Big Data revolution

Officina di Fisica

Settembre 2017

Contents

Chapter 1

Introduzione

Non solo una questione da prima pagina: La così detta "Big Data Revolution" è ora sulle bocche di tutti: come avremmo potuto tirarci fuori dal coro? Certamente la scelta dell'argomento è dovuta alla sua attuale rilevanza mediatica. Eppure per noi non è solo attualità, ma anche un'importante occasione per analizzare come sta evolvendo il metodo di ricerca scientifica e come si stanno trasformando i rapporti sociali, a partire dal lavoro, in relazione allo sviluppo delle nuove tecnologie. Come Officina questo è sempre uno dei nostri obbiettivi: farsi un'idea critica sulle relazioni che intercorrono fra scienza e società. In linea con tale spirito, speriamo che il lettore possa trovare in questo documento qualche spunto originale per guardare in modo diverso la realtà della scienza che lo circonda, ci teniamo a chiarire che la nostra trattazione degli argomenti non ha la pretesa di essere esaustiva, essa si presenta piuttosto come la condivisione in forma scritta di un percorso assembleare di autoformazione e riflessione.

-Cosa sono? Come spesso accade è utile in principio aver chiara la definizione. Per big data intendiamo tutte quelle collezioni di dati destrutturati le cui dimensioni superano le capacità di memorizzazione, gestione e analisi tipiche dei tradizionali sistemi per basi di dati. Generalmente le loro caratteristiche vengono riassunte dalle cinque "v": Volume, Velocità, Varietà, Viralità e Variabilità. Il volume è chiaramente riconducibile al peso in byte di questi ammassi di dati mentre la velocità è riferita alla generazione e all'accesso di essi. Varietà e variabilità consistono nel fatto che, avendo forme differenti in origine, non possono essere ordinati in una struttura prefissata. Infine la viralità esprime il loro potenziale di impatto sociale dovuto alla veloce diffusione delle informazioni che vi si estraggono.

Un Data Scientist non ci metterebbe più di un minuto a convincervi che è proprio su questa tecnologia che si basa il progresso ai giorni nostri: è su questo che si deve investire, dato che in molti e grandi lo stanno facendo, è

su questo che ci si deve specializzare, dato che sempre di più sono e saranno, secondo il trend attuale, gli esperti del settore richiesti. Trend che è positivo e in stabile crescita. Dal 2016 al 2020 è previsto che si passi da 1,1 a 2,3 ZB di traffico IP (1 ZB = 1012 GB). Qualcuno li ha definiti "il nuovo petrolio" e cercheremo nelle prossime pagine di capirne il motivo.

-Chi li produce? Noi tutti li produciamo. Quando ognuno di noi compie un'azione interfacciandosi con un dispositivo connesso alla rete, nulla va dimenticato: immediatamente viene generato un file in cui vengono salvati non solo i contenuti testuali o multimediali prodotti dalla nostra azione, ma anche tutte le informazioni ad essi connesse che prendono il nome di metadati. Perciò quando apriamo Facebook, messaggiamo su Whatsapp, facciamo una ricerca su Google o vediamo un video su Youtube, i server delle aziende appena citate registrano l'indirizzo IP del computer utilizzato, l'ora, il luogo e tutto ciò che fa da contesto alla nostra traccia digitale.

Il grado di consapevolezza rispetto a ciò da parte degli utenti risulta piuttosto variabile: da un parte siamo proprio noi ad accettare Termini e Condizioni di utilizzo dei servizi, dall'altra lo facciamo spesso senza sapere come i dati che generiamo vengono poi analizzati, e le deduzioni che ne vengono tratte. Dal momento che in questi server lo spazio di archiviazione dei dati è sufficientemente enorme, non c'è motivo per non conservarli. Quel che non strozza ingrassa e tutto, come vedremo, può sempre tornare utile.

Ma oltre a questi dati ci sono quelli prodotti da dispositivi non virtuali come le stazioni metereologiche, le webcam, i sensori del traffico, quelli di sicurezza, le immagini satellitari, i geo localizzatori, e tutti quegli oggetti, sempre più diffusi, il cui utilizzo viene integrato da una connessione ad Internet (Internet of Things **??**). Infine ultimi, ma non per importanza, vi sono i dati tradizionali, raccolti dalle agenzie pubbliche, dagli ospedali, dalle banche e dai laboratori di ricerca.

-E poi... come vengono utilizzati? Una volta prodotti e acquisiti i dati sarebbe un peccato non utilizzarli per qualche scopo pratico che implichi possibilmente un margine di guadagno. Ma a seconda di quale scopo si cerchi di soddisfare varia il processo di elaborazione, analisi, e interpretazione di questii. Ad esempio, sarà capitato a molti di voi di vedere Google suggerirvi una chiave di ricerca diversa da quella che avevate digitato, il classico "forse cercavi: ....?" Come fa Google a sapere meglio di te quello che vuoi trovare? La risposta è che Google impara da tutte le ricerche precedentemente fatte dai suoi utenti collezionando tutti i dati e metadati relativi a queste. Quindi quella che viene valutata è la coerenza della vostra ricerca con la Storia delle altre già effettuate.

Sicuramente, un'applicazione in cui Big Data giocano un ruolo fondamentale è quella della profilazione degli utenti e quindi di tutte le sue conseguenze.

L'esempio più semplice è quello della pubblicità mirata che possiamo esperire tutti i giorni considerando i banner che, ad esempio, dopo aver comprato un volo per Cuba, mostreranno durante giorni seguenti alternative di viaggio in offerta, guide economiche del Paese e busti di Fidel in plastica colorata. La profilazione tuttavia non si ferma ad un semplice resoconto delle abitudini dell'utente sulla rete ma può consistere in un vero e proprio tracciamento del suo profilo psicologico. Questa possibilità apre la strada a nuove modalità di orientamento dell'opinione pubblica e di propaganda elettorale di cui si è avuto un esempio concreto durante le ultime presidenziali negli USA **??**.

Diverso è il caso in cui si voglia con i dati a disposizione fare una predizione su una situazione futura. In questo caso essi non devono solo consentire una ricostruzione del contesto da cui sono stati estratti, ma devono altresì permettere di prefigurarsi come si modificherà tale situazione dopo un certo tempo. Questa capacità predittiva può avere differenti pretese che vanno dal semplice sondaggio, al controllo sociale sino all'effettiva necessità di fare previsioni su fenomeni fisici i cui modelli teorici non sono stati ancora sviluppati. Quest'ultima pretesa in particolare è al centro di un dibattito accademico molto accesso tutt'ora in corso **??**.

Abbiamo qui dato una panoramica generale sul tema dei big data delineandone le caratteristiche chiave e i principali utilizzi. Nei capitoli che seguono entreremo maggiormente in dettaglio su alcuni aspetti e questioni che sia da una prospettiva tecnico-scientifica che da un punto di vista sociale si collegano al fenomeno battezzato come "big data revolution".

Chapter 2

Scienza

Qui va una breve intro in cui si raccordano le tre sezioni.

2.1 Machine Learning

Quando vi sono dei modelli teorici questi schematizzano il problema, ne individuano le variabili rilevanti e le mettono in relazione in modo tale che se ne possa dedurre il risultato al variare del tempo. Quando invece questi mancano e non è noto né quali siano le caratteristiche dell'oggetto rilevanti da prendere in considerazione, né quale relazione intercorra fra queste e la predizione viene fatta in altro modo. Negli ultimi anni vasto e rapido è stato lo sviluppo di metodi informatici come il Machine Learning. Sotto questo nome si raccolgono moltissimi algoritmi di analisi diversi, ma essenzialmente le tecniche di processamento dei dati sono due: l'apprendimento supervisionato e quello non supervisionato. Di cosa si tratta? In cosa differiscono? Quel piccolo avverbio negativo sta ad indicare una differenza sostanziale: nel primo caso abbiamo dei dati con certe caratteristiche che sono divisi in un certo numero di classi. La macchina impara da questi dati le regole per suddividere i prossimi input nelle diverse classi. Se consideriamo come esempio individui con un profilo facebook possiamo raggrupparli nelle classi di "timidi" e "estroversi" in base a un nostro personale metro di giudizio. Successivamente prendiamo in considerazione per ognuno il numero degli amici che ha associati al suo profilo e facciamo sì che la macchina impari come questa caratteristica si associa alla suddivisione in classi. In questo modo chi il giorno successivo crea un nuovo profilo e manda un certo numero di richieste di amicizia è subito classificato sulla base di queste in timido o il suo contrario. Si tratta di un esempio banalizzante: è evidente che non basta il numero degli amici su facebook per definire quanto qualcuno è effettivamente

propenso all'interazione sociale. Ma se oltre al numero di contatti si considerassero anche il numero di post pubblicati, quelli commentati, i like messi a determinate pagine e il numero e tipo di gruppi a cui si è iscritti la classificazione ottenuta sarebbe forse meno discutibile. Nella seconda tecnica, quella dell'apprendimento non supervisionato, i dati non sono in partenza associati a classi, e quello che gli algoritmi devono riuscire a fare è definire dei gruppi di oggetti in base alle loro somiglianze. In altre parole si dice che devono svelare correlazioni intrinseche. sia quali siano le classi, sia come i dati sono associati a queste. Con queste tecniche si ottengono correlazioni fra i dati sufficienti a stimare, con un grado di accuratezza che dipende dalla quantità e qualità dei dati presi in esame.

2.2 Nuovo paradigma scientifico

Nell'articolo *"The end of theory"* pubblicato sulla rivista Wired nel 2007 l'autore preannuncia il fatto che le tecniche avanzate di analisi dei big data renderanno obsoleta l'approccio modellistico ai fenomeni naturali.

2.3 Big Data e biologia

Sempre di più e in modo crescente negli ultimi anni la biologia si serve di strumenti informatici per procedere nelle sue ricerche, tanto che ad oggi un biologo nella maggior parte dei casi ho ha queste competenze o lavora in equipe con chi ne ha. Eppure è proprio dal mondo biologico che ha preso ispirazione chi ha sviluppato gli algoritmi di intelligenza artificiale e dei neural network che ora non si può prescindere dall'utilizzare! Dal duemila con il Progetto Genoma Umano ad oggi i dati di ambito biologico e medico sono cresciuti esponenzialmente: analisi di sequenza di DNA, RNA, proteine, processi di regolazione, vie metaboliche.. Grazie ai vasti spazi di archiviazione è possibile conservare genomi di miliardi di paia di basi di migliaia di organismi, gli algoritmi di analisi permettono di estrarne velocemente informazioni, le tecniche di Machine Learning rendono possibile stabilire a quale malattia si è predisposti partendo da singoli polimorfismi del proprio DNA.. La ricerca biologica e quella medica paiono far veloci passi avanti grazie allo sviluppo dei Big data e delle tecnologie ad essi associate, ma come

Chapter 3

Profilazione e controllo sociale

La profilazione spezza le unghie

3.1 L'IoT e' la panacea dei casalinghi tecnofili

IoT e' una delle tante sigle che popolano il mondo dell'informatica, la si trova sui siti 'geek' o nelle rubriche delle riviste, e sta per Internet of Things.

L'idea dell'Internet of Things viene introdotta da Aston nel 1999, l'articolo originale e' disponibile a [**?**], e prospetta la possibilità di salvare tempo utile delle persone, lasciando ai computer il duro compito di esplorare da loro la realtà, senza attendere che sia un umano a fornire i dati da elaborare. Nell'arco degli anni, l'IoT, si è concentrato sulla domotica, e si riassume l'idea dicendo: "L'IoT è il tuo frigo che parla con la lavatrice", e riprendendo lo spunto dell'articolo di Aston offrire la possibilità di relazionarsi con gli oggetti attraverso il linguaggio naturale, come nei film di fantascienza. La domotica tuttavia è una scienza ben più antica e i suoi fini sono più concreti, l'ottimizzazione dei consumi e dei tempi in casa, giusto per capire il punto di partenza, quando si tratta di domotica si parla di sensori termici e orologi sparsi per casa che accendono i riscaldamenti o fanno partire la macchina del caffè all'ora della sveglia; queste tecnologie sono nate nel mondo analogico e spesso non necessitano un controllo remoto, piuttosto permettono di programmare e attuare delle routine domestiche in modo automatico. L'IoT si presenta come una radicale rivoluzione di questo mondo delle cose: ora, oltre ad essere programmabili e 'sensibili', gli oggetti divengono parlanti!

Basta aggiungere una connessione wifi al frigo, alla lavatrice o al folletto, ed ecco che è possibile comandarli da remoto o addirittura attraverso i social network. Questa 'rivoluzione' informatica e informatizzante si traduce in una sterminata rete di sensori che mappano la realtà e la traducono in un'immagine digitale, quando ci si riferisci a questi rinnovati oggetti si usa l'appellativo di Smart. Cerchiamo allora alcuni esempi per capire meglio quali oggetti e ambiti della vita possono essere investiti da questa ventata di innovazione digitale.

Facciamolo cercando di immaginare la vita di Mario, un cittadino tecnofilo, e abbastanza ricco da permettersi dei simpatici gadget: Si comincia con Bonjour [1], la sveglia intelligente che mette insieme informazioni su traffico, meteo e agenda per salvaguardare il più possibile il tuo sonno, a Bonjour dovrai fornire informazioni sulla tua quotidianità e elencare metodicamente i tuoi appuntamenti. Probabilmente la sveglia può twittare alla machinetta del caffe e dirgli di accendersi; quindi si esce e si sale sulla futuristisca SmartCar che impostasta la destinazione ci conduce nel più breve tempo possibile conoscendo il traffico e le abitudini degli altri utenti. Ovviamente Mario in-

[1] http://www.smartworld.it/tecnologia/bonjour-sveglia-smart.html//

dossa fitbit [2]] che monitora i parametri vitali e controlla che tutto sia nella norma, mandandoci un messagino quando siamo arrabiati, tenendo d'occhio la qualità del nostro sonno e assicurandoci un rapporto digitale con il nostro corpo. Fin qui sembra che sia tutto ok, che Mario abbia degli effettivi vantaggi in termini di efficienza e tempo risparmiato, ma non abbiamo considerato la possibilità che nella sua borsa ci sia Hidrateme, [3] la borraccia intelligente che ci ricorda di bere se abbiamo sete inviandoci una notifica, o che in casa disponga di qualche altra diavoleria come SMALT, il dispensatore di sale che controlla la quantità di minerale assunta. Molte di queste trovate sono elencate in https://weputachipinit.tumblr.com/.

Insomma parafrasando la simpatica storiella di Mario quello che ci preme dire è che se da un lato alcune innovazioni ci fanno risparmiare tempo perso e forse migliorano la qualità delle nostre giornate, dall'altro ci deresponsabilizzano e ci abituano a delegare l'attenzione e in alcuni casi la volontà a degli oggetti. E' solo una questione di principio la nostra scherzosa critica all'IoT? Perché se cosi fosse dovrebbe esser riservata ai singoli e lasciare a ciascuno la scelta di quanta attenzione porre nella proprie azioni quotidiane. Tuttavia ci sentiamo di sottolineare che l'alienazione di cui questa tecnologia è foriera è estremamente pervasiva e ci immerge in una spirale di dipendenza dagli oggetti che nella loro onnipresenza condizionano le nostre scelte e le nostre attività.

Tutto questo per non parlare dei possibili, e probabili, disagi che una vita iperinformatizzata nasconde. Uno dei più recenti casi è l'aggiornamento del firmware di Ilocks [4], che controllava la serratura di molte abitazioni e per un malfunzionamento ha lasciato fuori centinaia di utenti.

D'altro canto e' importante parlare di IoT perché l'atrofizzazione della scelta è solo la punta di un grande iceberg. La questione che ci preme discutere in questo paragrafo è che molti dei dati prodotti dai sensori dell'IoT finiscono, in un modo o nell'altro, in Internet. I modi in cui questi dati possono essere ottenuti sono molteplici, in alcuni casi sono le stesse aziende produttrici che li raccolgono, che questa clausola sia inserita nei Termini d'Uso o sia caldamente suggerita dal software al primo avvio conta poco; altre volte sono gli stessi utenti che li condividono perché spinti dall'edonismo della rete che vuole conoscere i risultati e mettere in competizione finanché gli sportivi della domenica. I dati ottenuti attraverso le SmartTv, i frigoriferi a scansione o i termostati fanno gola a molti, che siano essi soggetti economici

[2] https://www.fitbit.com/it/home[[

[3] $https://www.kickstarter.com/projects/582920317/hidrateme-smart-water-bottle?ref=category_recommended$

[4] `https://www.theregister.co.uk/2017/08/11/lockstate_bricks_smart_locks_with_dumb_firmware_upgrade/`

che pagano per ottenere profilare gli utenti e creare pubblicità o prodotti più appetibili, o che siano agenti più subdoli che si propongono di aggregare queste informazioni ottenute ai margini della legalità e una volta arricchite rivenderle a terze parti. I casi di leak ovvero di fuga di informazioni, infinem sono numerevoli, ci limitiamo a citarne uno che ha scosso l'opinione pubblica [5].

[5] https://motherboard.vice.com/en_us/article/pgwean/internet-of-things-teddy-bear-leaked-2-million-parent-and-kids-message-recordings

3.2 Embers

Sviluppato dal Descovery Analitics Centre della Virginia Polytechic Institute EMBERS è un progetto che dal 2012 predice ogni giorno 45-50 eventi di rilevanza sociale in molti paesi del Sud America. [1] I finanziamenti arrivano (22 milioni) dall'agenzia di intelligence di stato americana (IARPA) [6] in quanto parte del progetto OSI (Open Source Indicators, [7]), con una collaborazione attiva in termini di ricercatori e finanziamenti di molte università americane. Lavora utilizzando dati come tweets, pagine facebook , blog posts, ricerche di Google, Wikipedia, dati metereologici, indicatori finanziari ed economici, immagini satellitari. I dati utilizzati sono OpenSource, ovvero accessibili attraverso internet da qualsivoglia operatore, questi dati, che sono di fatto BigData, sono definiti dagli autori del progetto come *massivi, passivi* Instancabile il programma lavora 24h, 7su7, offrendo pronostici sugli eventi che sconvolgeranno i paesi posti sotto osservazione. I tipi di eventi prevedibili sono epidemie di malattie rare o di influenze, rivolte ed elezioni poilitche; ma gli autori del progetto sono, inaspettatamente, interessati alle ultime due classi di eventi. Nell'articolo già citato, con cui Embers si presenta al mondo, i ricercatori elencano i successi ottenuti nella previsione di eventi quali la primavera brasiliana del 2013, le violente proteste degli studenti venezuelani del 2014, le elezioni presidenziali di Panama e Colombia sempre del 2014.

EMBERS si presenta come il fiore all'occhiello della ricerca in casa Iarpa, infatti rispetto ai presistenti progetti (ICEWS, PITF), il sistema ha un'accuratezza elevata, fino ad indicare città, giorno e volume dell'assembramento di persone. Inoltre, l'utilizzo di motori per l'analisi e la produzione di testo naturale consente un certo livello di comprensione del fenomeno in questione, e finanché una narrazione dell'evento: Per capire come un marchingegno del genere possa funzionare entriamo nel dettaglio del sistema: Il processo di analisi di embers comprende quattro stadi:

- *Ingestion*: aquisizione dei dati OS dalle varie fonti elencate, per far ciò serve un sacco di spazio e delle connessioni molte veloci.
- *Enrichment*: qui i dati vengono 'migliorati', il testo viene processato e si tenta di iferire la città e altre informazioni sull'autore del post o del tweet.
- *Modeling* A questo punto viene compiuta l'analisi secondo i modelli che costituisconoil core di EMBERS. Qua avviene il miracolo big data: i dati vengono messi i relazione e si esplorano la semantica e il volume

[6] `https://www.iarpa.gov/`

[7] `https://www.iarpa.gov/index.php/research-programs/osi`

EMBERS forecasts that there will be a violent protest on February, 18th 2014 in Caracas, the capital city of Venezuela. We predict that the protest will involve people working in the business sector. The protest will be related to discontent about economic policies.
There were 5, 5, and 5 other similar warnings in last 2, 7 and 30 days, respectively.
The forecast date of the warning falls in week 7, which may have historical importance; this week is found to be statistically significant (pval=0.00461919415894, zscore=2.832, avg. count=57.25, mean=21.569 +/- 12.597)
Audit trail of the warning includes an article printed 2014-02-17.
Major players involved in the protest include Venezuelan opposition leader, students, President Nicolas Maduro, and Leopoldo Lopez.
Reasons: Protest against rising inflation and crime; Protestors want a political change; President Nicolas Maduro has accused US consular officials and right-wing.
Protests are characterized by: Venezuelan opposition leader spearheaded days of protest and calling for peaceful demonstration; Maduro accused official on 2014-12-16; Protests have seen several deadly street protests; Three people were killed on 2014-02-12; Demonstrations setting days of clashes; supporters to march to Interior Ministry on 2014-02-18.

Figure 3.1: An example narrative for a EMBERS alert message. Here, color red indicates named entities, green refers to descriptive protest related keywords. Items in blue are historical or real time statistics and those in magenta refer to inferred reasons of protest.

delle manifestazioni programmate. Gli algoritmi messi in campo sono i seguenti:

- Planned Protest Model; dai social sono identificati specifici segni di chiamate a eventi di protesta (con luogo e data);

- Dynamic Query expansion; usa twitter per identificare tempo e luogo di diffusione nell'uso di alcune parole chiavi legate alle proteste;

- Volume-based model si serve di molti dati di indicatori sociali, economici, politici.

- Cascade regression model modellizza le attività su Twitter che siano legate con organizzazioni e mobilitazioni;

- Baseline model usa un modello di stima a partire dallo storico degli eventi del GSR (a monthly catalog of events as reported in newspapers of record in 10 Latin American countries).

- *Selection* Integrazione dei risultati e presentazione delle predizioni finali, come in Fig. 3.1

Nella presentazione del progetto non si elude di affrontare anche le implicazioni etiche di questo! Si tratta di uno strumento che certamente può degenerare se nelle mani sbagliate, come per esempio quelle di un governo autoritario non democratico. Al contrario il popolo è salvo se EMBERS è utilizzato da un governo attento e premuroso come quello statunitense! Anzi, in queste circostanze è da considerare come sensore accurato degli umori dei

cittadini rispetto alle politiche governative, uno strumento capace di far sentire più forte la voce di tutti, di avvicinare palazzi del potere e mondo che li circonda.

Obbiettivi: -affinare sempre di più la precisione della previsione; -cercare di ridurre sempre più l'lemento umano necessario allo sviluppo dell'analisi, attualmente il suo ruolo maggiore è quello di generare il GSR (ancora una volta si esplicita quanto l'elemento umano sia il problema da eliminare per ottimizzare).

Bibliography

[1] Sathappan Muthiah, Anil Vullikanti, Achla Marathe, Kristen Summers, Graham Katz, Andy Doyle, Jaime Arredondo, Dipak K. Gupta, David Mares, Naren Ramakrishnan, Patrick Butler, Rupinder Paul Khandpur, Parang Saraf, Nathan Self, Alla Rozovskaya, Liang Zhao, Jose Cadena, and Chang-tien Lu. EMBERS at 4 years: Experiences operating an Open Source Indicators Forecasting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 205–214, 2016.