theFuture ofScience andEthics

Rivista scientifica a cura del Comitato Etico
della Fondazione Umberto Veronesi

Volume 1 **numero 2** ■ novembre 2016

Fondazione Umberto Veronesi
– per il progresso delle scienze

Articoli

Big Data e integrità nella ricerca: un punto di partenza

Big Data and research integrity: a starting point

SILVIA SCALZINI
silvia.scalzini@gmail.com

AFFILIAZIONE
Scuola Superiore Sant'Anna, Pisa

ABSTRACT

Dopo aver delineato i profili del fenomeno Big Data, l'articolo si concentra sulle questioni etiche che esso solleva nell'ambito della ricerca con l'intento di stimolare una più approfondita riflessione all'interno della comunità scientifica.

ABSTRACT

This work aims at identifying the ethical issues raised by the Big Data phenomenon in scientific research in order to promote and broaden the discussion in the scientific community.

KEYWORDS

Big Data
Big Data

Etica
Ethics

Integrità nella ricerca
Research integrity

Scienza dei dati
Data science

1. IL FENOMENO DEI BIG DATA[1]

Sebbene non vi sia una univoca definizione del fenomeno dei Big Data, si tratta senza dubbio di una rivoluzione nel modo di produrre ed usare la conoscenza (Metcalf, Keller, Boyd 2016) con un rilevante, immediato, impatto sulla società (Mayer-Schönberger, Cukier 2013)[2].

Big Data è, infatti, una locuzione che indica lo sviluppo di tecnologie capaci di conservare, combinare ed analizzare enormi volumi di dati provenienti da fonti eterogenee (European Data Protection Supervisor 2015)[3] ed «ottenere dal trattamento di questi dati, grazie ad algoritmi che sappiano interrogare la macchina in modo da avere da essa la risposta voluta o la informazione ricercata, una quantità ancora più sterminata di nuovi dati, che consentano nuove conoscenze ed analisi relative ai fenomeni naturali ed ai comportamenti umani» (Pizzetti 2016: 15, nota 21).

Nonostante l'espressione evochi in primo luogo l'aspetto dimensionale del fenomeno[4], il cuore è rappresentato «dalla nuova espansiva capacità di connettere, attraverso l'analisi algoritmica, *datasets* disparati, forgiando relazioni tra dati raccolti in differenti momenti e luoghi e per diverse finalità» (Metcalf, Keller, Boyd 2016: 5)[5]. Un tale uso delle informazioni permette di creare nuovi dati, nuova conoscenza, al fine di fare previsioni e risolvere problemi, tanto che al fenomeno Big Data e alla "data analytics" è stato riconosciuto un ruolo importante per la crescita economica, lo sviluppo ed il benessere (OECD 2015).

I Big Data consentono, ad esempio, di avere un'informazione accurata e finanche in tempo reale di ciò che accade nelle città e sono di grande rilevanza per lo sviluppo urbano sostenibile ed innovativo, comunemente indicato con il nome di "smart cities[6]". La raccolta e l'analisi di grandi quantità di dati, inoltre, stanno avendo un impiego sempre maggiore nel settore sanitario, dove la combinazione di dati di origine e natura diverse – come dati sanitari, clinici, ambientali, comportamentali – ha un impatto notevole sulla ricerca clinica e sulla prevenzione, cura e gestione delle malattie[7].

Date le potenzialità ed i benefici derivanti da un tale utilizzo dei dati, non sorprende la centralità del tema

della circolazione dei dati nell'ambito delle strategie per lo sviluppo economico europeo[8]. È, inoltre, avvertito il bisogno di inquadrare e regolare tale inedito fenomeno sia da un punto di vista giuridico[9] che etico[10] al fine di permetterne uno sviluppo bilanciato. In questo lavoro sarà analizzato il loro impiego nell'ambito della ricerca scientifica, esaminando in particolare i problemi che sorgono in materia di integrità nella ricerca[11]. La direzione che la scienza saprà imprimere al fenomeno Big Data influenza, infatti, la reputazione della comunità scientifica e la fiducia della società in ciò che essa potrà offrire.

2. I BIG DATA NELLA RICERCA SCIENTIFICA

L'evoluzione delle tecnologie dell'informazione e della comunicazione e gli strumenti forniti dai Big Data hanno offerto straordinarie potenzialità espansive alla ricerca scientifica in tutti i campi del sapere.

Si sono moltiplicate, infatti, le possibilità di produzione e raccolta di dati, grazie anche alla velocità di diffusione della conoscenza tra ricercatori e tra discipline diverse, aprendo nuovi orizzonti di collaborazione e nuove frontiere inesplorate di ricerca. Un esempio tra tutti è la collaborazione sempre più stretta tra scienziati sociali e scienziati dell'informazione[12]. L'ampiamento esponenziale della base di dati disponibili e la possibilità di raccogliere e mettere in relazione dati di natura, fonti e strutture diverse rappresenta una vera e propria opportunità per la ricerca scientifica. I nuovi potenti strumenti di analisi dei dati, inoltre, consentono non solo di scoprire correlazioni inaspettate ma anche di ottenere risposte più rapide ed esaustive, contribuendo a elevare il valore economico e la reputazione sociale della ricerca scientifica, laddove essa sia condotta responsabilmente. Da qui discende l'importanza del ruolo della branca del sapere che si occupa dell'analisi dei dati, la "scienza dei dati", e dello sviluppo di macchine ed algoritmi capaci di analisi sempre più complesse. La fiducia nei risultati della ricerca, infine, amplia le possibilità di applicazione degli stessi in altri campi del sapere e nei processi decisionali pubblici[13]. I benefici di tale nuovo modo di condurre la ricerca scientifica e le maggiori possibilità di collaborazione multidisciplinare spingono verso la promozione di modelli aperti di condivisione dei dati e della conoscenza scientifica. Secondo l'OECD (Organisation for Economic Co-operation and Development), ad esempio, i modelli di "open science" e "open data" consentono di affrontare sfide globali, come il cambiamento climatico o la salute della popolazione, attraverso un migliore coordinamento tra scienziati provenienti da tutto il mondo (OECD 2015: 301 e 302) ed aprono ad un maggiore collegamento con la società, potendo i cittadini contribuire in modo ancor più incisivo alla raccolta di dati utili alla ricerca scientifica (OECD 2015: 304 ss.)[14].

Il panorama qui brevemente descritto non è, tuttavia, esente da rischi che possono riguardare l'integrità nella ricerca. Se la raccolta e l'utilizzazione dei dati nella ricerca scientifica hanno da sempre sollevato questioni etiche, le caratteristiche dei Big Data pongono nuovi problemi, che meritano l'attenzione della comunità scientifica.

3. NUOVI INTERROGATIVI PER ASSICURARE L'INTEGRITÀ NELLA RICERCA

Recentemente è sorta una riflessione sulle direttrici da seguire per garantire responsabilità e correttezza da parte dei ricercatori nell'utilizzo dei potenti mezzi di ricerca forniti dalla *Big Data analytics.* Un *white paper* del *Council for Big Data, Ethics and Society* – ente che collabora con la *National Science Foundation* statunitense – ha evidenziato alcune «disgiunzioni tra i metodi della ricerca con i Big Data e i paradigmi etici oggi adottati nella ricerca» (Metcalf, Keller, Boyd 2016: 2)[15]. Secondo il *white paper* il passaggio ad una «algorithmic knowledge production» (Metcalf, Keller, Boyd 2016: 6) necessiterebbe, infatti, di una revisione delle tradizionali categorie dell'etica della ricerca. Ad essere modificato è essenzialmente il modo di intendere i dati di ricerca, i quali possono ora essere oggetto di infinite connessioni, utilizzazioni, re-utilizzazioni, e decontestualizzazioni rispetto ai contesti originari di misurazione e raccolta (Metcalf, Crawford 2016: 5). La enorme aggregazione di dati differenti conduce, inoltre, ad una de-individualizzazione del soggetto (umano) coinvolto nella ricerca, spostando il centro di interesse verso classi, gruppi, classificazioni di soggetti che condividono le stesse caratteristiche. Mentre le classificazioni di per sé possono apparire neutrali, alcune modalità di loro produzione o utilizzo potrebbero presentare problemi, tra cui limitazioni all'autonomia dei soggetti individuati come facenti parte di determinati gruppi e rischi di discriminazione (Comandé 2016). I risultati di ricerche che rivelano «informazioni scomode sui 'gruppi' potrebbero, in-

fatti, diventare un tema di grande rilevanza per l'etica della ricerca[16]» (Zwitter 2014: 5). Da qui discende la necessità di rispettare la dignità della persona, onde evitare classificazioni che finiscano per stigmatizzare, escludere e discriminare (European Data Protection Supervisor 2015).

Un ulteriore elemento di discontinuità che la ricerca tramite Big Data fa emergere consiste nella rilevanza di nuovi tipi di rischi e "pregiudizi" che i ricercatori dovrebbero considerare e conseguentemente evitare di causare, come la discriminazione individuale o di gruppo o l'invasione della *privacy* tramite – non la raccolta di dati personali ma – l'analisi combinata di *dataset* contenenti informazioni non personali o anonimizzate che congiuntamente, tuttavia, svelano informazioni sensibili di un soggetto (Crawford Schultz 2014). Alcuni studi dimostrano, infatti, che per rendere anonimi alcuni dati la mera de-identificazione dei soggetti non è sufficiente, ma sono necessarie cautele ulteriori (Giannotti, Pedreschi, Pentland, et al. 2012)[17].

Tali problemi sono particolarmente frequenti laddove l'analisi prenda in considerazione dati provenienti da *social network* (Zimmer 2010). Queste fonti di dati, inoltre, pongono in rilievo un altro elemento fondamentale per la conduzione responsabile della ricerca: il consenso informato. Sorge, infatti, il dubbio dell'adeguatezza sia dal punto di vista giuridico che etico del consenso prestato dal soggetto, quando i dati che questi (più o meno consapevolmente) fornisce siano staccati dal contesto (anche temporale) originario ed utilizzati per le più svariate finalità. Il consenso informato prestato all'inizio di una ricerca potrebbe non adeguatamente considerare tutti i possibili benefici e rischi associati all'uso dei dati oggetto di ricerca (Metcalf, Keller, Boyd 2016: 7). Tale profilo è stato enfatizzato a seguito di un noto esperimento sociale condotto sugli utenti *Facebook* e pubblicato sulla rivista scientifica *Proceedings of the National Academies of Science* (Kramer, Guillory, Hancock 2014). Attraverso la artificiosa modifica del feed algorithm di alcuni utenti *Facebook*, gli scienziati[18] ipotizzarono la sussistenza di contagio emotivo su larga scala tra gli utenti dei *social network.* Essi dimostrarono, infatti, che la valenza emotiva positiva o negativa dei *post* che apparivano sulla bacheca degli utenti influenzava la valenza emotiva dei post che questi ultimi a loro volta pubblicavano. La critica principale mossa a tale esperimento è stata quella per cui gli utenti i cui *feed* erano stati manipolati non avevano specificamente acconsentito a partecipare a tale ricerca né avevano avuto la possibilità di *opt-out*[19], ma il loro "consenso informato" era stato ritenuto sussistente sulla base delle condizioni generali di contratto e la *Facebook Data Use Policy* sottoscritti al momento di adesione ai servizi *Facebook*, sollevando una serie di interrogativi giuridici ed etici (Metcalf, Keller, Boyd 2016; Grimmelmann 2014). Il rapporto di fiducia tra i ricercatori e gli individui coinvolti nelle ricerche basato su un'informazione trasparente riguardo la conduzione delle relative attività è, infatti, un elemento fondamentale, da curare in tutti gli ambiti della ricerca scientifica e, in special modo, ove vi sia l'impiego di algoritmi che in modo non immediatamente intellegibile analizzino vastissime quantità di dati relativi a tali soggetti.

Se questi sono i principali interrogativi emersi in relazione ai nuovi mezzi a disposizione della ricerca, non devono essere tralasciati ulteriori aspetti capaci di inficiarne l'integrità. Sia nella fase di *knowledge discovery,* ovvero la raccolta e l'analisi dei dati, che nella fase di *application* delle correlazioni tra dati per fondare decisioni e fare previsioni (Waterman, Bruening 2014) i ricercatori devono adottare particolari cautele per garantire la qualità e l'affidabilità della ricerca. Maggiori sono le potenzialità ed i benefici, maggiori sono anche i rischi laddove la ricerca, con ogni mezzo condotta, presenti errori o colpevoli manomissioni.

La fase di *knowledge discovery*, se non condotta appropriatamente, rischia di produrre risultati non accurati, la cui applicazione a fini predittivi può creare danni all'immagine ed alla credibilità della scienza. Anche nell'ambito dei Big Data, infatti, la raccolta, la selezione ed il controllo sulle fonti dei dati devono essere condotte diligentemente per non inficiarne la qualità. La disponibilità dei dati oggetto di analisi, inoltre, deve essere preventivamente vagliata, al fine di non incorrere in violazioni dei diritti altrui[20]. L'analisi dei dati potrebbe poi comportare ulteriori rischi derivanti dalla incompleta comprensione di alcuni dati o dagli stessi processi di analisi (Waterman, Bruening 2014). Special-

mente nei casi di applicazione delle tecniche di *data mining* alle scienze sociali è stato evidenziato come tali strumenti non siano spesso in grado di cogliere in maniera soddisfacente la complessità e diversità delle dinamiche sociali (Giannotti, Pedreschi, Pentland, et al. 2012: 51).

Con riferimento all'ultimo profilo, è stato evidenziato che gli algoritmi automatizzati, nel seguire le istruzioni per filtrare e sistematizzare le informazioni, generano un prodotto finale che omette di rendere visibili gli elementi di «incertezza, interpretazione soggettiva, scelte arbitrarie, inconvenienti» (Rosenblat, Kneese, Boyd 2014) che possono essere emersi durante il processo di *knowledge discovery*[21]. Per questo motivo, è discussa l'opportunità di specificare con trasparenza almeno i rischi associati ad ogni fase descritta ed i margini di errore ragionevolmente attendibili da tali analisi.

La fase di *application* dei risultati della *knowledge discovery* a fini predittivi potrebbe risultare, invece, troppo invasiva a causa dei rischi sopra descritti di discriminazione o *informational privacy harm.* Si pone un quesito etico, infatti, sulla direzione e applicazione delle previsioni possibili grazie all'analisi dei Big Data. Utilizzare tale analisi per identificare la propensione di un individuo (o di un gruppo di individui) all'insorgenza di una determinata malattia al fine di prevenirla più efficacemente differisce molto dall' utilizzare la stessa analisi per determinare il rischio assicurativo o per la sottoposizione a determinati trattamenti (Waterman, Bruening 2014). È opportuno poi sottolineare che, differentemente dal modo tradizionale di inferire relazioni tra dati di ricerca prelevati dai propri contesti di riferimento, gli algoritmi e l'analisi dei Big Data non necessariamente tracciano relazioni causali tra dati (Comandé, 2016). La comprensione che una previsione possa basarsi su una relazione non intellegibile tra dati è fondamentale per gestire più coscientemente la fase applicativa.

Un ultimo aspetto da considerare consiste nella maggiore propensione verso modelli di "open science" e "open data" nella ricerca scientifica per aumentare i benefici derivanti dai nuovi mezzi di analisi (si veda *sopra*). Se da un lato ciò conduce ad una maggiore circolazione della conoscenza, maggiore controllabilità dei risultati scientifici, alla possibilità di una riutilizzazione dei dati per finalità differenti, dall'altro non devono essere sottovalutati i rischi derivanti dal potenziale "dual use"[22] dei dati o dalla necessità di riconoscere il contributo ed il lavoro dei ricercatori nella raccolta e misurazione dei dati poi messi a disposizione per altre ricerche ed analisi.

4. RIFLESSIONI CONCLUSIVE E POSSIBILI SVILUPPI

Le sintetiche considerazioni ora esposte, lungi dal delineare una vera e propria conclusione, fungono da punto di partenza e da stimolo per approfondire la riflessione nelle sedi opportune in un dialogo che non può che essere interdisciplinare ed aperto. La comunità scientifica deve essere in grado di cogliere le straordinarie opportunità che le tecnologie dell'informazione e della comunicazione offrono, limitandone tuttavia i rischi ed operando in una cornice che assicuri il rispetto dei principi, dei valori etici, dei doveri deontologici e degli standard professionali su cui si fondano la reputazione e l'immagine pubblica della scienza (Linee Guida per l'Etica della Ricerca 2016).

BIBLIOGRAFIA

- Caporale, Cinzia, Fanelli, Daniele (2016), «L'integrità nella ricerca, una questione di standard», in *The Future of Science and Ethics*, 1, 154-167.

- Comandé, Giovanni (2016), «Regulating algorithms regulation? First ethico-legal principles, problems and opportunities of algorithms», *working paper*, non ancora pubblicato.

- Conte, Rosaria (2016), «Big data: un'opportunità per le scienze sociali?», in *Sociologia e ricerca sociale*, 109, 18-27.

- Crawford, Kate, Schultz, Jason (2014), «Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms», in *Boston College Law* Review, 55 (93); NYU School of Law, Public Law Research Paper No. 13-64; NYU Law and Economics Research Paper No. 13-36. Disponibile all'indirizzo: https://ssrn.com/abstract=2325784, ultima consultazione 3 novembre 2016.

- European Data Protection Supervisor (2015), «Towards a New Digital Ethics. Data, Dignity and Technology», disponibile all'indirizzo https://secure.edps.europa.eu/ EDPSWEB/edps/EDPS/cache/offon- ce?lang=it, ultima consultazione 3 novembre 2016.

● Gelman, Andrew, Loken, Eric (2013), «The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p- hacking" and the research hypothesis was posited ahead of time», *working paper,* Department of Statistics, Columbia University, disponibile all'indirizzo http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking. pdf, ultima consultazione 6 dicembre 2016.

● Giannotti F, Pedreschi D, Pentland AP, Lukowicz P, Kossmann D, Crowley J, Helbing D (2012), «A planetary nervous system for social mining and collective awareness», in *The European Physical Journal Special Topics*, 214 (1), 49-75.

● Grimmelmann, James (2015), «The Law and Ethics of Experiments on Social Media Users», disponibile anche all'indirizzo: https://ssrn.com/abstract=2604168, ultima consultazione 10 novembre 2016.

● Kitchin, Rob (2014), «The Real-Time City? Big Data and Smart Urbanism», in *Geo Journal*, 79(1), 1-14.

● Kramer, Adam, Guillory, Jamie, Hancock, James (2014), «Experimental evidence of massive-scale emotional contagion through social networks», in *Proocedings of the National Academies of Science*, 111 (24), 8788-8790.

● Mayer-Schonberger, Viktor, Cukier, Kenneth (2013), *Big Data: A Revolution That Will Transform How We Live, Work and Think*, Boston - New York, Eamon Dolan/ Houghton Mifflin Harcourt.

● Metcalf, Jacob, Crawford, Kate (2016), *«Where are human subjects in Big Data research? The emerging ethics divide»*, in Big Data & Society, 3 (1).

● Metcalf, Jacob, Keller, Emily, Boyd, Danah (2016), «Perspectives on Big Data, Ethics, and Society», *White Paper of the Council for Big data, Ethics and Society*, disponibile all'indirizzo http://bdes.datasociety.net/, ultima consultazione 16 novembre 2016.

● OECD (2015), *Data driven innovation: Big Data for Growth and Well Being*, OECD Publications, Paris

● Olivieri, Gustavo, Falce, Valeria (a cura di) (2016), *Smart Cities e Diritto dell'innovazione*, Milano, Giuffré.

● Pizzetti, Francesco Maria (2016), *Privacy e il Diritto Europeo alla Protezione dei Dati Personali.* Dalla Direttiva 95/46 al nuovo Regolamento europeo, Torino, Giappichelli.

● Raghupathi, Wullianallur, Raghupathi, Viju (2014), «Big data analytics in healthcare: Promise and potential», in *Health Information Science & Systems*, 2(3), disponibile all'indirizzo https://hissjournal.biomedcentral.com/track/pdf/10.1186/2047- 2501-2 - 3? site=hissjournal. biomedcentral.com, ultima consultazione 5 novembre 2016.

● Richards, Neil, King, Jonathan (2014), «Big Data Ethics», in *Wake Forest Law Review*, disponibile all'indirizzo https://ssrn.com/ abstract=2384174, ultima consultazione 20 ottobre 2016.

● Rosenblat, Alex, Kneese, T, Boyd, Danah (2014), «Algorithmic Accountability, The Social, Cultural & Ethical Dimensions of "Big Data"», disponibile all'indirizzo: https:// ssrn.com/abstract=2535540 or http://dx.doi.org/10.2139/ ssrn.2535540, ultima consultazione 10 novembre 2016.

● Waterman, Krasnow, Bruening, Paula (2014), «Big Data analytics: risks and responsibilities», in *International Data Privacy Law*, 4 (2), 89-95.

● Zimmer, Michael (2010), *«"But the data is already public": on the ethics of research in Facebook», in Ethics of Information and Technology*, 12, 313.

● Zwitter, Andrej (2014), *«Big Data Ethics»,* in *Big Data & Society*, 1 (2), 1-6.

NOTE

1. Si ringraziano il Professor Giovanni Comandé, il Professor Francesco Maria Pizzetti ed i revisori anonimi per aver fornito utili spunti e prospettive di ricerca.

2. Secondo la definizione degli Autori «Big Data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship

between citizens and governments, and more».

3. Le nuove tecnologie della comunicazione e dell'informazione hanno portato ad un incremento esponenziale del volume dei dati generati, le cui fonti sono classificate in dirette (dati acquisiti attraverso tradizionali strumenti di sicurezza o sorveglianza, dove la tecnologia si concentra su una certa persona o un determinato luogo), automatiche dati prodotti in ragione del funzionamento di uno strumento, di un sistema o di un device) e volontarie (dati trasferiti volontariamente dagli utenti) e la cui natura si divide essenzialmente in personale e non personale. Cfr. (Kitchin 2014). In letteratura per l'emersione del fenomeno si veda anche Richards et al. (2014).

4. Molte definizioni di Big Data si concentrano essenzialmente sulle caratteristiche "volume", "velocity" e "variety of information". *Cfr. IT Glossary: Big Data, GARTNER*, www.gartner.com/it-glossary/big-data/ Recentemente alcune definizioni del mondo imprenditoriale inseriscono anche una quarta "V" ovvero "veracity", che si riferisce alla variabilità della qualità dei dati raccolti. Si veda, ad esempio, IBM Big Data & Analytics Hub, www.ibmbigdatahub.com/tag/587

5. Traduzione a cura dell'autrice.

6. In tema di "smart cities" *cfr.*, in particolare, Olivieri e Falce (2016). Sulla rilevanza dei Big Data nello sviluppo delle smart cities cfr. Kitchin (2014).

7. *Cfr.*, in particolare, OECD (2015) e (Raghupathi 2014).

8. *Cfr.* Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *A Digital Single Market Strategy for Europe* – COM (2015) 192 final.

9. Sulle questioni inerenti i profili di titolarità, appropriabilità, disponibilità e trasferibilità dei dati personali e non personali, cfr., *ex multis, Position Statement of the Max Planck Institute for Innovation and Competition of 16. August 2016 on the Current European Debate, data Ownership and Access to data*, disponibile all'indirizzo www.ip.mpg.de. Per gli aspetti più direttamente inerenti la *privacy* e la protezione dei dati personali cfr., ad esempio, lo *Statement del Gruppo di lavoro articolo 29 "on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU*" adottato il 16 settembre 2014 (WP 221).

10. Sulle questioni etiche, oltre ai contributi già citati, si veda (Zwitter 2014).

11. "Per integrità nella ricerca si intende l'insieme dei principi e dei valori etici, dei doveri deontologici e degli standard professionali sui quali si fonda una condotta responsabile e corretta da parte di chi svolge, finanzia o valuta la ricerca scientifica nonché da parte delle istituzioni che la promuovono e la realizzano". Cfr. *Linee guida per l'integrità nella ricerca*, elaborate nell'ambito delle attività della Commissione per l'Etica della Ricerca e la Bioetica del Consiglio Nazionale delle Ricerche (CNR) e pubblicate integralmente in questa Rivista nell'ambito dell'articolo di Caporale e Fanelli (2016).

12. Testimonia una tale evoluzione la diffusione di progetti di ricerca congiunti finanziati dall'Unione Europea. Ne sono esempi il progetto SoBig Data (European Laboratory on Big Data Analytics & Social Mining) finanziato nell'ambito del programma Horizon 2020, http://www.sobigdata.eu/ ed il progetto FuturICT (Participatory Computing for Our Complex World) finanziato nell'ambito del programma FT7, http://futurict.inn.ac/. Cfr. anche Giannotti et al. (2012); in tema si vedano anche le preoccupazioni di Conte (2016).

13. Tali evoluzioni sono ampiamente descritte in OECD, cit. capitolo 7, *Promoting data driven scientific research.*

14. In tema si veda anche il documento "Amsterdam Call for Action on Open Science" (2016).

15. Nonostante l'ordinamento statunitense abbia caratteristiche peculiari in materia di *Research Integrity,* si ritiene che alcuni dei temi sollevati siano comuni all'impiego dei Big Data nella ricerca *tout court.*

16. Traduzione a cura dell'autrice.

17. A tali istanze risponde (parzialmente) anche il comma 1 dell'art. 89 del Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sul-

la protezione dei dati). Tali profili rilevano, inoltre, non solo per tutelare i singoli individui a cui si riferiscono le informazioni raccolte, ma anche per non divulgare le caratteristiche dei dataset aziendali contenenti i profili dei clienti, al fine di non danneggiare le imprese annullandone i vantaggi competitivi.

18. Gli autori di tale studio erano un data scientist alle dipendenze di Facebook e due scienziati sociali affiliati alla Cornell University.

19. Si veda la *"Editorial Expression of Concern and Correction"* firmata dall' Editor in Chief Inder M. Verma e contenuta nel volume 111, n. 29 della stessa Rivista.

20. Alcuni dati potrebbero essere, ad esempio, inseriti in banche dati proprietarie, ostacolandone dunque l'estrazione o riferirsi a informazioni personali per cui è richiesto un preventivo consenso da parte del soggetto interessato. In tema sarà da approfondire lo studio delle deroghe previste dall'art. 89 del nuovo Regolamento (UE) 2016/679.

21. Segnalano, inoltre, il problema delle false correlazioni e comparazioni multiple dei dati Gelman e Loken (2013).

22. Il termine "dual use" si riferisce in questo contesto all'ambivalenza della conoscenza e al problema degli usi impropri dei risultati scientifici.

Fondazione Umberto Veronesi
– per il progresso delle scienze