BIG DATA E SANITÀ

Ovvero come la scienza dei dati sta rivoluzionando il nostro approccio alla salute

La storia moderna dei dati digitali inizia da internet e dal web. All'inizio degli anni Novanta internet era un piccolo regno dominato solo da scienziati e militari. Il web nacque proprio nel 1990 per facilitare la navigazione di internet e renderla disponibile anche ai comuni mortali che non erano né scienziati né militari. Nel 1990 anche i telefoni cellulari esistevano soltanto come esotici oggetti di tecnologia avanzata (e ingombrante!), che solo l'esercito poteva concepire di usare. In poco più di vent'anni questo scenario è cambiato in modo radicale e inaudito. La tecnologia mobile, internet, il web sono diventati così pervasivi da essere parte integrante del tessuto sociale che tiene insieme il consorzio umano. Molti diffidano di questa rivoluzione digitale che in pochi anni ha cambiato il mondo, ma le opportunità offerte da questa messe di dati generati dall'attività degli esseri umani in relazione alla tecnologia sono senza precedenti. Stiamo di fatto assistendo ad un nuovo rinascimento, in cui l'uomo è di nuovo al centro del sapere scientifico ed è la misura di tutte le cose. Lo specchio digitale, in cui l'umanità si riflette, permette di osservare l'umanità stessa con un'accuratezza mai vista prima. Grazie all'utilizzo di tecniche e metodologie provenienti dal dominio della *computer science* e del *data mining*[1], come ad esempio l'apprendimento automatico (*machine learning*, in inglese) o tecniche di elaborazione del linguaggio naturale (*natural language processing* o NLP, in inglese), è possibile identificare pattern, relazioni causali tra fenomeni precedentemente ignote, nuove conoscenze in modo automatico. Attualmente, la modellizzazione matematica e le tecniche di predizione possono essere applicate in un contesto di grande disponibilità di dati da sorgenti digitali e i modelli risultanti possono essere validati in tempo reale e su una scala mai vista prima.

Diversamente però da altri sistemi naturali, con il sistema uomo non si possono fare esperimenti. Si possono solo avere osservazioni nel passato, con un orizzonte temporale e una riproducibilità limitati e in un contesto esso stesso limitato. Inoltre, le sfide legate alla protezione della privacy degli individui che generano queste moli di dati sono ancora aperte. Nonostante le limitazioni, questo nuovo approccio, che il mondo accademico ha ribattezzato come 'scienza dei dati', sta portando una vera e propria rivoluzione nel campo delle scienze sociali ed anche e soprattutto nello studio di tutto ciò che riguarda la salute.

Nella maggior parte dei casi queste tracce digitali sono generate da attività che non hanno nulla a che vedere con studi epidemiologici o indagini legate alla sanità, ma il dettaglio e la qualità con cui questo cosiddetto *user generated content* viene prodotto è tale per cui è possibile utilizzarlo per studiare la salute della popolazione su scale fino ad ora impensabili.

Il caso di Google Flu Trends

Le storie di successo che vedono la scienza dei dati applicata al campo della sanità sono già numerose ed è ormai una pratica ben stabilita e accettata quella di utilizzare sorgenti digitali in ambito sanitario. Quella più famosa di tutte è probabilmente la storia di Google Flu Trends[2] e di come per anni, sin dal 2008, sia stato possibile utilizzare i dati del motore di ricerca di Google per raccogliere informazioni sui casi di influenza in giro per il mondo. Google Flu Trends ha mostrato per primo quali fossero le potenzialità dei *big data* in ambito sanitario, grazie alla possibilità di monitorare in tempo reale e su scala globale gli utenti che ricercano parole chiave associate alla malattia come, ad esempio, i sintomi. La grande innovazione di Google Flu Trends consisteva nel fatto che, grazie ad una *dashboard* (interfaccia) semplice e chiara, chiunque aveva la possibilità di utilizzare l'algoritmo di Google Flu Trends per accedere ai dati e osservare, Paese per Paese, la geografia della diffusione dell'influenza. Grazie a questi dati, Google Flu Trends forniva anche previsioni sull'andamento dell'influenza mostrando come, negli Stati Uniti, riuscisse ad essere in anticipo di diverse settimane sui dati divulgati dal Centers for Disease Control, l'istituto di sanità pubblica statunitense. La parabola di Google Flu Trends ebbe un arresto quando, nel 2012, le previsioni sull'andamento dell'influenza risultarono completamente errate. Un'analisi pubblicata su *Nature* del febbraio 2013[3], seguita da una ricerca approfondita di un team di studiosi di Northeastern University

e di Harvard, pubblicata nel marzo 2014 su *Science*[4] misero in discussione il modello utilizzato da Google Flu Trends e ne misero in luce tutte le limitazioni e i problemi[5].

I successi della scienza dei dati

Nonostante questa battuta di arresto legata al caso specifico di Google Flu Trends, i successi della scienza dei dati in ambito sanitario sono numerosi e in continuo aumento.
Nel caso dell'influenza stagionale, gli studi che hanno mostrato come sia possibile utilizzare sorgenti digitali per monitorare la diffusione geografica e temporale dell'influenza sono numerosi e si basano su sorgenti tanto eterogenee quanto abbondanti: da Twitter[6] a Wikipedia[7], per menzionarne solo alcuni. Più in generale nell'ambito delle malattie infettive, da quasi un decennio le sorgenti digitali vengono ampiamente utilizzate per portare avanti la cosiddetta *Digital Disease Detection*[8], che è diventata ormai un sottoambito ben delineato e completamente integrato (almeno in Paesi come gli Stati Uniti) della sanità pubblica. Già dal 2006, la possibilità di raccogliere ed elaborare informazioni dal web in maniera automatica ha portato alla creazione della piattaforma Healthmap[9], che la comunità epidemiologica internazionale ormai considera come il sistema di riferimento per la sorveglianza degli *outbreak* di malattie a livello globale.
Le storie di successo non riguardano soltanto le malattie infettive. Le sorgenti digitali possono essere usate anche per studiare, ad esempio, tramite dati da Facebook, la prevalenza di obesità in Paesi dove l'epidemia di obesità sta creando allarme[10]; tramite Twitter, la correlazione tra diffusione dei pollini e prevalenza di allergie nella popolazione[11]; parole chiave da motori di ricerca per la farmacovigilanza[12]; Twitter e depressione post parto[13]; Twitter e attitudine alla vaccinazione in relazione all'adozione del vaccino durante la pandemia di H1N1[14]. Uno dei più recenti e spettacolari successi della scienza dei dati applicata in ambito sanitario riguarda l'utilizzo di reti neurali e di tecniche di *deep learning*[15] applicate all'analisi di immagini di tumori della pelle per raggiungere capacità di diagnosi automatica paragonabili alle capacità diagnostiche di un dermatologo. Un gruppo di ricercatori di Stanford, nel febbraio del 2017, ha infatti pubblicato un articolo che spiega come dall'analisi di diverse centinaia di migliaia di foto di malattie della pelle, siano stati in grado di addestrare algoritmi di *deep learning* per diagnosticare in modo automatico patologie tumorali[16].

Le piattaforme di sorveglianza

Una particolare branca della scienza dei dati applicata alla sanità riguarda anche il diffondersi di piattaforme disegnate specificamente per raccogliere dati sanitari direttamente dalla popolazione, senza passare dai sistemi di sorveglianza istituzionali. Un esempio è rappresentato dalla decennale esperienza di piattaforme di sorveglianza partecipativa per l'influenza stagionale diffuse da diversi anni in Italia[17] e nel resto d'Europa[18] e che hanno portato alla creazione di esperienze simili anche in USA[19] e Australia[20]. Queste piattaforme si basano sull'attività di volontari che, durante tutta la stagione influenzale, riportano in modo regolare il proprio stato di salute e che permettono la raccolta di dati epidemiologici ad alta risoluzione geografica e temporale. Queste piattaforme, che combinano un approccio di *citizen science* alla sorveglianza epidemiologica, hanno avuto un successo così ampio e una rilevanza scientifica tale da essere in molti Paesi, come l'Italia con l'Istituto Superiore di Sanità e la Francia con l'INSERM, parte integrante del sistema nazionale di sorveglianza dell'influenza. Anche dati generati da comunità digitali di pazienti come PatientsLikeMe[21] sono ampiamente utilizzati dalla comunità scientifica per avere maggiori informazioni su specifiche patologie.

Dispositivi indossabili

Un'ulteriore sorgente di dati digitali che sta diventando sempre più rilevante in ambito sanitario è quella legata ai dati provenienti da sensori, indossabili e non, e da telefoni cellulari. Anche in questo ambito sono davvero numerosi gli esempi che si possono citare. Dati generati dagli individui che indossano sensori indossabili commerciali come FitBit o Jawbone possono ad esempio essere utilizzati per monitorare l'effetto di eventi sismici sul sonno degli individui[22].
Compagnie come Google ed Apple stanno equipaggiando i telefoni cellulari con app che permettono di tracciare da vicino l'attività fisica[23] e il sonno. Informazioni così dettagliate e personalizzate potranno poi in futuro essere usate per mettere a punto programmi di fitness ritagliati sulle caratteristiche

degli utenti. Prototipi di sensori indossabili sono largamente usati nella comunità accademico-scientifica anche per misurare le interazioni tra individui in ambientazioni a rischio di trasmissione di malattie infettive, come ospedali o scuole. Esempi molto conosciuti sono i sociometri dell'MIT[24] o i tag indossabili dell'esperimento italiano di Sociopatterns[25]. Questi ultimi sono stati utilizzati durante la pandemia di H1N1 nel 2009 per misurare le interazioni tra medici, infermieri, pazienti e visitatori all'interno dell'Ospedale pediatrico Bambino Gesù di Roma[26] per studiare i meccanismi di trasmissione dell'influenza e di malattie nosocomiali in ambito ospedaliero. Più recentemente sono stati utilizzati in ambientazioni a risorse limitate, come, per esempio, villaggi rurali in Kenya, per studiare le interazioni sociali tra nuclei familiari allargati tipici della società kenyota.

Conclusioni

In generale, alla luce di quanto visto fino ad ora, si potrebbe anche tentare una classificazione dei dati che possono essere utilizzati in sanità in base alla loro disponibilità e natura. Al di là di queste classificazioni, è comunque ormai chiaro alla comunità scientifica, al mondo industriale e alle istituzioni di sanità pubblica che internet, i social media, la tecnologia mobile hanno ormai più che rivoluzionato il mondo della sanità, creando la cosiddetta *digital health*, ovvero l'ambito cross-disciplinare che studia la convergenza tra tecnologie digitali e salute, assistenza sanitaria, benessere e società. In Italia, il Sole24H ha dedicato un blog permanente[27] alle tematiche di *digital health*. Le opportunità offerte da questo connubio tra scienza dei dati e mondo della sanità sono numerose e in continua crescita. Di pari passo con le opportunità, cresceranno anche le sfide riguardanti l'etica e la privacy che, in ambito sanitario, diventano ancora più rilevanti e che devono mettere in primo piano la salvaguardia degli interessi dei singoli individui come pazienti e come beneficiari delle opportunità offerte dai dati in sanità.

Daniela Paolotti
ISI Foundation, Institute for Scientific Interchange, Torino

NOTE E BIBLIOGRAFIA

1. Il *data mining* è l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di un sapere o di una conoscenza a partire da grandi quantità di dati (attraverso metodi automatici o semi-automatici) e l'utilizzo scientifico, industriale o operativo di questo sapere. Per approfondire https://it.wikipedia.org/wiki/Data_mining.
2. https://www.google.org/flutrends/about/
3. Butler D, When Google got flu wrong. Nature 2013; 494: 155-156.
4. Lazer D, Kennedy R, King G, Vespignani A, The parable of Google flu: traps in big data analysis. Science 2014; 343: 1203-1205.
5. http://cristinacenci.nova100.ilsole24ore.com/2014/04/06/google-flu-trends-big-data-senza-big-theory/
6. Lampos V, De Bie T, Cristianini N, Flu detector: tracking epidemics on Twitter. In: Balcázar JL, Bonchi F, Gionis A, Sebag M (Eds), Machine learning and knowledge discovery in databases. ECML PKDD 2010. Lecture Notes in Computer Science, vol. 6323. Springer, Berlin, Heidelberg.
7. Generous N, Fairchild G, Deshpande A et al, Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol 2014; 10 (11): e1003892.
8. Brownstein JS, Freifeld CC, Madoff LC, Digital disease detection-harnessing the Web for public health surveillance. N Engl J Med 2009; 360 (21): 2153-5, 2157.
9. http://www.healthmap.org/en/
10. Chunara R, Bouton L, Ayers JW, Brownstein JS, Assessing the online social environment for surveillance of obesity prevalence. PLoS ONE 2013; 8 (4): e61373.
11. Gesualdo F, Stilo G, D'Ambrosio A et al, Can Twitter be a source of information on allergy? Correlation of pollen counts with tweets reporting symptoms of allergic rhinoconjunctivitis and names of antihistamine drugs. PLoS One 2015; 10(7): e0133706.
12. Salathé M, Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. J Infect Dis 2016; 214 (Suppl 4): S399-S403.
13. De Choudhury M, Counts S, Horvitz E, Predicting postpartum changes in emotion and behavior via social media. CHI '13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 3267-3276.
14. Salathé M, Khandelwal S, Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011; 7 (10): e1002199.
15. Il *deep learning* (apprendimento profondo) è quel campo di ricerca dell'apprendimento automatico e dell'intelligenza artificiale che si basa su diversi livelli di rappresentazione, corrispondenti a gerarchie di caratteristiche di fattori o concetti, dove i concetti di alto livello sono definiti sulla base di quelli di basso. Per approfondire: https://it.wikipedia.org/wiki/Apprendimento_approfondito.
16. Esteva A, Kuprel B, Novoa RA et al, Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115-118.
17. https://www.influweb.it/
18. https://www.influenzanet.eu/
19. https://flunearyou.org/
20. http://www.flutracking.net/
21. https://www.patientslikeme.com/
22. https://jawbone.com/blog/napa-earthquake-effect-on-sleep/
23. https://www.wareable.com/sport/google-fit-vs-apple-health
24. https://www.forbes.com/forbes/2010/0830/e-gang-mit-sandy-pentland-darpa-sociometers-mining-reality.html.
25. http://www.sociopatterns.org/
26. Isella L, Romano M, Barrat A et al, Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. PLoS ONE 2011; 6 (2): e17144.
27. http://cristinacenci.nova100.ilsole24ore.com/