

(Big) Data Science e discipline storico-archeologiche: una sfida metodologica, tecnologica e culturale

Claudio Cortese, 4Science S.r.l., claudio.cortese@4science.it

Ci troviamo nella “Big Data Age”. Da diversi anni ormai il termine “Big Data” è entrato prepotentemente nel mondo dell’Information Technology, e nell’immaginario collettivo, per le nuove promettenti potenzialità, anche e soprattutto in termini economici, legate all’utilizzo di una nuova generazione di tecnologie e architetture in grado di estrarre valore dall’enorme mole di dati che viene continuamente prodotta negli ambiti più diversi. In questo senso, i più grandi attori del web stanno investendo miliardi di dollari in attività di ricerca e sviluppo di tecnologie per l’analisi di tali masse di dati.

In ambito scientifico i “Big Data” vengono visti come un’opportunità ancora più grande, qualcuno sostiene addirittura epocale: le nuove tecnologie e i moderni strumenti con cui gli studiosi di discipline diverse analizzano i fenomeni di loro interesse permettono, infatti, di produrre e conservare una quantità di dati di diversi ordini di grandezza superiore a quelli disponibili in precedenza (nel 2012 si stimava che il 90% dei dati presenti al mondo fosse stato prodotto nel biennio precedente – Silver 2012 -). In questo senso c’è anche chi ha affermato che il cosiddetto “data deluge” (diluvio di dati) renderà obsoleti alcuni dei concetti fondamentali su cui si è finora basato il metodo scientifico (modello, teoria, ipotesi, spiegazione), poiché, in presenza di una tale quantità di dati, i numeri sarebbero ormai in grado di “parlare da soli” (Anderson 2008).

Senza arrivare ad ipotizzare la definizione di un nuovo paradigma scientifico (Wilbanks 2009), sicuramente l’avere a disposizione e l’essere in grado di manipolare e analizzare enormi quantità di dati rappresenta un significativo progresso sia per la scienza sia per altri ambiti di ricerca, non in quanto abolirà la necessità di costruire, raffinare e verificare teorie, ma in quanto permetterà di formulare le ipotesi e di testarle in tempi infinitamente più rapidi e su un campione infinitamente più vasto rispetto a quanto avveniva in passato.

È proprio in quest’ottica che, da qualche anno, il concetto di “Big Data” ha iniziato ad affacciarsi anche tra gli storici e gli archeologi. Se, infatti, è vero che storia e archeologia non sono interessate, al momento, dal fenomeno del “data deluge”, esse stanno comunque assistendo ad una crescita dei dati a loro disposizione, grazie alla sempre maggiore diffusione di database, di riviste elettroniche, di digitalizzazioni del patrimonio culturale e di strumenti per l’estrazione e l’analisi dei dati (per una carrellata dei progetti più importanti emersi negli ultimi 10-15 anni si vedano Boonstra, Breure & Doorn 2004; Kristiansen 2014; Holm 2015). Per la prima volta, dunque, ci si trova di fronte alla possibilità di confrontare migliaia, se non milioni, di reperti archeologici, libri, opere d’arte, fonti archivistiche, ecc.

In particolare, in questi settori, l’interesse per la problematica dei “Big Data” è legato soprattutto alla possibilità di aggregare, trovare relazioni e analizzare in maniera integrata la molteplicità di dati che sono necessari per rispondere ai quesiti storici. La storia e l’archeologia moderne sono, infatti, spesso caratterizzate da un approccio multidisciplinare e dunque dalla varietà in termini di tipologia, formato, struttura e scala dei dati che, tra l’altro, possono essere patrimonio di istituzioni e gruppi di ricerca

differenti. Le ricerche storiche che forse più hanno segnato l'ultimo secolo, del resto, sono proprio quelle che sono riuscite ad utilizzare e integrare tutte le fonti disponibili, secondo i dettami della "lunga durata", della "storia globale" e della "storia comparativa", portati avanti, a partire dai decenni centrali dal secolo scorso, soprattutto da Fernand Braudel (1985, 1993, 2002) e dalla "Scuola delle Annales" (Le Goff 1980; Pomian 1980).

In questo senso, è stato più volte sottolineato il rapporto tra la crescente quantità di dati a disposizione e il rinnovato interesse per questo tipo di studi (Van Eijnatten J., Pieters T. & Verheul J. 2013; Kristiansen 2014; Larsson 2014; Guidi & Armitage 2015) in cui l'obiettivo è quello di riuscire a correlare una mole sempre più vasta di fonti diverse, per indagare meglio l'articolazione dei fenomeni storici e dei processi di trasformazione che hanno interessato la storia umana, come, ad esempio, sta già avvenendo con l'integrazione tra i dati paleoclimatici e quelli storici e archeologici (McCormick et al. 2012; Haldon et al. 2014).

Gli strumenti principali per affrontare questa sfida, come peraltro, qualunque progetto che richieda la gestione e l'analisi dei dati (a prescindere dalla loro quantità), sono il Data Management e la Data Science (dalla modellazione dei dati, al text e data mining, ai modelli predittivi, al machine learning, all'analisi del linguaggio naturale, alle simulazioni, all'intelligenza artificiale, alle analisi spaziali mediante Geographic Information System, alle tecniche di visualizzazione) che però devono tenere conto della peculiarità e delle caratteristiche dei dati pertinenti agli specifici domini di ricerca (Shennan 1996; Baxter 2003; Boonstra, Breure & Doorn 2004).

In ambito storico-archeologico, infatti, nella maggior parte dei casi, i dati non vengono generati da strumenti ma dagli autori che producono un'opera letteraria o un documento di archivio o dagli studiosi che descrivono un oggetto o un contesto archeologico o, ancora, da chi modella le cosiddette "metafonti" (Genet 1994); quindi, spesso, non sono "neutri" ma possono essere condizionati dalla persona, dal tempo e dal luogo in cui sono stati prodotti. Inoltre le informazioni su cui lavorano la storia e l'archeologia sono frammentarie, parziali e, in molti casi, distorte. Infine il "digitale" rappresenta solo una piccola parte del patrimonio, delle fonti e dei documenti su cui gli studiosi basano le loro interpretazioni e, anche il giorno in cui la maggioranza del patrimonio culturale sarà digitalizzata, essa sarà, comunque, spesso, solo un surrogato che non necessariamente potrà sostituire la visione diretta della fonte informativa originaria (sia essa una pentola in ceramica, un manoscritto o un dipinto).

Nell'analisi dei dati digitali è fondamentale, quindi, che essi non vengano considerati isolatamente ma congiuntamente a tutte le informazioni contestuali, digitali e non, necessarie a rispondere alle domande della ricerca. È necessario quindi, ad esempio, indagare e analizzare anche il contesto in cui è stato prodotto un documento/monumento (Foucault 1971; Le Goff 1978, 1980), ricostruire e tener conto dei processi formativi che hanno caratterizzato il deposito archeologico da cui i dati provengono (Schiffer 1996) o analizzare le associazioni contestuali che caratterizzano i documenti, i monumenti e i reperti, a diversi livelli e su scale diverse (Le Goff 1978; Hodder 1992), in un processo interpretativo che deve dare il giusto peso e cercare di spiegare anche le assenze, le lacune o i "silenzi" della storia, insomma i dati che non ci sono. È già stato sottolineato, del resto, con particolare riferimento all'ambito epigrafico (Lamé 2015), come proprio un approccio umanistico-informatico favorisca lo studio delle fonti primarie come "dispositivi", nel senso in cui il termine è utilizzato da Michel Foucault, in altre parole come testimonianze di una rete di sistemi eterogenei (sociale, economico, culturale, tecnologico, ecc.) le cui relazioni possono essere indagate attraverso di esse, mediante un'analisi globale e multidimensionale.

Tra l'altro, la gestione del contesto nell'analisi dei Big Data è considerata, anche al di fuori dell'ambito umanistico, una delle sfide fondamentali per arrivare a utilizzarli in maniera efficace, sfruttandone appieno il potenziale (Boyd & Crawford 2012). Importanza decisiva acquisiscono

dunque il background culturale e la capacità critica dello studioso, fondamentali contro il pericolo di una “decontestualizzazione” dei dati.

In quest’ottica, se un approccio “(Big) Data Driven” può essere auspicabile in ambito storico e archeologico, è necessario essere molto accorti dal punto di vista metodologico per evitare di cadere in modo acritico nel “feticismo del numero”. È quindi necessario che i data analyst e i data scientist, oggi, e in un futuro in cui auspicabilmente i Big Data diventeranno sempre più centrali anche nelle humanities, abbiano elevate competenze non solo di tipo informatico e statistico ma anche di dominio. Solo un esperto di dominio, infatti, è in grado di dire, ad esempio, se una correlazione statisticamente significativa è rilevante anche dal punto di vista storico o archeologico. Già alcuni decenni fa, trattando l’emergere della “storia quantitativa” ed evidenziando come ormai il dato (e non il fatto) costituisse l’unità di informazione fondamentale per lo storico, Jacques Le Goff (1980, p. 37) sottolineava come “la parte essenziale del lavoro storico resta, comunque, ancora da fare quando il calcolatore ha sfornato i suoi risultati” (in questo senso si vedano anche Gattiglia 2015; Holm 2015).

C’è e ci sarà, quindi, sempre più bisogno di storici e archeologi specializzati nella gestione, nell’integrazione e nell’analisi dei dati, discipline che, a livello universitario, non dovrebbero essere più relegate a laboratori di poche ore o, eventualmente, alla formazione post-laurea magistrale, ma che dovrebbero costituire una parte essenziale del bagaglio culturale di uno studente sin dal primo triennio universitario. Del resto i tempi dovrebbero essere ormai maturi perché l’informatica applicata alle discipline umanistiche, smetta di essere vista come una “disciplina ausiliaria”, che si può scegliere o meno di utilizzare, ma inizi ad essere considerata per quello che è, ovvero un fondamentale e irrinunciabile strumento di educazione alla modellazione e formalizzazione del ragionamento storico e archeologico, in grado di avere una ricaduta sulla metodologia di tali discipline in senso lato, e, tramite la creazione di “metafonti”, di valorizzare, anche mediante l’uso di adeguate tecniche quantitative, la rete di legami contestuali fondamentali per la comprensione dei dati (si vedano in questo senso, ad esempio, Gardin 1995 e 1996; Lamé 2015; Gattiglia 2015; Cortese 2016). Se le humanities potranno in futuro trarre benefici dalle possibilità offerte dai Big Data è una questione legata, dunque, anche a come l’università saprà adattare i suoi programmi a queste esigenze sempre più urgenti.

Un ulteriore passo necessario a favorire l’uso e, soprattutto, la diffusione dei metodi della (Big) Data Science e di una modalità di ricerca scientifica “Data Driven” nell’ambito delle discipline in esame è la disponibilità di infrastrutture e strumenti per l’integrazione, la condivisione, l’analisi e la conservazione dei dataset, e che, nello stesso tempo, rendano possibile interagire con fonti di dati esterne.

In questo senso, i VRE (Virtual Research Environment) potrebbero costituire la risposta alle necessità di un “ambiente” scalabile e sostenibile, finalizzato a gestire l’intero ciclo di vita del dato in modo collettivo. Si tratta di piattaforme che, da certi punti di vista, possono essere considerate un’evoluzione dei repository o delle Digital Library, in grado di mettere a disposizione degli studiosi strumenti di condivisione e software di analisi in un ambiente integrato, all’interno del quale i dati possono essere processati con una velocità molto maggiore rispetto a quanto può avvenire con un normale PC. Solo in Europa esistono ormai decine di progetti finalizzati a creare infrastrutture di questo tipo, alcune funzionali a gestire la ricerca relativa ad un singolo dominio, altre invece con scopi più generali (in particolare, per le Digital Humanities, si veda Blanke et al. 2010), che però, sono ancora utilizzate per lo più in ambito sperimentale e da gruppi ristretti di ricercatori, spesso provenienti dalle istituzioni che hanno partecipato allo sviluppo delle stesse. Il prossimo obiettivo dovrebbe essere, dunque, quello di riuscire a inserire gli strumenti tecnologici creati nell’ambito di questi progetti nella pratica quotidiana della ricerca. Ciò sarebbe tanto più utile in contesti come quelli umanistici dove è rara la disponibilità di avanzate infrastrutture tecnologiche e dove, nella maggior parte dei casi, i dati risiedono sui PC personali dei ricercatori. Sulla base di quanto detto in

precedenza, tali sistemi dovranno permettere agli studiosi di analizzare i dati, evidenziandone e valorizzandone le relazioni a diversi livelli e di esplicitare le loro interpretazioni rispetto alle dimensioni di variabilità significative e alla rete di legami contestuali che interessano le fonti storico-archeologiche. Dovranno dunque avere nella flessibilità del modello dei dati, oltre che negli strumenti di integrazione e di analisi, una caratteristica fondamentale.

La storia, l'archeologia, e, probabilmente, gli studi umanistici in generale si trovano, quindi, di fronte a una sfida metodologica (adattare al meglio i metodi della Data Science alle peculiarità dei propri dati, facendo tesoro degli strumenti teorici e metodologici messi a punto nel corso della storia degli studi), tecnologica (contribuire alla realizzazione di strumenti che rendano tali metodi più facilmente ed efficacemente utilizzabili da parte della comunità scientifica) e culturale (prendere coscienza del fatto che la Data Science deve entrare a pieno diritto nel percorso formativo dei giovani che si avvicinano a queste discipline). Solo se sarà in grado di vincere questa sfida, valorizzando dunque i portati del "patrimonio genetico" delle scienze storiche all'interno di un approccio di tipo nuovo, in grado di integrare il "tradizionale" lavoro ermeneutico e interpretativo dello storico e dell'archeologo e le più efficaci tecniche di gestione e analisi dei dati, la comunità degli studiosi nel suo complesso potrà trarre tutti i benefici insiti nella sempre crescente quantità di dati disponibili, che altrimenti rimarranno patrimonio, non condiviso, e forse nemmeno riconosciuto, solo dei gruppi più o meno ristretti che si occupano di "Digital Humanities".

Riferimenti bibliografici

Anderson C. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired Magazine*, June 23. Accessed November 7, 2016, <https://www.wired.com/2008/06/pb-theory/>

Baxter M. 2003. *Statistics in Archaeology*. London: Arnold.

Blanke T., Candela L., Hedges M., Priddy, M. & Simeoni F. 2010. "Deploying general-purpose virtual research environments for humanities research". *Philosophical Transactions of the Royal Society A* 368: 3813-3828.

Boonstra O., Beure L. & Doorn P. 2004. *Past, present and future of historical information science*. Amsterdam: DANS.

Boyd, D. & Crawford, K. 2012. "Critical Questions for Big Data". *Information, Communication and Society* 15:5: 662-679.

Braudel F. 1985. *Una lezione di storia*. Torino: Einaudi.

Braudel F. 1993. *Civiltà materiale, economia e capitalismo. Le strutture del quotidiano (secoli XV-XVIII)*. Torino: Einaudi.

Braudel F. 2002. *Civiltà e imperi del Mediterraneo nell'età di Filippo II*. Torino: Einaudi.

Cortese C. 2016. "Informatica applicata all'archeologia: 'disciplina ausiliaria' o fondamentale strumento di educazione alla modellazione e formalizzazione del ragionamento archeologico?". In *Archeologia classica e post-classica tra Italia e Mediterraneo. Scritti in ricordo di Maria Pia*

Rossignani, a cura di S. Lusuardi Siena, C. Perassi, F. Sacchi & M. Sannazaro, 571-576. Milano: Vita e Pensiero.

Foucault M. 1971. *Archeologia del sapere*. Milano: Rizzoli.

Gardin J. C. 1995. "L'analisi delle costruzioni scientifiche". In *L'architettura dei testi storiografici. Un'ipotesi*, a cura di J. C. Gardin & M. N. Borghetti, 25-110. Bologna: CLUEB.

Gardin J. C. 1996. "Cognitive issues in archaeology". *Archaeologia Polona* 34: 205-232.

Gattiglia G. 2015. "Think big about data: Archaeology and the Big Data challenge". *Archäologische Informationen* 38: 113-124.

Genet J. Ph. 1994. "Source, Métasource, Texte, Histoire". In *Storia & multimedia*, a cura di F. Bocchi & P. Denley, 3-17. Bologna: Grafis.

Guidi J. & Armitage D. 2015. *The History Manifesto*. Cambridge: Cambridge University Press.

Haldon J., Roberts N., Izdebsky A., Fleitman D., McCormick M., Cassis M., Doona O., Eastwood W., Elton H., Ladstatter S., Manning S., Newhard J., Nicoll K., Telelis I. & Xoplaki E. 2014. "The Climate and Environment of Byzantine Anatolia: Integrating Science, History and Archaeology". *Journal of Interdisciplinary History* XLV:2: 113-161.

Hodder I. 1992. *Leggere il passato. Tendenze attuali dell'archeologia*. Torino: Einaudi.

Holm P. 2015. "Climate Change, Big Data and the Medieval and Early Modern". In *Medieval or Early Modern. The Value of a Traditional Historical Division*, edited by R. Hutton, 70-85. Newcastle: Cambridge Scholars Publishing.

Kristiansen K. 2014. "Towards a new paradigm? The Third Science Revolution and its Possible Consequences in Archaeology". *Current Swedish Archaeology* 22: 11-34.

Lamé M. 2015. "Primary Sources of Information, Digitization Processes and Dispositive Analysis". In *Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem*, edited by F. Tomasi, R. Roselli Del Turco, A. M. Tammaro: ACM.

Larsson F. 2014. "The third science revolution and its possible consequences in archaeology. A personal reflection". *Current Swedish Archaeology* 22: 53-56.

Le Goff J. 1978. "Documento/Monumento". In *Enciclopedia*, 38-43. Torino: Einaudi.

Le Goff J. 1980. "La nuova storia". In *La nuova storia*, a cura di J. Le Goff, 7-46. Milano: Arnoldo Mondadori Editore.

McCormick M., Buntgen U., Cane M. A., Cook E. R., Harper K., Huybers P., Litt T., Manning S. W., Mayewsky, More A.M.F., Nicolussi K. & Tegel W. 2012. "Climate Change during and after the Roman Empire: Reconstructing the Past from Scientific and Historical Evidence". *Journal of Interdisciplinary History* XLIII:2: 169-220.

Pomian K. 1980. "Storia delle strutture". In *La nuova storia*, a cura di J. Le Goff, 81-110. Milano: Arnoldo Mondadori Editore.

Schiffer M.B. 1996. *Formation Processes of the Archaeological Record*. Salt Lake City: University of Utah Press.

Shennan S. 1996. *Quantifying Archaeology*. Edinburgh: University Press.

Silver N. 2012. *The Signal and the Noise*. New York. The Penguin Press.

Van Eijnatten J., Pieters T. & Verheul J. 2013. "Big Data for Global History. The transformative Promise of Digital Humanities". *Low Countries Historical Review*: 128-4: 55-77.

Wilbanks J. 2009. "I Have Seen the Paradigm Shift and It Is Us". In *The Fourth Paradigm. Data-Intensive Scientific Discovery*, edited by T. Hey, E. Tansley & K. Tolle, 209-214. Redmond, Washington: Microsoft Research.