IL COMPUTER COME MACROSCOPIO

Questo volume rappresenta la sintesi più attuale, ambiziosa e completa in lingua italiana sulla scienza sociale computazionale.
(dalla prefazione di *Claudio Cioffi-Revilla*)

L'esplosione degli strumenti per la gestione dei big data e la diffusione dei media digitali (Facebook, Twitter tra gli altri) si stanno traducendo in una sfida sempre più ardua alle nostre capacità di comprendere la contemporaneità.

Nell'ottica di chi oggi si occupa di analisi dei processi sociali, tanto nella ricerca accademica quanto in quella di mercato, il libro presenta nuovi strumenti concettuali (e operativi) per rendere conto di questa complessità.

Attraverso la metafora del macroscopio – il computer come strumento in grado di visualizzare processi estesi nel tempo e nello spazio – **il libro mostra l'approccio che lega insieme le scienze sociali con le metodologie informatiche: la scienza sociale computazionale** (*computational social science*).

Partendo dalla descrizione dei programmi di ricerca (dalla *network science* alle *digital humanities*), che hanno contribuito a integrare le scienze sociali con l'analisi computazionale, si arriva a impostare **i problemi relativi alle conseguenze etiche che sta ponendo questa collaborazione fra scienze sociali e scienze del computer**. La sovrabbondanza di dati è davvero un aiuto alla ricerca? Fino a che punto è legittimo usare archivi digitali che contengono informazioni sulla vita delle persone? Quali sono i nuovi volti del controllo e della sorveglianza che queste tecnologie portano con sé?

Sono solo alcune delle domande sollevate dallo scenario socio-tecnologico attuale e a cui è chiamata a rispondere una nuova generazione di ricercatori.

Davide Bennato, insegna Sociologia dei media digitali al Dipartimento di Scienze Umanistiche dell'Università di Catania. I suoi interessi di ricerca sono relativi all'analisi dei comportamenti collettivi nei social media, all'etica dei big data, al rapporto fra tecnologia e valori, ai modelli di comunicazione scientifica e tecnologica in rete.

Fra le sue pubblicazioni: *Le metafore del computer* (Meltemi 2003), *Sociologia dei media digitali* (Laterza 2011), *La dataveglianza di massa. Conseguenze etiche e relazionali delle scelte tecnologiche di Facebook* (in G. Greco, a cura, *Pubbliche intimità*, FrancoAngeli 2014), *Etica dei big data. Conseguenze sociali della raccolta massiva di informazioni* (in «Studi Culturali», 2014).

FrancoAngeli
La passione per le conoscenze

ISBN 978-88-917-1135-9

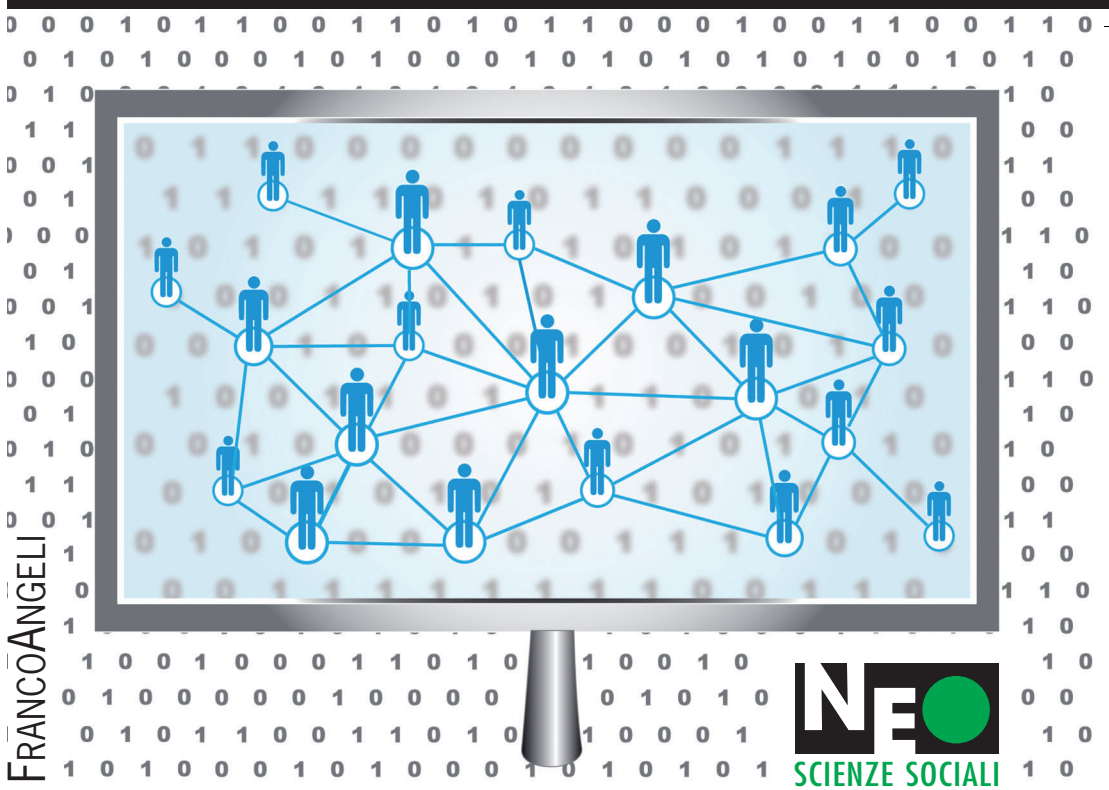
€ 18,00 (V)

666.3 D. BENNATO IL COMPUTER COME MACROSCOPIO

Davide Bennato

IL COMPUTER COME MACROSCOPIO

Big data e approccio computazionale per comprendere i cambiamenti sociali e culturali



FRANCOANGELI

NEO SCIENZE SOCIALI

Davide Bennato

IL COMPUTER COME MACROSCOPIO

Big data e approccio computazionale per comprendere i cambiamenti sociali e culturali

FRANCOANGELI

NEO SCIENZE SOCIALI

settore emergente è quello dei mercati di lavoro online (Berinsky, Huber, Lenz 2012; Mason, Suri 2012; Rand 2012). In questa tipologia di mercati esistono due tipologie di soggetti: chi domanda lavoro e chi offre lavoro. Chi domanda lavoro assegna arbitrariamente un prezzo a una serie di piccole attività ripetitive (micro task) che hanno bisogno delle persone per essere svolte (per esempio la classificazione di contenuti, il *tagging*[25], l'archiviazione). Chi offre lavoro si impegna a svolgere il micro task secondo le regole previste e il suo guadagno è proporzionale al numero di task svolti. In questo modo è possibile svolgere compiti ripetiti, impossibili (o costosi) da svolgere con sistemi automatici attraverso l'aiuto di un certo numero di persone. L'esempio principe in questo senso è Amazon Mechanical Turk[26], un mercato del lavoro online molto utilizzato per diverse attività come per esempio esperimenti nel campo delle scienze sociali o attività di ricerca nel campo dell'informatica.

2.2.7. Digital humanities e culturomics: il computer strumento dell'umanista

Il rapporto fra informatica e scienze umane – letteratura, filosofia, archeologia eccetera – è un rapporto molto antico dovuto al fatto che l'informatica è una disciplina che è basata tra le altre cose su due concetti chiave che sono il linguaggio (di programmazione) e la comunicazione (con una particolare tecnologia). Pertanto è possibile trovare connessioni tra computer e discipline umanistiche fin dai primi calcolatori. Anzi per certi versi è possibile considerare il computer una tecnologia appartenente alla stessa famiglia del libro e dell'orologio: una tecnologia caratterizzante, ovvero una tecnologia in grado di fornire legami metaforici (e non solo) con la cultura del periodo che l'ha prodotta (Bolter 1984). Non solo, è anche una tecnologia caratterizzata, ovvero il suo sviluppo tecnologico e sociale è figlio di una metafora, di un modo di descrivere il concetto di comunicazione, frutto della cultura del proprio tempo (Bennato 2002: 41-45). A ogni modo il rapporto stabile sia teorico che metodologico fra computer e scienze umane è frutto del programma di ricerca delle *digital humanities*

[25] Il tagging è lattività di assegnazione di una etichetta di testo (tag) a un contenuto mediale (testo, audio, video, foto).

[26] Amazon Mechanical Turk: www.mturk.com/mturk/welcome. Il nome si riferisce al celebre finto automa giocatore di scacchi di Wolfgang von Kempelen che nel XVII secolo ingannò le principali corti europee battendo a scacchi qualsiasi giocatore. Il trucco consisteva in un giocatore umano nascosto dentro l'automa.

(umanesimo digitale), i cui elementi teorici ed empirici, pur essendo fonte di ampi dibattiti, possono essere sintetizzati dalla seguente domanda: qual è il contributo che il computer e l'informatica possono dare all'avanzamento delle scienze umane?

Per certi versi la domanda è piuttosto semplice e – come spesso capita – la risposta non lo è. Per poter rispondere in maniera sistematica, è necessario distinguere due momenti che sono *humanities computing* (informatica umanistica) e le *digital humanities* (umanistica digitale). Per comodità, dato che la questione è fonte di ampio dibattito nella comunità scientifica italiana, useremo la terminologia internazionale usata in questo settore di ricerca.

L'humanities computing è un programma di ricerca il cui obiettivo è applicare le possibilità di calcolo e analisi rese possibile dal computer anche nell'ambito delle scienze umane, in particolare la letteratura (Gigliozzi 1997; Bozzi 2009; McCarthy 2003; Kirschenbaum 2010; Burdick *et al.* 2012; Perazzini 2013; Vanhoutte 2013; Vanhoutte *et al.* 2013a; 2013b).

Le humanities computing fanno riferimento al primo periodo della relazione fra informatica e scienze umane (dagli anni Quaranta agli anni Ottanta), periodo in cui la principale tecnologia di riferimento era il computer, sia nella versione *mainframe* (fino a gli anni Settanta) che nella versione personal computer (dagli anni Ottanta). Lo studioso che viene considerato come il fondatore di questo approccio è padre Roberto Busa, il quale, grazie anche all'aiuto tecnologico e finanziario di IBM nella figura del suo più famoso direttore – Thomas J. Watson – ha dato vita all'*Index Thomisticus*, ovvero lo studio delle concordanze linguistiche[27] nell'opera omnia di San Tommaso d'Aquino (Perrazzini 2013; Vanhoute 2013)[28]. Questo periodo è anche quello più importante per la creazione di una comunità di studiosi che si riconoscono nello studio delle scienze umane attraverso l'uso sistematico di strumenti computazionali. Nascono associazioni scientifiche (ALLC, Association for Literary and Linguistic Computing, Association for Computers and Humanities), riviste scientifiche (*Journal of Literary and Linguistic Computing*, *Computer and the Humanities*), monografie dedicate alla diffusione della disciplina (seminale quella di McCarty 2005), a cui fanno seguito i primi congressi internazionali (McCarthy 2003; Kirschenbaum 2010; Perazzini 2013; Vanhoutte 2013). Quasi tutti gli studiosi di questo primo periodo fanno risali-

[27] Tecnica di analisi che consiste nell'elenco delle parole di un testo ordinate alfabeticamente e contestualizzate per una migliore interpretazione (Gigliozzi 1997: 181-187).

[28] L'*Index Thomisticus* si presentava nella forma di una enorme collezione di schede perforate e ha attraversato tutte le principali tecnologie di archiviazione digitale. Dal 2005 è presente sul Web in versione consultabile sul sito www.corpusthomisticum.org/it.

re l'approccio delle humanities computing alle prime riflessioni su linguaggio, comunicazione e informatica facendo riferimento ad autori come Alan Turing (padre dell'informatica moderna) o Norbert Wiener (fondatore della cibernetica, la scienza della comunicazione e del controllo nell'uomo e nelle macchine). Le domande di ricerca di questa fase si concentrano essenzialmente su quella che possiamo considerare la componente algoritmica del testo che può essere definita come lo sviluppo degli strumenti software per l'analisi delle fonti (McCarty 2003). Come abbiamo già avuto modo di sottolineare per altri approcci, database e strumenti di analisi diventano centrali. Difatti questa fase può essere scomposta nei suoi due elementi chiave, quello relativo alla costruzione di database specializzati (elemento legato ai dati), e quello inerente alla costruzione degli strumenti di analisi (elemento legato ai tool) (Perazzini 2013). Questa fase sul finire degli anni Ottanta darà vita a una nuova componente, quella metalinguistica, il cui obiettivo è quello di sviluppare un sistema di marcatura (*tagging*) dei documenti digitali per scopi di analisi e di ricerca (McCarty 2003). Uno dei risultati più importanti di questa fase è il Text Encoding Iniziative (TEI), un consorzio che raccoglie diverse istituzioni internazionali per lo sviluppo di linee guida adatte alla codifica dei testi umanistici che ha proposto l'adozione di diverse tecnologie tra cui il SGML (Standard Generalized Markup Language) fino all'XML (eXtended Markup Language) (McCarty 2003; Vanhoutte 2013; Perazzini 2013). La terza componente delle humanities computing è la componente rappresentazionale, il cui obiettivo è la messa a punto di forme di manipolazione dei dati, centrate soprattutto sulla possibilità di visualizzare i testi analizzati in forme innovative, cosicché possano suggerire connessioni e linee di analisi non previste (McCarty 2003). Questa componente giungerà a maturità alla fine degli anni Ottanta, anche sulla spinta dell'emergere delle digital humanities.

Le digital humanities condividono con le humanities computing lo stesso programma di ricerca, con in più la componente legata alla diffusione di Internet e alla nascita del World Wide Web (Kirschenbaum 2010; Burdick *et al.* 2012; Perazzini 2013; Vanhoutte 2013). Lo slittamento di significato è attribuibile a diverse motivazioni. In primo luogo l'impatto sociale e culturale del Web che cambia sia le tecnologie a disposizione degli studiosi, sia la formazione di una base stabile di studiosi e ricercatori. Proprio rispetto a questo secondo punto c'è da dire che l'uso della rete in un'ottica di allargamento della comunità scientifica è avvenuto fin dalle origini da quando cioè nacque *Humanist* nel 1987, un forum online nato sulla rete BitNet dell'Università di Toronto dedicato alla discussione dei temi relativi al rap-

porto fra computer e scienze umanistiche (McCarty 2003)[29]. Oggi per esempio lo strumento principale di scambio di riflessioni online è sicuramente Twitter, emerso che canale di comunicazione privilegiato fra gli studiosi di digital humanities a partire dalla conferenza annuale di Filadelfia 2009 (Kirschenbaum 2010). L'emergere della nuova denominazione del settore non deve essere considerato un semplice restyling che segue la moda della rete, ma un cambiamento concettuale. Come l'ingresso del concetto di ipertesto. Per esempio fra le fonti di ispirazione indicati dagli studiosi come chiave per l'ampliamento dell'approccio e la nascita delle digital humanities ci sono sia i nomi chiave degli ideatori dell'ipertesto (McCarty 2003), da Vannevar Bush che immaginò una tecnologia elettromeccanica come Memex, a Ted Nelson che con il suo progetto Xanadu coniò il termine di ipertesto, fino a Douglas Engelbart, ideatore di un nuovo approccio al rapporto uomo-computer, di cui la testimonianza più evidente è l'invenzione del mouse[30]. Le digital humanities intese come ampliamento in chiave Web e Internet del programma di ricerca delle humanities computing nascono nel 2004 grazie al libro curato da Schreibman e collaboratori (Schreibman *et al.* 2004) che fa una rassegna piuttosto estesa e sistematica di che cosa fossero le digital humanities a partire dalla descrizione di progetti di ricerca, metodologie di analisi, problemi epistemologici, descrizione delle proprietà delle tecnologie informatiche a uso e consumo dei ricercatori in scienze umane e così via (Vanhoutte 2003)[31]. La dimensione computazionale – considerata troppo rigida – lasciava spazio alla componente digitale, aprendo così le porte a una concezione allargata di ricerca umanistica che comprende aree come i *game studies*, la letteratura *fandom*, le culture del *remix*, prima rubricate sotto altre discipline (*Internet studies*, *Web studies*) (Burdick *et al.* 2012). Inoltre la contaminazione fra digital humanities e studi sulla comunicazione online più vicini alle scienze sociali ha fatto si che alcuni temi chiave delle scienze sociali contemporanee siano entrati a far parte del bagaglio degli umanisti digitali. Per esempio il concetto di big data, anch'esso dovuto al sempre maggiore interesse che gli strumenti di ricerca – database, archivi, biblioteche digitali – hanno in questo settore (Burdick *et al.* 2012; Perazzini 2013). Bisogna dire che nel dibattito attuale sulle digital humanities, la questione della definizione di que-

[29] Il forum non è più attivo, ma le sue conversazioni sono state raccolte e sono disponibili a questo indirizzo Web: http://dhhumanist.org.

[30] Per una rassegna del processo sociale e culturale che trasformò il computer da strumento di calcolo a mezzo di comunicazione cfr. Bennato 2002: 69-79.

[31] Il libro è liberamente consultabile online all'indirizzo Web: www.digitalhumanities.org/companion.

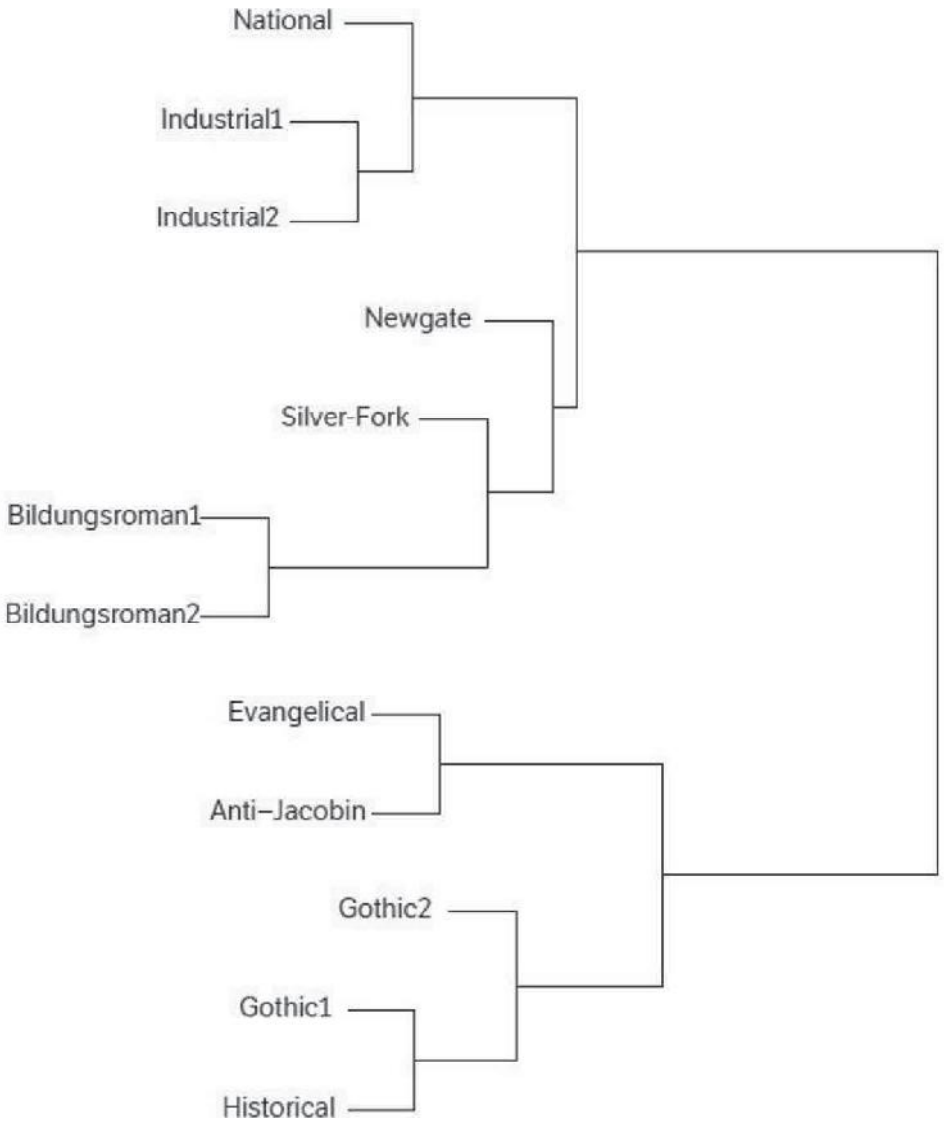
sto programma di ricerca – per quanto importante in un'ottica di demarcazione disciplinare – sta provocando una certa insofferenza nella comunità scientifica. Infatti molti ritengono che paper e libri scientifici che rispondono alla domanda "che cosa sono le digital humanities" siano un vero e proprio genere letterario (Kirschenbaum 2010), fino all'affermazione di alcuni studiosi: "se esistono due cose di cui l'accademia non ha bisogno, sono un altro libro su Darwin e un altro post sulla definizione delle digital humanities" (Gibbs 2011: 289).

A ogni modo, ciò che rende le digital humanities un programma di ricerca molto interessante in un approccio computazionale delle scienze sociali e umane è l'interesse verso il processo di sviluppo e uso dei modelli nel campo delle discipline umanistiche. In pratica se si volessero descrivere le due forme principali in cui c'è la collaborazione fra computer e scienze umane, potremmo distinguere fra la computazione per le scienze umane (impostazione strumentale) e la computazione nelle scienze umane (impostazione metodologica). Il processo di modellamento inteso dal punto di vista metodologico e di ricerca è stato particolarmente enfatizzato dagli umanisti informatici e digitali, tanto da essere considerato imprescindibile se si vuole questo approccio sia innovativo e porti a risultati interessanti (McCarty 2003; 2005). Un esempio emblematico in questo caso è il progetto dello Stanford Literary Lab (Lit Lab; http://litlab.stanford.edu) di Franco Moretti. Questo progetto ha come scopo un approccio guidato dai dati (*data driven*) nell'analisi del testo letterario. Moretti nel suo percorso di ricerca propone diverse innovazioni concettuali e metodologiche, tra cui una analisi del testo letterario attraverso una metodologia che lui chiama lettura a distanza (*distant reading*) in contrapposizione alla *close reading* (lettura ravvicinata) della critica letteraria di lingua inglese. Lo scopo è l'uso di diversi strumenti di scomposizione del testo – mappe, diagrammi, grafici – derivati da altre discipline – biologia, teoria dell'evoluzione, storia quantitativa – alla ricerca di modelli, schemi e *pattern* ricorrenti in grado di spiegare alcuni fenomeni letterari come la presenza di cicli narrativi, senza concentrarsi su un numero esiguo di testi, ma sul sistema letterario nel suo complesso, grazie all'aiuto di strumenti computazionali (Moretti 2005a; 2013). Un esempio emblematico in questo senso è la ricerca sul formalismo quantitativo (*quantitative formalism*) di Moretti e del suo gruppo (Allison *et al.* 2011).

In questo studio è stato possibile classificare automaticamente in vari sottogeneri letterari diversi testi, attraverso un software in grado di identificare le parole più usate in un romanzo (MFW: Most Frequent Words). Il software aveva "appreso" la struttura delle parole più ricorrenti in una serie di classici dei diversi generi letterari – da *Il monaco* di Matthew Gregory

Lewis per il genere gotico, *Ivanhoe* di Walter Scott per il genere storico, *Jane Eyre* di Charlotte Bronte per il romanzo di formazione e così via – e ha applicato quanto imparato su altri testi, riuscendo a classificare perfettamente, così come avrebbe fatto un critico letterario abituato alla lettura ravvicinata (fig. 2.8). Moretti non è nuovo a questo approccio che usa dati e rappresentazioni grafiche (Moretti 1997; 2005a; 2005b; 2013), negli ultimi tempi arricchito dall'approccio computazionale.

Fig. 2.8 – Dendrogramma dei generi letterari dei romanzi descritto usando la tecnica MFW



Fonte: Allison *et al.* (2011)

L'approccio che coniuga letteratura, strumenti digitali e utilizzo di database negli ultimi tempi si è unito con alcuni progetti legati ai big data, dando vita a una nuova linea di ricerca: la *culturomics*.

Il termine culturomics – italianizzato in culturomica dalla stampa italiana (Bazzi 2010; Pappalardo 2010; Longo 2011) – fa riferimento a un progetto di ricerca che mette insieme le digital humanities con il progetto di Google per la digitalizzazione dei libri (Google books[32]) e un gruppo interdisciplinare di matematici, biologi e linguisti dell'università di Harvard (Bohannon 2010; 2011). Il progetto consiste nella possibilità di studiare la curva di diffusione delle parole all'interno del corpus dei testi digitalizzati dal 1800 al 2000. Lo strumento messo a punto dal gruppo di ricerca che consente di studiare l'andamento nel tempo degli n-grammi, ovvero le stringhe di caratteri separati da uno spazio[33], prende il nome di Google n-gram[34]. La culturomics nelle intenzioni dei suoi ideatori – Jean Baptiste Michel ed Erez Aiden – è l'applicazione dei big data allo studio della cultura umana, ovvero attraverso l'interrogazione di enormi database relativi a libri, manoscritti, mappe, quotidiani e altri prodotti culturali è possibile studiare la cultura umana nella sua componente diacronica alla ricerca di pattern linguistici o altre strutture che potrebbero suggerire una interpretazione storica-culturale di diversi processi che usa le parole come unità di rilevazione (Michel *et al.* 2011; Aiden, Michel 2013), e dal punto di vista linguistico Google n-gram può essere descritto come un sofisticato strumento per la lessicologia computazionale (Perazzini 2012). Lo stesso articolo di Science con cui viene annunciato sia lo strumento che la disciplina fa degli esempi su come sia possibile fare analisi culturale a partire dall'analisi delle parole digitalizzate: l'evoluzione dei verbi irregolari, il cambiamento nell'uso di "grande guerra" sostituito da "prima guerra mondiale", l'assimilazione delle principali invenzioni tecnologiche del XIX e XX secolo, la curva di crescita della celebrità di diversi personaggi (Michel *et al.* 2011).

Un esempio interessante di analisi culturale attraverso l'andamento delle parole è quello della censura durante il Nazismo (fig. 2.9). Sulla base dei nomi di intellettuali indesiderati durante il Nazismo, è stato calcolato un indice di soppressione (*suppression index*)[35] per ogni persona della lista di esponenti dell'arte degenerata stilata dal bibliotecario del Reich Wolfgang Hermann e

[32] La pagina Web del progetto è: http://books.google.com.

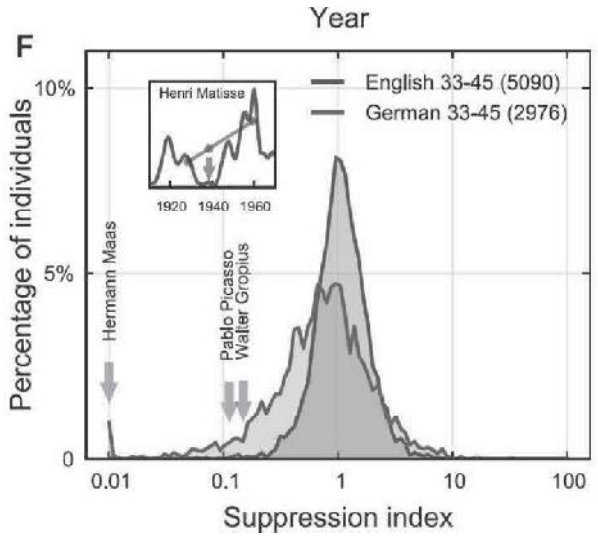
[33] Per esempio: banana, 1972 sono unigrammi (ovvero 1-grammi). Michael Jackson, Steven Spielberg sono digrammi (2-grammi). Jon Bon Jovi, Martin Luther King sono trigrammi (3-grammi) e così via.

[34] Il sito dove è possibile analizzare e visualizzare i dati è: https://books.google.com/ngrams.

[35] L'indice è calcolato dividendo la media delle citazioni ottenute dal personaggio in questione fra il 1933-1945 (periodo della Germania nazista) per la media della frequenza fra il 1925-1933 e la media fra il 1955-1965.

comparando la curva delle citazioni bibliografiche nel corpus di lingua inglese con la curva del corpus di lingua tedesca. Questa analisi mostra che il meccanismo di censura sviluppato dal Nazismo è stato sistematico. I dati mostrano che il 9,8% degli individui sono stati sottoposti a fortissima censura, come nel caso di Pablo Picasso e Walter Gropius, mentre 1,5% dei personaggi rimanenti mostrano nello stesso periodo una forte ascesa di visibilità: si tratta di sostenitori e fiancheggiatori del Nazismo (Michel *et al.* 2011).

Fig. 2.9 – Indice di soppressione degli intellettuali oppositori durante il periodo nazista



Fonte: Michel *et al.* (2011)

La culturomics è interessante in quanto frutto di una metodologia quantitativa derivata dalle tecniche matematiche della biologia evolutiva (Bohannon 2010) applicata a un dataset tipico dei big data, che archivia informazioni utili per lo studio della cultura (nel caso particolare libri) e quindi di enorme interesse per sociologi, letterati, storici. L'approccio fortemente quantitativo ma ricco di sfumature interpretative negli studi della cultura è tipico di altri orientamenti d'analisi come la *cultural analytics* (Manovich 2009). Ma tutti hanno lo stesso obiettivo: utilizzare dati e approcci computazionali sollevare il velo della storia sullo sviluppo della cultura.

2.3. Scienza sociale computazionale: modelli interdisciplinari per lo studio della complessità con un obiettivo sociale

Se il concetto di programma di ricerca ci ha aiutato a mettere in ordine alcuni dei più interessanti e promettenti approcci computazionali nelle