Big Data Research 2 (2015) 2–11

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Promises and Challenges of Big Data Computing in Health Sciences ☆

Tao Huang [b,1], Liang Lan [c,1], Xuexian Fang [a], Peng An [a,d], Junxia Min [d], Fudi Wang [a,*]

[a] Department of Nutrition, Research Center for Nutrition and Health, Institute of Nutrition and Food Safety, School of Public Health, School of Medicine, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Zhejiang University, Hangzhou, China
[b] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
[c] Institute for Infocomm Research, Singapore
[d] The First Affiliated Hospital, Institute for Translational Medicine, School of Medicine, Zhejiang University, Hangzhou, China

ARTICLE INFO

Article history:
Received 30 September 2014
Received in revised form 20 January 2015
Accepted 11 February 2015
Available online 18 February 2015

ABSTRACT

With the development of smart devices and cloud computing, more and more public health data can be collected from various sources and can be analyzed in an unprecedented way. The huge social and academic impact of such developments caused a worldwide buzz for big data. In this review article, we summarized the latest applications of Big Data in health sciences, including the recommendation systems in healthcare, Internet-based epidemic surveillance, sensor-based health conditions and food safety monitoring, Genome-Wide Association Studies (GWAS) and expression Quantitative Trait Loci (eQTL), inferring air quality using big data and metabolomics and ionomics for nutritionists. We also reviewed the latest technologies of big data collection, storage, transferring, and the state-of-the-art analytical methods, such as Hadoop distributed file system, MapReduce, recommendation system, deep learning and network Analysis. At last, we discussed the future perspectives of health sciences in the era of Big Data.

1. Introduction

The concept of Big Data is causing a world-wide buzz. Its successful applications in business [1], sciences and healthcare [2] have radically changed their traditional practices. The demand for Big Data analysis is increasing day by day. More than 200 colleges provide degrees with Data Science (http://colleges.datascience.community/, accessed on September 14, 2014).

But there are many misunderstandings about Big Data and even its definition is debatable. According to http://datascience.berkeley.edu/what-is-big-data/ (accessed on September 14, 2014), there are at least 43 different definitions. Generally speaking, people agree that Big Data should have four V's: (1) big **volume** of data; (2) **variety** of data type; (3) high **velocity** of data generation and updating [3]; and (4) big data creates big **value** [4]. The first three V's focused on data engineering, such as data collection, storage and transferring. The last V focused on data science, such as analytic and statistical methods, knowledge extraction and decision-making.

To start a Big Data project, several steps are suggested as shown in Fig. 1: First, the right problem should be chosen. There are three kinds of problems. The first kind of problem has already been solved with traditional method and there is no need to use big data technologies. The second kind of problem is impossible to be solved with current technologies. We should focus on the third kind of problem that is solvable with current big data technologies. Second, we need to generate the data by sensors, monitors, molecular profiling or extract the data from public databases/sources after setting up a practical goal. Third, we need to do data pre-processing to obtain clean and meaningful data. Data pre-processing is a critical step for the success of a Big Data project. A recent publication [5] showed that sample misalignment for eQTL (expression Quantitative Trait Loci) and mQTL (methylation Quantitative Trait Loci) studies will reduce the discovered associations by 2–7 folds. The quality control of data essentially determines the upper bound of the data product, i.e. garbage in garbage out. The clean data will be stored into database for the next step analysis. Fourth, the insight or knowledge will be discovered from the processed data through statistical analysis. At last, the analytic results will be presented to the end user as a report, an online recommendation or a decision-making. Visualization of data, such as networks/graphs and charts, make the analytic re-

☆ This article belongs to Comp, Bus & Health Sci.

* Corresponding author.
E-mail addresses: tohuangtao@126.com (T. Huang), lanliang@i2r.a-star.edu.sg (L. Lan), xuexianfang@zju.edu.cn (X. Fang), anpeng@sibs.ac.cn (P. An), junxiamin@zju.edu.cn (J. Min), fwang@zju.edu.cn, fudiwang.lab@gmail.com (F. Wang).
[1] Co-first author.

http://dx.doi.org/10.1016/j.bdr.2015.02.002
2214-5796/

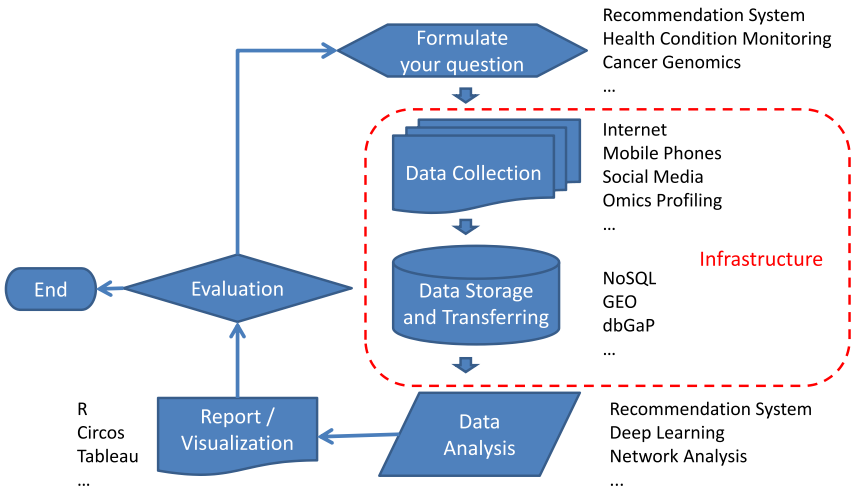

Fig. 1. The workflow of a typical Big Data project. Steps for Big Data projects: 1. Formulate your question; 2. Find the right ways (smart devices, Internet, hospitals ...) to collect your data; 3. Store the data; 4. Analyze your data; 5. Generate the analysis report with vivid visualization. 6. Evaluate the project: problem solved or start over.

sults easy to interpret and understand. If the results do not make sense, we need to reformulate our problems and start the steps over again.

In health sciences, there are many problems that can be addressed with big data technologies, such as recommendation system in healthcare, Internet based epidemic surveillance, sensor based health condition and food safety monitoring, Genome-Wide Association Studies (GWAS) and expression Quantitative Trait Loci (eQTL), inferring air quality using big data and metabolomics and ionomics for nutritionists.

To solve these problems, many advanced computational technologies will be used. We will cover the following technological perspectives: (1) Infrastructure of Big Data; (2) Analyzing of Big Data Results; and (3) Visualization of Big Data Results. And the future perspectives of health sciences in the era of big data will be discussed.

2. The Big Data studies in health sciences

Big Data technologies have many successful applications in biomedicine, especially in health sciences. For example, the data from search engines and social networks can help to gather people's reactions and monitor the conditions of epidemic diseases. It is worldwide real-time analysis and much faster than the official channels, such as CDC (Centers for Disease Control) and WHO (World Health Organization). Several cases studies will be elaborated in following paragraphs.

2.1. Recommendation system in health care

Recently, many researchers applied recommendation techniques to health information systems [5–7]. Duan et al. [6] proposed a nursing care plan recommendation system to provide clinical decision support, nursing education, and clinical quality control. It serves as a complement to existing practice guidelines. Hoens et al. [7] proposed a reliable privacy-preserved medical recommendation system. In this medical system, the patients can contribute their secured ratings of the physicians on different health conditions based on their satisfactions. This system can recommend a list of physicians who best suits their needs. Wiesner et al. [5] proposed a health recommendation system in the context of personal health record system. In their proposed Health Recommendation System (HRS), the items are non-confidential, scientifically proven or at least generally accepted medical information. The goal of HRS is to supply the users with medical information which is highly relevant to the patient's personal health record [8].

2.2. Internet based epidemic surveillance

At http://www.google.com/flutrends/, Google provides a tool called Google Flu Trends for real-time surveillance of influenza outbreaks [9]. Its assumption is that when the number of people have influenza symptoms, the searches for influenza related topics will increase [10]. Therefore based on Internet searches, the number of people with influenza symptoms can be estimated. The predictions made by Google Flu Trends were 7–10 days prior to the official CDC networks and their results were consistent [11].

For Chinese users, Baidu disease trend (http://trends.baidu.com/disease/) provided the province–city–county view of prevalence of several diseases include hepatitis, tuberculosis, venereal disease and influenza. What's more, its Big Data Trend product is open to ordinary users and therefore similar trends can be customized.

Twitter is a widely used social networking and news-sharing platform. The tweets reflected people's opinions and judgments about public event, especially the epidemic outbreaks [12]. Several methods were developed to monitor people's reaction to epidemic outbreak [12] and early disease syndrome based on Twitter [13]. The tweets involving H1N1 activity can be collected by searching key words, such as flu, influenza and H1N1. And the tweets involving public concern can also be filtered using keywords like travel, flight and ship for disease transmission, keywords like wash, hygiene and mask for disease counter measures. By studying the sequential tweets of H1N1 activity and public concern, the evolution pattern of public countermeasure can be revealed [12]. Similarly, by analyzing the early disease syndrome keywords, the risks of diseases such as cancer, flu, depression, aches/pains, allergies, obesity and dental disease, can be estimated [13].

2.3. Sensor based health condition and food safety monitoring

The integration of software and hardware, especially various sensors, create plenty amazing applications which monitor health condition and food safety. Many high-tech companies have launched their products, such Apple Watch from Apple (http://www.apple.com/watch/) which measures heart rate, Latin from Baidu (http://dulife.baidu.com/device/328) which measures body fat, MUMU from Baidu (http://dulife.baidu.com/device/330) which measures the blood pressure, Smart Chopsticks from Baidu which measures PH levels, temperature, calories and freshness of cooking oil [14]. Most such applications are based on well-established principles and have already been achieved with better accuracy or performance on larger instrument. The important meaning of

these products is that they can be easily used and their data can be automatically gathered and analyzed on the cloud. The gathered quantified data make the powerful Big Data analysis applicable and hidden patterns obvious.

2.4. GWAS and eQTL

With the drop down of genotyping cost in last decade, genome-wide association studies (GWAS) become more and more popular. It detects the association between common genetic variants (Minor Allele Frequency no less than 0.01) and phenotype trait [15]. To gain enough statistical power, usually thousands of samples or more are required [16]. Most GWAS results are included into GWAS databases, such as GWAS Catalog [17] (http://www.genome.gov/gwastudies/) and GWASdb [18] (http://jjwanglab.org/gwasdb). Based on GWAS Catalog [17] (accessed on September 28, 2014), there are 1961 GWAS of 1128 disease traits. 17,832 SNP-trait associations are discovered with p-value $< 10^{-5}$. Less stringent SNP-trait associations (p-value $< 10^{-3}$) can be found at GWASdb [18].

Many disease associated single-nucleotide polymorphisms (SNPs) were discovered by GWAS. For example, susceptibility variants for hepatitis B virus (HBV) infected hepatocellular carcinoma (HCC) were identified by Zhang et al. [19]. But the biological functions of these GWAS identified SNPs were not clearly, since many such SNPs are in intergenic region and the disease causal genes are very difficult to be exclusively pinned down [20]. This makes people rethink about how to improve the design GWAS and fully utilize GWAS results [16,21,22]. Expression Quantitative Trait Loci (eQTLs) could help interpret the mechanisms of disease associated SNPs who function through gene expression regulation. Several large eQTL databases, such as seeQTL [23], Genevar [24] and GTEx [25], are established to associate SNP and the gene expression it regulates. The association between genotype and gene expression can be measured with additive linear model [26], ANOVA model [26] or more sophisticated information-theoretic machine learning method [27].

2.5. Inferring air quality using Big Data

Air pollution would cause several health problems, such as, lung cancer [28], cardiovascular disease [29] and birth defects [30–32]. Now, air pollution is a significant issue in many developing countries. For example, Beijing reported 58 days of heavy air pollution in 2013 [33]. To monitor air quality helps to protect human health and control population. Traditional air quality monitoring methods need to establish and maintain physical monitoring stations. The coverage of monitoring is very limited because the expensive cost of establishing and maintaining them. Therefore, we may not know the air quality for the places that without monitoring stations.

Recently, several approaches are proposed to estimate air pollution from the perspective of big data. They try to inference air pollution information based on other data sources other than monitoring stations. Zheng et al. [34,35] inferred the air quality information in big cities in China by combining existing monitor stations data with other data sources, such as meteorology, traffic flow, human mobility, road networks and point of interests. This air quality monitory system can be publicly accessed at http://urbanair.msra.cn/. Mei et al. [36] estimate air quality from social media posts (e.g., Sina Weibo text context). Honicky et al. [37] suggested to attach sensors to GPS-enabled cell phones and used them to collect air pollution information. Chen et al. [38] introduced an indoor air quality monitoring system. They use an aerosol particle counter to detect indoor concentration of PM2.5.

2.6. Metabolomics and ionomics for nutritionists

The original goal of Metabolomics is to identify and quantify all metabolites within a system using Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS) [39]. But now the focus has shifted from identifying and quantifying metabolites to finding their biological associations with diseases [40]. It has been used to study Diabetes [41,42] and Toxicology [43]. MetaboLights [44] stored the metabolomics data from varies studies.

In recent year, Ionomics [45] is developed to study the ions, a special kind of metabolites, using ionomic profiling. It is first used in plant [46] and then in human [47]. Sun et al. [47] measured 17 ions (Ca, Cr, Cu, Fe, K, Mg, Mn, Mo, P, Re, S, Sb, Se, Sn, Sr, Ti, Zn) in 976 middle-aged Chinese men/women using plasma mass spectroscopy. Within the 976 participants, there are 516 overweight/obesity participants, 399 metabolic syndrome (MetS) participants and 122 type 2 diabetes (T2DM) participants. The metabolic abnormality associated ions or ion modules are identified with mutual information based method. It is found that some ions work individually while others tend to act together and form a functional module.

3. Infrastructure of Big Data

Even though goals of Big Data projects are different, they share some key patterns and use similar computational technologies. In this section, we will introduce these similar technologies, such as data collection, storage and transferring, from a computer science perspective.

3.1. Data collection

Large scale data could be collected from many differently data sources. In this section, we show four popular sources in generating large scale data.

Internet The early Big Data applications include online advertisement recommendation for search engines, product recommendation for e-commerce and potential friend recommendation for social networks. All these systems collect data from Internet, such as browsing history, search history, shopping history and social network. The Netflix Prize Challenge [48] was organized by Netflix in 2006 and its prize was one million dollars. It was a typical recommendation system challenge. The provided dataset included 480,000 users, 18,000 movies and 100 million ratings. The goal was to predict how much a user will love a movie.

Mobile phones The usage of mobile phones has been grown exceptionally in last decade. Furthermore, in past few years, smart phones remarkably embedded many sensors like GPS, accelerometer, gyroscope, camera, microphone, etc. The ubiquity of mobile phone and a large amount of data generated from embedded sensors offer new opportunities to characterize and understand user real-life behaviors (e.g., human mobility, interaction patterns). In Nokia Mobile Data Challenge [49], various information for 200 Nokia N95 phones was collected over a year. During this challenge, many interesting findings and multidisciplinary scientific advances were produced [49]. Mobile phones are also used frequently as data collection tools in the area of public health. Van Heerden et al. [50] used mobile phones to collect maternal health information from HIV-Positive Pregnant Woman in South Africa. Zhang et al. [51] showed that smart phones can be successfully used for household data collection on infant feeding in rural china.

Social media Online social networks such as Facebook, Twitter, LinkedIn have gained remarkable attentions in past few years. These online social networks are extremely rich in content. The tremendous amount of data available on online social networks provides unprecedented opportunities for data analytics in multiple areas. In the area of public health, online social network data are usually used for analyzing disease spreading/transmission [52,53]. The data from social networks could be collected by the following three ways [54]: (1) by retrieving data shared on social network websites; (2) by asking participants about their behavior; and (3) through deployed applications.

Biomedical data from hospital and scientific community Another important source for data is from hospital [55]. The patient data, such as CT (Computed Tomography) images, medical histories and genetic test data, is foundation for personalized medicine and large cohort study. In biomedical area, with the rapid development of sequencing and microarray technologies, tons of sequencing data and molecular profiling data were generated with low cost, high accuracy, fast speed and minimum sample requirement [56]. Some large projects, such 1000 genome (http://www.1000genomes.org) [57], ENCODE (Encyclopedia of DNA Elements, https://www.encodeproject.org/) [58], TCGA (The Cancer Genome Atlas, http://cancergenome.nih.gov/) [59], GTEx (Genotype-Tissue Expression, http://commonfund.nih.gov/GTEx) [25], accumulated large scale genomic data which are public available. These data sets provide benchmarks for method development and performance elevation of Big Data analysis.

3.2. Data storage and transferring

Since the data volume is huge, its storage and transferring are all quite different from traditional data analysis.

In traditional data management, SQL (Structured Query Language) based databases, such as MySQL (http://www.mysql.com/) and Oracle (http://www.oracle.com), were widely used to store and access the data. SQL queries the whole databases from disk and is slow. In Big Data era, NoSQL (often interpreted as **N**ot **O**nly **SQL**) was developed. It is fast and with low cost. It spreads and replicates the data on many different servers to achieve flexibility and reliability. NoSQL databases include MongoDB (http://www.mongodb.com/), Cassandra (http://cassandra.apache.org/) which is originated from Facebook, BigTable [60] which is developed by Google and only available to Google, and HBase (http://hbase.apache.org/) which is basically an open source version of BigTable.

For commercial cloud services, there are Windows Azure Service and Amazon Web Services (AWS). For academic research of biomedicine, especially gene expression and genotype data, the public available databases including Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/), the database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/gap/), Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra/), and TCGA Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu/).

Usually, the data collection, storage and analyzing are done or maintained by different teams or personals. To efficiently and safely transfer such huge amount of data is much difficult than it seems. Traditional tools, such as USB disks and CDs, are not good for Big Data sharing. Specialized big data transferring software, e.g. Aspera (http://asperasoft.com/), is developed to maximize the transfer speed and protect the context. Aspera is even used by US military to distribute defense videos and images [61]. For example, if an user would like to down data from dbGaP using Aspera, first, the user needs a dbGaP account which includes username and password, then a key with 100 random numbers and characters is generated for downloading a specific data file or selected files. In Aspera, the user needs to provide both the dbGaP account information and the random key to download the encrypted data stream. After the encrypted data is downloaded, the user needs to use a decrypt program and a separate decrypt code to decrypt the protected file into readable files. All these processes make the data transfer military-grade safe and prevent information leak. The transfer speed of Aspera could reach the upper limit of your Internet connection speed.

3.3. Computational technologies

Distributed and parallel systems are needed handle large scale data. Microprocessor technology has delivered impressive improvement in Central Processing Units (CPU) clock speed over the past decades. However, the CPU clock speed would not increase as rapidly as it used to because of its performance bottlenecks. To maintain the performance improvement in large-scale data processing, modern CPUs are becoming highly parallel. In the meanwhile, as stated in Amdahl's Law, improvements in processing speed must be matched by improvements in I/O [62]. A requirement of high I/O performance is a heavy use of local data storage. So, the data also needs to be distributedly stored as computation. Compared to single processor computation, it is hard to configure and maintain parallel and distributed hardware. And it is also difficult for software developer to write parallel and distributed software. In this section, we outline some recent progresses in parallel and distributed computing and storage.

3.3.1. Parallel computing

Nowadays, there are many different hardware architectures to support parallel computing. Two popular examples are: (1) Multicore computing: A multicore processor is a processor that includes multiple computing cores in a single chip; and (2) Cluster Computing: A cluster usually consists of set of inter-connected stand-alone computers that work together as a single integrated powerful computing resource. The individual computers with a cluster were connected by fast Local Area Networks (LAN).

3.3.2. Hadoop distributed file system

When dealing with the challenge of accessing big data that distributedly stored on a cluster, a cluster file system is a common solution. The cluster file system provides location transparent access to data files to the servers on the cluster. Hadoop Distributed File System (HDFS) [63] is a popular type of cluster file system which is designed for reliably storing large amount of data across machines in a large scale cluster. HDFS was originally derived from Google Files System (GFS) paper [64]. It provides an open source cluster file system similar to GFS.

HDFS utilizes master-slave architecture. HDFS logically separates the file system metadata and application data. The metadata is stored on a delicate computer named as NameNode (known as mater node in GFS) in HDFS. Application data was stored on other computers named as DataNodes (known as slave node in GFS). A data file is divided into one or more blocks and these divided blocks are replicated and stored across several DataNodes. All of the nodes contained with a Hadoop cluster are full connected and communicate with each other using TCP-based protocols. The NameNode maintains the file system namespace and the mapping of file blocks to DataNodes. When an HDFS client read a file, it first contacts the NameNode to get the locations of data blocks comprising the file and then reads these data blocks from closest DataNode.

Compared to traditional distributed file systems, HDFS have two important advantages: (1) Highly fault-tolerant. Unlike traditional

distributed file systems that use data protection mechanisms to make the data durable, HDFS provides replicated storage for massive data across multiple DataNodes and heartbeat for failure detection. When a DataNode fails, the NameNode will detect it and re-assign the work to other DataNodes; (2) Large Scale Data. The Hadoop clusters today can store **Petabyte** (**PB**) data. It supports high aggregate data bandwidth and scale to hundreds of nodes in a cluster.

3.3.3. MapReduce

MapReduce [65] is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. A popular open-source implementation of MapReduce framework is Apache Hadoop. The MapReduce is inspired by the map and reduce functions that are commonly used in functional programming. In a MapReduce framework, a programing consists of two primitive steps: Map() and Reduce(). In the Map() step, the master node takes the input and divides it into smaller sub-problems, and distributes them to slave nodes. A slave node may do this Map() procedure again for further dividing the problem. The slave node processes smaller problem and passes answer back to its master node. In the Reduce() step, the master nodes collects the answers and combines them together to form the final answer to the original problem that it wants to solve. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved when processing massive data on a large cluster. MapReduce greatly simplifies the task of writing a large-scale analysis on distributed data for many types of analysis.

4. How to analyze Big Data

Different types of Big Data require different analysis methods. A comprehensive list of analysis methods can be found at https://github.com/onurakpolat/awesome-bigdata. We choose three widely used analysis methods in computer science and biomedicine to share with the readers: (1) Recommendation System; (2) Deep Learning and (3) Network Analysis.

4.1. Recommendation system

Recommendation systems have become extremely common in recently years. They are applied successfully in various applications, such as recommending products on Amazon.com [66], movies recommendation by Netflix [67] and MovieLens [68], music recommendation by Last.fm and Pandora Radio, and news recommendation by VERSIFI Technologies [69].

In a recommendation system, there are two classes of entities: users and items. Let U be the set of all users and I be the set of all available items. Let R be the rating matrix, where $r(u,i)$ denotes the rating of an user u to an item i. The value of $r(u,i)$ indicates how a particular user likes a particular item. Usually, the rating matrix R (also known as utility matrix) is sparse, meaning that most entries in the matrix are "unknown". An "unknown" entry implies that we do not explicitly know a user's rating (i.e. preference) for an item. The goal of a recommendation system is to estimate the values of these "unknown" entries. Once the rating matrix is estimated, for each user, we can select top N items with highest ratings and recommended them to the user.

There are several ways to estimate the "unknown" entries in the rating matrix. Usually, recommendation systems are classified into three different categories based on their approaches to estimate the "unknown" ratings [70]: (1) Collaborative filtering [71] approaches, (2) Content-based filtering [72] and (3) Hybrid approaches. Collaborative filtering approaches are based on analyzing a large amount of data about users' rating, behaviors or activities over items. The items recommended to a user are those preferred by other similar users. Content-based filtering approaches are based on a description of the item and a profile of the user's preference. They recommend the items that are similar to the items that a user liked in the past. The hybrid approaches combine collaborative and content-based methods.

4.1.1. Collaborative Filtering

Collaborative Filtering methods estimate the unknown rating $r(u,i)$ based on the $r(u_j,i)$ assigned to item i by those user u_j who are similar to user u. There have been many collaborative systems developed in the academia and industry. The collaborative filtering algorithms are generally grouped into two classes: memory-based and model-based [73]. Memory-based algorithms [74] estimate the unknown rating based on the entire collection of previous rated items by the users. Formally, the value of an unknown rating $r(u,i)$ for user u on item i computed as an aggregation of rating of some other similar users for the same item i, One of the most common aggregation function is to use weighted sum shown as following

$$r_{u,i} = \sum_{u_j \in C} sim(u,u_j) r_{u_j,i}, \tag{1}$$

where C denotes a set of selected m users that are most similar to user u and $r(u_j,i)$ is the known rating of an user u_j for item i. To overcome the problem that different users may have different rating scale, Eq. (1) is modified to

$$r_{u,i} = \bar{r}_u + \sum_{u_j \in C} sim(u,u_j)(r_{u_j,i} - \bar{r}_{u_j}), \tag{2}$$

where $\bar{r}_u$ is the average rating of a user u. Both in (1) and (2), we need to compute the similarity between two users $sim(u,u_j)$. Usually, the similarity between two users is computed based on their ratings of the same items that they have previously rated. Two most popular metrics are *correlation* and *cosine-based*. There are also many extensions to the standard correlation and cosine-based techniques, such as default voting, weighted-majority prediction.

Instead of using some ad hoc heuristic rules (e.g. aggregate function) to estimate the "unknown" ratings, model-based algorithms use the previously ratings to learn a model, which is then used to predict "unknown ratings" [75]. There are many ways to build the model based on historical rating (i.e., training data), such as Bayesian networks, clustering methods and latent semantic models.

4.1.2. Content-based filtering

Contend-based filtering approaches estimate an "unknown" rating $r(u,i)$ based on ratings $r(u,j)$ of this user to the items $i' \in I$ that are "similar" to item i. The similarity between different items is estimated based on the descriptions of the items. For example, in music recommendation, the description of a song could contain: genres, singer, producer etc. The content-based recommendation approaches have its roots in information retrieval. In content-based recommendation system, keywords are used to describe the items. Similarity to text representation in information retrieval, Term Frequency/Inverse Document Frequency (TF-IDF) is used to represent the items. In content-based recommendation, various candidate items are compared with the items that a user liked in the past, and then the top "similar" items are recommended.

4.1.3. Hybrid approaches

To overcome the limitations for both collaborative filtering methods and content-based filtering methods, many hybrid recommendation systems are proposed. Recent research has demon-

strated that hybrid approaches can provide more accurate recommendation than pure collaborative or content-based filter methods. There are several ways to effective combine collaborative filter and contend-based filtering methods. A detailed review about recommender system can be found at [33].

4.2. Deep learning

Since 1980s, many machine learning methods, such as Neural Network (NN), Random Forest, Support Vector Machine (SVM), were developed to build classification models [76]. Then it was found that without better features, the performance of classifiers is difficult to improve [77,78]. Actually, the features determine the upper bound of a classifier's performance [79]. Since 2000, Transfer Learning [80], Manifold Learning [81] etc. were developed to learn the feature structure. However all these methods need hand craft and require experiences of tricks in practice [82]. In 2006, Hinton et al. published the historical paper "Reducing the dimensionality of data with neural networks" [83] which is considered as the beginning of deep learning. Unlike shallow machine learning models, deep learning uses Neural Network with many layers of hidden variables to automatically do feature learning, including pre-processing, feature extraction and feature selection. The unsupervised feature learning models in deep learning include Auto encoder [84], Restricted Boltzmann Machine [85] and Deep Boltzmann Machine [86]. More technical details and resources can be found at http://deeplearning.net/.

Deep learning has been quickly applied to many industrial projects and created big value, such as the speech recognition and Xbox from Microsoft, image recognition, Natural Language Processing (NLP) from Google. In China, Baidu Institute of Deep Learning (http://idl.baidu.com/) and Huawei Noah's Ark Lab (http://www.noahlab.com.hk/) also did many original works and launched products with Deep Learning technologies.

4.3. Network analysis

Many unstructured complex data can be organized as a graph. The graph theory provided solid theoretical foundation for network analysis. Several widely used network analysis methods will be briefly introduced in this section.

4.3.1. Topological analysis

In a graph, some nodes are more important than other. Such nodes are called hubs. The nodes of some sub-network are densely connected and form a module. The hub nodes can be defined by degrees and betweenness [87]. Degree is the number of connected edges in undirected network; in directed network, it can be further divided into in and out degree which correspond to the number of point out or point in edges. Many graph analysis were based degrees, such as k-core which decomposes the network into layers based on the degrees [88]. More basic concepts in graphs can be found in [89].

In signal analysis, betweenness is better than degrees since it considers the information flow across the whole network globally. For a graph $G = (V, E)$, where V are the nodes, E are the edges, the betweenness of node v is

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

where s and t are all the other network nodes, except node v, σ_{st} is the number of shortest paths between s and t, $\sigma_{st}(v)$ is the number of shortest paths that go through v. Betweenness has been proven to an important network features for proteins in protein–protein interaction network [90,91].

For module identifications, there are Cytoscape [92] apps, such as MCODE (Molecular Complex Detection) [93] and NeMo [94]. Basically, the nodes with high local neighborhood density are identified first and then expand from such nodes to get the densely connected regions, i.e., network modules [93].

4.3.2. Guilt-by-association

Guilt-by-association [95] is the basis of most protein function methods. It assumes that the functions of a protein are similar with the functions of its interaction neighbors. In general, it is true due to modularized nature of network [96]. Guilt-by-association can be modified into different forms. On weighted network, Hu et al. [97] calculated the weight sum of a protein's neighbors with certain functions, and then predicted the protein's most possible functions and most impossible functions based on the weight sum rank of all functions. This method fully utilized the edge weight information and its prediction was a ranked list of functions rather than just one function.

Another more subtle example is that the KEGG (Kyoto Encyclopedia of Genes and Genomes) and GO (Gene Ontology) enrichment score of a protein's neighbor can represent this protein's function accurately and robustly [90,98].

4.3.3. Shortest path analysis

The shortest path analysis from graph theory is partly corresponding to the concept of biological regulatory mechanisms. It is widely used to prioritize or identify novel diseases [99–102]. To do shortest path analysis, Dijkstra's algorithm [103] can find the shortest path between two nodes on the network. Based on shortest path, more flexible betweenness can be defined. For example, the betweenness among known disease genes can be a useful measurement for predicting novel disease genes [101,104,105] and the betweenness from virus protein to host protein can reflect the virus-host interactions and key proteins in virus infection [100,106].

Sometimes, the shortest path is too stringent. Even the paths that are slightly longer than the shortest one could provide very useful information, especially when the network is incomplete and noisy. Therefore, A^* search algorithm [107] is developed to find the 1st shortest paths, the 2nd shortest paths, ..., the k-th shortest paths. More detailed information can be found in [108].

4.3.4. Random Walk with restart

In recent year, there are more and more Random Walk with Restart (RWR) based disease gene prioritization method [109–114], even though sometimes its results are similar with shortest path analysis [115]. It simulates a random walker, who starts from m seed genes on the network and moves to its randomly chosen neighbors at each step [111]. In each tick of time, the state probabilities P_{t+1} at time $t+1$ is

$$P_{t+1} = (1-r)\, A' P_t + r P_0 \tag{4}$$

where P_t is state probabilities at time t, r is the restart probability, A' is the column-wise normalized adjacency matrix which represent the network structure, P_0 is the initial state probabilities which is a column vector with $1/m$ for the m seed genes and to 0 for other genes on the network.

This process is repeated until a steady-state is reached when the difference between two steps is smaller than a pre-defined cutoff. At last, each node on the network is assigned with a probability of the random walker reaching it [111].

An R [116] package RWOAG based the RWR method from Kohler et al. [111] is available at https://r-forge.r-project.org/R/?group_id=1126.

Table 1
The visualization tools for Big Data.

Tool	Data type	Web site	Advantage	Disadvantage
R	All types of data	http://cran.r-project.org/	Powerful, Thousands of packages, Rich documents Cross platform	Programming language, Difficult to learn
Cytoscape	Network	http://www.cytoscape.org/	Powerful, Graphical user interface, Easy to use	Click button, Time consuming
Circos	Circular	http://circos.ca/	Beautiful figures, Best for visualizing genome chromosomes, Widely used	Difficult to learn, Difficult to use
Gephi	Network	https://gephi.github.io/	Powerful, Graphical user interface, Easy to use	Click button, Time consuming
GraphViz	Network	http://www.graphviz.org/	Powerful	Command-line, Difficult to learn, Difficult to use
Tableau	Tables and Maps	http://www.tableausoftware.com/	Beautiful figures, Google map style	Commercial software

5. Visualization of Big Data results

Graphical presentation is the best way to intuitively get the meaning of the data and insight revealed by the analysis. There are many tools to visualize the Big Data as shown in Table 1.

As a generalized programming language, R [116] has about six thousand high quality packages (http://cran.r-project.org/web/packages/, accessed on September 14, 2014) that could achieve sophisticated functions. Its excellent help system and power functions make it the most widely used language in Data Science. For visualization R packages (http://cran.r-project.org/web/views/Graphics.html), ggplot2 and igraph provide general plot functions and network specific plot functions.

For network specific visualization, other choices include Cytoscape (http://www.cytoscape.org/) [92] which can save the network layout and provide rich visualization styles, Circos (http://circos.ca/) [117] which becomes the standard of visualizing genome chromosomes, Gephi (https://gephi.github.io/) [118] which is a java based application to create dynamic and hierarchical network graphs and GraphViz (http://www.graphviz.org/) which is a powerful command-line tool that can layout big network with massive numbers of nodes and edges. R provided wrappers for most visualization software, for example, RCytoscape [119] to Cytoscape, RCircos [120] to Circos, RGraphViz to GraphViz.

For general visualization which does not require programming skills, Tableau (http://www.tableausoftware.com/) is a good choice. Unlike previous ones, it is commercial software, but has try-it-free version. Its highlight feature is that it can visualize Google map style location data.

6. The future of health sciences

The changes that Big Data will bring to health sciences are much greater than most people estimated. Take the smart device for example. The health condition measurements from users will be stored, analyzed and shared in their cloud. The time course health measurement data with almost endless time points from millions of people will change the public health researchers from large number of nurses and doctors to computer scientist and few medical experts.

It will also change the usage of health science results. In pre-Big Data age, the health science result is usually a report and people will get its scientific meanings from Newspapers, TVs and Internets. It could take years or decades to educate the general population and build their health awareness. But in Big Data age, the scientific suggestions are personal and will be directly pushed to their smart devices. The personal suggestions are useful and therefore improve the compliance. The Information Push System makes sure the message can be delivered instantly and accurately. Science report could be as viral as viral videos.

The impact of health sciences will be amplified. The users not only get the suggestions which will create the will to change, but also the tools to full-fill that need. Precise product of exercise instrument, assistant for booking exercise field etc. could be provided. The seamless connection from science to business will bring enough financial support for the blooming of science research and actually improve the health of people.

Acknowledgements

The study was supported by research grants from National Natural Science Foundation of China grants (31030039, 31225013 and 31330036 to F.W.) and also by the Distinguished Professorship Program from Zhejiang University (to F.W.).

References

[1] A. McAfee, E. Brynjolfsson, Big data: the management revolution, Harv. Bus. Rev. 90 (2012) 60–66, 68, 128.
[2] M.M. Hansen, T. Miron-Shatz, A.Y. Lau, C. Paton, Big data in science and healthcare: a review of recent literature and perspectives, in: Contribution of the IMIA Social Media Working Group, Yearb. Med. Inform. 9 (2014) 21–26.
[3] R. Price, Volume, velocity and variety: key challenges for mining large volumes of multimedia information, in: Proceedings of the 7th Australasian Data Mining Conference, vol. 87, Australian Computer Society, Inc., Glenelg, Australia, 2008, p. 17.
[4] R. Leventhal, Trend: big data. Big data analytics: from volume to value, Healthc. Inform., Bus. Mag. Inf. Commun. Syst. 30 (2013) 12, 14.
[5] M. Wiesner, D. Pfeifer, Health recommender systems: concepts, requirements, technical basics and challenges, Int. J. Environ. Res. Public Health 11 (2014) 2580–2607.
[6] L. Duan, W.N. Street, E. Xu, Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterp. Inf. Syst. 5 (2011) 169–181.
[7] T.R. Hoens, M. Blanton, A. Steele, N.V. Chawla, Reliable medical recommendation systems with patient privacy, ACM Trans. Intell. Syst. Technol. 4 (2013) 1–31.
[8] L. Fernandez-Luque, R. Karlsen, L.K. Vognild, Challenges and opportunities of using recommender systems for personalized health education, Stud. Health Technol. Inform. 150 (2009) 903–907.
[9] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (2009) 1012–1014.
[10] H.A. Carneiro, E. Mylonakis, Google trends: a web-based tool for real-time surveillance of disease outbreaks, Clin. Infect. Dis., Off. Publ. Infect. Dis. Soc. Am. 49 (2009) 1557–1564.
[11] A.F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, R.E. Rothman, Influenza forecasting with Google flu trends, PLoS ONE 8 (2013) e56176.

[12] A. Signorini, A.M. Segre, P.M. Polgreen, The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic, PLoS ONE 6 (2011) e19467.
[13] M. Paul, M. Dredze, You are what you tweet: analyzing twitter for public health, Artif. Intell. (2011) 265–272.
[14] Y. Jie, Is your food safe? New 'smart chopsticks' can tell in: China real time, Wall Street J. (2014), http://blogs.wsj.com/chinarealtime/2014/09/03/is-your-food-safe-baidus-new-smart-chopsticks-can-tell/ (accessed on 2014/9/16).
[15] G. Zheng, Y. Yang, X. Zhu, R.C. Elston, Analysis of Genetic Association Studies, Springer, New York, 2012.
[16] P. Marjoram, A. Zubair, S.V. Nuzhdin, Post-GWAS: where next? More samples, more SNPs or more biology? Heredity 112 (2014) 79–88.
[17] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, H. Parkinson, The NHGRI GWAS catalog, a curated resource of SNP-trait associations, Nucleic Acids Res. 42 (2014) D1001–1006.
[18] M.J. Li, P. Wang, X. Liu, E.L. Lim, Z. Wang, M. Yeager, M.P. Wong, P.C. Sham, S.J. Chanock, J. Wang, GWASdb: a database for human genetic variants identified by genome-wide association studies, Nucleic Acids Res. 40 (2012) D1047–1054.
[19] H. Zhang, Y. Zhai, Z. Hu, C. Wu, J. Qian, W. Jia, F. Ma, W. Huang, L. Yu, W. Yue, Z. Wang, P. Li, Y. Zhang, R. Liang, Z. Wei, Y. Cui, W. Xie, M. Cai, X. Yu, Y. Yuan, X. Xia, X. Zhang, H. Yang, W. Qiu, J. Yang, F. Gong, M. Chen, H. Shen, D. Lin, Y.X. Zeng, F. He, G. Zhou, Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers, Nat. Genet. 42 (2010) 755–758.
[20] P.M. Visscher, M.A. Brown, M.I. McCarthy, J. Yang, Five years of GWAS discovery, Am. J. Hum. Genet. 90 (2012) 7–24.
[21] G.S. Yeo, Where next for GWAS? Brief. Funct. Genomics 10 (2011) 51.
[22] M.L. Freedman, A.N. Monteiro, S.A. Gayther, G.A. Coetzee, A. Risch, C. Plass, G. Casey, M. De Biasi, C. Carlson, D. Duggan, M. James, P. Liu, J.W. Tichelaar, H.G. Vikis, M. You, I.G. Mills, Principles for the post-GWAS functional characterization of cancer risk loci, Nat. Genet. 43 (2011) 513–518.
[23] K. Xia, A.A. Shabalin, S. Huang, V. Madar, Y.H. Zhou, W. Wang, F. Zou, W. Sun, P.F. Sullivan, F.A. Wright, seeQTL: a searchable database for human eQTLs, Bioinformatics 28 (2012) 451–452.
[24] T.P. Yang, C. Beazley, S.B. Montgomery, A.S. Dimas, M. Gutierrez-Arcelus, B.E. Stranger, P. Deloukas, E.T. Dermitzakis, Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies, Bioinformatics 26 (2010) 2474–2476.
[25] The Genotype-Tissue Expression (GTEx) project, Nat. Genet. 45 (2013) 580–585.
[26] A.A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations, Bioinformatics 28 (2012) 1353–1358.
[27] T. Huang, Y.-D. Cai, An information-theoretic machine learning approach to expression QTL analysis, PLoS ONE 8 (2013) e67899.
[28] O. Raaschou-Nielsen, Z.J. Andersen, R. Beelen, E. Samoli, M. Stafoggia, G. Weinmayr, B. Hoffmann, P. Fischer, M.J. Nieuwenhuijsen, B. Brunekreef, W.W. Xun, K. Katsouyanni, K. Dimakopoulou, J. Sommar, B. Forsberg, L. Modig, A. Oudin, B. Oftedal, P.E. Schwarze, P. Nafstad, U. De Faire, N.L. Pedersen, C.G. Ostenson, L. Fratiglioni, J. Penell, M. Korek, G. Pershagen, K.T. Eriksen, M. Sorensen, A. Tjonneland, T. Ellermann, M. Eeftens, P.H. Peeters, K. Meliefste, M. Wang, B. Bueno-de-Mesquita, T.J. Key, K. de Hoogh, H. Concin, G. Nagel, A. Vilier, S. Grioni, V. Krogh, M.Y. Tsai, F. Ricceri, C. Sacerdote, C. Galassi, E. Migliore, A. Ranzi, G. Cesaroni, C. Badaloni, F. Forastiere, I. Tamayo, P. Amiano, M. Dorronsoro, A. Trichopoulou, C. Bamia, P. Vineis, G. Hoek, Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE), Lancet Oncol. 14 (2013) 813–822.
[29] B.J. Lee, B. Kim, K. Lee, Air pollution exposure and cardiovascular disease, Toxicol. Res. 30 (2014) 71–75.
[30] Urban air pollution linked to birth defects, J. Environ. Health 65 (2002) 47–48.
[31] C.A. Hansen, A.G. Barnett, B.B. Jalaludin, G.G. Morgan, Ambient air pollution and birth defects in Brisbane, Australia, PLoS ONE 4 (2009) e5408.
[32] L.C. Vinikoor-Imler, J.A. Davis, R.E. Meyer, T.J. Luben, Early prenatal exposure to air pollution and its associations with birth defects in a state-wide birth cohort from North Carolina, birth defects research. Part A, Clin. Mol. Teratol. 97 (2013) 696–701.
[33] Xinhua, China imposes air quality targets, China.org.cn, 2014.
[34] Y. Zheng, F. Liu, H.-P. Hsieh, U-Air: when urban air quality inference meets big data, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Chicago, IL, USA, 2013, pp. 1436–1444.
[35] Y. Zheng, X. Chen, Q. Jin, Y. Chen, X. Qu, X. Liu, E. Chang, W.-Y. Ma, Y. Rui, W. Sun, A cloud-based knowledge discovery system for monitoring fine-grained air quality, in preparation, Microsoft Tech Report, http://research.microsoft.com/apps/pubs/default.aspx?id=210795, 2014.
[36] S. Mei, H. Li, J. Fan, X. Zhu, C.R. Dyer, Inferring air pollution by sniffing social media, in: IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM), 2014, pp. 534–539.
[37] R. Honicky, E.A. Brewer, E. Paulos, R. White, N-smarts: networked suite of mobile atmospheric real-time sensors, in: Proceedings of the Second ACM SIGCOMM Workshop on Networked Systems for Developing Regions, ACM, Seattle, WA, USA, 2008, pp. 25–30.
[38] X. Chen, Y. Zheng, Y. Chen, Q. Jin, W. Sun, E. Chang, W.-Y. Ma, Indoor air quality monitoring system for smart buildings, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, Seattle, Washington, 2014, pp. 471–475.
[39] J. Nielsen, M.C. Jewett, Metabolomics A Powerful Tool in Systems Biology, Springer, Heidelberg, 2007.
[40] M. Baker, Metabolomics: from small molecules to big ideas, Nat. Methods 8 (2011) 117–121.
[41] K. Suhre, C. Meisinger, A. Doring, E. Altmaier, P. Belcredi, C. Gieger, D. Chang, M.V. Milburn, W.E. Gall, K.M. Weinberger, H.W. Mewes, M. Hrabe de Angelis, H.E. Wichmann, F. Kronenberg, J. Adamski, T. Illig, Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting, PLoS ONE 5 (2010) e13953.
[42] J. Lu, G. Xie, W. Jia, Metabolomics in human type 2 diabetes research, Front. Med. 7 (2013) 4–13.
[43] T. Ramirez, M. Daneshian, H. Kamp, F.Y. Bois, M.R. Clench, M. Coen, B. Donley, S.M. Fischer, D.R. Ekman, E. Fabian, C. Guillou, J. Heuer, H.T. Hogberg, H. Jungnickel, H.C. Keun, G. Krennrich, E. Krupp, A. Luch, F. Noor, E. Peter, B. Riefke, M. Seymour, N. Skinner, L. Smirnova, E. Verheij, S. Wagner, T. Hartung, B. van Ravenzwaay, M. Leist, Metabolomics in toxicology and preclinical research, ALTEX 30 (2013) 209–225.
[44] R.M. Salek, K. Haug, P. Conesa, J. Hastings, M. Williams, T. Mahendraker, E. Maguire, A.N. Gonzalez-Beltran, P. Rocca-Serra, S.A. Sansone, C. Steinbeck, The MetaboLights repository: curation challenges in metabolomics, Database, J. Biol. Databases Curation 2013 (2013), bat029.
[45] I. Baxter, Ionomics: the functional genomics of elements, Brief. Funct. Genomics 9 (2010) 149–156.
[46] B. Lahner, J. Gong, M. Mahmoudian, E.L. Smith, K.B. Abid, E.E. Rogers, M.L. Guerinot, J.F. Harper, J.M. Ward, L. McIntyre, J.I. Schroeder, D.E. Salt, Genomic scale profiling of nutrient and trace elements in Arabidopsis Thaliana, Nat. Biotechnol. 21 (2003) 1215–1221.
[47] L. Sun, Y. Yu, T. Huang, P. An, D. Yu, Z. Yu, H. Li, H. Sheng, L. Cai, J. Xue, M. Jing, Y. Li, X. Lin, F. Wang, Associations between ionomic profile and metabolic abnormalities in human population, PLoS ONE 7 (2012) e38845.
[48] R.M. Bell, Y. Koren, Lessons from the Netflix prize challenge, SIGKDD Explor. 9 (2007) 75–79.
[49] J.K. Laurila, D. Gatica-Perez, I. Aad, B.J.O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, The mobile data challenge: big data for mobile computing research, in: Pervasive Computing, 2012, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.300.1092.
[50] A. van Heerden, S. Norris, S. Tollman, L. Richter, M.J. Rotheram-Borus, Collecting maternal health information from HIV-positive pregnant women using mobile phone-assisted face-to-face interviews in Southern Africa, J. Med. Internet Res. 15 (2013) e116.
[51] S. Zhang, Q. Wu, M.H. van Velthoven, L. Chen, J. Car, I. Rudan, Y. Zhang, Y. Li, R.W. Scherpbier, Smartphone versus pen-and-paper data collection of infant feeding practices in rural China, J. Med. Internet Res. 14 (2012) e119.
[52] A. Sadilek, H. Kautz, V. Silenzio, Predicting disease transmission from geo-tagged micro-blog data, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[53] A. Sadilek, H. Kautz, V. Silenzio, Modeling spread of disease from social interactions, in: Sixth AAAI International Conference on Weblogs and Social Media (ICWSM), 2012, http://www.cs.rochester.edu/~kautz/papers/Sadilek-Kautz-Silenzio_Modeling-Spread-of-Disease-from-Social-Interactions_ICWSM-12.pdf.
[54] F. Ben Abdesslem, I. Parris, T. Henderson, Reliable online social network data collection, in: A. Abraham (Ed.), Computational Social Networks, Springer, London, 2012, pp. 183–210.
[55] M. Stempniak, Beyond buzzwords: two state hospital associations collaborate around big data, Hosp. Health Netw. 88 (2014) 18.
[56] E.M. Bahassi, P.J. Stambrook, Next-generation sequencing technologies: breaking the sound barrier of human genetics, Mutagenesis 29 (2014) 303–310.
[57] G.R. Abecasis, D. Altshuler, A. Auton, L.D. Brooks, R.M. Durbin, R.A. Gibbs, M.E. Hurles, G.A. McVean, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.
[58] A user's guide to the encyclopedia of DNA elements (ENCODE), PLoS Biol. 9 (2011) e1001046.
[59] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, The cancer genome Atlas pan-cancer analysis project, Nat. Genet. 45 (2013) 1113–1120.
[60] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, Bigtable: a distributed storage system for structured data, in: OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, 2006.
[61] Aspera, Mission Possible: DVIDS well armed with Aspera for toughest military content distribution, in: Defense Video & Imagery Distribution System

(DVIDS), 2014, http://downloads.asperasoft.com/en/ecosystem/customer_showcase_1/Defense_Video_Imagery_Distribution_System_DVIDS_42 (accessed on 2014/9/20).
[62] N.R. Council, Frontiers in Massive Data Analysis, The National Academies Press, Washington, DC, 2013.
[63] K. Shvachko, K. Hairong, S. Radia, R. Chansler, The hadoop distributed file system, in: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010, pp. 1–10.
[64] S. Ghemawat, H. Gobioff, S.-T. Leung, The Google file system, SIGOPS Oper. Syst. Rev. 37 (2003) 29–43.
[65] J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clusters, Commun. ACM 51 (2008) 107–113.
[66] G. Linden, B. Smith, J. York, Amazon.com recommendations: item-to-item collaborative filtering, IEEE Internet Comput. 7 (2003) 76–80.
[67] Y. Koren, Tutorial on recent progress in collaborative filtering, in: Proceedings of the 2008 ACM Conference on Recommender Systems, ACM, Lausanne, Switzerland, 2008, pp. 333–334.
[68] B.N. Miller, I. Albert, S.K. Lam, J.A. Konstan, J. Riedl, MovieLens unplugged: experiences with an occasionally connected recommender system, in: Proceedings of the 8th International Conference on Intelligent User Interfaces, ACM, Miami, Florida, USA, 2003, pp. 263–266.
[69] D. Billsus, C.A. Brunk, C. Evans, B. Gladish, M. Pazzani, Adaptive interfaces for ubiquitous web access, Commun. ACM 45 (2002) 34–38.
[70] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (2005) 734–749.
[71] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, ACM, Chapel Hill, North Carolina, USA, 1994, pp. 175–186.
[72] M. Balabanović, Y. Shoham, Fab: content-based, collaborative recommendation, Commun. ACM 40 (1997) 66–72.
[73] J.S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., Madison, Wisconsin, 1998, pp. 43–52.
[74] A. Nakamura, N. Abe, Collaborative filtering using weighted majority prediction algorithms, in: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 1998, pp. 395–403.
[75] D. Billsus, M.J. Pazzani, Learning collaborative information filters, in: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 1998, pp. 46–54.
[76] N.M. Seel, Encyclopedia of the Sciences of Learning, Springer-Verlag, New York, 2012.
[77] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1627–1645.
[78] Z. Song, Q. Chen, Z. Huang, Y. Hua, S. Yan, Contextualizing object detection and classification, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2011, pp. 1585–1592.
[79] D. Parikh, C.L. Zitnick, Finding the weakest link in person detectors, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2011, pp. 1425–1432.
[80] P. Sinno Jialin, Y. Qiang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359.
[81] M. Balasubramanian, E.L. Schwartz, The isomap algorithm and topological stability, Science 295 (2002) 7.
[82] P. Gehler, S. Nowozin, On feature combination for multiclass object classification, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 221–228.
[83] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[84] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (2009) 1–127.
[85] G. Hinton, A practical guide to training restricted Boltzmann machines, in: G. Montavon, G. Orr, K.-R. Müller (Eds.), Neural Netw., Tricks Trade, vol. 7700, Springer, Berlin, Heidelberg, 2012, pp. 599–619.
[86] R. Salakhutdinov, G. Hinton, Deep Boltzmann machines, in: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), 2009, pp. 448–455.
[87] L.C. Freeman, Centrality in social networks: conceptual clarification, Soc. Netw. 1 (1979) 215–239.
[88] T. Huang, W. Cui, Z.S. He, L. Hu, F. Liu, T. Wen, Y. Li, Y. Cai, Functional association between influenza A (H1N1) virus and human, Biochem. Biophys. Res. Commun. 390 (2009) 1111–1113.
[89] T. Huang, J. Zhang, Z.P. Xu, L.L. Hu, L. Chen, J.L. Shao, L. Zhang, X.Y. Kong, Y.D. Cai, K.C. Chou, Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches, Biochimie 94 (2012) 1017–1025.
[90] T. Huang, P. Wang, Z.Q. Ye, H. Xu, Z. He, K.Y. Feng, L. Hu, W. Cui, K. Wang, X. Dong, L. Xie, X. Kong, Y.D. Cai, Y. Li, Prediction of deleterious non-synonymous SNPS based on protein interaction network and hybrid properties, PLoS ONE 5 (2010) e11900.
[91] Y. Wu, R. Jing, L. Jiang, Y. Jiang, Q. Kuang, L. Ye, L. Yang, Y. Li, M. Li, Combination use of protein–protein interaction network topological features improves the predictive scores of deleterious non-synonymous single-nucleotide polymorphisms, Amino Acids 46 (2014) 2025–2035.
[92] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.
[93] G.D. Bader, C.W. Hogue, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinform. 4 (2003) 2.
[94] C.G. Rivera, R. Vakil, J.S. Bader, NeMo: network module identification in cytoscape, BMC Bioinform. 11 (Suppl. 1) (2010) S61.
[95] S. Oliver, Guilt-by-association goes global, Nature 403 (2000) 601–603.
[96] A.L. Barabasi, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (2011) 56–68.
[97] L. Hu, T. Huang, X. Shi, W.C. Lu, Y.D. Cai, K.C. Chou, Predicting functions of proteins in mouse based on weighted protein–protein interaction network and protein hybrid properties, PLoS ONE 6 (2011) e14556.
[98] T. Huang, X.H. Shi, P. Wang, Z. He, K.Y. Feng, L. Hu, X. Kong, Y.X. Li, Y.D. Cai, K.C. Chou, Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks, PLoS ONE 5 (2010) e10972.
[99] T. Huang, Z. Xu, L. Chen, Y.D. Cai, X. Kong, Computational analysis of HIV-1 resistance based on gene expression profiles and the virus–host interaction network, PLoS ONE 6 (2011) e17291.
[100] T. Huang, J. Wang, Y.D. Cai, H. Yu, K.C. Chou, Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma, PLoS ONE 7 (2012) e34460.
[101] B.Q. Li, T. Huang, L. Liu, Y.D. Cai, K.C. Chou, Identification of colorectal cancer related genes with mRMR and shortest path in protein–protein interaction network, PLoS ONE 7 (2012) e33393.
[102] T. Huang, L. Liu, Q. Liu, G. Ding, Y. Tan, Z. Tu, Y. Li, H. Dai, L. Xie, The role of Hepatitis C Virus in the dynamic protein interaction networks of hepatocellular cirrhosis and carcinoma, Int. J. Comput. Biol. Drug Des. 4 (2011) 5–18.
[103] E.W. Dijkstra, A note on two problems in connexion with graphs, Numer. Math. 1 (1959) 269–271.
[104] B.Q. Li, J. Zhang, T. Huang, L. Zhang, Y.D. Cai, Identification of retinoblastoma related genes with shortest path in a protein–protein interaction network, Biochimie 94 (2012) 1910–1917.
[105] Y.M. Lee, L. An, F. Liu, R. Horesh, Y.T. Chae, R. Zhang, Applying science and mathematics to big data for smarter buildings, Ann. N.Y. Acad. Sci. 1295 (2013) 18–25.
[106] N. Zhang, M. Jiang, T. Huang, Y.D. Cai, Identification of Influenza A/H7N9 virus infection-related human genes based on shortest paths in a virus–human protein interaction network, Biomed. Res. Int. 2014 (2014) 239462.
[107] P.E. Hart, N.J. Nilsson, B. Raphael, A formal basis for the heuristic determination of minimum cost paths, IEEE Trans. Syst. Sci. Cybern. 4 (1968) 100–107.
[108] M. Jiang, Y. Chen, Y. Zhang, L. Chen, N. Zhang, T. Huang, Y.D. Cai, X. Kong, Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein–protein interaction network, Mol. BioSyst. 9 (2013) 2720–2728.
[109] K. Macropol, T. Can, A.K. Singh, RRW: repeated random walks on genome-scale protein networks for local cluster discovery, BMC Bioinform. 10 (2009) 283.
[110] Y. Li, J.C. Patra, Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network, Bioinformatics 26 (2010) 1219–1224.
[111] S. Kohler, S. Bauer, D. Horn, P.N. Robinson, Walking the interactome for prioritization of candidate disease genes, Am. J. Hum. Genet. 82 (2008) 949–958.
[112] R. Jiang, M. Gan, P. He, Constructing a gene semantic similarity network for the inference of disease genes, BMC Syst. Biol. 5 (Suppl. 2) (2011) S2.
[113] X. Chen, M.X. Liu, G.Y. Yan, Drug-target interaction prediction by random walk on the heterogeneous network, Mol. BioSyst. 8 (2012) 1970–1978.
[114] H. Shi, J. Xu, G. Zhang, L. Xu, C. Li, L. Wang, Z. Zhao, W. Jiang, Z. Guo, X. Li, Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes, BMC Syst. Biol. 7 (2013) 101.
[115] J. Wang, S. Zhang, Y. Wang, L. Chen, X.S. Zhang, Disease-aging network reveals significant roles of aging genes in connecting genetic diseases, PLoS Comput. Biol. 5 (2009) e1000521.
[116] R. Ihakaa, R. Gentleman, R: a language for data analysis and graphics, J. Comput. Graph. Stat. 5 (1996) 299–314.
[117] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S.J. Jones, M.A. Marra, Circos: an information aesthetic for comparative genomics, Genome Res. 19 (2009) 1639–1645.

[118] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: International AAAI Conference on Weblogs and Social Media, 2009, http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[119] P.T. Shannon, M. Grimes, B. Kutlu, J.J. Bot, D.J. Galas, RCytoscape: tools for exploratory network analysis, BMC Bioinform. 14 (2013) 217.

[120] H. Zhang, P. Meltzer, S. Davis, RCircos: an R package for Circos 2D track plots, BMC Bioinform. 14 (2013) 244.