Quale ruolo per la metrologia nel mondo dei big data?
Luca Mari, Dario Petri

Tutto_Misure, 4, 2017

[30.10.17]

Nel 2010, in un intervento alla conferenza Techonomy, l'allora CEO di Google Eric Schmidt parlò di "five exabytes of information created by the entire world between the dawn of civilization and 2003", e poi commentò: "now the same amount is created every two days" (cinque exabytes di informazione creata dall'intero mondo tra l'inizio della civilizzazione e il 2003; ora la stessa quantità è creata ogni due giorni) (www.youtube.com/watch?v=UAcCIsrAq70&t=479s). Naturalmente si tratta di stime, da considerare solo in riferimento agli ordini di grandezza implicati. Se la quantità fosse di 6 exabytes, invece che di 5, o i giorni fossero 3, invece che 2, non cambierebbe il messaggio (e perciò non è nemmeno importante accertare se Schmidt stesse parlando effettivamente di exabytes, 10^{18} bytes, o di exbibytes, 2^{60} bytes, una pur rilevante differenza di oltre il 15% – per i multipli binari delle unità si veda la definizione 1.17, 'multiplo dell'unità', del Vocabolario Internazionale di Metrologia, www.ceinorme.it/it/normazione-it/vim.html). Insomma, la nostra società sta generando informazione in quantità *molto* superiore rispetto a quello che è accaduto fino a un passato anche recente: è il fenomeno chiamato, con la solita efficace brevità della lingua inglese, *big data*. Dato che la misurazione è uno strumento di produzione di informazione, non è sorprendente che il tema del *ruolo della metrologia in un mondo di big data* cominci a essere preso in considerazione. Con questo articolo, proponiamo un, certamente parziale e dunque incompleto, inquadramento dell'argomento, nell'auspicio che altri contributi, anche ospitati in questa rubrica di Tutto_Misure, possano seguire.

Un primo riferimento che suggeriamo è il numero uscito recentemente della rivista IEEE Instrumentation & Measurement Magazine (vol. 20, n. 5, ottobre 2017), che contiene vari articoli proprio su "Big data in instrumentation and measurement". Se ne coglie la multiformità del tema e, nello stesso tempo, la novità, che rende i contenuti proposti ancora prevalentemente esplorativi. Comune ai diversi articoli è il riconoscimento dell'importanza di valutare la *qualità dei dati* di cui si dispone: ricordando il modello cosiddetto "delle 4 V", (presentato efficacemente per esempio in www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg), si caratterizzano i big data non solo per la loro quantità (volume), la loro disponibilità di tempo reale o quasi-reale (velocity), e le loro diverse tipologie (dati numerici ma anche testi, immagini, ecc: variety), ma anche e criticamente per l'affidabilità che si attribuisce loro (veracity), cosa che in ambito metrologico sarebbe modellizzata, valutata, ed espressa in termini di incertezza.
Certamente non tutti i dati sono il risultato di misurazioni, e dunque non in tutti i casi la qualità di tali dati è formalizzabile appropriatamente come un'incertezza di misura. Ma il principio metrologico, così chiaramente enunciato dalla Guida all'Espressione dell'Incertezza di Misura rimane valido: "Nel riportare il risultato di una misurazione di una grandezza fisica, è obbligatorio fornire una qualche indicazione quantitativa della qualità del risultato, cosicché gli utenti ne possano accertare l'attendibilità. Senza tale indicazione i risultati di misura non possono essere confrontati né tra di loro, né con valori di riferimento assegnati da specifiche o norme." (dalla sezione 0.1 della traduzione italiana, CEI UNI 70098-3:2016; l'originale inglese è accessibile da www.bipm.org/en/publications/guides/gum.html).

Alcuni Istituti Metrologici Nazionali hanno cominciato a studiare i big data, intesi come una dimensione emergente di un fenomeno ancora più generale e complesso, la *data science*, e quindi includendo nell'analisi anche l'internet of things, il cloud computing, ecc. L'obiettivo che plausibilmente muove queste attività è anche strategico: fino a che si è in tempo, fare i passi giusti per stabilire un ruolo appropriato per la metrologia in un mondo che si sta sviluppando e nel quale la gestione dell'informazione ha un ruolo centrale.

Il National Physical Laboratory (NPL), britannico, è particolarmente attivo su questi temi. Il recente numero 9 della sua rivista "Insights", www.npl.co.uk/upload/pdf/insights-09-big-data.pdf, è proprio dedicato al tema dei big data. Iniziativa ancora più importante, nel dicembre 2016 l'NPL ha organizzato il "UK Workshop on Data Metrology and Standards", a cui hanno partecipato un centinaio di persone, dell'NPL stesso ma anche da università e industria, e il cui ampio resoconto è accessibile qui: www.bigdata.cam.ac.uk/files/npl-industry-workshop-on-data-metrology-standards/npl-industry-workshop-on-data-metrology-standards-report.
A testimonianza del riconoscimento di una situazione ancora assai poco consolidata, il workshop è stato impostato in modo chiaramente bottom-up, con ampio spazio per i contributi dei partecipanti, ai quali è stato chiesto di identificare temi per progetti rilevanti e di proporre priorità. Sono prima di tutto interessanti i problemi ("needs and challenges") più frequentemente considerati come rilevanti per il contesto industriale, e tra questi:
– condizioni per ottenere informazione di alta qualità che sia utile per il decision making a partire da fonti di dati di diversa tipologia;
– strumenti per la valutazione della qualità dei risultati degli algoritmi di intelligenza artificiale che operano su dati in tempo reale;
– standardizzazione di metadati sui sensori, di metodi di memorizzazione di dataset ottenuti da sensori, di sistemi di crittografia per dati forniti da sensori;
– metodi per la propagazione dell'incertezza attraverso algoritmi di organizzazione ed elaborazione dei dati.
(Una nota: con un po' di dubbi sulla qualità del risultato, abbiamo provato a tradurre in italiano i testi, in inglese, dell'NPL: dovremmo invece rinunciarci e accettare, per esempio, che "data curation and analytics" è più comprensibile di "organizzazione ed elaborazione dei dati"?)
Per contribuire alla soluzione di questi problemi i partecipanti al workshop hanno infine votato i progetti considerati più importanti e nello stesso tempo comunque fattibili. Nell'elenco, in ordine decrescente di priorità, figurano i seguenti (il documento di resoconto contiene una preliminare, breve introduzione per ognuno di questi).
1. Norme e modelli di ottimizzazione per la qualità dei dati, in riferimento ad accuratezza, affidabilità, ecc.
2. Norme e specifiche tecniche per la tracciabilità di dati e metadati generati in ambienti complessi, per esempio catene logistiche estese.
3. Algoritmi e metodologie innovativi per l'integrazione di dati da fonti di diversa tipologia.
4. Metodi e statistiche per la stima dell'incertezza nell'elaborazione mediante modelli spazio-temporali di dati forniti da sensori.
5. Applicazioni di high-performance computing, big data, e sistemi cognitivi in ambito scientifico e ingegneristico.
6. Norme a proposito della sicurezza dei dati.
7. Strumenti per l'estrazione automatica o semi-automatica di informazione da documenti scientifici.
8. Strumenti per l'organizzazione e annotazione di dataset molto grandi.
9. Strumenti per l'ottimizzazione integrata di catene logistiche.
10. Strumenti e norme per tarare e accertare l'affidabilità di sensori attraverso internet.
11. Strumenti per migliorare la qualità dei dati derivanti da misurazioni e simulazioni.
12. Modelli per l'analisi e la previsione del rischio utilizzando dati forniti da fonti numerose e di diversa tipologia.
13. Strumenti innovativi di elaborazione dei dati.
14. Norme etiche e strumenti derivati adatti alla raccolta di grandi quantità di dati.
15. Modelli innovativi per la memorizzazione, l'accesso, e la diffusione di dati in sistemi distribuiti, compatibili con le legislazioni esistenti.
Come si vede, non tutti questi progetti hanno una chiara e riconoscibile connotazione metrologica: ma in una situazione complessa e dinamica come quella attuale relativa ai big data, in cui i confini disciplinari sono

molto sfumati, una ben definita "demarcazione dei territori" – ammesso anche che sia possibile – potrebbe non essere la strategia più appropriata, e potrebbe comunque penalizzare la metrologia, che è inerentemente un campo di conoscenza trasversale. Pare dunque una scelta oculata quella di coloro che propongono di studiare il problema della qualità dei dati anche non generati da misurazioni a partire dalle solide basi dell'esperienza e delle tecniche della misurazione (è in fondo questo il messaggio che abbiamo suggerito nel nostro articolo, "The metrological culture in the context of big data: managing data-driven decision confidence", in apertura del numero dell'IEEE I&M Magazine a cui abbiamo fatto riferimento sopra).

Praticamente contemporaneo al progetto dell'NPL è quello del National Institute of Standards and Technology (NIST), statunitense, attivato nell'autunno del 2015 e presentato nell'ampio articolo di B.J. Dorr e altri, "A new data science research program: evaluation, metrology, standards, and community outreach", pubblicato sull'International Journal of Data Science and Analytics, 1, 3–4, 177–197, 2016, accessibile qui: link.springer.com/article/10.1007/s41060-016-0016-z, che riprende ed espande quanto inizialmente presentato in "The NIST Data Science Initiative" (ieeexplore.ieee.org/document/7344805). Anche in questo caso, è evidente l'obiettivo di fondare, a partire dalla metrologia, un programma di ricerca per la *data science*, in riferimento a quattro dimensioni principali, ma ora identificate in modo top-down:
– progettare e realizzare un programma internazionale di valutazioni nel contesto della data science (il NIST intende in questo caso con 'valutazione' un complesso processo di acquisizione (inclusa la taratura degli strumenti di acquisizione), organizzazione, analisi, presentazione, ... di dati);
– sviluppare campioni di misura (measurement standards) per la data science;
– sviluppare un sistema per la gestione delle valutazioni (*evaluation management system*, EMS), inclusivo delle risorse computazionali e infrastrutturali necessarie;
– promuovere l'attivazione di una comunità di interesse, in cui *data scientists* possano collaborare in modo efficace coordinando le loro attività su classi di problemi analoghi.
Le sfide ("challenges") identificate dal NIST per questo programma di ricerca hanno molte analogie con quelle proposte dall'NPL: tracciabilità dei dati (in inglese "provenance"), eterogeneità dei dati, analisi predittiva dei dati, acquisizione automatica di conoscenza dai dati, riproducibilità di grandi quantità di dati, visualizzazione dell'informazione, incertezza dei dati, propagazione degli errori, riservatezza e sicurezza.

Un denominatore fondamentale, anche se implicito, comune a questi progetti pare essere il riconoscimento che la quantità dei dati, anche quando enorme, non sia comunque sufficiente per garantire la qualità dell'informazione che dai dati si può ottenere: occorrono modelli, e dunque conoscenza di contesto (con buona pace di Chris Anderson, che in un noto articolo del 2008 su Wired sostenne, certo anche con un po' di gusto per la provocazione, il contrario: "The end of theory: the data deluge makes the scientific method obsolete"; www.wired.com/2008/06/pb-theory). La misurazione può essere interpretata come uno strumento per acquisire sperimentalmente *dati* affidabili e quindi trasformarli in *informazione* e, se i modelli sottostanti sono sufficientemente ricchi, in *conoscenza* (siamo facendo riferimento alla "piramide dati-informazione-conoscenza-saggezza": en.wikipedia.org/wiki/DIKW_pyramid). In questo, la cultura metrologica può svolgere un ruolo davvero strategico nel mondo dei big data. Anzi: degli *smart data*, caratterizzati non più da 4, ma da 5 "V". Non solo volume, velocità, varietà, e veracità, ma anche *valore*, cioè effettiva utilità per le decisioni da prendere.