Grandi dati per piccoli bambini: intervista a Fosca Giannotti, scienziata di BigData

Fosca Giannotti "è una scienziata pioniera nel data mining orientato alla comprensione delle dinamiche della mobilità, nell'analisi dei social network, nell'elaborazione di tecniche di data mining che garantiscono la privacy. Ha coordinato decine di progetti europei e collaborazioni industriali. Oggi si occupa tra l'altro di SoBigData, l'infrastruttura di ricerca europea dedicata a BigData Analytics e Social Mining". Fosca è l'unica scienziata italiana e una delle sole tre europee della lista delle ricercatrici e innovatrici più influenti del settore: Le InspiringWomen in AI, BigData, Data Science, Machine Learning.
L'ECCE (Early Childhood Care and Education) è un programma dell'UNESCO dedicato alla cura dell'infanzia e dell'educazione di base che si inserisce come primo obiettivo dell'EFA (Education For All).
All'interno del progetto ECCE è stato creato nel 2010 un Comitato tecnico specifico con l'intento di superare la frammentazione esistente riguardante la cura della prima infanzia, in favore di una visione olistica del monitoraggio sull'educazione di base. Una delle priorità è di cercare di gestire l'enorme mole di dati attraverso tutti i componenti di sistema dell'ECCE.
La gestione, condivisione e produzione dei dati provenienti dai vari Paesi è alla base di una politica basata su decisioni operative e strategiche. Nazioni con una maggiore cultura relativa alla raccolta, gestione e utilizzo dei BigData, come la Svezia, sono riuscite, proprio attraverso i BigData a implementare il migliore sistema ECCE.
Per Kevin Maher, direttore della terapia intensiva pediatrica cardiologica dell'Ospedale di Atlanta i BigData in pediatria sono semplicemente l'evoluzione naturale del registro dati elettronico, una grande opportunità per poter estrarre valore dai dati che le istituzioni già normalmente raccolgono per curare meglio e con migliori outcome. L'analisi dei BigData può aiutare in maniera significativa i team medici nell'identificazione dei pazienti a rischio per eventi avversi significativi e per "acchiappare" gli indizi tranquilli che possono segnalarci in tempo l'avvento di problemi seri.
Nel nostro lavoro in terapia intensiva vediamo bambini in arresto e non posso non pensare che non ci siano stati indizi e informazioni 4-5 o 8 ore prima che ci potevano segnalare che il percorso del paziente stava prendendo la direzione sbagliata. Questo è il potenziale dei BigData: si possono associare un numero massivo e imponente di informazioni in millisecondi per poter assicurare la migliore cura.
Wikipedia ci offre la seguente definizione di BigData: "Il termine BigData ('grandi [masse di] dati' in inglese) indica genericamente una raccolta di dati così estesa in termini di volume, velocità e varietà da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore o conoscenza. Il termine è utilizzato in riferimento alla capacità (propria della scienza dei dati) di estrapolare, analizzare e mettere in relazione un'enorme mole di dati eterogenei, strutturati e non strutturati, allo scopo di scoprire i legami tra fenomeni diversi e prevedere quelli futuri".
Chiediamo a Fosca Giannotti di darci alcuni imput per una migliore comprensione di questa definizione.

I BigData come tracce digitali prodotte dalla pervasività delle tecnologie digitali in tutti i settori della società...

Nel mondo che abitiamo abbiamo l'opportunità di osservare da vicino e misurare il funzionamento della società attraverso i BigData, le briciole digitali che le nostre attività quotidiane lasciano per effetto del nostro uso dei sistemi ICT (Information Communication Technology). Briciole che registrano la nuda verità sui comportamenti individuali e collettivi con una precisione senza precedenti, in modo che le diverse dimensioni della nostra vita sociale trovano un'immagine riflessa nello specchio digitale: desideri, opinioni, stili di vita, movimenti, relazioni.
I nostri desideri, opinioni, sentimenti lasciano traccia nei social media a cui partecipiamo, nelle domande che facciamo ai motori di ricerca, nei tweet che inviamo e riceviamo, così come i nostri stili di vita lasciano traccia nei record dei nostri acquisti. I nostri movimenti lasciano traccia nelle traiettorie disegnate dai nostri smartphone e dai sistemi di navigazione delle nostre auto.
Anche le nostre relazioni sociali lasciano traccia nella rete dei nostri contatti telefonici e delle email e nei link di amicizia del social network preferito. Possiamo cominciare a esplorare la rete di relazioni che costituisce la nostra società, il tessuto sociale e la sua robustezza o debolezza.
I BigData sono il nuovo microscopio che rende "misurabile" la società. Come la scoperta di ogni nuovo microscopio o telescopio nel passato, i BigData stanno spingendo verso una nuova scienza dei dati o anche "data science", in grado di misurare e, in prospettiva, prevedere crisi economiche, epidemie e pandemie, diffusione di opinioni, distribuzione delle risorse economiche o energetiche, bisogni di mobilità.
Un'altra sorgente importante sono i cosiddetti "open data", cioè dati resi accessibili da parte delle pubbliche amministrazioni o organizzazioni varie. Buoni strumenti di "crawling" su rete possono realizzare velocemente raccolte massive di questi dati, oppure i servizi stessi mettono a disposizione delle funzionalità (API) per scaricarli.
Ma la grande pervasività di dati non servirebbe a niente se non fosse accompagnata dalla potenza di sintesi e trasformazione in conoscenza che i metodi di analisi mettono in gioco: algoritmi che riescono a raggruppare clienti con comportamenti simili o riconoscere moduli di proteine con funzioni simili nella rete biologica nelle interazioni tra proteine; algoritmi che riescono, dai molti esempi forniti dai dati, a imparare regole generali e modelli da usare per classificare come fraudolenta una pratica di rimborso, o un cliente come particolarmente proficuo, un paziente ammalato di una specifica patologia, un testo che esprime un sentimento positivo, un'immagine che rappresenta un gatto; algoritmi, che osservano e comprendono la dinamica delle opinioni; algoritmi che, concentrandosi sulle inter-

connessioni tra le entità, rivelano la complessità delle relazioni sociali, dei sistemi economici, delle reti biologiche e permettono di comprendere la dinamica che sta dietro alla diffusione delle opinioni o delle epidemie, o i meccanismi che stanno dietro a patologie complesse come tumori o disordini metabolici e far emergere le interconnessioni nascoste.
Focalizziamoci ora sui quattro attributi che caratterizzano i BigData e che identificano delle aree di criticità su cui si concentrano ricerca e tecnologia in questo ambito: Volume, Velocità, Varietà e anche Veracità (le 4 V). Volume si riferisce alle dimensioni di tali dati che per essere maneggiati hanno bisogno di tecniche di filtro e/o compressione, di stoccaggio e di elaborazione particolari. Le misure di riferimento attuali sono sulla scala sugli esabyte (1000^6) zetta-byte (1000^7), ma la cosa più importante è che stanno crescendo più della capacità di elaborarli. Velocità si riferisce alla dinamicità della produzione dei dati, per esempio le discussioni prodotte sui social media, o i dati azionari. Varietà si riferisce alle molteplici forme che possono avere i dati, non più tabelle strutturate, ma testo, video, audio, e combinazioni di queste forme. Infine Veracità si riferisce alla qualità del dato e alla sua affidabilità.
Al di là degli aspetti tecnici, lo stimolo che le varie discipline dovrebbero cogliere dai BigData è quello di usare quanto hanno raccolto nella loro massima dimensione storica, ma anche cogliere la sfida di integrarli con altre sorgenti dati, sia provenienti da altre parti delle loro organizzazioni, ma anche da quanto è disponibile esternamente.

Se i BigData sono in effetti, come citato da tutti gli esperti, l'"internet delle cose", nel senso che gli oggetti nel mondo tecnologico si interconnettono tra di loro e si relazionano, questo concetto trasposto nella pratica clinica e a livello politico-sanitario in che modo può essere utile ai pazienti?
Internet delle cose (IOT) è un ecosistema di tecnologie digitali che permette la realizzazione di ambienti di vita immersi da oggetti capaci di sentire e interagire con il mondo circostante (per esempio pazienti e/o medici) e uno degli effetti è la produzione di molti dati su queste interazioni e quindi l'abilitazione processi di estrazione di conoscenza che permettono di costruire un numero maggiore di osservazioni. L'esempio più elementare di IOT è il "fitbit" che interagisce con l'organismo e raccoglie parametri fisici elementari, e li restituisce all'utente in una forma utile, fino alla raccomandazione di stili di vita salutari. Se poi una "centrale" raccoglie i dati di tutti i "fitbit" può cominciare a osservare una popolazione e la sua salute, questo può servire al politico-sanitario. Sempre più i calciatori usano "fasce con sensori" durante l'allenamento e un'appropriata sintesi quantitativa (data da quei metodi analitici di cui abbiamo parlato prima) può servire al medico sportivo e al fisiatra per personalizzare le sedute di training sulle reazioni di un particolare atleta. Se poi una "centrale" raccoglie i dati di tutti le "fasce sportive" si può cominciare a osservare la salute dell'intera squadra, e questo può servire per cambiare l'intero training del gruppo.

La domanda seguente riguarda un argomento molto a cuore ai pazienti e a molti medici, ossia la indipendenza e la libertà individuale, oltre al problema della privacy. Come si conciliano questi aspetti con, per esempio le finalità di industrie o di aziende il cui scopo principale è il commercio? Una importante riflessione riguarda poi la provenienza dei dati per gli studi epidemiologici, dovrebbero essere utilizzati dati di pubblico dominio per i quali ci sia già un consenso al loro uso. Questo aspetto è attuabile considerando l'enorme mole di dati e le loro interrelazioni?
Certo, bisogna tenere conto della qualità dei dati e della loro rappresentatività. Certo, bisogna essere consapevoli delle grandi opportunità così come dei nuovi rischi: occorrono tecnologie a sostegno della privacy, occorre un "new deal" sui temi della privacy, della trasparenza e della fiducia, per far si che l'accesso alla conoscenza dei BigData sia bene pubblico per tutti. È necessario superare la fase attuale, in cui la maggior parte dei BigData interessanti sono tutt'altro che "open", ma chiusi nei database delle web corporation e sono sempre più la parte rilevante dei loro guadagni; questi problemi – qualità, privacy e proprietà dei BigData – sono decisivi.
Sicuramente questo nuovo percorso ha forti ripercussioni su molti aspetti etici: privacy e protezione dei dati personali (chi può accedere ai miei dati?), proprietà e sfruttamento dei dati (di chi sono i miei dati? Per quali scopi vengono usati?), trasparenza (chi può fare cosa con I miei dati?), consapevolezza e conoscenza di sé (come posso accedere alla conoscenza nascosta nei miei dati?), monopoli e asimmetrie (come controbilanciare il potere delle grandi corporation della conoscenza?). Nel nostro mondo interconnesso non possiamo permetterci di perdere l'opportunità offerta dai BigData, ma dobbiamo trovare un nuovo ecosistema socio-tecno-legale in cui la conoscenza sia un bene comune sicuro. La nuova direttiva europea sulla protezione dei dati GDPR entrata in vigore a maggio 2018 fa un ulteriore passo proprio in questa direzione.

Come possono i BigData riuscire a salvare vite o a predisporre programmi in ambito sanitario dai quali le popolazioni possano beneficiare in termini di salute?
I BigData non possono rimpiazzare il chirurgo, ma rendere più intelligente il robot che usa il chirurgo, o più efficace la comprensibilità delle immagini della risonanza che il chirurgo legge prima di decidere come procedere all'intervento, oppure aiutare a prevedere il picco dell'influenza e organizzare di conseguenza i turni in ospedale sulla base di ciò che la gente compra al supermercato: quest'ultimo non è un esempio a caso, ma un risultato scientifico del nostro laboratorio in collaborazione con gli epidemiologi digitali della Northeastern University di Boston.