessi dell'utente ecc.). La sostanziale differenza indotta dalla preesistenza di una finalità riguarda, invece, la qualità dei dati e il costo dell'interconnessione tra essi. Mentre infatti nell'esempio delle Olimpiadi, l'applicazione dello "schema" è tanto più efficace quanti più collegamenti è in grado di generare, magari senza troppo riguardo a valutazioni sulla qualità e pertinenza dei descrittori, nel caso in cui la finalità sia stabilita *a priori*, i soggetti che mettono in comune i loro dati dovranno farsi interamente carico della qualità dei dati conferiti, pena la mancata rilevanza dei collegamenti individuati e, in definitiva, il mancato conseguimento della stessa finalità. Inoltre, se in assenza di finalità ciò che rileva è il numero di collegamenti potenziali, allora il costo per generarli, con una delle tecniche richiamate, o sarà sostenuto da tecniche automatiche (hashing, machine learning, ecc.), ovvero distribuito sulla più ampia base possibile di utenti (crowdsourcing). Se invece la finalità è predeterminata, tutto il costo della descrizione dei dati dovrà essere sostenuto dai soggetti interconnessi, i quali dovranno sostenere preliminarmente il carico richiesto dalla definizione dei formati dei dati e dalla scelta della loro qualità. Uno scenario, come si intuisce, di scopo certamente modulabile in ragione della tipologia di soggetti interconnessi e dell'ambito in cui questi operano, ma comunque di portata più limitata.

Sull'efficacia dello "schema Big Data" applicato alla generalità dei dati, e in una così ampia varietà di casi, per scoprire corrispondenze tra fenomeni e per fornire una più compiuta descrizione del mondo, bisogna anche

considerare che siamo in presenza di un tipico fenomeno “non lineare” e che la calibrazione del potenziamento della capacità descrittiva dei dati, rappresentato dall’introduzione di nuovi attributi in ciascuno di essi, sarà un fattore determinante. Consideriamo il caso in figura 7 per averne una rappresentazione anche visiva. Immaginiamo di avere due fenomeni F1 e F2, che possono essere messi in relazione tra loro. Ipotizziamo che due dati D1 e D2 offrano la possibilità di collegare tra loro i fenomeni, ma che (in un contesto “Small Data” di partenza), la sfera di influenza dei dati D1 e D2 non sia tale da consentire l’applicazione dello “schema Big Data” (figura a). I dati D1 e D2 non consentono di collegare F1 e F2, ossia di ricomprendere i due fenomeni all’interno delle loro sfere di influenza, e il costo necessario per metterli in piena relazione (rappresentato dai tratti in rosso) è prevalentemente in capo al “ricercatore”. Ricorrendo ad un primo arricchimento dei descrittori dei due dati (figura b) allarghiamo significativamente la sfera d’influenza dei due dati, ma non al punto da consentire ancora di connettere i fenomeni. Il costo il costo per mettere i due fenomeni in relazione si riduce, ma non è annullato. A partire da questa ultima situazione, anche un piccolo ulteriore intervento sul potere descrittivo dei dati (figura c) consentirà di definire un’area ottenuta dall’unione delle due sfere attorno ai dati D1 e D2 che ricomprende entrambi i fenomeni F1 e F2 e ne ricostruisce interamente il nesso esistente. Ogni ulteriore accrescimento del potere descrittivo dei due dati diventa irrilevante per trovare il collegamento tra i fenomeni F1 e F2 e costituisce un puro costo per il “find engine”. Ne risulta un andamento qualitativamente rappresentabile dal grafico in figura 8, con una curva input-output (ovvero capacità di processing per l’arricchimento del dato vs. potere descrittivo dei dati) a incrementi marginali decrescenti, che potrà dare luogo nel tempo a forme di ottimizzazione dei profitti da parte del find engine che molta influenza avranno sull’applicazione di questo schema ai diversi casi concreti che si presenteranno (l’allocazione dei costi per l’arricchimento dei descrittori dei dati *in primis* rispetto ai benefici) e sul tipo di trama che presenterà un web “Big Data” (se totalmente e fittamente connesso, ovvero se più fittamente connesso, ma comunque per grandi “isole” tematiche o locali, ad esempio).

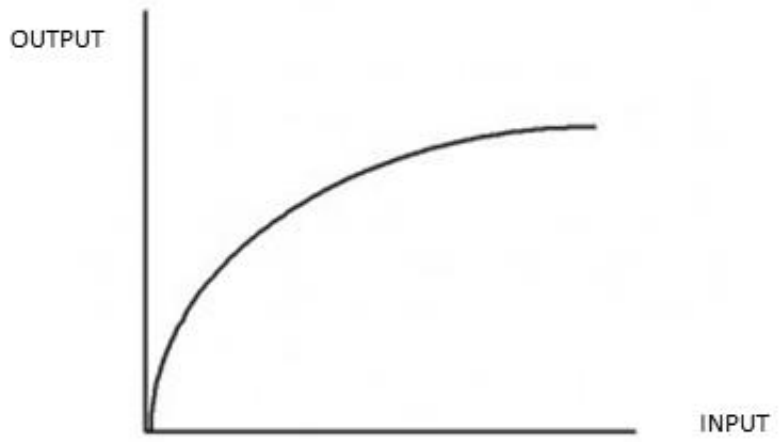


Figura 8. Curva input-output a incrementi marginali decrescenti

2. INTERNET E LE “COSE”

L’internet delle cose (*Internet of Things* o IoT) è una prospettiva tecnologica assai promettente per lo sviluppo di nuove applicazioni e servizi. Al momento, volendo riassumere le diverse proposte avanzate sulla IoT, abbiamo due schemi di riferimento: quello che considera la IoT come l’interconnessione, attraverso internet, di ogni sorta di oggetti e dispositivi, in modo da moltiplicare le opportunità di interlavoro tra di essi e creare così nuovi servizi e modelli di business [*url9*], e quello che vede nella IoT la concreta possibilità di una delocalizzazione della produzione delle "cose", che verrebbero riprodotte localmente mediante l’ausilio di printer 3D sempre più evolute, che si scambiano attraverso internet le informazioni su come costruire le “cose”, in modo da incrementare le opportunità di accesso a beni che sarebbero progettati, in definitiva, dagli stessi utilizzatori, e dunque seguendo criteri di massimizzazione di utilità e costo [R14].

Guardando alle “cose” in una prospettiva “Big Data”, occorre evidenziare come anche le “cose” siano in grado di generare dati potenzialmente utili a essere trattati secondo lo “schema Big Data”, e le relazioni tra le cose e tra i descrittori delle cose non sono meno interessanti da esplorare della relazione su cui ci siamo soffermati tra i dati e tra i descrittori dei dati. Non minori infatti sono gli stimoli a conoscere generati dalle “cose”, né gli interessi e bisogni che da questa conoscenza possono scaturire. Ogni dato infatti si riferisce a una “cosa”, e ogni “cosa” è portatrice di un significato: un particolare “stato” della “cosa”, la sua storia, l’uso che se ne può fare, l’insieme delle esperienze che hanno condotto alla sua realizzazione, l’insieme delle persone che la hanno realizzata, l’insieme delle altre “cose” di cui una “cosa” è composta, a loro volta portatrici di “stati”, storie, usi, esperienze, in un gioco di rimandi che può essere ripetuto potenzialmente all’infinito.

Le “cose”, in uno schema “Big Data”, diventano la porta di accesso ai significati di cui sono portatrici, rappresentabili per tramite di descrittori analoghi a quelli usati per i dati. Ciò rende le cose “trovabili” grazie allo stesso “schema Big Data” che, nell’esempio del capitolo 1, ha fatto emergere la connessione tra due immagini e successivamente l’ulteriore collegamento di entrambe con un terzo dato.

Possono valere per le “cose”, dunque, le stesse considerazioni fatte per l’arricchimento dei descrittori di dati. L’arricchimento delle “cose” potrà essere compiuto ancora una volta da un “find engine” in modo centralizzato e senza coordinamento tra le “cose” o tra i possessori delle “cose” impiegando le stesse tecnologie di processing attive e passive già introdotte. In una prospettiva “Big Data”, la quantità di “cose” e di significati da portare su internet è ancora enorme e gli spazi applicativi che si possono ipotizzare investono diversi ambiti sia economici, sia culturali. Ad esempio, la televisione non è ancora una “cosa” che sta su internet in “modalità Big Data”. Essa è riprodotta su internet (ossia, sono disponibili su internet tutti i canali delle televisioni), ma guardando un programma televisivo non è ancora possibile in modo fluido e senza dover “uscire” dal mezzo televisivo ed “entrare” in internet, “puntare” con la telecamera del proprio tablet lo schermo della TV e approfondire in tempo reale i contenuti che si stanno guardando: ad esempio, le storie dei protagonisti, gli altri contenuti collegati, i luoghi rappresentati, le musiche. Proseguendo con le esemplificazioni, i monumenti e le opere d’arte non sono “cose” che stanno su internet. Non è infatti possibile (salvo poche eccezioni) “puntare” con uno smartphone un monumento (magari classificato con un tag QR code) e apprenderne in modo fluido e immediato la storia. Né si riesce ancora a fare questa cosa (di nuovo, salvo poche eccezioni) con un prodotto enogastronomico, un libro o una pianta.

Sono tutte occasioni di conoscenza offerti dalle "cose", da cui potrebbero nascere nuovi interessi e bisogni e su cui è possibile articolare nuovi servizi. Le "cose" dotate di sensori, poi, potrebbero ancor più agevolmente dei dati generare descrittori di "stato" accurati (sul luogo, l'ora, le condizioni di funzionamento, l'ambiente circostante), prestandosi ad una standardizzazione di formati e significati, che renderebbe progressivamente più agevole la loro indicizzazione e l'azione del "find engine". L'uso combinato di queste tecnologie può consentire un accesso fluido a internet attraverso qualsiasi "cosa": la facciata di una chiesa o un dipinto di Caravaggio, il menu del ristorante in cui pranziamo o l'evento sportivo a cui stiamo assistendo allo stadio o in TV, in modo da aiutarci a trovare in ogni momento l'informazione che ci manca e da suscitare nuova conoscenza e nuovi interessi. Anche questo è Big Data.

Vista la varietà delle cose e le possibili relazioni che tra esse possono generarsi, fare previsioni oltre quanto si è detto su possibili nuovi scenari appare prematuro. Tuttavia, ciò che si può prevedere è che la modalità "Big Data" di fruizione delle "cose" avrà certamente impatti di natura sociale. Non è azzardato ipotizzare che per la descrizione in crowdsourcing di monumenti, luoghi, storie, ad esempio, un ruolo possa essere svolto da ampie fasce sociali che nel tempo richiederanno un progressivo avvicinamento o allontanamento dal mondo del lavoro, come studenti, giovani e anziani, a cui potranno essere offerte vere e proprie possibilità di lavoro nella descrizione delle "cose" a beneficio della collettività.

Senza fare ricorso alla fantasia nell'immaginare nuovi servizi e provando solo a ragionare su cosa potrebbe significare portare su internet le "cose" che adoperiamo comunemente, molte delle inefficienze che oggi patiamo (e alle quali siamo assuefatti, al punto da non accorgercene) potrebbero essere ridotte o annullate. Si pensi alla tempo di ricerca che ci è chiesto per imparare a usare gli oggetti: dover cercare un manuale è spesso una barriera all'adozione di nuove tecnologie. Barriera che potrebbe essere rimossa se la "cosa" non ci lasciasse da soli e ci accompagnasse al suo utilizzo. Anche la manutenzione degli oggetti ne beneficerebbe: un oggetto che è in grado di rivelare il suo "stato" di usura ci consentirebbe di approvvigionarci più efficientemente delle parti di ricambio, di richiedere assistenza o di assicurarci, in particolare se è lo stesso oggetto che attivamente innesca questo processo. Non si tratta di una sostituzione delle capacità umane di decidere, ma di un loro potenziamento, purché non ne risulti aumentata la complessità. Ma superare la complessità è problema tecnico e dunque risolvibile unilateralmente da parte di chi sviluppa le tecnologie. Esiste però una questione più complessa, legata al caso in cui i dati e le "cose" rivelino uno "stato" delle persone, che non può essere risolta unilateralmente e richiede l'interazione di molte parti. E questo ci porta al cuore del problema, che introduciamo nel prossimo capitolo.

3. Perché la privacy è importante

Trovare è bello. Bello scoprire le cose del mondo, meno piacevole che siano scoperte cose che ci riguardano o il codice per togliere l'antifurto di casa. Poiché ciascuno di noi può fare questo ragionamento, tutto lo scenario di benefici di cui abbiamo detto verrebbe a essere vanificato. Da queste semplici considerazioni, non è difficile comprendere la tensione che esiste tra sviluppo dei Big Data e privacy e su cui si articolerà nei prossimi anni il dibattito tecnologico e giuridico, alla ricerca di uno stato di equilibrio di lungo termine.

Anche l'esempio delle immagini e del film sulle olimpiadi, generatore di nuova conoscenza, se visto sotto la lente della privacy, desta molte questioni. Le due immagini caricate dai due uploader rivelano infatti caratteristiche di entrambi e ci si può chiedere a che titolo e sotto quali condizioni il find engine le tratti per derivarne

nuovi descrittori[1]. Inoltre, la proposta del film all'uploader 1 è pur sempre un'azione invasiva verso un utente, e anche qui ci si può chiedere cosa legittimi questa azione, che potrebbe, se non ben realizzata, essere percepita come una comunicazione non desiderata. In altri termini, nel rendere i dati "più grandi" e nel "trovare", che sono le due attività che caratterizzano il "find engine", non si può escludere che le "sfere di influenza dei dati" collidano le nostre "sfere private", ovvero implichino il trattamento dei nostri dati personali, o interessino dati relativi a "cose" che ci appartengono o che sono a noi intimamente legate. E, se ampliamo ancora lo sguardo, i dati e le "cose" in tal modo arricchite potrebbero consentire al find engine di trovare nuove relazioni tra dati e "cose" che investono persino la nostra salute o la sicurezza fisica dell'ambiente in cui viviamo, con ripercussioni rispetto alle quali una cautela ancor maggiore è ovviamente richiesta.

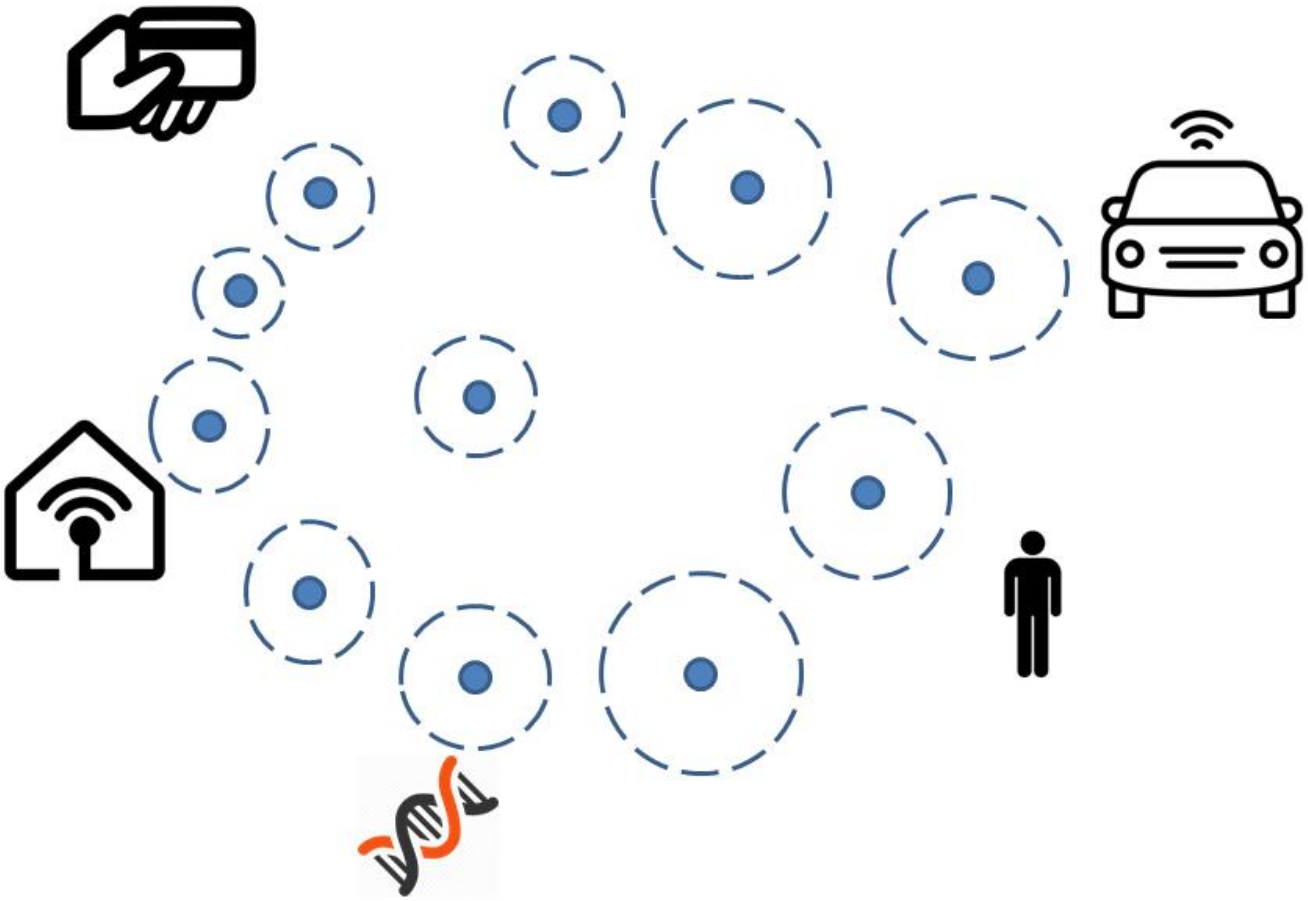


Figura 9. La separazione tra dati e "cose" in un contesto Small Data

[1] Un caso simile è già stato oggetto di una famosa sentenza della Corte di Giustizia Europea che ha valutato un interesse legittimo da parte di Google a trattare dati personali nell'ambito delle attività volte all'indicizzazione di contenuti presenti sul web [CGE14]

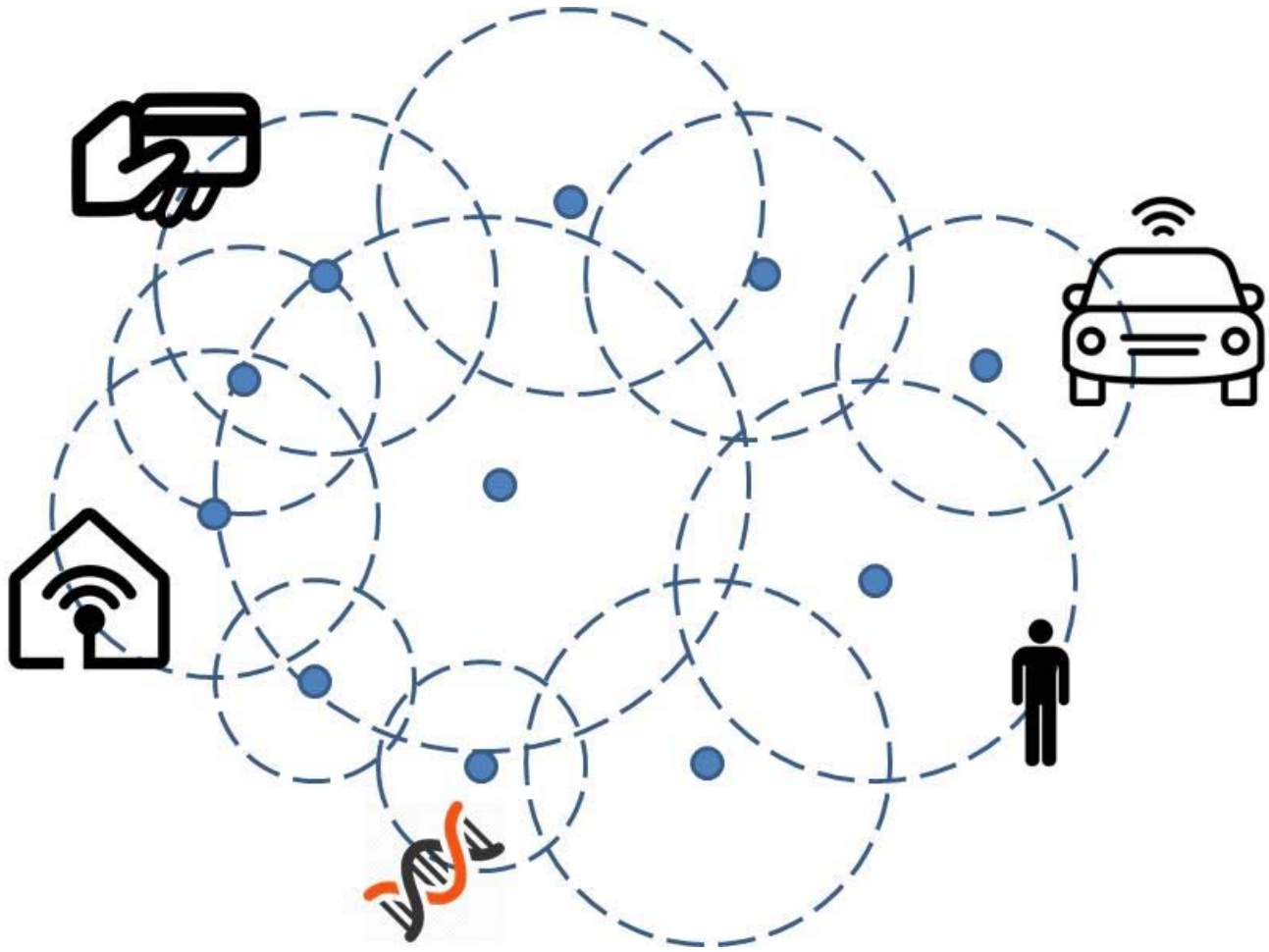


Figura 10. L'interconenssione tra dati e "cose" in un contesto Big Data

Affrontare e risolvere queste questioni è necessario per il successo stesso e la prosperità del modello Big Data, tanto che privacy e Big Data non devono apparire come due contendenti, ma come due alleati e garantire l'una è presupposto di stabilità per l'affermarsi dei benefici innegabili che deriveranno dall'altro [DDK15].

Proviamo a immaginare uno scenario in cui i Big Data dovessero "prevalere" sulla privacy, ovvero proviamo a immaginare una figura di find engine che sia in grado di portare alla luce ogni dato e ogni "cosa", in particolare ogni dato e ogni "cosa" a noi riferibile. Ne deriverebbe uno scenario non privo di inefficienze e inadatto a produrre benefici per le persone, e che metterebbe persino in discussione l'esistenza dello stesso "find engine".

Innanzitutto, questi meccanismi di arricchimento dei dati, che abbiamo denominato "schema Big Data", avvengono in un contesto di profonda asimmetria informativa bilaterale tra le parti, che se non considerato rischia di lacerare il rapporto tra utente (persona) e fornitore del servizio. Da una parte, infatti, il find engine, nel "trovare", decide associazioni tra dati e "cose" di cui le persone non sono consapevoli, dall'altro le persone potrebbero vedere esposti dati e "cose" la cui valenza semantica differisce da persona a persona ed è ignota al find engine. Un quadro conflittuale tra Big Data e privacy, nel quale la ragnatela di associazioni tra dati e "cose" venisse realizzata senza tener conto di queste asimmetrie (mitigando la prima e rispettando la seconda), ovvero in cui i Big Data dovessero "prevalere" sulla privacy, sarebbe soltanto generatore di sfiducia tra le parti. In uno scenario di conoscenza generalizzata e senza alcun filtro di ogni sorta di informazioni sul nostro conto, anche

quelle la cui valenza semantica ce le fa ritenere, invece, indisponibili per un osservatore esterno, possono emergere informazioni che ci espongono a conseguenze. Dunque se tutto fosse noto, alle persone verrebbe richiesto uno sforzo per difendere o ricostruire l'asimmetria informativa persa. In altri termini si genererebbe un ambiente più conflittuale. Sino ad oggi, questo genere di dinamiche hanno riguardato prevalentemente persone che hanno scelto di condurre una vita "in pubblico", e molti sono i casi di personalità pubbliche che hanno dovuto compiere sforzi e impiegare risorse per ripristinare le loro verità o mitigare distorsioni percepite sulle loro vite, non sempre peraltro con successo [*url1*0]. Solo raramente, fino a oggi, la rete ha "trovato" una persona annullando l'asimmetria informativa che questa aveva costruito tra sé e il mondo. Sono casi relativamente rari (anche se i numeri non sono trascurabili e sono in aumento) e quando si verificano, molte serie questioni si presentano per affrontarli. Problemi di motivazione personale, innanzitutto: non tutte le persone la cui asimmetria informativa con il mondo è violata sono infatti sufficientemente motivate, o dispongono delle necessarie risorse o competenze per provare a ripristinarla. Quindi, problemi di giurisdizione, in quanto il contesto tecnologico tende ad annullare le distanze e con esse la tradizionale categoria di legge applicabile. Infine, problemi legati all'efficacia delle misure: esistono infatti tipi di violazioni difficilmente ripristinabili (non si può infatti ordinare ad alcuno di dimenticare una vicenda che preferivamo restasse riservata). In uno scenario di Big Data "trionfante" sulle ragioni della privacy, ciascuno di noi avrebbe una ragione documentabile per essere considerato "buono" e una ragione documentabile per essere considerato "cattivo". Il giudizio finale di "buoni" o "cattivi" dipenderebbe anche dalla forza e dalla capacità di ciascuno di far prevalere le une sulle altre. In altri termini, molte relazioni che intratteniamo verrebbero portate innaturalmente su un piano ideologico e la forza richiesta per far propendere l'ago sul "buono" o sul "cattivo" rischierebbe di prevalere su ogni altra considerazione. Forse, persino una semplice operazione, come comprare un qualsiasi bene da un qualsiasi negoziante, verrebbe ad essere subordinata ad una valutazione "morale" su di esso in base al tipo di esposizione che le rete ne dovesse restituire [C15].

Proviamo, per ciò che è possibile intuire, a fare qualche esempio concreto di Big Data senza il filtro della privacy. Tutti avremmo un buon motivo (e documentato) per ricevere un trattamento differenziato. In situazioni del genere, il limite tra differenziazione e discriminazione potrebbe diventare molto sottile. Ad alcuni ad esempio potrebbe capitare di dover combattere per l'accesso negato a un bene o un servizio. Oppure, di soccombere davanti ad una prestazione negata per "non essere rientrati nei canoni", non avendo la forza di opporsi a questa valutazione [O10]. Altri potrebbero vedersi associato un prezzo per un bene in ragione del proprio "stato", dal quale potrebbe essere praticamente impossibile sottrarsi, vedendosi sottrarre per intero il proprio surplus economico, ossia il differenziale tra valore e prezzo che è, di nuovo, una forma di asimmetria informativa tra noi e il mercato per difendere il nostro potere economico [O09]. Altri potrebbero non riuscire a cambiare il fornitore di un servizio per via di una personalizzazione spinta dei beni o dei servizi che, di fatto, innescherebbe un lock-in senza via di uscita [*url1*1]. Altri ancora potrebbero entrare nel circolo delle persone dai gusti simili che sarebbero aiutate dalla rete a trovare sempre e soltanto i contenuti più vicini ai loro gusti, venendone quasi vincolati e perdendo per strada il piacere della scoperta casuale [P11]. Nella scoperta delle "cose" interconnesse attraverso internet, poi, a qualcuno potrebbe capitare di vedere esposto il codice che blocca il motore del modello della propria macchina, oppure che apre il modello delle porte di tutti gli hotel della catena alberghiera dove sta per trascorrere le vacanze [FJP16]. C'è chi ha addirittura ipotizzato che a partire dalla conoscenza della sequenza del DNA di ciascuno, che rischia di diventare un altro dato "trovabile" su di noi, potrebbero essere sintetizzati dei virus letali in laboratorio capaci di agire solo sulla persona a cui quel DNA appartiene, come la più perfetta delle armi che non lascia alcuna traccia ambientale [*url1*2]. Sono, alcuni di questi, scenari più da

science fiction che da sviluppo di policy rigorose per disciplinare lo sviluppo di una tecnologia, tuttavia essi hanno un pregio: ci fanno capire che dalla personalizzazione possono certamente venire benefici, ma che ognuno, se il processo che condurrà ai Big Data non sarà sufficientemente ponderato, potrà avere il suo fastidio personalizzato e che, in ragione del contesto, questo fastidio potrà diventare una mancata opportunità, una discriminazione, un danno se non addirittura di più.

Siamo certi che tutto questo ci piacerebbe? Appare improbabile che possa realizzarsi uno scenario Big data che lasci questi effetti indesiderati contro i soggetti verso cui intende dispiegare i benefici. Le componenti tecnologiche necessarie per ottenere il beneficio personalizzato sono le stesse capaci di realizzare il fastidio personalizzato. "Trovare" come risultato dell'azione di un mediatore è una esternalità, ossia un risultato al quale noi non contribuiamo o contribuiamo marginalmente, e se questa sarà una esternalità positiva, ossia una scoperta, o negativa, ossia una esposizione indesiderata, dipende unicamente dalla condotta del mediatore. Detto in altri termini, la conciliazione tra le prospettive di sviluppo offerte dai Big Data e il rispetto delle persone e delle loro naturali asimmetrie informative come forma di difesa verso il mondo si traduce in un tema di neutralità. Se in un contesto Big Data, devolveremo alla rete una parte della nostra facoltà di "trovare" dati e "cose", ogni transazione diventerà una transazione mediata e questa mediazione o sarà rispettosa delle nostre asimmetrie informative, ossia neutrale rispetto a esse, oppure si trasformerebbe in un tentativo forzoso di condizionarci nelle scelte, generatore di quel tipo di esternalità negative a cui accennavamo, e dunque non si realizzerebbe.

Cosa sia un atteggiamento neutrale è molto difficile da definire. Comprendiamo bene cosa è non neutrale, ma sulla effettiva neutralità di una scelta resta sempre una indeterminatezza. Se, da un ipotetico motore di ricerca che fornisca solo due risultati, digitando la query "Italia" dovessimo vederci restituiti come risultati la Torre Eiffel e l'Empire State Building, potremmo legittimamente arguire che qualcosa nel meccanismo di indicizzazione non funziona correttamente (o perché non è corretta l'informazione, o perché non è "corretto" il motore di ricerca). Ma se dovessimo vederci restituiti come primo risultato la Fontana di Trevi e come secondo risultato la Torre di Pisa, potremmo dire che questa scelta è neutrale? Forse, un albergatore di Roma sarebbe contento del risultato, certamente non uno di Venezia, che vedrebbe come effetto della scelta operata dal motore di ricerca un minore afflusso di turisti nella sua città. Nell'internet attuale (l'internet "Small Data"), il motore di ricerca è di fatto deresponsabilizzato da questa scelta [FTC13], in quanto i collegamenti tra i dati sono scelti dagli uploader e l'aspetto quantitativo che determina il posizionamento di un collegamento tra i risultati di una query è legato al numero di visite ai siti, che il motore di ricerca può solo misurare e non determinare. In un'internet Big Data, in cui il find engine avrà un ruolo più attivo nell'individuare connessioni tra i dati e le "cose", altri fattori determineranno le scelte, non solo in ordine al posizionamento di un risultato, ma sulla rilevanza finale per ciascuno del contenuto associato a un risultato. In altre parole, nel passare dal "cercare" al "trovare" non è più possibile invocare una piena deresponsabilizzazione del mediatore. Certamente parte della scelta sarà determinata dai dati stessi e dalle "cose". Quanto semanticamente più ricca sarà infatti la loro descrizione, tanto più netto sarà l'indirizzo che i dati e le cose forniranno al "ricercatore" sull'oggetto finale della sua ricerca (il dato o la "cosa" ultima da trovare). Quindi anche le cose avranno una loro "saggezza", una "wisdom of the things", da esprimere. Ma, più verosimilmente, al find engine sarà richiesta una maggiore responsabilità e una maggiore "wisdom of the mediator" rispetto a quanto accade oggi, in particolare se i dati e le "cose" sono riferibili alle persone e se il risultato della personalizzazione vuole essere un beneficio e non appartenere alla categoria delle esternalità negative.

Ma come asseverare questa la neutralità sull'utilizzo dei dati personali? Certamente, preoccupandosi di misurare in modo sempre più oggettivo [CDM14] la rilevanza dei dati e delle "cose" che vengono trovate per mezzo della rete. Ma, con un'ottica ex-ante, non vi è modo più efficace che lasciare che le persone si esprimano sul tipo di trattamento che esse consentono sui loro dati. E il modo più efficace che esiste per raggiungere questo obiettivo è rimettere il controllo sui dati alle persone che li hanno generati o a cui si riferiscono, come pure dei dati generati dalle "cose". Questo controllo si chiama privacy. In sua assenza, le persone tenderebbero naturalmente a creare in modo alternativo quella barriera di asimmetria informativa che le protegge dal mondo, ad esempio rinunciando a usare i servizi o, magari, disseminando il web di dati falsi (fake). È nell'interesse del find engine promuovere l'inclusione e evitare fake che distorcerebbero la realtà, facendoci perdere questa enorme opportunità di conoscenza.

Questa possibile deriva avrebbe un impatto enorme sulla qualità complessiva dei dati e sulla rilevanza dei risultati di una ricerca. Poiché lo strumento tecnologico è essenzialmente quantitativo, infatti, il find engine può "trovare" dati e "cose" unicamente come effetto di misure di correlazione, ovvero tramite stime di frequenze congiunte di fenomeni. Quanto più frequentemente due fenomeni si verificano insieme in un'osservazione, tanto maggiore è la probabilità che essi siano legati tra loro da una legge sottostante. Probabilità, però, non certezza. È infatti ben nota la differenza tra correlazione e causalità: l'una semplicemente osserva e misura l'occorrenza congiunta dei fenomeni, senza cercare spiegazioni né attribuire relazioni di causa-effetto, l'altra cerca invece di spiegarne il nesso distinguendo proprio chi sia causa e chi conseguenza[2]. Prendere l'esistenza di una correlazione tra due dati per una relazione causa-effetto, attribuendo ad un termine dalla relazione il ruolo di causa e all'altro quello di effetto, non sempre è un'operazione corretta e non pochi sono i casi in letteratura di fenomeni palesemente disgiunti eppure solo fortuitamente correlati [S54]. Questa creazione di collegamenti errati tra dati e "cose" diventerebbe molto evidente se dovesse realizzarsi la deriva verso una crescente presenza di dati distorti, volontariamente immessi dalle persone in rete come forma di difesa e segnale di sfiducia nei confronti di servizi ritenuti invasivi o dagli effetti addirittura pericolosi. Con la conseguenza che i dati o le "cose" trovate dalla rete sarebbero sempre meno rilevanti per il ricercatore, innescando in tal modo una reazione negativa capace di compromettere seriamente la prosperità del modello Big Data. Improbabile che possa realizzarsi uno scenario Big data che lasci questa indeterminatezza sulla qualità dei dati e che non risolva nel lungo termine le questioni legate alle responsabilità e neutralità del find engine sull'uso dei dati e la generazione di connessioni tra essi.

Infine, il più efficace "find engine" è un find engine invisibile e che rende invisibili, non caricandone sull'utenza il costo, tutti i passaggi intermedi dal "cercare" al "trovare", *in primis* l'arricchimento, l'indicizzazione e la classificazione, arrivando direttamente alla consegna del dato o della "cosa" che ci serve. Naturalmente, ciò pone problemi di conoscibilità del soggetto che ci offre il servizio. Non si può ipotizzare di non sapere chi "trova" dati e "cose" per noi. Ma pone anche problemi di remunerazione: come fare percepire ad un utente un servizio invisibile? Come fare per ricondurre l'azione del "trovare" e tutta la complessità qui soltanto accennata a un soggetto che è tanto più efficiente quanto meno si mostra. Il modello di business del tradizionale motore di ricerca si fonda sulla discontinuità: la presentazione dei risultati di una query, la rivendita di parole chiave, la presenza visibile di spazi pubblicitari sono fasi distinte dalla presentazione del contenuto indicizzato. Siamo ancora distanti dall'avere individuato un modello di remunerazione per il find engine. Tuttavia per la sua esistenza economica è necessario che il suo "marchio" e la sua azione siano riconoscibili. La trasparenza sulla titolarità di un

[2] Forse l'evoluzione rispetto allo schema Big Data?

trattamento che è un caposaldo della privacy, diventa perciò nello "schema Big Data" un fattore determinante anche come strumento per stare sul mercato, un elemento a favore della prosperità stessa del find engine.

La privacy come vediamo non è dunque soltanto importante, è necessaria.

4. I principi della protezione dei dati personali

Nell'effettuare l'arricchimento nella descrizione dei dati e delle "cose" che ha luogo nello "schema Big Data", il find engine dovrà dunque attenersi a una serie di principi ben codificati dalle leggi vigenti in ambito comunitario, la Direttiva 46/95, e nazionale, il Codice in materia di protezione dei dati personali, e ulteriormente rafforzati dal nuovo Regolamento europeo in materia di protezione dei dati personali, che le sostituirà e che sarà definitivamente applicabile in via diretta in tutti i Paesi UE a partire dal 2018. Per la finalità di questo testo sarà sufficiente una sintesi di tali principi, rimandando il lettore interessato a pubblicazioni che approfondiscono il quadro storico e giuridico che ha portato a questo importante risultato [P16].

Come detto, è nell'interesse di tutti che questi principi siano conosciuti e applicati per la stessa prosperità del modello Big Data. Innanzitutto, è bene precisare che la nozione di dato personale è piuttosto ampia e ricomprende sia dati direttamente identificativi (il nome, l'indirizzo, il codice fiscale e così via), sia dati indirettamente identificativi, ossia riconducibili a dati direttamente identificativi attraverso codici o tabelle di passaggio intermedie. Oggi, è ormai interpretazione universalmente accettata [WP2914] considerare che si è in presenza di un trattamento di dati personali ogni volta che è possibile isolare un soggetto in un database, ovvero se è possibile collegare il dato che si sta trattando a dati relativi allo stesso soggetto presenti in diversi database, o ancora se è possibile dedurre, con probabilità significativa, una caratteristica di quel soggetto dal trattamento di un dato, per prendere decisioni che riguardano una specifica persona, anche senza che sia nota la sua identità.

Un altro concetto fondamentale per l'applicazione del quadro giuridico dell'UE in materia di protezione dei dati è quello di titolare del trattamento, o controller. Un controller è l'entità che determina le finalità e gli strumenti impiegati per il trattamento di dati personali. A causa di queste responsabilità, i titolari del trattamento hanno obblighi specifici. Innanzitutto, l'obbligo di rispettare i principi relativi alla qualità dei dati: i dati personali devono infatti essere trattati in modo corretto, indicando con ciò che essi non soltanto devono fornire una rappresentazione corretta della persona (ossia devono essere aggiornati e non contenere errori), ma che non dovrebbero mai essere trattati senza che l'individuo ne sia realmente consapevole. Inoltre, il rispetto del principio di finalità, il quale implica che i dati personali possono essere trattati solo per finalità determinate, esplicite e definite prima che il trattamento dei dati avviene. Ogni ulteriore scopo incompatibile con tali finalità originali è illecito ai sensi del diritto dell'Unione. Poi, i dati trattati devono essere rigorosamente quelli che sono necessari per lo scopo specifico perseguito dal titolare del trattamento. Questo è il principio di necessità. Inoltre, i dati personali possono essere impiegati dal titolare per un periodo di tempo necessario per lo scopo per il quale sono stati trattati, al termine del quale non devono essere più disponibili.

Non basta la definizione della finalità per effettuare un trattamento. Per potere essere trattati dal titolare, i dati personali necessitano di un presupposto di legittimità del trattamento. In pratica, le basi giuridiche che possono essere rilevanti nel contesto dei Big Data sono il consenso, l'adempimento di obblighi contrattuali e l'interesse legittimo del titolare. Il consenso è lo strumento di controllo più efficace nella disponibilità delle persone. Per essere valido, esso deve essere dato liberamente (la persona interessata deve avere la possibilità di accettare o rifiutare il trattamento dei suoi dati personali), informato (la persona interessata deve avere le

informazioni necessarie sul trattamento, in modo da potersi formare un giudizio preciso), specifico (l'espressione della volontà deve riguardare gli scopi per cui i dati sono trattati) e inequivoco (è necessaria un'azione positiva che segnali senza ambiguità la volontà della persona interessata prima che il trattamento abbia luogo). In taluni casi, il trattamento è legittimo se è necessario per l'esecuzione di un contratto, nella misura in cui vi è un legame diretto e oggettivo tra il trattamento e le finalità delle prestazioni contrattuali. In terzo luogo, il trattamento di dati personali è consentito se necessario per il perseguimento di un interesse legittimo del titolare o di una terza parte, purché tali interessi non prevalgano sui diritti fondamentali della persona interessata (ossia se ne consentono l'esercizio).

I dati personali non possono essere estorti all'interessato con un accesso forzato ai suoi dispositivi. È infatti necessario che la persona interessata acconsenta all'accesso ai dati memorizzati nel proprio dispositivo, a meno che tali dati non siano strettamente necessari al fine di fornire un servizio esplicitamente richiesto.

Esistono poi specifici obblighi di trasparenza per i titolari. Trasparenza sulla propria identità, sulle finalità del trattamento, sulla presenza di eventuali ulteriori destinatari dei dati, sulla sussistenza dei diritti di accesso ai dati (per chiederne la cancellazione o l'aggiornamento, ad esempio) e sul diritto di opporsi al trattamento. La disponibilità e la chiarezza di queste informazioni è un prerequisito per la validità del consenso e, come si è detto, è uno strumento molto importante per tener conto delle asimmetria che possono verificarsi nel rapporto tra utenti e fornitori di servizi. Ogni titolare è inoltre pienamente responsabile per la sicurezza dei dati, che si concretizza nella predisposizione di adeguate misure tecniche e organizzative per proteggere i dati personali quali, in particolare, la realizzazione di controlli per limitare l'accesso ai dati.

A questi principi "tradizionali" si aggiungono con il nuovo Regolamento nuovi diritti per l'interessato. Il diritto alla portabilità dei dati, ad esempio, introdotto con lo scopo di rafforzare il diritto di accesso degli utenti. L'interessato può infatti ricevere i dati personali forniti a un titolare in un formato strutturato, comunemente usato e *machine readable*, e ha il diritto di trasmettere tali dati ad un altro titolare. Il diritto a poter essere dimenticati (*right to be forgotten*), che con le limitazioni a cui si accennava (non si può obbligare nessuno a dimenticare), si pone l'obiettivo di creare un atteggiamento più responsabile e cauto nei casi di comunicazione di dati da un titolare a terze parti. In particolare, questo nuovo diritto prevede l'obbligo per il controller che ha comunicato dati a terzi di prendere ogni misura necessaria per informarli della richiesta di cancellazione da parte di un interessato al fine di pervenire alla cancellazione delle copie presso questi ultimi.

Il Regolamento introduce anche un nuovo strumento di sensibilizzazione per i titolari del trattamento, che dovrebbe facilitare il rispetto degli obblighi in materia di protezione dei dati, la valutazione d'impatto (o Data Protection Impact Assessment). Questa valutazione deve contenere una descrizione dei trattamenti previsti, una valutazione dei rischi per la privacy e le misure previste per affrontare tali rischi. È interessante notare anche come siano introdotti nuovi meccanismi di co-titolarità (co-controllership), proprio con lo scopo di affrontare le complessità della catena del valore dei Big Data. Nel caso in cui due o più controller determinino congiuntamente le finalità e gli strumenti per il trattamento, infatti, uno soltanto di essi potrà agire come punto di contatto per l'esercizio dei diritti, mentre le rispettive responsabilità in materia di privacy potranno essere e stabilite in modo trasparente tra i diversi titolari.

5. Le nuove sfide per l'applicazione dei principi

Se il presupposto è la necessità della protezione di dati personali per la stessa prosperità dei Big Data, la sfida che si porrà è come dare applicazione a questi principi, e non già se essi siano in grado di resistere all'onda d'urto dei Big Data. Si tratta di una sfida, indotta dai cambiamenti dello scenario tecnologico, che potrà richiedere anche decisi cambi nel modo in cui negli anni ne è stata data attuazione.

In un contesto *data intensive* come quello dei Big Data, la complessità dei trattamenti è certamente destinata a crescere. Rendere le persone pienamente consapevoli di ogni trattamento in corso sui loro dati sarà anch'essa un'operazione complessa. L'approccio puramente quantitativo e testuale con cui le informative sono state pensate sin dalle origini non è idoneo allo scenario che abbiamo descritto. La necessità di trasparenza dovrà essere perseguita con modalità che puntino all'attenzione della persona, molto più di quanto sia capace di fare un testo scritto, in particolare se al testo si richiede completezza. Ciò richiederà una vera e propria progettazione dell'informazione che si articoli su diversi canali (testo, video, audio) e che tenga conto del contesto in cui può trovarsi la persona (la disponibilità di tempo, il terminale impiegato, la fase del servizio in cui si trova e l'eventuale presenza di nuovi rischi), in modo da catturarne sempre l'attenzione, offrendo in ogni momento opportunità di approfondimento [SDB15].

Occorre poi osservare che sul possibile arricchimento del potere descrittivo di un dato, anche personale, che è alla base dello "schema Big Data", questo non deve essere sempre considerato come una nuova finalità incompatibile con quella originale per cui il dato è stato immesso in rete. Piuttosto che imporre un requisito di compatibilità, infatti, il quadro giuridico vigente e futuro vieta l'incompatibilità tra finalità, lasciando uno spazio di flessibilità e manovra su cui è bene riflettere per l'impiego di dati per finalità "non incompatibili". Quindi, al fine di valorizzare le notevoli possibilità che i Big Data offrono, sarà essenziale valutare accuratamente le finalità ulteriori, rimuovendo ogni arbitrio o ambiguità. In questo esercizio, aspetti come i tempi del trattamento e l'impatto sulle persone interessate dovranno essere considerati, per comprendere se per ogni nuova finalità siamo in presenza di scopi realmente distinti e incompatibili o, per esempio, di due diverse fasi temporali dello stesso trattamento dei dati, ovvero di una nuova finalità non incompatibile con la precedente.

Anche l'istituto del consenso richiederà nuove riflessioni. Una volta definiti i principi, non vi è alcuna specifica prescrizione nel quadro normativo su come l'azione positiva che significa il consenso debba configurarsi. Si tratta di un problema di neutralità della tecnologia, ed è solo una questione di creatività. Gli spazi a disposizione dell'industria per soluzioni *user friendly* sono ampi, tenuto conto che sempre più i dispositivi saranno dotati di una varietà di sensori (accelerometri, giroscopi, telecamere, microfoni, funzioni di rilevamento touch, software di rilevamento del movimento, ecc) e applicazioni per l'elaborazione dei segnali, capaci di raccogliere ogni precisa azione dell'utente e di interpretarla in modo non ambiguo come un consenso libero e specifico, senza interrompere la "user experience" nella fruizione di un servizio.

Lo stesso può dirsi per l'esercizio degli altri diritti. Più in generale, per ciò che l'evoluzione delle tecnologie consente di prevedere, anche per i casi in cui il consenso non sarà applicabile, molte efficaci possibilità saranno disponibili alle persone interessate per l'esercizio dei loro diritti (di accesso, di opposizione, di portabilità ecc.). La tecnologia, la stessa tecnologia che consente il dispiegamento delle potenzialità dei Big Data, offre notevoli strumenti anche per un esercizio dei diritti a prova dei tempi. Non ci potrà essere una tecnologia di "nuova generazione" applicata ai Big Data e una di "vecchia generazione" per la privacy.

Approfondimento 1

Sull'inefficacia di un approccio alla conoscenza puramente quantitativo, è il caso di menzionare l'arguto paradosso sul metodo induttivo elaborato dal logico Carl Gustav Hempel negli anni '40, noto come paradosso del corvo nero. Egli osservava che per cercare la verità dell'affermazione «tutti i corvi sono neri», per logica, esistono due modi equivalenti di procedere: o osservare tutti i corvi, constatando che tutti sono neri, o osservare tutti gli oggetti non-neri, constatando che tra essi non vi è nessun corvo.

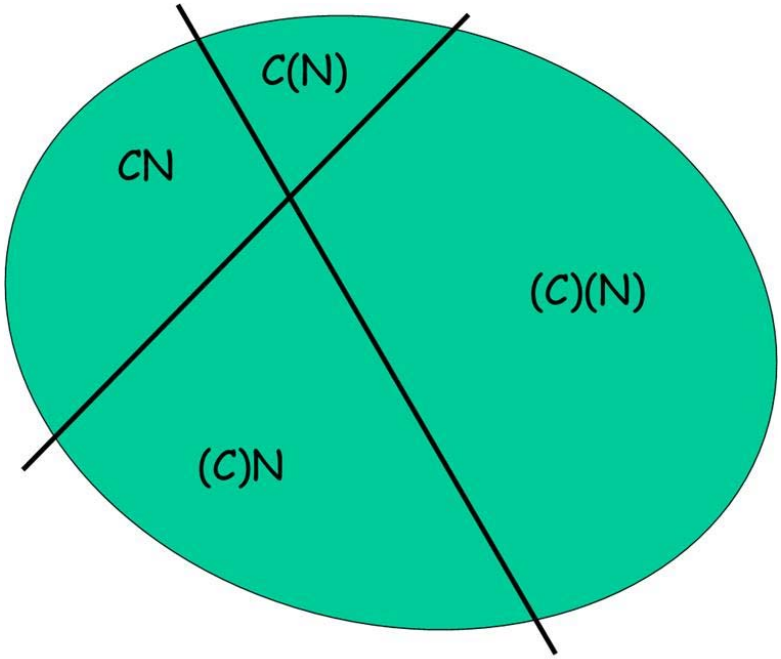


Figura a1 Si osservi la ripartizione dell'universo in oggetti che sono Corvi C e oggetti che non sono Corvi (C) - in parentesi la negazione dell'attributo – e in oggetti che sono Neri N e oggetti che non sono Neri (N), suddivisi nelle quattro classi CN, C(N), (C)N, (C)(N).

La relativa scarsità dei corvi potrebbe farci propendere per ricercare una conferma della legge generale «tutti i corvi sono neri» nell'osservazione più semplice e immediata che una mela rossa non è un corvo, né lo è un limone giallo, né nessun altro oggetto non nero. Nel seguire questa "scorciatoia" paradossale, in realtà, non considereremmo che il numero di oggetti non neri è molto più elevato del numero di corvi, e l'osservazione di una mela rossa aumenta molto meno il nostro grado di conferma della legge generale «tutti i corvi sono neri» rispetto all'osservazione di un corvo nero. Seguire unicamente l'approccio quantitativo (più informazione immediatamente disponibile), in definitiva, può allontanare dalla conoscenza della legge generale. L'efficienza, in questo caso, dovrebbe invece essere perseguita nel ridurre i tempi di ricerca dei corvi.

6. Cosa è privacy by design

Trovare le persone attraverso internet nella maniera sin qui descritta, ossia attraverso collegamenti tra i dati, oggi è ancora un'eccezione. Non ci sono noti i dati personali e le vicende delle persone che incontriamo casualmente per strada, né sono trovabili le cose che appartengono alle persone per accedervi e controllarle. Quando fenomeni simili si verificano, fortuitamente o come effetto di azioni deliberate, essi destano molto clamore, tanto da richiedere l'intervento delle Autorità di protezione dei dati e, come è successo, della Corte di Giustizia Europea, che devono affrontare questioni molto complesse che, lo si è detto, riguardano aspetti di giurisdizione ed efficacia dei rimedi. Nello "schema Big Data", trovare sarà la norma e non tutti i casi in cui dovessero essere coinvolti dati personali potranno essere affrontati ex-post con una sentenza della Corte di Giustizia. Il nuovo Regolamento offre una prospettiva molto pragmatica per affrontare ex-ante questo rischio, introducendo il principio di privacy by design, che obbliga il titolare, nel momento in cui determina le finalità e le modalità del trattamento, tenuto conto dei rischi del trattamento, dello stato dell'arte delle tecnologie e dell'ambito di applicazione, di mettere in atto misure tecniche e organizzative adeguate per integrare nel trattamento le necessarie garanzie per tutelare i diritti degli interessati.

Integrare le tutele nel trattamento. È questo il passaggio chiave, al quale due significati possono essere attribuiti: o trattare i dati in modo da minimizzarne l'uso per perseguire una finalità, al punto da non ritenersi più necessario un trattamento di dati personali, ovvero trattarli in modo da incrementarne la sicurezza, rafforzando la confidenzialità del dato. Sono due distinte forme di tutela integrate nel trattamento, che richiedono un approfondimento, in ragione delle conseguenze significativamente diverse a cui la loro applicazione conduce.

6.1. Il concetto di anonimizzazione e di dato anonimizzato

Può essere a questo punto utile chiedersi se sia possibile da un punto di vista giuridico e tecnico effettuare trattamenti su dati "minimizzati" fino al punto da non consentire l'identificazione diretta o indiretta di una persona. In altri termini, se sia possibile ipotizzare, e concretamente effettuare, trattamenti su dati anonimizzati, come sia possibile pervenire a dati anonimizzati, e quale sia per la persona il tipo di tutela integrata nel trattamento.

A questo proposito, tanto la Direttiva 95/46 quanto il Regolamento, nel riferirsi all'anonimizzazione, sono concordi nell'offrirci una linea guida di lungo termine, asserendo che per determinare se una persona è identificabile, è opportuno prendere in considerazione l'insieme dei mezzi che possono essere ragionevolmente utilizzati dal titolare o da altri per identificare una persona e che i principi della tutela non si applicano a dati resi anonimi in modo tale che la persona interessata non è più identificabile.

Partiamo da queste premesse e proviamo a enucleare alcuni aspetti chiave che ci aiutano a costruire una definizione concettuale di anonimizzazione. Innanzitutto, occorre osservare che le considerazioni sull'effetto dell'anonimizzazione, ovvero la non identificabilità della persona in relazione al dato anonimizzato, non possono essere disgiunte dalla valutazione dei mezzi nella disponibilità di chi (titolare o altro soggetto) provi a utilizzare il dato anonimizzato per identificare la persona. A proposito dei mezzi, poi, è bene osservare che, volendo seguire le indicazioni della Direttiva e del Regolamento, viene a determinarsi ripartizione dei mezzi in due classi; i mezzi "ragionevolmente utilizzabili" e, dobbiamo ritenere, quelli "irragionevolmente utilizzabili". In tal modo, se disponendo del dato anonimizzato e soltanto con l'ausilio di mezzi "irragionevolmente utilizzabili" è possibile identificare la persona, il dato anonimizzato non rientra nell'ambito di applicazione della legge. Natural-

mente, quanto minore sarà il numero di elementi potenzialmente identificativi presenti nel dato anonimizzato, tanto maggiore sarà lo sforzo necessario all'identificazione della persona per chi utilizza quel dato e conseguentemente tanto più esiguo l'insieme dei mezzi idonei allo scopo, al punto che volendo procedere in questa "sottrazione" si raggiungerà un momento in cui l'insieme dei mezzi idonei a disvelare l'identità della persona diventa "irragionevolmente utilizzabile". In questa costruzione è dunque di tutta evidenza l'importanza della metodologia di "sottrazione": essa deve essere tale da rendere l'impiego di qualsiasi mezzo –utilizzabile in combinazione con il dato anonimizzato, per fini identificativi – una opzione inutile o irragionevole.

Così concettualmente posto, il processo di anonimizzazione ha in definitiva l'obiettivo di pervenire ad una nuova rappresentazione del dato, il dato anonimizzato, che, alle condizioni sopra richiamate, non rientra nell'ambito di applicazione della direttiva. L'anonimizzazione costituisce, in altri termini, un trattamento successivo di dati personali rispetto a quello effettuato per la finalità originaria perseguita dal titolare, per il quale deve ovviamente sussistere una idonea base giuridica (consenso, adempimento di un obbligo contrattuale, legittimo interesse). In relazione al fine ulteriore (di pervenire, cioè, ad una rappresentazione anonimizzata dei dati), perché non si sconfini nell'insieme dei trattamenti incompatibili, non consentiti dal quadro normativo, questo dovrà soddisfare il requisito di compatibilità [WP2913] che deve tener conto, in particolare, del rischio di impatti indesiderati sull'interessato derivanti dal trattamento ulteriore e delle salvaguardie poste in essere dal titolare per scongiurarne l'evenienza. Ma, più concretamente, quando si scongiura ogni impatto sulla persona all'esito del processo di anonimizzazione? L'orientamento dei Garanti europei è su questo punto unanime [WP2914]: un efficace processo di anonimizzazione scongiura impatti sulla persona se è in grado di impedire a chiunque impieghi un insieme di dati anonimizzati, in combinazione con i mezzi "ragionevolmente utilizzabili" di cui può disporre, 1) di isolare una persona in un gruppo, 2) di collegare un dato anonimizzato a dati riferibili ad una persona presenti in un distinto insieme di dati e 3) di dedurre da un dato anonimizzato nuove informazioni riferibili a una persona.

Su un piano più operativo, questi obiettivi possono essere perseguiti mediante l'applicazione, anche congiunta, di diverse tecniche di anonimizzazione (su cui ci soffermeremo nel seguito) raggruppabili in due categorie: la *distorsione* (o *randomizzazione*) e la *generalizzazione* dei dati, entrambe concepite per introdurre un grado di incertezza, misurabile in termini probabilistici, sull'attribuzione di un dato anonimizzato a un soggetto determinato. La distorsione è una famiglia di tecniche che modifica la veridicità dei dati al fine di eliminare, ove possibile, il legame che esiste tra il dato puntuale e la persona. Se, infatti, i dati sono resi sufficientemente incerti, ad esempio mediante l'aggiunta di "rumore" statistico ai loro valori, ovvero operandone una differente attribuzione casuale ai diversi interessati cui si riferiscono, essi possono non più essere riferiti a una persona specifica, a tal punto da trasferire in taluni casi questa incertezza persino sulla stessa presenza di un dato riferibile ad un interessato all'interno di un database [D06]. La generalizzazione rappresenta la seconda famiglia di tecniche di anonimizzazione e consiste nel diluire gli attributi, ossia gli elementi costitutivi dei dati delle persone interessate, modificandone la scala o ordine di grandezza (vale a dire, una regione anziché una città, un mese anziché una settimana, ad esempio). L'incertezza in questo caso è legata al fatto che quanto più lasca è la scala dei valori degli attributi, tanto maggiore è il numero di interessati potenzialmente riferibili a un certo attributo "generalizzato", in modo da rendere via via meno probabile l'attribuzione del dato alla persona.

Naturalmente, per l'impiego di entrambe le tecniche, si pone un problema di utilità del dato all'esito del processo di anonimizzazione. Nel caso delle tecniche di distorsione, se il rumore prevale rispetto al dato utile, questo diventa, oltre che incerto (ossia non riferibile ad alcuno), inaccurato e inidoneo a qualsiasi tipo di analisi. È

dunque necessaria una calibrazione della distorsione, in ragione dell'uso che si vorrà fare del dato anonimizzato. Nel caso del ricorso a tecniche di generalizzazione, se la scala è troppo lasca, il dato rischia di perdere ogni valenza semantica, diventando inidoneo a esprimere qualsiasi nesso di causalità utile a descrivere un fenomeno.

Con riferimento all'impiego di queste tecniche occorre inoltre precisare che, in linea di principio, disponendo di un dato anonimizzato non può essere mai scongiurato il rischio che esso sia arbitrariamente associato ad una persona. Tuttavia, se il processo di anonimizzazione è correttamente applicato, la verosimiglianza di tale attribuzione è del tutto assimilabile a quella di una attribuzione casuale effettuabile anche in assenza del dato anonimizzato, e se una decisione viene presa su quella persona in base a tale attribuzione, quest'ultima dovrà essere considerata alla stregua di un evento in alcun modo riconducibile alle caratteristiche del dato ottenuto tramite il processo di anonimizzazione.

Si è detto come la valutazione dell'efficacia del processo di anonimizzazione debba comportare anche una considerazione dei mezzi. Questi potranno di volta in volta consistere in mezzi economici, informazioni, risorse tecnologiche, competenze nonché tempo disponibile, e ogni valutazione se essi siano in effetti "ragionevolmente utilizzabili" non potrà non tener conto di elementi soggettivi, che possono variare in ragione del contesto. La disponibilità di certe risorse, che può essere "irragionevole" per taluni, potrà infatti non esserlo per altri soggetti. Dunque questa valutazione, che è componente necessaria del test di compatibilità della finalità ulteriore perseguita da ogni processo di anonimizzazione, dovrà essere effettuata caso per caso. Tra gli aspetti da considerare in questo esercizio vi dovrà essere primariamente il livello di motivazione di eventuali soggetti interessati ad associare il dato anonimizzato ad una persona. Inoltre, si dovrà tener conto della natura dei dati originali e della riferibilità dei dati a specifiche tipologie di interessati, che per questa stessa caratteristica possono essere più facilmente identificabili. Ulteriori elementi da tenere presenti dovranno essere l'applicazione, da parte del titolare che effettua l'anonimizzazione, di idonee misure di sicurezza, o di vincoli contrattuali che possono limitare la "visibilità" dei dati anonimizzati, ad esempio a soli utilizzatori in possesso di specifiche credenziali di accesso e sulla base di riconosciute esigenze a conoscere il dato anonimizzato. Anche la trasparenza sulle tecniche di anonimizzazione adottate costituisce un importante elemento di valutazione, così come, se del caso, la metodologia di campionamento impiegata.

Va sottolineato come un fattore in grado di compromettere significativamente le tutele introdotte con l'anonimizzazione sia la disponibilità di dati ausiliari riferibili ad una persona a cui collegare il dato anonimizzato. Poiché la quantità di informazioni, anche pubblicamente disponibili, è destinata a crescere nel tempo, un mezzo oggi valutato irragionevole, in considerazione dell'informazione ausiliaria attualmente disponibile, potrà non essere giudicato tale in successive valutazioni, anche tenuto conto dell'evoluzione delle tecnologie. Pertanto, la considerazione sui mezzi non deve essere vista come una valutazione *una tantum*, ma come un'operazione che deve essere oggetto di un riesame periodico in ragione dei nuovi rischi connessi alla crescente disponibilità di mezzi tecnici a basso costo (il cloud computing ad esempio), all'accessibilità pubblica sempre maggiore di altre banche dati e alle competenze tecniche utilizzabili.

Un processo di anonimizzazione che si basi su tecniche (di distorsione o generalizzazione dei dati) riconosciute dalla comunità scientifica internazionale e che tenga conto degli aspetti contestuali idonei a valutare l'irragionevolezza dei mezzi è dunque, a tutti gli effetti, strumento di tutela integrato nel trattamento, così come richiesto dal principio di privacy by design introdotto dal nuovo Regolamento. Inoltre, considerata la finalità

ulteriore perseguita dal processo di anonimizzazione, di impedire la re-identificazione della persona mediante l'uso di ogni mezzo ragionevole, tale finalità diventa compatibile con *qualsiasi* finalità iniziale originalmente e legittimamente perseguita dal titolare, prestandosi a promuovere un riuso dei dati ampio e trasversale, come è nella *ratio* del modello Big Data.

6.2. La pseudonimizzazione come misura di sicurezza

Un'ulteriore e distinta opzione di tutela integrata nel trattamento è offerta dai processi di pseudonimizzazione del dato. Si tratta dell'applicazione di un insieme di tecniche sulle quali è bene soffermarsi, proprio per sottolinearne la differenza concettuale rispetto al processo di anonimizzazione e le rispettive differenti finalità di tutela perseguite, dal momento che dall'erronea interpretazione della pseudonimizzazione come forma di anonimizzazione le tutele ipotizzate possono risultare fortemente compromesse, ove non vanificate. La pseudonimizzazione infatti consiste nel sostituire un attributo, solitamente univoco, di un dato con un altro, ugualmente univoco e solitamente non immediatamente intellegibile. Questo accorgimento può rendere più complessa l'identificazione, richiedendo mezzi anche onerosi per la riferibilità del dato alla persona, ma mantiene inalterato il quadro di certezze nella concatenazione dei passaggi necessari per l'attribuzione del dato pseudo-anonimo a quest'ultima. Si tratta di un risultato in un certo senso opposto a quello che si prefigge l'anonimizzazione, basata proprio sull'introduzione di incertezze nell'attribuzione di un dato ad un soggetto determinato. In altri termini, l'associazione biunivoca tra dato e persona non è modificata in alcun modo dalla pseudonimizzazione e il dato pseudo-anonimo, una volta impiegato in combinazione con tutti i mezzi necessari per effettuare la sostituzione di attributi a ritroso, è inequivocamente riferibile alla persona. Ciò non accade con un processo di anonimizzazione ben congegnato: sia con l'applicazione delle tecniche di distorsione, sia di generalizzazione, infatti, la riferibilità del dato anonimizzato alla persona diventa, lo si ribadisce, verosimile quanto una attribuzione casuale. All'esito di un processo di pseudonimizzazione, la persona potrebbe essere ancora identificata in maniera indiretta e di conseguenza la pseudonimizzazione, riducendo l'intellegibilità di un insieme di dati relativi comunque a una persona interessata, rappresenta, se ben realizzata, unicamente una misura di sicurezza utile, ma non un metodo di anonimizzazione.

Il risultato della pseudonimizzazione può essere indipendente dal dato iniziale (come accade nel caso di un valore casuale assegnato a un attributo del dato) o può essere calcolato a partire dal valore originale di un attributo o insieme di attributi, ad esempio mediante l'applicazione di una tecnica crittografica.

Il titolare è dunque chiamato ad una accurata valutazione preliminare del tipo di tutela integrata nel trattamento che si prefigge. Se, in ipotesi, dall'uso del dato trattato si generano comunque specifiche conseguenze sulla persona che richiedono il mantenimento dei diritti di accesso sul dato (ad esempio per ragioni connesse alla garanzia sulla sua qualità), allora il titolare sta procedendo *naturaliter* ad un trattamento di pseudonimizzazione, che si pone proprio l'obiettivo di impedire ogni incertezza sull'attribuzione del dato.

Si tratta, dunque, di un tipo di tutela integrata nel trattamento concettualmente diversa rispetto a quella perseguita dal processo di anonimizzazione. In questo caso infatti, la tutela introdotta con la pseudonimizzazione è volta a garantire la confidenzialità del dato, non più immediatamente intellegibile, ma anche, come avviene nel caso dell'applicazione di tecniche crittografiche, a garantirne l'integrità contro manipolazioni anche accidentali. Nel caso dell'anonimizzazione la tutela è invece volta a impedire, a meno di dover ricorrere a mezzi irragione-

volmente utilizzabili, la riferibilità del dato a una persona. L'una è tutela, ovvero misura, di sicurezza, l'altra di privacy.

7. Conclusioni

La privacy by design si presenta dunque come uno strumento operativo introdotto dal nuovo Regolamento per offrire alle persone nuove forme di tutela "integrate nei trattamenti", nonché per offrire ai titolari la possibilità di accompagnare la progettazione dei servizi con opportunità di tutela "solidali" allo sviluppo stesso del servizio. Questo approccio appare molto appropriato per un contesto *data intensive* come quello dei Big Data. Vedere le tutele come "contrapposte" ai trattamenti (ossia esigibili in tempi distinti dal trattamento, o con strumenti diversi da quelli impiegati per ricevere un servizio), lo si è detto, non è la sola prospettiva possibile per un esercizio dei diritti a prova dei tempi. Le tutele sono necessarie per la prosperità stessa di queste nuove applicazioni e poter intervenire per tempo modificando le "sfere di influenza" dei dati in modo tale da non generare collisioni con dati personali o con "cose" che ci appartengono, o la cui compromissione può esporci a rischi, è la prospettiva di lungo termine che il nuovo quadro giuridico ci offre.

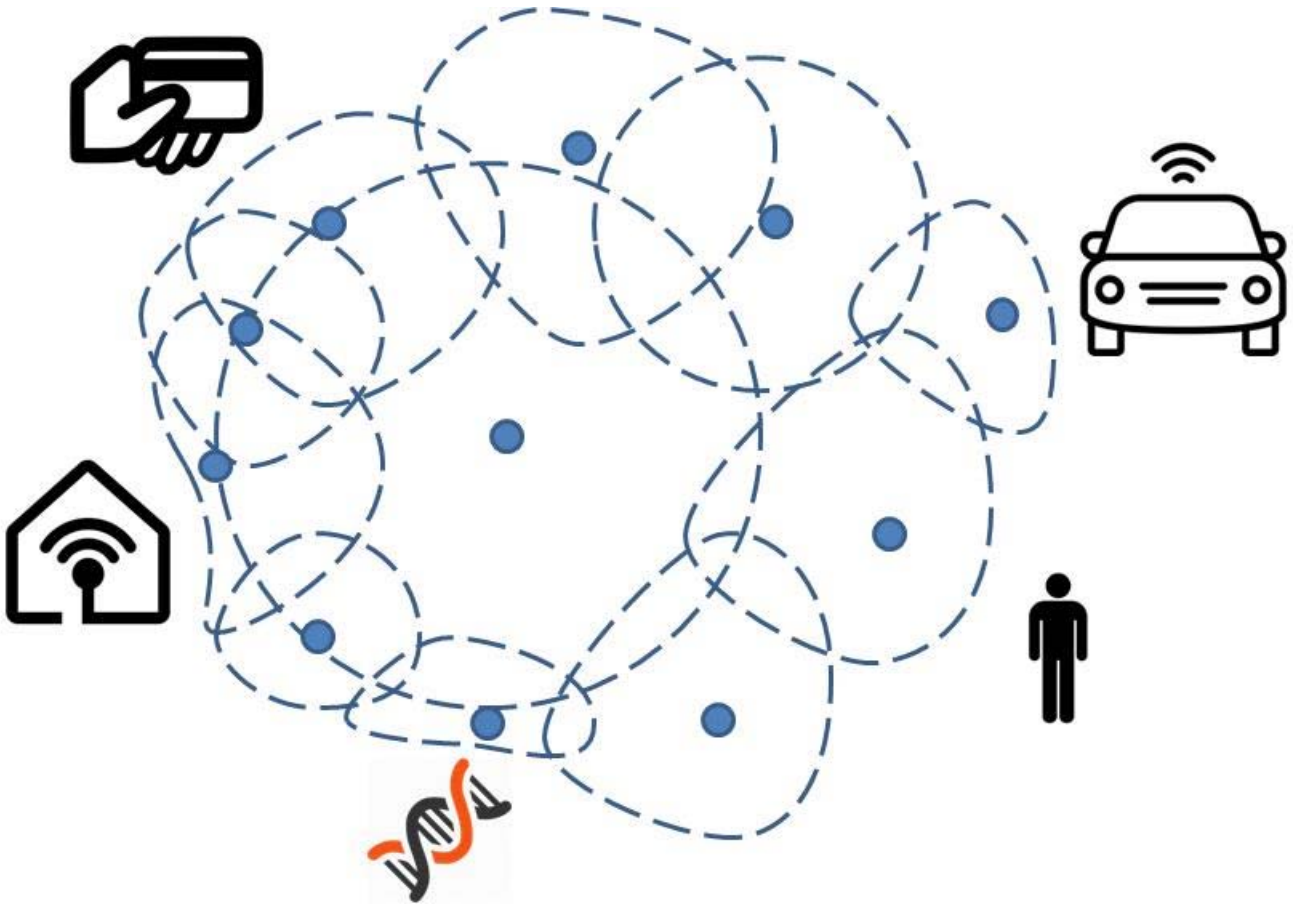


Figura 11. La privacy by design: modificare le "sfere d'influenza" per tenere dentro i dati e fuori le persone

"*Gli oggetti sono in certo modo doppi*", ci dice Leopardi traguardando con l'immaginazione i secoli. Oggi lo sono per davvero e la loro parte nascosta è potenzialmente rivelatrice di connessioni e generatrice di conoscenza. Nel "*secondo genere di obbietti sta tutto il bello e il piacevole delle cose*", ma anche, purtroppo, di insidie e di rischi. Affrontiamo la sfida di valorizzare il bello e il piacevole (e l'utile), mitigando il più possibile i rischi, con la consapevolezza che la privacy by design è una buona idea che nel tempo troverà la sua strada, e che un bagaglio di strumenti analitici comincia già a emergere per realizzare un'ingegneria su questo nuovo principio. Nei prossimi capitoli introdurremo le tecniche disponibili per l'anonimizzazione e la pseudonimizzazione dei dati, fornendo strumenti utili per la valutazione di casi concreti e per la progettazione di soluzioni specifiche.

8. Bibliografia

[C15] R. Calo, *Privacy and Markets: A Love Story*. University of Washington School of Law. Legal Studies Research Paper No. 2015-26, 2015

[CGE14] Sentenza della Corte (Grande Sezione) 13 maggio 2014, *«Dati personali – Tutela delle persone fisiche con riguardo al trattamento di tali dati – Direttiva 95/46/CE – Articoli 2, 4, 12 e 14 – Ambito di applicazione materiale e territoriale – Motori di ricerca su Internet – Trattamento dei dati contenuti in siti web – Ricerca, indicizzazione e memorizzazione di tali dati – Responsabilità del gestore del motore di ricerca – Stabilimento nel territorio di uno Stato membro – Portata degli obblighi di tale gestore e dei diritti della persona interessata – Carta dei diritti fondamentali dell'Unione europea – Articoli 7 e 8».*

[CDM14] P. Coucheney, G. D'Acquisto, P. Maille, M. Naldi and B.Tuffin, *Influence of search neutrality on the economics of advertisement-financed content*, ACM Transactions on Internet Technology, Volume 14, Issue 2-3, Article 10, October 2014

[D06] C. Dwork, *Differential Privacy*, 33rd International Colloquium on Automata, Languages and Programming. Venice, Italy,:pag. 1-129–16 July 2006

[DDK15] G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.A. de Montjoye, A. Bourka, *Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics*. ENISA report, 2015

[FJP16] E. Fernandes, J. Jung, A. Prakash, *Security Analysis of Emerging Smart Home Applications*, 37th IEEE Symposium on Security and Privacy, (S&P 2016), San Jose, May 2016.

[FTC13] *Statement of the Federal Trade Commission Regarding Google's Search Practices In the Matter of Google Inc. FTC File Number 111-0163* January 3, 2013

[O09] A. Odlyzko, *Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets*,. Review of Network Economics, vol. 8, no. 1, pp. 40-60, March 2009

[O10] P. Ohm, *When Network Neutrality Met Privacy*, Communications of the ACM, Vol. 53 No. 4, Pages 30-32, 2010

[OS99] A. V. Oppenheim, R. W. Schafer, J.R. Buck, *Discrete-time signal processing*, Prentice Hall, 1999

[P11] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Press New York, 2011

[P16] F. Pizzetti, *Privacy e il diritto europeo alla protezione dei dati personali. Dalla Direttiva 95/46 al nuovo Regolamento europeo*, Giappichelli 2016

[R14] J. Rifkin, *La società a costo marginale zero. L'Internet delle cose, l'ascesa del Commons Collaborativo e l'eclissi del capitalismo*, Milano, Mondadori, 2014

[S54] H.A. Simon, *Spurious Correlation: A Causal Interpretation*, Journal of the American Statistical Association, Volume 49, Issue 267, 1954

[SBD15] F. Schaub, R. Balebako, A.L. Durity, L.F. Cranor, *A Design Space for Effective Privacy Notices*, Eleventh Symposium On Usable Privacy and Security (SOUPS), 1-17, 2015

[url1] *https://en.wikipedia.org/wiki/Hash_function*

[url2] *https://en.wikipedia.org/wiki/Machine_learning*

[url3] *https://en.wikipedia.org/wiki/Stylometry*

[url4] *https://en.wikipedia.org/wiki/Fingerprint_(computing)*

[url5] *https://en.wikipedia.org/wiki/Acoustic_fingerprint*

[url6] *https://en.wikipedia.org/wiki/Digital_video_fingerprinting*

[url7] *https://en.wikipedia.org/wiki/Crowdsourcing*

[url8] *https://en.wikipedia.org/wiki/CAPTCHA*

[url9] *http://www.iso.org/iso/internet_of_things_report-jtc1.pdf*

[url10] *https://en.wikipedia.org/wiki/Streisand_effect*

[url11] *http://www.parlamentari.org/wp-content/uploads/2015/11/2015-Odlyzko-Slides-The-end-of-privacy-and-the-seeds-of-capitalism_s-destruction.pdf*

[url12] *http://www.theatlantic.com/magazine/archive/2012/11/hacking-the-presidents-dna/309147/*

[WP2913] *WP29, Opinion 03/2013 on Purpose Limitation*, *http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf*, (pag. 23-27)

[WP2914] WP29, *Opinion 05/2014 on Anonymization Techniques*, *http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf*