Oltre i big data:
banche dati pubbliche e private alla prova del *deep learning*

Un paio d'anni fa, prenotando – come al solito all'ultimo momento – una vacanza con mia moglie ci siamo trovati, per varie ragioni, con le carte di credito bloccate o con credito insufficiente, per cui siamo arrivati in albergo senza sapere se la prenotazione fosse stata confermata o meno.
Per fortuna alla reception ci han detto che Booking aveva garantito per noi, essendo ottimi clienti, e che certamente avremmo risolto il problema all'arrivo – come infatti è stato – per cui è andato tutto liscio e ci siamo goduti due belle settimane in Costa Azzurra.
Ho voluto partire da questo episodio non per parlare delle mie vacanze (vi risparmio quindi le foto ed i filmini), bensì per commentare che la grande stima che Booking ha per noi è certamente influenzata dal fatto che mia moglie, avvocato, viaggia spesso all'estero e deve anticipare le spese di viaggio sostenute per conto dello Studio, inserendo quindi il rimborso nella fattura del mese successivo, dando pertanto l'impressione di un "tenore di vita vacanziero" decisamente superiore a quello che ci si può aspettare da una famiglia con il nostro reddito.
In realtà, anche nell'ipotesi che Booking avesse accesso ai dati fiscali e che il suo algoritmo sia così sofisticato da rendersi conto che la maggior parte dei nostri viaggi non sono di piacere bensì di lavoro, penso che non cambierebbe nulla rispetto al *rating* attribuitoci: tutto sommato, all'agenzia importa che uno viaggi molto, paghi puntualmente e non crei problemi con richieste particolari o lamentele continue, dubito che siano interessati alla provenienza ultima del denaro. Tuttavia, se queste informazioni e criteri di valutazione fossero inseriti in un sistema di analisi più esteso o fossero condivisi con altri, a partire dell'Agenzia delle Entrate, risulterebbe un'anomalia, facilmente spiegabile ad un addetto (ma è ragionevole che io debba giustificare a qualcuno quanto spendo in viaggi?), ma non immediatamente comprensibile per un sistema automatico.
Tutta questa premessa per far rilevare che stiamo costruendo sistemi sempre più complessi e capillari di raccolta e condivisione dei dati, utili per tante applicazioni, ma sempre più difficili da interpretare e sempre più a rischio di errori o fraintendimenti, che possono diventare via via più gravi quanto più gli algoritmi ed i sensori diventano diffusi, invasivi ed automatici.
Avere a disposizione molti dati non è automaticamente sinonimo di possedere molte informazioni utili!
Il fatto che i sistemi di analisi debbano essere automatici è scontato: si stanno generando e raccogliendo volumi di informazioni ingestibili dagli esseri umani. Le videocamere invadono le nostre città, ma è impossibile che ci sia h 24 una persona che le tenga tutte sotto controllo. O ci si accontenta di tenerle accese per deterrente (o per sfizio) e cancellare la ripresa quando tutto va bene e vedere a posteriori cos'è andato storto in caso di problemi, oppure saranno gli algoritmi a dover sovrintendere alla ripresa e decidere se e come intervenire.
Sono già in fase di sviluppo ed operativa più o meno avanzata sistemi che misurano lo stress, magari per impedire che un pilota si metta alla cloche sotto ricatto perché gli han rapito la famiglia, che rilevano bagagli abbandonati in aeroporti e stazioni, che osservano comportamenti sospetti dei passeggeri o dei clienti di un supermercato. Saranno anche utili per prevenire incidenti e reati, ma la loro diffusione rappresenta certo una grossa incognita sulle libertà personali e sulla spontaneità dei comportamenti. Anche perché un algoritmo che – opportunamente – impedisce di postare su Facebook foto di bambini nudi non sa distinguere una foto di alto valore storico ed emotivo[1], per cui, analogamente, dubito che questi software sappiano distinguere se il pilota è stressato perché gli hanno rapito il figlio per fargli schiantare l'aereo, oppure perché teme che la moglie abbia scoperto la tresca con

[1] http://www.ilsole24ore.com/art/mondo/2016-09-09/gaffe-facebook-foto-simbolo-vietnam--125657.shtml?uuid=ADdP9hHB

la hostess, senza particolari implicazioni sul volo. D'accordo, nel caso specifico come passeggero preferisco un falso positivo che un suicidio con me a bordo, ma non vorrei che in banca scattasse la procedura antirapina perché sono entrato un po' alterato dopo che il barbiere mi ha sbagliato il taglio di capelli (per quel che mi riguarda parlo in via molto ipotetica, ovviamente). D'altra parte il comportamento e lo stress sono anche molto soggettivi: ci sono persone che si emozionano facilmente e mostrano nervosismo per nulla ed altre che restano fredde anche davanti ad un evento drammatico. Inoltre alcuni eventi possono risultare stressanti per circostanze o novità, ma cessano di essere un problema quando diventano abituali: pensiamo allo stress di parlare in pubblico, che di solito è molto consistente alle prime esperienze e si annulla con la pratica. Immagino che un rapinatore seriale possa entrare in banca con un livello di stress decisamente più basso rispetto ad un neofita del crimine, per cui il punteggio che gli viene attribuito dal sistema può non essere così elevato da far scattare l'allarme. Questo perché, in ultima analisi, un software ragiona in termini matematici e può soltanto attivare o meno una determinata procedura sulla base di un numero calcolato attraverso i dati di *input* e l'algoritmo di analisi. Un software preposto a fare scattare un allarme, a lanciare un missile o a far evacuare un aeroporto lo fa se l'insieme degli i*nput*, pesati secondo delle regole imposte, gli fanno raggiungere una determinata soglia. Al massimo ci possono essere più soglie, ad esempio un software che controlla i passeggeri in una stazione della metro può avere una prima soglia di normalità, una soglia di attenzione, superata la quale richiede una valutazione umana, una soglia di richiesta d'intervento delle forze dell'ordine ed una soglia di evacuazione immediata, ma comunque tutto si riduce ad un valore numerico, ad un punteggio.
Esattamente come Booking ha valutato che siamo ottimi clienti perché evidentemente abbiamo superato una determinata soglia di spesa, di pagamenti precisi, di mancate lamentele. Almeno è quanto immagino, perché quali parametri vengano considerati e con quali pesi vengano computati non è dato saperlo; questo è un aspetto che trovo particolarmente inquietante, sul quale vorrei tornare alla fine.
Sinora però abbiamo esaminato il caso di una valutazione singola, puntuale: Booking si interessa alle mie vacanze, il software di controllo della metro mi prende in carico da quando entro in una stazione a quando esco da un'altra, un supermercato verifica che non commetta danni o taccheggi (oppure analizza il mio comportamento a scopi di marketing), ma si disinteressano a quel che accade prima o dopo, a chi sono, a quali sono i miei pensieri ed a tutto ciò che non è direttamente legato alle opportunità di incrementare le vendite, di mantenere la sicurezza o di propormi le offerte più adatte al mio modello di vacanza. In questo senso un algoritmo "classico", basato su istruzioni matematiche più o meno sofisticate, ma comunque definite e ripetibili, è più che sufficiente.
Già i social ed i motori di ricerca devono svolgere una profilazione più profonda, e d'altra parte hanno anche più dati sui quali lavorare. Ma come ha fatto Facebook a capire che una foto ritraeva un bambino nudo? Come fa Google Street View a cancellare automaticamente i volti delle persone (e delle statue!)? Mentre i pur sbandierati sistemi di analisi neurale sono ancora ai primi passi, se non decisamente da sviluppare, un nuovo metodo di analisi sta dando i primi frutti applicativi: il *deep learning*, che sta entrando in una fase applicativa o comunque di sperimentazione avanzata. In pratica, non si installa nel computer un algoritmo fisso, bensì il programma contiene una serie di informazioni e di strumenti d'analisi di base, poi impara da solo, migliorandosi con l'esperienza, come avviene per gli esseri umani. Come facciamo ad intuire l'età di una persona? Avendo visto centinaia di visi, di figure ed associandole ad un'età in base ad una serie di dettagli, spesso neppure percepiti in modo cosciente. Naturalmente quando qualcuno esce da questi schemi sbagliamo la valutazione. Oppure riusciamo a capire la funzione di un oggetto per similitudine e per contesto: se andiamo a casa di qualcuno e ci offrono da bere, sappiamo riconoscere un bicchiere anche se quelli di casa nostra sono diversi per forma e per colore. Si sta cercando di replicare questo approccio anche nei computer, sviluppando sistemi che apprendano attraverso lo

studio di centinaia o migliaia di esempi e che possano quindi effettuare analisi non puramente basate su una soglia fissa, bensì su un insieme più ampio (ed aleatorio) di parametri. Per far capire che un oggetto è una chiave, si fanno passare migliaia di immagini diverse, spiegando quali contengono la chiave e quali no; terminato l'apprendimento (che in realtà non termina mai, migliorandosi sempre) il computer potrà riconoscere le chiavi anche in contesti mai visti oppure se hanno forme non identiche a quelle esaminate in precedenza. Avrà un comportamento simile a quello di un essere umano: un bravo agente della sicurezza, infatti, è in grado di cogliere piccoli segnali, sguardi, comportamenti anomali, scremandoli dai comportamenti innocui di gente magari strana ma non animata da cattive intenzioni. Se e quando un software riuscirà a farlo, è tutto da dimostrare, ma intanto molti istituiti di ricerca ci stanno lavorando. Peraltro, se ci riusciranno, otterranno software che sbaglieranno esattamente come un essere umano.

Mentre si sta estendendo la raccolta dei dati, si stanno accumulano quante più informazioni possibili e si collegano fra di loro in rete, condivise a livello di Stato, di Europa, di alleati, sta nascendo anche un modo diverso di sfruttare questa massa di informazioni. Di ogni cittadino sapremo cosa consuma al supermercato, quali auto guida, cosa gli piace leggere, che siti guarda, quando e con chi viaggia, e ne sapremo magari predire il comportamento. La prima domanda è: serve a debellare il terrorismo? Peraltro ho trovato il titolo del convegno – non me ne voglia Marco – un po' limitativo. E' vero che il terrorismo più di altre attività illecite si nutre di Internet: reclutamento, auto-indottrinamento, campagne mediatiche, comunicazione fra cellule e mandanti. E' vero che, per definizione, esso ha il proposito di portare il terrore, costringendoci quindi a cambiare il nostro stile di vita, ma, da un punto di vista statistico, ha un impatto tutto sommato risibile sul numero di morti o sulla probabilità di essere coinvolti: nel 2014, in tutto il mondo, 32.658[2] vittime per terrorismo, contro 1.250.000[3] per incidenti stradali. In certe zone di certe città non ci si può andare perché si viene rapinati, non per il timore della bomba di un estremista. A casa ho paura che entrino i ladri, non uno jihadista. Nelle terre di mafia e camorra un kamikaze è l'ultimo dei problemi. Quindi per me il punto centrale non è tanto di combattere il solo terrorismo, che la storia italiana ha dimostrato essere tutto sommato più facile da estirpare che non la criminalità organizzata, quanto di capire se e come la raccolta più o meno indiscriminata di dati può essere davvero d'aiuto nella lotta contro la criminalità, di qualunque matrice. In Europa c'è ora un registro dei passeggeri che volano sui nostri aerei[4]. Secondo il direttore della nostra intelligence è uno strumento utile, anche perché alcuni terroristi, per non dare nell'occhio, si siedono accanto sull'aereo in modo da potersi parlare senza dare l'impressione di un incontro concordato[5]. Lui ne sa certamente più di me, ma, a parte il fatto che un aereo di linea non mi sembra propriamente il luogo ideale per complottare senza che nessuno senta, mi viene il dubbio che ora detti individui si incontrino su un pullman, su un treno o su un traghetto, vanificando buona parte del lavoro svolto da questo registro. Forse è più utile all'ufficio imposte, per verificare se il reddito dichiarato è in linea con il tenore di vita.

Ma immaginiamo un qualcosa di ancora più esteso. Dopo un attentato, ad esempio una bomba nella metro, si scopre sempre, a posteriori, che gli attentatori avevano fatto dei sopralluoghi per pianificare l'attacco. Potrà anche servire ad individuare ulteriori complici, ma si tratta di un'analisi tutto sommato sterile, in quanto il fatto è già avvenuto. Se il sistema fosse così esteso e capillare da analizzare il comportamento non solo all'interno della metro ma da quando uno esce di casa a quando ci ritorna, si potrebbero prevenire questi attentati?

---

[2] http://www.lastampa.it/2015/11/24/esteri/le-vittime-del-terrorismo-nel-mondo-in-un-anno-sono-aumentate-dell-per-cento-tWAuzG3rtr67zu3mc5JifM/pagina.html

[3] http://www.who.int/gho/road_safety/mortality/traffic_deaths_number/en/

[4] http://www.futuro-europa.it/20281/europa/pnr-registro-passeggeri-europei.html

[5] http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/4612330

I nuovi sistemi di *deep learning* o gli ancor più nuovi software di analisi neurale sarebbero più efficaci? Probabilmente no, perché uno che fa due passi, entra nella metro, va alla moschea e torna a casa non mostra un comportamento sospetto. L'attentatore di Nizza abitava lì, più che fare sopralluoghi andava e veniva da casa sua. Difficilmente un software anche sofisticato potrà distinguere chi viaggia con la testa da un'altra parte da chi osserva il territorio con l'occhio del soldato.
Però sistemi di analisi evoluta possono attirare tanto i privati quanto lo Stato, che ha l'interesse e la necessità di raccogliere ed analizzare dati, al fine di individuare le tendenze, pianificare le risorse e gli investimenti, oltre a prevenire crimini ed attacchi esterni. Dalle più o meno banali tendenze demografiche (se la popolazione invecchia si dovranno costruire meno asili e più case di riposo) ai più complessi flussi migratori legati a guerre, carestie, mutamenti climatici o al semplice mutare delle aspirazioni e delle opportunità fornite dalle nuove tecnologie ed alle fluttuazioni del mercato del lavoro, la necessità di analisi previsionale sono infinite.
I primi passi in questa direzione provengono dall'ANPR, la nuova Anagrafe della Popolazione Residente che raccoglierà (con almeno un anno di ritardo) tutti i dati anagrafici dei residenti in Italia (italiani o stranieri) e di tutti gli italiani residenti all'estero. Il passo successivo è di associare a queste schede anagrafiche altri dati riguardanti il cittadino: il suo nucleo familiare, lo stato di salute, le abitazioni, i mezzi di trasporto, i titoli di studio, la posizione lavorativa e previdenziale, come illustrato sul sito dedicato alla nuova carta d'identità elettronica, recentemente aggiornato[6].
A parte il fatto che, con l'occhio del demografico, si nota che mancano i dati elettorali e di leva, c'è anche un progetto già definito per includere anche gli atti dello stato civile, con grande sollievo di chi nasce a Palermo si sposa a Roma e quando va a vivere a Milano non deve scrivere al Comune o mandare amici e parenti nel luogo dell'evento per avere un certificato (almeno per i nuovi eventi, per lo storico si vedrà).
D'altra parte l'Agenzia delle Entrate già ha assorbito catasto urbano e conservatoria, già il vecchio sistema INA-SAIA consentiva (più o meno) di condividere i dati anagrafici fra Enti convenzionati, il nuovo intreccio delle banche dati è la naturale conseguenza della possibilità di costruire algoritmi sofisticati per comprendere più in profondità il comportamento e le tendenze dei singoli e delle masse. Certo, sapere con dieci anni d'anticipo che una zona andrà svuotandosi mentre un'altra sarà in crescita consente di progettare abitazioni ed infrastrutture in modo adeguato; incrociare i dati di reddito con i dati della motorizzazione, del catasto e magari dell'agenzia viaggi consente di scovare gli evasori e magari anche i prestanome, mettere in rete i dati giudiziari dei vari procedimenti consente ad un giudice di valutare adeguatamente quello che può essere un piccolo reato se preso singolarmente, ma diventare un problema serio se ripetuto molte volte sul territorio.
In questo senso i sistemi di *deep learning* potranno dare certamente un contributo. Essi infatti funzionano bene (o meglio, funzioneranno bene, perché al momento sono quantomeno acerbi) proprio quando si tratta di applicare valutazioni puntuali e flessibili su una quantità immensa e piuttosto fluida di dati, come può essere appunto un sistema di raccolta pubblico.

Ma torniamo allora alla domanda di prima: è davvero utile questa raccolta ed analisi dei dati per combattere la criminalità? In parte certamente sì, facendo pesca a strascico qualcosa prima o poi si prende per forza, quindi la domanda giusta è: lo sforzo umano ed economico per effettuare la raccolta dati - e le conseguenti rinunce in termini di privacy e libertà personali - valgono questo sforzo? In termini più giuridici: stiamo rispettando il principio di proporzionalità?
Questo è decisamente più difficile da sostenere, anche perché molto spesso gli attentatori

[6] http://www.cartaidentita.interno.gov.it/servizi-italia-it/

erano già persone note ai servizi di sicurezza[7], a volte addirittura denunciate, magari dai loro stessi familiari, e lasciate libere di agire, anche perché in democrazia c'è questo fastidioso limite di non poter arrestare nessuno che non abbia (ancora) commesso reati.
Certo, un'analisi automatica basata su tecniche di *deep learning*, quando saranno sufficientemente evolute, potrebbe aiutare appunto nell'evidenziare comportamenti anomali che si possono rilevare solo unendo un'analisi estesa del comportamento ad un minimo di intuito, dote che sinora non ha contraddistinto gli elaboratori. Rilevare che una persona che va a lavorare con la linea blu e raggiunge il bar degli amici con la linea rossa non ha ragioni di prendere la linea verde solo per osservare le stazioni e tornare indietro può prevenire un attentato, ma può anche rendere sospetto un giovane che voleva incontrare una ragazza, ma per timidezza ha rinunciato all'ultimo momento.
Vogliamo vivere in una società che ci osserva in ogni momento ed analizza ogni nostro comportamento, valutandolo secondo criteri sconosciuti ai più? Vogliamo che il nostro modo di agire non sia spontaneo, ma subordinato al rientrare in determinati parametri, magari neppure chiarissimi?
Dicevo all'inizio che posso solo immaginare in parte quali criteri abbia usato Booking per attribuirmi un rating, ma l'algoritmo ed i parametri reali sono noti solo ai vertici dell'azienda. E mi resta pure il dubbio che algoritmi complessi, come quelli di Facebook o Google, siano oramai così complicati da sfuggire persino agli stessi programmatori[8]. Col *deep learning* il problema sarà ancora più esteso: a fronte di un algoritmo di base identico due calcolatori diversi acquisiranno esperienze diverse, per cui a fronte della stessa situazione potranno arrivare a conclusioni diverse, riproducendo peraltro l'imprevedibilità del comportamento umano che porta a superare la frontiera indenne o meno a seconda del doganiere che trovi.
Per inciso, il fatto che finora un calcolatore operasse in modo rigido e prevedibile comportava vantaggi e svantaggi, esattamente come comportava vantaggi e svantaggi l'essere umano, la cui imprevedibilità si accompagnava all'intelligenza. Speriamo che i nuovi sistemi massimizzino i vantaggi, combinando lo stakanovismo di un calcolatore con l'intelligenza umana, piuttosto che massimizzando i problemi, ovvero l'imprevedibilità umana con la stupidità di un computer. Rischiamo dunque di trovarci con una società pesantemente controllata, con sistemi che valutano il nostro comportamento, magari anche con il lodevole intento di prevenire il crimine, ma che giudicano secondo parametri non del tutto conosciuti e di dubbia affidabilità. In passate edizioni avevo sempre sostenuto l'uso dell'informatica ai fini di giustizia, ma avevo anche sostenuto la necessità che dette applicazioni siano impiegate in modo controllato. E, sarò retrogrado, ma preferirei che il controllo ultimo avvenga ad opera di un essere umano, come peraltro prescritto dal nuovo Regolamento europeo per il trattamento dei dati personali[9].
Con questo non voglio dire che la ricerca debba rinunciare a sviluppare i nuovi metodi ed il nuovo modo di approcciare le analisi. Anche perché c'è sempre la possibilità che il nemico faccia lo stesso, ovvero che un terrorista applichi tecniche di *deep learning* per individuare autocivetta, agenti sotto copertura, per cui la conoscenza dei sistemi può certamente dare qualche vantaggio strategico.
Neppure voglio dire che questo tipo di analisi verrà davvero utilizzata dallo Stato, stavo solo ipotizzando un futuro possibile, neppure troppo futuribile.
Voglio dire che, come tutto ciò che riguarda i rapporti fra cittadini e Pubbliche Amministrazioni, questi sistemi devono essere usati solo se effettivamente utili ed efficaci, se non c'è un altro sistema meno invasivo, e soprattutto se possono essere impiegati in modo trasparente, sotto il controllo democratico e popolare.

---

[7] www.garanteprivacy.it, ibidem
[8] cfr. http://www.lescienze.it/news/2016/10/22/news/apprendimento_profondo_scatola_nera_problemi_spiegazione-3282154/?ref=nl-Le-Scienze_28-10-2016&refresh_ce
[9] Punto (71) delle premesse