

Big data e Open data

N°2-Giugno 2017

[Di Cosa Parliamo Quando Parliamo Di...](#)

di Gea Arcella, Notaio

Quotidianamente siamo immersi in un mare di informazioni e noi stessi ne produciamo continuamente: il nostro essere interconnessi, l'uso continuo degli strumenti informatici ha reso tutto questo materiale sempre più trattabile per le finalità più disparate e ciascuno di noi non sempre è in grado di governare la **mole di dati** che vengono estratti dalle nostre interazioni informatiche.

Le foto, i post, le email, la musica che ascoltiamo, i nostri commenti, uniti a dati di geo-localizzazione, le nostre ricerche, perfino i download rimasti incompleti sono diventati il nuovo "petrolio" del terzo millennio: dati che analizzati producono altri dati e consentono profilazioni dei comportamenti e analisi di mercato sempre più penetranti. In una parola, "i big data".

Cosa sono i Big data

Big data è il termine utilizzato per descrivere una raccolta di dati così estesa in termini di volume, velocità e varietà da richiedere **tecnologie e metodi di analisi specifici** al fine di poter estrarre da essi ulteriore valore in termini informativi (fonte [Wikipedia](#)).

Il progressivo aumento della dimensione dei data set è legato alla necessità di analizzare un unico grande insieme di dati, da cui ricavare informazioni aggiuntive rispetto a quelle che si potrebbero ottenere analizzando dei data base parziali e separati, anche se cumulativamente contenenti la stessa quantità totale di dati: in pratica le relazioni e le inferenze tra i dati sono un **ulteriore valore aggiunto** ricavabile da questa tipologia di data set.

Un esempio possono essere le analisi condotte per sondare gli "umori" dei mercati e dei consumi: esse richiedono l'individuazione delle principali tendenze sociali in un dato momento storico e vengono effettuate attraverso l'analisi e l'incrocio del fiume di informazioni che transitano attraverso Internet.

I Big data, dunque, mettono insieme dati provenienti potenzialmente da **fonti eterogenee**: non soltanto dati strutturati - come quelli presenti nei database - ma anche dati non strutturati, come immagini, email, dati di geo-localizzazione o geo-referenziati, informazioni prese dai social network.

Il volume dei Big data è misurabile dell'ordine degli zettabyte, ovvero miliardi di terabyte: pertanto per la loro analisi è necessaria una potenza di calcolo parallelo e massivo enorme, sviluppata con strumenti dedicati che vanno "eseguiti" su decine, centinaia o anche migliaia di server contemporaneamente.

Le caratteristiche dei Big data

Le principali caratteristiche dei big data possono essere così individuate:

- **volume:** rappresenta la dimensione effettiva del dataset; in astratto la quantità di dati che è possibile raccogliere potrebbe apparentemente rappresentare un problema: in realtà l'utilizzo di risorse in cloud e la virtualizzazione delle "macchine" aiutano nella gestione della grande quantità di dati disponibili, semplificando i processi di raccolta, immagazzinamento e accesso ai dati;
- **velocità:** si riferisce alla velocità con cui vengono generati ed analizzati i dati; infatti le analisi dei dati vengono effettuate in tempo reale o quasi;
- **varietà:** concerne le varie tipologie di dati, provenienti da fonti diverse (strutturate e non);
- **veridicità:** fa riferimento all'attendibilità dell'informazione.

La business intelligence

Fino a qualche tempo fa l'analisi delle informazioni a fini commerciali apparteneva alla c.d. business intelligence; l'emergere di una definizione autonoma dei big data mette in evidenza le differenze con tale branca del marketing, soprattutto con riferimento ai dati ed al loro utilizzo:

- la business intelligence utilizza la **statistica descrittiva** con dati ad alta densità di informazione per misurare cose, rilevare tendenze, ecc., cioè utilizza data set limitati, dati puliti e modelli semplici;
- l'analisi dei big data utilizza la **statistica inferenziale** e concetti di **identificazione di sistemi non lineari**, per dedurre leggi (regressioni, relazioni non lineari, ed effetti causali) da grandi insiemi di dati e per rivelare i rapporti, le dipendenze, e effettuare previsioni di risultati e comportamenti, cioè utilizza data set eterogenei (non correlati tra loro), dati raw (grezzi, privi di metadati) e modelli predittivi complessi.

Cosa sono gli Open data

Concettualmente diversi sono gli open data, che vengono definiti come: «un contenuto o un dato da chiunque utilizzabile, riutilizzabile e ridistribuibile, soggetto, al massimo, alla richiesta di attribuzione e condivisione allo stesso modo».

Il concetto di open data nasce nell'ambito delle teorie statunitensi sull'**Open Government**, quale mezzo per assicurare nell'agire amministrativo trasparenza, conoscibilità e controllo da parte dei cittadini.

La partecipazione ha come presupposto necessario la conoscenza e questa può essere assicurata attraverso gli open data, ovvero attraverso la condivisione delle informazioni.

Ma quali sono le **tipologie** di open data?

Essi sono generalmente identificati in: dati cartografici, genetici, chimici, matematici e scientifici (incluse le relative formule), dati medici e delle bioscienze, dati anagrafici, dati governativi, ecc.

La fonte degli Open data è il patrimonio di conoscenza pubblico, comune a tutti i cittadini, che la Pubblica Amministrazione utilizza e rielabora per la gestione dei processi relativi all'assolvimento dei suoi compiti istituzionali e di servizio; pertanto in un **meccanismo circolare e virtuoso** i dati pubblici posti a base delle decisioni della PA vengono rimessi a disposizione della comunità sia per consentire la verifica dell'operato della medesima PA ma anche come *sapere* da riutilizzare.

Le caratteristiche degli Open data

Così come abbiamo cercato di individuare le principali caratteristiche dei big data, di seguito vediamo quelle degli Open data:

- **valore:** sono dati abilitanti, ovvero che mettono il cittadino nelle condizioni di disporre degli strumenti conoscitivi indispensabili per poter prendere decisioni o comunque valutare le decisioni prese dall'Amministrazione;
- **veridicità:** sono informazioni attendibili perché validate in ambito pubblico;
- **utilizzabilità:** sono dati forniti in modo da permetterne il riutilizzo e la redistribuzione. Ciò comprende la possibilità di combinarli con altre basi di dati; tutti devono essere in grado di usare, riutilizzare e redistribuire i dati, nessuna restrizione né commerciale né legata a fini determinati è ammessa.

Le finalità e gli utilizzi

Sicuramente gli Open data sono un **sottoinsieme dei Big data**, che per la loro estensione li ricomprendono, ma profondamente diverse sono le finalità e gli utilizzi che caratterizzano queste due grandi categorizzazioni dei dati:

- i Big data vengono raccolti anche l'insaputa dell'interessato, per profilarne i gusti e le tendenze, e vengono utilizzati a **fini privatistici** e di analisi del mercato,
- gli Open data sono dati pubblici raccolti nell'abito dell'azione della P.A., essi devono essere disponibili, riutilizzabili, e vengono messi a **disposizione della comunità** per incentivare la partecipazione alla gestione della cosa pubblica.

Il valore dell'informazione pubblica

Consapevolmente o meno, continuamente cediamo pezzi della nostra personalità consentendo alle aziende di capire le nostre opinioni e le nostre preferenze semplicemente navigando in internet o utilizzando gli strumenti social.

La pervasività di queste analisi ha portato il 30 maggio 2017 l'AGCOM, l'Antitrust e il Garante privacy a dare il via a **un'indagine congiunta** sui Big data.

Concorrenza, libertà di informazione e protezione della vita privata sono i grandi temi sotto la lente delle tre autorità di garanzia.

Eppure come singoli spesso viviamo dei paradossi: il miraggio di qualcosa di gratuito ci porta a consentire continue invasioni nella nostra sfera privata, mentre siamo talvolta insofferenti verso l'informazione pubblica, intesa come quel complesso di dati e notizie presenti nei registri di pubblicità legale, anch'essa parte essenziale degli open data.

Recentemente della Corte di giustizia dell'Unione europea, con la sentenza del 9 marzo 2017, relativa alla causa C-398/15 – ha ribadito il valore dell'informazione pubblica in relazione alla richiesta di un soggetto che chiedeva l'oscuramento di alcuni suoi dati personali collegati al fallimento di una società (tecnicamente chiedeva il diritto all'oblio) affermando che: “... **la pubblicità del Registro delle imprese ricopre una funzione pubblica essenziale**, in quanto garantisce la certezza del diritto nelle relazioni tra le società e i terzi e tutela gli interessi di questi ultimi rispetto alle società di capitali, dal momento che queste offrono come unica garanzia per i terzi il proprio patrimonio sociale; per garantire il soddisfacimento di tali finalità, la pubblicità legale deve consentire ai terzi di conoscere gli atti essenziali della società interessata, determinate informazioni che la concernono e in particolare le generalità delle persone che hanno il potere di obbligarla.”

Il principio di diritto espresso dalla Corte europea è particolarmente importante per la salvaguardia della completezza ed esaustività dei pubblici registri: l'aver ribadito che esistono delle **esigenze superiori rispetto agli interessi dei singoli**, che permangono nonostante il decorso del tempo, significa che la pubblica fede, connessa a prerogative statuali, è riconosciuta come un valore; così come un valore per l'intera comunità è tutta l'informazione pubblica, soprattutto se a contenuto legale.

Totalmente diverso l'approccio della stessa Corte - con la nota sentenza c.d. **Google Spain** - rispetto ai grandi provider privati che dei dati personali fanno commercio anche con finalità non del tutto trasparenti e qualche volta lesive dell'identità personale, le quali non giustificano la permanenza dell'informazione senza limiti di tempo, anzi la pervasività delle analisi sui Big Data che abbiamo descritto devono portare ad un ripensamento della regolamentazione sul tracciamento e la raccolta delle informazioni personali in rete ed attraverso la rete.