IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received June 6, 2019, accepted June 19, 2019, date of publication June 28, 2019, date of current version July 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925706

Visualizing Literature Review Theme Evolution on Timeline Maps: Comparison Across Disciplines

CHUNLEI TANG[1], (Senior Member, IEEE), YANGYONG ZHU[2], SUZANNE V. BLACKLEY[1], JOSEPH M. PLASEK[1], MEIHAN WAN[2], LI ZHOU[1], JING MA[3], AND DAVID W. BATES[1]

[1]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02120, USA
[2]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai 201203, China
[3]Department of Population Medicine, Harvard Medical School, Boston, MA 02125, USA

Corresponding author: Jing Ma (jing_ma@harvardpilgrim.org)

This work was supported in part by the National Natural Science Foundation of China under Project U1636207, and in part by the Shanghai Science and Technology Development Fund under Grant 16JC1400801, Grant 17511105502, and Grant 17511101702.

ABSTRACT Data-driven visualization techniques can be utilized to enhance the literature review process across different disciplines. In our work, 910 articles were retrieved using keyword search from bibliographic databases of two different disciplines (computer science: DBLP and medicine: MEDLINE) between 2001 and 2016. These articles' titles were processed using dynamic latent Dirichlet allocation to generate a set of themes/topics, which were subsequently classified and assigned to regions in a spatiotemporal geographical map. Resulting data visualizations from both repositories were manually reviewed by independent annotators. The results from the DBLP and MEDLINE were comparable and, taken together, suggest potential benefits of increased future interaction amongst multidisciplinary fields. Our findings indicate that spiral timeline maps have the potential to help researchers acquire or compare knowledge efficiently without prior domain knowledge.

INDEX TERMS Review, information storage and retrieval, research, medical informatics, epidemiology, MEDLINE, data visualization.

I. INTRODUCTION

Biomedical literature reviews allow investigators to quickly and easily familiarize themselves with a new area of study. For example, using literature data to track past trends in public health informatics can lead to a better understanding of epidemiological factors and circumstances. However, current methods for conducting literature reviews primarily employ manual review for the full range of required tasks (including content analytics, Medical Subject Headings (MeSH) indexing, paper selection, and information retrieval) making these reviews labor-intensive to perform.

Multiple visualization tools [1]–[5] already exist for creating timelines (or storylines) of literature to help new investigators rapidly acquire knowledge of a topic in a field. A disadvantage of many storyline mining approaches is their inability to clearly display complex narratives. A timeline makes it possible to construct a dynamic chronological tracking model of literary narratives in an evolving field. However, existing timelines are mostly constructed as static straight lines, potentially making it difficult to present several themes and theme words simultaneously in chronological order.

We previously developed a supervised method (called the regional classifier [6]) to visualize research theme words under different themes in a timeline map to show chronological development of the focused research topics. When combined with a geographical map, timelines can depict temporal patterns of events with respect to their spatial attributes. The classifier partitions theme words pretreated by semantic analysis tools into a real map such that words appearing in the same time period appear in the same region. The resulting visual displays were similar when starting with a specific research subject as when starting with a well-known scholar of that subject, suggesting that both strategies are equally viable starting points for researchers new to a particular field.

In this study, we apply an optimized version of the classifier to investigate survey- and questionnaire-related studies, which are widely used in multidisciplinary research including public health, medicine, and computer science. Thus, a literature review on questionnaire-related studies is an interdisciplinary study. The relevant studies in the field of computer science were collected from the Computer Science Bibliography (DBLP). DBLP can be accessed at https://dblp.uni-trier.de. A parallel comparison was conducted on PubMed/MEDLINE. The aim of this study was to elicit a comparison of research and technology development across

The associate editor coordinating the review of this manuscript and approving it for publication was Walter Didimo.

the two disciplines (i.e., medicine and computer science) specifically related to surveys and questionnaires.

II. MATERIALS AND METHODS

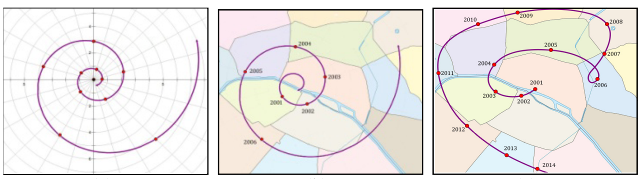
This section describes the regional classifier (Section II.A) and presents an enhanced literature review process incorporating the regional classifier as well as the regional classifiers validation (Section II.B).

A. THE REGIONAL CLASSIFIER

We propose utilizing timeline maps combined with dynamic Latent Dirichlet Allocation (LDA) [7] and the direct grid method [8] to produce a regional classifier [6].

A literature review is a process of exploring one or more research themes that appeared in publications of a specific area. A research theme (e.g., technology development and application scenario) can be expressed as a small set of words attributable to this theme. The evolution of each theme may be expressed visually in the relationships (or patterns) among relationships and patterns among publications over time via a timeline. We define these relationships as follows.

Definition 1 (Timeline Map [5], [6]): A timeline map T is a pair (G, Π), where $G = (V, e)$ is a directed graph and Π is a set of paths (research paths) to be built along the timeline in G.



(a) (b) (c)

FIGURE 1. logarithmic spiral timeline (a), examples of a logarithmic spiral's segments overlaid onto the Paris municipal map (b), and a different ordering of red points learned from a customized lemniscate shape (c).

In our previous work, we used a logarithmic spiral $\gamma = a \cdot e^{b\theta}$ as the timeline (**Figure 1a**), where the radius γ is a monotonic continuous function of the angle θ, and a, b are arbitrary positive real constants. It is easy to cut equal points on a logarithmic spiral timeline based on the same angle measure in degrees (e.g., 75°). Thus, the classification problem is converted since each aliquot point is relaxed to an area (**Figure 1b**).

Definition 2 (Regional Classifier [6]): A regional classifier is a classifier which assigns theme words in the same time period to their corresponding regions in a spiral geographical map.

We envision that researchers may have their own preferences regarding the shape of connectivity of research paths. To meet this need, we can reassign time periods to different regions on the map learned from different shape definition. For example, **Figure 1c** which uses the lemniscate (i.e., an infinity symbol ∞) displays in a different chronological order compared to the spiral.

B. STEPS IN THE LITERATURE REVIEW PROCESS

We adapted the four steps (literature search, classifier use, annotation studies, and literature review; see the Figure 5) of developing a data-driven literature review for questionnaire-based studies in the disciplines of computer science and medicine using DBLP and MEDLINE.

1) LITERATURE SEARCH

We conducted a semi-structured search of DBLP and MEDLINE using grouped search terms designed for maximal retrieval of relevant studies. DBLP's search used the terms "survey" and "questionnaire data" to identify a total of 437 records from DBLP. MEDLINE's search used three MeSH terms ("questionnaires," "epidemiology," and "epidemiologic methods") with filter conditions set to exclude articles related to clinical trials and reviews (as neither type is likely to describe novel methods). The results were limited to articles written in English and published between 2001 and 2016. After excluding eight duplicate records, 429 and 481 paper titles were included in this study in the DBLP and MEDLINE datasets, respectively.

2) CLASSIFIER USE

While the regional classifier can be applied to any part of a scientific paper (e.g., titles, abstracts, methods, results, or the whole article), we use article titles for simplifying evaluation. We assumed each title conveyed the major themes of the reported study. The LDA part of our classifier retrieves themes from article titles expressed as a bag of words. We also incorporated external words to improve coherence of LDA by building a Markov Random Field on the latent topic layer of LDA to encourage words labeled as similar to share the same topic label. For example, given a set of titles related to the word "association," the sense of this word can be ambiguous, referring either to an organizational relationship or an analytical method. The target meaning can be disambiguated by considering nearby words; for example, "association" accompanied by words such as "rule" or "analysis" would likely imply an analytical method. The regional classifier generated a fixed number of theme words for each dataset. These words were partitioned into regions on the map corresponding to the year they first appeared. We then connected words according to their themes to form research paths. To track the words related to a specific theme, we considered only a word's first appearance to track the introduction of new technologies; using high-frequency words is another possible (and valid) choice that may be useful in tracking the hype of new technologies.

We conducted a baseline comparison of the merits/demerits of the different classifiers on the task of analyzing the evolution of ideas. For this analysis, we utilized support vector machines (SVM). To measure the accuracy of SVM classifiers, our target is the years labeled on all the theme words produced by the dynamic LDA model on the DBLP dataset. The task involved the following steps: 1) constructing a "the number of theme words $\times$ year" matrix with the row that labels one of words and column that labels each year from 2001 to 2016; 2) filling the matrix by calculating the number of occurrences of each keyword among all the keywords corresponding to a particular year (utilized as our training set); 3) using Pearson correlation to calculate each pair of words' similarity to extend the "the number of theme words × year" matrix into a "the number of theme words × the number of theme words" matrix; and 4) applying principal component analysis (PCA) decomposition to decrease the dimensionality into a "the number of theme words × 2" matrix with the two columns as the x-axis and y-axis, such that each keyword can be presented as a coordinate on a map. We assume that the coordinates of the entire map are in a certain range (e.g., [0, 250]), and put all the coordinates of the map into the test set. We implemented our SVM models in sci-kit learn using the Linearsvc and support vector classification (SVC) function to see the effect of different kernels (e.g., linear, polynomial, and radial basis function (RBF)) on the classification.

3) ANNOTATION STUDIES

Evaluating the classifier's results poses some significant challenges. For example, a map of the findings is more than just a set of lines; there is information in its structure as well. We used human annotators' judgments to assess the effectiveness of our classifier. Each map was manually reviewed by at least two annotators. There were three annotators with computer science backgrounds. They did not confer with each other or with the authors. To facilitate examination of inter-rater reliability, we asked annotators to evaluate the results using a list of guidelines that consider both the correctness and the generalizability of the classifier [6]: (1) a good research clue should contain as many research paths as possible; (2) each path should contain only one theme; (3) the information in one path should come from as many different articles as possible; (4) partial overlap is not allowed, as a keyword cannot belong to two paths; and (5) generalization is allowed, as keywords (and their linguistic variants) may describe the same person, fact, or event. The three annotators reviewed all article titles in each dataset to judge whether the maps covered the important concepts of each research topic. Our annotators then performed two tasks. First, they manually summarized differences between the maps produced by the classifier. Second, if a theme word (or the associated theme word) occurred in both datasets, the annotators labeled it with the year it first appeared in the other data set.

4) LITERATURE REVIEW

Full-text articles highlighted by the classifier's output were included in the review. These reviews were conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement. Two pairs of authors independently identified original articles based on the classifier's results. We assessed three sources of bias: performance bias, detection bias, and reporting bias, along with the use of intention-to-treat analysis. Discrepancies were rechecked by a third author and consensus achieved by discussion. Two pairs of authors independently identified a total of 60 full-text articles (55 from DBLP and 5 from MEDLINE) reporting technology development and medical applications for review based on classifier suggested articles corresponding to the oldest article within a new theme. Each pair of authors screened 30 articles.

III. RESULTS

A. ANNOTATION RESULTS FOR THE REGIONAL CLASSIFIER

Figure 2 shows a total of five themes generated from 429 article titles extracted from DBLP using the terms "survey" and "questionnaire data" between 2001 and 2016. Figure 2 has a total of 150 theme words. We categorized the five themes into two groups based on interactions amongst these themes: technology development (RED LINE, KHAKI LINE, and GREEN LINE) and application scenario (PURPLE LINE and BLUE LINE).

First, the **RED LINE** is the main theme of technology development in computer science, with 55 theme words. The RED LINE can be observed transitioning from issues of digitizing questionnaire items (e.g., answer (2001), Likert (2007)) for ordinary computing devices, and entire questionnaires for new electronic devices (e.g., touchscreen (2004), iPad (2011)), to questionnaire data analysis.

Second, the **KHAKI LINE** (27 theme words) evolves from traditional data analysis (e.g., analysis (2002), statistical (2003)) to data mining techniques (e.g., clustering (2008), neural network (2011)).

Third, the **GREEN LINE** (29 theme words) emphasizes issues related to survey studies in computer science. 41.4% (n = 12) theme words regarding properties of computation (e.g., reflections (2001), credibility (2004), security (2007)) appeared only in DBLP.

Fourth, the **PURPLE LINE** (33 theme words) indicates application scenarios unrelated to medical research, from which two theme words (i.e., physicians (2005) and medicine (2005)) are excluded according to the annotators' definition of this theme. These demonstrate that the application of technology to questionnaire-related studies has gained popularity across general science domains. These applications include the social sciences, education, production, employee or people satisfaction, and so on.

Finally, the **BLUE LINE** (34 theme words) diverged from the PURPLE LINE in 2005 with two theme words, "physicians" and "medicine." The BLUE LINE indicates application scenarios in medicine and clinical research, which have experienced increased growth rates beginning in 2013.

Figure 3 and **Figure 4** provide a parallel comparison of DBLP and MEDLINE and were generated from 481 article titles. Figure 3 is based on all 43 theme words of Figure 2. Open dots indicate theme words from MEDLINE labeled with the year of their first appearance. For words that appeared in MEDLINE earlier than in DBLP, the year of their first MEDLINE appearance is shown in parentheses. In Figure 4, the annotators labeled in square brackets a total of 17 identical or similar theme words (among a total of 108) found in MEDLINE that also appear in DBLP,

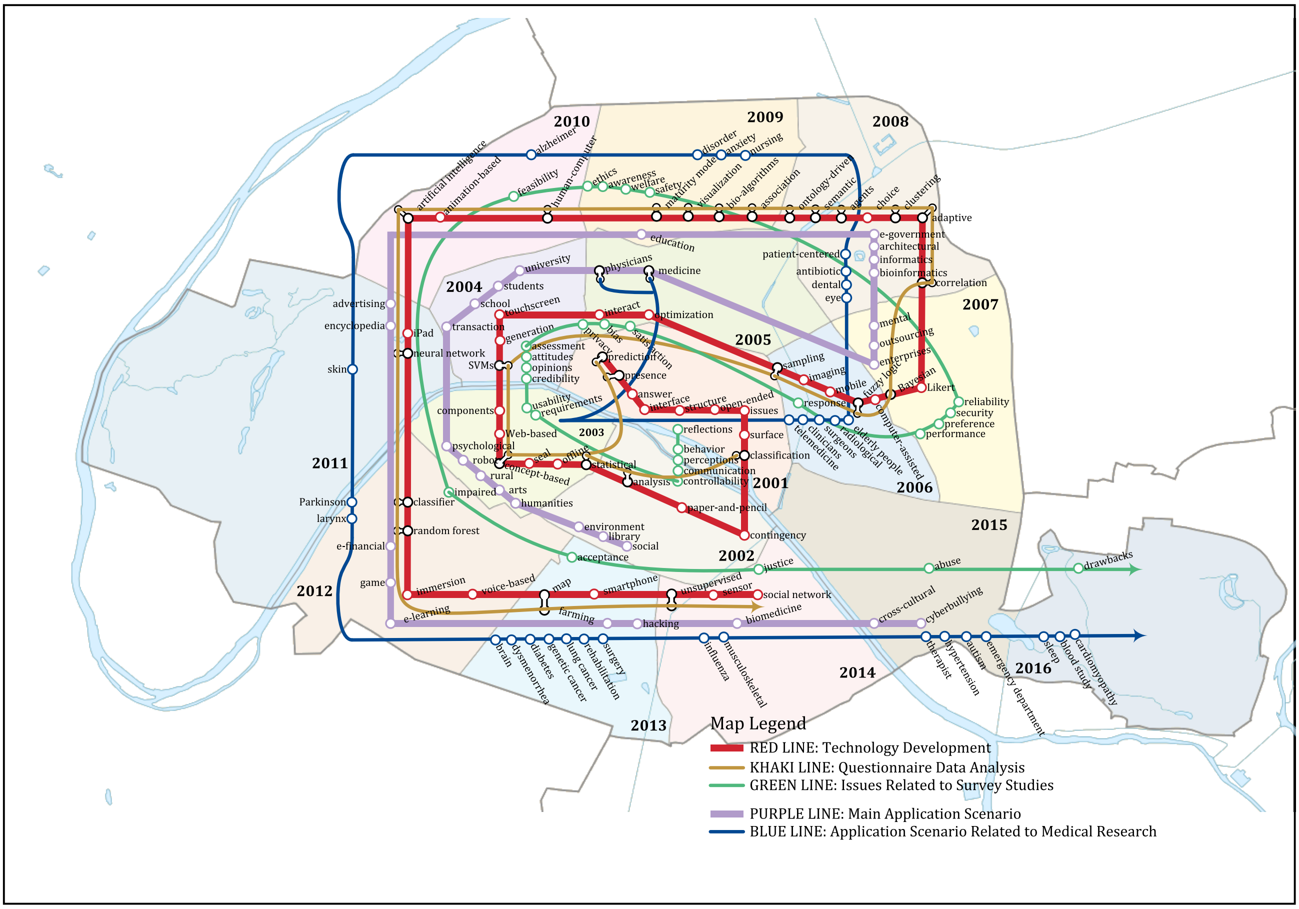


FIGURE 2. Chronological tracking of five themes generated among 429 article titles extracted from the DBLP database using the terms "Questionnaire Data" and "Survey" between 2001 and 2016.

and further defined four themes: 1) application scenario in epidemiology (BLUE LINE) with 108 words, 2) medical focus (**Subtheme 1**) with 67 theme words, 3) lifestyle research (**Subtheme 2**) with 40 theme words, and 4) drug research (**Subtheme 3**) with seven theme words.

B. BASELINE CLASSIFIER COMPARISON

In sum, we collected 150 keywords on the DBLP dataset with the year in which they first appeared corresponding to five themes (see Figure 2) into a training set that was used as input into SVMs for baseline comparison. Figure 6 shows the best classification results for the 4 SVMs. Specifically, each region shown on the map (with random colors) denotes an aggregation of keywords with the same predicted year. We observed that 1) the keywords had a lot of overlap, 2) the contour lines of different classes of the predicted year was not apparent, 3) the distribution of Figure 6d (i.e., SVC with the polynomial kernel) is a diffuse annular extension from the center, which confirmed the reliability of the regional classifier.

C. SUMMARY OF EVIDENCE FROM THE CLASSIFIER

For technology development, Figure 2 exhibits two phases: Phase 1 (RED LINE excluding KHAKI LINE) is focused on methodologies related to survey method digitization and data storage. Studies in phase 1 were mainly published between 2001 and 2006, although one third of theme words (nine out of 27, or 33.3%) can be found after 2007. To some extent, these nine theme words were related to electronic equipment released in 2010 or later. For example, the earliest version of Apple's iPad was released in 2010; the same theme word "iPad" appears in DBLP in 2011. Phase 2 (KHAKI LINE) studies published between 2007 and 2016 were mainly focused on data mining technologies. Nearly half of the phase 2's theme words (12 out of 27, or 44.4%) are concentrated between 2007 and 2009.

As shown in Figure 3, only about half of the theme words (or their synonyms/linguistic variants) (43 out of 83, or 51.8%) found in DBLP also appear in MEDLINE. Of these, 11 words (RED LINE excluding KHAKI LINE in 2001-2006) and 14 words (KHAKI LINE in 2007-2016) were related to Phase 1 and Phase 2, respectively. Our annotators' comments on these words show that the majority (17 out of 25, or 68.0%) appeared first in a later year in MEDLINE compared to DBLP. For example, the theme word "web-based" first appeared in 2003 [9] in DBLP, seven years before its first appearance (2010) in MEDLINE [10], despite the existence of some related words (i.e., Internet mail (2001) and Web survey (2004)). Another example is "classification," which appeared in 2001 in DBLP but did not appear in

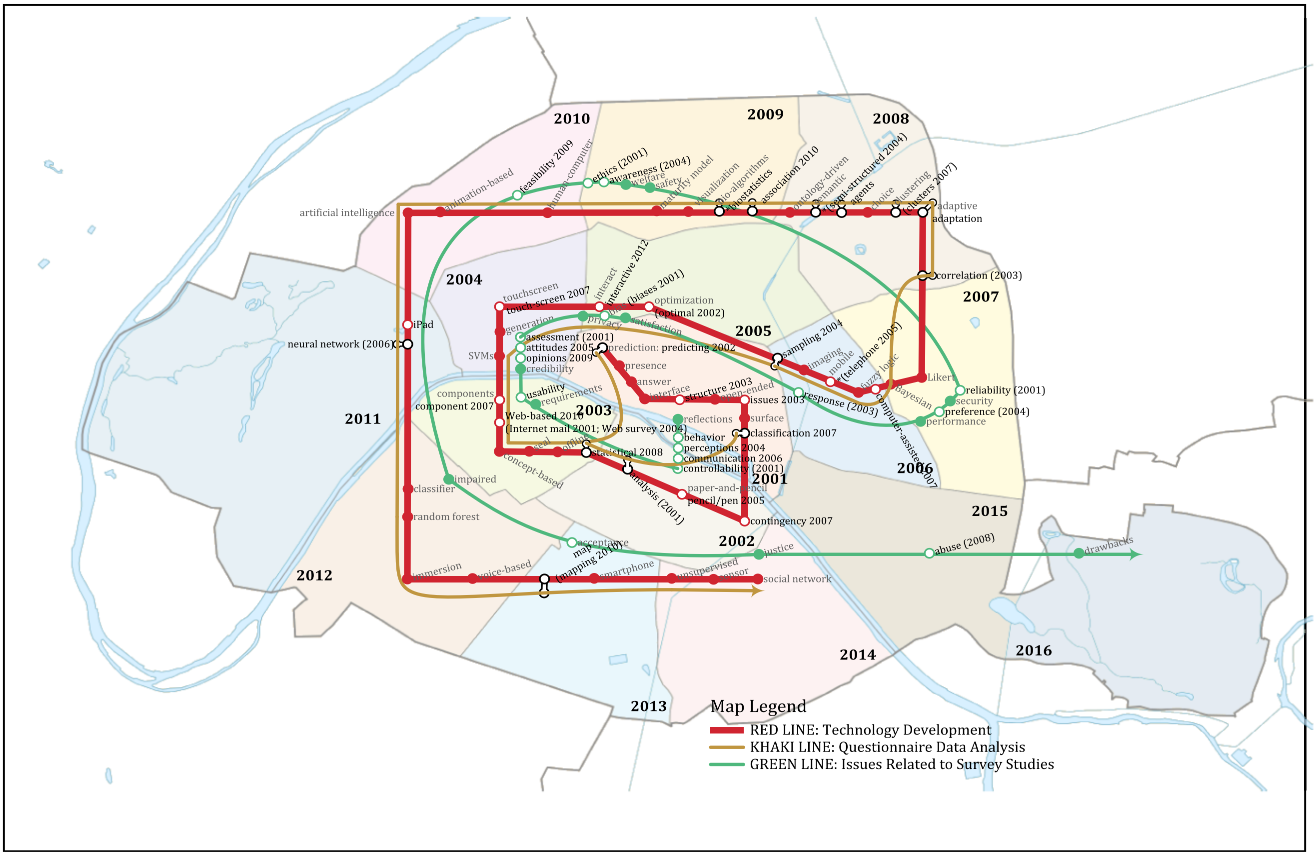


FIGURE 3. Chronological tracking of three themes generated among 481 article titles extracted from the MEDLINE database. Open dots indicate theme words from MEDLINE labeled with the year of their first appearance. Words that first appeared in MEDLINE earlier than in the DBLP database are shown in parentheses.

MEDLINE until 2007. Thus, Figure 3 reveals that medical research has been developing behind computer science by about five to eight years in terms of these technological themes.

For the application scenario, Figure 2 (PURPLE LINE) revealed that computer science technology developments were generally about five to eight years ahead of their applications in medical research. More specifically (KHAKI LINE compared to BLUE LINE), theme words pertaining to data mining techniques are highly concentrated in 2008 and 2009; however, a similar composition of applications in medical research did not manifest until 2013.

The findings from Figure 2 compared to Figure 3 & 4 show that computer science and medicine have entirely different emphases concerning questionnaire-related studies. Computer science focuses on the properties of computation (see GREEN in Figure 2, e.g., reflections (2001), security (2007), safety (2009)), whereas medical research tries to understand the nature of people (see GREEN LINE in Figure 3, e.g., ethics (2001), behavior (2002), attitudes (2005)).

There are other examples (see BLUE LINE in Figure 2 & 4) indicating differences in technological solutions between the two disciplines. The first is the first appearance of "eye," which occurred in MEDLINE in 2002 [11] and in DBLP in 2008 [12]. MEDLINE's "eye" involved questionnaire development, while DBLP's was about an application based on anthropometric and questionnaire data using support vector machines (SVMs) with eye fixations as features. The other is the theme word "sleep". DBLP's first occurrence was in 2016 [13] where the study integrated fuzzy set theory and decision trees to predict obstructive sleep apnea patterns, whereas insomnia [14] first appeared in MEDLINE in 2001 in a study involving telephone interviews.

IV. DISCUSSION

Over the last three decades, computer science and technology research has undergone two phases of development [15]. The first phase primarily involves studies of data collection methods, storage capacity and efficiency, and networking and communication by computing devices. Phase 2 is a modern descendant of the increasing power of data handling by more efficient and scalable tools, ranging from the study of information access, to knowledge discovery. The boundaries between these phases are fuzzy. Tracing this evolution may facilitate better understanding of the interaction between computer science and medical research and support more intentional acceleration of novel data technologies' adoption in medical research.

A. QUALITY OF EVIDENCE

The map's geometric shape represents relationships between papers in a way that captures developments in

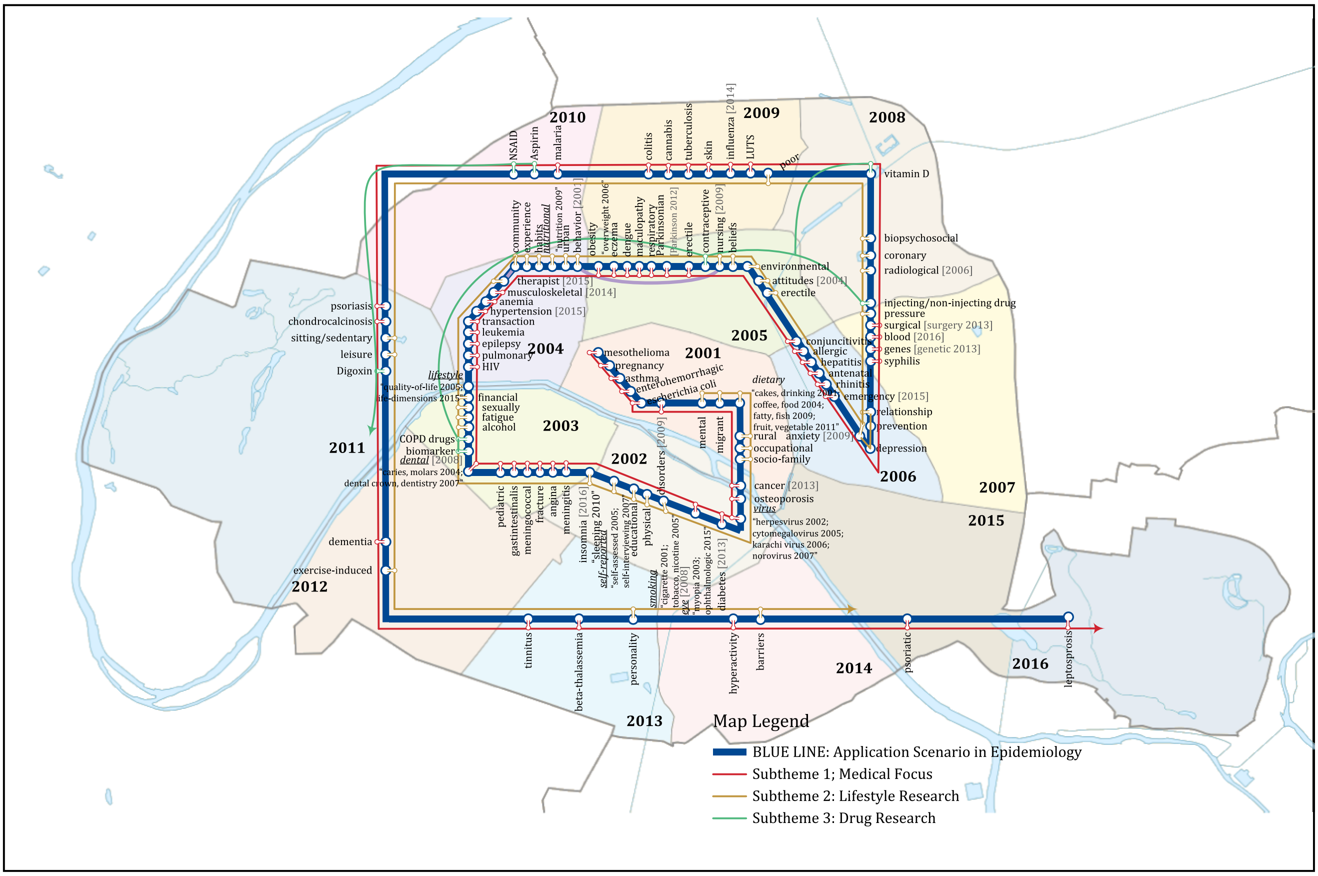


FIGURE 4. Chronological tracking of four themes generated among 481 article titles extracted from MEDLINE using three MeSH terms, "Questionnaires," "Epidemiology," and "Epidemiologic methods", between 2001 and 2016. Words that are the same as or similar to theme words that can be found in DBLP are labeled in square brackets.

specific themes. These intuitive representations make it easy to understand each theme's evolution. Although several annotators were needed to help evaluate the results, the classifier reduced the need for manual intervention, such as screening articles to avoid selection and attrition bias. Meanwhile, the annotators' workload was minimal; they were only required to review articles' titles and theme words produced by the classifier, and no prior knowledge of medicine was necessary. Our findings have the potential to help accelerate medical research by filtering relevant research ideas and to help users acquire knowledge more efficiently. Another important contribution is the ability to show chronological development of a selected research topic.

B. TECHNOLOGY DEVELOPMENT IN DBLP

Figure 2 displays a clear division in 2006 between two technology phases (Phase 1 and Phase 2). This represents two reversal processes, namely, from information to data in the field of information technology and from data to information in the field of data technology. Information technology is the process of producing data (via digitization and digital preservation), which stores information generated in the real world by human minds in digital form as data. Data technology is the process of analyzing data, which extracts useful information from data to support a wide range of applications.

The first phase of technology development in DBLP consisted of digitization of survey methods and data storage using information technology methods, which occurred between 2001 and 2006. Phase 1 in DBLP can be divided into three two-year periods: survey method digitization (2001-2002), questionnaire data preservation (2003-2004), and comparative analysis of computerized vs. paper-based survey databases (2005-2006). Survey method digitization is the conversion of traditional paper-based questionnaires into an electronic copy, such as by using a data matrix to store a questionnaire concerning employment, personal economics, computer skills, and disability level of handicapped computer specialists in Poland [16]. Although digitization is often seen as preservation, this is insufficient as file formats may not be readable in the future [17]. Preservation includes storage [18], [19], formatting [20], and visualization issues of computer-based questionnaires [21], that is, keeping information available and usable for future generations, requiring much more complex actions. Also, for digital materials (e.g., PDA [22], Web-based [9], touch screen [23], mobile [24], iPad [25], smartphone [26], and even stores in the cloud), there is a substantial difference between digitization and preservation. Notably, this phase was sparsely represented within MEDLINE.

The current phase of technology development in DBLP consists of questionnaire data analysis using data technology, which began in 2007. Phase 2 in DBLP can be further divided into 3 sub-phases. The earliest subphase, ending in 2008, involved automatically generating questionnaires, especially those involving a question bank and adaptive choice questions [27]. The second sub-phase involved analysis of a specific item within a questionnaire. Questionnaires [28], [29] are instruments or procedures that ask one or more questions and can include items from five main categories: binary responses (e.g., agree/disagree), rating scales (e.g., Likert response scales), multiple or single selection, open-ended comments, and non-question components (e.g., invitation, introduction, closing). Questionnaire analysis can involve analyzing Likert scales [30], [31], text-mining open-ended answers [32]–[38], handling missing data [39]–[42], drawing inferences according to existing answers [43], and cross-analyzing answers to obtain the psychology of audiences [36], [44]. Starting in 2010, questionnaire data analysis began to focus on related factors from non-questionnaire data (e.g., image, voice) to questionnaire data [45], [46]. The first two sub-phases involved conventional analyses that make it difficult to consider multiple data types simultaneously; the third sub-phase is questionnaire data mining. From a technical viewpoint, questionnaire analysis gradually progressed from conventional statistics to basic binary classification (e.g., SVMs [47], naïve Bayes [13], decision trees [48], random forests [49], [50]), to text analytics (e.g., semantic [51], ontology-driven [52], [53]) and, most recently, to data mining techniques (e.g., association rules [54], [55], clustering [56], [57]). We found that the analysis of innovation was concentrated between 2008 and 2009, converging with the computerization of questionnaires (whose keywords all appeared before 2007 and included ''interface'', ''surface'', ''structure'', ''offline'', ''seal'', ''concept-based'', ''components'', ''generation'', ''interact'', ''optimization'', ''imaging'', and ''computer-assisted'').

The quantity and complexity of today's questionnaire data raises a host of technical challenges not previously addressed within the studies identified as part of this literature review. For example, one challenge of analyzing heterogeneous data is to address the problem of composite hypotheses (e.g., normal, lognormal, Poisson) when integrating questionnaire data from multiple sources. However, traditional qualitative and quantitative methods are limited in their ability to cope with a multivariate explanation. Our classifier produced an intuitive representation of theme words that make it easy to understand this evolution.

C. APPLICATION SCENARIOS IN MEDICAL RESEARCH

DBLP literature covers a broad mix of research content, including: marketing [58], online shopping [59], service industries [60], advertising [61], online games [62], and medicine [13], [47], [63], [64]). Epidemiologic methods cover numerous aspects of medicine (108 theme words in total), such as specific medical conditions, lifestyle research, and drug research, suggesting that the development of innovative research methods is critical to enhancing medical and public health research (Figure 4).

D. GENERALIZABILITY

Questionnaire-related studies are generalizable and exist in multiple domains including public health, medicine, and computer science. A literature review of questionnaire-related studies might involve a large quantity of articles. Our approach was able to enhance the literature review process by classifying, tracking, and visualizing research trends automatically. The regional classifier partitioned theme words preprocessed by semantic analysis tools into a real map such that words that appear in the same time frame also appear in the same region. The geometric shapes on the map represent relationships between articles in a way that captures the development of specific themes. The maps' intuitive representations made it easy to understand each theme's evolution. Although we required several annotators to help us evaluate the results, the classifier reduced the need for manual intervention, such as screening or reviewing articles to avoid selection and attrition bias. Meanwhile, the annotators' workload was minimal; they were only required to review the articles' titles and the theme words produced by the classifier and no prior domain knowledge was necessary.

The efficiency of the literature review process can be enhanced through the automatic reclassification of articles. For example, the regional classifier divided the theme ''application scenario in epidemiology'' into three subclasses ''research on medical specialists,'' ''lifestyle research,'' and ''drug research'' (Figure 4). Such a reclassification based on an additional training dataset gave us a considerable amount of new information. Given such information, we envision that class reclassification is one possible breakthrough that will improve the quality, integrity, and applicability of clinical and public health research via data mining or machine learning techniques. The regional classifier uses new tools (e.g., graph theory) as a basis for the automatic partitioning of populations into successively smaller groups that are at increased risk of disease.

This study has several limitations. We only created maps based on article titles; future research may consider other parts of publications, such as the structured abstract (e.g., objectives, methods, results), which might offer more detailed information for a review. In addition, we might consider not only the first appearance of theme words, but also the year in which they appeared most frequently.

The quantity and complexity of questionnaire data constantly increases. It seems likely that data mining and machine learning techniques will have major implications in the future development of survey methodology. Key steps in the future of questionnaire-based medical research include: (1) studying the disease or pathology and being innovative in data selection, approach, and populations of interest

TABLE 1. Five Themes with their theme words generated from DBLP using the terms "Survey" or "Questionnaire Data" between 2001 and 2016.

Themes	Theme Words
Technology Development (RED LINE excluding KHAKI LINE) *27 words*	(2001) answer, interface, structure, open-ended, issues, surface; (2002) contingency, paper-and-pencil; (2003) offline, seal, Web-based, components; (2004) generation, touchscreen; (2005) interact, optimization; (2006) imaging, mobile; (2007) Likert; (2008) choice; (2010) animation-based; (2011) iPad; (2012) immersion, voice-based; (2013) Smartphone; and (2014) sensor, social network.
Questionnaire Data Analysis (KHAKI LINE) *27 words*	(2001) prediction, presence, classification; (2002) analysis; (2003) statistical, concept-based; (2004) SVMs; (2006) sampling, fuzzy logic; (2007) Bayesian; (2008) correlation, adaptive, clustering, agents, semantic, ontology-driven; (2009) association, bio-algorithms, visualization, maturity model; (2010) human-computer, artificial intelligence; (2011) neural network; (2012) classifier, random forest; (2013) map; and (2014) unsupervised.
Issues Related to Survey Studies (GREEN LINE) *29 words*	(2001) reflections; (2002) behavior, perceptions, communication, controllability; (2003) requirements, usability; (2004) credibility, opinions, attitudes, assessment; (2005) privacy, bias, satisfaction; (2006) response; (2007) performance, preference, security, reliability; (2009) safety, welfare, awareness, ethics; (2010) feasibility; (2012) impaired; (2013) acceptance; (2014) justice; (2015) abuse; and (2016) drawbacks.
Application Scenario Unrelated to Medical Research (PURPLE LINE excluding two theme words: (2005) physicians, medicine) *33 words*	(2002) social, library, environment; (2003) humanities, arts, psychological, industrial, rural, robot, transaction; (2004) school, students; (2005) university; (2007) enterprises, outsourcing, mental; (2008) bioinformatics, informatics, architectural, e-government; (2009) education, teachers, pupil; (2011) advertising, encyclopedia; (2012) e-financial, game, e-learning; (2013) farming, hacking; (2014) biomedicine; (2015) cross-cultural, cyberbullying.
Application Scenario Related to Medical Research (BLUE LINE) *34 words*	(2005) physicians, medicine; (2006) telemedicine, clinicians, surgeons, radiological, elderly people; (2008) eye, dental, antibiotic, patient-centered; (2009) nursing, anxiety, disorder; (2011) skin; (2012) Parkinson, larynx; (2013) brain, dysmenorrhea, diabetes, genetic cancer, lung cancer, rehabilitation, surgery; (2014) influenza, musculoskeletal; (2015) therapist, hypertension, ophthalmic, autism, emergency department; (2016) sleep, blood study, cardiomyopathy.

Note that underlined theme words in DBLP also appear in MEDLINE.

(i.e., class reclassification) to identify data most likely to help solve a clinical puzzle; (2) automating hypothesis generation and evaluation processes to reveal relationships between agents, vectors, modes of transmission, incubation periods, and host susceptibility (e.g., genomics) [65]; and (3) utilizing machine intelligence algorithms to evaluate and monitor epidemics to determine optimal approaches to prevention and disease management.

V. CONCLUSION

We conducted a data-driven exploratory literature review and identified efficiencies by augmenting that process with the regional classifier. For example, we identified two phases of technology development in questionnaire-based research, focusing on the shift from information technology methods to data technology methods. The regional classifier has the potential to help users acquire knowledge efficiently through an efficient filtering and visualization of relevant research ideas.

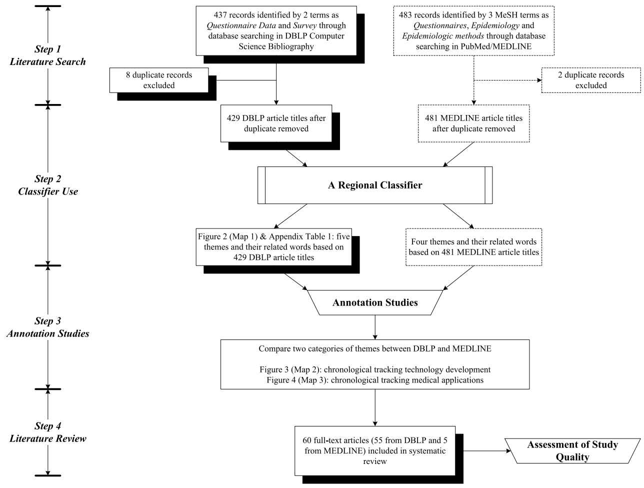


FIGURE 5. Flowchart of methods including four steps: Literature search, classifier use, annotation studies, and review. The boxes with dotted outlines show a parallel comparison in MEDLINE.

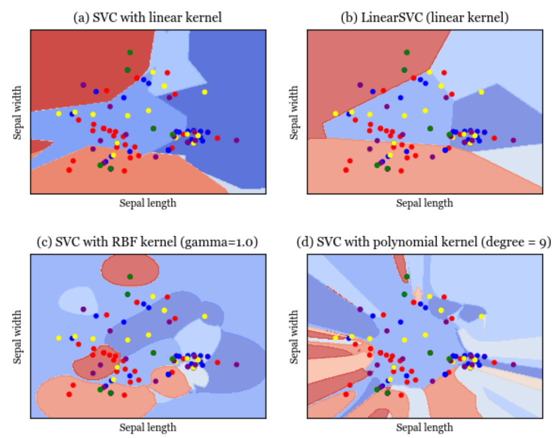


FIGURE 6. Four SVM classifiers' results on the DBLP dataset compared with our classifier, each region with random colors denotes a class of predicted year. Note that a total of 150 color points that show five themes is in lines with Figure 2 and Table 1.

APPENDIX

See Table 1 and Figures 5 and 6.

ACKNOWLEDGMENT

The authors would like to thank Haohan Zhang, Yuxuan She, and Lu Ruan for their help with the annotation.

REFERENCES

[1] K. Börner, *Atlas of Science: Visualizing What We Know*. Cambridge, MA, USA: MIT Press, 2010.
[2] R. Swan and D. Jensen, "Timemines: Constructing timelines with statistical models of word usage," in *Proc. KDD Workshop Text Mining*, 2000, pp. 73–80.

[3] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang, "Evolutionary timeline summarization: A balanced optimization framework via iterative substitution," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2011, pp. 745–754.
[4] K. P. Hewagamage, M. Hirakawa, and T. Ichikawa, "Interactive visualization of spatiotemporal patterns using spirals on a geographical map," in *Proc. IEEE Symp. Vis. Lang.*, Sep. 1999, pp. 296–303.
[5] D. Shahaf, C. Guestrin, and E. Horvitz, "Metro maps of science," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1122–1130.
[6] C. Tang, H. Zhang, K. H. Lai, Y. She, Y. Xiong, and L. Zhou, "Developing a regional classifier to track patient needs in medical literature using spiral timelines on a geographical map," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, Nov. 2017, pp. 864–869.
[7] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120.
[8] W. Hackbusch, *Multi-Grid Methods and Applications*. Berlin, Germany: Springer, 2003.
[9] S. Milne, L. Gibson, P. Gregor, and K. Keighren, "Pupil consultation online: Developing a Web-based questionnaire system," in *Proc. Conf. Interact. Design Children*, 2003, pp. 127–133.
[10] M. M. H. J. van Gelder, R. W. Bretveld, and N. Roeleveld, "Web-based questionnaires: The future in epidemiology?" *Amer. J. Epidemiology*, vol. 172, no. 11, pp. 1292–1298, 2010.
[11] D. Globe, A. Schoua-Glusberg, S. Paz, E. Yu, S. Preston-Martin, S. Azen, and R. Varma, "Using focus groups to develop a culturally sensitive methodology for epidemiological surveys in a Latino population: Findings from the Los Angeles Latino Eye Study (LALES)," *Ethnicity Disease*, vol. 12, no. 2, pp. 259–266, 2002.
[12] M. Nakayama and Y. Takahasi, "Estimation of certainty for responses to multiple-choice questionnaires using eye movements," *Trans. Multimedia Comput., Commun., Appl.*, vol. 5, no. 2, 2008, Art. no. 14.
[13] K.-J. Wang, K.-H. Chen, S.-H. Huang, and N.-C. Teng, "A prognosis tool based on fuzzy anthropometric and questionnaire data for obstructive sleep apnea severity," *J. Med. Syst.*, vol. 40, no. 4, p. 110, 2016.
[14] M. M. Ohayon, "Methodology of a study on insomnia in the general population," *L'Encephale*, vol. 28, no. 3, pp. 217–226, 2002.
[15] C. Tang, *The Data Industry: The Business and Economics of Information and Big Data*. Hoboken, NJ, USA: Wiley, 2016.
[16] W. Szczesny, "Grade correspondence analysis applied to contingency tables and questionnaire data," *Intell. Data Anal.*, vol. 6, no. 1, pp. 17–51, 2002.
[17] S. Perry, "Digitization and digital preservation: A review of the literature," *SLIS Student Res. J.*, vol. 4, no. 1, p. 4, 2014.
[18] S. Booth-Kewley, G. E. Larson, and D. K. Miyoshi, "Social desirability effects on computerized and paper-and-pencil questionnaires," *Comput. Hum. Behav.*, vol. 23, no. 1, pp. 463–477, 2007.
[19] A. Kano and J. Read, "Interchangeability of computer and paper based questionnaires in gathering computer experience data from young children," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 2397–2400.
[20] A. B. Smith, G. Velikova, E. P. Wright, P. Lynch, and P. J. Selby, "Computer-assisted questionnaires may facilitate collection of quality-of-life (QOL) data: At a cost," *Comput. Hum. Behav.*, vol. 22, no. 6, pp. 991–1000, 2006.
[21] J. Z. Zhan and S. Matwin, "Privacy-preserving data mining in electronic surveys," *IJ Netw. Secur.*, vol. 4, no. 3, pp. 318–327, 2007.
[22] P. Boissy, K. Jacobs, and S. H. Roy, "Usability of a barcode scanning system as a means of data entry on a PDA for self-report health outcome questionnaires: A pilot study in individuals over 60 years of age," *BMC Med. Inform. Decis. Making*, vol. 6, no. 1, p. 42, 2006.
[23] P. J. Blignaut, "Computerized self-administered questionnaires on touch-screen kiosks: Do they tell the truth?" in *Proc. CHI Extended Abstracts Hum. Factors Comput. Syst.*, 2004, pp. 1183–1186.
[24] T. Robinson, T. Cronin, H. Ibrahim, M. Jinks, T. Molitor, J. Newman, and J. Shapiro, "Smartphone use and acceptability among clinical medical students: A questionnaire-based study," *J. Med. Syst.*, vol. 37, no. 3, p. 9936, 2013.
[25] K. Gaunt, F. M. Schmitz, and M. Stolze, "Choose popovers over buttons for iPad questionnaires," in *Proc. IFIP Conf. Hum.-Comput. Interact.* Berlin, Germany: Springer, 2011, pp. 533–540.
[26] J. Axup and S. Viller, "Sampling mobile opinion: A contextual postcard questionnaire study," *First Monday*, vol. 11, no. 9, 2006. [Online]. Available: https://journals.uic.edu/ojs/index.php/fm/article/view/1402/1320
[27] J. Abernethy, T. Evgeniou, O. Toubia, and J.-P. Vert, "Eliciting consumer preferences using robust adaptive choice questionnaires," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 2, pp. 145–155, Feb. 2008.
[28] R. M. Groves, F. J. Fowler, Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, *Survey Methodology*. Hoboken, NJ, USA: Wiley, 2011.
[29] J. H. Abramson and Z. H. Abramson, *Research Methods in Community Medicine: Surveys, Epidemiological Research, Programme Evaluation, Clinical Trials*. Hoboken, NJ, USA: Wiley, 2011.
[30] M. Slater and M. Garau, "The use of questionnaire data in presence studies: Do not seriously Likert," *Presence*, vol. 16, no. 4, pp. 447–456, Aug. 2007.
[31] A. Gosavi, "Analyzing responses from likert surveys and risk-adjusted ranking: A data analytics perspective," *Procedia Comput. Sci.*, vol. 61, no. 7, pp. 24–31, 2015.
[32] A. Hiramatsu, H. Oiso, S. Tamura, and N. Komoda, "Support system for analyzing open-ended questionnaires data by culling typical opinions," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2004, pp. 1377–1382.
[33] D. Kelly, P. B. Kantor, E. L. Morse, J. Scholtz, and Y. Sun, "Questionnaires for eliciting evaluation data from users of interactive question answering systems," *Natural Lang. Eng.*, vol. 15, no. 1, pp. 119–141, 2009.
[34] M. Kimura, "The method to analyze freely described data from questionnaires," *J. Adv. Comput. Intell. Intell. Inform.*, vol. 13, no. 3, pp. 268–274, 2009.
[35] Y. Uchida, T. Yoshikawa, T. Furuhashi, E. Hirao, and H. Iguchi, "Extraction of important keywords in free text of questionnaire data and visualization of relationship among sentences," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Aug. 2009, pp. 1604–1608.
[36] J. Mažgut, M. Paulinyová, and P. Tiňo, "Using dimensionality reduction method for binary data to questionnaire analysis," in *Proc. Int. Doctoral Workshop Math. Eng. Methods Comput. Sci.* Berlin, Germany: Springer, 2011, pp. 146–154.
[37] D. Kobayashi, T. Yoshikawa, and T. Furuhashi, "Visualization and analytical support of questionnaire free-texts data based on HK Graph with concepts of words," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jun. 2011, pp. 1339–1343.
[38] J. R. Van Ginkel, L. A. Van der Ark, K. Sijtsma, and J. K. Vermunt, "Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation," *Comput. Statist. Data Anal.*, vol. 51, no. 8, pp. 4013–4027, 2007.
[39] J. Dauwels, L. Garg, A. Earnest, and L. K. Pang, "Tensor factorization for missing data imputation in medical questionnaires," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 2109–2112.
[40] N. Yamaguchi, M. Wu, M. Nakata, and H. Sakai, "Application of rough set-based information analysis to questionnaire data," *J. Adv. Comput. Intell. Intell. Inform.*, vol. 18, no. 6, pp. 953–961, Nov. 2014.
[41] L. Garg, J. Dauwels, A. Earnest, and K. P. Leong, "Tensor-based methods for handling missing data in quality-of-life questionnaires," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1571–1580, Sep. 2014.
[42] W. Wan, H. Xu, W. Zhang, X. Hu, and G. Deng, "Questionnaires-based skin attribute prediction using Elman neural network," *Neurocomputing*, vol. 74, no. 17, pp. 2834–2841, 2011.
[43] P. Busch and D. Richards, "Modelling tacit knowledge via questionnaire data," in *Proc. Int. Conf. Formal Concept Anal.* Berlin, Germany: Springer, 2004, pp. 321–328.
[44] S. L. France and W. H. Batchelder, "Unsupervised consensus analysis for on-line review and questionnaire data," *Inf. Sci.*, vol. 283, pp. 241–257, Nov. 2014.
[45] A. Verikas, A. Gelzinis, M. Bacauskiene, M. Hållander, V. Uloza, and M. Kaseta, "Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders," *Artif. Intell. Med.*, vol. 49, no. 1, pp. 43–50, 2010.
[46] R. Chen, T. Chen, Y. Chien, and Y. Yang, "Novel questionnaire-responded transaction approach with SVM for credit card fraud detection," in *Proc. Int. Symp. Neural Netw.* Berlin, Germany: Springer, 2005, pp. 916–921.
[47] A. C. Constantinou, N. Fenton, W. Marsh, and L. Radlinski, "From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support," *Artif. Intell. Med.*, vol. 67, pp. 75–93, Feb. 2016.
[48] A. Verikas, M. Bacauskiene, A. Gelzinis, and V. Uloza, "Monitoring human larynx by random forests using questionnaire data," in *Proc. 11th Int. Conf. Intell. Syst. Design Appl.*, Nov. 2011, pp. 914–919.
[49] M. Bacauskiene, A. Verikas, A. Gelzinis, and A. Vegiene, "Random forests based monitoring of human larynx using questionnaire data," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5506–5512, 2012.

[50] K. R. Larsen, D. Nevo, and E. Rich, "Exploring the semantic validity of questionnaire scales," in *Proc. 41st Annu. Hawaii Int. Conf. Syst. Sci.*, Jan. 2008, p. 440.

[51] N. Tamura, T. Terashita, and K. Ogasawara, "A pilot study measuring changes in student impressions before and after clinical training using a questionnaire based on the semantic differential technique," *Stud. Health Technol. Inform.*, vol. 192, p. 1038, Jan. 2012.

[52] M. Bouamrane, A. Rector, and M. Hurrell, "Gathering precise patient medical history with an ontology-driven adaptive questionnaire," in *Proc. 21st IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2008, pp. 539–541.

[53] Y.-L. Chen and C.-H. Weng, "Mining fuzzy association rules from questionnaire data," *Knowl.-Based Syst.*, vol. 22, no. 1, pp. 46–56, 2009.

[54] S. H. Burton, R. G. Morris, C. G. Giraud-Carrier, J. H. West, and R. Thackeray, "Mining useful association rules from questionnaire data," *Intell. Data Anal.*, vol. 18, no. 3, pp. 479–494, 2014.

[55] L.-H. Wang, C.-F. Hong, and C.-L. Hsu, "Closed-ended questionnaire data analysis," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.* Berlin, Germany: Springer, 2006, pp. 1–7.

[56] Y. Watanabe, T. Yoshikawa, and T. Furuhashi, "Study on analysis of questionnaire data based on interactive clustering," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Aug. 2009, pp. 98–103.

[57] K. Inagaki, T. Yoshikawa, and T. Furuhashi, "A study on extraction of minority groups in questionnaire data based on spectral clustering," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2014, pp. 988–993.

[58] T. Terano and Y. Ishino, "Interactive knowledge discovery from marketing questionnaire using simulated breeding and inductive learning methods," in *Proc. 2nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 279–282.

[59] G. Xirong and H. Yang, "Typology of consumers' risk perceptions in online shopping: an empirical approach based on the questionnaire data," in *Proc. Int. Conf. E-Bus. E-Government (ICEE)*, May 2010, pp. 2188–2191.

[60] T. Ishigaki, Y. Motomura, M. Dohi, M. Kouchi, and M. Mochimaru, "Knowledge extraction from a computational consumer model based on questionnaire data observed in retail service," *Int. J. Syst. Service-Oriented Eng.*, vol. 1, no. 2, pp. 40–54, 2010.

[61] A. Cymbala and M. Owczarczuk, "Variables importance in questionnaire data on advertising," *Expert Syst. Appl.*, vol. 38, no. 11, pp. 14218–14224, 2011.

[62] J. Frommel, K. Rogers, J. Brich, D. Besserer, L. Bradatsch, I. Ortinau, R. Schabenberger, V. Riemer, C. Schrader, and M. Weber, "Integrated questionnaires: Maintaining presence in game environments for self-reported data acquisition," in *Proc. Annu. Symp. Comput.-Hum. Interact. Play*, 2015, pp. 359–368.

[63] A. Verikas, A. Gelzinis, M. Bacauskiene, V. Uloza, and M. Kaseta, "Using the patient's questionnaire data to screen laryngeal disorders," *Comput. Biol. Med.*, vol. 39, no. 2, pp. 148–155, 2009.

[64] E. S. V. Eijk, M. W. Bettink-Remeijer, R. Timman, and J. J. V. Busschbach, "From pen-and-paper questionnaire to a computer-assisted instrument for self-triage in the ophthalmic emergency department: Process and validation," *Comput. Biol. Med.*, vol. 66, pp. 258–262, Nov. 2015.

[65] L. H. Kuller, "Epidemiology: Then and now," *Amer. J. Epidemiol.*, vol. 183, no. 5, pp. 372–380, 2016.

CHUNLEI TANG is a research associate at Harvard Medical School. Dr. Tang, author of *The Data Industry: The Business and Economics of Information and Big Data*, is a recognized advocate for the Data Economy. Tang's work begins with an innovative data-mining method built upon her doctoral work and applicable to various fields (e.g., finance, economics, insurance, bioinformatics, and sociology). She then endeavored to better match data products with their marketing capabilities by tackling business innovation. Inspired by business, she implemented a theoretical innovation for Data Science and proposed the (now-mainstream) definition of the data industry in 2013, stating that the data industry aims to bridge the gap between data science and economics. Her current focus is on applying innovation in healthcare, and she believes precision medicine is an application of the data industry.

YANGYONG ZHU is a notable data scientist from China, who pioneered the concept of "dataology". Zhu is a professor of computer science in data mining at Fudan University and serves as the chairman of the academic committee for Fudan's School of Computer Science. Zhu is also a founding director of the Shanghai Key Laboratory of Data Science and the co-founder of Fudan's Institute for Data Industry. So far, the laboratory is the only provincial key laboratory in China. Zhu's work begins in the application of data mining to various fields (e.g., finance, economics, insurance, bioinformatics, sociology) from 1996, and since 2004 has focused on data science, particularly at the theoretical level. After activities and publications, Zhu is recognized as one of the vital global advocates in data science.

SUZANNE V. BLACKLEY is a computational linguist and has been a member of the MTERMS research lab at Brigham and Women's Hospital since 2015. Her work focuses on clinical applications of natural language processing and machine learning.

JOSEPH M. PLASEK is a health data scientist and clinical informatician. Joseph joined the MTERMS research lab at Brigham and Women's Hospital in 2010. His research focuses on clinical applications of natural language processing, machine learning, and dynamic systems theory.

MEIHAN WAN is a postgraduate majoring in computer science at Fudan University. Her research area is data mining, representation learning and clinical applications using machine learning.

LI ZHOU is an Associate Professor at Harvard Medical School and a Lead Investigator at Brigham and Women's Hospital (BWH). Her research interests are in the areas of data science, natural language processing (NLP), medication safety, and clinical decision support. She has served as PI and co-investigator on numerous collaborative, multiple-site research grants. She directs the MTERMS lab at BWH. Her team has designed and implemented a generic and highly configurable NLP system known as MTERMS, which has been effectively applied to various clinical domains and documents.

JING MA is an associate professor and the director of the China Program for Health Innovation & Transformation at the Department of Population Medicine and Harvard Pilgrim Health Care Institute at Harvard Medical School. Ma has extensive experience in large scale longitudinal population research in the United States. Ma's research interests have been focusing on biological, lifestyle and environmental risk factors and their interactions that related to risk and progression of noncommunicable diseases, especially cancer, drawing on the data of Physicians' Health Study and Nurses' Health Study, among others. She has also been closely involved in fostering international research collaborations with multiple cohorts among many countries.

DAVID W. BATES is an American-born physician, biomedical informatician, and professor, who is internationally renowned for his work regarding the use of health information technology to improve the safety and quality of healthcare, in particular by using Clinical Decision Support. Dr. Bates has done especially important work in the area of medication safety. He began by describing the epidemiology of harm caused by medications, first in hospitalized patients and then in other settings such as the home and nursing homes. Subsequently, he demonstrated that by implementing computerized physician order entry (CPOE), medication safety could be dramatically improved in hospitals. This work led the Leapfrog Group to call CPOE one of the four changes that would most improve the safety of U.S. healthcare. It also helped hospitals to justify investing in electronic health records and in particular, CPOE.