COLLANA PISIA

La governance dei dati pubblici Testi, contesti e politiche pubbliche

a cura di
Maria Stella Righettini e Stefano Sbalchiero

PADOVA UP
PADOVA UNIVERSITY PRESS

PISIA

Collana del Master in Innovazione,
progettazione e valutazione
delle politiche e dei servizi

Prima edizione 2022 Padova University Press

Titolo originale *La governance dei dati pubblici. Testi, contesti e politiche pubbliche*

Università degli Studi di Padova
via 8 Febbraio 2, Padova
www.padovauniversitypress.it

Progetto grafico: Padova University Press
Impaginazione: Padova University Press

In copertina: disegno di Giulia David

ISBN 978-88-6938-318-2

La governance dei dati pubblici. Testi, contesti e politiche pubbliche

Come usare i dati testuali a supporto della capacità di policy, della capacità amministrativa e della qualità dei servizi pubblici

a cura di

Maria Stella Righettini
Stefano Sbalchiero

PADOVA UP

Quando i dati sono i testi. Approcci e procedure per l'analisi dei dati testuali

Stefano Sbalchiero[1]

Text as data, Text mining, Analisi statistica dei testi.

Introduzione

«Text as data» (Gilardi et al. 2018) è una formula, più che un semplice espediente linguistico, oggigiorno entrata nel linguaggio della ricerca sociale e rappresenta, utilizzando un'affermazione che potrebbe sembrare perfino troppo esplicita, solamente un punto di arrivo di un percorso la cui caratteristica intrinseca è certamente una irrefrenabile pluralità, sia rispetto ai metodi e le tecniche sia per quanto riguarda settori scientifico -disciplinari differenti. E questo per almeno tre ragioni fondamentali. La prima si riferisce all'ambito metodologico in quanto, nonostante siano presenti approcci ben consolidati nel tempo, non solo i contesti di ricerca risultano mutevoli, ma con loro pure le tecniche e gli strumenti che gli studiosi mettono in campo al fine di estrarre informazioni dai testi per restituirle sotto una nuova forma. La seconda ragione rimanda, invece, agli ambiti disciplinari: chi si dovrebbe occupare dei testi come dati? Quali sono le discipline interessate? La risposta non può che essere, a sua volta, provvisoria. Molte sono le discipline che storicamente si sono interessate

[1] Sociologo, Dipartimento di Filosofia, Sociologia, Pedagogia applicata (FISPPA), Università degli Studi di Padova.

all'analisi dei testi, come ad esempio la linguistica, l'informatica, la statistica, per nominarne solamente alcune, e molte altre sono oggi le discipline che nutrono un certo interesse per le possibilità offerte dall'analisi dei dati testuali. La terza questione pertiene invece all'oggetto stesso della discussione: i discorsi, i testi, le parole. Se da un lato quanto detto pocanzi permette di comprendere i vantaggi, riconosciuti da più parti, di un approccio interdisciplinare al variegato universo di testi e parole che quotidianamente permeano la nostra realtà sociale, dall'altro lato l'utilizzo del termine 'discorso' non è casuale e improprio: l'etimologia stessa rimanda a 'scorrere', con il senso figurato di muoversi da una cosa all'altra. I discorsi sono, quindi, comunicazioni in movimento che non solo caratterizzano le nostre attività quotidiane e sociali, ma partecipano a pieno titolo alla produzione e condivisione di senso e significato che gli attori sociali attribuiscono a pratiche, siano esse soggettive o collettive. *Les Mots et les Choses*, direbbe Foucault (1966), Le parole e le cose. Se a questo si aggiunge la gran mole di informazioni con le quali ognuno di noi si interfaccia ogni giorno, dai social media ai flussi informativi, dalla carta stampata al web, dai *post* ai siti istituzionali, l'effettiva quantità di dati prodotti, ma anche utilizzati, o utilizzabili, è abnorme. Nell'epoca dei *big data*, altra formula difficile da definire in modo univoco, ulteriore elemento di complessità deriva dal rapido e costante cambiamento che investe la natura stessa dei rapporti tra media, *new media*, processi di digitalizzazione e modalità di fruizione e gestione dei contenuti. Entro tali processi, la partecipazione degli utenti ha innescato, infatti, l'irreversibile modifica sia nella produzione, attraverso anche la digitalizzazione, sia nell'interpretazione di testi. Per fare soltanto un esempio, basti pensare agli esiti della cultura definita *grassroot*, generata dall'utente: una stessa idea può essere espressa in una forma differente, attraverso la molteplicità e la divergenza degli strumenti e dei supporti disponibili, pur mantenendo i medesimi contenuti (Jenkins 2007). Che dire, poi, dei processi di digitalizzazione e dell'importanza che rivestono oggigiorno in ambiti quali l'amministrazione digitale, o *e-government*, entro cui assume particolare rilevanza il fenomeno della data governance, com'è stata definita nell'introduzione. Al di là delle problematiche relative alla gestione digitalizzata della pubblica amministrazione, tra i fenomeni che più interessano in questa sede vi sono, senza ombra di dubbio, tutte quelle azioni che rientrano a pieno titolo entro tali processi di cambiamento organizzativo e che in misura differente richiedono di mettere in campo strumenti e competenze diversificate. Saper organizzare e trattare la documentazione resa disponi-

bile, al fine di estrarre informazioni e, così facendo, ottimizzare non solo il lavoro del funzionario e degli enti preposti, ma offrire supporto allo sviluppo di politiche e servizi. Come già anticipato, se i dati in possesso delle pubbliche amministrazioni possono essere suddivisi, seppure con uno sforzo classificatorio, tra dati numerici – ad esempio il numero di delibere, di piani strategici o di sentenze e così via – e dati testuali, allora questi ultimi possono essere considerati strategici nella misura in cui vi sia la possibilità di accedere ai contenuti per comprenderne le implicazioni rispetto allo sviluppo di politiche e di supporto alle decisioni. Si pensi all'analisi di programmi, piani e delibere che a vari livelli, locali o sovralocali, molto ci dicono rispetto alle traiettorie progettuali intraprese, oppure all'analisi della documentazione inerente alla valutazione o agli strumenti di *accountability* con cui gli enti preposti comunicano verso l'esterno la propria identità rispondendo alle domande 'chi sono', 'cosa hanno fatto', 'cosa si accingono a fare'.

Difronte a scenari di questo tipo, l'assunto incontrovertibile dal quale muovere per una considerazione che faccia eco alla valorizzazione del dato testuale, accanto a quello numerico, è il seguente: a fronte della complessità, disponibilità e accessibilità al rapido flusso di parole che scorrono, è crescente l'esigenza di riuscire a intercettare e comprendere dinamiche veicolate attraverso i testi tramite la sintesi e la restituzione di dati testuali, avendo la possibilità di trovare dei riscontri oggettivi. Senza ombra di dubbio, questa considerazione pone alcune questioni fondamentali sia rispetto agli approcci e ai frame teorici, sia per quanto riguarda gli strumenti da adottare. Da un punto di vista dello scienziato sociale, infatti, il tipo di approccio teorico potrebbe essere parzialmente differente da quello del linguista, così come l'interesse dello statistico nei confronti dei testi può prevedere l'utilizzo di strumenti diversi da quelli considerati maggiormente utili dal politologo. Tuttavia, questo ragionamento non deve essere fuorviante: va ribadito, in linea con quanto anticipato, che l'approccio interdisciplinare non è solo auspicabile, ma fondamentale nel panorama teorico dell'analisi dei testi. In questa ottica, l'analisi dei testi moderna e supportata da software e tecniche sempre più raffinate vede linguisti, *computer scientist*, statistici, psicologi, politologi e sociologi attivare proficue collaborazioni al fine di sviluppare nuovi strumenti e percorsi di indagine in grado di rispondere alle specifiche quanto diverse esigenze di ricerca. È persino inutile sottolineare che questo dipende fortemente dalle domande che vengono poste, vale a dire dagli obiettivi dell'indagine. In questo caso, i due concetti di affidabilità, vale a dire la

capacità di produrre misurazioni tra di loro coerenti, e quello di validità, cioè il grado in cui uno strumento misura effettivamente ciò che si vuole misurare, risultano centrali nella valutazione aprioristica delle scelte sia di approcci sia di strumenti. Se per esempio l'intento del ricercatore fosse quello di studiare l'evoluzione del linguaggio amministrativo in una prospettiva storica, non risponderebbe al criterio di validità l'applicazione ai testi di un algoritmo di *stemming*, che riduce alla sua forma radice la parola, dato che verrebbero perse preziose informazioni e peculiarità linguistiche. Dall'altro lato lo *stemming* potrebbe produrre risultati più consistenti e validi volendo analizzare i temi maggiormente affrontati dalle pubbliche amministrazioni, anche in ottica comparativa, attraverso documenti programmatici relativi all'economia circolare o al cambiamento climatico. Allo stesso modo, per lo studio delle rappresentazioni di determinate categorie di attori sociali nella documentazione prodotta da un ente locale, potrebbe essere fondamentale ricorrere a specifici strumenti di analisi linguistica, come il *tagging* grammaticale, per lo studio delle forme verbali ad esse associate. In questo caso, se l'utilizzo del verbo 'supportare' riferito a una categoria specifica di beneficiari risultasse sottoutilizzato rispetto a 'contrastare' evidentemente le conclusioni sarebbero differenti, così come 'riconoscere' e 'disapprovare' avrebbero un peso diverso in seno al dibattito riferito alle politiche di risparmio energetico. In quest'ottica, l'utilizzo del dato testuale, accanto al dato numerico, può costituire un terreno di incontro privilegiato tra differenti discipline e dar luogo ad ambiti di sperimentazione continua, in grado di stimolare gli strumenti che abbiamo a disposizione rispetto alle crescenti sollecitazioni provenienti dalla realtà sociale.

Al di là di questi esempi, molte sono le problematiche che caratterizzano la discussione sui metodi e le applicazioni per analizzare materiale testuale o estrarre informazioni dai testi, come si vedrà di seguito. Per il momento basterà dire che quando facciamo riferimento all'analisi dei dati testuali questa non può che costituirsi come interdisciplina o, detto in altri termini, più che rappresentare un'isola essa non può che configurarsi come una terra di mezzo tra saperi fortemente interrelati.

Nei paragrafi seguenti, a partire da questa riflessione, il tentativo sarà quello di collocare l'analisi dei testi, secondo diversi orientamenti, proponendo alcuni indirizzi di indagine che interessano le scienze umane e sociali e che hanno trovato una diretta applicazione nei singoli lavori che verranno presentati a seguire.

Contesti e dibattiti: Text Mining, *Digital Methods* e *Big Data*

Quello che oggi conosciamo con il nome di text mining è il risultato, senza ombra di dubbio, della crescente bibliografia e del moltiplicarsi di studi soprattutto in ambito linguistico che hanno incrociato l'informatica e lo sviluppo tecnologico (Bolasco 2005, Sbalchiero 2018). Fortemente connotata dalla collaborazione tra ambiti disciplinari differenti, tale espressione può essere considerata come un settore entro il quale la linguistica computazionale, la matematica e la statistica, così come le scienze sociali, informatiche e *computer science* hanno creato un terreno fertile alle collaborazioni e agli sviluppi di tecniche e metodi. Non può tra l'altro sfuggire l'assonanza tra text mining e *data mining*: si tratta, infatti, di quell'insieme di metodi e tecniche il cui fine è la produzione di conoscenza attraverso l'estrazione di informazioni da grandi quantità di dati tramite metodi automatici. Gli esiti delle ricerche sono sotto gli occhi di tutti e hanno trovato espressione sia nella ricerca di base sia nella ricerca applicata e industriale: il riconoscimento vocale, la traduzione automatica, i sistemi informativi così come la gestione della conoscenza per mezzo delle tecnologie dell'informazione da parte di imprese, il così detto *knowledge management* (Giuliano et al. 2008). Ed è proprio in questo contesto che le tecniche di text mining si sono moltiplicate nel tentativo, in primo luogo, di affrontare i problemi connessi all'analisi e alla gestione di elevate quantità di testi liberi, ovvero non strutturati, con l'obiettivo di estrarre e ricavare da queste collezioni di documenti, in secondo luogo, informazioni utili per organizzare e alimentare database di grandi dimensioni. Il successo di queste tecniche, oltre all'ambito meramente industriale e aziendale, è dovuto alla possibilità di applicarle a qualsiasi tipo di testo non strutturato, ovvero con una struttura che non può essere identificata, come ad esempio le pagine web, agenzie stampa, archivi online e così via, stimolando l'interesse di molte discipline (Sanger et al. 2007). Fra le tappe che hanno segnato lo sviluppo di tale ambito, possiamo annoverare gli studi pionieristici di tipo lessico-testuale (Luhn 1959) e focalizzati sull'indicizzazione. Da questo punto di vista, l'*information retrieval* e *l'information extraction*, rispettivamente il recupero di documenti tramite *query* di ricerca e l'assegnazione di categorie ai documenti per facilitarne la consultazione, hanno non solo permesso di gestire grandi corpora testuali, ma gettato le basi per lo sviluppo di tecniche sempre più sofisticate (Bolasco 2005). L'attenzione è stata anche posta maggiormente sulla possibilità di una vera e propria descrizione dei testi attraverso il linguaggio matematico-statistico, o ancora attraverso innovative strategie per

recuperare le informazioni attraverso, ad esempio, il *clustering* gerarchico automatico dei documenti (Jardine et al. 1971). Non solo, ma attraverso l'utilizzo di dizionari elettronici e lessici di frequenza, ricorrendo quindi anche a meta-informazioni linguistiche, alcuni studi pionieristici diedero vita alla linguistica computazionale (Busa 1974-1980), espressione che ha poi portato a riflessioni sulla disciplina e a coniare il termine informatica umanistica (Orlandi et al. 2003), utilizzato nella lingua italiana anche come traduzione di *digital humanities*:

> «l'informatica umanistica costituisce il punto di contatto tra scienze umane e scienze esatte: ragionando sui caratteri comuni delle diverse discipline umanistiche e formalizzando le procedure necessarie per condurre la ricerca nei diversi ambiti, propone l'integrazione dei due mondi superando la semplice applicazione di tecnologie avanzate a settori delle scienze» (Celentano et al. 2004, p. 44).

Ed è proprio a partire dai numerosi studi che si sono succeduti a partire dagli anni '70 e '80, assieme agli sviluppi dell'informatica, che si assiste al moltiplicarsi di tentativi e allo sviluppo di algoritmi nell'ambito, soprattutto, dell'intelligenza artificiale e del *machine learning* applicati ai dati testuali e alla ricerca di tipo linguistico (Porter 1980, Berger et al. 1996). Il resto è storia recente, e il passo dal *data mining* al text mining è risultato breve. In sostanza, non può certamente sorprendere che l'interrogativo rispetto a che cosa si intenda oggi con *digital humanities* ci dice molto rispetto all'evoluzione di una disciplina piuttosto recente, il cui sviluppo è stato accompagnato dalla rapida moltiplicazione di gruppi di ricerca, conferenze, pubblicazioni, riviste e così via, ma che contemporaneamente necessita ancora di una definizione largamente condivisa (Gold 2012). Questo per dire che la sua definizione, in modo simile ad altre discipline, non può rappresentare uno stadio dello sviluppo e maturazione della disciplina stessa (Arthur et al. 2014). Infatti, all'interno di tale dibattito, è emersa anche, sotto forma di critica, la necessità di una riflessione maggiormente profonda capace di cogliere i motivi per cui, a fronte della sempre maggior diffusione e utilizzo di strumenti digitali e risorse di text mining, la difficoltà di collegamento tra l'umanistica digitale e ricerca *mainstream* in ambito umanistico risulta tutt'oggi evidente. In questo senso, Patrik Juola (2008) parla di un vero e proprio abbandono, percepito dai ricercatori che si occupano di *digital humanities*, da parte della più generale comunità umanistica. Questo sarebbe dovuto, da un lato, alla constatazione che tale comunità nel suo complesso non è sempre a conoscenza degli strumenti sviluppati dai professionisti in metodi digitali e,

dall'altro lato, perché il *mainstream* umanista tenderebbe a non prendere sul serio molti dei risultati delle ricerche prodotte e ottenute con metodi e strumenti propri delle *digital humanities* (*ivi*, p.73). Senza voler entrare nei dettagli di tale dibattito, ma volendo fornire una possibile interpretazione, si potrebbe anche aggiungere, in continuità con quanto detto, che il problema della definizione di tale ambito di ricerca è assimilabile al problema della demarcazione dei confini tra le discipline scientifiche (Gieryn 1983, 1995). Una demarcazione non tanto verticale, vale a dire tra ciò che può essere definito scientifico o meno, quanto invece orizzontale tra ambiti disciplinari differenti e/o tra specializzazioni entro la stessa disciplina. Dovrebbe risultare chiaro, seguendo questo ragionamento, che nel caso delle *digital humanities* il paradosso tra la crescente diffusione di questo termine ombrello, il suo inconfutabile successo nel panorama odierno, fino alla denuncia di mancato riconoscimento, potrebbe in realtà essere visto come un punto di forza, trattandosi di un orientamento fortemente interdisciplinare, capace di generare nuove prospettive di analisi anche senza catturare l'attenzione della ricerca *mainstream* (Prescott 2012). Questo perché, nella ricerca sociale odierna, i ricercatori devono confrontarsi con quesiti epistemologici e metodologici da un lato rispetto ai fenomeni, a volte inediti, che necessitano ancora di essere definiti e, dall'altro lato, che riguardano i presupposti stessi che stanno alla base dello sviluppo di nuove pratiche di ricerca attraverso strumenti a loro volta nuovi o innovativi. A questo proposito, un'ulteriore nota a margine rispetto a quanto accennato può essere utile per collocare il text mining entro i complessi scenari della ricerca odierna. Il tema *Big Data*, che secondo il modello delle tre V (Laney 2001), sono caratterizzati da alta varietà (fortemente eterogenei), velocità (acquisiti ed elaborati in tempo reale) e volume (ordini di grandezza in termini di misurazione), necessita di un constante confronto interdisciplinare dal momento che i database raccolgono tipi di dati che possono essere strutturati o non strutturati e quindi offrire nuovi modi di fare scienza sociale (Roberts et al. 2016). In effetti, quando si parla di text mining tale concetto si accompagna di sovente a quello di *big data* (Delmastro et al. 2019). Questo perché, come già anticipato, le procedure di text mining riguardano principalmente l'analisi automatizzata di grandi quantità di testo, espressione questa che viene associata al concetto di *big data*, a volte in modo fuorviante: con esso ci si riferisce a vaste raccolte di dati che per le loro caratteristiche e dimensioni non possono essere archiviate e analizzate con strumenti convenzionali e tecniche informatiche standard. I *big data*, dunque, possono

certo fare riferimento a vari tipi di dati, che comprendono database di grandi dimensioni, ma anche *file* audio e video, così come grandi quantità di testo, per lo più non strutturato. Ma tale espressione può essere correttamente utilizzata quando l'insieme di questi dati e la loro definizione, siano essi testuali o di altra natura, supera una soglia che non solo non viene definita in modo unanime in letteratura, ma soprattutto risulta talmente complessa e di dimensioni tali da richiedere nuove tecniche e strumenti per estrarre, elaborare, gestire e restituire le informazioni in una nuova forma (De Mauro 2019). Tale constatazione porta sicuramente a una considerazione di questo tipo: nella ricerca sociale e nell'ambito della ricerca umanistica, ciò significa mettere a punto strategie di ricerca, nel nostro caso applicate ai testi, facendo ricorso a procedure e software in grado di estrarre informazioni e renderle disponibili all'interpretazione con approcci adeguati rispetto alla dimensione del materiale empirico oggetto di indagine. Di più, possiamo affermare che il nodo cruciale è relativo alla capacità dei ricercatori di promuovere una costante unione tra scienze umane e tecnologiche, e tra le discipline umanistiche stesse, e che il successo di tali approcci negli anni a venire dipenderà notevolmente sia dalla diffusione della pratica scientifica in esse svolta, sia nella capacità di rappresentare e interpretare la realtà sociale, vale a dire selezionare ed includere metodi, orientamenti e caratteristiche conoscitive utili (in quel momento) per raggiungere dei fini pragmatici, legittimando così l'autorità culturale e le pretese di conoscenza in esse rivendicate.

Approcci, metodi e tecniche

Tra i numerosi approcci per l'analisi dei testi possono essere distinti diversi orientamenti, i quali, per caratteristiche tecniche e per obiettivi conoscitivi, risultano utili ai fini della presente ricostruzione. Un primo orientamento è, a livello generale, collocabile entro la tradizione dell'analisi del contenuto nella versione moderna, l'analisi automatica del contenuto, supportata da software, in cui uno degli obiettivi è l'identificazione degli argomenti principali presenti in una collezione di testi. Altri approcci, invece, non si pongono tanto l'obiettivo di fare analisi del contenuto, come viene tradizionalmente intesa, quanto invece di analizzare i testi secondo procedure e tecniche utili, tra le altre cose, alla classificazione dei testi e all'analisi di documenti a seconda della 'distanza' esistente tra di loro. Tra tutte le possibilità offerte dall'analisi testuale e dal text mining, gli approcci e gli strumenti presentati, come si vedrà, sono risultati particolarmente utili per esplorare grandi quantità di documenti, estrarre le

informazioni principali, analizzare i principali contenuti o classificare in modo automatico i testi, anche in assenza di un sistema di classificazione preesistente.

Prima di addentraci nei principali orientamenti e dettagli tecnici relativi ai software e alle procedure, può risultare utile chiarire alcuni termini che verranno utilizzati in seguito. Quando parliamo di *corpora* facciamo riferimento a una collezione di testi:

> «il materiale testuale oggetto delle analisi prende il nome di corpus e si configura come una collezione di testi. Il corpus raccoglie testi coerenti con gli scopi perseguiti dalla ricerca e questa coerenza è valutabile solo discrezionalmente. Nello studio dell'intera opera di un autore i testi costituenti il corpus possono essere, per esempio, le singole opere inedite e/o inedite di cui si conosce l'esistenza; nello studio di un romanzo i singoli capitoli; nell'analisi dei risultati di un'indagine con intervista a domande aperte le trascrizioni dei colloqui [...] nell'analisi di annate di stampa i quotidiani (o i settimanali o mensili ecc.) pubblicati» (Tuzzi 2003, p. 29).

Il *vocabolario*, uno dei principali strumenti prodotti attraverso i software che vedremo, si presenta come una lista ordinata, generalmente per frequenza decrescente, di *parole*, ovvero *forme grafiche*, alle quali è associata la frequenza con cui si presentano nel corpus. Una forma grafica altro non è che una sequenza di caratteri appartenenti all'alfabeto, delimitata da due separatori, come lo spazio (il *blank*) e i segni di interpunzione. La forma grafica, dunque, viene utilizzata per identificare quella che nel linguaggio comune viene definita parola. A sua volta la forma grafica può essere definita in due differenti modi: come *word type*, quando facciamo riferimento alla lista di parole diverse presenti nel vocabolario; come *word token*, invece, quando intendiamo le occorrenze presenti nel vocabolario.

L'analisi automatica del contenuto

Il primo approccio cui possiamo fare riferimento è la *topic detection* nella versione che possiamo annoverare nel panorama degli strumenti e dei metodi automatici per l'analisi del contenuto (Sbalchiero 2018). Anche se l'analisi del contenuto, da un punto di vista storico, non è certamente recente (Losito 1993, Tuzzi 2003), basti pensare ai lavori di Lasswell sulla *Propaganda Technique in the World War* (Lasswell 1927) e sull'analisi quantitativa del linguaggio politico (Lasswell 1949), o al lavoro di Thomas e Znaniecki riguardante l'analisi delle lettere nel celebre *Il contadino po-*

lacco in Europa e in America (Thomas et al. 1918-1920), per citarne soltanto alcuni, il percorso di affinamento dell'approccio ha conosciuto fasi alterne (Krippendorff 1983). Interessante, da questo punto di vista, sono le numerose considerazioni e le riflessioni metodologiche che si sono susseguite negli anni rispetto ai modi di fare ricerca, ben interpretate da Sorokin (1956), che utilizzò il termine 'quantofrenia' facendo emergere un problema rilevante anche dal punto di vista del metodo: l'eccessiva rigidità imposta da un percorso di ricerca che sotto l'egida dell'oggettività scientifica finiva per trascurare, se non perdere totalmente, la ricchezza qualitativa della ricerca. Accanto a questo, lo sviluppo tecnologico, dei computer e dei *software* portò ad accrescere l'interesse verso l'analisi dei contenuti: se da un lato si potevano gestire molti più dati, è soprattutto nel campo degli studi linguistici che emersero approcci non soltanto orientati al mero aspetto quantitativo, e quindi orientamenti di tipo lessicale (Tuzzi 2003). In particolare, in seno alla scuola francese (Beaudouin 2016), e grazie ai lavori e alle innovazioni metodologiche introdotte da Jean P. Benzécri (1982), lo sviluppo dell'approccio lessico-testuale superava da un lato i limiti dell'analisi delle sole frequenze e si orientava, dall'altro lato, all'analisi delle relazioni tra più variabili adottando una prospettiva multidimensionale, come nell'analisi delle corrispondenze lessicali (Benzécri 1992). Tutte queste premesse ci conducono al nocciolo della questione: la necessità di gestire grandi quantità di testi, attraverso analisi del contenuto maggiormente complesse, e quindi la necessità di una sintesi tra aspetti quantitativi e approfondimenti qualitativi al fine dello sfruttamento statistico dei dati testuali (Lebart et al. 1988, Bolasco 1999).

Ed è in questa cornice che si situa l'algoritmo utilizzato per i lavori che seguiranno. Si tratta del così detto metodo Reinert (1983, 1990), implementato nel software Alceste *(Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte)* e, più di recente, disponibile nella versione *R-based* di Iramuteq *(Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires),* (Ratinaud 2014).

Per quanto riguarda i fini del presente contributo, è utile sottolineare che il metodo Reinert risulta di particolare interesse considerando l'analisi del contenuto che vede la collaborazione tra linguisti, *computer scientist*, statistici, psicologi, politologi e sociologi nello sviluppo e utilizzo di nuovi strumenti e percorsi di indagine in grado di rispondere alle specifiche quanto diverse esigenze di ricerca. In particolare, i concetti di affidabilità (la capacità di produrre misurazioni tra di loro coerenti e indipendenti dal ricercatore), e quello di validità (la capacità di uno stru-

mento di misurare effettivamente ciò che ci si propone di misurare), sono risultati centrali nel tentativo di coniugare, nello stesso strumento, rigore scientifico e approfondimenti qualitativi nell'analisi automatica dei contenuti. Detto in modo differente, l'algoritmo ideato da Reinert si pone l'obiettivo di integrare quantità e qualità attraverso un percorso di ricerca valido ed efficacie, facendo ricorso, da un lato, all'approccio moderno lessico-testuale, ma, dall'altro lato, perseguendo fini e obiettivi conoscitivi che valorizzano anche l'aspetto qualitativo.

A livello operativo, la procedura consta di diverse fasi successive così come vengono implementate dal software Iramuteq (Ratinaud 2014, Sbalchiero 2018). Il corpus testuale, composto da diversi testi, viene ridotto in porzioni di testo che coincidono con una frase, un enunciato o un paragrafo delimitato da punteggiatura, e che prendono il nome di *Elementary Context Units* (ECU). In questo modo, il corpus viene organizzato in modo automatico dal software in porzioni di testo di lunghezza simile. Ad un passo successivo, l'algoritmo identifica le co-occorrenze delle parole in ogni ECU attraverso la costruzione di una matrice di contingenza *parole x ECU* (Tab. 1)

Tab. 1 Esempio di una tabella *(parole x ECU)* e relativo dendrogramma.

	Parola 1	*Parola 2*	*Parola 3*	*…*	*Parola N*	
ECU 2	1	0	1	0	0	classe 1
ECU 4	1	0	1	0	1	
ECU 1	0	1	0	1	1	classe 2
ECU 3	1	1	0	1	1	
ECU 5	0	1	0	1	1	classe 3
ECU 6	0	1	0	1	1	
ECU 7	0	1	0	1	1	

Tale matrice, organizzata come tabella presenza *vs* assenza, dove "0" corrisponde all'assenza mentre "1" alla presenza della parola nella porzione di testo, viene costruita come base per analizzare la similarità delle ECU, che viene identificata tramite una procedura di *clustering*. Tale procedura rileva, in modo gerarchico e attraverso la distanza del χ^2 tra le classi, quelli che vengono definiti mondi lessicali (Reinert 1983) o classi semantiche (Ratinaud et al. 2012, 2015, Smyrnaios et al. 2017). Il dendrogramma, come risultato principale, viene costruito a passi successivi: all'inizio dell'analisi vengono create due classi, ognuna delle quali raggruppa le ECU che rispecchiano un contenuto lessicale simile, che con-

dividono cioè parole e che, di converso, si differenziano maggiormente l'una dall'altra, cercando in questo modo di ridurre al minimo le parole in comune. Si procede quindi con ulteriori ripartizioni delle ECU per classi fino a quando risultano sufficientemente omogenee da non poter essere ulteriormente disaggregate. I risultati ottenuti sono interessanti per diverse ragioni. Da un lato, la classe semantica (Reinert 1993), interpretabile come una variabile latente (Reinert 1998, pp. 292-293), è caratterizzata da uno specifico vocabolario di parole tra di loro associate, che co-occorrono nelle stesse porzioni di testo e che, dall'altro lato, rendono particolarmente efficacie e intuitivo il processo di interpretazione delle classi, ovvero dei principali argomenti – *topics* – presenti nel corpus analizzato. Nella fase di interpretazione, inoltre, per facilitare il compito dei ricercatori, tramite l'utilizzo dei valori di associazione del χ^2 delle parole con la classe, il software permette di estrarre le porzioni di testo (ECU) maggiormente significative per dar conto di quella classe semantica. In questo senso, l'algoritmo riporta, per ognuna delle porzioni di testo classificate dal metodo, un valore che corrisponde alla media dei valori del χ^2 delle parole significative e che sono associate a quella classe. Di conseguenza, la significatività di una porzione di testo per dar conto di uno specifico argomento è tanto maggiore quanto più contiene un numero di parole che risultano significative per quella classe e quindi in grado di rappresentarla in termini di contenuto. Infatti, l'identificazione dei principali argomenti e quindi dei contenuti è l'obiettivo principale dell'algoritmo, che identificando i mondi lessicali partecipa all'individuazione e alla costruzione di vocabolari di parole co-occorrenti che li caratterizzano. Detto in modo differente, l'esito della classificazione delle porzioni di testo (ECU) in un insieme di classi che includono parole rilevanti per la classe indica che tanto più le ECU sono simili, quanto più condividono parole. Va altresì specificato che l'elenco delle parole più significative e che meglio rappresentano un mondo lessicale – il vocabolario dei *topics* – viene identificato tramite l'associazione del χ^2 tra le parole e le classi. In questo senso, la soglia di significatività del χ^2 viene identificata tramite il *p-value*: le parole che presentano un *p-value* < 0.0001 saranno quindi considerate più significative rispetto a quelle che hanno un valore superiore alla soglia *p-value* $\geq 0{,}05$.

Come si può notare nell'esempio (Tab. 2), la procedura di *clustering* evidenzia gruppi tematici omogenei e che presentano contenuti simili interpretabili osservando le forme significativamente associate ad ogni classe.

Tab. 2 Esempio di vocabolari delle classi semantiche. Le parole vengono ordinate per valore di associazione decrescente (χ^2) e con una significatività del *p-value* < 0.0001.

Classe 1	χ^2	**Classe 2**	χ^2	**Classe 3**	χ^2
ambiente	253,609	energia	206,603	Mobilità	113,497
territorio	134,013	sostenibilità	149,65	Urbana	105,728
Turismo	117,431	rinnovabili	124,805	Impatto	93,7
promozione	96,304	risorse	114,543	Sociale	75,603
:	:	:	:	:	:
politiche	34,247	alternative	46,763	Scelte	45,272
ambientali	30,867	futuro	41,353	Policy	42,827

Infine, i risultati dell'analisi del contenuto prodotti dall'algoritmo possono essere utilizzati per valutare il grado di associazione delle classi semantiche con le modalità di altre variabili per rispondere, ad esempio, alla domanda 'chi dice che cosa' (incrociando i *topics* ottenuti con le testate giornalistiche, in caso di analisi di articoli, oppure con i paesi, in caso di analisi comparativa di documenti istituzionali), oppure ancora con la variabile anno, qualora sia necessario verificare la presenza di determinati argomenti in una prospettiva longitudinale. Anche in questo caso, le relazioni delle classi semantiche individuate con le altre variabili sono basate sui valori di associazione del χ^2 che misurano in che modo tali variabili sono associate con le classi semantiche individuate.

Infine, tra le varie opportunità offerte dal software, vi è anche l'analisi delle corrispondenze, uno strumento classico per l'analisi dei dati testuali, utilizzata in diversi lavori che seguiranno. L'analisi delle corrispondenze (Greenacre 2007, Lebart et al. 1984, Tuzzi 2018) è particolare caso di *Principal Component Analysis* (PCA) applicata a una tabella di contingenza che incrocia le parole (righe) per le modalità della variabile (nelle colonne). Tale tecnica ha l'obiettivo di trasformare le frequenze delle parole in coordinate su un sistema di assi cartesiani multidimensionali: in questo modo, è possibile visualizzare modalità di variabili e parole traducendo il concetto di similarità in una specifica distanza euclidea (Murtagh 2005) e quindi in piani cartesiani (Sbalchiero et al. 2016). L'analisi delle corrispondenze applicata ai testi, dunque, si basa sui profili lessicali e mira a mettere in evidenza le relazioni tra modalità delle variabili (come, ad esempio, gli anni, oppure i documenti prodotti da diverse istituzioni e così via) e le parole. Le posizioni delle modalità delle variabili sui piani sono quindi determinate dal grado di somiglianza dei profili lessicali: ne consegue che quelle presenti nello stesso quadrante del piano rispecchia-

no contenuti simili in termini di parole ed è quindi possibile interpretare le posizioni reciproche (modalità delle variabili e parole) e ricostruire i principali contenuti entro una mappa.

Text mining: *topic detection*

Se l'estrazione dei *topics* con il metodo Reinert rientra propriamente nel panorama dei metodi per l'analisi automatica dei contenuti, strumento molto utile, come si è visto, al fine dell'analisi statistica dei testi, molti sono i recenti progressi nella tecnologia hardware e software che hanno incentivato lo sviluppo e la diffusione di algoritmi di text mining (Aggarwal et al. 2012). In particolare, ai fini del presente contributo, tra i numerosi approcci e algoritmi di *topic modelling* (Gilardi et al. 2021), sviluppati nell'ambito dell'informatica e del *Machine Learning* e che godono di un grande successo nelle applicazioni di text mining, troviamo la *Latent Dirichelet Allocation* (LDA), utilizzata anche in alcuni lavori a seguire. Il successo di modelli probabilistici è basato sul fatto che non solo possono essere potenzialmente applicati a qualsiasi collezione di testi, anche di vasta dimensione, ma che possono estrarre argomenti latenti secondo una logica diversa da quella che abbiamo descritto pocanzi. Nello specifico, l'LDA è un modello generativo probabilistico, presentato in uno studio pubblicato da David Blei e colleghi (2003), a cui sono seguite altre varianti basate su di esso, come il *dynamic topic model* (Blei et al. 2006), il *correlated topic models* (Blei et al. 2007) o la *structural topic model* (Roberts et al. 2014). Tali sviluppi rappresentano modelli adattati a esigenze specifiche, come ad esempio la necessità di mettere in relazione i *topics* rilevati con altre variabili, ma in ogni caso l'LDA rimane una pietra miliare nel contesto dei *topic models*. L'assunto dell'LDA è che ogni testo che compone un corpus può essere rappresentato da un insieme di *topics* latenti, onde per cui un documento viene considerato come una combinazione probabilistica di questi argomenti, ognuno dei quali è caratterizzato da una specifica distribuzione di parole (Griffiths et al. 2004). Si tratta, in questo caso, di un modello generativo in quanto, non potendo osservare direttamente i *topics*, vengono utilizzati i dati a disposizione, ovvero le parole, per ricostruire, a posteriori, la struttura latente del corpus analizzando i testi e le parole che lo compongono. In altre parole, questo processo, probabilistico e generativo, evidenzia l'interazione tra i documenti osservati e i *topics* latenti al fine di analizzare la probabilità che un documento contenga informazioni su un argomento sulla base della distribuzione delle parole contenute nel testo. Se, quindi, ogni testo è costituito da un numero T di

topics ($j = 1..T$), che a loro volta sono caratterizzati da parole specifiche, ognuno di questi *topics* può essere rappresentato come una distribuzione di probabilità sul vocabolario. Di conseguenza,

> «se abbiamo T topics, possiamo calcolare la probabilità della parola i-esima in un dato documento come:
>
> $$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$
>
> dove z_i è la variabile latente che indica il topic da cui è stata tratta la i-esima parola e $P(w_i|z_i=j)$ è la probabilità della parola rispetto al j-esimo topic. $P(z_i=j)$ fornisce la probabilità di scegliere una parola dai topics j nel documento, che varierà tra i diversi documenti. Intuitivamente, $P(w|z)$ indica quali parole sono importanti per un topic, mentre $P(z)$ è la prevalenza di quei topics all'interno di un documento» (Griffiths et al. 2004, p. 5228, trad. nostra).

Possiamo dire che i risultati principali ottenuti con questo approccio sono da un lato liste di parole associate ai *topics* e, dall'altro lato, quelli che risultano essere i documenti più significativi per i *topics* rilevati. L'LDA può essere applicata a un corpus di testi attraverso l'utilizzo del '*topicmodels*' *package* (Grün et al. 2011) disponibile in R (R development core team 2016), che permette di implementare l'algoritmo proposto da Blei in ambiente *open source*. Per ulteriori approfondimenti si rimanda alla bibliografia (Blei et al. 2009), anche se un ulteriore aspetto vale la pena di essere sottolineato. Come si è visto, trattandosi di un modello generativo, l'LDA ricostruisce (ovvero genera) i documenti del corpus assegnando il peso probabilistico di ogni *topic* ai documenti e, per ogni *topic*, la distribuzione delle parole, con la possibilità di evidenziare quelle con il livello di probabilità più elevato, vale a dire maggiormente rilevanti, per lo stesso *topic*. Ora, rispetto all'applicazione dell'algoritmo, è stato dimostrato che la *topic detection* funziona abbastanza bene con testi brevi, come ad esempio *abstract* di articoli, *post* sui *social media*, articoli di giornale, articoli scientifici, articoli di Wikipedia, per citarne alcuni, e il motivo è che intuitivamente forniscono informazioni concise sui contenuti principali. Cosa accade quando applichiamo la *topic detection* a testi lunghi? Di fatto, quando l'analisi viene applicata a testi lunghi, sorgono alcuni problemi (Michel et al. 2011). In particolare, è difficile dedurre la prevalenza di un *topic* che sia coerente con un lungo testo, ad esempio un libro, perché di solito contiene una gamma di argomenti diversi. Non solo, ma il problema dell'adattamento del modello all'analisi è cruciale

perché l'algoritmo LDA richiede che il numero dei *topics* sia specificato a priori: inutile dire che tale parametro influisce sui risultati dell'analisi. A questo proposito, un recente contributo (Sbalchiero et al. 2020) si inserisce in questo dibattito attraverso una serie di esperimenti che applicano l'LDA ai testi lunghi, cercando di gettare nuova luce sulla complessa relazione tra la lunghezza dei testi e la determinazione del parametro rispetto al miglior numero di *topics* da rilevare in un corpus. L'esito del lavoro sottolinea come la lunghezza delle porzioni di testo è una variabile per spiegare il cambiamento nella determinazione del miglior numero di *topics* da rilevare in un corpus. Di conseguenza, seguendo i risultati proposti, viene formulata la regola Sbalchiero-Eder, sotto forma di modello matematico (*ivi*, p. 1103), che mette in relazione il numero ottimale di *topics* da estrarre in un corpus e la determinazione della dimensione delle porzioni di testo utile all'organizzazione del corpus: dato un corpus, il numero migliore di *topics*, come parametro necessario all'implementazione dell'LDA, è inversamente proporzionale alla lunghezza delle porzioni di testo, ovvero più grandi sono le porzioni di testo, minore è il numero di *topics* da rilevare.

La classificazione e il concetto di distanza

Particolarmente interessante risulta il concetto di distanza quando applicato allo studio di collezioni di testi. Utilizzata nell'ambito della stilometria, ovvero l'analisi statistica dello stile linguistico e letterario, tra le varie applicazioni possibili molto diffuso è il suo utilizzo nell'ambito dell'attribuzione d'autore (Tuzzi et al. 2018). Il presupposto è che le caratteristiche di ogni autore, ovvero il suo stile, sarebbero rintracciabili e quantificabili attraverso lo studio dei testi al fine di distinguere, ad esempio, un autore da un altro (Holmes 1998, Joula 2006). Se, ad esempio, l'obiettivo principale di un'analisi fosse quello della classificazione dei testi tramite procedure di *clustering* (Berry 2004), ovvero un tipo specifico di classificazione di documenti finalizzata a raggruppare in *cluster* testi simili, separati da quelli dissimili a formare gruppi distinti, occorre una distanza (Sbalchiero et al. 2016).

Il concetto di distanza può essere inteso, quindi, nei termini di una misura di somiglianza in grado di valutare fino a che punto ogni testo può essere considerato simile o diverso da un altro.

A livello intuitivo, la distanza intertestuale, ovvero tra due testi, si basa sulla seguente assunzione: se due testi sono identici, tutte le parole compaiono nei due testi con la stessa frequenza e la distanza risulta pari a zero. Si considerino, ad esempio (Tab. 3), i seguenti due testi: testo A *(cambiare le carte in tavola)*, testo B *(la classe non è acqua*).

Tab. 3 Esempio di distanza tra due testi A e B.

Forma	testo A	testo B	$f_{i,A} - f_{i,B}$
Acqua	0	1	1
cambiare	1	0	1
Carte	1	0	1
classe	0	1	1
è	0	1	1
in	1	0	1
La	0	1	1
Le	1	0	1
non	0	1	1
Tavola	1	0	1
	5	5	10

$$d(A,B) = \frac{10}{5+5} = \frac{10}{10} = 1$$

Ne consegue che la distanza raggiunge il massimo teorico 1 quando due testi non hanno nessuna parola in comune, ovvero presentano una distanza massima. Viceversa, se due testi sono identici, tutte le parole compaiono nei due testi con la stessa frequenza e la distanza, in questo caso, risulta pari a 0. Per chiarire quanto detto, può risultare utile un richiamo alla distanza intertestuale di Labbé (Labbé et al. 2001, Labbé 2007, Tuzzi 2010, Cortelazzo et al. 2013), che tra le numerose misure di distanza disponibili (Rudman 1998, Trevisani et al. 2020), risulta di assoluto interesse.

Nello specifico, l'elenco delle parole, e il numero di occorrenze corrispondenti, rispecchia quello che può essere definito il profilo lessicale di ciascun documento. La distanza di Labbé è basata su una somma di differenze tra le frequenze delle parole presenti nei testi. In altri termini,

> «data una coppia di testi A e B di dimensione N_A e N_B con $N_A \leq N_B$, la loro distanza d è:

$$d(A,B) = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f^*_{i,B} \right|}{2N_A}$$

dove $V_{A\cup B}$ rappresenta il vocabolario di A e B e la frequenza $f_{i,B}$ di ogni parola i nel testo più grande B viene ridotta in base alla dimensione del documento più breve A per mezzo di una semplice proporzione

$$f^*_{i,B} = f_{i,B} \frac{N_A}{N_B}$$

La distanza tra A e B è uguale alla distanza tra B e A, cioè la distanza è simmetrica, e la distanza tra ogni testo e se stesso è pari a zero e, più in generale, se due testi contengono le stesse parole con la stessa frequenza, la loro distanza è zero. Se due testi non hanno parole in comune, sono separate da una distanza pari a uno» (Sbalchiero et al. 2016, p. 1339).

La traduzione del concetto di distanza entro la ricerca empirica può risultare di estremo di interesse, come mostrano alcuni lavori che seguiranno, per esempio, per l'analisi dei profili lessicali su corpora di documenti di programmazione o testi istituzionali, ottenendo così la distanza intertestuale tra i documenti analizzati al fine di verificare direttrici comuni tra i testi e il grado di recepimento di direttive più generali. Tra i software utilizzati per applicare questo tipo di analisi va segnalato il pacchetto 'stylo', *stylometry* (Eder et al. 2016), che risulta interessante perché combina analisi diverse e numerose misure di distanza in ambiente R (R development core team 2016).

Considerazioni a margine

Per ricomporre gli innumerevoli argomenti affrontati entro un quadro unitario è opportuno chiarire alcune questioni a margine che, per quanto provvisorie, sono necessarie nel tentativo di rispondere al seguente quesito: quale tipo di conoscenza è possibile ottenere attraverso ricerche che utilizzano metodi di analisi statistica dei dati testuali e tecniche di text mining? La risposta non può che essere di questo tipo: dipende dalla domanda. Si tratta di una delle questioni che attraversa l'insieme dell'esperienza del lavoro scientifico, dalla raccolta del materiale empirico alla restituzione dei risultati, passando dalla formulazione del percorso di indagine. Quanto detto è di fondamentale importanza per riuscire a comprendere i lavori che seguiranno e, soprattutto, per distinguere un percorso di ricerca finalizzato alla *descrizione* da un orientamento votato alla *comprensione* dei fenomeni (Sbalchiero 2021). La risposta a questa domanda, in effetti, segna la differenza tra descrivere, nel senso di riassumere un insieme di testi o, di converso, operare una rigorosa analisi di questi testi

attraverso la costruzione di evidenze empiriche ed elementi interpretativi in grado di cogliere quel fenomeno. Un'affermazione di questo tipo potrebbe sembrare banale, ma non lo è affatto: implica il costante misurarsi con i significati che orientano la ricerca e costituisce uno spartiacque tra il mero gioco intellettuale e l'attività scientifica. È chiaro che formulare delle strategie entro un percorso di ricerca significa operare delle scelte e rappresentarsi, fin dall'inizio, il percorso stesso che, a sua volta, dovrà essere coerente agli obiettivi conoscitivi che ci si prefigge di raggiungere. In tal senso,

> «un progetto di ricerca che possa definirsi rigoroso, e la relativa riflessione metodologica, devono saper coniugare un orientamento strategico alle tecniche, alle analisi e alle eventuali verifiche empiriche che il ricercatore ritiene più adatti. In tal senso non esiste il metodo migliore o la tecnica migliore, ma esistono scelte più o meno pertinenti, sulla base della loro appropriatezza e coerenza rispetto al problema selezionato» (Sbalchiero 2021, p. 108).

Un ulteriore aspetto è degno di nota: i lavori che seguiranno si collocano entro esperienze di ricerca attraverso l'uso di strumenti che cambiano a seconda delle finalità perseguite. Questo punto è estremamente rilevante perché, va ribadito con forza, l'utilizzo di metodi e tecniche non può prescindere da una riflessione sul disegno della ricerca e, quindi, sulla domanda di ricerca che coinvolge specifiche scelte pratiche e operative. Formulare una domanda di ricerca rispetto a un fenomeno significa chiedersi non soltanto come possa essere affrontato e definito, ma anche dove ci debba portare la ricerca che stiamo conducendo. Questo significa, allo stesso tempo, mantenere aperta la possibilità di apportare delle modifiche qualora determinate scelte presuppongano una riconfigurazione stessa del processo di ricerca oppure, come di sovente capita, quando si intravede la possibilità di operare ulteriori approfondimenti coniugando diversi approcci. Ad esempio, l'esito della *topic detection*, ottenuta con l'LDA, potrebbe essere utilizzata per approfondimenti tematici di determinati *topics* tramite il metodo Reinert, avendo a disposizione una classificazione di testi organizzati per argomenti. In questo caso, l'esito sarà un focus specifico su un argomento che verrà ulteriormente approfondito attraverso l'estrazione e l'analisi dei mondi lessicali caratterizzanti quel *topic*. Seguendo questo ragionamento, le possibilità che si aprono sono innumerevoli. Una buona regola rimane, comunque, quella di descrivere le scelte fatte durante il percorso di ricerca in modo tale che gli strumenti e gli approcci utilizzati risultino complementari, e non alternativi, anche al lettore. In

questo senso, di nuovo, è sempre necessario valutare caso per caso, sulla base di quali sono gli obiettivi perseguiti, qual è l'approccio più adeguato, nella consapevolezza che l'analisi automatica di testi non dovrebbe essere vista come un'alternativa ai più tradizionali approcci di tipo qualitativo, piuttosto li integra, mettendo a disposizione una vasta gamma di strumenti software e statistici che continuamente offrono nuove possibilità di ricerca. A fronte degli sviluppi nell'analisi statistica dei testi e del text mining, assieme alla crescente accessibilità a grandi collezioni di materiale empirico digitale, che ampliano notevolmente le opportunità di ricerca, si può affermare che i diversi percorsi costituiscono un proficuo terreno di incontro tra ricercatori di diversa estrazione. Infatti, anche quando il materiale impiegato nella ricerca ben si adatta all'elaborazione statistica, occorre tenere a mente che necessita di conoscenze, competenze e intuizioni che nella fase di interpretazione dei risultati rimangono, per ora, di stretta competenza umana.