Introduzione

Il tema del corretto uso e della protezione dei dati personali è sempre più rilevante nella società in cui viviamo. Ne facciamo esperienza in molti ambiti, da quello ricreativo a quello professionale, dai rapporti con la pubblica amministrazione a quelli con i molti venditori di beni e servizi che acquistiamo online (libri, biglietti, vacanze etc.): nulla di ciò che quotidianamente facciamo, utilizzando online e offline strumenti tecnologici, sarebbe possibile senza l'impiego dei nostri dati (l'indirizzo di casa, o del luogo di lavoro, per la consegna di un pacco, i dati della carta di credito per il pagamento, il nostro numero di telefono, l'indirizzo email e così via). Inoltre (come ben sa chi si occupa direttamente delle tecnologie impiegate per trattare i nostri dati) la quantità di dati, così come il numero di soggetti che li trattano, sono cresciuti nel tempo, e questa tendenza alla crescita non è un effetto transitorio, ma è destinata a mantenersi e, per ciò che ci dicono tutte le analisi più autorevoli [1], a intensificarsi.

C'è infatti una domanda crescente di dati personali, alimentata dalla disponibilità di molti terminali, dalla capillarità e capacità della rete e dalla mobilità delle persone. Ma c'è anche un crescita, forse ancora maggiore, di offerta di dati personali, legata allo sviluppo e alla diffusione di applicazioni che sui dati fondano il loro successo (reti sociali, applicazioni video, giochi online, piattaforme di commercio elettronico, cloud computing, internet delle cose), alla migliore efficacia nella distribuzione dei beni fisici, e alla personalizzazione del processo produttivo [2]. Si parla di Big Data per indicare questo fenomeno.

D'altro canto, ormai da venti anni, il quadro giuridico europeo riconosce alle persone il diritto alla protezione dei propri dati personali [3], codificato in una serie di principi (finalità, necessità, proporzionalità, ad esempio), oltre che in una serie di obblighi per chi vuole trattare i nostri dati (trasparenza, sussistenza di legittimità, tra i più importanti). Il fatto poi che questo diritto sia un diritto fondamentale per le persone e che il nuovo quadro giuridico (che sarà in vigore a partire dal 2018 con il nuovo Regolamento Generale sulla protezione dei dati personali [4], che disciplinerà la materia, non più in ambito nazionale ma in un contesto continentale, probabilmente per i prossimi venti anni) abbia confermato e talora anche rafforzato le tutele per le persone, ci porta a concludere che sia ineludibile affrontare la questione Big Data e la questione privacy congiuntamente.

Ma, a nostro avviso, l'approccio per affrontare questi temi non può essere più soltanto, o prevalentemente, di natura giuridica. Il legislatore europeo, intro-

ducendo il principio di Privacy by Design, ha infatti attribuito alla componente tecnologica un ruolo di "tutela", aggiuntivo alle tutele tradizionali previste dal precedente quadro giuridico. Come si legge nell'articolo 25 del nuovo Regolamento infatti

Tenendo conto dello stato dell'arte e dei costi di attuazione, nonché della natura, dell'ambito di applicazione, del contesto e delle finalità del trattamento, come anche dei rischi aventi probabilità e gravità diverse per i diritti e le libertà delle persone fisiche costituiti dal trattamento, sia al momento di determinare i mezzi del trattamento sia all'atto del trattamento stesso il titolare del trattamento mette in atto misure tecniche e organizzative adeguate, quali la pseudonimizzazione, volte ad attuare in modo efficace i principi di protezione dei dati, quali la minimizzazione, e a integrare nel trattamento le necessarie garanzie al fine di soddisfare i requisiti del presente Regolamento e tutelare i diritti degli interessati.

Si tratta di una possibilità nuova, e di grande rilievo, offerta alle persone e a coloro che trattano i dati delle persone di vedere nei trattamenti anche una forma di tutela. Trattandosi di un principio nuovo, tutto è da fare e ogni domanda è lecita. Cosa è un trattamento di tutela? Come si realizza? Si può misurare il livello di tutela introdotto? Rispondere a queste domande non è semplice, ma è necessario, come diremo. Il dibattito su questi temi è già vivo; molte proposte sono state avanzate, prevalentemente dalla comunità scientifica, e stanno progressivamente diventando strumento per i decisori.

Non si può quindi separare il piano normativo da quello tecnico, e già oggi la giurisprudenza include riferimenti di carattere tecnico che fino a qualche anno fa erano dominio esclusivo degli specialisti di formazione ingegneristica.

Il libro si pone quindi l'obiettivo di raccogliere le principali idee sulle modalità tecniche disponibili per integrare le tutele nei trattamenti, sforzandosi di renderle accessibili anche a lettori con un bagaglio di conoscenze prevalentemente giuridiche sui temi della protezione dei dati personali, oppure tecnico-economiche, ma magari non ancora del tutto formato su questi temi nuovi. Il libro è quindi rivolto a chi si occupa di tecnologie, ma a diversi livelli e con una diversa formazione: chi le progetta, chi le impiega per realizzare servizi, chi ne disciplina l'uso all'interno delle aziende o in ambito pubblico, chi prende decisioni strategiche su investimenti e piani di sviluppo. Tutti costoro, prima o poi, si accorgeranno che una delle componenti di tali tecnologie, ciò che le fa funzionare nel modo in cui essi le pensano e noi tutti le utilizziamo è proprio quella particolare categoria di dati, riferibili agli utilizzatori di tali tecnologie, di cui non è (e sempre meno sarà) possibile fare a meno: i dati personali. Mostreremo che la crescita nella disponibilità di questi dati non è soltanto una questione quantitativa, ma determina un cambiamento radicale di tipo cognitivo. La quantità di dati cresce da sempre e da sempre le tecnologie sono capaci di far fronte a questa crescita con un incremento delle prestazioni, ossia delle capacità trasmissive delle reti e di *processing* delle macchine. Ma oggi la prospettiva dei Big Data può costituire un vero e proprio cambiamento di paradigma, in cui algoritmi sempre

più sofisticati sono capaci di fornirci l'informazione che ci serve attraverso la scoperta di relazioni significative all'interno di una mole crescente di dati.

La struttura che abbiamo dato al libro prevede un ampio capitolo in cui si descrive il *quid pluris* dei Big Data perché su di essi si possa fondare una reale aspettativa di novità. Intimamente connesso al tema dei Big Data è il rapporto tra i dati e *le cose*. Affronteremo l'argomento senza intenti di demonizzazione: si tratta obiettivamente di grandissime opportunità di "crescita", in ogni campo, dalla crescita economica a quella culturale. Tuttavia, mostreremo, questa opportunità ci espone a dei rischi, che bisogna conoscere e fronteggiare. Quindi, nella parte centrale del libro, introdurremo i fondamenti delle tecniche ad oggi disponibili per fronteggiare questi rischi, occupandoci di anonimizzazione dei dati e della loro pseudonimizzazione. Si tratta di tecniche relativamente nuove che richiedono la conoscenza di un substrato minimo di strumenti analitico-matematici, prevalentemente derivati dalla teoria della probabilità, di cui daremo le basi e che approfondiremo in alcuni riquadri separati dalla "trama" principale del libro, per non affaticare il lettore con questioni più spiccatamente tecniche, ma offrendo allo stesso tempo a chi è interessato la possibilità di una conoscenza più dettagliata. Infine, ci occuperemo di sicurezza e di investimenti nella sicurezza, mostrando come l'applicazione del principio di Privacy by Design possa consentire di raggiungere più facilmente l'obiettivo di una maggiore sicurezza delle informazioni e come, letta alla luce di questo nuovo principio, la sicurezza possa dare un impulso nuovo al raggiungimento di un obiettivo cruciale per l'affermazione dei Big Data, quale è quello legato alla qualità dei dati. Il libro si chiude con un indice analitico per argomenti, che consentirà una lettura del testo per temi o parole chiave, e con una bibliografia selezionata per chi vorrà ulteriormente approfondire i temi trattati con le più recenti pubblicazioni in area tecnologica e giuridica. Essendo un libro rivolto a lettori con diverse formazioni, vari percorsi di lettura possono essere suggeriti. Così, un lettore con una formazione più giuridica potrà concentrarsi maggiormente sulle questioni di scenario delineate nei capitoli 1 e 4, trovando nei capitoli 2 e 3 la possibilità di un approfondimento su questioni di natura più tecnica e progettuale. Viceversa, un lettore con una formazione più tecnica potrà acquisire familiarità con le metodologie di privacy by design introdotte nei capitoli 2 e 3, e avere un quadro d'insieme sui contesti in cui queste metodologie trovano applicazione dalla lettura dei capitoli 1 e 4.

Desideriamo ringraziare il prof. Franco Pizzetti per avere creduto nel cambio di prospettiva indotto dal nuovo Regolamento, incoraggiando sin dall'inizio questo lavoro, che vuole portare a più stretto contatto le due anime, giuridica e tecnologica, della protezione dei dati personali in modo che possano comprendersi sempre di più, anziché divergere per via della complessità dello scenario tecnologico, e per aver scritto la prefazione del libro.

Siamo infine debitori ai colleghi e amici revisori per la loro accurata rilettura del testo, assai preziosa. In particolare, desideriamo ringraziare il dott. Antonio Caselli, l'ing. Serena D'Acquisto, il dott. Onofrio Fanelli.

Saremo lieti di ricevere dai lettori commenti, suggerimenti o segnalazioni di errori ai seguenti indirizzi di posta elettronica

dacquisto@ing.uniroma2.it
maurizio.naldi@uniroma2.it

DISCLAIMER: Nello scrivere questo libro, gli autori hanno attinto dalla loro personale esperienza di ricercatori e hanno beneficiato dalla possibilità di confrontare le loro idee con molti colleghi nei tavoli di lavoro e nelle conferenze internazionali in cui questi argomenti sono dibattuti. Per il ruolo svolto dagli autori, molte di queste discussioni hanno nel tempo dato luogo a pubblicazioni scientifiche e hanno generato spunti anche per la formulazione di documenti istituzionali. Tuttavia è bene far presente che quanto scritto in questo libro è unicamente opinione degli autori e non potrà essere attribuito alle rispettive Istituzioni e Università in cui svolgono la loro attività.

Capitolo 1
Big Data e protezione dei dati personali

SOMMARIO: 1.1. Cosa è *Big* nei *Big Data*? – 1.2. Internet e le *cose*. – 1.3. Perché la privacy è importante. – 1.4. I principi della protezione dei dati personali. – 1.5. Le nuove sfide per l'applicazione dei principi. – 1.6. Cosa è privacy by design. – 1.6.1. Il concetto di anonimizzazione e di dato anonimizzato. – 1.6.2. Il concetto di pseudonimizzazione e di dato pseudonimo.

> 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*

1.1. Cosa è *Big* nei *Big Data*?

Una maniera consolidata di descrivere il fenomeno dei Big Data è costituita dal cosiddetto paradigma delle tre V, dalle iniziali delle tre caratteristiche fondamentali: Volume, Velocità e Varietà. A queste tre V se ne aggiunge spesso una quarta, definita come Veracità (*Veracity*) [1].

La prima V (Volume) si riferisce al fatto che il gran numero di dispositivi oggi connessi alla rete fornisce una mole di dati che mai era stata disponibile nella storia precedente dell'umanità. Si stima che ogni giorno vengano prodotti nel mondo 2,5 miliardi di miliardi di byte di nuovi dati (sotto le più svariate forme: documenti Word, file pdf, immagini, video, etc.). Ma non è solamente la quantità di dati prodotti che fa la differenza, bensì anche la facilità (e l'economicità) della loro raccolta. La disponibilità di questi dati, connessa alla sempre maggiore potenza di elaborazione dei computer (e di dispositivi come

[1] Una infografica che descrive molto bene le quattro V si trova sulla pagina *http://www.ibmbigdatahub.com/infographic/four-vs-big-data*.

smartphone e tablet che non sono classificati direttamente come computer, ma in realtà hanno capacità di elaborazione di tutto rispetto), permette di analizzare qualsiasi fenomeno (sociale, economico, etc.) non più parzialmente, ovvero come avveniva in passato, anche recente, semplicemente mediante un campione rappresentativo[2], ma globalmente, ovvero considerando *tutti i dati relativi al fenomeno*. Per esempio, supponiamo che un operatore telefonico voglia condurre un'indagine sull'utilizzo degli sms da parte dei suoi utenti. In precedenza, l'operatore avrebbe scelto (mediante una procedura di campionamento statistico) un insieme di utenti (con caratteristiche di età, distribuzione geografica, etc. analoghe a quelle dell'insieme dei suoi utenti) ed avrebbe analizzato le azioni di quegli utenti (quanti sms inviano, in che ore del giorno, e così via). Oggi, lo stesso operatore può conteggiare **tutti** gli sms inviati da **tutti** gli utenti ed avere così una visione completa del fenomeno. Inoltre, la scelta di un campione era tipicamente orientata ad una particolare finalità; una nuova analisi per una finalità diversa richiedeva un nuovo campionamento, con un ovvio impatto sui tempi e sui costi dell'analisi. La disponibilità di tutti i dati li rende potenzialmente riutilizzabili per altre finalità future.

La seconda V (Velocità) si riferisce al fatto che i dati vengono ormai prodotti con continuità, in maniera dinamica e non più statica. Per esempio, ogni autovettura contiene ormai un gran numero di sensori, che monitorano continuamente tutti gli aspetti del suo funzionamento e accumulano i relativi dati. I dati relativi alle nostre abitudini di consumo vengono continuamente aggiornati. Siamo quindi in presenza di dati forniti sotto forma di *flusso* di informazioni, rilasciate ad una certa velocità, ovvero di uno *stream* di dati. L'elaborazione deve quindi essere continuamente aggiornata sulla base dei nuovi dati. Derogare da questa necessità può portare ad errori. Un esempio clamoroso è costituito dalla predizione della diffusione di casi di influenza, svolta da Google qualche anno fa. Nel 2009 un nuovo ceppo influenzale, derivato dall'influenza aviaria, venne individuato e ci si interrogò subito sulla possibile dinamica della sua diffusione e sulla possibilità di una pandemia (che avrebbe richiesto uno sforzo di contenimento di dimensioni enormi). Una tale analisi, condotta tradizionalmente sulla base delle statistiche raccolte dai medici di famiglia, avrebbe richiesto tempi abbastanza lunghi rispetto alle necessità di una pianificazione degli interventi (a causa dei ritardi con cui un paziente si reca dal dottore e del successivo tempo necessario affinché i dati pervengano dal medico di famiglia all'ente centrale responsabile per l'analisi) e sarebbe risultata comunque incompleta (molti non ricorrono alle cure del medico per ciò che classificano come una semplice influenza). Google individuò allora una tecnica di predi-

[2] Per campione rappresentativo si intende una piccola parte dell'insieme che costituisce oggetto dell'indagine, scelta mediante una procedura di *campionamento statistico* che garantisce che il campione sia rappresentativo dell'insieme, ovvero che da quel campione si possano trarre indicazioni valide per tutto l'insieme. Si tratta dell'operazione che si effettua per esempio nei sondaggi politici o commerciali, in cui si chiede il parere di un piccolo numero di persone per estrapolare conclusioni relative a tutta la popolazione di elettori o di consumatori.

zione basata su **tutte** le richieste di informazioni inviate al motore di ricerca riguardanti l'influenza, e sulla possibilità di provare un gran numero di modelli diversi per descrivere l'evoluzione del fenomeno, selezionando poi quello migliore [5]. Il metodo si rivelò capace di individuare con precisione il picco di casi di influenza del 2009.

La terza V sta per Varietà. Con ciò si intende il fatto che i dati arrivano ormai in una varietà di formati e da una varietà di fonti. Si tratta di dati pubblicati sulle pagine Facebook, oppure di dati originati da sensori, o ancora di video pubblicati su YouTube, oppure di tweet lanciati su Twitter. L'informazione disponibile cresce quindi grazie a contributi variegati, ma nel contempo diviene più complessa la sua gestione.

Infine la quarta V, talvolta associata alle prime tre per definire compiutamente il fenomeno dei Big Data, sta per *Veracity*, ovvero per veridicità (o qualità) dei dati. Infatti l'eterogeneità stessa delle sorgenti rende più complesso l'accertamento della correttezza dei dati. Ci si deve aspettare che l'aumento dei dati comporti anche un aumento dell'incertezza, perché alcuni di questi non saranno corretti.

Questa definizione del fenomeno, basata sulle 3 o 4 V, è ormai abbastanza consolidata. Ma se riflettiamo sui passaggi di un semplice processo di formazione della conoscenza, legato per esempio alla scrittura di questo stesso libro che voi leggete, possiamo sin da subito cogliere altri significati del termine "big" e farci un'idea, forse più immediata e intuitiva, 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. 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 quando

affrontiamo particolari "concetti", ci fossero automaticamente 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 *medium* (ad esempio, la conoscenza è avvenuta tramite una serie di collegamenti su internet, mentre 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 libro, 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 libro. Ma tanto l'albergo, quanto questo testo, 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 queste 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", mentre noi (ancora) troviamo successivamente ciò che ci serve. Questo passaggio, con i rischi e costi associati, è ancora interamente nostro.

Cosa manca per compiere quest'ultimo passo? Serve una rappresentazione dei dati adeguata rispetto allo scopo: servono dati "più grandi", *bigger data* e non soltanto *more data*.

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 più "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 le due immagini riportate in Figura 1.1 e 1.2.

Figura 1.1: Primo termine di una relazione Big Data

Figura 1.2: 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 la-

scia 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 esaminare insieme i due eventi a cui si riferiscono.

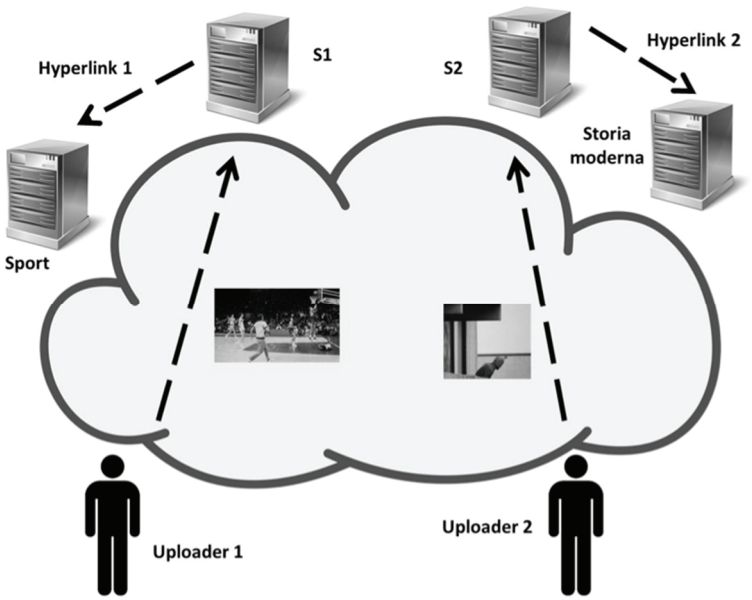

Figura 1.3: Relazione tra dati in una rete Small Data

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 in Figura 1.3. Ciascuno dei due soggetti che caricavano i

dati (o uploader) decideva a quali altri dati ciascuna immagine fosse collegata, secondo criteri unilateralmente stabiliti. Nell'esempio in figura, l'uploader 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 le 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 era 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.

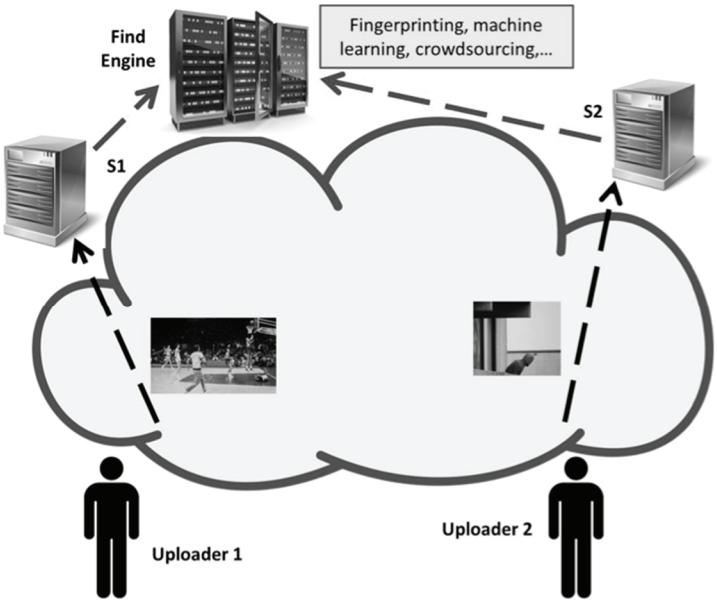

Figura 1.4: Relazione tra dati in una rete Big Data

In "modalità Big Data", il motore di ricerca – sempre più "find engine" e non soltanto "search engine" – effettua un'elaborazione 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 in Figura 1.4. Ma dov'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.

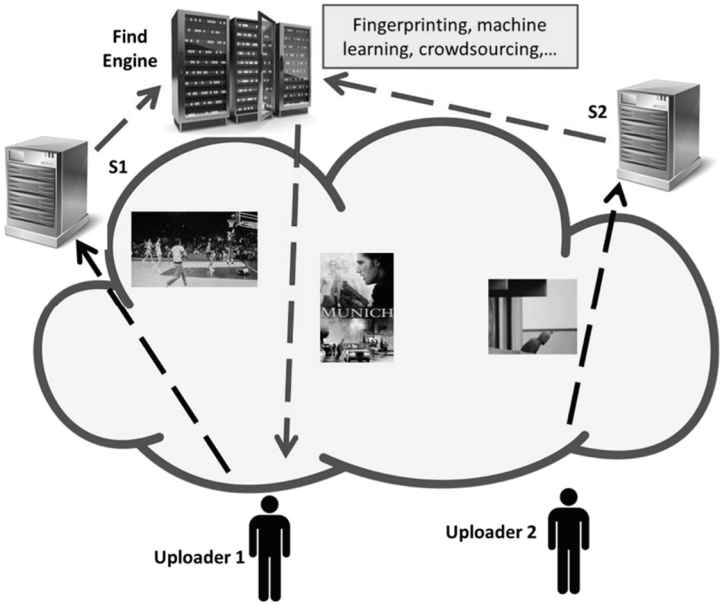

Figura 1.5: La generazione di nuova conoscenza in una rete Big Data

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 approfondimento 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 1.5. 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

(come illustrato in Figura 1.6), che per semplicità e comodità suddivideremo in due sezioni: la sezione "Small Data" a sinistra, la sezione "Big Data" 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 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 nella sezione "Big Data" della tabella).

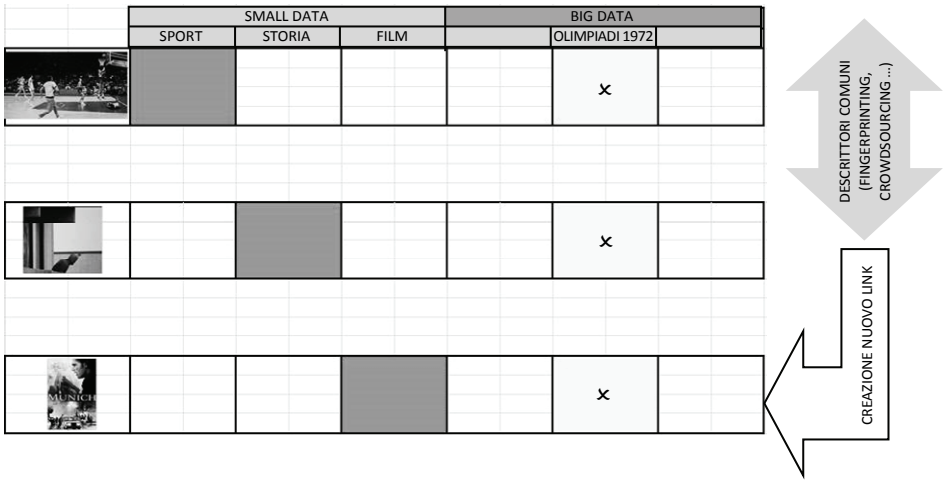

Figura 1.6: I trattamenti di dati in uno schema Big Data

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 effettuate in modo automatico e senza intervento umano, altre invece, che coinvolgono la sfera dei "significati" da attribuire ai dati, sono evidente-

mente di natura attiva e non possono prescindere dall'intervento umano. Rientrano nella categoria delle tecniche passive le seguenti metodologie di processing:

– *machine learning*[3], ovvero l'individuazione automatica delle categorie a cui il dato appartiene. Si distingue in *unsupervised machine learning* e *supervised machine learning*. Nel primo caso le categorie non sono pre-identificate, ma vengono definite dall'algoritmo a partire dall'analisi di tutti i dati. Nel secondo caso le categorie sono state già definite, e l'obiettivo è assegnare un nuovo dato alla categoria corrispondente. Si tratta di una tecnica idonea ad essere impiegata in un'ampia gamma di situazioni per dati testuali o multimediali (v. Box 1.1);

– *hashing*[4], 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.

Box 1.1. Il Machine Learning

Il machine learning, o apprendimento automatico, è un settore dell'informatica che sviluppa algoritmi e metodi per consentire ai computer di apprendere specifici compiti direttamente dai dati, senza la necessità di una esplicita, dedicata programmazione. La disciplina nasce intorno agli anni '50 del secolo scorso, come parte degli studi sull'Intelligenza Artificiale e si sviluppa fino agli anni '70, grazie anche ai primi successi sperimentali delle reti neurali (si veda l'infografica in Figura 1.7). Negli anni 2000, lo sviluppo della miniaturizzazione dei circuiti elettronici e di algoritmi efficienti per suddividere le attività di calcolo su più processori in parallelo porta alla creazione di reti neurali sempre più complesse, come nel caso delle reti cosiddette "deep learning" contenenti centinaia o migliaia di strati di neuroni artificiali, i cui successi ne sanciscono una prima diffusione commerciale. Oggi gli algoritmi di Machine Learning sono utilizzati dai più importanti fornitori di servizi online, da Google a Facebook, da Amazon a Netflix, per migliorare i risultati delle ricerche Web, individuare le notizie o i prodotti di maggiore interesse per l'utente. A gennaio del 2016 gli algoritmi di Machine Learning sviluppati da Google sono stati sulle pagine dei principali giornali in quanto un sistema di questo tipo ha battuto il campione umano Lee Sedol al gioco del Go, l'ultimo gioco da tavolo per il quale gli umani continuavano a giocare meglio delle macchine.

I problemi che possono essere trattati con successo usando algoritmi di Machine Learning sono molteplici ma per comodità possono essere suddivisi in quattro macro gruppi:

– classificazione: i dati in ingresso devono essere suddivisi in due o più classi, e l'algoritmo deve produrre un modello che assegna ingressi non noti ad una di queste classi (etichetta);

– regressione: l'output del modello non è un valore discreto (etichetta) ma un

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

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

valore continuo. Un esempio di utilizzo è la stima del valore di un immobile sulla base delle sue caratteristiche (es. metri quadri, numero vani, posizione, etc.);

– clustering: simile alla classificazione, con la differenza che nel caso del clustering non è nota a priori la suddivisione in gruppi dei dati per l'addestramento, ma tale suddivisone emerge autonomamente proprio mediante l'apprendimento;

– riduzione dimensionale: l'algoritmo deve essere in grado di mappare correttamente input multidimensionali in uno spazio con un minor numero di dimensioni. Un esempio di utilizzo di questo tipo di algoritmi è la ricerca di documenti correlati per argomento.

Un'altra importante distinzione per gli algoritmi di Machine Learning è invece relativa alla modalità di apprendimento. La distinzione qui è fra tre metodi di apprendimento:

– apprendimento supervisionato: l'algoritmo viene "addestrato" presentando coppie di input-output note;

– apprendimento non supervisionato: è lasciato all'algoritmo il compito di dividere dati in gruppi. La definizione di gruppi è il fine dell'attività;

– apprendimento rinforzato: l'algoritmo interagisce con un sistema in tempo reale e deve cercare di raggiungere un obiettivo predefinito. Ad esempio, acquisire la capacità di guidare un veicolo, oppure apprendere a giocare un videogioco contro un avversario.

Danilo Benedetti

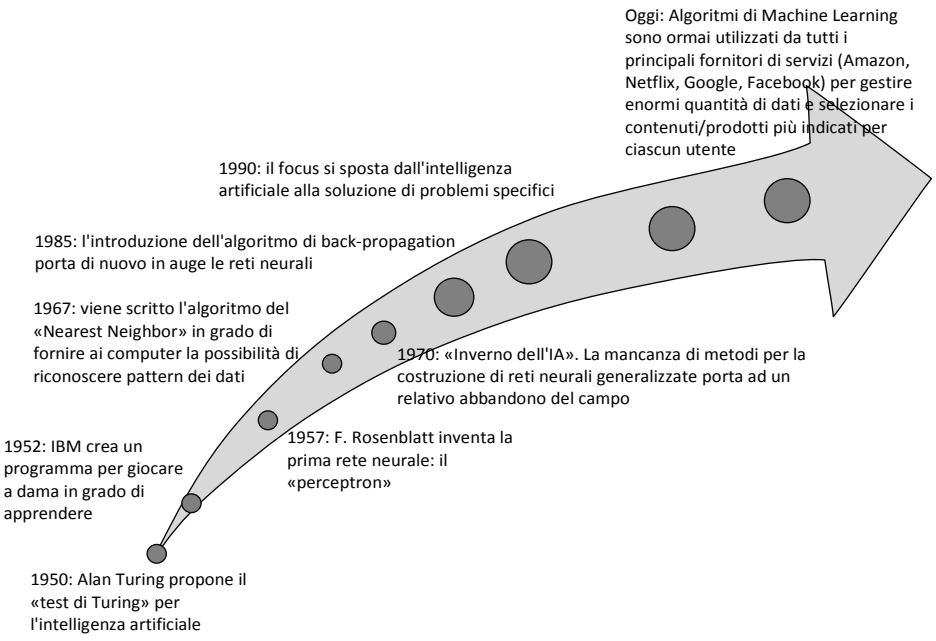

Figura 1.7: L'evoluzione del Machine Learning

– trasformata di Fourier [6], 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[5], 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:

– la classificazione semantica di un contenuto per parole chiave o mediante indicatori[6], (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*[7], 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 etc.), ovvero l'analisi dei suggerimenti e feedback forniti sulle traduzioni automatiche, o ancora l'uso dei risultati dei test CAPTCHA[8] 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 il web attualmente è costruito, ossia su collegamenti tra i dati scelti unilateralmente da chi li immette in rete.

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. 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

---

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

[6] *https://en.wikipedia.org/wiki/Fingerprint_(computing)*, *https://en.wikipedia.org/wiki/Acoustic_fingerprint*, *https://en.wikipedia.org/wiki/Digital_video_fingerprinting.*

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

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

che ascoltiamo. Sono tutti risultati di questa accresciuta interconnessione tra dati che si "sovrappone" alla rete di relazioni tra dati decisa da chi li ha immessi per la prima volta su internet, in modo da arricchirla e potenziarne le capacità di spiegare il mondo. Con uno sguardo prospettico (e, ovviamente, con qualche inevitabile problema legato alla difficoltà 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 (dati che già oggi le macchine fotografiche aggiungono alla 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 corsi personalizzati 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 [7], 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) saranno rimosse, rimettendo alla rete l'onere di cercare i collegamenti o trovare soluzioni, e l'uso di internet sarà 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.

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 rientrino tra gli interessi dell'utente etc.). 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, etc.), 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 preliminarmente farsi carico della 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.

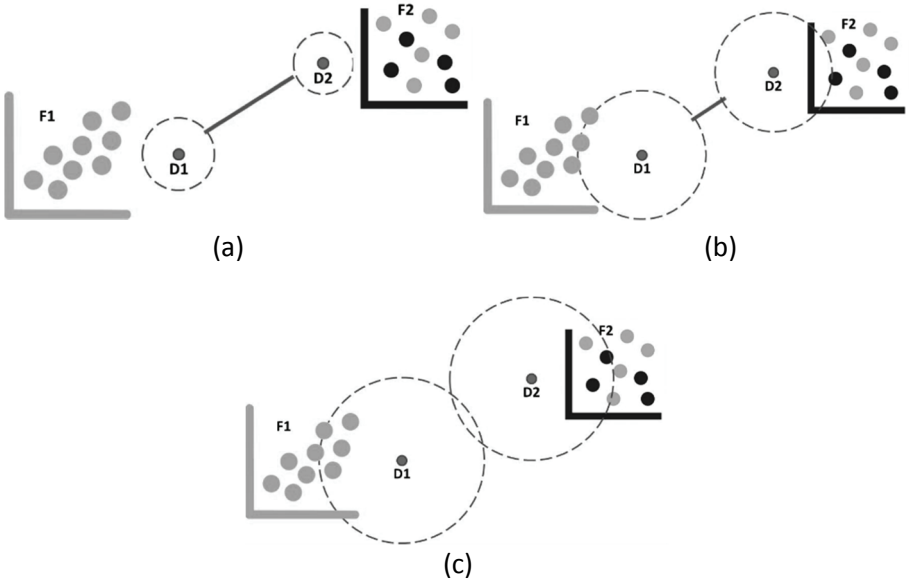

Figura 1.8: Big Data come fenomeno non lineare

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 1.8 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 1.8a). I dati D1 e D2 non consentono di collegare F1 e F2, ossia di ricomprendere i due fenomeni all'interno delle loro reciproche sfere di influenza, e

il costo necessario per metterli in piena relazione (rappresentato dai tratti continui) è prevalentemente in capo al "ricercatore". Ricorrendo ad un primo arricchimento dei descrittori dei due dati (Figura 1.8b) allarghiamo significativamente la sfera d'influenza dei due dati, ma non al punto da consentire ancora di connettere i fenomeni. 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 1.8c) 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. Ne risulta un andamento qualitativamente rappresentabile dal grafico in Figura 1.9, con una curva input-output (ovvero capacità di processing per l'arricchimento del dato vs. potere descrittivo dei dati) a incrementi marginali decrescenti [8], che potrà dare luogo nel tempo a forme di ottimizzazione che molta influenza avranno sull'applicazione di questo schema ai diversi casi concreti che si presenteranno (in primis, la questione dell'allocazione dei costi per l'arricchimento dei descrittori dei dati) 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).

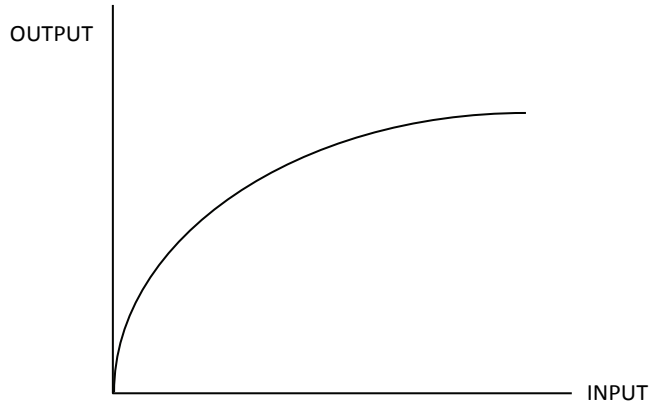

Figura 1.9: Curva input-output a incrementi marginali decrescenti

1.2. 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