Ricerca di lavoro e metodologie di web data mining

Il profilo del Data scientist nelle inserzioni on-line

di Achille P. Paliotta

Abstract: La *Data science* (DS) è un fenomeno emergente, soprattutto negli Stati Uniti, tanto da essere considerata la professione simbolo dell'epoca attuale. In questo contesto, la richiesta di "Data scientist" sembra essere elevata, in costante crescita da alcuni anni, superando anche quella dello statistico (Google Trends), mentre in Italia si iniziano a vedere solo alcuni indizi embrionali nel campo della ricerca di personale qualificato *on-line*. Qui viene presentato uno studio di caso sperimentale di estrazione di dati, non strutturati, da un motore di ricerca verticale (Indeed.com) con l'obiettivo di individuare le competenze/abilità del profilo professionale del "Data scientist". Tale sperimentazione vuole essere la prosecuzione, con altre modalità, di una delle rilevazioni più datate dell'ISFOL, la *Domanda di lavoro qualificato* (DLQ). La metodologia utilizzata ha mostrato di essere efficace e può, pertanto, essere applicata a rilevazioni riguardanti una mole maggiore di dati. L'indagine ha permesso di individuare e ricostruire il profilo di "Data scientist" facendo uso del modello tripartito delle competenze ISFOL. I risultati ottenuti mostrano, infine, una significativa differenza tra annunci scritti in italiano e quelli in inglese soprattutto per quel che riguarda le competenze tecnico professionali.

Parole chiave: Annunci di lavoro on-line; Data science; Professioni innovative

"Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician"

Josh Wills, *@josh_wills*, 3 mag 2012

Introduzione

Nell'ottobre 2012 la rivista, nota a livello internazionale, "The Harvard Business Review" (Davenport & D.J. Patil, 2012) dedicava un articolo, che si sarebbe rivelato essere seminale, a una nuova professione, lo "scienziato dei dati" ("Data scientist"), il quale veniva così sommariamente descritto: "it's a high-ranking professional with the training and curiosity to make discoveries in the world of big data". La coniazione del termine veniva retrodatata al 2008 e accreditata a Jeff Hammerbacher e a D.J. Patil, i quali si fecero promotori dei primi gruppi di lavoro di *Data science* (DS) a Facebook Inc. e LinkedIn Inc., rispettivamente. Il titolo dell'articolo qui citato rimandava direttamente a una citazione di Hal Varian, il quale aveva "etichettato" lo "scienziato dei dati" "the sexiest job of the 21st century" e così articolava le sue convinzioni in materia. "I think statisticians are part of it, but it's just a part. You also want to be able to visualize the data, communicate the data, and utilize it effectively. But I do think those skills—of being able to access, understand, and communicate the insights you get from data analysis—are going to be extremely important. Managers need to be able to access and understand the data themselves"[1]. Tale etichetta rimaneva, in seguito, strettamente legata alla professione, sia nell'immaginario collettivo sia nella pubblicistica degli addetti ai lavori.

In questo brevissimo lasso di tempo, la DS diveniva un fenomeno emergente, soprattutto negli Stati Uniti, insieme ai *big data*, tanto da assurgere a iperbole comunicativa (alla stregua di un vero e proprio *hype,* e come tale recensito dalla Gartner Inc.)[2]. Come conseguenza diretta di questo successo si iniziava a parlare, assai presto, della carenza di figure professionali adeguatamente formate (*skill shortage*) e, di conseguenza, dell'inevitabile investimento in attività formative[3], al fine di poter far fronte alle richieste di reclutamento, oramai pressanti, da parte delle imprese (Provost & Fawcett, 2013; Bughin, 2016, p.12)[4]. Diverse ricerche svolte da primarie società di consulenza, tra altre, sembravano confermare, tali assunti.

In Italia, invece, la tematica solo adesso ha raggiunto la grande stampa di opinione e, aspetto sicuramente più rilevante, solo ora inizia ad essere considerata una delle

[1] *Hal Varian on how the Web challenges managers* in McKinsey & Company Insights & Publications, January 2009, *http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers.*

[2] Heather Levy, *What's New in Gartner's Hype Cycle for Emerging Technologies 2015*, 20 October, 2015, *http://www.gartner.com/smarterwithgartner/whats-new-in-gartners-hype-cycle-for-emerging-technologies-2015/*.

Cfr. anche, William Vorhies, *Big Data Falls Off the Hype Cycle*, 17 August, 2015, *http://www.datasciencecentral.com/profiles/blogs/big-data-falls-off-the-hype-cycle.*

[3] I corsi di DS figurano al secondo posto nella graduatoria dei certificati rilasciati dalla più grande piattaforma di MOOC's al mondo, con 18 milioni di iscritti globali, Coursera Inc. «We dug into the professional skills learners were most eager to gain this year, as defined by the Specializations with the highest rates of course completers posting their Certificates to LinkedIn. The result is our first-ever list of the "Top 10 Most Coveted Coursera Certificates"», *Our Top 10 Most Coveted Certificates of 2015*, Coursera Blog, 16 dicembre 2015, *http://blog.coursera.org/post/135338805637/our-top-10-most-coveted-certificates-of-2015.*

[4] «The major performance impact of big data resides in the close complementarity between big data IT investment and labour skills».

leve strategiche principali da parte delle imprese ma la situazione è, invero, tutta da indagare. Un passo significativo, dell'ultimo rapporto dell'Associazione italiana per l'Information Technology (Assinform), viene dedicato alla carenza dei profili nel settore e alla necessità di sviluppo di attività formative al riguardo. "La trasformazione digitale rischia infatti di esser frenata anche dalla carenza di competenze. Sta crescendo il *gap* tra domanda e offerta di profili specializzati nelle nuove tecnologie ICT e nei nuovi *business* digitali. Ci sono mezzo milione di posizioni di lavoro disponibili che non si riesce a coprire per mancanza di skills. È urgente intervenire sul sistema della formazione, creando così nuove opportunità non solo per il sistema, ma per centinaia di migliaia di giovani"[5].

Delle tante verifiche empiriche che è possibile avere dell'emergenza di tale fenomeno una delle più attendibili rimane, senz'altro, quella del reclutamento di tale profilo, da parte del sistema produttivo nazionale. In questo modo, l'asserito *mismatch* può essere confermato o meno e, oggigiorno, una tale indagine può essere svolta anche in maniera davvero innovativa rispetto, ad esempio, alla tradizionale "Domanda di lavoro qualificato" (DLQ) dell'ISFOL (Paliotta, 2014), basata sulle inserzioni di lavoro a mezzo stampa, pubblicate sui principali quotidiani italiani. Grazie alla diffusione delle nuove tecnologie (e alle correlate inedite capacità computazionali dei processori attuali) nonché alla proliferazione dei siti di incontro tra domanda ed offerta di lavoro (Paliotta, 2015) si può applicare, a tale rilevazione sperimentale, tutta la panoplia dei nuovi strumenti analitici propri della DS. Per questa ragione si è deciso di rilevare le inserzioni di lavoro pubblicate sui siti *web*, facendo uso del maggiore motore di ricerca verticale, Indeed.com. Dopo aver estratto le informazioni utili dal sito si sono applicate tutte le tecniche di pulizia e trattamento dei dati al fine di identificare le caratteristiche distintive di tale profilo in termini di competenze/abilità. Riguardo al *framework* teorico, in questa sede, si farà uso del modello delle competenze tripartito messo a punto dall'ISFOL il quale distingue tra competenze di base, trasversali e tecnico professionali.

Nel corso di questo testo, di carattere prettamente esplorativo, si cercherà di rispondere, dunque, alle seguenti domande: 1) le tecniche automatizzate di estrazione dei dati dalla rete sono proficuamente utilizzabili per tale tipo di indagini?; 2) quali sono le competenze/abilità principali che le imprese stanno attivamente cercando per questi profili?; 3) vi sono differenze sostanziali tra il "Data scientist" e alcuni profili affini, quali il "Business analyst" e il "Data analyst"?.

Il testo è strutturato nel modo seguente. *In primis*, verrà delineata l'emersione di questa nuova figura professionale nel contesto statunitense, in quanto è ancora di carente diffusione in quello nazionale. Nel secondo paragrafo verrà descritta la metodologia dell'indagine di campo, soffermandosi soprattutto sulla fase di pulizia dei dati (*data cleaning*) mentre nel terzo si presenteranno i principali risultati.

[5] Assinform, *Rapporto Assinform 2015*, Comunicato stampa, *http://www.rapportoassinform.it/Rapporto-Assinform/Comunicati-Stampa/Mercato-Digitale-Italiano-15-Nei-Primi-Sei-Mesi-E-Previsioni-Riviste-Al-Rialzo-Per-Lintero-2015-Ma-Non-Basta-Per-La-Ripresa_1.kl.*

Il rapido sviluppo della DS negli Stati Uniti

Si può sostenere che la DS sia sostanzialmente una diretta filiazione della *Computer science* (CS): una disciplina accademica, sviluppatasi intorno agli anni Sessanta, basata sui linguaggi di programmazione, sui compilatori, sui sistemi operativi, ecc.. Nel decennio successivo, gli algoritmi si aggiunsero, quale importante aspetto della teoria, con l'obiettivo di rendere "intelligenti" i primi *computers*. Nel corso degli ultimissimi anni, invece, grazie alla crescente disponibilità di enormi basi di dati, soprattutto grazie a fenomeni quali il *web* 2.0, alla nascita dei dispositivi mobili, al *cloud computing*, nonché ai *social networks*, ecc.., pian piano iniziava a prendere forma una nuova area disciplinare. All'inizio il termine era piuttosto vago, già utilizzato da alcuni precursori, alquanto indeterminato, e veniva usato perlopiù per indicare molte cose giustapposte, spesso anche dissimili tra loro. L'argomentazione centrale di John Wilder Tukey (1915-2000) era che un'inedita area disciplinare chiamata "*Data analysis*" sarebbe presto divenuta una nuova *scienza* e non una mera branca della statistica. "For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt.... All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data" (Tukey, 1962:1).

Il baricentro semantico del termine "Data scientist" rimaneva, tuttavia, sempre fortemente centrato sul concetto di dato e su tutte le tecniche di trattamento dello stesso: la quantità di dati, disponibili per la manipolazione, costituisce il vero valore aggiunto rispetto agli anni precedenti. Non v'è dubbio alcuno, difatti, che oggi vi sia un "diluvio" di dati. Grazie ai sensori dei cellulari, ai dispositivi dell'*Internet of things* (IoT), alle "*data-driven apps*" e alle tecnologie indossabili (*wereables*) la rete è piena di dati "rilasciati" dagli utenti, in modo più o meno volontario, una sorta di "feedback loop", in cui questi ultimi contribuiscono allo sviluppo ulteriore dei prodotti utilizzati: è la nascita della DS secondo l'opinione di Mike Loukides[6]. In buona sostanza, grazie alla tecnologia "intelligente", il mondo viene sempre più "mappato", misurato, "registrato" in *bits* digitali, "immagazzinato" e sempre più una parte della vita quotidiana, di crescenti fasce di popolazione, trova una sua ragion d'essere nel mondo virtuale. E i dati disponibili sono un sottoprodotto di questa crescente, correlata, esistenza digitale. Un'enorme quantità di dati disponibili che è d'uopo chiamare, oggigiorno, *big data* (O'Reilly, 2015). In questo contesto, molte imprese hanno compiuto elevati investimenti in nuove tecnologie per immagazzinare, analizzare, costruire *reports* e vi-

[6] «The thread that ties most of these applications together is that data collected from users provides added value. (...) The users are in a feedback loop in which they contribute to the products they use. That's the beginning of data science. (...) Increased storage capacity demands increased sophistication in the analysis and use of that data. That's the foundation of data science», Mike Loukides, *An O'Reilly Radar Report. What is Data Science?*, *http://cdn.oreilly.com/radar/2010/06/What_is_Data_Science.pdf*.

sualizzare i dati ma, ancora oggi esse sono fortemente dipendenti da come tutti questi strumenti vengono utilizzati da esperti, in grado di estrarre le informazioni utili, a fini aziendali, da un'elevata congerie di dati.

Le evoluzioni più recenti si incentrano, per lo più, sul tentativo di arrivare, in tempi brevi, a soluzioni automatizzate, "scalabili", le quali possono interpretare i dati, trovare corrispondenze inedite tra i diversi fenomeni e metterli a disposizione delle strategie aziendali. L'intelligenza artificiale, la *deep learning*, il *web* semantico e intelligente (Shroff, 2013; Workman, 2016) sono tutti campi di sviluppo futuro, in questo senso, e dall'altro lato, ogni posizione organizzativa dove si spendono tempi considerevoli in compiti ripetitivi è la candidata ideale ad essere sostituita da dispositivi "intelligenti", ovvero a rischio di automatizzazione (Paliotta, 2016). Infine, i costanti progressi nel campo della generazione dei linguaggi naturali avanzati (*natural language processing*, NLP) porterà tali *smart machines* a una sorta di interazione sempre più personalizzata e quasi umana (*human-like intelligence*), di cui gli assistenti vocali attuali (Amazon Echo, Apple Siri, Google Now, Microsoft Cortana) costituiscono una mera fase embrionale: si arriverà, assai presto, a un'intersezione tra linguaggi naturali, *machine learning*, *big data* e *artificial intelligence* (AI).

In questo contesto di frontiera, si sviluppa, dunque, la figura professionale del "Data scientist", la quale negli Stati Uniti e nei paesi economicamente più sviluppati, costituisce oramai una realtà incontestabile. Secondo i dati di Google Trends la ricerca di informazioni su questo profilo professionale, a partire dal settembre 2013, per la prima volta, superava quello dello statistico[7]. Da qui i relativi fabbisogni professionali e i ricorrenti proclami di *mismatch* tra domanda ed offerta di lavoro. Tra i primi e più autorevoli, uno studio, molto citato, della McKinsey & Company Inc. prevedeva, nel 2011, che entro il 2018, gli Stati Uniti avrebbero dovuto affrontare una carenza di personale qualificato, da 140.000 fino a 190.000 unità, in possesso di "*deep analytic skills*" nonché di 1,5 milioni di "*managers*" e "analisti" con il *know-how* di base per utilizzare al meglio l'analisi derivante dai *big data*[8].

Per quanto riguarda la situazione d'oltreoceano vale qui riportare anche alcune stime di Tara Sinclair relative alle informazioni desumibili dalle inserzioni *on-line* pubblicate dal più grande sito aggregatore di posti vacanti al mondo (*www.indeed.com*), un portale che è disponibile in oltre 150 paesi, in 28 lingue, un tipico motore di ricerca verticale, con più di 180 milioni di visitatori unici al mese. Ebbene, da questi dati si evince che, di fatto, tale professione, non solo non esisteva tre anni fa ma adesso è una delle occupazioni meglio pagate negli Stati Uniti (117.000 dollari di salario medio nel

[7] <https://www.google.com/trends/explore#q=statistician%2C%20data%20scientist>.

[8] «There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions», McKinsey & Company, *Big data. The next frontier for innovation, competition, and productivity* in McKinsey & Company Insights & Publications, January 2009, *http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.*

2015)[9]. Tale cifra trova conferme anche in un altro studio dove la retribuzione media del profilo (104.000 dollari) è tra le più alte all'interno della *software industry* (King & Magoulas, 2015:6)[10].

L'indagine svolta da Indeed mostra come nell'ambito dei *software job titles,* il "Data scientist" occupa la prima posizione insieme al "Software architect", seguiti dal "Software engineer", dal "Mobile engineer" e dal "Mobile developer" (queste ultime tre con una retribuzione media di 102.000$). I primi dieci posti di questa graduatoria sono completati dallo "UI/UX Developer" (99.000$), dal "Software developer" e dal "Front-end developer" (ambedue con un salario medio di 95.000$), dal "Web developer" (87.000$) e buon ultimo, con un certo distacco, dal "Data analyst" (62.000$).

Da un'iniziale analisi di tale graduatoria, sembrerebbe esserci una certa sovrapposizione tra le figure professionali che hanno come parola chiave *software* e che vengono qui declinate come "architetti", "ingegneri" e "sviluppatori" rispettivamente, ma le disuguaglianze salariali suggeriscono che si possono fare alcune distinzioni tra di esse. Le differenze principali sono che i "Software architects" sono più spesso coinvolti in attività di responsabilità, di gestione manageriale e di *leadership* complessiva dei progetti, e ciò spiega il loro stipendio più elevato, mentre i "Software engineers" e i "Software developers" apportano, in genere, il loro contributo solo su alcune parti di tali progetti. Ed ancora, nonostante le stime retributive simili, ci sono notevoli differenze tra i "Mobile engineers" e i "Mobile developers". I primi sono più propensi a costruire la struttura di *back-end* di un'applicazione mobile, mentre i secondi sono più focalizzati sul *front-end.* I primi si distinguono anche dai "Software engineers" in quanto, questi ultimi, fanno maggiormente uso di competenze tecnico professionali specialistiche quali i linguaggi di programmazione e quelli inerenti i *databases* mentre gli altri elaborano soprattutto sistemi operativi mobili quali IOS, Android ed altri.

Se questo è il contesto d'oltreoceano, cosa si può dire di quello italiano? Innanzitutto che mancano le informazioni per cui non è possibile fare una comparazione puntuale con i dati sulle retribuzioni su riportati ma sono insufficienti anche le informazioni generali relative alle caratteristiche demografiche, alla diffusione di tale profilo nel mercato del lavoro nazionale, allo *skill-set* posseduto e così via. Trattandosi di una figura innovativa ciò è anche comprensibile per cui, qui, per inferire qualche informazione di quadro generale, atta a contestualizzare la successiva indagine di campo, si farà ricorso alla piattaforma professionale LinkedIn e alla stima dei posti vacanti reperibili *on-line* su Indeed.com. Ebbene, facendo una ricerca generica riguardo a "Data scientist", sul sito statunitense di Indeed, al 18 febbraio 2016, si ottengono come risultato 22.178 posti vacanti mentre quelli ricavabili sul sito italiano sono 46 offerte di lavoro comprensive anche di *stages* e tirocini. Già solo questi semplici numeri dicono tutta la distanza tra i due mercati occupazionali, almeno per quanto riguarda tale profilo. Ma si può ve-

[9] Tara Sinclair, *Beyond the Talent Shortage. How Tech Candidates Search For Jobs,* Indeed Report, *http://blog.indeed.com/hiring-lab/beyond-the-global-talent-shortage/.*

[10] «The median annual base salary of the survey sample is $91,000, and among US respondents is $104,000. These figures show no significant change from last year. The middle 50% of US respondents earn between $77,000 and $135,000».

rificare se la stessa situazione si registra andando a visionare i profili degli addetti ai lavori pubblicati su LinkedIn. Sempre al 18 febbraio 2016, sul *network* professionale, erano presenti 30.235 "Data scientist", ovvero il numero degli *accounts* che contengono, nella posizione lavorativa attuale, questo *job title* e non nel riepilogo (in questo caso i risultati sarebbero stati più numerosi). Riguardo ai vari paesi, gli Stati Uniti, come si può facilmente immaginare, vi giocano la parte del leone con 15.075 profili, di cui 4.064 a San Francisco Bay Area e 1.936 a New York City, seguiti dal Regno Unito (2.102) e dall'India (2.009). A livello continentale, la Francia è presente con 1.512 *accounts* seguita dalla Germania (913), dalla Spagna (638) e dall'Italia, con 362 profili.

Come si vede da questi scarni indizi la situazione italiana è connotata da una sostanziale carenza sia di profili, sia di informazioni conoscitive sul fenomeno oggetto di studio. Per tale ragione si è deciso di rilevare direttamente tali dati mediante uno studio di caso sperimentale così come verrà illustrato qui di seguito.

Metodologia e trattamento informatico del dato rilevato

La ricerca si è svolta mediante quattro fasi principali. La prima è stata quella della selezione dei siti di ricerca di personale e pur potendo fare diverse distinzioni tra gli stessi quali siti aziendali, siti istituzionali, *job boards*, ecc.. (Paliotta, 2015) ci si è avvalsi sostanzialmente di una sola tipologia, i motori di ricerca specialistici o verticali (*vertical job search engine*), stante il loro carattere onnicomprensivo. Questi siti raccolgono, difatti, annunci di lavoro da migliaia di siti, incluse le bacheche di lavoro, i quotidiani in rete, le società di intermediazione e le pagine lavoro delle imprese private. Tra tutti questi *job websites* si è scelto il più importante a livello globale, ovvero "Indeed.com", analizzato nella sua versione italiana (*www.it.indeed.com*)[11].

La seconda fase è stata quella della raccolta delle informazioni relative alle inserzioni che contenessero la figura professionale del "Data scientist". In generale, l'attività di *web data mining* si deve confrontare con diversi formati tecnologici (XML, JSON, HTML5, AJAX): nonostante gli ultimi formati del *web*, quali JSON e XML, permettano di estrarre dei dati "ben educati" vi sono ancora moltissimi siti scritti in linguaggi quali HTML e AJAX i quali necessitano di diverse operazioni con cui estrarli e predisporli in un formato adatto all'analisi. A questo scopo possono servire "librerie" come quelle di Beautiful Soup, scritta in Python, o altri "pacchetti", scritti in R, per risolvere i problemi pratici dovuti alla strutturazione delle pagine *web*. Si è deciso, inoltre, di svolgere una ricerca "ristretta" di tale figura, ovvero tra apici "" in modo da restringere il campo a quelle

[11] La nota metodologica dell'indagine Indeed, citata in precedenza, è utile per circoscrivere tale fonte di dati, anche nel caso di studio qui indagato. «The job posting data on Indeed includes millions of jobs from thousands of sources. It is important to note that Indeed job postings do not reflect the precise number of jobs available in the labor market, as an opening may be listed on more than one website and could remain online for a period of time after it has been filled. Moreover, employers sometimes use a single job posting for multiple job openings. However, the data do represent a broad measure of each job title's share of job openings in the labor market», *http://blog.indeed.com/hiring-lab/beyond-the-global-talent-shortage/*.

inserzioni che contenessero solo tale profilo. Dal sito italiano di Indeed.com si sono estratti, pertanto, tutti gli annunci pubblicati, in un determinato periodo (circa un mese), consultati il 12 febbraio 2016, ovvero 31 inserzioni, di cui 13 scritte in italiano e 18 redatte in inglese.

La terza fase è stata quella della pulizia del dato (*data cleaning*). Solo dopo aver effettuato delle tecniche di *scraping* si è potuto iniziare ad affrontare il nodo della qualità dei dati. Gli annunci di lavoro, pubblicati in rete, costituiscono, difatti, un *corpus* molto vasto, solo in apparenza facilmente collazionabile ma non immediatamente fruibile per l'analisi, tanto che tale fase ha richiesto circa il 70% di tutto il lavoro complessivo. I dati provenienti dalla rete sono, difatti, caratterizzati da un'alta eterogeneità di contenuti e da un basso livello di strutturazione. Come visto, non vi sono solo problemi legati ai diversi formati tecnologici ma soprattutto dall'essere gli annunci espressi in linguaggio naturale e, dunque, di per sé poco formalizzati. I testi non strutturati necessitano, difatti, di essere pre-processati prima di essere trattati. In questa fase, si sono applicate tecniche di *text mining* quali la *tokenization*, l'eliminazione delle *stop-words* e la riscrittura degli errori di battitura (*spell-errors*). Gli spazi bianchi in eccesso, i caratteri speciali e i segni di punteggiatura sono stati eliminati durante la fase di tokenizzazione. I *tokens* ottenuti sono stati, poi, controllati con una lista di *stop-words* e, queste ultime, successivamente eliminate perché sono parti del discorso presenti frequentemente nei testi (articoli, pronomi, ecc.) ma non utili nell'analisi testuale e negli algoritmi di classificazione.

La quarta, ed ultima fase, è stata quella del trattamento del dato mediante vari *softwares* di analisi, tutti caratterizzati dall'essere *open source*. Oltre R e Python un'ulteriore fase dell'analisi è stata svolta mediante l'ausilio del pacchetto *software* IRaMuTeQ (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires)[12] il quale permette di indagare i "mondi lessicali stabilizzati" delle inserzioni mediante un metodo statistico di analisi dei dati. Da un *corpus* testuale, il *software* permette di effettuare una prima analisi dettagliata del suo vocabolario e di creare un dizionario delle parole con le loro radici e frequenze. In seguito, per frazionamenti successivi divide il testo in segmenti omogenei contenenti un numero sufficiente di parole e, quindi, procede ad una classificazione dei segmenti identificando le opposizioni più forti. Questo metodo permette di estrarre delle classi di senso, costituite dalle parole e dalle frasi più significative, le quali rappresentano le idee e i temi dominanti del corpo testuale, il quale è qui costituito dall'insieme delle inserzioni estratte dal *web*.

[12] IRaMuTeQ, creato da Pierre Ratinaud, rappresenta, di fatto, la versione libera, *open source*, di ALCESTE (Analyse des Lexèmes Co-occurrents dans les Enoncés Simples d'un Texte) sviluppato da Max Reinert.

In questo *software*, l'ultimo disponibile (version 0.7 alpha2), la lemmatizzazione avviene a partire dal proprio dizionario, senza disambiguazione e consiste nel rapportare i verbi all'infinito, i nomi al singolare e gli aggettivi al maschile singolare. La maggior parte delle analisi effettua una distinzione tra forme "piene" (o forme attive, quali verbi, nomi, aggettivi, avverbi) e parole "outils" (o "forme supplementari", quali pronomi, congiunzioni, certi avverbi e certi verbi frequenti). Solo le forme «attive» vengono utilizzate per alcune analisi come la classificazione.

L'unità di analisi presa in esame è stato il singolo annuncio di lavoro. Le inserzioni pubblicate in rete, con le correlate competenze richieste ai candidati (*job description*), da considerarsi alla stregua di un genere di "comunicazione organizzativa" (Yates & Orlikowski, 1992:323)[13], ovvero degli "organizational artifacts" (Rafaeli & Oliver, 1998:343)[14], costituiscono un'area di ricerca sempre più significativa grazie alla possibilità di far uso di una grande mole di dati. I metodi tradizionali di raccolta, difatti, non sono più sufficienti per un'analisi esaustiva di tali dati. Per questa ragione, nelle ultime due decadi, si è sviluppato, in maniera vieppiù crescente, sia il campo di ricerca disciplinare sia le applicazioni pratiche di *data mining*. Queste ultime si sono velocemente diffuse tra gli addetti ai lavori, e nel più generale campo della *business community*, portando all'estensione di tecniche di rilevamento che oggigiorno possono riguardare, solo a titolo esemplificativo, i motori di ricerca *ad hoc* (*crawler*), la classificazione automatizzata di documenti *web*, l'analisi dei *web log*, la messa a punto di *query* intelligenti, l'implementazione di modelli predittivi, ecc.. Nel corso di questo lavoro si farà riferimento a una forma peculiare di tecnica, definibile come *web data mining*, usata per navigare attraverso pagine in rete per estrarre informazioni presenti, in forma non strutturata, negli annunci di lavoro.

Le informazioni desumibili dalle inserzioni *on-line* sono sostanzialmente le seguenti: 1) nome della posizione lavorativa; 2) definizione sintetica e riassuntiva; 3) livello gerarchico; 4) compiti principali (mansioni lavorative, in dettaglio); 5) caratteristiche richieste per svolgere i diversi compiti lavorativi, in termini di qualifiche, competenze, conoscenze, abilità, esperienze, tratti caratteriali, ecc..

Di tutte queste parti le più significative, ai fini dell'analisi, sono quelle relative alle competenze e abilità. In termini generali, la *competenza* può essere intesa come una comprovata capacità di utilizzare conoscenze, abilità e capacità personali, in situazioni di lavoro e nello sviluppo professionale e personale. La competenza può essere, poi, racchiusa all'interno del più generale costrutto di "ruolo lavorativo". Quest'ultimo si declina, difatti, in tre caratteristiche o "aspettative di ruolo" (competenze, obiettivi e compiti (*tasks*), che descrivono, con minuzia di particolari, le azioni da compiere di ogni singola attività professionale. Altre dimensioni possono essere incluse nel costrutto di ruolo lavorativo, ma tendono a risultare genericamente secondarie. Tra queste si possono includere i *pre-requisiti* del titolo di studio e degli anni di esperienza. In alcuni, rari, casi anche gli *atteggiamenti* possono risultare dimensioni richieste ad un candidato (ad esempio, la disponibilità, la pazienza, la cortesia, ecc..) nonché la motivazione. In quest'analisi, si farà uso, in una versione semplificata, del modello ISFOL delle competenze (Di Francesco, 1998), il quale distingue tra:

[13] «Embedded social process that over time produces, reproduces, and modifies particular genres of communication».

[14] «Employment ads can be seen as organizational artifacts that enact or celebrate organizational processes (…) Employment ads may therefore be a medium that connects individuals, groups, occupations, and organizations».

- *competenze di base* (conoscenze di carattere generale e capacità tecniche fondamentali per l'occupabilità e il diritto di cittadinanza che tutte le persone dovrebbero avere. Queste competenze fanno riferimento alla dimensione culturale generale di un individuo)[15];
- *competenze tecnico professionali* (sono altamente specifiche, connesse ad un contenuto lavorativo e si identificano in mestieri e ambiti disciplinari. Sono costituite dalle conoscenze e dalle tecniche operative specifiche di una certa attività che il soggetto deve presidiare per poter "agire con competenza");
- *competenze trasversali* (capacità trasversali, vale a dire non connesse a una specifica attività o posizione lavorativa e che possono essere, pertanto, applicate in più ambiti lavorativi e di vita. Queste competenze, quindi, appaiono come strategie generali, riferite all'ambiente, flessibili e modificabili)[16].

Nel corso del testo, tale *framework* è stato applicato al *corpus* delle inserzioni, come griglia di analisi, sia per rintracciare alcune caratteristiche di fondo ma soprattutto, dopo aver svolto la fase di trattamento dei dati, per dar conto dei risultati della ricerca.

La rilevazione on-line delle job vacancies: Skills e Job description

Come detto in precedenza, dal sito italiano di Indeed.com si sono estratte 31 inserzioni, di cui 13 scritte in italiano e 18 redatte in inglese. Già solo da questo dato si vede come tale figura sia strutturalmente connotata dalla sua "matrice" anglofona, accertato che gli annunci pubblicati direttamente in lingua inglese sono più numerosi di quelli scritti in italiano, pur essendo riferiti a posti di lavoro con sede in Italia. Ciò dà conto dell'internazionalizzazione spinta dei mercati del lavoro nazionali sia perché le imprese, multinazionali o meno, operano in tutti i mercati sia perché la conoscenza della lingua inglesa è oramai data per scontata in gruppi di lavoro globali. Anche il numero delle inserzioni stesse è molto ridotto rispetto a quanto si potesse preventivare all'inizio. A questo riguardo, v'è da aggiungere che sono state svolte anche *query* su altri importanti siti di ricerca *on-line* quali Monster, Adecco, InfoJobs, Randstad e i risultati sono stati anche minori (in alcuni casi non si è rilevata nessuna inserzione).

La localizzazione geografica degli annunci, distinti per regione, è la seguente: Lombardia (23); Piemonte (2); Lazio (2); Emilia Romagna (1); Veneto (1); Liguria (1); Estero (1). Ciò mostra, assai bene, come il segmento delle imprese innovative italiane, almeno per quanto riguarda questa figura, sia localizzata quasi esclusivamente a Milano e regione.

[15] Nel dettaglio, esse sono: 1) lingua inglese; 2) informatica di base; 3) organizzazione aziendale; 4) diritto del lavoro e sindacale; 5) tecniche di ricerca attiva del lavoro; 6) economia di base.

[16] Le competenze trasversali identificate da ISFOL sono: *diagnosticare* (le proprie competenze e attitudini; i problemi); *relazionarsi* (comunicare, lavorare in gruppo, negoziare); *affrontare* (potenziare l'auto apprendimento, affrontare e risolvere problemi, sviluppare soluzioni creative).

Le *competenze di base* riguardano essenzialmente la conoscenza delle lingue e qui ci si può solo riferire, almeno nel caso degli annunci redatti in italiano, alla conoscenza della lingua inglese la quale viene richiesta in 9 casi su 13. Per quel che concerne il *pre-requisito* del titolo di studio viene indicata la laurea, nelle seguenti discipline (con l'indicazione della relativa numerosità): Statistica (6); Matematica (6); Informatica (5); Ingegneria matematica (4); Ingegneria informatica (4); Fisica (4); Economia (2).

Per quel che concerne le inserzioni redatte in inglese, la lingua italiana e inglese viene considerata obbligatoria mentre per quel che riguarda il titolo di studio gli annunci fanno riferimento al Bachelor degree, al Master e al PhD nelle seguenti discipline: Statistics (5); Computer science (4); Mathematics (3); Engineering (2); Physics (2); Electrical engineering (1); Management engineering (1); Applied mathematics (1); Medicine (1); Actuaries (1); Biomedical engineering (1); Robotics (1), Organic chemistry (1); Biology (1); Food science (1). Rispetto alle inserzioni redatte in italiano qui i titoli di studio sono maggiormente dettagliati e coprono un'area disciplinare più ampia. Si può anche inferire, forse, che essi non riguardano solo le tradizionali aree della statistica, informatica e ingegneria ma altre che dicono relazione a diversi domini di conoscenza (*domain expertise*) a indicare un'estensione delle tecniche di DS ad altre e più numerose attività lavorative.

Le *competenze tecnico professionali* sono state articolate nelle seguenti aree specifiche: 1) *big data* e *cloud computing*; 2) *Databases*; 3) Pacchetti di analisi statistica; 4) OS e linguaggi di programmazione; 5) Altre (*machine learning*, *data mining*, pacchetti di Microsoft Office). A causa della loro rilevanza, nella vita professionale quotidiana del "Data scientist", ad esse è stato accordato uno spazio di approfondimento maggiore, rispetto alle altre due tipologie di competenze (di base e trasversali).

Come messo in luce dai risultati di quest'analisi, presentati in tabella 1, in generale, per quanto riguarda tutti gli annunci rilevati, il "Data scientist" deve necessariamente saper manipolare i dati. Oltre a saper lavorare con un pacchetto di analisi statistica (R (perlopiù nelle inserzioni in italiano), SAS, SPSS o Matlab), deve saper programmare in diversi linguaggi (Python (soprattutto negli annunci in inglese), Java (Javascript), C++, Scala, Perl), saper utilizzare un linguaggio di elaborazione dei dati connesso a un *database* (Structured query language (SQL) e saper utilizzare alcuni programmi quali MONGODB, Oracle, POSTGRESQL), deve saper far uso di tecniche di *machine learning*, *data mining* e NLP oltre ai tradizionali pacchetti Microsoft (Excel, Access, Office) e, infine, deve spesso confrontarsi con le problematiche connesse ai *big data* e al *cloud computing* (Hadoop, Spark, Hive, Cassandra, Impala, Pig, MapReduce, Cloudera, HBase).

È chiaro che un professionista con tutte queste competenze, possedute al massimo grado, può essere considerato una sorta di "unicorno" professionale, assai difficile da rintracciare e ancor meno da assumere. Ciò potrebbe ingenerare, poi, dall'altro lato, in chi non fosse in possesso di tutte queste *skills*, la percezione di non poter intraprendere una carriera in tale ambito lavorativo. Quello che si può dire, al riguardo, è che non bisogna essere altamente specializzati in tutte queste competenze tecnico professionali quanto piuttosto avere la flessibilità e la capacità di muoversi, in maniera "agile", tra di esse. Ad esempio, non si tratta di essere un "Database administrator" altamente qualificato quanto, piuttosto, di conoscere le funzioni basilari del linguaggio SQL; non

di essere un provetto programmatore quanto di essere bravo a scrivere, o almeno a recuperare dalla rete, uno *script* adeguato alla bisogna; non di padroneggiare tutte le tecniche statistiche quanto, piuttosto, di essere bravo ad applicare quelle relative al fenomeno oggetto di analisi.

Dalla tabella 1 si vede, molto bene, come vi sia una marcata differenze tra gli annunci redatti in lingua italiana e quelli in lingua inglese. Innanzitutto, in quelli in lingua inglese vi è un maggior dettaglio delle competenze richieste ai potenziali candidati mentre in quelli in italiano il ventaglio di tali *skills* si riduce all'essenziale. Potrebbe essere solo un caso, legato a queste inserzioni, le quali si sono rilevate in un determinato periodo, oppure potrebbe confermare quanto detto in precedenza, in riferimento ai titoli di studio, vale a dire di una maggiore ricchezza descrittiva ed espositiva degli annunci redatti in inglese.

Al momento di tirare le somme, di quanto sin qui sostenuto, in termini di individuazione delle *skills* tecnico professionali, si può dire che i "Data scientist" devono essere in grado, grazie al loro *set* di competenze/abilità di riuscire a concatenare insieme gli *scripts* di linguaggi diversi, di comprendere e maneggiare con destrezza i loro carichi di lavoro flessibili completando i diversi e complessi flussi di lavoro richiesti nelle fasi operative di un progetto. Devono essere in grado di costruire, manipolare ed interfacciarsi con piattaforme informative innovative quali le "information platforms or dataspaces" secondo la definizione di Jeff Hammerbacher le quali sono simili ai tradizionali *data warehouses* seppur strutturalmente differenti[17]. Devono riuscire a districarsi tra diversi *databases*, rendere le fasi lavorative "scalabili" con l'obiettivo di non rallentarle troppo, saper immagazzinare quantità enormi di dati e, dunque, conoscere i *databases* non relazionali (NOSQL) i quali sono ben rappresentati da Cassandra e HBase. Questo grande ammontare di dati presuppone, poi, altre tecnologie di *big data* quali Hadoop e Spark. All'inizio, lo stesso Hadoop è stato fondamentale nell'abilitare pratiche di lavoro "agili"[18], termine che nello sviluppo del *software* viene utilizzato per indicare tempi di ciclo di produzione molto veloci e una più stretta interazione tra sviluppatore ed utente finale. La creazione di modalità *cloud* quali Amazon Elastic MapReduce ha permesso, infine, di poter far testare, in un tempo computazionale relativamente breve, e a costi sostenibili, molteplici *datasets* e diversi algoritmi favorendo così l'esplosione, nonché la susseguente volgarizzazione, di tecniche di *machine learning* oltre il ristretto confine degli specialisti.

[17] «They expose rich APIs, and are designed for exploring and understanding the data rather than for traditional analysis and reporting. They accept all data formats, including the most messy, and their schemas evolve as the understanding of the data changes», citato in Mike Loukides, *What is data science? The future belongs to the companies and people that turn data into products*, 2 June 2010, *https://www.oreilly.com/ideas/what-is-data-science?log-in*.

[18] «Hadoop goes far beyond a simple MapReduce implementation (of which there are several); it's the key component of a data platform. It incorporates HDFS, a distributed filesystem designed for the performance and reliability requirements of huge datasets; the HBase database; Hive, which lets developers explore Hadoop datasets using SQL-like queries; a high-level dataflow language called Pig; and other components» (*ivi*).

Le competenze *trasversali* sono, invece, quelle che meno distinguono i due *datasets*, quello degli annunci in italiano e quello in inglese, essendo equamente distribuite anche per quanto riguarda la numerosità delle stesse, a dire il vero con numeri molto bassi. Si sono rintracciate, sostanzialmente, le competenze preventivabili in fase di impostazione della ricerca, ovvero le seguenti: capacità di analisi (analitiche); risoluzione dei problemi (*problem solving*); lavoro di squadra (*team*); lavoro per progetto; doti di *leadership*; orientamento al risultato; condivisione degli obiettivi; capacità comunicative, doti relazionali; gestione di progetti complessi, ecc..

Il *software* IRaMuTeQ è stato utilizzato al fine di avere una verifica, di carattere statistico-testuale, di quanto sin qui sostenuto. Dopo avere "importato" il *corpus* da analizzare, ovvero l'insieme delle inserzioni rilevate dalla rete, IRaMuTeQ ha generato le seguenti informazioni, definite statistiche generali: numero di testi: 31; occorrenze: 10.082; numero di forme: 2.398; *hapax*: 1.176 (11,66% di occorrenze; 49,04% di forme;); media di occorrenza per testo: 325,23%[19].

Dopo le statistiche generali[20] sono state prodotte le "forme attive", ovvero quell'insieme di parole atte a rappresentare un "monde lexicaux stabilisés" (Reinert, 2008:983)[21]. Le "tracce del mondo lessicale" di questo *corpus*, inerente le inserzioni, basato sulle forme attive, vengono presentate nella tabella 2 così come la *word cloud* (Fig. 1). Senza poter approfondire tali aspetti, per ragioni di spazio, vale qui dire che tutte queste analisi confermano il quadro generale sin qui delineato, anzi hanno il pregio di rendere di più immediata comprensione tale lettura, grazie alla resa grafica delle stesse.

In ultimo, si vuole qui rispondere, almeno in prima battuta, alla domanda sulla reale innovatività e sulle differenze/similitudini del profilo del "Data scientist" con quelli, affini, del "Business analyst" e del "Data analyst". Dopo aver svolto quest'indagine, di carattere prettamente esplorativo, pare verosimile affermare, nel caso della professione oggetto di studio, che si tratta di una nuova, ed emergente, professione di cui, ad esempio, non v'è ancora traccia nel pur completo, e assai dettagliato, *Occupational Outlook Handbook, 2014-15 Edition* (O*Net) del Bureau of Labor Statistics statunitense e, ancor meno, nell'italiana e più datata, Classificazione delle professioni CP2011

[19] Le statistiche generali fornite dal *software* sono le seguenti: 1) il numero di testi (oppure dei segmenti di testo); 2) le occorrenze (il numero totale delle parole); 3) il numero delle forme presenti; 4) il numero di *hapax* (parole singole che sono presenti nel corpo testuale); 5) la media di occorrenza per testo: numero occorrenze/numero testi.

[20] Dal corpo testuale generale sono stati estratti 2 *sub corpus* relativi agli annunci in italiano e in inglese. Le statistiche delle inserzioni scritte in italiano sono le seguenti: numero di testi: 13; occorrenze: 2.783; numero di forme: 867; *hapax*: 459 (16,49% di occorrenze; 52,94%di forme;); media di occorrenza per testo: 214,08%.

Si riportano anche le statistiche degli annunci in inglese: numero di testi: 18; occorrenze: 6.943; numero di forme: 1.512; *hapax*: 699 (10,07% di occorrenze; 46,23%di forme;); media di occorrenza per testo: 385,72%.

[21] Max Reinert ha così definito queste «tracce del mondo lessicale»: «Dans l'activité langagière, les mots pleins constituent, selon nos hypothèses, des traces possibles des contenus de nos activités. Ils ne sont pas les signifiants mais bien des traces possibles de ce contenu en acte» (*ivi*).

(con la correlata Nomenclatura e classificazione delle unità professionali, NUP06). La rilevanza acquisita oggigiorno, poi, da tale profilo, quale figura paradigmatica presso la pubblica opinione e gli addetti ai lavori, è indubitabile. Essa è assai maggiore, ad esempio, anche di alcuni costrutti teorici, veri e propri modelli ideal-tipici, quali il "Symbolic analyst", divulgato da Robert Bernard Reich (Reich, 1991) negli anni Novanta, oppure l'assai conosciuto, soprattutto in ambito scientifico e tra gli addetti ai lavori, "Knowledge worker" coniato da Peter Ferdinand Drucker (1909-2005) nel 1957 (Drucker, 1959, 1999:135)[22].

Anche un veloce raffronto qualitativo, in termini di competenze/abilità, con i due *job titles* affini, "Business analyst" e "Data analyst", fa emergere le caratteristiche distintive dello "scienziato dei dati". Per quanto riguarda il confronto con il "Data analyst", fino a poco tempo fa, queste due figure venivano spesso considerate un tutt'uno, nel linguaggio comune. Basti qui riportare solo questo *tweet*, del 16 novembre 2012, di Pierogi emoji (*@edmundjackson*), "What is a 'Data scientist'? An analyst who lives in California". A parte l'ironia, lo "scienziato dei dati" emerge, invece, come un ruolo che comporta una gestione e una direzione strategica maggiore di quella del "Data analyst" e la presenza di tante competenze tecnico professionali, nel suo *skill-set*, quali algoritmi (relativi alle tecniche di *machine learning*, forse l'ambito più innovativo del 2015), *data mining*, linguaggi di programmazione (Python e Java ma anche R), *text mining*, *search personalization*, *neural network*, AI, *recommendation engines*, NLP, sembra confermare tale ipotesi. Si può mettere in evidenza, inoltre, che gli "analisti dei dati" provengono spesso da un *background* relativo all'ambito *business* mentre, nell'altro caso, gli "scienziati" sono maggiormente associati con la gestione di insiemi complessi di informazioni (*big data*) come sin qui sostenuto.

Se si vuole chiarire, invece, cosa distingue la DS dalla *Business intelligence* (BI) si può dire che quest'ultima si basa perlopiù sull'utilizzo di un *data warehouse* e su una qualche forma di reportistica, ovvero di cruscotto (*dashboard*). In questo senso, i dati servono ad alimentare continuamente il *database* aziendale e vengono utilizzati per rispondere ad una specifica questione, di solito legata a un tipico contesto di *management* mentre il concetto di DS è molto più ampio, non legato a una serie di tecniche specializzate, seppur implica una buona conoscenza delle stesse. Inoltre, nell'ambito della BI, gli "analisti" non si aspettano di utilizzare, a livello personale, le proprie analisi e i *data products* poiché questi sono sempre legati a un contesto aziendale.

Queste differenze, qui sinteticamente discusse, abbisognano, comunque, di maggiori approfondimenti, possibili solo con ulteriori ricerche di campo *ad hoc*, da svolgersi prossimamente, facendo uso della stessa metodologia, in modo da avere dei confronti maggiormente omogenei tra tutti e tre i profili professionali.

[22] "The most valuable asset of a 21st-century institution, whether business or non-business, will be its knowledge workers and their productivity." Qui non è possibile, per ragioni di spazio, approfondire tale tematica e neppure presentare una carrellata delle figure professionali proposte quali paradigmatiche nel corso dei decenni passati.

Conclusioni

I risultati ottenuti con questa rilevazione sperimentale sono stati diversi. A livello metodologico, la tecnica di estrazione dei dati non strutturati dal *web*, qui utilizzata, ha confermato tutta la sua efficacia e può, pertanto, essere applicata a rilevazioni riguardanti una mole maggiore di dati. Altri punti a favore di una rilevazione *on-line*, relativa alle *job vacancies*, è che la rete permette di tener conto della velocità del processo evolutivo dell'universo delle professioni sia riguardo al numero complessivo delle stesse – si assiste oramai quasi quotidianamente alla comparsa di nuove occupazioni, o quanto meno a una variazione nominalistica delle stesse – sia alle loro caratteristiche essenziali, in termini di competenze/*skills*, e questa è forse una caratteristica ancor più interessante. Entrambi questi aspetti sono possibili, al giorno d'oggi, perché si rinuncia a una classificazione *ex ante* e si decide, invece, di raccogliere in rete i dati relativi alla domanda di lavoro così come essa viene espressa dal sistema produttivo. In definitiva, ciò è quanto si è fatto, in questa rilevazione sperimentale, con una professione, innovativa ed emergente, quale il "Data scientist". A questo riguardo, l'ulteriore sviluppo metodologico dovrà necessariamente riguardare quello della qualità dei dati raccolti in nome di una migliore capacità di estrarre delle informazioni di valore dalla rete. Del resto, con le tecniche di *machine learning* attuali, sempre più potenti e sofisticate, ci sarà ancora più spazio per tali strategie di classificazione automatizzata di testi estratti dalla rete. Altri sviluppi sono quelli della creazione, sempre maggiore, di *application program interface* (API) e dello sviluppo del *web*, semantico ed "intelligente", sempre più funzionale all'estrazione dei dati in maniera appropriata.

Per quel che concerne la ricostruzione del profilo del "Data scientist" è da evidenziare un profondo iato tra gli annunci scritti in italiano e quelli in inglese. A livello di individuazione delle competenze di tale profilo, si è visto come il "*Data scientist*" sia connotato da uno *skill-set* inedito e innovativo. Si sono, poi, appena iniziate a delineare le principali differenze tra questi e le figure affini del "Business analyst" e del "Data analyst", queste ultime da fare oggetto di prossime rilevazioni di campo.

A livello generale di diffusione e di rilevanza della DS appare chiaro come, nel contesto nazionale, la diffusione di tale tematica sia ancora in una fase embrionale, a differenza degli Stati Uniti. Nel contesto d'oltreoceano, difatti, il valore di mercato dello "scienziato dei dati", dopo essere cresciuto tanto, anche a livello retributivo, sembra far registrare ulteriori segnali di crescita. In più, si inizia a considerare la DS al livello di un approccio sistemico grazie alla crescente diffusione dei dati (democratizzazione degli stessi, massificazione, volgarizzazione) e il conseguente accesso di ogni posizione organizzativa agli stessi (*marketing*, finanza, HR, aree della produzione). Ciò comporterà la necessità di avere personale qualificato con delle *basic skills* in questo campo. Ciò potrebbe portare, oltre che al reclutamento di "Data scientist" veri e propri, anche alla ricerca di personale in possesso di una basilare *data-driven culture*. A questo personale non verrà richiesto di avere competenze specializzate quanto di avere, nondimeno, le capacità "basiche" nel maneggiare e analizzare i principali dati inerenti il proprio lavoro, con un certo livello di competenza, e di fornire le risposte adatte agli *input* aziendali. Queste figure professionali avranno bisogno, pertanto, di essere ri-qualificate per poter avere una comprensione generale dei risultati derivanti dalle analisi condotte con le tecniche della DS.

Della carenza di tali competenze tecniche si iniziano a vedere già alcuni indizi significativi, sia nel campo della ricerca di personale qualificato *on-line* che della formazione, sia quella tradizionale universitaria che quella innovativa, basata sulle piattaforme *massive open on-line courses* (MOOC's). L'aver acquisito un titolo formativo, difatti, nel più breve tempo possibile, in questo campo, potrebbe essere decisivo per un più veloce inserimento lavorativo, e ciò sta dando luogo ad una corsa all'acquisizione di una credenziale educativa, da spendere in un mercato del lavoro sempre più globale. Sono iniziati, dunque, a nascere nuovi professionisti che qualificano essi stessi come "scienziati dei dati" e di ciò se ne può vedere un piccolo campionario sul *network* professionale LinkedIn. Un ulteriore indizio è, infine, dato dalla considerazione che tale professione figura tra quelle più ricercate e pagate, almeno negli Stati Uniti, e che essa ha, in parte, soppiantato altre che avevano, sin qui, goduto di ampia popolarità quali i "Software engineers" oppure i "Data analyst". Resiste, invece, il "Business analyst" grazie alla sua maggiore declinazione operativa nell'ambito del *management* e dei processi aziendali.

Un'ultima notazione si può fare rispetto alle dimensioni dei due mercati, quello statunitense e quello nazionale i quali sembrano trovarsi agli antipodi: lo stesso andamento complessivo del mercato italiano rimane ancora fragile forse perché si è "ancora distanti dalla velocità di trasformazione digitale che occorrerebbe per recuperare il *gap* che ancora ci separa dagli altri paesi guida, e che condiziona la nostra capacità di competere e creare nuova occupazione", così scrive il presidente di Assinform, a margine della presentazione del Rapporto 2015.

In conclusione, il *gap* tecnologico ed occupazionale sembra allargarsi e non diminuire e se la situazione non dovesse cambiare, in un prossimo futuro, si tratterebbe dell'ennesima occasione persa dal Paese nel cercare di avere una forza lavoro assai qualificata, in grado di far fronte alle tematiche più innovative (Industria 4.0, IoT, *big data*, *machine learning* e DS), correndo così il rischio di perdere ulteriori possibilità di sviluppo, non solo occupazionali, in un segmento del mercato del lavoro oramai di dimensioni globali, e altamente competitivo.

Figura 1. Word cloud delle inserzioni scritte in inglese e italiano, rispettivamente

Tabella 1. *Skills* tecnico-specialistiche negli annunci italiani e inglesi relative al "Data scientist"

Cloud Softwares	Inserzioni in italiano	Inserzioni in inglese
Apache Hadoop	2	8
Apache Spark	1	8
Apache Hive	1	6
Apache Cassandra	0	5
Cloudera Impala	0	5
Apache Pig	1	4
Redis	0	4
Apache Flink	0	4
MapReduce	1	1
Cloudera	1	1
Apache HBase	0	1

Databases	Inserzioni in italiano	Inserzioni in inglese
SQL	3	13
MONGODB (NOSQL)	1	6
Oracle	0	3
POSTGRESQL	0	1

Sw Statistica	Inserzioni in italiano	Inserzioni in inglese
R	6	3
SAS	4	3
MatLab	3	2
SPSS	2	1

OS e linguaggi di programmazione	Inserzioni in italiano	Inserzioni in inglese
Python	3	10
Java (Javascript)	5	5
C C++	3	3
Linux	0	2
Unix	0	2
Scala	0	2
Perl	1	1

Fonte: ISFOL, elaborazioni dell'autore

Tabella 2. Le prime 20 forme attive nelle inserzioni scritte in italiano e in inglese, rispettivamente

data_scientist	25	experience	60
conoscenza	23	datum	55
big_data	19	work	42
matematico	16	customer	38
informatico	15	business	37
esperienza	15	skill	35
dato	13	technology	34
candidato	13	analytics	33
analisi	13	solution	32
giorno	12	team	31
buono	12	knowledge	30
sede	11	support	27
ricerca	11	develop	27
requisito	11	product	26
processo	11	model	24
ingegneria	11	analysis	23
azienda	11	process	21
statistico	10	project	20
business	10	data_scientist	20
ambito	10	scientific	19

Riferimenti bibliografici

Bughin Jacques, *Big Data, Big Bang?*, "Journal of Big Data", vol. 3, art. 2, pp. 14, 2016.

Davenport Thomas H., Jeanne G. Harris, Robert Morison, *Analytics at Work. Smarter Decisions, Better Results*, Harvard Business Review Press, Boston (MA), 2016.

Davenport Thomas H., D.J. Patil, *Data Scientist. The Sexiest Job of the 21st Century*, "Harvard Business Review", October, pp. 70-76, 2016.

Di Francesco Gabriella (a cura di), *Unità Capitalizzabili e crediti formativi. Metodologie e strumenti di lavoro*, ISFOL, Franco Angeli, Milano, 1998.

Peter Ferdinand Drucker, *The Landmarks of Tomorrow*, Harper and Row, New York (NY), 1959.

Peter Ferdinand Drucker, *Management Challenges for the 21st Century*, Harper Collins, New York (NY), 1999.

King John, Roger Magoulas, *2015 Data Science Salary Survey. Tools, Trends, What Pays (and What Doesn't) for Data Professionals*, O'Reilly, Sebastopol (CA), 2015.

ISFOL, *Competenze trasversali e comportamento organizzativo. Le abilità di base per il lavoro che cambia,* Franco Angeli, Milano, 1994.
McAfee Andrew, Erik Brynjolfsson, *Big Data. The Management Revolution*, "Harvard Business Review", October, 2012, pp. 60-68.
Paliotta Achille Pierre, *Una serie storica lunga trent'anni. Caratteristiche distintive e peculiarità metodologiche* in Michele Cuppone, Anna Mocavini, Achille Pierre Paliotta, Giulio Rauco, *La domanda di lavoro qualificato. Le inserzioni a modulo nel 2009. Trent'anni di rilevazioni ISFOL-CSA,* ISFOL, Research Paper, Roma, 2014, pp. 8-21.
Paliotta Achille Pierre, *Where the Jobs Are. Diffusione, tipologie e caratteristiche dei job websites negli USA e in Italia*, "Osservatorio ISFOL", V (2015), n. 4, pp. 133-153.
Paliotta Achille Pierre, *La "difesa" del lavoro dalle macchine intelligenti*, in "Il Sussidiario.net", quotidiano *on-line,* 12 febbraio 2016.
Provost Foster, Tom Fawcett, *Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking*, O'Reilly, Sebastopol (CA), 2013.
O'Reilly, *Big Data Now. 2014 Edition,* O'Reilly, Sebastopol (CA), 2015.
Rafaeli Anat, Amalya L. Oliver, *Employment Ads. A Configurational Research Agenda*, "Journal of Management Inquiry", VII (1998), n. 4, pp. 342-358.
Reich Robert Bernard, *The Work of Nations. Preparing Ourselves for 21st Century Capitalism*, Vintage Press, New York (NY), 1991.
Reinert Max, *Mondes lexicaux stabilisés et analyse statistique de discours*, Paper JADT 2008. 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon (FR), March 12-14, 2008, pp. 981-993.
Shroff Gautam, *The Intelligent Web. Search, Smart Algorithms, and Big Data*, Oxford University Press, Oxford (UK), 2013.
Tukey John Wilder (1915-2000), *The Future of Data Analysis*, "The Annals of Mathematical Statistics", 1962, p. 67.
Workman Michael (ed.), *Semantic Web. Implications for Technologies and Business Practices*, Springer International Publishing, Cham (CH), 2016.
Yates JoAnne, Wanda J. Orlikowski, *Genres of Organizational Communication. A Structurational Approach to Studying Communication and Media*, "The Academy of Management Review", v. XVII, n. 2, April, 1992, pp. 299-326.

Sitografia: Nel corso del testo sono riportati molti *links* a pagine *web* i quali non vengono qui riportati per non appesantire troppo la bibliografia: tutti questi, quando non espressamente indicato, sono stati visitati tra fine febbraio e inizio aprile 2016.

Per citare questo articolo: Achille P. Paliotta, *Ricerca di lavoro e metodologie di web data mining. Il profilo del Data scientist nelle inserzioni on-line*, "Osservatorio Isfol", VI (2016), n. 3, pp. 131-149.