Il business della vendita di dati, pratiche e conseguenze nell'etica sociale

Silvia Crafa, Alessandro Zangari

Sommario

L'abbondanza e la facilità di accesso ad informazioni relative ad entità e utenti del Web e alle loro abitudini, hanno condotto al proliferare di compagnie, note con il nome di data broker, *che sfruttano questa ricchezza, raccogliendo ed elaborando tali informazioni per venderle ad altre aziende. La crescente rilevanza che l'utilizzo dei servizi Internet sta assumendo nella società di oggi, unita al progresso degli strumenti di analisi dei* big data, *rende possibile ai data broker ottenere informazioni sempre più rilevanti, specifiche e sensibili. Si evidenzia dunque la necessità di porre degli obiettivi di trasparenza verso il pubblico, nonché di comprendere le implicazioni etiche dei servizi offerti da queste compagnie.*

Abstract

The wealth of digital information and personal data freely accessible on the Web has encouraged many companies, known as data brokers, *to exploit such resources. These entities gather and process large amounts of information to later sell them to other companies. Today, the combined relevance of Internet services in people's lives and the advancements in* big data *analysis allows data brokers to retain larger and larger quantities of remarkably precise and sensitive personal data. It becomes therefore necessary to enforce the principles of transparency and ethical usage of consumer information.*

Keywords: Data broker - Data segments - Data onboarding - Predictive policing - Big Data Ethics - Digital footprint

1. Introduzione

Il ruolo di mediatore che la tecnologia digitale ricopre all'interno delle ordinarie attività umane produce un costante flusso di dati generati dai dispositivi utilizzati quotidianamente, come carte di credito, smartphone e dispositivi indossabili. Sebbene non sia difficile rendersi conto che questi dati possano essere raccolti da alcune aziende, non è immediato comprendere la portata di questa pratica e capirne lo scopo e la rilevanza, sia economica che sociale.

Le informazioni raccolte possono sembrare innocue, ma in verità questi dati, estratti per periodi di tempo significativi e integrati con altre informazioni su abitudini e stato sociale (es. stipendio percepito, cronologia di navigazione, informazioni su abbonamenti), possono fornire un quadro complessivo molto preciso su ogni individuo. La nostra identità è infatti definita in funzione delle esperienze che viviamo, luoghi in cui lavoriamo o interagiamo con gli altri, interessi, famiglia e persone che ci circondano. Proprio queste informazioni, parte integrante della nostra individualità, alimentano un'industria multi-miliardaria, il cui valore stimato supererà nel 2020 i 200 miliardi di dollari [1].

Questo valore deriva in gran parte dalla vendita di *dati personali*, che il vigente Regolamento Generale per la Protezione dei Dati (GDPR) definisce "qualsiasi informazione riguardante una persona fisica identificata o identificabile", dove "si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale" [2, art. 4]. Un dato è un identificatore indiretto, e pertanto comunque considerato *dato personale*, se è possibile risalire all'identità dell'interessato utilizzando tale dato congiuntamente ad altre informazioni [3]. In certi contesti, anche i *dati comportamentali*, sono da considerarsi dati personali; questi dati sono mantenuti sotto forma di una cronologia di azioni che un utente, anonimo o identificato, ha compiuto durante la sua permanenza su un sito web, o più in generale rappresentano un'indicazione di quale comportamento ha avuto un individuo in un determinato contesto [4]. Dati di questo tipo possono includere la cronologia delle pagine visualizzate, dei click effettuati, delle dinamiche di utilizzo del mouse, che possono indicare dove un visitatore di un sito web ha esitato a cliccare, o su quali prodotti si è soffermato maggiormente [5].

Peraltro, non è detto che queste informazioni riguardino solo attività online: ad esempio, i rivoluzionari supermercati Amazon Go non dispongono delle tradizionali "casse", ma tengono invece traccia degli acquisti dei consumatori utilizzando centinaia di telecamere e addebitano gli acquisti direttamente sull'account collegato [6]. Un sistema di questo tipo è concepito per la raccolta di innumerevoli dati sul comportamento dei singoli consumatori di fronte agli scaffali, che possono essere elaborati tramite *analitiche* (glossario) e utilizzati all'interno di altri prodotti. Tutti questi dati, una volta estratti, possono anche essere venduti, scambiati e integrati con quelli di altre aziende partner.

La psicologa sociale S. Zuboff chiama questo scenario *mercato dei behavioural futures,* per sottolineare quanto l'accumulazione sia mirata alla comprensione del comportamento e della socialità delle persone, caratteristiche che possono così essere sfruttate per accrescere il coinvolgimento degli utenti o la fedeltà verso un marchio [7].

Gli snodi di questa rete di scambio sono aziende chiamate *data broker*, che operano raccogliendo e acquistando dati da diverse fonti, processandoli e vendendoli ai propri clienti. Il potere, sotto forma di conoscenza, in grembo a queste aziende, le rende ideali per lanciare campagne pubblicitarie o per avviare macchine di propaganda che possono avere rilevanti conseguenze sul tessuto sociale e non si limitano quindi ad influenzare la vita online dei consumatori.

2. Le fonti dei dati

Il business dei data broker si regge sull'aggregazione di grandi moli di dati, chiamati *big data***,** che possono essere ottenuti con facilità monitorando le tracce digitali degli individui.

Il modello di business più comune per la raccolta di dati è quello basato sull'offerta di servizi o prodotti gratuiti, che essendo utilizzati da milioni di persone, consentono ai fornitori degli stessi di ottenere un enorme quantità di dati comportamentali, spesso raccolti senza un consenso informato da parte degli utenti [7]. Non sempre però tali informazioni provengono esclusivamente dal tracciamento online. Una relazione del 2014 della Federal Trade Commission [8], che ha analizzato le pratiche di raccolta e utilizzo dei dati di 9 data broker, ha catalogato le fonti di dati in 3 categorie: fonti governative, commerciali e pubbliche.

È riportato che, fra le *fonti governative*, l'U.S. Census Bureau forniva informazioni demografiche su quartieri urbani, tra cui informazioni sulle etnie, età, occupazione e livello educativo degli abitanti, mentre amministrazioni locali e tribunali fornivano informazioni sulle licenze professionali, proprietà, ipoteche e fascicoli giudiziari.

Le *fonti pubbliche* includono blog e social network, fra cui LinkedIn, dove sono raccolte informazioni di utenti che hanno impostato bassi livelli di protezione del proprio profilo. I dati raccolti includono potenzialmente tutto ciò che è visibile pubblicamente, inclusi post, immagini, e condivisioni.

Le *sorgenti commerciali* sono aziende che vendono dati ai data broker, e fra queste si annoverano case editrici, che forniscono informazioni sugli abbonamenti dei propri lettori, istituti finanziari, negozi, che forniscono informazioni più o meno specifiche sugli acquisti effettuati, come la categoria dell'oggetto acquistato (es. scarpe di alta moda, cibo per cani, prodotti sanitari), l'importo pagato e il metodo di pagamento. Infine, in molti casi, diverse informazioni sono acquistate da altri data broker, che possono averle ottenute direttamente o indirettamente da fonti terze. Ad esempio, IDMERIT, un'azienda attiva nel business dei dati, dichiara di avere a disposizione informazioni su utenti in oltre 90 paesi, provenienti da governi, liste elettorali, registri delle tasse, uffici anagrafe, servizi postali e aziende di telecomunicazioni [9].

2.1 Un esempio: i dati di Acxiom

La maggioranza dei data broker, fornendo l'accesso ai dati come un servizio, rendono possibile accedervi tramite *richieste HTTP* (glossario), ed espongono delle interfacce (API) per integrarle nelle applicazioni software. Una di queste è Acxiom, che offre molteplici servizi, fra cui risoluzione dell'identità e servizi cloud di analisi dei dati (analitiche). L'azienda dichiara di possedere fino a 10'000 attributi per ogni individuo, con una copertura di 2.5 miliardi di persone in oltre 60 paesi, fra cui l'Italia [10]. Collegandosi al portale per sviluppatori è possibile sfogliare pagine di documentazione che riportano quali attributi possono essere disponibili per ogni consumatore [11]. Si tratta di informazioni di ogni tipo, da quelle demografiche, a dati sulle abitazioni, come date di costruzione e insediamento, proprietà, valore catastale, e anche dettagli riguardo all'edificio, come categoria urbanistica (appartamento, villa, ecc.), materiale di costruzione, il numero di stanze da letto, tipologia di riscaldamento, metratura e molto altro.

Tra i pacchetti di dati disponibili vi è anche quello dedicato ai dati finanziari, che contiene informazioni riguardo ad entrate annue, propensione agli investimenti, situazione finanziaria, frequenza di utilizzo della carta di credito, mentre in altri pacchetti sono indicate informazioni sull'affiliazione politica, religiosa, l'appartenenza etnica, i canali di acquisto preferiti e le abitudini di shopping. I valori degli attributi possono essere cifre o categorie precise, intervalli di valori e in alcuni casi categorie probabilistiche (es. "likely", "very likely").

Acxiom offre la possibilità di ottenere un *portrait*, ovvero un ritratto di un consumatore arricchito con tutte le informazioni sopra citate in possesso dell'azienda. È anche specificato che in mancanza di alcune informazioni, i campi sono riempiti con i valori statisticamente più frequenti in base ai parametri di input forniti. Pertanto il "ritratto" fornito non sarà sempre preciso, ed in alcuni casi costituisce una rappresentazione statistica approssimativa dell'individuo [12].

Va puntualizzato che tale documentazione descrive un'interfaccia standard, e non vi è quindi garanzia che tali attributi siano disponibili per tutti i 2.5 miliardi di consumatori profilati da Acxiom.

3. La trasformazione dei dati

I dati raccolti e venduti senza apportarvi modifiche sono chiamati *dati grezzi*. Questi possono essere lavorati per estrarne informazioni raffinate da incorporare in prodotti venduti dall'azienda. Ad esempio, se una persona visita spesso un sito di vendita di scarpe ed è abbonata a riviste di alta moda, può essere dedotto il suo interesse per le scarpe di alta moda. Attributi di questo tipo, inferiti a partire da quelli aggregati dall'azienda, sono definiti *dati derivati*.

Nelle piattaforme di *segmentazione dei consumatori* i dati, sia grezzi che derivati, sono organizzati per creare *segmenti di dati*, ovvero categorie che raggruppano i consumatori con precise caratteristiche. Questi segmenti possono essere prodotti definendo funzioni di similarità fra consumatori utilizzabili all'interno di algoritmi di *apprendimento non supervisionato* (glossario). In secondo luogo è possibile estrarre un sottoinsieme di consumatori che hanno acquistato una certa categoria di prodotti, e utilizzare le caratteristiche in comune per costruire un modello predittivo da applicare a nuovi consumatori al fine di determinare la probabilità che acquistino quella tipologia di prodotti, avvalendosi di algoritmi di *apprendimento supervisionato* (glossario) [8].

I segmenti di dati dividono quindi i consumatori in gruppi in base alle necessità dei clienti dei data broker. Nello studio della FTC, viene evidenziato come alcuni segmenti siano focalizzati su minoranze o sullo stato socio-economico: ad esempio la categoria "Resolute Renters" è descritta come composta da "consumatori tra i 30 e 40 anni, single e senza figli, frequenti affittuari e al livello più basso di guadagni e peso sociale", oppure "Tradition & Timecards" composta da "lavoratori di età media 53 anni, che rappresentano gli ispanici meno acculturati che risiedono in aree urbane" [8]. Altre categorie sono "Credit Worthiness", "Allergy sufferer" e "Financially Challenged". Acxiom offre un servizio di segmentazione dei consumatori chiamato Personicx, che include categorie come "Work & Causes", che nello studio della FTC era descritta come "Persone con bassi guadagni tra i 45-55 anni, che vivono in unità abitative multi familiari", "Rural Parents" e "Metro Strivers", mentre fra i segmenti riguardanti lo status finanziario compaiono "Comfort Zone", "Online Influencers" e "Budget Optimists" [13].

Viene altresì specificato che nel costruire i modelli per l'assegnazione delle categorie finanziarie di Acxiom non vengono utilizzati dati direttamente riguardanti età, sesso, etnia, colore della pelle, religione, origini geografiche e stato coniugale [14].

4. Tipologie di prodotti venduti

I prodotti venduti dai data broker sono di tre tipologie principali: i prodotti di mitigazione dei rischi, ricerca di persone e marketing [8]. Alla prima categoria appartengono prodotti di prevenzione di frodi, come quelli di verifica dell'identità. Questi sono utilizzati per confermare l'identità di un individuo durante lo svolgimento di operazioni per le quali è richiesto accertarsi della stessa, come ad esempio transazioni bancarie.

In un recente articolo, Bloomberg stima a 12.8 miliardi di dollari il valore di questo business, proiettato nel 2024 [15]. Tra i maggiori fornitori di tali servizi, vengono citate le aziende Experian, Equifax, IDMERIT e TransUnion, dove alcune di queste vendono anche prodotti di altre tipologie. IDMERIT fornisce strumenti di verifica dell'email, del numero di telefono, dell'età, che i clienti possono utilizzare per verifiche in tempo reale delle informazioni inserite, anche per adempiere agli obblighi etici e legali in materia *know-your-customer*, come ad esempio i controlli di identità previsti dalla legge anti-terrorismo PATRIOT negli Stati Uniti [8]. Queste informazioni possono essere combinate con quelle ottenute dai prodotti di monitoraggio di biometriche comportamentali, come dinamiche di utilizzo del mouse, di digitazione da tastiera, l'inclinazione del dispositivo, geolocalizzazione, e in alcune situazioni anche riconoscimento vocale e analisi della pressione corporea [16]. Questi prodotti sono indicati per istituti finanziari, piattaforme di scambio di criptovalute, agenzie governative e piattaforme di gioco online. Altri prodotti di rilevamento frodi sono indirizzati al rilevamento di informazioni false inviate all'interno di form, per esempio riguardanti il proprio reddito [8, 17] o per il rilevamento di utilizzi non autorizzati delle carte bancarie agli sportelli ATM [18].

I prodotti per l'individuazione di persone consentono, con una ricerca, di aggregare le informazioni associate ad un individuo disponibili pubblicamente in milioni di siti web. Alcuni di questi prodotti sono utilizzabili pubblicamente come PeekYou [19], mentre altri sono rivolti ad un uso esclusivamente aziendale. TLOxp è un prodotto di questa categoria offerto da TransUnion. Offre numerose funzionalità e fornisce accesso a numeri di telefono, informazioni anagrafiche, dati riguardanti famiglia, lavoro e istruzione, fascicoli giudiziari, nickname, link a foto e post inseriti su social media, informazioni da siti di incontri e portali di e-commerce [20].

La terza e più redditizia categoria è quella dei prodotti per il marketing, il cui obiettivo è fornire ai clienti dei data broker, la possibilità di raggiungere i consumatori con messaggi pubblicitari mirati. Fra i servizi per il marketing diretto vi sono la vendita di *liste di marketing* (es. una lista di indirizzi di famiglie con bambini in una certa zona geografica) e il *data append*, attività che consente al cliente di integrare i dati di cui l'azienda è in possesso con le informazioni disponibili al data broker. Per esempio l'azienda NextMark vende liste di marketing di consumatori raggruppati in base a certe caratteristiche, in alcuni casi piuttosto sensibili: sono acquistabili liste di persone con disabilità [21], che soffrono di malattie mentali [22] o raggruppate in base ad attributi sulla situazione finanziaria [23] e età [24]. Si tratta di liste che contengono diversi milioni di elementi, la maggioranza provenienti da USA e Canada e reperite da sondaggi e database proprietari.

4.1 Il data onboarding

Una delle pratiche più diffuse finalizzata al marketing online viene definita *onboarding*, e consente di combinare dati online e offline, ovvero non relativi alle attività in rete, per ottenere una più completa conoscenza dei consumatori e rendendo possibile l'utilizzo dei dati per campagne *cross-channel* (glossario) personalizzate. Dati offline possono essere la lista di acquisti effettuata in un centro commerciale o la cronologia della posizione geografica.

L'applicazione più ovvia di questa strategia è la visualizzazione di annunci pubblicitari personalizzati, ma è anche possibile utilizzarla per la personalizzazione dei contenuti, come video e feed di notizie [25, 26].

Un cliente può quindi chiedere al data broker di trovare consumatori con particolari caratteristiche, per esempio indicando a quali segmenti di clienti desidera mostrare pubblicità personalizzata (per esempio "Allergy sufferer") o può fornire la propria lista di clienti a cui vorrebbe mostrare contenuti pubblicitari. Tali dati sono incrociati con quelli a disposizione del data broker fra cui identificatori online e identificativi di dispositivi posseduti dai consumatori (es. PC, tablet, smartphone).

I cookie sono esempi di identificatori online molto utilizzati, ma in realtà meccanismi di tracciamento più sofisticati possono risalire al proprietario di un dispositivo anche se i cookie sono stati cancellati o disabilitati. Sui dispositivi mobili vengono frequentemente usati gli *id pubblicitari*, forniti dai sistemi operativi mobili per questo preciso scopo, mentre altri identificatori sono l'*indirizzo IP*, l'*indirizzo MAC* (glossario) o molto semplicemente l'identificativo del proprio account se si effettua l'accesso ad un sito [27]. Una volta a conoscenza dei dispositivi utilizzati dai consumatori e in possesso dei loro dati online e offline, il data broker può mostrare contenuti pubblicitari per conto del cliente, e potrebbe fornire le informazioni identificative ai circuiti pubblicitari, solitamente privandole delle *personally identifiable information* (PII), come nome e indirizzo [28].

4.2 Software data-driven nei processi decisionali per il controllo sociale

Uno dei settori applicativi più recenti dell'economia dei dati è quello del *predictive policing*. L'obiettivo è quello di determinare le aree urbane a maggior rischio di criminalità e ottimizzare di conseguenza l'allocazione delle risorse per il controllo del territorio.

Software creati a questo scopo sono IBM Blue Crush [29], HunchLab [30] e PredPol. Quest'ultimo è in grado di effettuare previsioni del crimine su aree ampie 500 x 500 piedi (circa 150 metri) [31]. In particolare il software prevede la tipologia di crimine, data e orario, utilizzando algoritmi di apprendimento automatico [32] e assumendo che reati commessi in un determinato momento abbiano maggiori probabilità di verificarsi nello stesso luogo in futuro, o in luoghi molto vicini [33].

Una recente ricerca ha inoltre analizzato il funzionamento di Harm Assessment Risk Tool, un prodotto di *valutazione del rischio* sviluppato dall'Università di Cambridge in collaborazione con il dipartimento di polizia di Durham per supportare la polizia nelle decisioni di custodia [34].

HART utilizza un algoritmo di apprendimento automatico chiamato *random forest* che combina le predizioni di 509 *alberi di decisione* [35]. Un albero di decisione classifica un input sulla base della validità di una serie di regole if-then. Ogni nodo dell'albero verifica la validità di condizioni su un attributo dell'input, e ogni arco uscente da un nodo rappresenta un possibile valore dell'attributo (ad esempio per l'attributo "genere" i valori possono essere "maschio" o "femmina", per l'età si possono utilizzare intervalli di valori come "0-18", "19-30" e così via). Dalla radice, ogni esempio è classificato sulla base degli attributi, fino a raggiungere le foglie dell'albero, che fornisce in output la predizione finale per quell'esempio. Il modello utilizzato in HART classifica una persona colpevole di reato sulla base della probabilità che ne commetta un altro, assegnando una di 3 possibili etichette: "high risk" se prevede che la stessa persona commetta un nuovo crimine grave nei successivi 2 anni, "moderate risk" se prevede un crimine non grave nei successivi 2 anni e "low risk" se non prevede crimini nei prossimi 2 anni.
Nella tabella 1 sono indicati alcuni degli attributi utilizzati in HART per produrre la classificazione finale.

Attributo	Descrizione
CustodyAge	Età del soggetto alla commissione del reato
Gender	Maschio o femmina
InstantAnyOffenceCount	Conteggio dei crimini commessi fino al presente
InstantViolenceOffenceBinary	Un valore binario per indicare se il crimine ha natura violenta
InstantPropertyOffenceBinary	Un valore binario per indicare se si tratta di reato di proprietà
CustodyPostcodeOutwardTop24	I primi caratteri del codice postale dell'incriminato
CustodyMosaicCodeTop28	Uno dei 28 codici che descrive le caratteristiche socio-geo demografiche più comuni nella Contea di Durham
FirstAnyOffenceAge	L'età del soggetto al primo reato commesso

Tabella 1
Alcuni attributi utilizzati all'interno di HART

In particolare, secondo Big Brother Watch, l'attributo "CustodyMosaicCodeTop28" mostrato in tabella si riferisce ai dati forniti da Mosaic [34], un prodotto di Experian per la segmentazione del mercato che "classifica individui e famiglie in gruppi e tipologie dettagliate permettendo di raggiungerli con comunicazioni rilevanti" [36]. Le informazioni socio-demografiche generate da Mosaic includono categorie come "Disconnected Youth" [34], "Platinum Prosperity", "Tough Times" e molte altre [37].

5. Conclusioni

L'utilizzo di strumenti di predictive policing e di valutazione automatica del rischio solleva alcuni problemi di natura etica, riguardo alla capacità e all'opportunità di effettuare scelte così delicate tramite algoritmi, il cui funzionamento rimane in parte opaco.

Nonostante in Europa il regolamento GDPR vieti l'utilizzo non consensuale di questi strumenti per decisioni esclusivamente automatizzate che incidano significativamente sulla vita di persone [2, art. 22], resta comunque la possibilità che decisori umani siano influenzati dalle criticità di tali metodologie.

Talvolta vi è la errata percezione che un algoritmo ben progettato fornisca autonomamente una soluzione corretta e imparziale. Al contrario, il risultato ottenuto dipende in larga parte dalle capacità tecniche del personale, sia nella definizione dell'algoritmo, che nella conoscenza del dominio applicativo necessaria per una preparazione dei dati che garantisca risultati significativi e non ambigui [38].

Innanzitutto un algoritmo di apprendimento automatico non può considerare tutte le possibili relazioni fra i dati, ma ha bisogno di *assunzioni*, indicate con il nome di *bias algoritmico*, senza il quale l'apprendimento non sarebbe possibile. Quindi necessariamente l'algoritmo sarà sbilanciato a favore di alcune *ipotesi* (glossario) piuttosto che altre.

In secondo luogo, l'algoritmo impara una relazione presente all'interno dei dati che gli sono forniti durante la fase di allenamento, nella quale apprende una rappresentazione di essi utile a fornire output corretti. Se però all'interno di questi dati è incorporato uno stereotipo sociale discriminatorio, l'algoritmo potrebbe imparare a riprodurlo. A titolo di esempio, consideriamo un software per effettuare una pre-selezione dei curriculum vitae inviati ad una azienda per aggiudicarsi un posto di lavoro, e supponiamo che sia stato allenato sui dati dei dipendenti migliori dell'azienda, in modo da garantire una priorità di colloquio più alta ai candidati con maggiore affinità. Se i dipendenti aziendali sono per il 90% uomini, l'algoritmo potrebbe imparare un'ipotesi che discrimina fortemente i candidati di sesso femminile. In questo caso l'algoritmo ha individuato una *correlazione* fra individui di sesso maschile e migliori dipendenti e ha inferito che con alta probabilità il candidato ideale deve essere maschio, sebbene in realtà non ci sia alcun rapporto di *causalità* fra i due attributi.

Allo stesso modo nei prodotti di policing come PredPol, l'utilizzo di attributi quali il codice postale o il quartiere di residenza potrebbe alimentare un circolo vizioso, dove persone che vivono in aree ad "alto rischio" sono sottoposte ad un controllo più accurato, portando alla luce un maggior numero di reati rispetto ad altre aree [34], rinforzando quindi il condizionamento esistente. Andrew Ferguson, docente di legge all'Università di Washington D.C., sottolinea che mancano prove e ricerca scientifica sull'efficienza e l'efficacia dei sistemi di predictive policing, aggiungendo che è necessaria una validazione esterna di questi strumenti per comprendere il significato e la valenza della metodologia [31].

Più in generale, un rischio della raccolta intensiva e dell'utilizzo non regolato dei dati personali è la crescita smisurata dell'importanza della *digital footprint*, ovvero l'identità digitale di ciascuna persona. Questo può condurre all'adesione della società ad una cultura della conformità, nella quale gli individui, consapevoli del monitoraggio e dell'importanza del loro comportamento, lo modificano in funzione di quello considerato più corretto per i propri obiettivi. Per esempio è stato riportato che diverse università americane controllano i profili social dei candidati per decidere se accettarli [39]. Questo può essere un incentivo, per i futuri studenti, a cambiare le proprie abitudini online per apparire conformi al profilo considerato "migliore" per uno studente.

Tijmen Schep, un *privacy designer* e critico tecnologico, definisce questo effetto *social cooling*, che promuove l'abitudine di avversione al rischio e una maggiore rigidità e uniformità sociale, che soffoca le qualità individuali [40]. Questa tendenza è dovuta anche al trattamento statistico dei dati, che evidenzia il comportamento della maggioranza, ma penalizza le divergenze comportamentali, che diventano anomalie, più difficili da comprendere e da gestire.

Va tuttavia sottolineato che l'economia dei dati e i prodotti *data-driven* non hanno solo conseguenze negative: un prodotto di prevenzione di frodi potrebbe segnalare una transazione in corso come illegittima e bloccare un tentativo di furto di denaro, mentre la ricezione di un'offerta pubblicitaria personalizzata potrebbe non essere sempre sgradita. D'altro canto è necessario lo sviluppo di una maggior consapevolezza da parte dei consumatori del valore dei propri dati e la promozione di una sensibilità da parte delle aziende nel garantire un utilizzo corretto e sostenibile di queste risorse.

Glossario

Analitiche

Software per l'analisi dei dati che utilizzano modelli matematici al fine di effettuare predizioni, ottenere nuove informazioni e generare raccomandazioni. Strumenti di analitica possono analizzare i dati utilizzando tecniche di data mining, machine learning, pattern matching, sentiment analysis, analisi di reti e grafi, simulazione e reti neurali. [41, 42]

Apprendimento non supervisionato

Paradigma di apprendimento automatico, dove, dato un insieme di esempi, l'obiettivo è trovare regolarità e pattern che siano veri sull'intero dominio. In questo caso gli esempi non devono essere pre-classificati da un esperto umano, quindi in questo senso non è necessaria supervisione.

Apprendimento supervisionato

Paradigma di apprendimento automatico, nel quale un algoritmo, dato un insieme di dati pre-classificati, apprende una descrizione generale degli stessi che cattura il contenuto informativo utile per effettuare predizioni o elaborazioni su nuovi dati.

Targeting cross-channel

L'insieme di pratiche con cui è possibile raggiungere un particolare consumatore con messaggi di marketing personalizzati su tutti i suoi dispositivi (es. smartphone, tablet, PC) e servendosi di diversi canali (es. email, telefono, pubblicità online). [43]

Ipotesi (machine learning)

Termine con cui è indicata una funzione che un algoritmo di machine learning è in grado di apprendere. L'insieme di funzioni che l'algoritmo può imparare è definito *spazio delle ipotesi*. [44]

Richiesta HTTP

Una richiesta di informazioni inviata da un client (come un web browser) ad un server utilizzando il protocollo HTTP (Hyper Text Transfer Protocol). [45]

Indirizzo IP

Etichetta numerica assegnata ad ogni dispositivo connesso ad una rete di computer che utilizza il protocollo IP (Internet Protocol) per la comunicazione e lo scambio di informazioni. [46]

Indirizzo MAC

Il Media Access Control address (o indirizzo MAC) è un identificatore univoco assegnato ad un dispositivo connesso ad una rete ed è spesso è assegnato dal produttore dell'hardware. [47]

Bibliografia

[1] Press, G. (2017). 6 Predictions For The $203 Billion Big Data Analytics Market, *Forbes*, https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/#251c7ce32083 (ultimo accesso ottobre 2019).

[2] (2016). "Regolamento (UE) 2016/679 del Parlamento Europeo e del Consiglio", *Gazzetta ufficiale dell'Unione europea*.

[3] https://protezionedatipersonali.it/dato-personale (ultimo accesso ottobre 2019).

[4] https://www.indicative.com/blog/what-is-behavioral-data-and-behavioral-analytics/ (ultimo accesso ottobre 2019).

[5] https://conversionxl.com/conversion-optimization/mouse-tracking-and-heat-maps/ (ultimo accesso ottobre 2019).

[6] https://www.aboutamazon.it/innovazioni/amazon-go (ultimo accesso ottobre 2019).

[7] Naughton, J. (2019). "'The goal is to automate us': welcome to the age of surveillance capitalism", *The Guardian*, https://www.theguardian.com/technology/2019/jan/20/shoshana-zuboff-age-of-surveillance-capitalism-google-facebook (ultimo accesso ottobre 2019).

[8] Ramirez, E., Brill, J., Ohlhausen, M.K., D. Wright, J., McSweeny, T. (2014). "Data Brokers A Call for Transparency and Accountability", *Federal Trade Commission report*.

[9] https://www.idmerit.com/global-coverage/ (ultimo accesso ottobre 2019).

[10] https://www.acxiom.com/what-we-do/data/ (ultimo accesso ottobre 2019).

[11] https://developer.myacxiom.com/code/api/data-bundles/main (ultimo accesso ottobre 2019).

[12] https://developer.myacxiom.com/code/api/endpoints/portrait (ultimo accesso ottobre 2019).

[13] https://developer.myacxiom.com/code/api/data-bundles/level1bundle/demographics (ultimo accesso ottobre 2019).

[14] https://developer.myacxiom.com/code/api/data-bundles/bundle/personicxFinancial (ultimo accesso ottobre 2019).

[15] https://www.bloomberg.com/press-releases/2019-05-14/identity-verification-market-worth-12-8-billion-by-2024-exclusive-report-by-marketsandmarkets (ultimo accesso ottobre 2019).

[16] https://www.idmerit.com/behavior-monitoring/ (ultimo accesso ottobre 2019).

[17] https://www.dnb.com/resources/what-is-data-validation.html (ultimo accesso ottobre 2019).

[18] Rahman, M.M., Saha, A.R. (2019). "A Comparative Study and Performance Analysis of ATM Card Fraud Detection Techniques", *Journal of Information Security*, 10, 188-197, https://doi.org/10.4236/jis.2019.103011.

[19] https://www.peekyou.com/ (ultimo accesso ottobre 2019).

[20] https://www.tlo.com/social-media (ultimo accesso ottobre 2019).

[21] https://lists.nextmark.com/market?page=order/online/datacard&id=189752 (ultimo accesso ottobre 2019).

[22] https://lists.nextmark.com/market?page=order/online/datacard&id=243005 (ultimo accesso ottobre 2019).

[23] https://lists.nextmark.com/market?page=order/online/datacard&id=142394 (ultimo accesso ottobre 2019).

[24] https://lists.nextmark.com/market?page=order/online/datacard&id=315873 (ultimo accesso ottobre 2019).

[25] https://liveramp.com/blog/customer-data-onboarding/ (ultimo accesso ottobre 2019).

[26] https://www.braze.com/docs/help/best_practices/news_feed/ (ultimo accesso ottobre 2019).

[27] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-are-identifiers-and-related-factors/ (ultimo accesso ottobre 2019).

[28] https://www.lotame.com/back-basics-data-onboarding/ (ultimo accesso ottobre 2019).

[29] https://www.ibm.com/ibm/history/ibm100/us/en/icons/crimefighting/transform/ (ultimo accesso ottobre 2019).

[30] https://www.hunchlab.com/ (ultimo accesso ottobre 2019).

[31] Haskins, C. (2019). "Dozens of Cities Have Secretly Experimented With Predictive Policing Software", *Vice*, https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software (ultimo accesso ottobre 2019).

[32] https://www.predpol.com/about/ (ultimo accesso ottobre 2019).

[33] https://www.predpol.com/technology/ (ultimo accesso ottobre 2019).

[34] Scantamburlo, T., Charlesworth, A., Cristianini, N. (2019). "Machine Decisions and Human Consequences", *University of Bristol*.

[35] Oswald, M., Grace, J., Urwin, S., C. Barnes, G. (2018). "Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality", *Information & Communications Technology Law*, 27, 223-250.

[36] https://www.experian.co.uk/business/marketing/segmentation-targeting/mosaic/ (ultimo accesso ottobre 2019).

[37] https://www.experian.com/assets/marketing-services/brochures/mosaic-brochure.pdf (ultimo accesso ottobre 2019).

[38] Tsiptsis, K., Chorianopoulos, A. (2011). Data Mining Techniques in CRM: Inside Customer Segmentation, John Wiley & Sons.

[39] Moody, J. (2019). "Why Colleges Look at Students' Social Media", *U.S. News & World Report*, https://www.usnews.com/education/best-colleges/articles/2019-08-22/why-colleges-look-at-students-social-media-accounts (ultimo accesso ottobre 2019).

[40] https://www.socialcooling.com/ (ultimo accesso ottobre 2019).

[41] https://dictionary.cambridge.org/dictionary/english/analytics (ultimo accesso ottobre 2019).

[42] https://www.gartner.com/it-glossary/advanced-analytics/ (ultimo accesso ottobre 2019).

[43] https://www.oracle.com/marketingcloud/resources/cross-channel-marketing.html (ultimo accesso ottobre 2019).

[44] Mitchell, T. M. (1997). *Machine Learning*, McGraw-Hill Education.

[45] https://developer.mozilla.org/en-US/docs/Web/HTTP/Messages (ultimo accesso ottobre 2019).

[46] https://en.wikipedia.org/wiki/IP_address (ultimo accesso ottobre 2019).

[47] https://en.wikipedia.org/wiki/MAC_address (ultimo accesso ottobre 2019).

Biografie

Silvia Crafa è ricercatrice confermata presso il Dipartimento di Matematica dell'Università di Padova. La sua attività di ricerca in ambito informatico si focalizza sui metodi formali per l'analisi dei sistemi concorrenti e dei linguaggi di programmazione. Studia inoltre l'impatto sociale delle tecnologie digitali in ottica interdisciplinare. È autrice di numerose pubblicazioni su prestigiose riviste internazionali e collabora con diverse istituzioni straniere. È membro del Laboratorio Nazionale CINI su Informatica e Società ed è stato membro del working group Informatics Europe e EU-ACM per la definizione di un libro bianco sugli algoritmi di decisione automatica.

Email: silvia.crafa@unipd.it

Alessandro Zangari è studente del Corso di Laurea Magistrale in Informatica presso l'Università degli Studi di Padova. Ha conseguito la laurea triennale in Informatica nel 2018 e ha proseguito gli studi interessandosi all'ambito dell'Intelligenza Artificiale, del Machine Learning e di come queste tecnologie influenzino la vita umana.

Email: alessandro.zangari@studenti.unipd.it