Ingegneria della conoscenza

I 'Data Scientist' una risorsa per l'Intelligence

ANTONIO TETI

Secondo un sondaggio elaborato dalla statunitense EMC Corporations[1] (EMC Data Scientist Study), nei prossimi anni la figura più ricercata dalle aziende e dalle organizzazioni governative, sarà quella del Data Scientist, uno "scienziato dei dati" capace di trasformare in "conoscenza" l'immenso universo di informazioni prodotte quotidianamente a livello mondiale. La sua valenza è già considerata nevralgica per molte organizzazioni...

Il concetto di ingegneria della conoscenza

La produzione e la gestione delle informazioni è forse l'argomento che, nei secoli, più di altri ha alimentato la discussione e lo scontro tra gli esseri umani. Umanisti, psicologi e scienziati come Shannon, Wiener, von Bertalanffy, Turino, hanno condotto, nei decenni, importanti studi e ricerche sui mezzi di comunicazione, che hanno generato importanti scoperte nei settori della logica matematica, della teoria dell'informazione, della cibernetica e della teoria dei sistemi. Con l'avvento e il successivo sviluppo degli elaboratori elettronici, l'informazione ha assunto, ancor di più, un ruolo fondamentale nella vita dell'uomo, grazie alla possibilità di elaborare, filtrare e incrociare le inarrestabili quantità di dati fagocitate dai media.

Quindi l'analisi e lo studio delle notizie, come elemento pregnante su cui basare previsioni, analisi e decisioni, diventa un'attività di primaria importanza, indipendentemente dal settore o dalla tipologia di organizzazione in cui l'individuo si trova a operare.

La scomposizione delle informazioni, l'evoluzione incessante delle potenzialità dei computer e lo sviluppo della rete Internet, hanno permesso di giungere alla definizione di un nuovo settore scientifico, in cui studiare le integrazioni delle conoscenze, l'utilizzo dei sistemi informatici per l'elaborazione e la *raffinazione delle informazioni per generare un sistema di conoscenza avanzato*, fruibile per una molteplicità di scopi e obiettivi.

Il vantaggio dell'utilizzo dei sistemi informatici risulta fondamentale per la possibilità di effettuare sofisticate elaborazioni di informazioni che, in altro modo, richiederebbero l'utilizzo di ingenti risorse umane e di complessi livelli

[1] www.emc.com/utilities/globalsiteselect.jhtml?checked=true

di specializzazione. L'articolato lavoro di "ingegnerizzazione" delle informazioni, risulta inglobato all'interno di un settore scientifico ben definito: *l'ingegneria della conoscenza*.

Coniato nel 1977 da Edward Feigenbaum[2], il termine *knowledge engineering*, in realtà, identifica quel ramo dell'intelligenza artificiale che si occupa della progettazione, realizzazione e gestione di *sistemi basati su conoscenze* (Knowledge Based System, KBS), oppure di sistemi informatici in grado di sfruttare le informazioni contenute in una base di conoscenza (Knowledge Base, KB) attraverso delle procedure automatiche di ragionamento. Nello stesso filone scientifico è possibile identificare anche i *sistemi esperti*[3] (Expert Systems).

Solitamente, nell'immaginario collettivo, con il termine "sistema basato sulle conoscenze", si identifica un sistema informatico in grado di elaborare e ottimizzare le informazioni contenute in un *database di conoscenza* (informazioni) utilizzando delle specifiche *procedure informatizzate di ragionamento logico*. Al giorno d'oggi, con lo stesso termine, possiamo identificare quelle applicazioni software realizzate non solo per automatizzare l'analisi delle informazioni, ma capaci anche di assemblarle attraverso un sistema di tipo logico, basato sulla *semantica*. In tal modo, senza l'intervento e il lavoro costante dell'uomo, è possibile costruire e conservare dei *contenitori informativi intelligenti*, fruibili per molteplici scopi.

Nel libro "*Knowledge and the Flow of Information*", pubblicato nel 1981, il filosofo Frederick Irwin Dretske, noto per i suoi contributi nello studio dell'epistemologia e filosofia della mente, asserì che "*L'informazione è un bene capace di produrre conoscenza: è il veicolo di un segnale che può far scattare un processo di apprendimento. La conoscenza è credenza prodotta dall'informazione*". Dall'affermazione di Dretske è possibile cogliere due aspetti sostanziali:

- l'informazione può produrre conoscenza;
- la conoscenza assume la connotazione della verità (credenza).

In altri termini, attraverso l'analisi delle informazioni è possibile produrre una conoscenza su di uno specifico oggetto/soggetto o evento, che può assumere, per la persona che assimila queste informazioni, un riferimento determinante per stabilire delle azioni, dei comportamenti o delle metodologie da

[2] Edward Albert Feigenbaum. Statunitense, nato nel New Jersey nel 1936, è un informatico che ha lavorato per lunghi anni nel settore dell'intelligenza artificiale. Soprannominato "father of expertsystem", realizzò una tesi di dottorato alla Carnegie Mellon University su EPAM (Elementary Perceiver and Memorizer), uno dei primi modelli di apprendimento e memorizzazione implementato come un programma per computer.

[3] Sistema esperto. È una branca dell'intelligenza artificiale e identifica un programma che tenta di riprodurre le prestazioni di una o più persone esperte in un determinato settore.

adottare per raggiungere uno specifico obiettivo. La conoscenza consente di comprendere la realtà che ci circonda ma, anche, di formulare delle previsioni sulle possibili evoluzioni del futuro; in funzione di ciò, l'individuo è messo nelle condizioni di poter operare in modo da apportare delle modificazioni sulla realtà, affinché produca effetti sugli eventi futuri. Quindi, il processo di trasformazione delle informazioni in conoscenza, rappresenta un passaggio strategicamente importantissimo per chiunque, sia per stabilire la corretta percezione della realtà che lo circonda, sia per le azioni che può esercitare per modificare, a proprio vantaggio, la realtà che lo circonda. Lo stesso Dretske identifica la conoscenza come *informazione disponibile per l'azione*. Ma solo in funzione dell'acquisizione di "conoscenze riscontrate" l'individuo può agire in maniera razionale e può, quindi, basare le sue azioni su elementi reali, verificati e ragionati.

Tuttavia, in considerazione della sterminata quantità di dati disponibili, l'individuo può procedere all'analisi intelligente delle informazioni solo se si avvale di tecnologie informatiche avanzate e della rete Internet.

È, altresì, vero che, anche se i computer possono gestire enormi quantità di informazioni, solo una modesta parte di esse può essere elaborata in maniera intelligente, dato che la maggioranza dei file contenuti nelle memorie degli elaboratori è costituita da documenti che utilizzano un linguaggio naturale, il cui significato si rivela spesso impenetrabile dalle comuni applicazioni software. Appare, quindi, evidente che i computer, pur essendo risolutivi per molteplici attività, in questo caso dimostrino tutta la loro incapacità nello sfruttamento intelligente delle informazioni in loro possesso. Pertanto, per implementare un sistema di ingegnerizzazione delle informazioni, si rende necessaria ancora la presenza umana, soprattutto per la costruzione di quel *work flow processes*, appositamente studiato per l'identificazione delle *fasi utili* fruibili per la realizzazione di un sistema di consapevolezza cognitiva.

Tipologie e fonti di conoscenza

Prima di procedere all'analisi delle peculiarità di un sistema di conoscenza, è opportuno soffermarsi sulle *metodologie di apprendimento* che concorrono alla formazione del contenitore informativo su cui si andrà ad operare.

Una prima distinzione va compiuta tra *knowing that* (sapendo che) e *knowing how* (sapere come) e, cioè, tra *ciò che si conosce* e *ciò che si sa fare*. In realtà fu il filosofo britannico Gilbert Ryle[4], a coniare questi termini per distinguere *ciò che l'individuo conosce*, in funzione dell'apprendimento culturale e sociale individuale, *rispetto alle proprie abilità operative* acquisite in contesti e condizioni diverse.

[4] G. Ryle, "The Concept of Mind", Chicago: The University of Chicago Press, 1949; traduzione italiana: Il concetto di mente, Editori Laterza, Bari, 2007, ISBN 978-88-420-7482-3.

Sulle conoscenze riconducibili all'apprendimento culturale e sociale dell'individuo, è possibile distinguere varie modalità di acquisizione:

- *Conoscenze basate su concetti*. Deriva dalla conoscenza di concetti posseduti che consentono di interpretare la realtà. Ad esempio, se pensiamo ad un'automobile, immaginiamo un veicolo che incorpora un motore o un meccanismo propulsivo in grado di imprimere una forza motrice.

- *Conoscenze basate su regole*. Si basa sulla conoscenza di norme, più o meno definite, che regolano il funzionamento dell'intero pianeta. Ad esempio, se il combustibile si esaurisce il veicolo cessa di funzionare, così come risulta ovvio che l'avvelenamento di un essere umano provoca la sua morte.

- *Conoscenze basate sulla conoscenza di fatti*. Si fonda sulla conoscenza delle notizie. Ad esempio, l'individuo conosce il percorso e gli orari dell'autobus, perché lo utilizza per recarsi al lavoro ogni mattina.

Per quanto concerne le *fonti di conoscenza*, sappiamo che le stesse possono provenire da *conoscenze dirette* (esperienze personali, informazioni acquisite direttamente dall'individuo), dal *ragionamento* (elaborazione di molteplici informazioni che sono elaborate e intersecate tra loro, elaborazione di immagini, scenari, filmati, discussioni, ecc.), dalla *comunicazione* (l'utilizzo di linguaggi, strumenti e metodologie di comunicazione per trasferire informazioni).

Le fonti servono principalmente per *raccogliere* e *memorizzare* le notizie, ma per poterle "lavorare" è indispensabile catalogarle secondo schemi e criteri che devono garantire la loro massima e immediata fruizione per trasformarle in una *base di conoscenza*.

Ma il valore della conoscenza è, soprattutto, legato alla possibilità di saper interpretare in maniera corretta la realtà, di comprendere velocemente le situazioni e gli scenari, di saper interpretare i fatti della vita quotidiana, anche quelli apparentemente irrilevanti. Da ciò deriva una capacità di prevedere le evoluzioni della realtà e dei possibili scenari che potrebbero verificarsi nei contesti di maggiore interesse. Più affidabili saranno le previsioni, maggiori saranno le capacità previsionali e le azioni preventive per gestire in maniera ottimale i possibili eventi futuri. Quindi, le azioni preventive si traducono in azioni miranti alla modificazione della realtà, per produrre degli effetti che vadano a proprio vantaggio.

Dal reperimento delle informazioni al Knowledge Engineering

Il termine *ingegneria della conoscenza* identifica, solitamente, quel settore scientifico che si dedica all'integrazione della conoscenza con i sistemi informatici avanzati, per consentire di risolvere problematiche complesse e sofisti-

cate che richiedono il supporto di personale con particolari livelli di specializzazione. Nondimeno all'interno di questa branca della scienza è possibile identificare molteplicità di filoni scientifici, che vanno dall'ingegneria del software all'intelligenza artificiale, dall'utilizzo intelligente delle basi di dati alla fruizione delle più avanzate tecniche di *data mining*, fino a toccare aspetti di logica matematica e di discipline che interessano le scienze cognitive.

L'ingegneria della conoscenza, sin dai primi anni '70, ha prodotto delle modificazioni in molti ambienti lavorativi e, in particolar modo, in ambito industriale. Grazie all'espansione dell'informatica, è stato possibile sviluppare delle applicazioni in grado di riprodurre il lavoro di personale specializzato in diversi settori e attività, sviluppando sistemi informatici molto performanti meglio noti con il termine di *sistemi esperti* (expert system). In ambito tecnologico, con il termine "sistema esperto" ci si riferisce ad un'applicazione software in grado di risolvere problematiche complesse che possono interessare settori diversi. Ad esempio, per l'analisi di enormi volumi di dati, per il controllo o la diagnostica di giganteschi complessi industriali o per la progettazione di impianti, è possibile ricorrere a questi sistemi, affidando ad un unico dispositivo hardware/software il controllo e la gestione di corposi e complessi processi lavorativi.

La differenza sostanziale tra i sistemi esperti e le altre applicazioni fruibili per agevolare il lavoro dell'uomo, risiede nella complessità delle tecnologie messe in campo per questi sistemi di tipo *expert*, che sono in grado di assicurare l'esibizione di tutti i passaggi logici (e automatici) che sottostanno alle decisioni dell'uomo. In sostanza, sono in grado di gestire in assoluta autonomia il lavoro dell'uomo, utilizzando le informazioni storiche accumulate per migliorare le procedure utilizzate. È proprio in questa capacità che si ravvisano elementi di *intelligenza artificiale*. Con queste premesse, è comprensibile come un sistema esperto possa essere adattato anche per attività di *data intelligence*.

Va chiarito che in un sistema esperto convivono tre elementi:

- *La base di conoscenza*. Contiene tutte le regole logiche e le procedure di cui si serve il sistema per il suo funzionamento;

- *Il motore inferenziale*. È il cuore del sistema e si occupa della gestione della base di dati e della fruizione intelligente dei dati contenuti;

- *L'interfaccia utente*. È lo strumento che permette all'utente di interagire con il sistema attraverso un'interfaccia software semplificata.

Come già evidenziato, in molti casi l'utilizzo di questi sistemi ha consentito di agevolare e rendere più semplici i processi di automazione di numerose attività complesse, tuttavia le applicazioni di inferenza non riescono a soddisfare tutte le esigenze dell'uomo, come nel caso della gestione dei fabbisogni

informativi. Ciò è dovuto ai costi e alla complessità che avvolgono la fase di progettazione e realizzazione di un sistema esperto, ma anche all'investimento in risorse umane che può rivelarsi particolarmente gravoso in termini di costi. Grazie all'avvento della rete Internet, sul finire degli anni '90, si comincia a ragionare in termini di sistemi esperti dedicati alla gestione delle informazioni. La possibilità di creare degli enormi database di dati e notizie da elaborare, per consentire di produrre informazioni complete e strutturate (conoscenza), ha indotto le comunità scientifiche internazionali a concentrare gli studi e le ricerche sullo sviluppo di sistemi informativi, basati soprattutto sul Web. Parallelamente, si assiste allo sviluppo di un nuovo filone scientifico, in cui la ricerca e l'implementazione di sistemi di *knowledge engineering*, in grado di consentire l'accentramento dei dati in Rete e l'interoperabilità dei programmi e delle tecnologie web, conduce allo sviluppo di un modello architetturale di ricerca intelligente delle informazioni: nasce il *web semantico*.

Intelligence delle informazioni in Rete: il web semantico

Le masse tendono a considerare il World Wide Web come un gigantesco contenitore di testi collegati in qualche modo tra loro. La maggiore peculiarità del Web è la sua universalità, che lo rende uno straordinario strumento di informazione ma, nel contempo, anche di straordinaria disinformazione. Basato sul funzionamento di *link ipertestuali*, il Web consente di accedere ad una miriade di dati e notizie, molte delle quali possono rivelarsi anche di scarso valore o addirittura inutili per le ricerche dei naviganti del Cyberspazio. È opportuno ricordare che le notizie immesse in Rete sono prodotte da un'utenza mondiale e, quindi, suscettibili di personalizzazioni o condizionamenti di ogni genere. Pertanto il cybernauta non può sentirsi al riparo da informazioni false, alterate o addirittura fuorvianti.

Per l'accesso alle informazioni nel Cyberspazio, utilizziamo i motori di ricerca, straordinari strumenti di accesso alla conoscenza che hanno rivoluzionato il concetto stesso di *accesso alle informazioni*. Tuttavia, l'irrefrenabile impulso dell'individuo di fagocitare dati e notizie, può essere pilotato da esigenze momentanee e può anche cambiare nel tempo, oppure può non essere preciso nelle ricerche e, di conseguenza, la navigazione cognitiva può rivelarsi fallace e completamente inutile. I cybernauti cercano le informazioni scandagliando il Web, basando la ricerca in funzione delle loro esperienze cognitive personali, che li porta spesso a saltellare da un portale all'altro, senza una logica di ricerca ben definita. Ma sono soprattutto le capacità di rievocazioni mnemoniche su *parole* o *espressioni chiave* ad influire notevolmente sui percorsi di ricerca. Ad esempio, non è insolito per un *web navigator* l'accesso "pilotato" a portali in cui è sicuro di trovare le informazioni desiderate, oppure l'accesso a siti il cui aspetto grafico sembra offrire garanzie in termini di disponibilità di notizie vere e affidabili.

Comunque sia, gli strumenti basilari per la ricerca delle informazioni di interesse, sono sempre riconducibili allo stesso metodo: l'utilizzo di specifiche *parole* o *espressioni chiave*. Per queste motivazioni, nel 2001, fu Tim Berners-Lee[5] ad ipotizzare, attraverso l'utilizzo delle pagine ipertestuali HTML[6], la creazione di applicazioni in grado di interpretare il contenuto dei documenti ipertestuali per elaborazioni basate sull'*aspetto semantico*[7]. L'idea si rivelò vincente e fu così che nacque il *semantic web*, un nuovo modo di ricercare le informazioni sul Web, basato non solo sulla possibilità di intercettare delle particolari *keywords* inserite nei documenti, ma in grado di utilizzare applicazioni specialistiche capaci di costruire reti di relazioni e complesse connessioni documentali secondo logiche personalizzabili.

Nel web semantico agiscono *agenti intelligenti*, applicazioni in grado di comprendere il significato dei testi presenti sui siti web e di accompagnare l'utente, senza deviazioni inutili, verso l'informazione desiderata. Un agente intelligente può essere anche predisposto per sostituirsi all'utente nello svolgimento di alcune operazioni di ricerca. Un *intelligent agent* deve principalmente garantire:

- la comprensione esatta del contenuto dei documenti presenti in Rete;
- la creazione di percorsi di ricerca guidati e l'acquisizione di informazioni attinenti alle indicazioni fornite dall'utente. È essenziale la *funzione di guida* che l'agente deve esercitare per consentire al cybernauta di raggiungere *solo* le informazioni di interesse;
- la navigazione *logica* sui portali web, per evitare di accedere a siti che non contengono le informazioni richieste e per collegare in maniera strutturata i dati acquisiti durante la navigazione.

La *funzione semantica*, costruita su apposite regole, deve essenzialmente basarsi sul significato delle parole in un particolare contesto, degli insiemi di parole riconducibili ad uno specifico campo di interesse e delle frasi e dei testi che trattano gli stessi argomenti. Pertanto l'applicazione semantica deve mettere in relazione le espressioni linguistiche con quello che il contenuto di tali espressioni "vuole significare". Per questo motivo il processo più importante è rappresentato da quello dell'interpretazione corretta del significato delle parole all'interno dei documenti.

[5] Timothy John Berners-Lee. È il famoso informatico britannico inventore, con Robert Cailliau, del World Wide Web.

[6] HTML (Hyper Text Markup Language). È il linguaggio usato per i documenti ipertestuali disponibili nel Web.

[7] Semantica. È quella parte della linguistica che studia il significato delle parole, degli insiemi delle parole, delle frasi e dei documenti. Essa ha delle strette attinenze con altre discipline scientifiche, come la semiologia, la logica, la psicologia, le comunicazioni, la filosofia del linguaggio.

Va rilevato che ciò che indirizza la nostra mente nella ricerca delle informazioni desiderate, è il cosiddetto *dominio di conoscenza*, cioè l'insieme di tutti i termini che il nostro sistema cerebrale collega all'informazione desiderata. I termini memorizzati nella nostra mente sono elaborati da processi cognitivi che li porta a legarli tra loro. Ad esempio, i termini "macchina", "automobile", "vettura", "veicolo" sono accomunati dalla stessa associazione di significato e lo strumento che ci consente di effettuare questi collegamenti cerebrali è la *competenza linguistica*, su cui il teorico della comunicazione Noam Chomsky ha elaborato numerosi ed approfonditi studi[8].

Per questo motivo i motori di ricerca fruibili in Rete stanno ottimizzando le loro funzioni per consentire di raffinare le ricerche degli utenti (ad esempio Google consente di visualizzare anche i link che contengono i sinonimi della parola ricercata). Sempre più spesso, però, le "parole chiave" non sono sufficienti a garantire una ricerca accurata e completa. Per questo motivo è indispensabile focalizzare le proprie ricerche sui "concetti". Il modello di ricerca basato sul concetto, si fonda sulla capacità del nostro cervello di organizzare delle ragnatele di significati di parole, che concorrono alla formazione di *mappe geografiche* in cui, ad esempio, le città sono assimilabili ai termini su cui effettuare la ricerca, e le strade corrispondono ai collegamenti con altri termini attinenti a quello iniziale. Questi collegamenti sono spiegabili con il concetto *prossimità semantica*, per il quale un insieme di parole e/o documenti, in funzione dell'utilizzo di metriche speciali, risulta simile per significato o contenuto semantico. Queste analogie tra termini, che conducono alla *conoscenza*, sono conosciute come *reti semantiche* e permettono di simulare le competenze linguistiche dell'individuo. Un sistema di ricerca basato su reti semantiche è in grado di ricercare e analizzare tutti i documenti presenti nel Web, in cui è possibile identificare delle analogie con il significato reale dei termini ricercati. Questa metodologia sfrutta il vantaggio dell'adozione di un sistema di ricerca raffinata, basata su concetti e domìni di conoscenza.

Tuttavia, anche il web semantico non è completamente esente da "difetti di funzionamento" e, in alcuni casi, può accusare qualche problema di efficienza. Facciamo un esempio: supponiamo di condurre una ricerca sul termine "posata" attribuendo alla parola il concetto di utensile da tavola (la forchetta, il coltello o il cucchiaio). Il motore di ricerca effettua un'analisi "grezza" sulla presenza di documenti che contengono questa parola. Il termine identifica anche una particolare tipologia di scrittura (scrittura italiana posata), il participio passato femminile di posare, l'aggettivo femminile di posato, un cognome di una persona, una marca di abbigliamento e il nome di un ristorante.

A questo punto dovremmo aggiungere altri termini che ci possano consentire di raffinare la ricerca in atto. Se, poi, aggiungessimo altre parole, l'elenco dei links proposti dal motore di ricerca aumenterebbe in maniera esponen-

[8] John R. Searle, "Chomsky's Revolution in Linguistics", The New York Review of Books, June 29, 1972.

ziale e saremmo costretti a modificare l'applicazione web-semantica per raffinare ulteriormente le metodologie di ricerca.

Per questo motivo risulta indispensabile la realizzazione di applicazioni software che siano in grado di analizzare la Rete, grazie all'utilizzo di *reti semantiche avanzate*, che contengono database lessicali che operano su base concettuale. Questi sistemi sono in grado di mettere in relazione centinaia di migliaia di termini, anche di lingue diverse, per comprendere in maniera univoca i significati dei documenti esaminati, agendo quindi come riduttori delle ambiguità e delle anomalie concettuali dei testi esaminati.

Questi sistemi si basano sull'implementazione di un *modello ontologico*, cioè uno schema concettuale complesso, capace di comprendere i criteri scelti dall'individuo per la pertinenza dei documenti ricercati dal sistema. Ovviamente un sistema basato sul *modello ontologico*[9], per la sua realizzazione e il successivo utilizzo, necessita di professionalità specifiche, con competenze in molteplici settori e preferibilmente in possesso di un corposo bagaglio di esperienze. L'individuo torna ad assumere un ruolo determinate, in funzione dell'alta professionalità posseduta. La figura professionale è quella dello specialista delle informazioni, il solo in grado di comprendere il complesso universo delle metodologie di ricerca intelligente delle informazioni, ed il solo in grado di gestire adeguatamente le applicazioni informatiche capaci di generare conoscenza.

Una figura indispensabile: il Data Scientist

Come abbiamo compreso, la valorizzazione delle informazioni, quale base per la costruzione di conoscenza, è una *mission* basilare per qualsiasi organizzazione o azienda impegnata quotidianamente a misurarsi in un mondo in continua evoluzione. I dati costituiscono il fulcro cognitivo su cui si basa il potere decisionale dell'individuo, la corretta interpretazione delle stesse è essenziale. Secondo uno studio condotto nel 2011, realizzato da IDC[10] e commissionato da EMC[11], nel giro di pochi anni saranno creati, a livello planetario, circa 1,8 zettabyte di dati (uno zettabyte equivale a 1000 miliardi di gigabyte): un vero e proprio "universo di informazioni". Questo sovraccarico informativo

[9] Ontologia è una delle branche fondamentali della filosofia e concentra la sua attenzione sullo studio dell'essere umano, della realtà in cui vive, delle categorie fondamentali dell'essere e delle sue relazioni. Nel settore delle tecnologie informatiche, un'ontologia è una rappresentazione formale, condivisa ed esplicita della concettualizzazione di un dominio di interesse. In altri termini, serve per descrivere il modo in cui diversi schemi vengono combinati in una struttura dati contenente tutte le entità rilevanti e le loro relazioni in un dominio.

[10] IDC. Azienda multinazionale specializzata in studi e analisi di mercato nel settore IT e delle telecomunicazioni (www.idc.com).

[11] EMC. EMC Corporation è un'azienda statunitense che sviluppa, fornisce e supporta infrastrutture per l'Information and Communication Technology (www.emc.com).

imporrà, entro il 2020, a tutte le aziende a livello mondiale, l'ampliamento delle rispettive dotazioni di computer utilizzati per la memorizzazione dei dati, per un valore complessivo di circa dieci volte quello attuale.

La notizia ha creato non poco scompiglio nelle aziende e organizzazioni di tutto il mondo, che hanno ammesso la loro totale impreparazione nell'affrontare questa sfida. I timori dei responsabili IT sono stati evidenziati anche da un'indagine condotta da Gartner[12] che ha sottolineato come la crescita continua delle informazioni rappresenti la problematica più avvertita dai responsabili delle infrastrutture IT. Circa il 47% degli intervistati la reputa tra i primi tre elementi di criticità. Va sottolineato che tra le preoccupazioni che serpeggiano tra i responsabili IT[13] di tutto il pianeta, c'è anche quella riconducibile ai crescenti costi energetici che dovranno sostenere le imprese per soddisfare la "fame" di energia di cui necessitano i moderni Data Center. Ma ciò che preoccupa maggiormente i CIO (Chief Information Officer) è proprio l'inarrestabile crescita dei dati, che renderà ancora più difficoltosa la "gestione intelligente" delle informazioni. In funzione di ciò le aziende saranno costrette ad arruolare i *data scientist*, figure strategiche che avranno il compito di trasformare questo *mare magnum* di informazioni in un erogatore di conoscenza. Ma quali sono le peculiarità e le responsabilità che dovrebbero caratterizzare una figura come questa? In che modo può gestire, in maniera ottimale, enormi ed eterogenei database di dati?

Tra le diverse specialità, alcune sono da considerarsi essenziali per questa singolare figura, in particolare la capacità di:

- individuare gli algoritmi migliori per le operazioni di data mining;
- individuare i criteri di analisi di maggiore rilevanza;
- sviluppare nuove metodologie di gestione e ottimizzazione dei dati (data conditioning);
- gestire, estrapolare, presentare e distribuire i dati e di trasformarli in conoscenza;
- identificare nuove tipologie di database analitici in funzione del tipo di data mining utilizzato;
- identificare strumenti di analisi di tipo "high-end", che sono più predittivi e fruibili dalle organizzazioni (ad esempio, per la prevenzione delle frodi o per effettuare previsioni sull'andamento dei mercati e della concorrenza);
- individuare le problematiche legali, in funzione della trattazione di dati riservati o informazioni protetti dalla privacy;
- possedere competenze statistiche, matematiche, metodologie di calcolo, calcolo delle probabilità, e digital processing.

[12] Gartner Inc. Azienda multinazionale statunitense specializzata in analisi, ricerche e eventi nel settore ICT (www.gartner.com).

[13] Information Technology.

Tuttavia, le competenze finora evidenziate non sono da considerarsi esaustive. Il data scientist, ad esempio, deve essere in grado di vagliare attentamente le informazioni che giungono da *fonti informative* diverse, prima di decidere quali possano essere giudicate "utili" per le sue ricerche. Così come dovrà essere in grado di incrociare preliminarmente i dati giunti in suo possesso (provenienti da molteplici fonti), con particolare attenzione a quelli che provengono dalla Rete, come i social network, blog, web server, o dalle registrazioni online. Deve essere in grado di gestire dati di maggiore complessità (come quelli geospaziali) e dovrà utilizzare algoritmi di ricerca più raffinati, capaci di scansionare (data mining) immensi database di terabyte di dati in tempi relativamente brevi; dovrà, anche, essere in grado di selezionare lo strumento di *business intelligence* più adeguato, per eseguire le analisi richieste dall'organizzazione per la quale lavora.

Anche se potrà sembrare apparentemente inconsueto, egli deve possedere anche una mentalità orientata alle arti e alla creatività, per far si che possa elaborare visioni sulle metodologie di gestione intelligente delle informazioni e perfino sul loro possibile utilizzo per finalità diverse da quelle originarie.

Se consideriamo come la necessità di *consulenti di social media* sia cresciuta con la nascita dell'era dei social network, non c'è da stupirsi se nel giro di pochi anni, in funzione della imminente esplosione dei dati, il data scientist assumerà il ruolo del professionista più ricercato al mondo. Thornton May, antropologo culturale e futurista, descrive questa particolare figura addirittura come "*l'eroe dei tempi futuri*".

Apparentemente le molteplici competenze del Data Scientist, potrebbero sembrare "eccessive" ma, contrariamente a quanto potrebbe sembrare, non è poi così difficile riuscire a identificare dei professionisti che ne siano in possesso. Infatti va rilevato che lo scienziato dei dati deve, soprattutto, eccellere in alcuni aspetti caratteriali, come la creatività, la curiosità e la determinazione nel saper affrontare situazioni nuove e particolarmente complesse. Egli deve sentirsi ispirato all'organizzazione e alla distribuzione di informazioni che siano utili per l'organizzazione per la quale lavora. La sua vera *mission* è trasformare dati in valore. Per una maggiore comprensione del livello di intersecazione delle competenze, ricorriamo ad un classico diagramma di Eulero-Venn[14] (o semplicemente di Venn), in cui le diverse discipline (o aree di competenza) si intersecano tra loro creando il recinto ottimale in cui si sviluppa la scienza dei dati (figura 1). Le diverse competenze, se unite ad altre, possono accrescere il loro valore in termini di potenziamento delle informazioni, ma è altresì vero che possono anche produrre risultati che potrebbero essere considerati "pericolosi". In altri termini, le aree di competenza devono intersecarsi in maniera *intelligente*, in modo da produrre dei processi di elaborazione infor-

[14] Diagramma di Eulero-Venn. È un diagramma che rappresenta graficamente un insieme di elementi racchiusi all'interno di una linea chiusa e non intrecciata. Può anche essere definito come un grafico utilizzato per rappresentare un'algebra di insiemi.

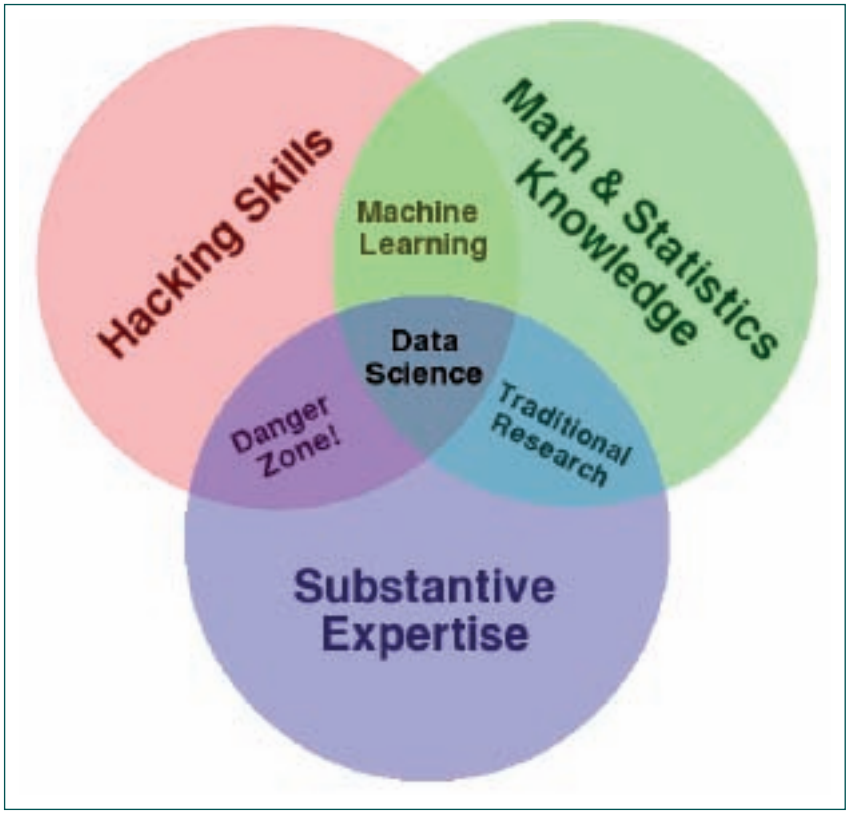

Figura 1 - *Data Science nel diagramma di Venn*
(fonte: Drew Conway, "Data Science in the U.S. Intelligence Community", Vol. 2 N. 4, IQT Quarterly Spring 2011)

mativa utile per la generazione di conoscenza. La *Danger Zone* riportata in figura 1, ipotizza una zona di rischio (pericolo), che può derivare dalla commistione tra le esperienze personali e la competenza nel settore dell'*hacking*. Ad esempio, non è consigliabile inserire un esperto di informatica, con precedenti di crimini informatici, in un processo di analisi di informazioni riservate.

Dal Data Scientist all'Intelligence Data Scientist

La caccia ai Data Scientist è iniziata già da alcuni mesi. Aziende multinazionali, come ad esempio l'American Express, da giugno 2012, stanno concentrando la loro attenzione sulla creazione di team di esperti nella gestione delle informazioni[15]. Per la nota azienda statunitense che opera nel settore finanzia-

[15] http://blogs.cio.com/business-intelligence/17184/desperately-seeking-data-scientists

rio, il reclutamento di questi professionisti sembra addirittura "fondamentale per il futuro" dell'azienda. I vertici aziendali giudicano il lavoro di questi professionisti come parte centrale della "trasformazione digitale" che investirà l'organizzazione nel prossimo futuro. Indicativo è anche il compenso annuo stabilito per questa figura: 160.000 dollari USA.

Anche in Russia l'attenzione sul problema della gestione del "big data" non ha tardato a manifestarsi. In questo caso l'annuncio è di febbraio 2012[16] e riguarda l'apertura di un nuovo centro di ricerca in cui si svilupperanno tecnologie di analisi di dati con particolare riferimento per alcuni settori, come la bioinformatica e l'efficienza energetica. L'accordo è stato raggiunto tra EMC[17] e Skolkovo Foundation[18] e il centro sarà ospitato nell'InnovationHub della fondazione moscovita. Il centro svilupperà metodologie di analisi di informazioni anche per altri settori, come la medicina molecolare, la farmaceutica, la biomedicina, le biotecnologie industriali, la geopolitica, ecc..

Della necessità di introdurre esperti in gestione delle informazioni se ne sono accorte anche alcune Agenzie di Intelligence.

"*Do you have a passion for creating data-driven solutions to the world's most difficult problems? The CIA needs technically-savvy specialist to organize and interpret Big Data to inform US decision makers, drive successful operations, and shape CIA technology and resource investments*"[19]. Questo è il messaggio che campeggia su una delle pagine web dedicate al reclutamento di personale del portale della CIA (Central Intelligence Agency) in cui sono elencate tutte le diverse professionalità ricercate dall'Agenzia. Il messaggio è chiaro: anche nel settore dell'intelligence, per interpretare correttamente gli enormi database che quotidianamente accumulano dati, è indispensabile un Data Scientist.

Tra i passaggi descrittivi delle diverse peculiarità richieste, risulta singolare il seguente passaggio "*Se si dispone di esperienza in analisi dei dati, informatica, matematica, statistica, economia, ricerca operativa, calcolo delle scienze sociali, finanza quantitativa, ingegneria o di altri campi di analisi dei dati, considerate una carriera come scienziato dati alla CIA*". È, inoltre, possibile rilevare come l'Agenzia attribuisca un'attenzione particolare alle capacità intrinseche e caratteriali di questo professionista. "*Come Data Scientist alla CIA si lavorerà con hardware e software avanzati, con tecniche per sviluppare algoritmi di calcolo e con metodi statistici per trovare modelli e relazioni per grandi volumi di dati. In funzione della missione globale della CIA, l'agenzia ha accesso a particolari set di dati che possono essere analizzati in un unico ambiente di calcolo. I candidati prescelti avranno acuto intuito tecnico, creatività, iniziativa ed una mente curiosa. I Data Scientist sono tenuti a comunicare le loro conclusioni, in termini chiari, ad un pubblico eterogeneo e a diventa-*

16 www.01net.it/emc-dalla-russia-con-big-data/0,1254,1_ART_145976,00.html)

17 www.emc.com/utilities/globalsiteselect.jhtml?checked=true

18 www.sk.ru/en/Model/AboutFund.aspx

19 www.cia.gov/careers/opportunities/science-technology/data-scientist.html

re esperti attraverso la formazione continua, la partecipazione a conferenze accademiche e tecniche, e la collaborazione con la comunità dell'intelligence".

I data scientist sono chiamati anche a svolgere un ruolo di interfacciamento (a scopo formativo-informativo), con strutture tecniche e verticistiche dell'Agenzia, senza escludere i referenti di altre strutture governative. La sede di lavoro indicata è quella di Washington, DC, cioè il centro del potere del Governo. E non è certo un caso.

Nel 2011 il giornalista Kimberly Dozier, dell'Associated Press, durante una visita condotta presso l'Open Source Center della CIA[20], ha l'opportunità di verificare alcune delle attività condotte dai primi data scientist reclutati dall'Agenzia. L'attività primaria consiste nella raccolta continua di dati e informazioni prodotte nel Cyberspazio nella loro "lingua madre" (cioè in tutte le lingue parlate nel mondo), spaziando dall'arabo al cinese mandarino. In seguito, si procede con l'analisi dei dati raccolti: si indagano post sospetti, tweet di utenti arrabbiati, email contenenti frasi strane in documenti chiari, minacce e insulti nei social network, blog anomali o superficiali gestiti da utenti improbabili. Poi, si incrociano le informazioni "attenzionate", con giornali locali, conversazioni telefoniche, foto, immagini geostazionarie, posta elettronica, sms, mms, ecc.. Successivamente, si provvede alla costruzione di "scenari ricercati", studi e rapporti utili per i più alti livelli della Casa Bianca, come nel caso del documento in cui si illustrava lo stato d'animo del popolo della regione pakistana in cui è avvenuto il blitz, all'indomani della cattura e dell'uccisione di Osama Bin Laden. Nell'articolo di Dozier è significativa l'affermazione di un funzionario della CIA, Doug Naquin, che dichiara come la sua squadra di analisti avesse previsto la rivolta verificatasi lo scorso anno in Egitto senza, tuttavia, riuscire a prevedere il periodo esatto in cui si sarebbe sviluppata. Il funzionario rassicura anche che i data scientist stanno lavorando proprio in questa direzione, cioè poter ottenere dei modelli previsionali in grado di stabilire perfino il momento in cui si verificherà un determinato evento.

A gennaio 2012 anche il Federal Bureau of Investigation (FBI) annuncia di essere alla ricerca di data scientist[21]. La richiesta del Bureau di queste figure, nasce dall'esigenza di monitorizzare l'attività dei più popolari *social media* utilizzati nel mondo (con un'attenzione particolare a Facebook e Twitter) e di individuare soluzioni software in grado di migliorare le attività di analisi e di intelligence. L'obiettivo è di realizzare un sistema in grado di ricercare in Rete notizie e dati utili ad informare le autorità sulle possibili future minacce e sui nuovi rischi emergenti capaci di influire sugli scenari interni del paese. Il sistema dovrebbe consentire una "stratificazione di dati da correlare", cioè la possibilità di sommare più informazioni locali (immagini di telecamere predisposte sul territorio, mappe di percorsi stradali e autostradali, ubicazione di installazioni strategiche e a rischio, luoghi in cui si sono verificati atti terroristi-

[20] www.huffingtonpost.com/2011/11/04/cia-open-source-center_n_1075827.html

[21] www.centrifugesystems.com/news/fbi-requesting-social-media-analytics-solutions

ci, ubicazione di personaggi a rischio, ecc.) a dati di tipo predittivo, in grado di elaborare modelli previsionali. Queste tecniche di *data display* e *data integration* vengono utilizzati già da tempo dall'autorità federale quando opera su casi particolarmente difficili e complessi. Un tipico esempio è costituito dall'utilizzo di alcune applicazioni informatiche che acquisiscono ed elaborano informazioni sulle frodi condotte negli USA, al fine di individuare le tendenze caratteriali dell'individuo che commette questa tipologia di crimine.

Anche il GCHQ (Government Communications Headquarters), struttura governativa di intelligence britannica, meglio nota con il termine di British Intelligence Agency, è alla ricerca di questi scienziati[22] che vengono, però, definiti con un termine un po' glamour: Information Specialist.

Pur conservando le specificità tipiche di uno specialista dei dati, l'Information Specialist, secondo l'Agenzia britannica, deve garantire delle conoscenze nei seguenti campi:

- Open Source Research
- Information Management
- Information Governance
- Information Risk Management
- Information Legislation
- Electronic Documentation and Records Management
- Intranet Management
- Collaboration Technology
- Knowledge Management

Identificate come delle "best practices", il possesso di queste competenze, oltre che per l'Intelligence, è considerato essenziale per l'infrastruttura informatica di qualsivoglia azienda.

Se ci spostiamo verso Oriente, seguendo il percorso che ci conduce alla scoperta di quei paesi che hanno compreso la strategicità di questi new scientist, non possiamo che soffermarci in quella nazione che forse prima di tutte ha ben compreso l'importanza del loro contributo: la Cina.

Soprattutto le aziende, ma anche alcune strutture governative, sono alla ricerca di *Chief Data Scientist* e *Senior Data Scientist*[23]. È indicativo il fatto che siano alla ricerca di responsabili di gruppi di "scienziati di dati", aspetto che lascia intendere che abbiano già da tempo in funzione "team" di professionisti concentrati sulla raffinazione delle informazioni. E non c'è da stupirsi se nella Repubblica Popolare Cinese sia stato creato, sin dal 2007, il *Center for Data Science and Dataology*[24], un Centro in cui si effettuano ricerche su teorie, data mining, metodi e tecnologie per l'analisi dei dati nel Cyberspazio. Come si

[22] http://www.gchq-careers.co.uk/Jobs/Information-specialists.html

[23] http://topic.csdn.net/u/20120830/17/99a40264-9541-4961-b970-e35e0a58c37c.html

[24] http://datascience.fudan.edu.cn/s/98/t/316/main.htm

evince da un'intervista rilasciata da una docente e ricercatrice della struttura, il Centro di ricerche focalizza l'attenzione, soprattutto, sulle tecniche di analisi di dati multisettore, come la finanza, l'economia, le assicurazioni, la bioinformatica e la sociologia. Il Centro si compone di sette unità:

1. The Data Resource Service Office
2. Dataology and Data Science Research Lab
3. New Economy Development Strategy Research Lab
4. Bio-Medical Data Research Lab
5. Brain Informatics Research Group
6. Intelligent Transportation Data Research Group
7. Financial Data Research Group

Tutto il personale della struttura è rigidamente selezionato in funzione delle competenze e delle specificità di impiego. Periodicamente svolgono dei seminari e convegni (anche a livello internazionale) e la prima monografia realizzata dai suoi ricercatori, dal titolo *Dataology and Data Science*, risale al 2009.

Ma l'aspetto che maggiormente assume rilevanza nella struttura ubicata a Shanghai, è la varietà e la particolarità dei campi di ricerca. Ad esempio, si conducono studi sulle tecniche di data mining per il sequenziamento del gene, sull'analisi delle informazioni sui sistemi di trasporto intelligence, sulle motivazioni che conducono alla creazione di virus informatici e, perfino, sugli aspetti psicologici che influiscono sui mercati azionari. Insomma, stiamo parlando di metodologie di analisi ed elaborazione *intelligente* di tutte le informazioni disponibili.

Ma il progetto di ricerca più importante del Centro è quello dello studio della teoria fondante della scienza di dati (Theory of Data Science). Secondo il Professor Yangyong Zhu, uno dei responsabili della struttura, il termine "data science" identifica la scienza dei dati nel Cyberspazio e si compone di due *dimensioni chiave*: la prima è di fornire un metodo di investigazione che i ricercatori chiamano *Scientific Research Method with Data*, fruibile per le scienze naturali e sociali; l'altra è quella della ricerca di fenomeni e leggi di *data nature*.

Quest'ultimo termine, si riferisce al complesso dei dati disponibili nel Cyberspazio, che riflettono la natura e i comportamenti umani. In altri termini, si tratta di indentificare le informazioni giudicabili tangibili e credibili da tutte quelle che non hanno riferimenti diretti su contesti reali (meglio identificati come *dati spazzatura*). Secondo il Professor Zhu, nei secoli, si sono verificate due "*esplosioni di dati*". La prima con l'invenzione e la fabbricazione della carta, la seconda con l'invenzione del computer e del Web.

Inoltre, egli asserisce che il data mining e l'analisi dei dati, sono solo delle tecnologie relative al trattamento dei dati e che non possono essere considerate come una vera "scienza". Nel Data Science viene data una forte enfasi so-

[25] http://whatsthebigdata.com/2012/06/30/the-data-science-interview-yun-xiong-fudan-university/

prattutto alle teorie sulla gestione dei dati, più che sulle tecnologie utilizzate. Quindi la sua *mission* è quella di studiare i fenomeni e le leggi che governano la natura dei dati (*data nature*). Le ricerche cinesi hanno condotto allo studio dell'*ontologia dei dati*, coniando un nuovo termine: Dataology.

Le aree fondamentali del Dataology sono:

- la teoria fondante della scienza dei dati e la Dataology;
- i metodi di ragionamento logico sui dati e le sperimentazioni;
- le teorie e metodi di funzionamento nel Dataology (ontologia dei dati applicata al comportamento, al sistema nervoso umano, alle motivazioni per la produzione di informazioni, ecc.);
- i metodi e le tecnologie che utilizzano e sfruttano i dati come risorsa.

La crescente quantità di dati non può essere più gestita con sistemi informatici e con metodologie di analisi che risalgono a qualche decennio fa. L'immagazzinamento di informazioni in enormi database, in cui lavorano ininterrottamente applicazioni strutturate per le ricerche basate su "keywords", non ha più alcun senso e ben poca utilità reale. Occorrono nuove tecniche per trattare i dati in maniera intelligente. Alcune strutture che operano nell'Intelligence stanno già sviluppando nuove teorie e metodologie di ricerca e trattamento intelligente delle informazioni. Nel giro di pochi anni, assisteremo alla nascita di gruppi di ricerca che si occuperanno della misurazione dei dati, dell'algebra dei dati, della somiglianza delle informazioni e si svilupperanno algoritmi in grado di definire la veridicità delle informazioni.

Queste innovazioni produrranno effetti considerevoli soprattutto nel settore dell'Intelligence. In questo momento sono già diverse le applicazioni software in grado di utilizzare la *semantic intelligence* per sfruttare i vantaggi della ricerca semantica e del linguaggio naturale. Tuttavia, pur essendo utilissime per la trasformazione di informazioni in conoscenza, non soddisfano ancora appieno le esigenze derivanti dalle attività di intelligence, che notoriamente sono caratterizzate da informazioni articolate e diverse e che necessitano di analisi e di intersecazioni di particolare complessità.

Per questo motivo è essenziale procedere allo sviluppo di nuovi algoritmi in grado di ampliare i possibili significati e interpretazioni delle informazioni acquisite, con l'implementazione di nuove regole e schemi di intersecazione dei dati. A tal riguardo, è evidente che il settore dell'Intelligence che maggiormente sarebbe interessato ad implementazioni verticalizzate sulla ricerca delle informazioni, è quello dell'Open Source Intelligence (OSINT). Come abbiamo potuto comprendere, la progettazione, la realizzazione e la gestione di nuovi sistemi di elaborazione intelligente delle informazioni, non può prescindere dalla presenza di personaggi che siano in possesso di specifiche e complesse competenze. Pertanto, l'*Intelligence Data Scientist* assume un ruolo fondamentale per qualsiasi organizzazione, ma si rivela vitale per le Agenzie di Intelligence, che devono affrontare le sfide del prossimo futuro per garantire

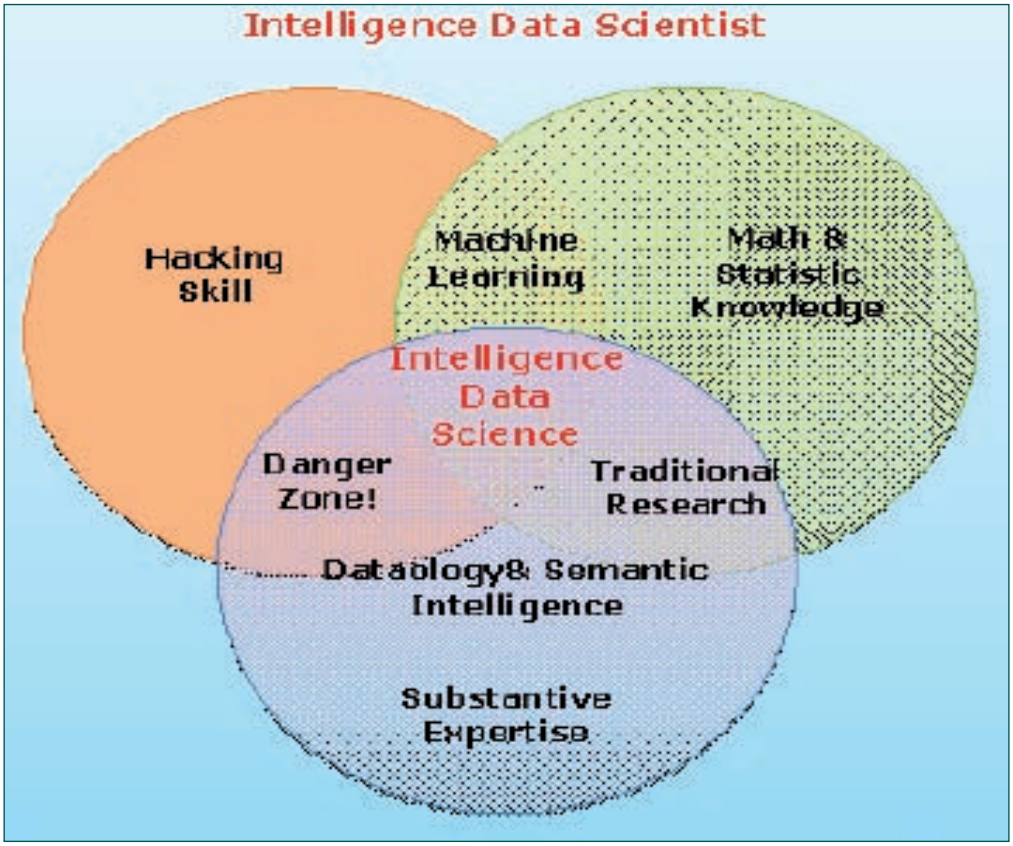

Figura 2 - *Intelligence Data Science nel diagramma di Venn*

la sicurezza interna ed esterna del Paese. La proiezione verso uno scenario lanciato verso il massiccio utilizzo di tecnologie informatiche e sull'utilizzo della Rete, crea i presupposti per la configurazione di un nuovo filone scientifico: l'*Intelligence Data Science* (figura 2).

In un futuro imminente, rappresenterà il settore scientifico in cui confluiranno tutti gli scienziati di dati specializzati in Intelligence, che avranno il compito di produrre le "conoscenze" necessarie per salvaguardare la sicurezza e gli interessi dei rispettivi paesi di appartenenza. Di conseguenza, e per non giungere impreparati alle sfide del prossimo futuro, le Agenzie di intelligence devono al più presto attivare dei processi di acquisizione e formazione miranti all'inserimento di queste nuove figure di esperti nella gestione intelligente dei dati.

George Bernard Shaw asserì che "*la scienza è sempre imperfetta. Ogni volta che risolve un problema, ne crea almeno dieci nuovi*". Forse non è del tutto falso, tuttavia, l'importante è cercare di non andarle incontro impreparati…

Bibliografia

- Lei Liu, Hui Zhang, Jianhui Li, Runqiang Wang, Luqing Yu, Jianjun Yu and Peisen Li, "*Building a Community of Data Scientist: an explorative analysis*", Data Science Journal, Volume 8, 24 October 2009
- John R. Searle, "*Chomsky's Revolution in Linguistics*", The New York Review of Books, June 29, 1972 (www.chomsky.info/onchomsky/19720629.htm)
- Drew Conway, "*Data Science in the U.S. Intelligence Community*", Vol. 2 N. 4, IQT Quarterly Spring 2011
- Mike Loukides, "*What is Data Science*", An O'Reilly Radar Report, O'Reilly Media, 2010

Per approfondimenti l'autore suggerisce…

What is Data Science?
Autore: Mike Loukides
Editore: O'Reilly Media, 2012

Chomsky's Revolution in Linguistics
Autore: John Searle
Editore: The New York Review of Books, 1972