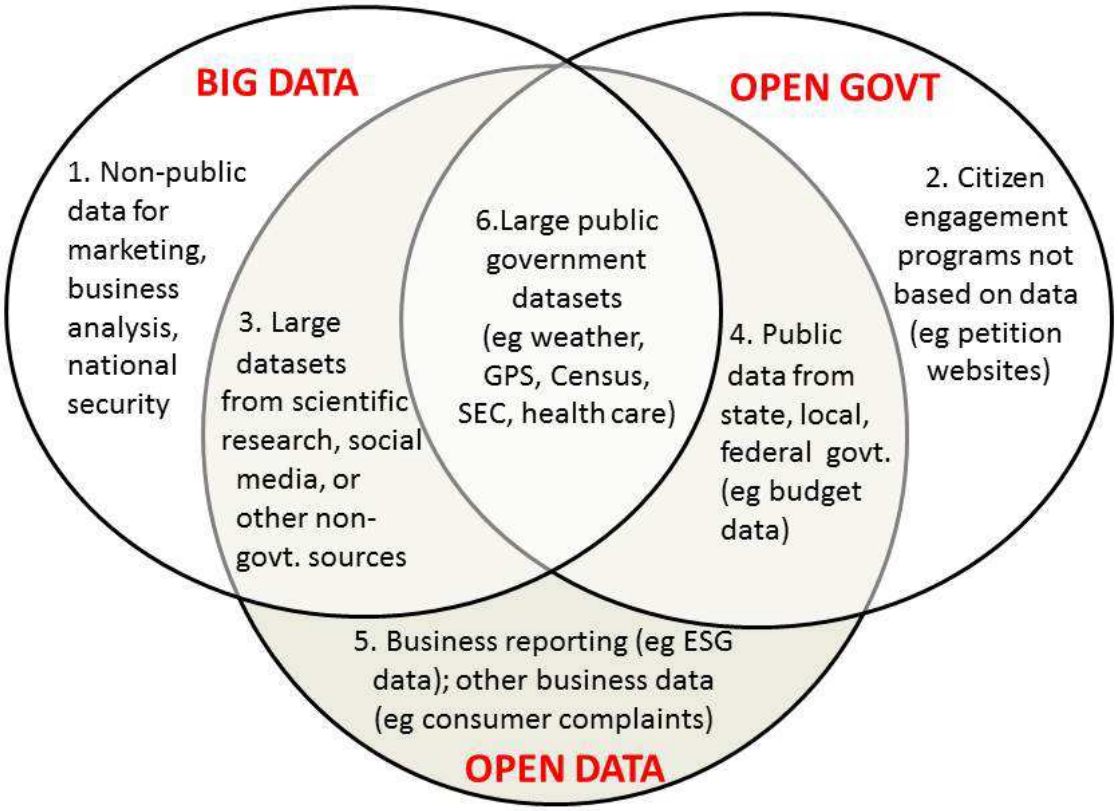
Letture di approfondimento su Big e Open data

Big data e open data: connessioni e differenze



1

From http://www.opendatanow.com/

Joel Gurin, Founder and Editor, OpenDataNow.com

08 November 2013

As I've worked on my upcoming book, Open Data Now – to be published by McGraw-Hill on January 10 – I've had to think through and explain how Open Data, Big Data, and Open Government are related to each other. Lately I've seen a number of others, like the authors of the new McKinsey Open Data report (see page 4), try to map the territory in similar ways. The Open Data community is producing a lot of Venn diagrams these days, with a lot of colorful overlapping circles. (Some also deal with the use of Personal Data, but that's one circle too many for me.)

For my own contribution to the discussion, I'm proposing the model shown here. To all Open Data wonks: Please take a look, comment, and add your own ideas. We're at a stage where we need to define more precisely what we're talking about. This may help.

My starting point is the evolving understanding of these three areas. Big Data essentially describes very large datasets, but that's a somewhat subjective judgment that depends on technology: today's Big Data may not seem so big in a few years when data analysis and computing technology improve. Open Government is a combination of ideas: it includes collaborative strategies to engage citizens in government; government releasing data about its own operations, like federal spending data; and government releasing data that it collects on issues of public interest, such as health, environment, and different industries.

While others have produced thoughtful and comprehensive descriptions of Open Data, some refer largely to open government data, which I think of as a category of Open Data (though perhaps the most important one). I've described Open Data as accessible public data that people, companies, and organizations can use to launch new ventures, analyze patterns and trends, make data-driven decisions, and solve complex problems. All definitions of Open Data include two basic features: The data must be publicly available for anyone to use, and it must be licensed in a way that allows for its reuse. Open Data also should be relatively easy to use, although here there are gradations of "openness." And there's general agreement that Open Data should be free of charge or cost just a minimal amount.

Starting with those basic descriptions, the intersection of these three concepts defines the six subtypes of data shown on the diagram. (There's no separate category for the intersection of Big Data and Open Government – anything in that category is also Open Data.) Here are characteristic examples of each, referring to the numbers above.

1. Big Data that's not Open Data. A lot of Big Data falls in this category, including some Big Data that has great commercial value. All of the data that large retailers hold on customers' buying habits, that hospitals hold about their patients, or that banks hold about their credit-card holders, falls here. It's information that the data-holders own and can use for commercial advantage. National security data, like the data collected by the NSA, is also in this category.

2. Open Government work that's not Open Data. This is the part of Open Government that focuses purely on citizen engagement. For instance, the White House has started a petition website, called We the People, to open itself to citizen input. While the site makes its data available, publishing Open Data – beyond numbers of signatures – is not its main purpose.

3. Big, Open, Non-Governmental Data. Here we find scientific data-sharing and citizen science projects like Zooniverse. Big data from astronomical observations, from large biomedical projects like the Human Genome Project, or from other sources realizes its greatest value through an open, shared approach. While some of this research may be government-funded, it's not "government data" because it's not generally held, maintained, or analyzed by government agencies. This category also includes a very different kind of Open Data: the data that can be analyzed from Twitter and other forms of social media.

4. Open Government Data that's not Big Data. Government data doesn't have to be Big Data to be valuable. Modest amounts of data from states, cities, and the federal government can have a major impact when it's released. This kind of data fuels the participatory budgeting movement, where cities around the world invite their residents to look at the city budget and help decide how to spend it. It's also the fuel for apps that help people use city services like public buses or health clinics.

5. Open Data – not Big, not from Government. This includes the private-sector data that companies choose to share for their own purposes – for example, to satisfy their potential investors or to enhance their reputations. Environmental, social, and governance (ESG) metrics fall here. In addition, reputational data, such as data from consumer complaints, is highly relevant to business and falls in this category.

6. Big, Open, Government Data (the trifecta). These datasets may have the most impact of any category. Government agencies have the capacity and funds to gather very large amounts of data, and making those datasets open can have major economic benefits. National weather data and GPS data are the most often-cited examples. U.S. Census data, and data collected by the Securities and Exchange Commission and the Department of Health and Human Services, are others. With the new Open Data Policy, this category will likely become larger, more robust, and even more significant.

The map is still evolving, but this is a start. Let's get the discussion going. I look forward to hearing your thoughts.

What is Open?

Open Knowledge Foundation

https://okfn.org

What is Open?

The Open Definition gives full details on the requirements for ‘open’ data and content. Open data are the building blocks of open knowledge. Open knowledge is what open data becomes when it’s useful, usable and used.

The key features of openness are:

Availability and access: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.

Reuse and redistribution: the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.

Universal participation: everyone must be able to use, reuse and redistribute — there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

What kinds of open data?

There are many kinds of open data that have potential uses and applications:

- Cultural: Data about cultural works and artefacts — for example titles and authors — and generally collected and held by galleries, libraries, archives and museums.
- Science: Data that is produced as part of scientific research from astronomy to zoology.
- Finance: Data such as government accounts (expenditure and revenue) and information on financial markets (stocks, shares, bonds etc).
- Statistics: Data produced by statistical offices such as the census and key socioeconomic indicators.
- Weather: The many types of information used to understand and predict the weather and climate.
- Environment: Information related to the natural environment such presence and level of pollutants, the quality and rivers and seas.
- Transport: Data such as timetables, routes, on-time statistics.

Why open data?

Why should data be open? The answer, of course, depends somewhat on the type of data. However, there are common reasons such as:

- Transparency. In a well-functioning, democratic society citizens need to know what their government is doing. To do that, they must be able freely to access government data and information and to share that information with other citizens. Transparency isn’t just about access, it is also about sharing and reuse — often, to understand material it needs to be analyzed and visualized and this requires that the material be open so that it can be freely used and reused.

- Releasing social and commercial value. In a digital age, data is a key resource for social and commercial activities. Everything from finding your local post office to building a search engine requires access to data, much of which is created or held by government. By opening up data, government can help drive the creation of innovative business and services that deliver social and commercial value.
- Participation and engagement – participatory governance or for business and organizations engaging with your users and audience. Much of the time citizens are only able to engage with their own governance sporadically — maybe just at an election every 4 or 5 years. By opening up data, citizens are enabled to be much more directly informed and involved in decision-making. This is more than transparency: it's about making a full "read/write" society, not just about knowing what is happening in the process of governance but being able to contribute to it.

La faccia umana dei Big data

L'uso dei dati come bene comune aiuterà a capire meglio le realtà più complesse. E a valutare le nostre decisioni

Dino Pedreschi
NOVA24

Il Sole 24 ore

16 Novembre 2014

Prendiamo l'intera fascia sub-sahariana del continente africano. Immaginiamo di disporre dei dati lasciati dagli abitanti di quel territorio attraverso i loro cellulari, con cui si telefona, ci si scambia sms, ci si connette alla rete per accedere a servizi di micro-credito o partecipare sui social media. Immaginiamo di avere accesso a questa impronta digitale per lunghi periodi.
Ecco, ci sono usi assai sorprendenti di queste sorgenti di dati. È possibile, per esempio, costruire mappe dettagliate della mobilità degli abitanti a diverse scale geografiche. Comprendere i pattern di spostamento a livello di singole città, intere regioni o nazioni. Comprendere i flussi giornalieri casa-lavoro. Ragionare sul futuro di infrastrutture, strade, ferrovie, simulando gli scenari più sostenibili e promettenti, in Paesi in cui le decisioni per lo sviluppo possono ancora essere prese. Scoprire per tempo fenomeni migratori inconsueti, comprenderne le ragioni, intervenire subito in caso di emergenze umanitarie.
È possibile anche misurare caratteristiche delle reti sociali e di mobilità che rappresentano il livello di sviluppo economico delle diverse aree. E quindi monitorare il benessere e il suo andamento nel tempo, a livello regionale o cittadino. Anticipare i cambiamenti, positivi o negativi. Avere il polso dello stato di salute di un territorio, prevedere una carestia o una crisi economica che mette a rischio una comunità rurale. Capire più a fondo e più rapidamente le cause scatenanti di un cambiamento in atto, incrociando le informazioni dei social media o di Earth observation dei satelliti. È possibile ragionare sulla diffusione di epidemie, come Ebola. Usare i dati sulla mobilità locale, regionale e internazionale insieme alle informazioni cliniche per mettere a punto modelli che prevedono l'andamento dell'epidemia, se e quando interesserà ogni regione, con quale intensità. Quali scenari aspettarci in seguito ad azioni di profilassi, vaccinazione, restrizioni sui viaggi.
Non si tratta di futurologia ma di esperimenti concreti, perseguiti nei laboratori di data science nel mondo. La D4D challenge (Data for Development), ad esempio, lanciata da Orange Telecom nel 2013 condividendo un grande dataset legato alla telefonia mobile in Costa D'Avorio, ha suscitato una competizione sull'uso creativo di questi dati. Il nostro laboratorio di Big Data analytics e Social mining a Pisa ha sviluppato modelli di traffico grazie a questi dati, rendendo possibili simulazioni che richiedono indagini campionarie altrimenti insostenibili.
Insieme a data scientist di varia estrazione geografica e scientifica stiamo lavorando per definire misurazioni del benessere, della demografia cittadina, dei flussi di mobilità attraverso i big data. Scienziati come Alex Vespignani sviluppano modelli epidemiologici predittivi basati su sistemi complessi e big data. László Barabási, pioniere della scienza delle reti (nell'articolo a fianco in realtà aumentata, da leggere mediante l'app

NòvaAJ, Ndr), della sorprendente prevedibilità del comportamento umano che emerge dai dati. Gli scienziati europei ragionano sulle infrastrutture necessarie per la ricerca e l'innovazione nella big data analytics: l'analogo del Cern per i dati e la conoscenza. Fosca Giannotti ci aggiorna sulla visione del white paper risultante. L'Onu ha sostenuto iniziative come Global Pulse per monitorare la salute del pianeta, e proprio in questi giorni lancia il rapporto «Mobilising the Data Revolution for Sustainable Development», una strategia per condividere e usare i dati per il bene comune. Gli uffici statistici di mezzo mondo si stanno attrezzando per produrre statistiche ufficiali più dettagliate e tempestive con i big data: Emanuele Baldacci (sempre qui di fianco, Ndr) ci aggiorna su questa linea.
C'è ancora tanta strada da fare, ma una cosa è certa: big data non è solo profiling, target marketing estremo e spionaggio del Datagate. C'è una faccia umana dei big data, il loro uso etico e aperto come bene comune, che ci aiuterà a comprendere meglio il mondo complesso e interconnesso che abitiamo, e la portata delle nostre decisioni individuali e collettive.
Dino Pedreschi è professore ordinario di Informatica all'Università di Pisa

Ancora pochi big data nelle aziende europee

Un’indagine condotta su 1.651 imprese dell’Ue rivela un basso tasso di utilizzo e comprensione delle potenzialità di uno strumento che ha ancora molti margini di applicazione

6 nov. 2014

http://www.wired.it/

I big data possono creare nuove opportunità di business per le aziende? Sì, peccato che siano ancora in pochi a sapere cosa siano e a cosa servano. Stando ai dati di un’indagine condotta nel 2013 da IDC, società di ricerche di mercato e consulenza in ambito IT, su 1.651 imprese dell’Unione europea (Italia, Germania, Francia, Spagna e Regno Unito), più della metà delle aziende (il 53%) non ha adottato soluzioni per elaborare grandi quantità di informazioni non strutturate. In base alla ricerca, presentata durante il convegno IDC Big Data & Analytics Conference 2014 a Milano, il 15% delle compagnie si dichiara addirittura non familiare con l’argomento. In altre parole, non ha un’idea precisa di cosa siano.

C’è però chi riconosce invece in questi grossi volumi di dati un’enorme possibilità per comprendere meglio le esigenze dei propri clienti e migliorare prodotti e servizi. Il 24% delle imprese intervistate ha acquistato un’infrastruttura per catturare, scoprire ed analizzare big data e analytics e il 7% ha pianificato di dotarsi di strumenti di questo tipo entro 24 mesi.

I comparti che finora in Europa hanno scommesso di più sui big data sono: telecomunicazioni e media (che hanno scelto questi sistemi soprattutto per interagire meglio con l’utente finale e capire dove si trovano i clienti), servizi finanziari (che usano questi dati principalmente per algoritmi di trading e per profilare i clienti) e manifattura (che li sfrutta principalmente per analizzare le transazioni e interpretare meglio le informazioni provenienti dall’Internet delle cose). Tra i settori che invece prevedono di investire di più in big data entro il 2015 c’è l’healthcare, a dimostrazione che i big data possono essere utili per migliorare i servizi forniti al paziente, analizzare le evoluzioni di malattie e sindromi e anticipare eventuali azioni correttive.

E in Italia? Facendo riferimento a un campione più ristretto (100 imprese), dalla ricerca di IDC emerge che il 30% delle aziende ha puntato sui big data. Se il numero di compagnie prese in esame fosse stato simile a quello relativo allo studio sulle imprese continentali, il valore italiano sarebbe stato più alto rispetto agli altri Paesi Ue. Tuttavia, l’indagine mette in rilievo che il grado di maturità riguardo alla conoscenza e all’utilizzo di questi strumenti è ancora basso. Delle 30 aziende su 100 che hanno adottato sistemi per la gestione dei big data, ben 18 li hanno impiegati in progetti pilota e sperimentazioni tecniche limitate. Le imprese che invece non si sono dotate di servizi ad hoc hanno legato il loro no a diverse cause. Prima tra tutte, la carenza di competenze specifiche, seguita dalla non sufficiente comprensione dell’utilità dei big data e dai prezzi delle tecnologie, ritenuti troppo elevati.

I settori che in Italia hanno più investito nei big data sono la finanza, i trasporti, il retail e la pubblica amministrazione (che li utilizza soprattutto in ottica open data oppure per la lotta all’evasione fiscale). Per

quanto solo una minoranza delle imprese abbia compreso il potenziale di queste tecnologie, gli investimenti nel settore stanno crescendo a ritmo sostenuto e si prevede che l'incremento sarà ancora maggiore nei prossimi anni. Se nel 2013 le aziende italiane hanno investito 148,6 milioni di euro in soluzioni per big data, nel 2018 si arriverà a quota 373,3 milioni, diretti soprattutto all'acquisto di software (36%), servizi IT (24%) e strumenti per lo storage (24%).

Tutti i falsi miti dell'Open data

Gli analisti stimano un potenziale di 3 miliardi di dollari. Ma ad oggi le previsioni economiche non si sono realizzate

Maurizio Napolitano

NOVA24

Il Sole 24 Ore

26 ottobre 2014

Una delle continue promesse dell'open data è legata al mondo imprenditoriale. Studi prodotti da diversi attori, fra cui vale la pena citare aziende di consulenza come Deloitte e McKinsey, hanno stimato una ricchezza potenziale di 3 miliardi di dollari (o più) in vari settori. Su questo la Commissione europea sta investendo con tutti gli strumenti in proprio possesso. Il motto «trasformiamo l'informazione del settore pubblico in oro» ha dato vita ad alcune riforme (ad esempio. la direttiva Psi) e alla destinazione di fondi per incentivare il riuso e il rilascio di open data.

Al momento però queste previsioni non sembrano realizzarsi, in quanto il percorso è ancora molto lungo.

Open data vuol dire semplicemente rendere disponibili i dati a chiunque per qualunque scopo. Si tratta quindi di abilitarne il riuso dandone il permesso e usando tecnologie abilitanti. Un concetto tanto semplice quanto complesso perché si tratta di un cambiamento culturale.

Nello scenario dell'open data il ruolo della pubblica amministrazione è fondamentale: ha il mandato di raccogliere i dati per fare il bene comune, e dato che i dati sono un bene comune digitale, il fatto di renderli disponibili a chiunque per qualunque scopo viene visto come un dovere.

Dietro questo però c'è un lavoro non indifferente fatto di riorganizzazione dei processi, introduzione di nuove tecnologie e necessità di incrementare competenze digitali.

Nello scenario specifico poi della Pa il tema dell'open data è maggiormente indirizzato verso il concetto di trasparenza restringendo quindi gli scenari imprenditoriali.

In realtà l'open data è una opportunità per tutti e le stesse aziende aprendo i dati possono trarne vantaggio. Aprire dati permette diversi scenari quali: avere una ottima vetrina per mostrare cosa si è in grado di offrire, creare nuove sinergie con terzi che riescono a farne un uso diverso, aumentare la fiducia dei propri clienti permettendo di tracciare la propria produzione, coinvolgere comunità, individuare nuovi talenti da coinvolgere nelle proprie attività eccetera.

Esistono poi comunità che, per risolvere alcune necessità, creano dei beni comuni digitali come il caso di OpenStreetMap il cui fine è quello di avere una banca dati geografica dell'intero pianeta.

Questo progetto che coinvolge oltre 2 milioni di persone nel mondo si dimostra ormai come un ottimo esempio di progetto open data da dove nascono tantissimi prodotti di interesse per tantissimi settori.

E qui un altro mito da sfatare: l'open data non è strumento esclusivo del mondo Ict o, peggio, destinato solo al mercato delle applicazioni per smartphone.

I dati sono alla base della gerarchia della conoscenza, e spesso sono utilizzati per prendere decisioni. Si tratta di una lunga filiera che attraversa vari macropassaggi quali: l'acquisizione, la pulizia e trasformazione, l'analisi e la presentazione. Ciascuno di questi passaggi può dare vita a nuovi scenari di mercato. L'accesso ai dati permette a molti liberi professionisti di elaborare strategie, produrre report, pianificare eccetera. L'informatica è lo strumento attraverso cui si elabora tutto questo per il semplice fatto che sono archiviati in formato digitale. Allo stato attuale il concetto di open data è pertanto ridotto a dati della pubblica amministrazione il più delle volte utilizzati per tracciare la trasparenza di governo e distribuiti con formati obsoleti, ma, in realtà, il potenziale che può sprigionare è molto alto e necessita di uno sforzo più ampio. Il salto di qualità avviene quando i dati sono resi disponibili in maniera automatizzata, aggiornati tempestivamente e distribuiti non solo come file ma anche attraverso servizi. Gli open data, anche quelli attualmente disponibili, rivestono comunque un ruolo fondamentale per creare una infrastruttura di base su cui elaborare nuovi scenari. Un esempio molto esaustivo viene dai dati territoriali. Questi, per loro natura, sono di interesse trasversale a moltissimi settori e richiedono anche un aggiornamento costante per cui, più si è vicini al luogo che i dati descrivono, e maggiore sarà la loro qualità. Questo è un compito affidato alla pubblica amministrazione che, però, il dato lo gestisce all'interno del compito che deve svolgere. Le aziende possono inserirsi nella filiera andando a rielaborare il dato, rendendolo più facile nel riuso, rappresentandolo in maniera diversa e costruendo sopra nuovi scenari. Un altro ruolo però, ancora più importante, degli open data è quello di essere strumenti necessari per migliorare le competenze digitali. Le elaborazioni dei dati, l'integrazione fra di loro e tutte le "magie" dell'intelligenza artificiale che vanno sotto i nomi di machine learning, data mining, pattern matching, linked data... necessitano di dati per essere apprese. Per fare un semplice esempio il corso "big data" del Coursera, il famoso servizio di formazione online gratuito, si basa sull'elenco dei passeggerei del Titanic. Di certo un file non così "big" come si pensa, ma da cui è possibile apprendere il necessario per migliorare le proprie competenze digitali e diventare dei data scientist.

L'open data ha il solo scopo di convidere la conoscenza, pertanto di stimolare il riuso e quindi di permettere a chiunque di migliorare le proprie competenze e, pertanto, conquistare nuovi mercati.

Big Data: tutto quello che c’è da sapere sulla professione del Data scientist

di Giuditta Mosca

NOVA24

Il Sole 24 Ore

26 ottobre 2014

Quella del data scientist è stata definita dall'economista Hal Ronald Varian «la professione più sexy del futuro», laddove l'aggettivo assume l'accezione di «interessante».

Cosa fa un data scientist

Come suggerisce il nome, analizza dati per fornire al management le informazioni utili ad assumere decisioni e disegnare strategie. Per lunghi anni si è parlato dell'importanza dei dati, ora nasce l'esigenza di saperne fare buon uso. Benché si possa credere che la figura del data scientist sia appropriata solo alle grandi aziende, un simile profilo si rivolge a qualsiasi realtà, dalle Pmi alle multinazionali. Di norma viene inquadrato tra i manager, anche dal punto di vista della retribuzione, proprio perché è con gli altri manager che deve dialogare. È una figura professionale nuova e, in qualche modo, ancora da definire. Lo scienziato dei dati non è solo un'analista, non è solo uno stratega del business, non è solo un marketer così come non è solo un information manager. Il frutto delle sue analisi copre trasversalmente tutti i reparti di un'azienda, trasformando i dati in informazioni comprensibili affinché per i vertici le strategie da assumere siano chiare e in qualche modo obbligate. Ciò si adatta anche alle Pa. Dino Pedreschi, professore ordinario di Informatica

all'Università di Pisa, descrive lo scienziato dei dati come: «Una figura che deve avere più competenze. La prima è sapere gestire, acquisire, organizzare ed elaborare dati. La seconda competenza è di tipo statistico, ovvero il sapere come e quali dati estrarre, la terza capacità è una forma di storytelling, il sapere comunicare a tutti, con diverse forme di rappresentazione, cosa suggeriscono i dati». Non basta quindi una formazione in statistica, in economia o in informatica, tutte doti utili alla figura del data scientist ma che necessitano di essere mixate sapientemente.

Perché c'è bisogno di data scientist

La risposta in due sostantivi: produttività e cambiamenti. Da una parte cambiano i modelli di business delle aziende, così come cambiano le loro politiche economiche e i mercati e, dall'altra parte, vige la necessità di aumentare produttività e profitti. Un esempio reale arriva da Mario Alemi, data scientist italiano (laureato in fisica): «Le email personalizzate in base ai gusti letterari dei clienti hanno generato, nei negozi, il 27% delle vendite in più di quelle conseguite con le email generiche». Un'indagine McKinsey rileva che, negli Usa, mancano tra i 140 e i 190mila data scientist, ciò testimonia quali prospettive possa avere la professione. «Quella del data scientist sarà nei prossimi anni tra le figure più ricercate nel mondo del lavoro – continua Pedreschi – e sono sempre di più le università che preparano percorsi post-universitari aperti a tutti i curriculum».

Quale formazione è necessaria

Ci sono decine tra atenei e centri studio che offrono formazione specifica. Giuseppe Ragusa, direttore del master in Big Data Analytics della Luiss, in collaborazione con Oracle, riassume così le qualifiche necessarie per abbracciare la professione: «Il data scientist è un animale a tante teste, deve avere tre set di skill: una preparazione informatica molto solida, una buona comprensione degli aspetti tecnologici e allo stesso tempo è un conoscitore degli aspetti aziendali. Una figura professionale dotata di competenze trasversali e capace di relazionarsi con il management dell'azienda». Anche Dino Pedreschi apporta la sua esperienza di docente universitario e parlando del master in Big Data Analytics e Social Mining, dell'Università di Pisa, che partirà a febbraio 2015 spiega: «Stiamo organizzando un master apposito che si rivolge a laureati di qualsiasi provenienza, perché non ci sono requisiti stretti in ingresso, se non la voglia di mettersi alla prova con tutte le competenze necessarie, in collaborazione con il mondo industriale».

La situazione in Italia

I poli mondiali sono Usa e Uk, laddove nei primi anni del Duemila si erano già create metodologie e procedure. Alle nostre latitudini le aziende cominciano a concepire la necessità di una simile figura e cercano di formarla al proprio interno. Nel frattempo, dice Ragusa, le imprese chiedono alle università i dati di chi frequenta i corsi. Perché l'Italia stenta a carburare lo spiega Alemi: «La nostra cultura è prettamente umanistica, siamo sempre un passo indietro quando si parla di discipline scientifiche, ma sono ottimista, questo gap verrà colmato nei prossimi 5-10 anni». Il data scientist lavora con i big data ed è in questa direzione che bisogna muoversi; le anagrafiche sono il primo patrimonio di un'azienda, concetto ancora non del tutto consolidato in Italia e questo, da solo, spiega già gran parte dell'handicap che abbiamo in materia di scienza dei dati.

Cosa aspettarsi dal futuro

Una rilevazione voluta da Emc Data Science segnala che l'assenza di risorse uomo sufficientemente preparate e aziende non strutturate per il data science si equivalgono, entrambe con il 32%, nell'elenco dei principali freni allo sviluppo sia della professione, sia della crescita dell'intero settore che, ancora lontano da misurazioni di tipo economico, è comunque una costola del comparto dei big data il quale, secondo Gartner, varrà 26miliardi di dollari entro la fine del 2015.

Data Scientist, il professionista più sexy del XXI secolo

Un po' matematico, un po' informatico. Deve capire di statistica, saper usare un foglio excel ed aiutare l'azienda per la quale lavora a prevedere il futuro. Non è uno sciamano ma un nuovo professionista, lo 'scienziato dei dati'. Che negli Usa è tra i più contesi

http://www.rainews.it/dl/rainews/articoli/data-scientist-il-lavoro-sexy-50712477-5a83-4682-8f73-522428eb3281.html

di Celia GuimaraesRoma

26 marzo 2014

Hal Varian, chief Economist di Google, in un'intervista al New York Times l'ha definito "il lavoro più sexy del XXI secolo".

Il 'Data scientist', con competenze trasversali in statistica, matematica e informatica, è tra le figure più ricercate del mondo del lavoro, con una domanda che supera di gran lunga la disponibilità di candidati.

Esperti in formazione citano una recente ricerca secondo la quale l'80% delle offerte di lavoro rimane insoddisfatta per mancanza di personale. Ovviamente stiamo parlando del mercato Usa, dove il termine 'Data scientist' è stato coniato.

In Italia ancora no

"La retribuzione invece è un aspetto sconosciuto", racconta Marco Russo, consulente e formatore di questa nuova professione (Business Intelligence consultant and trainer, recita il suo profilo). "Il valore dipende dal mercato, oggi non ce ne sono [professionisti] ma la domanda in Italia è bassissima. Negli Usa cominciano ad esserci ricerche su questo ruolo, lì il valore di mercato è più alto perché c'è domanda ma non c'è ancora offerta". Negli Stati Uniti le offerte vanno da 70 a 300 mila dollari l'anno, anche se la retribuzione-tipo si aggira sui 100 mila dollari.

Chiamatelo analista

Marco Russo ci ha raccontato gli aspetti meno conosciuti di questa professione. "Data scientist è un termine creato per identificare una figura professionale che ha un percorso all'interno dell'azienda, spesso casuale e non formalizzato".

Più che 'scienziato', Russo preferisce chiamarlo analista, che può lavorare in un'azienda commerciale come in un'università come in una Onlus. La caratteristica del suo lavoro è quella di analizzare dati che oggi sono disponibili a ritmo esponenziale sia per volume che per tipo, ai quali vanno aggiunti canali del tutto nuovi come ad esempio i social network.

Si tratta di dati complessi e il Ds "deve avere competenze trasversali, non è un informatico ma deve conoscere le basi dei modelli di dati, deve avere conoscenze di base di statistica e matematica, deve persino saper usare un foglio excel per incrociare dati". Mettiamo, per fare un esempio semplice, che debba creare delle tabelle con i dati degli amici e le loro squadre di calcio, oppure l'attendibilità di un sondaggio pubblicato da un giornale.

Predizione, non premonizione

Per fare il suo lavoro si serve di tecnologia, matematica e statistica per presentare dati che formulano un'ipotesi. "E anche il modo di presentarli, di rendere graficamente la visualizzazione è importante",

sottolinea Russo. Perché una volta terminato, il lavoro servirà a supportare le conclusioni di chi nell'azienda ha potere decisionale.

Ecco perché il Ds fornisce elementi predittivi ("in termini tecnici significa usare algoritmi predittivi e data mining"), per fare previsioni su andamenti. L'esempio classico, secondo Russo, è quello dei profili utilizzati da Amazon per proporre prodotti ai clienti.

Un lavoro sexy

Il Chief Economist di Google, Hal Varian, che cosa intendeva parlando di professione 'sexy' (anche se in realtà parlava di statistici)? "Più che sexy è attraente, challenging, una sfida. Porta il Data scientist in una selva oscura dove c'è l'incertezza di un mondo nuovo dove i dati hanno a che fare con il comportamento delle persone. Ma è un lavoro molto duro incrociare dati", ricorda Marco Russo.

In pratica si lavora in quattro fasi: estrazione dati; valutazione della loro qualità ("spesso non sono compatibili o danno troppo margine d'errore"); creazione di un modello; interrogazione dei i dati per il test di validità. Tutto questo per arrivare ad una sintesi che dica, ad esempio, "il 20% dei giovani di età x fanno [la cosa] y".

Una carriera per giovani?

Dipende dal giovane, dice Marco Russo. Non ci sono percorsi, l'esperienza va fatta in ambito aziendale, magari con degli stage mentre si studia. Il corso universitario più vicino alle competenze del Data scientist è Ingegneria gestionale, che però ha finalità diverse. Il Data scientist invece pensa e fa le cose in modo autonomo, e la sua autorevolezza arriva con il tempo. In definitiva, il buon Data scientist è quello che sa 'inventarsi', farsi manager di se stesso.

The GovLab Selected Readings on the Economic Impact of Open Data

Posted on July 31, 2014 by Kevin Van Nguyen in

http://thegovlab.org/

GovLab ha realizzato un'interessante selezione di 22 articoli e report sul (sempre dibattuto) tema dell'impatto economico degli Open Data.

As part of an ongoing effort to build a knowledge base for the field of opening governance by organizing and disseminating its learnings, the GovLab Selected Readings series provides an annotated and curated collection of recommended works on key opening governance topics. In this edition, we explore the literature on The Economic Impact of Open Data. To suggest additional readings on this or any other topic, please email biblio@thegovlab.org.

Open data is publicly available data – often released by governments, scientists, and occasionally private companies – that is made available for anyone to use, in a machine-readable format, free of charge. Considerable attention has been devoted to the economic potential of open data for businesses and other organizations, and it is now widely accepted that open data plays an important role in spurring innovation, growth, and job creation. From new business models to innovation in local governance, open data is being quickly adopted as a valuable resource at many levels.

Measuring and analyzing the economic impact of open data in a systematic way is challenging, and governments as well as other providers of open data seek to provide access to the data in a standardized way. As governmental transparency increases and open data changes business models and activities in many

economic sectors, it is important to understand best practices for releasing and using non-proprietary, public information. Costs, social challenges, and technical barriers also influence the economic impact of open data.

These selected readings are intended as a first step in the direction of answering the question of if we can and how we consider if opening data spurs economic impact.

Selected Reading List:

Carla Bonina — New Business Models and the Values of Open Data: Definitions, Challenges, and Opportunities. – Paper provides an introduction to open data and open data business models, evaluating their potential economic value and identifying future challenges for the effectiveness of open data

John Carpenter and Phil Watts — Assessing the Value of OS OpenData™ to the Economy of Great Britain – Synopsis – A study examining the economic impact of the OS OpenData initiative to the economy of Great Britain.

Capgemini Consulting. — The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data. Capgemini Consulting – Paper analyzes trends in open government data interventions among different countries with goal of identifying best practices for stimulating economic impact and creating economic value.

Deloitte — Open Growth: Stimulating Demand for Open Data in the UK. – Explores emerging data-driven business models and its potential to stimulate demand for open data in the UK economy.

Nicholas Gruen, John Houghton and Richard Tooth — Open for Business: How Open Data Can Help Achieve the G20 Growth Target — Assesses exiting literature, in-depth case studies, and proposes key strategies for institutions to open data to spur economic development and growth.

Felipe I Heusser — Understanding Open Government Data and Addressing Its Impact (draft version) – Early research on open data initiatives and its economic impact in developing countries.

Alex Howard — San Francisco Looks to Tap into the Open Data Economy – This article examines San Francisco's use of open data in municipal governance.

Noor Huijboom and Tijs Van den Broek — Open Data: An International Comparison of Strategies — This paper examines five countries and their open data strategies, identifying key features, main barriers, and drivers of progress for of open data programs.

James Manyika, Michael Chui, Diana Farrell, Steve Van Kuiken, Peter Groves, and Elizabeth Almasi Doshi —Open Data: Unlocking Innovation and Performance with Liquid Innovation — Focuses on quantifying the potential value of open data in critical domains of the global economy.

Alida Moore — Congressional Transparency Caucus: How Open Data Creates Jobs — Summary of the March 24th briefing of the Congressional Transparency Caucus on the need to increase government transparency through adopting open data initiatives for job creation.

Andrew Stott —Open Data for Economic Growth— Examines five archetypes of businesses using open data, and provides recommendations for governments trying to maximize economic growth from open data.

Annotated Selected Reading List (in alphabetical order)

Bonina, Carla. New Business Models and the Values of Open Data: Definitions, Challenges, and Opportunities. NEMODE 3K – Small Grants Call 2013. http://bit.ly/1xGf9oe

In this paper, Dr. Carla Bonina provides an introduction to open data and open data business models, evaluating their potential economic value and identifying future challenges for the effectiveness of open data, such as personal data and privacy, the emerging data divide, and the costs of collecting, producing and releasing open (government) data.

Carpenter, John and Phil Watts. Assessing the Value of OS OpenData™ to the Economy of Great Britain – Synopsis. June 2013. Accessed July 25, 2014. http://bit.ly/1rTLVUE

John Carpenter and Phil Watts of Ordnance Survey undertook a study to examine the economic impact of open data to the economy of Great Britain. Using a variety of methods such as case studies, interviews, downlad analysis, adoption rates, impact calculation, and CGE modeling, the authors estimates that the OS OpenData initiative will deliver a net of increase in GDP of £13 – 28.5 million for Great Britain in 2013.

Capgemini Consulting. The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data. Capgemini Consulting. Accessed July 24, 2014. http://bit.ly/1n7MR02

This report explores how governments are leveraging open data for economic benefits. Through using a compariative approach, the authors study important open data from organizational, technological, social and political perspectives. The study highlights the potential of open data to drive profit through increasing the effectiveness of benchmarking and other data-driven business strategies.

Deloitte. Open Growth: Stimulating Demand for Open Data in the UK. Deloitte Analytics. December 2012. Accessed July 24, 2014. http://bit.ly/1oeFhks

This early paper on open data by Deloitte uses case studies and statistical analysis on open government data to create models of businesses using open data. They also review the market supply and demand of open government data in emerging sectors of the economy.

Gruen, Nicholas, John Houghton and Richard Tooth. Open for Business: How Open Data Can Help Achieve the G20 Growth Target. Accessed July 24, 2014, http://bit.ly/UOmBRe

This report highlights the potential economic value of the open data agenda in Australia and the G20. The report provides an initial literature review on the economic value of open data, as well as a asset of case studies on the economic value of open data, and a set of recommendations for how open data can help the G20 and Australia achieve target objectives in the areas of trade, finance, fiscal and monetary policy, anti-corruption, employment, energy, and infrastructure.

Heusser, Felipe I. Understanding Open Government Data and Addressing Its Impact (draft version). World Wide Web Foundation. http://bit.ly/1o9Egym

The World Wide Web Foundation, in collaboration with IDRC has begun a research network to explore the impacts of open data in developing countries. In addition to the Web Foundation and IDRC, the network includes the Berkman Center for Internet and Society at Harvard, the Open Development Technology Alliance and Practical Participation.

Howard, Alex. San Francisco Looks to Tap Into the Open Data Economy. O'Reilly Radar: Insight, Analysis, and Reach about Emerging Technologies. October 19, 2012. Accessed July 24, 2014. http://oreil.ly/1qNRt3h

Alex Howard points to San Francisco as one of the first municipalities in the United States to embrace an open data platform. He outlines how open data has driven innovation in local governance. Moreover, he discusses the potential impact of open data on job creation and government technology infrastructure in the City and County of San Francisco.

Huijboom, Noor and Tijs Van den Broek. Open Data: An International Comparison of Strategies. European Journal of ePractice. March 2011. Accessed July 24, 2014. http://bit.ly/1AE24jq

This article examines five countries and their open data strategies, identifying key features, main barriers, and drivers of progress for of open data programs. The authors outline the key challenges facing European, and other national open data policies, highlighting the emerging role open data initiatives are playing in political and administrative agendas around the world.

Manyika, J., Michael Chui, Diana Farrell, Steve Van Kuiken, Peter Groves, and Elizabeth Almasi Doshi. Open Data: Unlocking Innovation and Performance with Liquid Innovation. McKinsey Global Institute. October 2013. Accessed July 24, 2014. http://bit.ly/1lgDX0v

This research focuses on quantifying the potential value of open data in seven "domains" in the global economy: education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance.

Moore, Alida. Congressional Transparency Caucus: How Open Data Creates Jobs. April 2, 2014. Accessed July 30, 2014. Socrata. http://bit.ly/1n7OJpp

Socrata provides a summary of the March 24th briefing of the Congressional Transparency Caucus on the need to increase government transparency through adopting open data initiatives. They include key takeaways from the panel discussion, as well as their role in making open data available for businesses.

Stott, Andrew. Open Data for Economic Growth. The World Bank. June 25, 2014. Accessed July 24, 2014. http://bit.ly/1n7PRJF

In this report, The World Bank examines the evidence for the economic potential of open data, holding that the economic potential is quite large, despite a variation in the published estimates, and difficulties assessing its potential methodologically. They provide five archetypes of businesses using open data, and provides recommendations for governments trying to maximize economic growth from open data.

Ainsley Sutherland also contributed to this post.

More than economics: The social impact of open data

Alex Howard states that focusing on publishing open data with economic value shouldn't preclude or take too much focus away from digitizing and releasing data with other societal value.

By Alex Howard July 31, 2014

http://www.techrepublic.com/

The national governments of the US, the UK, and other G7 nations have been focusing more attention on the economic value of open data, as opposed to broader societal benefits.

While pointing to evidence that open data fuels economic activity is a good rationale for the release of relevant data sets, it's far from the only impact that releasing government data can have upon the world. As I've explored in past columns, publishing open data can increase resilience against climate change, offer insight into healthcare costs and outcomes, protect consumers, and fuel accountability and transparency.

If national governments are going to invest time, money, and public attention on releasing data, they should also focus upon releases that have social benefits as well as economic outcomes. Last week, looking for fresh examples, outcomes, and emerging issues around these issues, I attended a forum on the social impact of open data hosted by the Center for Data Innovation in Washington, DC. (Video of the event is embedded below.)

If you watch, you'll hear remarks on the social impact of open data (PDF) by Maureen Ohlhausen, Commissioner of the Federal Trade Commission (FTC), followed by a panel discussion between Daniel Castro, director of the Center for Data Innovation, Sandra Moscoso, deputy program manager at the World Bank, Brian Rayburn, lead data scientist at Symcat, and Emily Shaw, national policy manager at the Sunlight Foundation.

"Often, when people talk about government and data they focus on government as a consumer of information and how government should or should not be limited in the data it can collect and use," said Ohlhausen. "We have an entire section of constitutional law dedicated to that topic."

But there is another aspect of government data that isn't discussed as much, except perhaps by the people in this room: Government as a producer of data. Federal, state, and local governments generate and store 3 massive amounts of data about themselves, about us, and about the world around us. Even before the very first U.S. census report, government has been producing large -- and increasing -- amounts of data. Government produces many types of data: Personal data, such as social security earnings, tax information, unemployment filings, and voter registration; societal data such as demographics, employment estimates, and economic indicators; and impersonal or scientific data, such as weather and climate measurements and geolocation data. There is great potential in applying powerful new big data tools to the rich troves of government data. The private sector could use the wide range of government-produced data to reveal new insights into difficult problems in nearly every area of human endeavor.

The discussion took place in the context of a new section of Data.gov that profiles companies that use government data as a way of demonstrating the impact of its publication. The profiles raised a few eyebrows this past April, when the federal open data platform only featured examples of the economic impacts of open data. As readers of this column know, free publishing government data in a machine-readable format, under an open license, can have salutary economic outcomes ranging from real estate, health, transit, energy, consumer finance, and weather.

Over the course of the event, the panelists advocated for the release of open data that benefit citizens, not just startups and established businesses. To put it simply, beyond rationales of increased efficiency, reduced costs, increased productivity, and economic growth that will spur the release of new data, there's considerable potential for open data releases to extend to positive social justice, environmental, educational, public safety and health outcomes.

Ohlhausen outlined a role for the FTC in regulating and guiding the publication of open data and its use in data analysis:

"By understanding the limits of big data and emphasizing the need for human judgment in the use of such tools, the FTC can help tamp down hype over big data," she said. "The FTC can help create a healthier regulatory atmosphere by critically evaluating the claims of both the pop-science promoters of big data as a 'magic bullet' solution and the naysayers who fear massive consumer harm from all-knowing algorithms. A realistic understanding of big data's potential will help the agency to identify and focus on actual harms to consumers, if they occur."

The return on investment for open government goes beyond making government institutions and services more transparent, and the people that run them more accountable for the use of taxpayer dollars: In systems of governance that are of the people, for the people, and by the people, open government provides access to information about how those people are being governed and new opportunities to participate in that governance. That means that focusing on publishing open data with economic value shouldn't preclude or take too much focus away from digitizing and releasing data with other societal value.

There's also potential for increased risks to privacy, security, and discrimination, if rules, regulations, norms, ethics, and a careful approach to enterprise inventories, digitization, and data publishing aren't undertaken as part of the process, or fuel the creation of applications and services that favor people who already are privileged in society. Ohlhausen spoke to those issues in her remarks:

"Obviously many -- perhaps even the majority -- of government data sets have nothing to do with 'personally identifiable information,'" she said.

Open access to many scientific and economic data sets, for example, raises no privacy risks. However, opening other useful data sets may raise some privacy concerns. For example, applying big data techniques to government health data or education records could help address the most pressing societal issues we face, but people understandably worry about how such information is used and shared. The FTC can guide other government agencies on how to open access to data while mitigating privacy risks through aggregation, de-identification, use-based limitations, and other techniques. Furthermore, the FTC must continue to explore how to resolve the tension between the promise of big data and certain Fair Information Practice Principles such as notice and purpose limitation and data minimization, which, strictly applied, could hinder big data's promise.

I published a series of tweets during the event with pictures, links, and references to cited research, projects, and services during the event.

[…]

During the brief question and answer period that followed, I had an opportunity to question the FTC Commissioner about the agency's open data practices and took it. (You can watch her answers here.) To her credit, Ohlhausen followed up on Twitter with answers to my questions.

In her replies, she shared a link to the FTC's open government plan and examples of newly released datasets, including a .csv of consumer complaints that she references. I found that the data didn't include the name of individual companies, only aggregates by industry.

I hope the FTC takes a proactive approach to converting any data that it still publishes in PDFs as structured data online, leading by example, and uses Freedom of Information Act requests to prioritize future releases, including which companies are subject to the most complaints.