IL RUOLO DEL BIG DATA ANALYTICS E MACHINE LEARNING NELLA SICUREZZA

Michele Colajanni

L'acquisizione di dati digitali da molteplici sorgenti, insieme ai miglioramenti delle capacità computazionali, consente di affrontare alcune sfide della sicurezza attraverso approcci basati su Big Data analytics e machine learning. Escludendo eccessive aspettative nell'automatizzazione di tutti i processi di difesa, si analizzano le principali tecnologie, le problematiche di governance e le possibili ricadute nell'ambito della prevenzione e dell'individuazione di contromisure, laddove strumenti informatici semi-autonomi possono trovare applicazione, ovvero richiedere ulteriori ricerche e verifiche sull'efficacia.

BIG DATA, MACHINE LEARNING E DEEP LEARNING

Un'informazione ricca e accurata rappresenta le fondamenta della sicurezza, e l'informatica ha sempre costituito un fattore determinante, oggi imprescindibile in tale direzione. Dalle applicazioni scientifiche e militari a quelle gestionali, i progressi dell'informatica sono trainati da un medesimo obiettivo: supportare l'uomo nell'acquisizione di dati grezzi e nella loro rapida elaborazione al fine di ottenere informazioni a valore aggiunto e, in tal modo, aumentare la conoscenza così da poter assumere le decisioni con maggiore consapevolezza.

Le applicazioni sono innumerevoli, ma il contributo dell'informatica nella riduzione dell'incertezza è tuttora rilevante. È solo dall'ultima decade dello scorso millennio che, grazie alla diffusione d'internet e all'avvento del web, l'informatica ha elaborato nuovi metodi per la diffusione dell'informazione, del commercio, delle comunicazioni e delle interazioni sociali. Tali aspetti, interessanti dal punto di vista della sicurezza e dell'intelligence, non costituiscono il focus di questo articolo, se non per aver accresciuto le opportunità di elaborare enormi quantità di dati in tempo reale.

Gran parte delle applicazioni informatiche sta diventando una combinazione di Big Data e machine learning e la sicurezza non può prescindere da tale evoluzione. Il Big Data analytics ha determinato un'innovazione dirompente rispetto alle tradizionali modalità di gestione dei dati strutturati, così come il deep learning sta modificando radicalmente le potenzialità dell'intelligenza artificiale. Questi due spartiacque vanno colti nella loro essenza sinergica prima di studiarne le potenzialità applicative nell'ambito della sicurezza.

Il termine Big Data tende a essere fuorviante in quanto è spesso confuso con la sua traduzione immediata di «grande quantità di dati», trascurando che l'aspetto più innovativo è costituito dai sistemi di gestione e di analisi di tali volumi ed eterogeneità. Le tecniche e le metodologie del Big Data analytics superano la classica modalità di gestione su *database* S*ql*, *data warehousing* e D*ata* M*ining*, per tendere verso procedure di analisi online mediante tecnologie altamente scalabili dette, per differenziarsi dal passato, NoSql[1].

Trainato dalle esigenze applicative delle grandi società Over The Top, lo spazio di competenza del Big Data analytics è identificabile in una combinazione di dimensione dei dati e dei relativi tempi per elaborarli. Come illustrato nella figura 1, se occorre analizzare molteplici Terabyte nell'ordine di pochi minuti, si è nel contesto dei Big Data. Viceversa, se trattando analoga quantità di dati è ammissibile ottenere risposte nel giro di qualche giorno, si possano applicare tecniche tradizionali di data warehouse e Data Mining che, essendo ben consolidate, preservano la loro validità.

L'intelligenza artificiale è un antico mito-obiettivo dell'informatica sin dai tempi dei padri fondatori, Turing e Von Neumann. Tuttavia, per molti decenni, l'artificiale ha prevalso sull'intelligenza in quanto l'apprendimento si è basato su materie consolidate, quali la statistica, la logica e la semantica, che gli informatici hanno saputo tradurre in algoritmi basati su costrutti e logiche noti. L'intelligenza artificiale, pur nelle molteplici applicazioni, è rimasta nell'ambito di un machine learning algoritmico con logiche deterministicamente controllabili. È solo dall'inizio di questo decennio che, complice la disponibilità di un'enorme potenza elaborativa e di grandi quantità di dati, si sta affermando un insieme di tecniche ispirate al funzionamento multilivello dei neuroni della neocorteccia, detto deep learning[2]. Queste reti emulano molteplici livelli di neuroni artificiali ed elaborazioni non lineari e sono addestrate, in modo supervisionato o meno, su vasti insiemi di dati. In tal modo, sono in grado di ottenere 'conoscenza automatica' senza che l'uomo controlli interamente il processo logico mediante cui ci si è arrivati[3].

Il software 'impara' in modo autonomo a riconoscere pattern nella rappresentazione digitale ed è da considerare un fattore dirompente nella storia dell'informatica e, quindi, dell'uomo. Le sue potenzialità applicative presentano margini ancora in via di esplorazione. Per l'analisi diagnostica medica, il deep learning non è più una frontiera della ricerca ma un business potenziale, la cui diffusione è limitata più da normative che dai casi di successo.

1. Sullivan 2015; Harrison 2015.

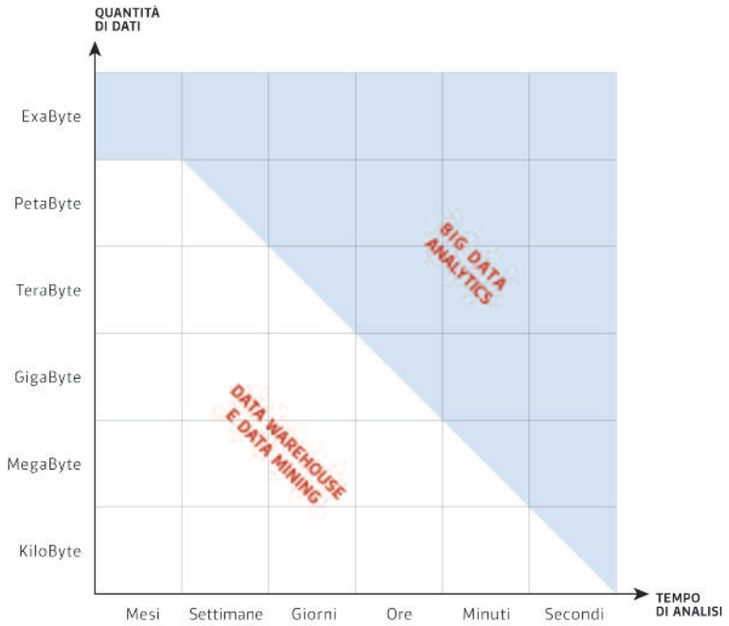

Figura 1. Ambito di competenza del Big Data analytics.

2. Goodfellow 2017.
3. Knight 2017.

Applicazioni in ambito economico e finanziario sono dietro l'angolo. I successi di Google Deep mind in *AlphaGo*, ritenuto il gioco basato sull'informazione perfetta più complesso del mondo[4], e di *Libratus* (Carnegie Mellon) nonché di *DeepStack* (Università di Alberta e di Praga) nel poker professionistico, gioco a informazione imperfetta, rappresentano pietre miliari che, insieme alla dichiarazione del capo progetto di Libratus («We didn't tell Libratus how to play poker. We gave it the rules of poker and said 'learn on your own'»), denotano le possibilità di applicare il deep learning a contesti di incertezza: dalle scelte terapeutiche alla guida autonoma, dalle pianificazioni strategiche[5] alle trattative commerciali, fino al supporto nei negoziati[6].

GOVERNANCE DELLA SICUREZZA NELL'ERA DIGITALE

Dal paragrafo precedente si evince che le applicazioni del Big Data analytics non possono prescindere dal machine learning e, parimenti, analisi rapide di grandi quantità di dati richiedono algoritmi sempre più efficaci. È la loro integrazione che può rendere pervasiva l'applicazione di Big Data e machine learning in molteplici ambiti della sicurezza. Tale diffusione porta a indubbi vantaggi, ma sta creando aspettative forse oltre il lecito, cambiando i rapporti di forza all'interno di tutte le organizzazioni.

Per non limitarsi all'ambito della sicurezza, sembra che oggi non si possa far politica nel mondo Occidentale senza schiere di spin doctor che, con strumenti di Big Data e machine learning, modelli di microtargeting e chiari obiettivi, decidono le priorità delle agende, le campagne pubblicitarie, le dichiarazioni da effettuare e il loro impatto[7]. Lo stesso sta avvenendo nel mondo aziendale, dove la business intelligence, supportata da algoritmi in grado di analizzare in tempo reale enormi quantità di dati, sta scalando dal livello marketing e commerciale verso il livello strategico. È questo trend che ha portato alla recente dichiarazione di Jack Ma, fondatore e presidente di Ali Baba: «Technology will make many Ceo irrelevant in the not-too-distant future. In 30 years, a robot will likely be on the cover of Time Magazine as the best Ceo».

4. Silver 2016.
5. Hartford 2016.
6. Papangelis 2015.
7. Helbing 2017.

Provocazione, fantascienza o realtà?

È interessante valutare se vi siano analoghe opportunità e rischi nell'ambito della sicurezza, in quanto le modalità per prendere decisioni basandosi sui dati sono analoghe e indipendenti dal contesto applicativo:

- identificazione delle sorgenti dati;
- acquisizione dati;
- elaborazione dati mediante algoritmi;
- analisi informazioni per arrivare alla conoscenza (o a una minore incertezza).

Il primo errore da evitare è aspettarsi che i dati parlino da soli, anche perché ciò indurrebbe alla bulimia informativa: poiché è più facile acquisire dati digitali in modo automatico, si ritiene utile ottenerne il più possibile. Il primo punto è fuorviante perché, senza avere un'idea dell'ago, è impossibile identificarlo; il secondo è altrettanto errato, in quanto avere un pagliaio sempre più grande non facilita la ricerca dell'ago. È una scelta politica consapevole e competente determinare le corrette dimensioni del pagliaio, anche perché le sorgenti che potrebbero alimentarlo aumentano a ritmi inauditi.

Su un punto le statistiche delle società Emc e Cisco convergono: nel 2016 siamo entrati nell'era degli Zettabyte anche prima del previsto. In particolare, la quantità di dati creata e copiata annualmente raddoppia ogni anno e raggiungerà i 44 Zettabyte nel 2020[8]. Analogamente, il traffico internet ha sorpassato la soglia dello Zettabyte nel 2016 e si stima che raggiungerà i 2,3 Zettabyte annuali nel 2020[9]. Tali cifre, al di fuori della nostra comprensione, fanno intuire che l'escalation del Big Data deve essere gestita in una prospettiva di breve-medio termine, e non solo dal punto di vista tecnico.

Come nella politica, nel business e nella finanza anche nell'ambito della sicurezza si stanno rimodulando i rapporti tra ruoli decisionali e ruoli tecnici, e tali modifiche vanno governate. Nella figura 2 si propone un modello di gestione del Big Data analytics dove si pone al centro l'uomo, cui spetta la definizione precisa dell'ambito e degli obiettivi, ma anche la valutazione politica di ciascun passo tecnico seguente. Anche le successive fasi, apparentemente tecniche, in realtà comportano decisioni fondamentali che andrebbero assunte in collaborazione tra ruoli tecnici e decisionali. Ad esempio, escluso che abbia senso ottenere tutto, quali sorgenti informative è conveniente utilizzare? E, passando alla fase successiva, quali modalità di acquisizione sono possibili, lecite, economicamente vantaggiose? Ancora più importante è la collaborazione nelle fasi di raffinamento dei dati grezzi (nella figura 2, indicizzazione ed elaborazione dati e analisi informazioni).

8. Emc 2014.
9. Cisco 2016.

I processi intermedi, basati su iterazioni successive, necessitano di un continuo riscontro fra i due principali attori: chi comprende gli aspetti politici e chi domina gli aspetti tecnici. In alcuni contesti, questo è sempre stato chiaro. In altre organizzazioni si stanno delegando ai ruoli tecnici decisioni che sarebbero riconducibili all'ambito di competenze manageriali, con danno per entrambi.

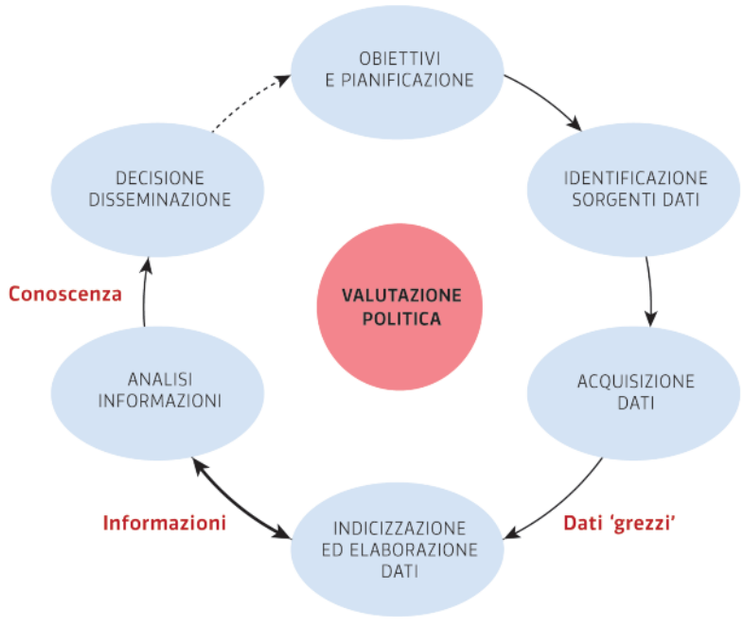

Figura 2. Ciclo di gestione e analisi dei dati.

TECNOLOGIE INFORMATICHE SEMI-AUTONOME PER LA SICUREZZA

Gli sviluppi di Big Data analytics e di machine learning aprono nuove prospettive anche nel campo delle tecnologie per la difesa autonoma o semiautonoma. In realtà, i primi strumenti automatici sono stati utilizzati dagli attaccanti e tale tendenza è in crescita. In uno studio di Radware, il 90% delle 300 grandi società intervistate aveva subìto un attacco cyber e la metà di questi era stato condotto mediante sistemi automatici[10]. Un altro ruolo determinante è svolto dalla produzione automatica di malware derivato da ceppi originali o ricombinandone parti. Grazie a tali mezzi, gli attaccanti riescono a produrre centinaia

10. Radware 2016.

di migliaia di nuovi campioni di malware al giorno. Di fronte a una simile capacità di fuoco, anche i difensori hanno necessità di potenziare gli strumenti semiautomatici per la difesa. Prendendo in considerazione le fasi principali della sicurezza informatica (prevenzione, individuazione dell'attacco, risposta emergenziale, ripristino e contromisure: figura 3), è necessario evidenziare quali attività e obiettivi potrebbero essere supportati o automatizzati da sistemi informatici basati su Big Data analytics e sul processo descritto nella figura 2. Il fil rouge che caratterizza i sistemi di difesa automatizzati è la capacità di essere eseguiti a velocità, intensità e continuità non raggiungibili da esperti umani. Le funzionalità difensive automatizzate più mature che stanno apparendo sul mercato includono soprattutto la fase di prevenzione e d'individuazione. Le attività di risposta emergenziale e di ripristino sono tuttora basate sull'impegno prevalente di esperti del settore, mentre si cominciano a intravedere sistemi di *forensic* semiautomatizzati, ma con supervisione umana. Interessante è la fase delle contromisure (già analizzata in precedenza) che potrebbe spingersi fino ad azioni di ritorsione.

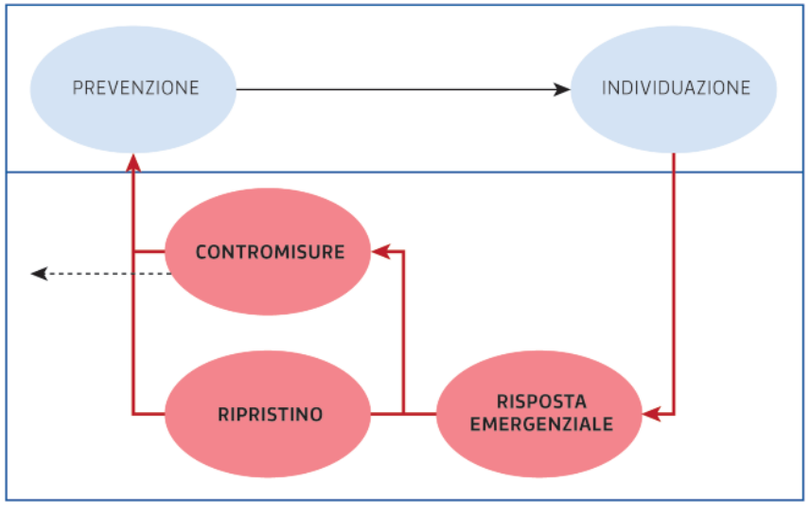

Figura 3. Le macro fasi della sicurezza.

PREVENZIONE

La fase di prevenzione è, per sua natura, intelligence, basata sull'acquisizione e analisi di grandi quantità di dati eterogenei da molteplici fonti. Da un lato, il crescente utilizzo di strumenti crittografici per le comunicazioni sta riducendo il potenziale delle modalità di acquisizione dei dati tradizionali. Dal-

l'altro, la diffusione del web, dei social network e del broadcasting radio-televisivo via internet ha portato l'open source intelligence a livelli di semiautomazione incomparabili rispetto al «Io leggo libri» di Turner/Redford nell'indimenticabile film. Oltre alle sorgenti aperte vi sono innumerevoli fonti grigie e dark. Ed è in tali contesti che trova applicazione la *cyber threat intelligence*, termine legato ai Big Data e divenuto popolare quanto impalpabile nei suoi confini applicativi[11]. Nella pratica, tali attività d'intelligence sono svolte soprattutto all'esterno del perimetro del cliente e sono finalizzate a identificare le minacce prima che possano concretizzarsi in attacchi. La raccolta dati per la cyber threat intelligence va da informazioni pubblicamente disponibili ad attività sotto copertura nei mercati sotterranei utilizzati per acquistare e vendere malware e nuovi strumenti di attacco, fino a sistemi automatici in grado di rilevare nuove vulnerabilità, tipologie di malware e lancio di campagne di social engineering mediante decine di migliaia di sensori distribuiti in tutto il mondo. A eccezione dello storico progetto Honeynet, gestito da un'organizzazione senza scopo di lucro, le attività di cyber threat intelligence sono di competenza di poche grandi aziende di sicurezza che operano su scala mondiale. Per essere efficace, la cyber threat intelligence deve essere perseguita in modo continuativo e, attesa la mole di dati da gestire e l'importanza della tempistica dei risultati, le tecniche di Big Data analytics risultano fondamentali.
Più recenti sono le applicazioni della threat intelligence all'interno e alla catena del valore di un'organizzazione, nonché l'utilizzo di strumenti simili per valutare minacce di tipo economico-finanziario, di protezione del marchio e della reputazione, dell'affidabilità di clienti e fornitori, e delle società dello stesso settore per scopi di competitive intelligence. Dal punto di vista collaborativo, conoscere in anticipo gli attacchi che hanno coinvolto industrie e organizzazioni dello stesso settore è uno degli indicatori più utili per individuare probabili attacchi che potranno essere subìti dalla propria.

11. Gartner rimane a un alto livello di astrazione, limitandosi a definire la cyber threat intelligence come «la conoscenza basata su prove, inclusi il contesto, i meccanismi, gli indicatori, le implicazioni e le raccomandazioni, relativamente a una minaccia esistente o emergente per le risorse dell'organizzazione, che può essere utilizzata per adottare decisioni utili per rispondere in anticipo a tale minaccia o pericolo».

INDIVIDUAZIONE

A differenza della sicurezza fisica, l'individuazione degli attacchi cyber è un problema di per sé. Se si escludono quelli di sabotaggio, i Denial-of-Service e i ricatti, le attività svolte dagli attaccanti a scopo d'intrusione, esfiltrazione dati e installazione di malware sono tese a eludere le difese basate sull'identificazione di modelli e tecniche di attacco noti. L'inefficacia degli antimalware tradizionali è ben nota ed è causata dalla produzione giornaliera di centinaia di migliaia di nuovi pattern che non possono essere integrati nell'antivirus con analoga rapidità, e dalla prevalenza di malware multilivello, polimorfico e metamorfico. Sono in corso diverse attività di ricerca per lo sviluppo di tecniche d'individuazione del malware basate sul comportamento e di agenti software in grado di riconoscere vulnerabilità del software interagendo autonomamente con vari componenti, nonché di progettare e applicare automaticamente patch a vulnerabilità identificate.
In continua crescita sono le ricerche e i prodotti di accertamento degli attacchi basati sull'identificazione di anomalie. L'idea di fondo è quella di stabilire il comportamento normale di un sistema e/o di un utente, in modo da individuare eventuali deviazioni causate dalle attività degli attaccanti. Questi approcci si basano sulla capacità di costruire e mantenere un modello comportamentale di sistemi informatici complessi con comportamenti umani tendenzialmente erratici. Un obiettivo simile richiede l'adozione di analisi avanzate in tempo reale su enormi quantità d'informazioni. Di conseguenza, ogni soluzione d'individuazione delle anomalie deve basarsi su una piattaforma scalabile di Big Data in cui i metodi di analisi siano derivati da diversi settori della statistica e del machine learning e adattati per essere applicati al dominio della sicurezza.

CONTROMISURE

La fase delle molteplici contromisure da adottare è la più aperta a uno studio multisettoriale in quanto potrebbe giungere fino ad azioni ritorsive nel campo cyber o addirittura fisico.
In realtà, le soluzioni attuate si limitano alla raccolta d'informazioni da fonti aperte e chiuse, mirate alla geolocalizzazione della fonte di attacco, alle interpretazioni per similarità e dissimilarità degli strumenti adottati, possibilmente integrate con informazioni derivanti da analisi geopolitiche ed economiche.

Proposte più aggressive si spingono fino alla possibilità di raccogliere informazioni sulla fonte apparente di attacco interagendo automaticamente con essa. In tal modo si potrebbe addivenire alla rilevazione di sue vulnerabilità, così da risalire ai livelli più alti della catena. Si ipotizzano, addirittura, forme futuristiche di reazione, di ritorsione automatica, in cui contromisure reattive siano adottate a seguito di attività di ricognizione sulla fonte apparente di un attacco, in cui molteplici strumenti di Big Data analytics si integrino con database di informazioni storicizzate.

In questo settore i limiti non sono tecnologici ma dettati da leggi nazionali, regolamenti internazionali e, soprattutto, dall'impossibilità di ricondurre gli attacchi in modo certo ai rispettivi autori, a causa della natura di internet e del software.

Nonostante gli sforzi, non c'è alcun caso dove si possa parlare di 'pistola fumante' che inchiodi il presunto colpevole. Neanche alcuni tra gli eventi più famosi, dove le responsabilità sono date per certe[12], si basano su prove garantite da terza parte. Sebbene l'attribuzione certificata sia ancora lontana o non dichiarabile, è il momento di estendere al dominio cyber i dibattiti sulle armi autonome, decisioni senza supervisione e relative implicazioni legali ed etiche. Questioni che esorbitano dall'economia di questo articolo, ma che meritano un approfondimento in ragione del frenetico sviluppo di Big Data, di machine learning e del recente deep learning.

CONCLUSIONI

La rilevazione di frodi, il filtraggio di spam della posta elettronica, l'interpretazione del linguaggio naturale, i sistemi di profilazione e raccomandazione commerciale, l'analisi dei video sono migliorati in modo sorprendente grazie al progresso del Big Data analytics combinato con il machine learning, reso possibile dalla continua crescita della potenza computazionale a costi sempre più ridotti. Le aspettative nell'applicazione di tali tecnologie alla sicurezza semiautonoma di grandi organizzazioni e del paese sono da valutare nella loro corretta dimensione. È pur vero che l'informatica che batte l'uomo nei giochi più complicati è una pietra miliare della ricerca, ma il gioco, sebbene complesso, si basa su regole rispettate

12. Ad esempio: Stuxnet, attributo a Usa e Israele; Shamoon, all'Iran; svariati Apt alla Russia, alla Cina e alla Corea del Nord.

dai giocatori; la vita, no. Finché le applicazioni informatiche saranno di supporto e supervisionate dall'uomo non vi saranno problemi, ma criticità sono già in vista[13]. È opinione diffusa che l'uomo abbia raggiunto i propri limiti, e alcuni miglioramenti di scala lineare potranno avvenire solo attraverso la collaborazione.

La tecnologia, mantenendo una crescita esponenziale, indurrà a domande del tipo: fino a quando i decision-maker e i governi accetteranno di limitare le potenzialità dell'informatica adeguandole alla lentezza, inaccuratezza, incuranza e indifferenza umane?

In altre parole, fino a quando l'informatica sarà di supporto e non prenderà decisioni in nostra vece? Fra dieci anni, di fronte a una diagnosi, ci fideremo più del medico o dell'informatica? E quale interpretazione di uno scenario economico, sociale, militare riterremo più valida? Senza tendere al catastrofismo, c'è da porre la giusta attenzione, in quanto l'impatto delle risposte potrebbe essere dirompente quanto altri fattori (quali cambiamenti climatici, flussi migratori, discriminazioni economiche, proliferazione nucleare) presenti in tutte le agende

13. Knight 2017.

BIBLIOGRAFIA

Cisco, *Visual Networking Index*, 2016.
Emc, *Digital Universe report*, 2014.
I. Goodfellow – Y. Bengio – A. Courville, *Deep Learning*, MIT Press, 2017.
G. Harrison, *Next Generation Databases: NoSql, NewSql, and Big Data*, Apress, 2015.
J.S. Hartford – J.R. Wright – K. Leyton-Brown, *Deep learning for predicting human strategic behavior*, Atti della conferenza *Neural Information Processing Systems*, Barcellona 2016.
D. Helbing et al., *Will democracy survive big data and artificial intelligence?*, «Scientific American» (febbraio 2017).
W. Knight, *The dark secret at the heart of* AI, «MIT Technology Review» (maggio-giugno 2017).
A. Papangelis – K. Georgila, *Reinforcement learning of multi-issue negotiation dialogue policies*, Atti della conferenza Sigdial 2015, Prague 9/2015.
Radware, *Global application and network security report – 2015-2016*, 2016.
D. Silver et al., *Mastering the game of Go with deep neural networks and tree search*, «Nature» (2016) 529.
D. Sullivan, *NoSql for Mere Mortals*, Addison-Wesley Professional, 2015.