Dinov *GigaScience* (2016) 5:12
DOI 10.1186/s13742-016-0117-6

GigaScience

REVIEW Open Access

Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data

Ivo D. Dinov

Abstract

Managing, processing and understanding big healthcare data is challenging, costly and demanding. Without a robust fundamental theory for representation, analysis and inference, a roadmap for uniform handling and analyzing of such complex data remains elusive. In this article, we outline various big data challenges, opportunities, modeling methods and software techniques for blending complex healthcare data, advanced analytic tools, and distributed scientific computing. Using imaging, genetic and healthcare data we provide examples of processing heterogeneous datasets using distributed cloud services, automated and semi-automated classification techniques, and open-science protocols. Despite substantial advances, new innovative technologies need to be developed that enhance, scale and optimize the management and processing of large, complex and heterogeneous data. Stakeholder investments in data acquisition, research and development, computational infrastructure and education will be critical to realize the huge potential of big data, to reap the expected information benefits and to build lasting knowledge assets. Multi-faceted proprietary, open-source, and community developments will be essential to enable broad, reliable, sustainable and efficient data-driven discovery and analytics. Big data will affect every sector of the economy and their hallmark will be 'team science'.

Keywords: Big data, Analytics, Modeling, Information technology, Cloud services, Processing, Visualization, Workflows

Background

This article outlines some of the known barriers, intellectual and computational challenges, and opportunities in the area of big healthcare data (BHD). A blend of 'team science', open-source developments, engagement of diverse communities, innovative education and hands-on training will be essential to advance the field of biomedical research [1]. Technical problems, substantial resource costs, and the intellectual demands of handling, processing and interrogating BHD are barriers to advancement and progress. At present, a canonical framework for representation, analysis and inference that is based on incongruent, multi-source and multi-scale biomedical data does not exist. After two decades of rapid computational advances, a tsunami of data and substantial scientific discoveries, urgent unmet needs remain for (near) real-time predictive data analytics, (semi) automated decision support systems and scalable technologies for extracting valuable information, deriving actionable knowledge and realizing the huge potential of BHD.

The pillars of complexity science in healthcare include the diversity of health-related ailments (disorders) and their co-morbidities, the heterogeneity of treatments and outcomes and the subtle intricacies of study designs, analytical methods and approaches for collecting, processing and interpreting healthcare data [2]. In general, BHD has complementary dimensions - large size, disparate sources, multiple scales, incongruences, incompleteness and complexity [3]. No universal protocol currently exists to model, compare or benchmark the performance of various data analysis strategies. BHD sizes can vary, although complexity studies frequently involve hundreds to thousands of individuals, structured and unstructured data elements, and metadata whose volume can be in the 'mega-giga-tera' byte range. Such data often arise from multiple sources and can have many different scales,

Correspondence: statistics@umich.edu
Statistics Online Computational Resource (SOCR), Health Behavior and Biological Sciences, Michigan Institute for Data Science, University of Michigan, 426 N. Ingalls, Ann Arbor, MI 49109, USA

which makes modeling difficult. Finally, the complexity of the data formats, representations, sampling incongruences and observation missingness further complicates the data analysis protocols [4].

There are four phases in the analysis of BHD. The first phase is always to recognize the complexity of the process and understand the structure of the observed data as its proxy. Next comes the representation of BHD that should accommodate effective data management and computational processing. The last two phases of BHD analytics involve data modeling (including embedding biomedical constraints) and inference or interpretation of the results.

Innovative scientific techniques, predictive models and analytics need to be developed to interrogate BHD and gain insight about patterns, trends, connections and associations in the data. Owing to the unique characteristics of BHD, studies relying on large and heterogeneous data trade off the importance of traditional hypothesis-driven inference and statistical significance with computational efficiency, protocol complexity and methodological validity.

Strategies, techniques and resources

Structured and unstructured BHD

A key component of the complexity of BHD is the fact that most of the data is often unstructured, which means that in their raw format they are mostly qualitative or incongruent; this lack of congruence effectively stifles the ability to computationally process BHD [5, 6]. Examples of such unstructured data include raw text (such as clinical notes), images, video, volumetric data, biomedical shape observations, whole-genome sequences, pathology reports, biospecimen data, etc. Text mining [7], image or sequence analysis [8] and other preprocessing techniques [9, 10] need to be used to give structure to this unstructured raw data, extract important information or generate quantitative signature vectors. For example, text preprocessing can use statistical parsing [11], computational linguistics [12, 13] and machine learning [14] to derive meaningful numerical summaries. Information extraction approaches, such as entity recognition [15], relation extraction [16], and term frequency and inverse document frequency techniques [17, 18], provide mechanisms to extract structured information from unstructured text. Figure 1 shows an example of text parsing and semantic interpretation of clinical notes to obtain structured data elements that enable subsequent quantitative processing and statistical inference.

In the past decade, a sustained effort has been made to develop data standards, controlled vocabularies and ontologies for structural or semantic representations of data and metadata [19–22]. Specific examples of successful representation platforms for biomedical and healthcare data include minimum information standards. Examples of such standards include minimum information for biological and biomedical investigations (MIBBI) [23], minimum information about a microarray experiment (MIAME) [24], minimum information requested in the annotation of biochemical models (MIRIAM) [25], and core information for metabolomics reporting (CIMR) [26]. Examples of effective solutions and data standards developed and supported by various consortia include investigation/study/assay (ISA) [27], Clinical Data Interchange Standards Consortium (CDISC) [28], proteomics mass spectrometric data format (mzML)

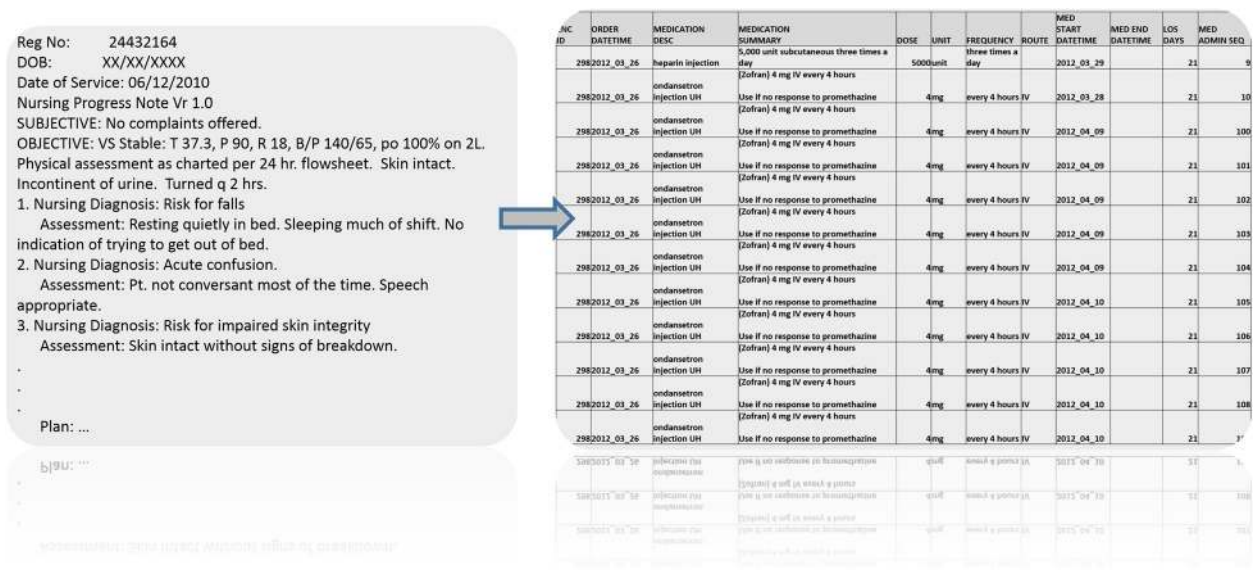


Fig. 1 An example of parsing and interpreting unstructured clinical notes (left) and deriving structured data elements (right)

[29], and the nuclear magnetic resonance spectroscopy for metabolomics data markup language (nmrML) [27]. Powerful controlled vocabularies enable annotation, integration and servicing of millions of names, concepts and meta-data (e.g. diseases, conditions, phenotypes), and their relationships, in dozens of biomedical vocabularies, such as medical subject headings (MeSH) [30], gene ontology (GO) [31], and systematized nomenclature of medicine-clinical terms (SNOMED CT) [32]. Finally, there is a broad spectrum of domain-specific biomedical modeling standards, such as predictive model markup language (PMML) [33], XML format for encoding biophysically based systems of ordinary differential equations (CellML) [34], systems biology markup language (SBML) [35, 36], neural open markup language (NeuroML) [37] and tumor markup language for computational cancer modeling (TumorML) [38]. These architectures enable mathematical modeling and representation of biological constraints, and also promote machine-learning applications through the use of meta-learning schemes, data mining, boosting or bagging [39]. In a similar way, imaging, volumetric and shape-based observations can be preprocessed (e.g. by application of inhomogeneity correction [40], surface modeling [41], feature segmentation [42], etc.) to generate simpler biomedical morphometry measures, or biomarkers, that can be used as proxies of the raw unstructured data [43–46]. In general, summarizing data involves extractive or abstractive approaches for attaining structured information that is computationally tractable. Natural language processing (NLP) [47] is commonly used in healthcare, finance, marketing and social research as an abstractive summarization or a classification technique. Audio analytics (e.g. large-vocabulary continuous speech recognition) [48, 49] provide a mechanism for preprocessing and analyzing unstructured speech or sound data to facilitate subsequent extraction of structured information. Similarly, video content analysis (VCA) [50] can be used to monitor, analyze and extract summary information from live or archived video streams. In addition, such video analytics provide a valuable tool for longitudinal surveying, monitoring and tracking objects in 3D scenes.

Graph networks

Social media applications, biomedical and environmental sensors, and municipal and government services provide enormous volumes of data that can carry valuable information. However, the informational content of such data might be hidden from plain view, entangled or encoded, which obfuscates the extraction of structured data and their interpretation in the networking context in which they were acquired. Content-based social analytics [51] focus on user-provided data in diverse social media platforms, wearables, apps and web services. Social data are always voluminous, unstructured, noisy, dynamic, incomplete and often inconsistent. In addition to the rudimentary challenges of managing such complex data, researchers encounter problems related to continuous software updates, technological advances (e.g. wearables), web server patches and product feature changes occurring during social studies.

Social network analytics [52] aim to harmonize, aggregate and synthesize structural attributes by using automated (unsupervised) [53] or semi-supervised algorithms [54] for data processing, discovery of relationships, or pattern extraction [55] among the participating social data entities. Social network modeling represents the data as a set of nodes (observations) and edges (relations between observations) that reflect the study participants and the associations within the network. Activity networks are a type of social graphs in which the nodes are either data elements or cases (participants) and the edges represent the actual interactions between pairs of nodes. Examples of interactions include dependencies (causal or relational) in which active relationships might be directly relevant to the network analysis. Social graphs are an alternative in which edges connecting pairs of nodes only signify the existence of a loose connection or weak link between the corresponding entities. Social graphs are useful to identify communities, clusters, cohorts or hubs. In scale-rich graphs, the connections between the nodes are uniformly random. Whereas in scale-free networks, the distribution of degrees of connectedness follows a power law with the increase in the number of nodes. Several powerful graphing methods exist for rendering, interrogating and visualizing complex network data [56–59]. Two network visualization examples are shown in Fig. 2.

Community discovery graph methods [60, 61] facilitate the implicit extraction of harmonious subgraphs within a network. Similar to clustering, community detection provides the means to summarize large networks, uncover intrinsic patterns or behaviors and predict critical properties of the network [62, 63]. Graph-based data mining can be used to partition networks into disjointed subgraphs (subnetworks, or hubs) on the basis of node similarity or distance measures. To model, evaluate and understand the influence of various nodes (actors) or edges (relations) in a social network we can use social influence analysis [64, 65]. As actions and behaviors of individuals within a social network affect others to varying degrees, assessing the joint influence of all participants on the entire community provides quantitative information about the strength of the network connections [66]. Social influence analysis captures the importance of nodes in the network and the stability, dynamics and efficiency of the entire social biosphere, and enables the modeling of influence diffusion through the network. Examples of specific approaches include linear threshold modeling and independent cascade modeling [67]. Various quantitative measures describing

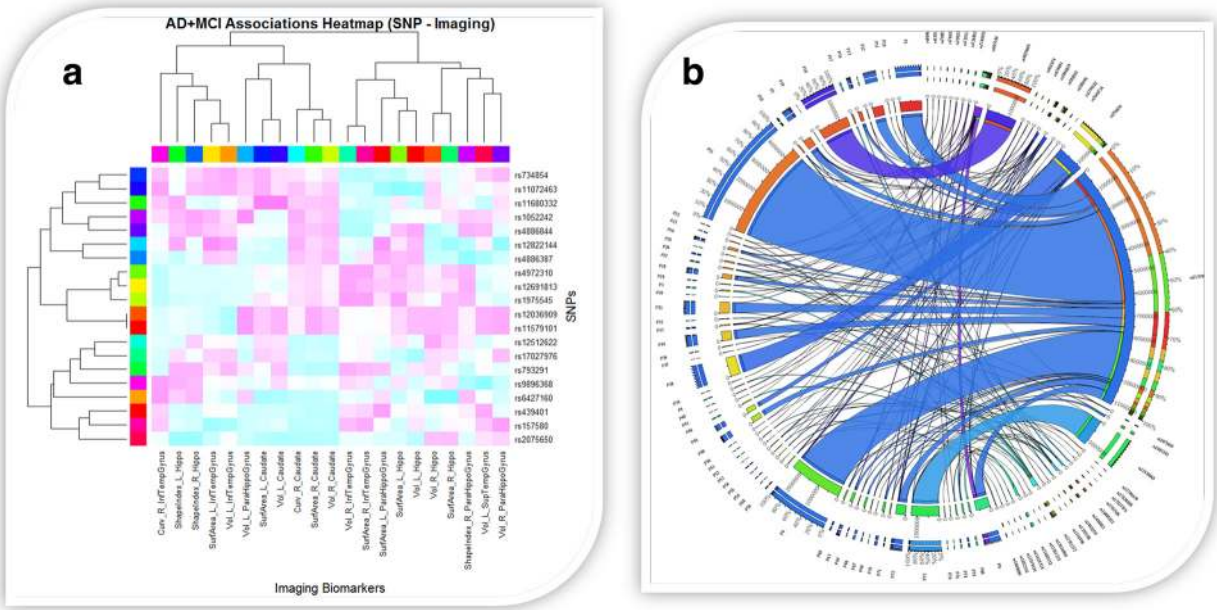


Fig. 2 Examples of rendering complex network data. **a** A heatmap of neuroimaging-derived measures associated with individual phenotypes and genotypes [193]. **b** A Circos connectogram showing the associations (types and strengths) between genomics (single nucleotide polymorphisms) and neuroimaging (morphometry measures of brain regions) biomarkers [204]

the social network characteristics can be defined [68]. Examples include measures of centrality (e.g. degree, betweenness, closeness, eigenvector or Katz centrality), graph distance measures (e.g. graph distance matrix, vertex eccentricity, graph radius), transitivity (e.g. graph reciprocity, global clustering coefficient, mean clustering coefficient), similarity (e.g. mean neighbor degree, mean degree connectivity, vertex dice similarity), etc. [69–71].

An important problem in social network research is predicting prospective linkages between the existing nodes in the graph network [72, 73]. The structure of social networks is mostly dynamic and continuously morphs with the creation of new or destruction and modification of existing nodes or edges. Understanding the internal network organization might enable the prediction of the dynamics or evolution of the network. Naturally observed networks, such as the internet, social networks, air-transportation networks and metabolomics networks, frequently share similar structural properties [74]. They are scale-free (with the fraction of network nodes with k connections to other nodes following asymptotically a power law, $P(k) \sim k^{-\gamma}$, for large k, with a power parameter typically $2 < \gamma < 3$) [75], and exhibit small-world features (all nodes, even non-neighbors, can be reached from every other node through a short sequence of steps. The six degrees of separation theory suggests that a chain of friendships between people can be made to connect any two humans in a maximum of six connections [76]. For example, network link prediction aims to estimate the chance of an interaction between entities and assess the influence among nodes in the network at a prospective time point [72]. Link prediction can also be used to examine associations in networks and to develop network decision support systems [77]. Network medicine is another example of a successful graph theoretic application [78], which uses functional interdependencies between cellular and molecular components to examine disease networks in situations in which several genes, multiple intracellular interactions and various tissue and/or organ systems jointly explain human pathology. Such networks enable the systematic exploration of molecular, environmental and genetic complexity for specific disease pathways and phenotypes.

Classification

A plethora of algorithms, techniques and software tools are available for automated or semi-automated segmentation, clustering and classification of complex data [79–81]. Unsupervised machine-learning methods can be used to uncover patterns (or item sets) in numeric or categorical multivariate data [82, 83]. Bayes belief networks enable prediction, classification and imputation of missing values, and can be used to generate network representations of conditional dependencies among a large number of variables [84]. Deep learning is useful for complex unlabeled datasets and encapsulates machine-learning algorithms for organizing the data hierarchically and exposing the most important features, characteristics and explanatory

variables as high-level graph nodes [85]. Ensemble methods combine the results from many different algorithms that vote in concert to generate increasingly accurate estimates. Compared with the results of any single algorithm or technique across the space of all possible datasets, ensemble methods provide highly effective predictive outputs [86]. Single-class classifiers are based on logistic regression and enable us to assess whether a data point belongs to a particular class. These classifiers can be useful in studies involving multiple cohorts in which the research interest is in identifying only one of many possible outcomes [87–89].

Gaussian mixture modeling (GMM) represents an unsupervised learning technique for data clustering that uses expectation maximization to generate a linear mixture of clusters of the full dataset on the basis of univariate Gaussian (normal) distribution models for each cluster [90, 91]. Fig. 3 illustrates an example of using GMM to dynamically segment a 3D structural brain volume image into white matter, gray matter and cerebrospinal fluid. GMM algorithms typically output sets of cluster attributes (means, variances and centroids) for each cluster that enable us to quantify the differences and similarities between different cohorts. Random forests represent a family of decision-tree classification methods that produce a 'forest of trees' representing alternative models by iteratively randomizing one input variable at a time and learning whether the randomization process actually produces a more or less accurate classification result [92]. When the results are less or more optimal, compared to the results of the previous iteration(s), the variable is either removed from, or included into, the model at the next iteration, respectively.

K-nearest neighbors (kNN) classification algorithms [93–95] include the *K*-means methods for data clustering [96] and *K*-itemsets techniques [97] for association mining. These iterative methods partition a given dataset into a fixed user-specified number of clusters, *K*, which can be used to identify outliers as well as index, search, or catalog high-dimensional data. The local linear embedding method [98] is an example of a manifold learning method that aims to discover real, yet low-dimensional, topological shapes or patterns in the data [99]. Globally, the Euclidian representations of such shape manifolds can be warped and twisted. However, their intrinsic metric is locally homeomorphic to a lower-dimensional Euclidean distance measure [100]. For instance, consider the embedding in 3D of the 2D manifold representing the cortical surface of the human brain [101]. Cortical activation can be difficult to examine in 3D (because of the topology of the cortical surface); however, using the 2D manifold coordinates we can represent activation as data attributes anchored at vertices on the cortical surface. Another example is 3D data that live on a complex 2D hyperplane representing the linear associations of three variables representing the three natural base coordinates of the data [102, 103].

The different machine-learning (or statistical-learning) methods [104] are divided into supervised approaches (in which the goal is to use a training set that includes already classified data to draw inference or classify prospective, testing, data) [105] and unsupervised approaches (whose main task is to identify structure, such as clusters, in unlabeled data) [106]. Semi-supervised learning-based classification methods attempt to balance performance and precision using small sets of labeled or annotated data and a much larger unlabeled data collection [107]. Support vector machines (SVM) are powerful supervised machine-learning techniques for data classification [108] that use binary linear classification. SVM partition data vectors into classes on the basis of *a priori* features of the training data. SVM operate by constructing an optimal hyperplane (i.e. a maximum-margin hyperplane in a transformed feature vector space) that divides the high-dimensional dataset into

Fig. 3 Example of using expectation maximization and Gaussian mixture modeling to classify stereotactic neuroimaging data [91]

two subspaces to maximize the separation of the clusters (for example, normal versus pathological cases). Boosting machine-learning methods create highly accurate prediction rules by combining many weak and inaccurate rules, associations or affinities detected in a (large) dataset [14, 109]. Adaptive boosting is one example in which the algorithm iteratively exploits misclassified examples from previous learning iterations and assigns them higher weights in the next round, which explains the adaptive influence, or iterative re-weighting, that is the signature feature of this method [110].

As the complexity of machine-learning algorithms can increase exponentially with the volume of the data, alternative model-based techniques, like generalized linear models (GLMs), may be more appropriate as they are computationally efficient and applicable for classifying extremely large datasets, [111, 112]. Using parallel processing [113], bootstrap sampling [114] and algorithm optimization [112, 115] can substantially improve the efficiency of all machine-leaning methods [116]. Compared with learning-based classification methods, such as SVM and boosting, the efficiency of GLMs in analyzing big data is rooted in their more simplistic linear modeling and regression estimation that make use of observed explanatory variables to predict the corresponding outcome response variable(s).

Examples of unsupervised quantitative data exploration and data mining algorithms for unlabeled datasets include association mining [117], link analysis [118], principal or independent component analyses (PCA/ICA) [119, 120] and outlier detection [102]. PCA projects high-dimensional data into a subspace of reduced dimension spanned by a family of orthonormal principal component vectors that maximize the residual variance not already present in the previous components. In practice, mutual orthogonality of the principal components might be a too strong assumption. Additionally, PCA relies on second-order statistics to estimate the covariances between the observed variables, which implies that the features that are generated might only be sensitive to second-order effects. Correlation-based learning algorithms such as PCA are designed to account for the amplitude spectra of data but largely ignore their phase spectra. This might limit their ability to characterize datasets with informative features that are modeled by higher-order statistics (e.g. skewness, kurtosis, etc.). ICA provides linear models for non-Gaussian data by generating components that are statistically independent. ICA model representations use blind source separation to capture the core structure of the data, which facilitates feature extraction and cohort separation. ICA is computationally efficient and applicable for data mining problems involving recovering statistically independent features from data assumed to represent unknown linear mixtures of attributes.

Association mining represents another class of machine-learning algorithms applicable to large categorical data. This approach is mostly focused on discovering frequently occurring coherent associations among a collection of variables and aims to identify such associations on the basis of their frequencies of co-occurrence relative to random sampling of all possibilities. Link analysis aims to assign class labels to data elements on the basis of various link characteristics derived from iterative classification, relaxation labeling or other methods. Using link-based distance measures between entries we can generate associations expressing relative quantitative assessments of the between-element link associations in the entire dataset, extrapolate these patterns as network links, deduce novel plausible links and mine the collection. Many outliner detection methods exist for quantitative or qualitative detection of measurement errors, atypical observations, abnormal values or critical events [121].

Incompleteness

Missing data arise in most complex data-driven inquiries [122]. To handle incomplete data, knowledge about the cause of missingness is critical [123]. If data are missing completely at random (MCAR), the probability of an observation being missing is the same for all entities [124]. In these situations, throwing out cases with missing data does not bias the final scientific inference. However, if the pattern of data missingness is not completely at random, such as when non-response rates are different in different subpopulations, the probability of observing an entity might be variable and we need to model, impute or correct for the missing values to obtain unbiased inference. We can model the process of missingness via logistic regression, in which the outcome variable equals 1 for observed cases or 0 for unobserved entities. When an outcome variable is missing at random (MAR), we can still exclude the missing cases as unobserved; however, the regression model should control for all the variables that affect the probability of missingness (e.g. object characteristics or subject demographics) [125]. Another common cause for incomplete data is missingness that depends on some specific unobserved predictors. Missingness not at random (MNAR) suggests that the incompleteness of the data depends on information that is not available, i.e., unobserved information may predict the missing values [126]. For instance, an aggressive cancer intervention can have side effects that make patients more likely to discontinue the treatment. Side effects and 'discomfort' associated with an intervention can be difficult to measure, which can lead to incomplete data due to MNAR. In such cases, we have to explicitly model the incompleteness of the data to avoid inferential bias. In certain situations, missingness can depend on the unobserved entity itself, that is, the probability of missingness depends on the

missing variable [127]. For example, if younger adults are less likely to enroll in healthcare plans, case censoring may be in effect due to aging and we must account for the related missing-data by including more predictors in the missing-data model – that is, bring the process of missingness closer to MAR.

Exploratory data analytics

Countless examples show the equivalence of a 'word' to a 'thousand pictures' [128] and its pseudo-converse that equates a 'picture' to a 'thousand words' [129]. Protocols for image parsing to text description (I2T) generate text from still images (or video streams) [130]. Conversely, exploratory data analytics transform text (tables) into figures (images) that represent a synthesized view of the information contained in the ASCII data. This duality of representation of complex information is also directly demonstrated by the homology between time-space and frequency (Fourier) representations of multidimensional data [131, 132]. Visual exploratory and explanatory analytics are critical components of any study of complex data. Such tools facilitate the graphical 'storytelling' of the properties and characteristics leading to, or explaining, BHD discoveries.

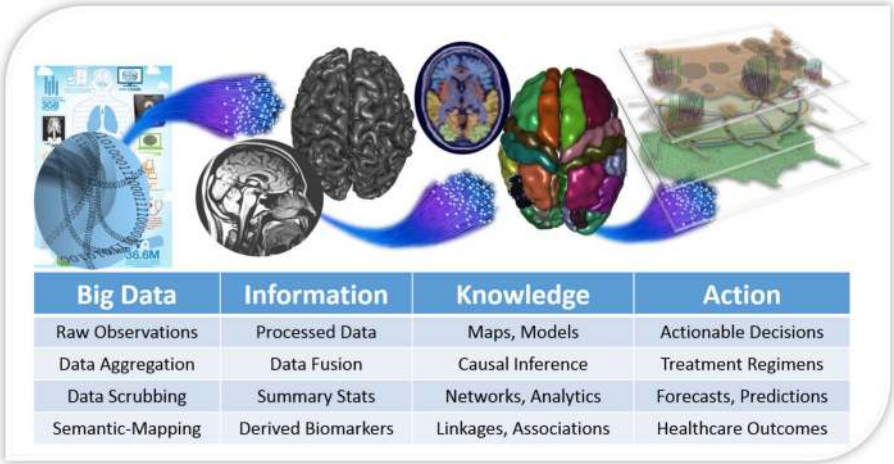
Data profiling is a collection of exploratory data analytic methods that facilitates quick and effective identification of some basic data characteristics [133]. Profiling evaluates the information content, intrinsic structure and quality of the data and explores variable relationships within them. Examining frequency distributions of different data elements provides insight into the type, center, spread and shape of each variable. Cross-variable analysis can also expose embedded value dependencies and discover overlapping or correlated features among the entities. Motion charts [134] are an interactive mechanism for mapping variables to different graphical widgets, which facilitates the dynamic traversal (playing the chart) across a time dimension. Typically, motion charts facilitate on-the-fly transformation of quantitative and qualitative information contained in multivariate data to expose relevant and actionable knowledge about the interplays among multiple data elements. ManyEyes data visualization [135] enables users to generate graphical displays of their own data. Socrata [136] enables the servicing and sharing of dynamic data via a user-friendly and cost-effective interface. D3 is a modern JavaScript platform for developing dynamic data visualizations. The Cytoscape visualization suite [56] enables exploration of network and tabular data. Several dashboard platforms exist (e.g. Tableau [137], SOCR MotionCharts [134] and SOCR Dashboard [138]) for interrogation of complex, structured or unstructured multisource data. Data Wrangler [139] includes mechanisms for manipulating, transforming, filtering and visualizing incongruent data.

Choosing the right statistical methodology

In terms of selecting appropriate statistical tests, the most important question is: 'What are the main study hypotheses and specific goals?' In some cases no *a priori* testable hypothesis exists; the investigator just wants to 'see what is there'. For example, in a study investigating the prevalence of a disease, there is no hypothesis to test, and the size of the study is determined by how accurately the investigator wants to determine prevalence. If no hypothesis exists, then no corresponding statistical test are conducted. It is important to decide *a priori* which hypotheses are confirmatory (that is, whether we are testing some presupposed relationship), and which are exploratory (whether they are suggested by the data). No single study can support a whole series of hypotheses. There are a number of strategies to determine the most appropriate statistical tests and often alternative approaches need to be investigated. As there is no unique, complete, and consistent ontological hierarchy to guide practitioners, consultations with experts are useful. An example of a table of frequently used study designs and appropriate corresponding statistical analysis approaches is available online [140].

Predictive analytics

Large and complex clinical datasets require data-specific and study-specific analytic protocols for managing raw data, extracting valuable information, transforming the information to knowledge, and enabling clinical decision-making and action that are evidence-based (Fig. 4) [138]. Various methods exist to predict future outcomes or forecast trends using retrospective and current data. Predictive analytics are useful in all scientific inquiries or research explorations. Anticipating future failures or systemic changes using multi-source data streams that generate hundreds or thousands of data points is critical in decision-making, whether when buying a stock, preparing for natural disasters, forecasting pandemics, projecting the course of normal or pathological aging or anticipating the behavior of social groups. Predictive analytics aim to uncover patterns and expose critical relations in phenomena using the associations between data elements detected in the observed process. Two generic types of predictive analytics techniques exist: model-based or model-free. Predictive time series analyses can use moving averages to build a model using historical or training data and extrapolate the trend predicted by the model into the future. Multivariate regression methods [141, 142] represent variable interdependencies between predictors and responses in terms of some base functions (e.g. polynomials) whose coefficients capture the influence of all variables on the outcomes and facilitate forward predictions. Alternatively, machine-learning techniques [143], classification theory [144] and network analytics [145, 146] can be used for



Big Data	Information	Knowledge	Action
Raw Observations	Processed Data	Maps, Models	Actionable Decisions
Data Aggregation	Data Fusion	Causal Inference	Treatment Regimens
Data Scrubbing	Summary Stats	Networks, Analytics	Forecasts, Predictions
Semantic-Mapping	Derived Biomarkers	Linkages, Associations	Healthcare Outcomes

Fig. 4 A schematic illustrating the big healthcare data analytic pipeline in a neuroscientific context, including data management, mapping, processing, interpretation and inference [138]

model-free (semi) unsupervised data mining, hierarchical clustering [147], pattern recognition [148], fuzzy clustering [149] or trend identification [150]. The type of outcome variables affects the analytical techniques used to study the process. For example, multilinear regression [151] is applicable for analyzing continuous outcome variables, whereas random forest classification [92] and logistic regression [152] can be applied to analyze discrete outcome variables.

Contemporary data science and analytic research demand innovative predictive forecasting and statistical methods that are capable of dealing with the complexity of big data that are prevalent in biomedical studies [153, 154]. Classical statistical methods are based on conventional complete data and specific *a priori* statistical significance assumptions. Scientific inference often depends on small data samples from a specific population with some assumptions on their distribution. To examine the significance of a particular relationship, statistical results are typically contrasted against random chance. Finally, data-driven findings might be generalized as a conclusion applied to the entire (unobserved) population. There are substantial differences in the sample attributes of traditional studies and big data studies. The latter are characterized by incompleteness, incongruency, multi-source elements, multiple scales, excessive heterogeneity, and enormous size. Big data samples frequently represent a substantial fraction of the entire population [155, 156]. This process trades off exactness and stability with completeness and consistency of the proxy observations. Thus, in BHD studies, the classical notion of statistical significance morphs into scientific inference that is based on joint modeling of all elements of big data using exploratory, classification, and pattern-tracking methods. Other essential distinctions exist between standard statistical analysis methods and advanced data analytics techniques [157]. Computational efficiency, data management, validation and reproducibility need Big-Data-specific, agile and scalable algorithms and models to obtain reliable inference on complex and heterogeneous data. The heterogeneity [158], noise concentration [3], spurious correlations [159], incidental endogeneity (hidden correlations between data elements and error terms) [160], and variable latency [161] that characterize big data also demonstrate the major challenges associated with handling, modeling and information extraction of BHD.

Data heterogeneity reflects the unavoidable differences in population characteristics, data formatting and type variability [162]. Big data always include heterogeneous data elements where small sub-samples might capture specific cohorts that include outliers or extreme data. An important property of big data that makes them useful is the population coverage of the data, asymptotically with the increase of the sample size. This enables us to model, stratify, and understand the heterogeneity of multiple sub-cohorts in the population. At the same time, noise concentration may creep in due to the aggregation of heterogeneous data elements and the accumulation of individual error terms into the joint big data analysis. Developing predictive big data analytic models requires simultaneous estimation of multiple parameters, model coefficients or likelihoods. In this joint processing, error estimates might compile (noise aggregation can be linear or non-linear in terms of the number of variables) and thus dominate the variable effect sizes or obfuscate the true effect of a parameter included in the model.

Spurious effects refer to data elements that are not associated in reality but that, owing to data complexity, are falsely determined to be significantly correlated [163]. For example, correlation coefficients between independent

random variables can increase with the increase of the data size, incongruences in noise levels or the presence of latent variable effects. Another important factor in all Big Data analytic studies is the 'curse of dimensionality', which arises in dealing with high-dimensional data. This paradox is not present in traditional low-dimensional datasets. In high-dimensions many numerical analyses, data sampling protocols, combinatorial inferences, machine learning methods, or data managing processes are susceptible to the 'curse of dimensionality'. Increases of data dimensionality (including a larger number of data elements) leads to parallel, and faster, increases of the space volume containing the observed data, thus, the actual points of data into the high-dimensional space appear to be drifting apart (distances between data points increases). The sparsity between points, even for big data, affects all quantitative analytic methods, as the corresponding statistical inference depends explicitly on the stability of 'distance' metrics [164]. The reliability of the statistical inference relies on balancing the volume of data (number of observation points) that needs to grow exponentially with the number of dimensions in which the data are embedded. In a high-dimensional space, objects may appear to be farther apart and artificially dissimilar, which affects data structuring, organization, modeling and inference. However, in big data studies, this problem of increased dimensionality and the associated challenges of interpreting data from multiple sources trades off with the potential for reduced bias, increased level of unique and heterogeneous population characteristics captured and broader interpretation of results.

Incidental endogeneity is a property that violates the common regression technique assumption that requires the independent (explanatory) variables to be independent of the error term (model residuals) [159]. Many parametric statistical methods depend on this assumption, as presence of incidental endogeneity allows potentially strong dependences between some predictors and the residuals that render the techniques possibly unreliable or underpowered. In traditional studies involving standard datasets the exogeneity assumption is usually met, that is, no acute incidental endogeneities occur. However, in BHD analyses, the expectation is that incidental endogeneity may be ubiquitous [165]. A difference exists between spurious effects and incidental endogeneity: the former refers to pseudo-random relationships, whereas the latter refers to natural intrinsic associations between the explanatory variables and the model residual error term.

Data harmonization and fusion

When interpreting the information content of large and heterogeneous data, the processes of extraction of patterns, trends and associations demand considerable insights, computational power and analytical tools. Raw and derived data might come from multiple unrelated sources, and latent effects or multivariate correlations might complicate data interrogation. Traditional databases have bottlenecks in ingesting, retrieving and processing vast amounts of heterogeneous data. Modern structured query language (SQL) and NoSQL databases [166, 167], platforms for extract-transform-load processing [168] and cloud-based services [169–171] are improving the human and machine interfaces to BHD. Incongruent data often arrive from disparate digital sources, which can represent orthogonal, co-linear, or causally related information. Solid foundation for analytical and computational representation of big data is important. Alternative data representation schemes, canonical models or reference frameworks that facilitate data harmonization and integration across different granularity scales, encoding protocols, measurement types, phenotypes and formats are being developed [172, 173]. In practice, data incongruity can be due to the lack of such a common data representation architecture. Incompatibility of data elements is ubiquitous and unavoidable in most studies of real health data that rely on data-driven inference or evidence-based decision-making. Variable transformations, data imputations, low-dimensional modeling, and joint analyses all depend on a common scheme for effective representation of complex BHD. The implicit data harmonization necessary to enable subsequent data integration and processing is predicated on successful wrangling and fusion of incongruous data elements.

Services and infrastructure

The MapReduce model from Google provides an attractive mechanism for parallel processing and *ad hoc* inference for large and heterogeneous datasets [174, 175]. A pair of functions, a mapper and a reducer, split real-world computational tasks (e.g. data cleaning, modeling, machine learning, filtering, aggregation, merging, etc.) into manageable scalable pieces that can be independently completed in parallel using separate parts of the (Big) data. These tasks could be performed on separate (but connected) machines even under failing node conditions. Hadoop is an open-source implementation of MapReduce [176]. The open-source Apache http://spark.apache.org/ enables distributed computing for large and complex datasets. Spark and MapReduce are linearly scalable and fault-tolerant; however, Spark can be up to 100 times faster for certain applications and provides rich and intuitive machine interfaces (e.g. application program interfaces in Python, Java, Scala and R) to support data abstraction and a wide spectrum of computing-intensive tasks, interactive queries, streaming, machine learning and graph processing.

PMML [177] is an XML-based language for describing, assembling and sharing predictive models learned within a data mining process that facilitates computational

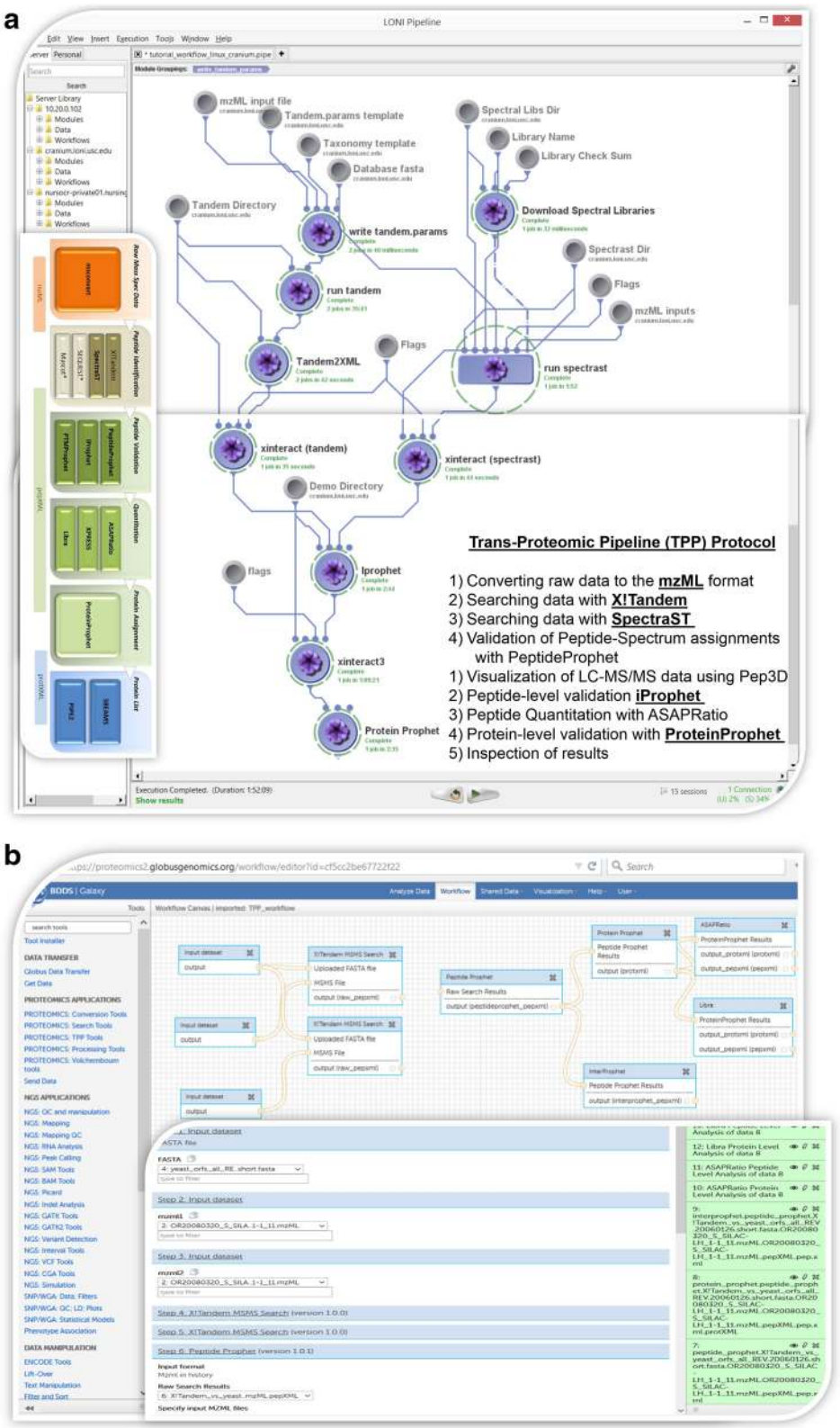


Fig. 5 Two alternative end-to-end cloud-based solutions to the Trans-Proteomic Pipeline (TPP) protocol that apply advanced data modeling, processing and visualization methods to process mass spectroscopy datasets using multiple cloud web services. **a** Implementation of the TPP computational protocol in the Pipeline environment. **b** Implementation of the TPP computational protocol in the Globus Galaxies environment [205]

processing (machine-to-machine communication and distributed manipulation). DataMining-as-a-Service (DMaaS) [178], DecisionScience-as-a-Service (DSaaS) [179], Platform-as-a-Service (PaaS) [180], Infrastructure-as-a-Service (IaaS) [181] and Software-as-a-Service (SaaS) [182] are all examples of cloud-based data, protocol and infrastructure services enabling reliable, efficient and distributed data analytics. R packages [124, 147], KNIME [183], WEKA [184], RapidMiner [185] and Orange [186] include hundreds of powerful open-source algorithms and software tools for high-throughput machine learning, data mining, exploration, profiling, analytics and visualization.

Figure 5 provides an example of a high-throughput end-to-end computational protocol in which several of such cloud web services are used. This example illustrates the implementation of the Institute for Systems Biology Trans-Proteomic Pipeline (TPP), which applies advanced data modeling, processing and visualization to the search and process datasets using multiple engines [187]. The dual Pipeline-based and Galaxy-based solutions are alternative service-oriented protocols that yield the same results using vastly different computational platforms. Many similar examples that use the Imaging Data Archive services [188, 189], Parkinson's Progression Markers Initiative services [190, 191], Galaxy computational services [192], Pipeline client-server infrastructure [45, 193, 194] and proteomics services [195] are available online [196, 197].

Various national and international big data science initiatives have emerged as a response to sizeable financial support from government agencies, philanthropic organizations and industry partners to develop platforms enabling 'open-science', data sharing, collaborative development and transdisciplinary engagement. For example, in the USA, the National Institutes of Health funded 11 National big data to Knowledge Centers (BD2K) [198] and several satellite BD2K activities. In Europe, the Virtual Physiological Human initiative [199], the European Life-sciences Infrastructure for Biological Information [200] and the Translational Information & Knowledge Management Services [201] have secured resources to build and use open-source translational data, tools and services (e.g. tranSMART [202]) to tackle challenging problems.

Conclusions

In the biomedical and healthcare community, managing, processing and understanding BHD pose substantial challenges that parallel enormous opportunities in understanding human conditions in health and disease, across location, time, and scale. Although no unique blueprint or perfect roadmap exist, the characteristics of the data, the underlying model assumptions, the computational infrastructure demands, and the application scope all have vital roles in the choices about how to guide, handle and analyze such complex data. The field of Big-Data-driven research discoveries bridges various scientific disciplines, advanced information and communication technologies, and multiple sources, and is rapidly evolving. We have outlined big data challenges, identified big data opportunities and presented modeling methods and software techniques for blending complex healthcare data and contemporary scientific approaches. We give examples of several techniques for processing heterogeneous datasets using cloud services, advanced automated and semi-automated techniques and protocols for open-science investigations. New technologies are still necessary to improve, scale and expedite the handling and processing of large data that are increasing in size and complexity [193]. At the same time, substantial methodological progress, powerful software tools and distributed service infrastructure are already in place to enable the design, simulation and productization of the future computational resources necessary to support the expected avalanche of data [203]. Big data analytics are likely to encounter some setbacks and some great advances in the next decade. Additional public, private and institutional investments in data acquisition, research and development, and computational infrastructure, along with education, will spur the involvement of bright young minds to tackle the huge big data challenges, reap the expected information benefits and assemble knowledge assets. Balancing proprietary, open-source and community commons developments will be essential for broad, reliable, sustainable and efficient development efforts. The influence of big data will go beyond financing, high-tech and biomedical research. Big data will be likely to touch every sector of the economy and their signature feature will be rapid on-demand team science.

Abbreviations

BHD: big healthcare data; GMM: Gaussian mixture modeling; ICA: independent component analysis; PCA: principal component analysis; SVM: support vector machines.

Competing interests

The author declares that he has no competing interests.

Acknowledgments

This work was partially supported by National Science Foundation grants 1416953, 0716055 and 1023115, and by the National Institutes of Health grants P20 NR015331, U54 EB020406, P50 NS091856 and P30 DK089503. Many colleagues (from the Statistics Online Computational Resource, the Big Data for Discovery Science Center and the Michigan Institute for Data Science) have provided encouragement and valuable suggestions. Journal editorial comments and reviewer's critiques substantially improved the manuscript.

Received: 8 September 2015 Accepted: 9 February 2016
Published online: 25 February 2016

References

1. Alberts B et al. Rescuing US biomedical research from its systemic flaws. Proc Natl Acad Sci. 2014;111(16):5773–7.
2. McMurty A. Reinterpreting interdisciplinary health teams from a complexity science perspective. Univ Alberta Health Sci J. 2007;4(1):33–42.

3. Bollier D, Firestone CM. The promise and peril of big data. Communications and Society Program. Washington: Aspen Institute; 2010.
4. Dipnall JF et al. Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers. Methods. 2014;69(3):237–46.
5. Caballero, I., M. Serrano, and M. Piattini, A Data Quality in Use Model for Big Data, in Advances in Conceptual Modeling, M. Indulska and S. Purao, Editors. 2014, Springer. p. 65–74.
6. Chen, E.S. and I.N. Sarkar, Mining the Electronic Health Record for Disease Knowledge, in Biomedical Literature Mining. 2014, Springer. p. 269–286.
7. Feldman, R. and J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data. 2006: Cambridge University Press.
8. Almeida JS. Sequence analysis by iterated maps, a review. Brief Bioinform. 2014;15(3):369–75.
9. Chen CP, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inform Sci. 2014;275:314–47.
10. Khan N et al. Big data: survey, technologies, opportunities, and challenges. Scientific World Journal. 2014;2014.
11. Agerri R et al. Big data for Natural Language Processing: A streaming approach. Knowledge-Based Systems. 2014;79:36–42.
12. Wu X, Fan W, Peng J, Zhang K, Yu Y. Iterative sampling based frequent itemset mining for big data. International Journal of Machine Learning and Cybernetics 2015;6(6):875–882.
13. Riezler S. On the problem of theoretical terms in empirical computational linguistics. Computational Linguistics. 2014;40(1):235–45.
14. Alpaydin, E., Introduction to machine learning. 2014: MIT press.
15. Tang Z, Jiang L, Yang L, Li K, Li K. CRFs based parallel biomedical named entity recognition algorithm employing MapReduce framework. Cluster Computing. 2015;18(2):493–505.
16. Gui, F., et al. Social relation extraction of large-scale logistics network based on mapreduce. in Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on. 2014. IEEE.
17. Kim, J., et al., Noise Removal Using TF-IDF Criterion for Extracting Patent Keyword, in Soft Computing in Big Data Processing. 2014, Springer. p. 61–69.
18. Aggarwal, C.C. and C.K. Reddy, Data clustering: algorithms and applications. 2013: CRC Press.
19. Smith B et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251–5.
20. Tenenbaum JD, Sansone S-A, Haendel M. A sea of standards for omics data: sink or swim? J Am Med Inform Assoc. 2014;21(2):200–3.
21. Toga A, Dino ID. Sharing big biomedical data. J Big Data. 2015;2(1):7.
22. Ivanovi M, Budimac Z. An overview of ontologies and data resources in medical domains. Expert Systems Appl. 2014;41(11):5158–66.
23. Taylor CF et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol. 2008;26(8):889–96.
24. Brazma A et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001;29(4):365–71.
25. Novere NL et al. Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol. 2005;23(12):1509–15.
26. Taylor CF. Standards for reporting bioscience data: a forward look. Drug Discov Today. 2007;12(13):527–33.
27. Salek RM, Haug K, Steinbeck C. Dissemination of metabolomics results: role of MetaboLights and COSMOS. GigaScience. 2013;2(8.10):1186.
28. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007;14(6):687–96.
29. Martens L et al. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics. 2011;10(1):R110. 000133.
30. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32 suppl 1:D267–70.
31. Côté RG et al. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics. 2006;7(1):97.
32. Ochs C et al. A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships. J Am Med Inform Assoc. 2015;22(3):628–39.
33. Kliegr, T., S. Vojí , and J. Rauch. Background knowledge and PMML: first considerations. in Proceedings of the 2011 workshop on Predictive markup language modeling. 2011. ACM.
34. Nickerson, D.P., et al., Using CellML with OpenCMISS to simulate multi-scale physiology. Frontiers in bioengineering and biotechnology, 2014. 2(79): p. 10.3389/fbioe.2014.00079.
35. Hucka M et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003;19(4):524–31.
36. Smith LP et al. SBML and CellML translation in Antimony and JSim. Bioinformatics. 2014;30(7):903–7.
37. Cannon RC et al. LEMS: a language for expressing complex biological models in concise and hierarchical form and its use in underpinning NeuroML 2. Frontiers in Neuroinformatics. 2014;8.
38. Johnson, D., J. Cooper, and S. McKeever. TumorML: Concept and requirements of an in silico cancer modelling markup language. in Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. 2011. IEEE.
39. Peng Y, Wang G, Wang H. User preferences based software defect detection algorithms selection using MCDM. Inform Sci. 2012;191:3–13.
40. Irarrazabal P et al. Inhomogeneity correction using an estimated linear field map. Magn Reson Med. 1996;35(2):278–82.
41. Malladi R, Sethian JA, Vemuri BC. Shape modeling with front propagation: A level set approach. Pattern Analysis and Machine Intelligence. IEEE Transactions. 1995;17(2):158–75.
42. Bajaj C, Yu Z, Auer M. Volumetric feature extraction and visualization of tomographic molecular imaging. J Struct Biol. 2003;144(1):132–43.
43. Ashburner J, Friston K. Voxel-based morphometry. 2007.
44. Ho AJ et al. Comparing 3 T and 1.5 T MRI for tracking Alzheimer's disease progression with tensor-based morphometry. Hum Brain Mapp. 2010;31(4): 499–514.
45. Dinov I et al. Neuroimaging Study Designs, Computational Analyses and Data Provenance Using the LONI Pipeline. PLoS One. 2010;5(9):e13070. doi:10.1371/journal.pone.0013070.
46. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage. 2000;11(6):805–21.
47. Chowdhury GG. Natural language processing. Ann Rev Inform Sci Technol. 2003;37(1):51–89.
48. Vacher, M., et al., Development of audio sensing technology for ambient assisted living: Applications and challenges, in Digital Advances in Medicine, E-Health, and Communication Technologies. 2013, IGI Global. p. 148.
49. Huijbregts, M., R. Ordelman, and F. de Jong, Annotation of heterogeneous multimedia content using automatic speech recognition, in Semantic Multimedia. 2007, Springer. p. 78–90.
50. Dimitrova N et al. Applications of video-content analysis and retrieval. IEEE Multimedia. 2002;9(3):42–55.
51. Agrawal, D., et al., Big Data in Online Social Networks: User Interaction Analysis to Model User Behavior in Social Networks, in Databases in Networked Information Systems. 2014, Springer. p. 1–16.
52. Aggarwal, C.C., An introduction to social network data analytics. 2011: Springer.
53. Almeida JS, Prieto CA. Automated unsupervised classification of the Sloan Digital Sky Survey stellar spectra using k-means clustering. Astrophysical J. 2013;763(1):50.
54. Gan H et al. Using clustering analysis to improve semi-supervised classification. Neurocomputing. 2013;101:290–8.
55. Basirat, A., A.I. Khan, and H.W. Schmidt, Pattern Recognition for Large-Scale Data Processing, in Strategic Data-Based Wisdom in the Big Data Era, J. Girard, Editor. 2015, IGI Global. p. 198.
56. Ono K, Demchak B, Ideker T. Cytoscape tools for the web age: D3. js and Cytoscape. js exporters. F1000Research. 2014;3:143–5.
57. Reimann, M., et al., Visualization and Interactive Analysis for Complex Networks by means of Lossless Network Compression, in Computational Network Theory: Theoretical Foundations and Applications, M. Dehmer, F. Emmert-Streib, and S. Pickl, Editors. 2015, John Wiley & Sons.
58. Le Meur, N. and R. Gentleman, Analyzing biological data using R: methods for graphs and networks, in Bacterial Molecular Networks, J. van Helden, A. Toussaint, and D. Thieffry, Editors. 2012, Springer. p. 343–373.
59. Freeman, L.C., Social Network Visualization, in Computational Complexity, R. Meyers, Editor. 2012, Springer. p. 2981–2998.
60. Zhu Z, Wang C, Ma L, Pan Y, Ding Z. Scalable community discovery of large networks. in Web-Age Information Management, 2008. WAIM'08. The Ninth International Conference on. Zhangjiajie: IEEE; 2008.
61. Satuluri, V., S. Parthasarathy, and Y. Ruan. Local graph sparsification for scalable clustering. in Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. 2011. ACM.

62. Newman M. Communities, modules and large-scale structure in networks. Nat Phys. 2012;8(1):25–31.
63. Mitra B, Tabourier L, Roth C. Intrinsically dynamic network communities. Computer Networks. 2012;56(3):1041–53.
64. Abrahamse W, Steg L. Social influence approaches to encourage resource conservation: A meta-analysis. Glob Environ Chang. 2013;23(6):1773–85.
65. Wang C et al. Dynamic social influence analysis through time-dependent factor graphs. in Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. Kaohsiung: IEEE; 2011.
66. Sivakumar B, Woldemeskel FM. A network-based analysis of spatial rainfall connections. Environ Model Software. 2015;69:55–62.
67. Kempe, D., J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003. ACM.
68. Kennedy DP et al. The analysis of duocentric social networks: A primer. J Marriage Fam. 2015;77(1):295–311.
69. Dem ar U, patenková O, Virrantaus K. Identifying critical locations in a spatial network with graph theory. Transactions in GIS. 2008;12(1):61–82.
70. Brandes, U. and T. Erlebach, Network analysis: methodological foundations. Vol. 3418. 2005: Springer Science & Business Media.https://books.google.com/books?id=VIMSPClafakC
71. Berry MW et al. Identifying influential edges in a directed network: big events, upsets and non-transitivity. J Complex Networks. 2014;2(2):87–109.
72. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. J Am Society Inform Scie Technol. 2007;58(7):1019–31.
73. Backstrom, L. and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. in Proceedings of the fourth ACM international conference on Web search and data mining. 2011. ACM.
74. Ostriker JP, Naab T. Theoretical challenges in understanding galaxy evolution. Physics Today. 2012;65(8):43–9.
75. Holme P, Kim BJ. Growing scale-free networks with tunable clustering. Physical Rev E. 2002;65(2):026107.
76. Travers J, Milgram S. An experimental study of the small world problem. Sociometry. 1969;32(4):425–43.
77. Kim, Y. and J. Srivastava. Impact of social influence in e-commerce decision making. in Proceedings of the ninth international conference on Electronic commerce. 2007. ACM.
78. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
79. Chilali O et al. A survey of prostate modeling for image analysis. Comput Biol Med. 2014;53:190–202.
80. Galinsky VL, Frank LR. Automated segmentation and shape characterization of volumetric data. Neuroimage. 2014;92:156–68.
81. Norouzi A et al. Medical image segmentation methods, algorithms, and applications. IETE Tech Rev. 2014;31(3):199–213.
82. Kodratoff, Y. and R.S. Michalski, Machine learning: an artificial intelligence approach. Vol. 3. 2014: Morgan Kaufmann. https://books.google.com/books?hl=en&lr=&id=vHyjBQAAQBAJ
83. Le QV. Building high-level features using large scale unsupervised learning. in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. Vancouver: IEEE; 2013.
84. Henrion, M., et al., Automated construction of sparse Bayesian networks from unstructured probabilistic models, in Uncertainty in Artificial Intelligence 5, R. Shachter, Kanal, LN, Henrion, M, Lemmer, JF, Editor. 2014, Elsevier. p. 295.
85. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2015;61:85–117.
86. Lihu, A. and . Holban, A review of ensemble methods for de novo motif discovery in ChIP-Seq data. Briefings in bioinformatics, 2015: p. doi: 10.1093/bib/bbv022.
87. Khan SS, Madden MG. One-class classification: taxonomy of study and review of techniques. Knowledge Eng Rev. 2014;29(03):345–74.
88. Menahem E, Rokach L, Elovici Y. Combining one-class classifiers via meta learning. in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. New York: ACM; 2013.
89. Lin W-J, Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinform. 2012;14(1):13–26.
90. Tian G et al. Hybrid genetic and variational expectation-maximization algorithm for Gaussian-mixture-model-based brain MR image segmentation. Information Technology in Biomedicine. IEEE Transact. 2011;15(3):373–80.
91. Dinov, I., Expectation Maximization and Mixture Modeling Tutorial. Statistics Online Computational Resource, in UCLA: Statistics Online Computational Resource. 2008 (Accession Date: Jan 15, 2016), UCLA: Los Angeles, CA, http://escholarship.org/uc/item/1rb70972.
92. Rodriguez-Galiano V et al. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogrammetry Remote Sensing. 2012;67:93–104.
93. Denoeux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. Syst Man Cybern IEEE Trans. 1995;25(5):804–13.
94. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. Syst Man Cybern IEEE Transact. 1985;SMC-15(4):580–5.
95. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surveys (CSUR). 1999;31(3):264–323.
96. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010;31(8):651–66.
97. Knobbe, A.J. and E.K. Ho. Maximally informative k-itemsets and their efficient discovery. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. ACM.
98. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290(5500):2323–6.
99. Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci. 2003;100(10):5591–6.
100. Shi Y, Sun B, Lai R, Dinov I, Toga A. Automated sulci identification via intrinsic modeling of cortical anatomy. in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010. Beijing: MICCAI; 2010.
101. Shi Y et al. Direct cortical mapping via solving partial differential equations on implicit surfaces. Med Image Anal. 2007;11(3):207–23.
102. Aggarwal, C.C., Linear Models for Outlier Detection, in Outlier Analysis. 2013, Springer. p. 75–99.
103. Ge SS, He H, Shen C. Geometrically local embedding in manifolds for dimension reduction. Pattern Recogn. 2012;45(4):1455–70.
104. Fritzke B. Growing cell structures—a self-organizing network for unsupervised and supervised learning. Neural Netw. 1994;7(9):1441–60.
105. Caruana, R. and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. in Proceedings of the 23rd international conference on Machine learning. 2006. ACM.
106. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning. 2001;42(1–2):177–96.
107. Cherniavsky, N., et al., Semi-supervised learning of facial attributes in video, in Trends and Topics in Computer Vision, K. Kutulakos, Editor. 2012, Springer. p. 43–56.
108. Hearst MA, Dumais P, Susan T, Osman E, Platt J, Scholkopf B. Support vector machines. Intell Syst Appl IEEE. 1998;13(4):18–28.
109. Vapnik, V. Boosting and Other Machine Learning Algorithms. in Machine Learning Proceedings 1994: Proceedings of the Eighth International Conference. 2014. Morgan Kaufmann.
110. Gavinsky D. Optimally-smooth adaptive boosting and application to agnostic learning. J Machine Learn Res. 2003;4:101–17.
111. McCulloch C. Generalized linear models. Vol. 95. Alexandria: ETATS-UNIS: American Statistical Association; 2000.
112. McCulloch, C., Neuhaus, JM, Generalized linear mixed models, in Encyclopedia of Environmetrics. 2013, John Wiley & Sons.
113. Hwang, K., Dongarra, J, Fox, GC, Distributed and cloud computing: from parallel processing to the internet of things. 2013: Morgan Kaufmann.
114. Wang S, Li Z, Zhang X. Bootstrap sampling based data cleaning and maximum entropy SVMs for large datasets. in Tools with Artificial Intelligence (ICTAI). Athens: IEEE; 2012.
115. Fernández M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res. 2012;40(10):e77–7.
116. He Y et al. Support vector machine and optimised feature extraction in integrated eddy current instrument. Measurement. 2013;46(1):764–74.
117. Zaki MJ. Scalable algorithms for association mining. Knowledge Data Eng IEEE Transact. 2000;12(3):372–90.
118. Lu, Q. and L. Getoor. Link-based classification. in Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). 2003. Washington, DC.
119. Jolliffe, I., Principal component analysis. 2005: Wiley Online Library.
120. Comon P. Independent component analysis, a new concept? Signal Process. 1994;36(3):287–314.

121. van den Boogaart, K.G. and R. Tolosana-Delgado, Zeroes, Missings, and Outliers, in Analyzing Compositional Data with R. 2013, Springer. p. 209–253.
122. Jagadish H, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel Jignesh M, Ramakrishnan R, Shahabi, Cyrus. Big data and its technical challenges. Commun ACM. 2014;57(7):86–94.
123. Little, R.J. and D.B. Rubin, Statistical analysis with missing data. 2014: John Wiley & Sons.
124. Jamshidian M, Jalal SJ, Jansen C. Missmech: an R package for testing homoscedasticity, multivariate normality, and missing completely at random (mcar). J Stat Software. 2014;56(6):1–31.
125. Cheema JR. A Review of Missing Data Handling Methods in Education Research. Rev Educ Res. 2014;84(4):487–508.
126. Moreno-Betancur M, Rey G, Latouche A. Direct likelihood inference and sensitivity analysis for competing risks regression with missing causes of failure. Biometrics. 2015;71(2):498–507.
127. Afrianti, Y., S. Indratno, and U. Pasaribu. Imputation algorithm based on copula for missing value in timeseries data. in Technology, Informatics, Management, Engineering, and Environment (TIME-E), 2014 2nd International Conference on. 2014. IEEE.
128. Doumont J-L. Verbal versus visual: A word is worth a thousand pictures, too. Technical Commun. 2002;49(2):219–24.
129. Pinsky LE, Wipf JE. A picture is worth a thousand words. J Gen Intern Med. 2000;15(11):805–10.
130. Yao BZ et al. I2t: Image parsing to text description. Proceedings IEEE. 2010; 98(8):1485–508.
131. Candès EJ, Wakin MB. An introduction to compressive sampling. Signal Process Magazine, IEEE. 2008;25(2):21–30.
132. Folland, G.B., Fourier analysis and its applications. Vol. 4. 1992: American Mathematical Soc. https://books.google.com/books?id=ix2iCQ-o9x4C
133. Naumann F. Data profiling revisited. ACM SIGMOD Record. 2014;42(4):40–9.
134. Al-Aziz J, Christou N, Dinov I. SOCR Motion Charts: An Efficient, Open-Source, Interactive and Dynamic Applet for Visualizing Longitudinal Multivariate Data. JSE. 2010;18(3):1–29.
135. Viegas FB et al. Manyeyes: a site for visualization at internet scale. Visual Comput Graph IEEE Transact. 2007;13(6):1121–8.
136. Erickson JS et al. Open Government Data: A Data Analytics Approach. IEEE Intell Syst. 2013;28(5):19–23.
137. Nandeshwar, A., Tableau data visualization cookbook. 2013: Packt Publishing Ltd.
138. Husain S, Kalinin A, Truong A, Dinov ID. SOCR Data dashboard: an integrated big data archive mashing medicare, labor, census and econometric information. J Big Data. 2015;2(13):1–18.
139. Kandel, S., Paepcke, A, Hellerstein, J, Heer, J. Wrangler: Interactive visual specification of data transformation scripts. in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011. ACM.
140. SOCR. SOCR Protocol for Choosing Appropriate Statistical Methods. 2016 [cited 2016; Available from: http://socr.umich.edu/Applets/ChoiceOfTest.html
141. Bray, J.H. and S.E. Maxwell, Multivariate analysis of variance. 1985: Sage.
142. McIntosh AR, Mi ic B. Multivariate Statistical Analyses for Neuroimaging Data. Annu Rev Psychol. 2013;64:499–525.
143. Eom J, Zhang B. PubMiner: machine learning-based text mining for biomedical information analysis. Genomics Inform. 2004;2(2):99–106.
144. Friedman, SD, Hyttinen T, Kulikov V. Generalized descriptive set theory and classification theory. American Mathematical Soc. 2014:230(1081). DOI: http://dx.doi.org/10.1090/memo/1081
145. Joshi A, Joshi SH, Leahy RM, Shattuck DW, Dinov I, Toga AW. Bayesian approach for network modeling of brain structural features. in Medical Imaging 2010: Biomedical Applications in Molecular, Structural, and Functional Imaging. San Diego: Proc. SPIE; 2011.
146. Li R et al. Large-scale directional connections among multi resting-state neural networks in human brain: A functional MRI and Bayesian network modeling study. Neuroimage. 2011;56(3):1035–42.
147. Le S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. J Stat Software. 2008;25(1):1–18.
148. Bishop, C.M., Neural networks for pattern recognition. 1995: Oxford University press.
149. en, Z., New Trends in Fuzzy Clustering, in Data Mining in Dynamic Social Networks and Fuzzy Systems, V. Bhatnagar, Editor. 2013, IGI Global. p. 248.
150. Nohuddin PN et al. Trend mining in social networks: from trend identification to visualization. Expert Syst. 2014;31(5):457–68.
151. Harris, R.J., A primer of multivariate statistics. 2014: Psychology Press.
152. Hosmer, D., Lemeshow, S, Sturdivant, RX, Applied logistic regression. 2 ed. 2013: John Wiley & Sons.
153. Bohlouli, M., et al., Towards an integrated platform for big data analysis, in Integration of practice-oriented knowledge technology: Trends and prospectives. 2013, Springer. p. 47–56.
154. Kaisler, S., et al. Big data: Issues and challenges moving forward. in System Sciences (HICSS), 2013 46th Hawaii International Conference on. 2013. IEEE.
155. Leonelli S. What difference does quantity make? On the epistemology of Big Data in biology. Big Data Soc. 2014;1(1):2053951714534395.
156. Pinheiro, C.A.R. and F. McNeill, Heuristics in Analytics: A Practical Perspective of what Influences Our Analytical World. 2014: John Wiley & Sons.
157. Larose, D.T., Discovering knowledge in data: an introduction to data mining. 2014: John Wiley & Sons.
158. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev. 2012;90:61–8.
159. Fan J, Han F, Liu H. Challenges of big data analysis. Nat Sci Rev. 2014;1(2): 293–314.
160. Mathur, A., et al. A new perspective to data processing: Big Data. in Computing for Sustainable Global Development (INDIACom), 2014 International Conference on. 2014. IEEE.
161. Wang, Y. and H. Yu. An ultralow-power memory-based big-data computing platform by nonvolatile domain-wall nanowire devices. in Proceedings of the International Symposium on Low Power Electronics and Design. 2013. IEEE Press.
162. Patiño J et al. Accounting for data heterogeneity in patterns of biodiversity: an application of linear mixed effect models to the oceanic island biogeography of spore-producing plants. Ecography. 2013;36(8):904–13.
163. Anderson DR, Burnham KP, Gould WR, Cherry S. Concerns about finding effects that are actually spurious. Wildlife Society Bulletin. 2001;29(1):311–316.
164. Spinello L, Arras KO, Triebel R, Siegwart R. A Layered Approach to People Detection in 3D Range Data. in Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta: AAAI Press; 2010.
165. Grolinger, K., et al. Challenges for mapreduce in big data. in Services (SERVICES), 2014 IEEE World Congress on. 2014. IEEE.
166. Cattell R. Scalable SQL and NoSQL data stores. ACM SIGMOD Record. 2011; 39(4):12–27.
167. Gudivada V, Rao D, Raghavan VV. NoSQL Systems for Big Data Management. in 2014 IEEE World Congress on Services (SERVICES). Anchorage: AK IEEE; 2014.
168. El Akkaoui, Z., et al. A model-driven framework for ETL process development. in Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP. 2011. ACM.
169. Rimal, B.P., E. Choi, and I. Lumb. A taxonomy and survey of cloud computing systems. in INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on. 2009. Ieee.
170. Baun, C., et al., Cloud computing: Web-based dynamic IT services. 2011: Springer Science & Business Media.
171. Buyya R et al. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Comput Syst. 2009;25(6):599–616.
172. Agarwal, P., G. Shroff, and P. Malhotra. Approximate incremental big-data harmonization. in Big Data (BigData Congress), 2013 IEEE International Congress on. 2013. IEEE.
173. Shroff, G., et al. Prescriptive information fusion. in Information Fusion (FUSION), 2014 17th International Conference on. 2014. IEEE.
174. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107–13.
175. Lämmel R. Google's MapReduce programming model—Revisited. Sci Comput Program. 2008;70(1):1–30.
176. Holmes, A., Hadoop in practice. 2012: Manning Publications Co.
177. Grossman R et al. The management and mining of multiple predictive models using the predictive modeling markup language. Inform Software Technol. 1999;41(9):589–95.
178. Chen, T., J. Chen, and B. Zhou. A System for Parallel data mining service on cloud. in Cloud and Green Computing (CGC), 2012 Second International Conference on. 2012. IEEE.
179. Granville, V., Developing Analytic Talent: Becoming a Data Scientist. 2014: John Wiley & Sons.
180. Ananthakrishnan R, Chard K, Foster I, Tuecke S. Globus platform-as-a-service for collaborative science applications. Concurrency and Computation. Pract Experience. 2014;27(2):290–305.

181. Manvi SS, Shyam GK. Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. J Network Comput Appl. 2014;41:424–40.
182. Allen B et al. Software as a service for data scientists. Commun ACM. 2012;55(2):81–8.
183. Berthold MR et al. KNIME: The Konstanz Information Miner. In: Preisach C et al., editors. Data Analysis, Machine Learning and Applications. Berlin Heidelberg: Springer; 2008. p. 319–26.
184. Hall M et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter. 2009;11(1):10–8.
185. Hofmann, M. and R. Klinkenberg, RapidMiner: Data mining use cases and business analytics applications. 2013: CRC Press.
186. Podpečan V, Zemenova M, Lavrač N. Orange4WS environment for service-oriented data mining. Comput J. 2011;55(1):82–98.
187. Deutsch EW et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics. 2010;10(6):1150–9.
188. Neu, S., Valentino, DJ, Ouellette, KR, Toga, AW. Managing multiple medical image file formats and conventions. in Proceedings of SPIE Medical Imaging 2003:PACS and Integrated Medical Information Systems. 2003. San Diego, CA.
189. Neu S, Valentino DJ, Toga AW. The LONI Debabeler: a mediator for neuroimaging software. Neuroimage. 2005;24(4):1170–9.
190. Frasier M et al. Biomarkers in Parkinson's disease: a funder's perspective. Biomarkers. 2010;4(5):723–9.
191. PPMI. Parkinson's Progression Markers Initiative. [cited 2016; Available from: http://www.PPMI-info.org.
192. Goecks J et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
193. Dinov ID, Pestroyan P, Liu Z, Eggert P, Hobel S, Vespa P, Woo Moon S, Van Horn JD, Franco J and Toga AW. High-Throughput Neuroimaging-Genetics Computational Infrastructure. Frontiers in Neuroinformatics. 2014;8(41):1–11.
194. LONI. The Pipleine Environment. 2016; Available from: http://Pipeline.loni.usc.edu.
195. Slagel J, Mendoza L, Shteynberg D, Deutsch EW, Moritz RL. Processing shotgun proteomics data on the Amazon Cloud with the Trans-Proteomic Pipeline. Mol Cell Proteomics. 2014;14(2):399–404.
196. LONI. Pipeline Library Navigator. 2016; Available from: http://pipeline.loni.usc.edu/explore/library-navigator.
197. Galaxy. The Galaxy Pipeline Project. 2016; Available from: https://GalaxyProject.org.
198. NIH. Big Data to Knowledge (BD2K) Initiative. 2014; Available from: http://BD2K.nih.gov.
199. VHP. Virtual Physiological Human Initiative 2016; Available from: http://www.vph-institute.org.
200. ELIXIR. European Life-sciences Infrastructure for Biological Information 2016; Available from: http://www.ELIXIR-europe.org.
201. eTRIKS. Translational Information & Knowledge Management Services 2016; Available from: http://www.eTRIKS.org.
202. Athey, B., Braxenthaler, M, Haas, M, Guo, Y, tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. AMIA Summits on Translational Science Proceedings, 2013: p. 6–8.
203. Philip Chen CL, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inform Sci. 2014;275:314–47.
204. Moon S, Dinov ID, Zamanyan A, Shi R, Genco A, Hobel S, Thompson, PM, Toga, AW. Alzheimer's Disease Neuroimaging Initiative, Gene Interactions and Structural Brain Change in Early-Onset Alzheimer's Disease Subjects Using the Pipeline Environment. Psychiatry Investigation. 2015;12(1):125–35.
205. Madduri, R., et al., The Globus Galaxies platform: delivering science gateways as a service. Concurrency and Computation: Practice and Experience, 2015. doi:10.1002/cpe.3486.