BIG DATA TECHNOLOGY IN PARTICULAR BIOLOGICAL APPLICATIONS: AN ANALYSIS

Subrata Paul[1] and Shivnath Ghosh[1]*

Department of CSE, Brainware University, Barasat, West Bengal, India

**Corresponding Author. E-mail:shivghosh.cs@gmail.com*

Received on: February 28, 2023 | Accepted on: March 15, 2023 | Published on: March 30, 2023

Abstract

Big data technologies have been employed in a variety of fields since their inception, with their use within biomedical and health-care informatics research enhancing at an astonishing speed. At an extraordinary speed and scale, massive amounts of biological and clinical data have indeed been produced and gathered. Big data applications open up new avenues for discovering new knowledge and developing innovative techniques for enhancing health-care quality. The use of big data in health care is a rapidly expanding field, with numerous recent findings and techniques emerging in recent years. The authors will examine and address big data applications in four major biomedical subfields in this paper: (1) bioinformatics, (2) clinical informatics, and (3) imaging informatics, and (4) public health informatics.

Keywords: *Big data, biomedical, healthcare, clinical informatics, public-health informatics.*

1. Introduction

Biologists are now part of the big-data club. As an outcome of the development of high-throughput genomic information, life researchers have started to deal with enormous amounts of data, experiencing obstacles with managing, handling, and moving information that have been earlier the area of astronomers and high-energy physicists. Big data refers to large or complicated data sets that conventional information processing applications cannot handle. Analysis, capture, data collection, search, communicating, storage, exchange, visualisation, query processing, upgrading, and information privacy are all obstacles that such data sets face. The term commonly applies to the application of predictive analytics or other sophisticated analytics methodologies that retrieve value from information, rather than a specific dimensions of data set [1]. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk. From just a scientific point of view, big data is characterized by the gathering, collection, and evaluation of information sets with an elevated degree of complexity or dimensions. In alongside data volume, the definition of big data includes data variety, velocity, and value. As a result, data handling and evaluation employing conventional methods and instruments is becoming challenging. Access to data and assessment are influenced by information systems (IT). Information is deemed large when its reusability adds to the creation of new insights. Given this, IT has been going to support new scientific models and recently founded interdisciplinary research areas.

Data sets are rising exponentially in component since they are progressively assembled by cheap and innumerable information-sensing mobile devices, aerial (remote sensing), software logs, cams, microphones, radio-frequency identification (RFID) audience, and wireless sensor networks [2] and [3]. Big data typically describes information sets that are larger than the capacity of popular computer tools to grasp, collate, handle, and analyse information in a reasonable amount of time [4]. Big data has increased the demand of information management specialists in that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. Even the usage of big data has been found its extensive use in various biological aspects which we shall be discussing throughout this paper.

Although the algorithms and modelling techniques are comparable, the user interfaces of conventional analytics techniques as well as those in use for big data are significantly different; conventional medical analytics tools have grown to be extremely user-friendly and straightforward. Big data analytics tools, on the contrary hand, are incredibly complicated, necessitating comprehensive programming and the implementation of a diverse set of skills. They surfaced haphazardly, as mostly opensource development platforms and tools, and thus lack the assistance and user-friendliness of vendor-driven patented technology tools. As Figure 1 indicates, the complexity begins with the data itself.

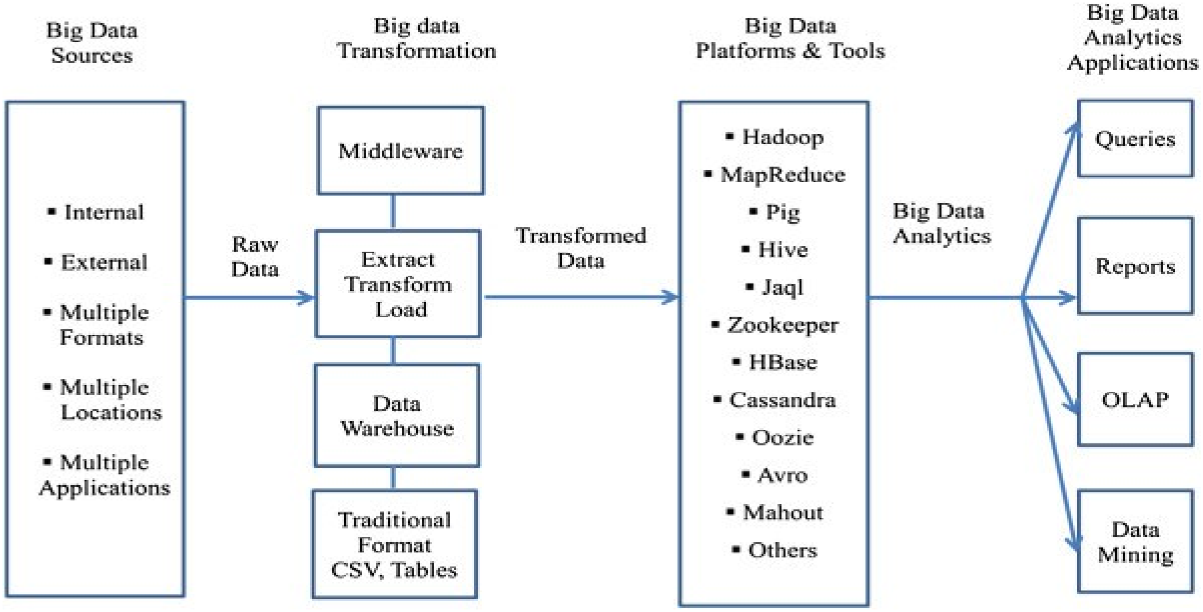

Figure 1 Complexity with the big data

1.1 Challenges in Big Data

This is the point where big data can change the way life sciences research is conducted. In this case, big data can integrate gene sequencing data with relevant proteomic and metabolomic information on one platform. This might appear to be a simple solution to the problem, it is essential to keep in mind that it would necessitate the integration of data from many different sources in such a manner that scientists can effectively analyse and interpret it. Regretfully, there has been an increasing scarcity of technology innovations capable of dealing with the enormous scale and variety of information. Moreover, the big data solution demanded by the biosciences industry must additionally be competent not just of managing the enormous quantity of information currently accessible, additionally maintaining up with the increasing amount of information that would be published each and every day.

Over 200,000 clinical studies are actually ongoing, including 21,000 drug elements, 1,357 unique drugs, 22,000 genes, and hundreds of millions of proteins. There are different types and levels and experimentations inside every one of these research fields that generate a variety of information. Furthermore, over 24 million medical and scientific articles have been published to date, with an approximated 1.8 million fresh publications becoming reported every year.

Chosen to take as a whole, any solitary researcher would struggle to extract all of this information. Researchers are lacking numerous chances to get their hands-on knowledge that might contribute to their individual research endeavours because the average research scientist interprets somewhere around 250 and 300 articles per year.

2. Big data features:

We've all learned of the "three Vs" of big data: volume, variety, and velocity. However, Inderpal Bhandar, Chief Data Officer at Express Scripts, stated during his demonstration at the Big Data Innovation Summit in Boston that there is a few supplementary "Vs" that IT, business, and data scientists must be worried about, most important of which is big data veracity. Validity and volatility are two other big data "Vs" getting praise at the summit. The 6Vs of big data are summarised below.

a. **Volume:** The term "big data" implies massive amounts of data. Employees were previously responsible for data generation. Now since information is produced by equipment, networks, and human interaction on frameworks including social media, the quantity of data that can be analyzed is enormous. However, Inderpal claims that the amount of information is not as problematic as other Vs, such as veracity.

b. **Variety** refers to the numerous data types and sources that exist, both unstructured and structured. Designers used to save information from databases and excel sheets. Data now consists of emails, photos, videos, monitoring equipment, PDFs, audio, and so on. This wide range of unstructured information creates difficulties for information storage, quarrying, and assessment. Jeff Veis, VP Solutions at HP Autonomy presented how HP is helping organizations deal with big challenges including data variety.

c. **Velocity**: The rate upon which data flows in through sources such as business operations, machineries, connections, and social interaction with things such as social media websites, portable devices, etc. is referred to as velocity. The information flow is enormous and constant.

If investigators and business owners can manage the speed, real-time data could indeed assist them in making precious judgements that would provide competitive advantages and return on investment. Inderpal believes that sample selection information may assist with problems such as volume and velocity.

d. Veracity: Big Data Veracity refers to a prejudices, loudness, and irregularities in information. Is the information being storable and extracted relevant to the issue under investigation? Once especially in comparison to items like volume and velocity, Inderpal believes that the most challenging thing throughout data analysis is veracity. When planning one's big data strategy, make certain that your employees and associates are working together to maintain your information clean and your procedures are in place to avoid "dirty data" from accruing in your systems.

e. Validity: The question of validity, such as the question of the veracity of big data, depends on whether the information is reliable and precise for such intentional usages. Trying to make rational decisions necessitates the use of obviously real information. Phil Francisco, VP of Product Management at IBM, has spoken about IBM's big data strategy as well as the techniques they provide for assisting with data veracity and validity.

f. Volatility: The term "big data volatility" relates as far as how lengthy data is legitimate and ought to be stored. In this world of instantaneous data, users need to ascertain as to what point the information ceases to be important for the present analysis.

Big data clearly deals with issues beyond volume, variety and velocity to other concerns like veracity, validity and volatility which is clearly depicted in details in Figure 2[5]. The figure also shows the levels and volume of data as distributed in the three different features.

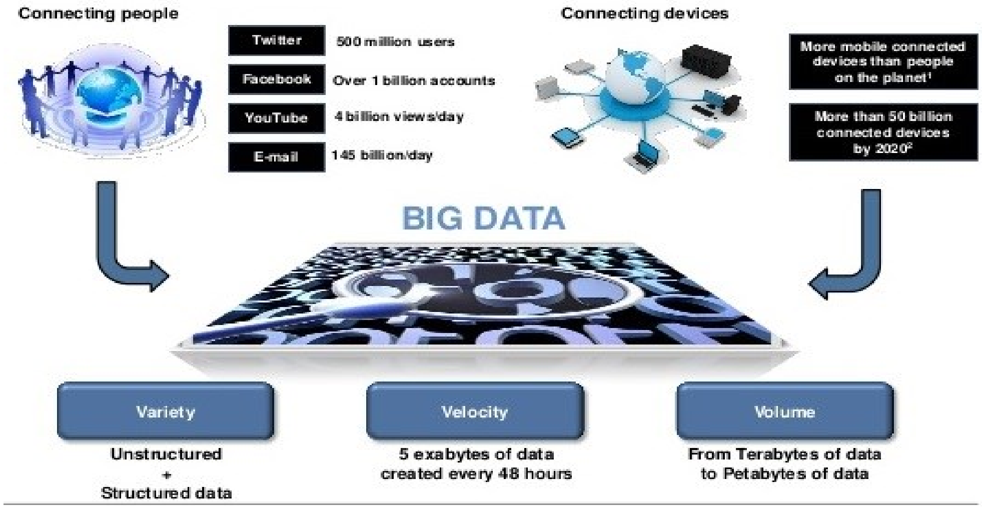
Connecting people
Twitter
500 million users
Facebook
Over 1 billion accounts
YouTube
4 billion views/day
E-mail
145 billion/day
Connecting devices
More mobile connected devices than people on the planet¹
More than 50 billion connected devices by 2020²
BIG DATA
Variety
Unstructured + Structured data
Velocity
5 exabytes of data created every 48 hours
Volume
From Terabytes of data to Petabytes of data

Figure 2 A brief overview on big-data features.

3. Features of Big Data and sparsest solution in high confidence set

3.1 Heterogeneity

Even though individual characteristics are large, Big Data enhances our capacity for identifying similarities in an inhabitant. One instance is whether drinking a glass of wine lessens the likelihood of contracting particular illnesses. When there is a lot of numerical data noise, the population constructions can get entombed. Despite this, big sample sizes allow statistical methods to uncover such hidden patterns. [6].

3.2 Computation

Large-scale estimation is critical in big data analysis. High-dimensional enhancement isn't only costly, but also instability in data processing, with slow convergence. Because of instability and computation cost, techniques involving incremental distortions of large matrices are impractical. Modular and steady high-dimensional numerical process deployments must be decided to seek. This strongly depends on mathematical instinct, large-scale monitoring, and fine-grained optimization [7].

3.3 Spurious correlation

High dimensionality is characterised by spurious connection. It pertains to parameters which aren't conceivably associated but have a large sample correlation [8].

3.4 Incidental endogeneity

Extraneous association is also caused by high dimensionality. Contributing factors that are roughly comparable to the responding are collected by researchers. Since there are numerous predictor variables, a few of the variables can indeed be coincidently associated with the remaining noise. This can result in prototype inaccuracy and inaccurate gene or SNP selection for recognising molecular pathways or biological affiliations [9].

3.5 Noise accumulation

When a system depends on the assessment of distinct variables, residuals can add up. Sound accrual is more drastic in high-dimensional statistical data, and it may ultimately overwhelm the fundamental signals [10].

4. Big data analytics in healthcare

The volume of health data is likely to rise significantly in the coming years [11]. Furthermore, healthcare reimbursement frameworks are evolving, with purposeful use as well as performance-based payments arising as critical technology considerations in the current healthcare setting. Even though profit isn't really and shouldn't be the key motive, healthcare providers must obtain the methods available, infrastructural facilities, and methodologies to effective and efficient marketing big data or potentially lose millions of dollars in revenue and profits [12].

4.1 Advantages to healthcare

Healthcare organisations that range from single-physician office spaces and multi-provider associations to major hospital networks and accountable care organisations can benefit significantly by digitising, incorporating, and successfully utilising big data [13]. Possible advantages involve trying to detect diseases previously in their progression, so they may be handled more readily and successfully; trying to manage particular health of individuals and populations; and designed to detect fraud in healthcare more quickly and effectively. Innumerable issues are possible with big data analytics. Definite advancements or consequences, such as length of stay (LOS), patients who are going to select elective surgery, patients who might likely not gain from surgery, and problems, can be anticipated and/or guesstimated relying on immense quantities of historical information, patients at danger for serious conditions; people at risk for sepsis, MRSA, C. difficile, and perhaps other hospital-acquired ailments; ailment or progression of the disease; patients at risk for disease state progression; ailment or increased infection possible causes; and potential co-morbid circumstances (EMC Consulting).

Figure 3 shows the various ways in which big data can be used in the healthcare industry. It finds its applications right from step 1 i.e split to the final step i.e theme or keyword analysis.

Figure 4 have presented an in-detail explanation of the solutions being provided by big-data in various areas.

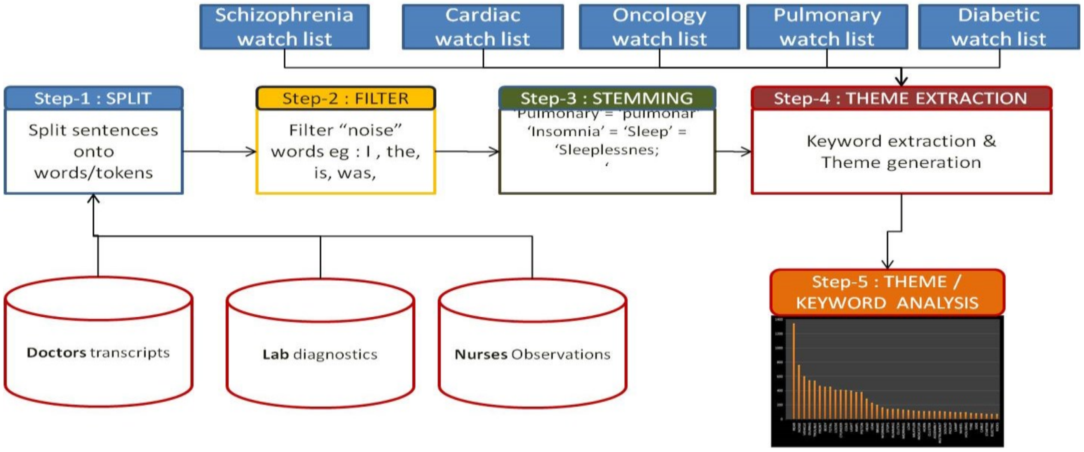

Figure 3 Healthcare use case diagram

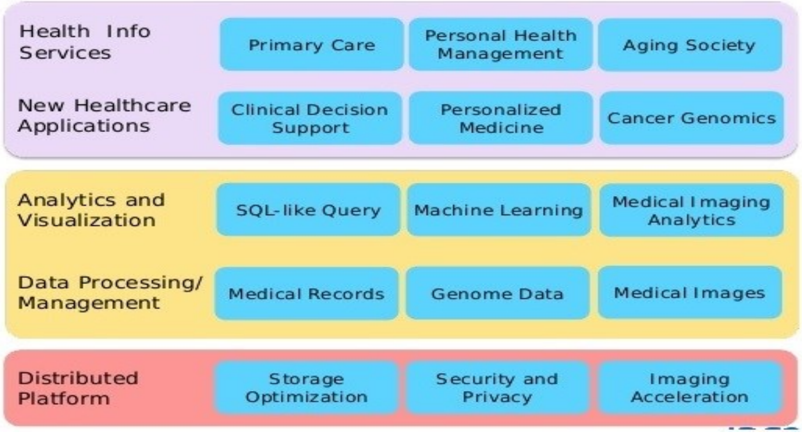

Figure 4 Big Data solutions for healthcare

Furthermore, [14] implies that big data analytics in healthcare could indeed help with:

1. *Evidence-based medicine:* the use of both structured and unstructured information (EMRs, financial and operational data, clinical data, and genomic data) to contest therapies with consequences, anticipate clients at danger of illness or admittance, and deliver more effective service.
2. *Genomic analytics*: end up making gene genome sequence extra effective and expense, and incorporate epigenetic analysis into routine medical care decision-making as well as the rising patient medical record [15].
3. *Pre-adjudication fraud analysis:* evaluate huge numbers of assertion queries quickly in hopes of minimizing fraud, waste, and abuse.
4. *Device/remote monitoring:* acquisition and analyze information from in-hospital and at-home gadgets at all times for safety assessment and adverse reaction prediction.
5. *Patient profile analytics:* Use analytics tools on patient statuses (e.g., fragmentation and forecasting) to identify people who would advantage from appropriate medical care or lifestyle modification, such as those with a high likelihood of getting a particular illness (e.g., diabetes) and therefore would advantage from preventative medicine [14].

4.2 Big data sources in healthcare

Big data in healthcare can emerge from both intrinsic (e.g., electronic medical records, systems for clinical decision-making, CPOE, etc.) and external entities, mostly in various formats (flat files,.csv, interpersonal tables, ASCII/text, etc.) and taking up residence in numerous places (regional in addition to different medical providers' sites) in innumerable heritage as well as other implementations (transaction processing applications, databases, etc.). Among the source materials and types of data are:

a. Clickstream and communication data from Facebook, Twitter, LinkedIn, blogs, and other social networking sites. It may additionally encompass health insurance internet sites, mobile applications, and so on [11].

b. Data exchanged between machines: observations from sensing devices, metres, as well as other vital sign gadgets [11].

c. Health care assertions as well as other billing information are increasingly accessible in semi-structured and unorganised formats [11].

d. Fingerprints, genetic factors, handwriting, retina scanning, x-rays and other imaging techniques, blood pressure, pulse and pulse-oximetry interpretations, as well as other similar kinds of information are examples of biometric data [11].

e. Human-generated data: unstructured and semi-structured information including such EMRs, physicians' notations, e - mails, and documents [11].

The multidisciplinary big data analytics team in healthcare creates a "notion declaration" in Step 1. It is the initial attempt to determine the significance of such a project. The notion declaration is accompanied by an explanation of the importance of the venture. The healthcare provider will recognise trade-offs in aspects of possible alternatives, expense, expandability, and so forth. Once a notion declaration has been accepted, the squad can move on to Stage 2, proposition advancement. More information is provided here. Numerous issues are addressed relying on the notion declaration. The stages in the technique are then clearly laid out and executed in Step 3. The notion declaration is divided into a number of suppositions. The variables or indicators are recognised concurrently. Platform and device assessment and choice is a critical step at this point. As previously stated, there are numerous choices available, such as AWS Hadoop, Cloudera, and IBM Big Insights. The information will then be subjected to various big data analytics methodologies. Just one difference between this procedure and routine predictive analysis is that the methodologies are expanded to large data sets. Big data analytics provides insight throughout a number of iterations and what-if analyses. With this knowledge, sound decisions can be made. Step 4 involves testing, validating, and presenting the models and their research results to stakeholders for action. Implementation is a staged approach with feedback loops built in at each stage to minimize risk of failure. A detailed diagrammatic view has been provided in Figure 5 demonstrating these steps.

Step 1	Concept statement
	• Establishment of the requirement of big data analytics project in healthcare on the basis of “4Vs”.
Step 2	Proposal
	• Definition of the problem that is being lectured?
	• Important and motivation behind the problem?
	• The need for following big data analytics approach?
	• Background material
Step 3	Methodology
	• Proposal
	• Variable selection
	• Information collection
	• ETL and data transformation
	• Platform/tool selection
	• Conceptual model
	• Analytic methods
	-Association, clustering, classification, etc.
	• Consequences & understanding
Step 4	Positioning
	• Assessment & authentication
	• Testing

Table 1 Outline of big data analytics in healthcare methodology

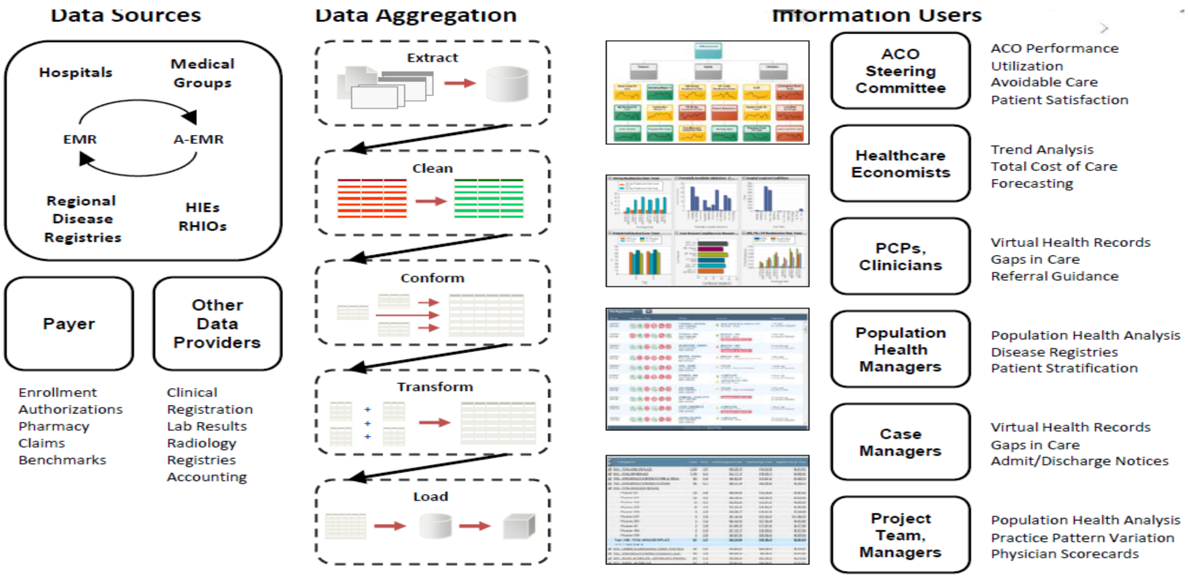

Figure 5 An in-detail diagrammatic view of the application of Big data in Healthcare

6. Big Data Application in Biomedical Research

Big data is a new framework and ecosystem in biomedical informatics that converts specific instance research into large-scale, data-driven investigation. The characteristics of big data are widely accepted to be characterised by three main components, popularly known as the "3 Vs": volume, variety, and velocity. First and foremost, the amount of information in biomedical informatics disciplines is growing at a rate that is exponential. [16-22] The Proteomics DB [23], for example, encompasses 92% (18,097 of 19,629) of recognised genetic mutations catalogued in the Swiss-Prot database. The volume of information in Proteomics DB is 5.17 TB. From 2009 to 2012, the promotion of the HITECH Act [24] roughly doubled the rate of adoption of electronic health records (EHRs) in hospitals to 44%. Electronically stored data from millions of patients have already been gathered and might potentially improve healthcare services and expand research prospects [25, 26].In addition, medical imaging (eg, MRI, CT scans) produces vast amounts of data with even more complex features and broader dimensions.

6.1 Big Data Technologies in Biomedical Research

Biomedical researchers are confronted with new obstacles in storages, trying to manage, and evaluating enormous quantities of information. [27] Big data's features necessitate utilizing influential and unique innovations to retrieve useful data and allow more comprehensive health-care solutions. We discovered technology tools used together in the majority of the cases reported, including such artificial intelligence (AI), Hadoop [28], and data mining tools.

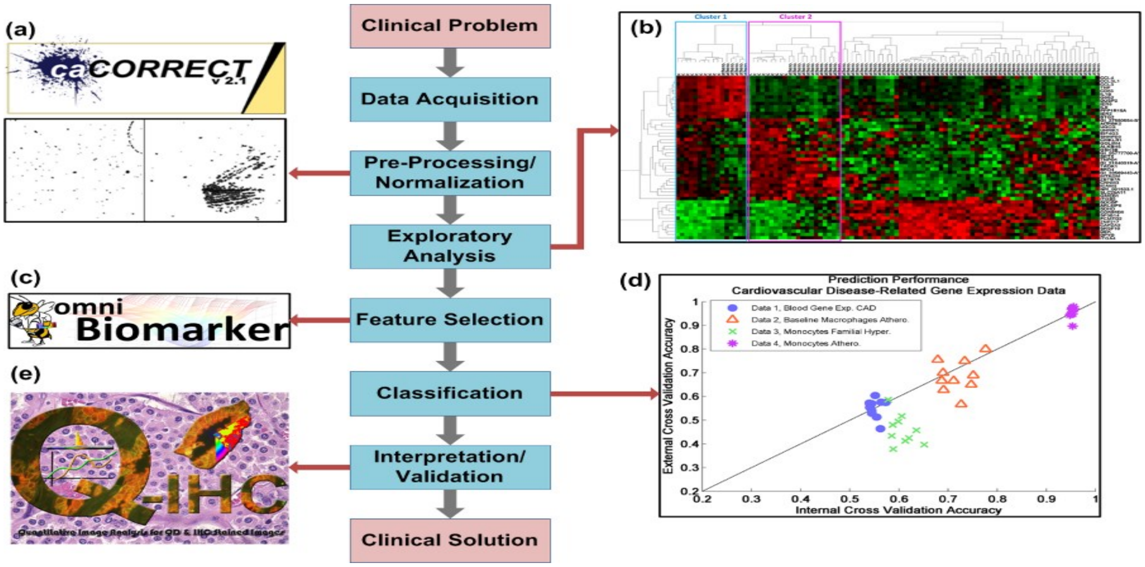
(a)
caCORRECT v2.1
Clinical Problem
Data Acquisition
Pre-Processing/ Normalization
Exploratory Analysis
Feature Selection
Classification
Interpretation/ Validation
Clinical Solution
(b)
Cluster 1
Cluster 2
(c)
omni Biomarker
(d)
Prediction Performance
Cardiovascular Disease-Related Gene Expression Data
Data 1, Blood Gene Exp. CAD
Data 2, Baseline Macrophages Athero.
Data 3, Monocytes Familial Hyper.
Data 4, Monocytes Athero.
External Cross Validation Accuracy
Internal Cross Validation Accuracy
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
(e)
Q-IHC
Quantitative Image Analysis for QD & IHC Stained Images

Figure 6 Stepwise diagram explaining the application of big data in biomedical research

Parallel computing is a vital infrastructure for trying to manage big data tasks. It is able to perform computational programs concurrently on a machine cluster, or supercomputer. Novel parallel computing designs, including such Google's MapReduce [29], have indeed been put forward in recent years for a new big data facility. Hadoop [28], a usable MapReduce tool for decentralised data management, was just released by Apache. Data access to clustered servers is supported simultaneously via the Hadoop Distributed File System (HDFS). Hadoop-based facilities may additionally be thought of as platforms for cloud computing, as they enable both central data storage and remote monitoring via the Internet.

As being such, cloud computing presents an innovative model for communicating customizable computational power over a network [30], and it can function as a facility, framework, and/or software to provide an integrated approach. Moreover, cloud computing can boost system's performance, quickness, and versatility by reducing the requirement to sustain either software or hardware capabilities and requiring limited resources for framework scheduled maintenance like setup, arrangement, and checking. Cloud technologies are at the heart of several new big data applications.

6.2 Biomedical Research methodology

Bioinformatics study analyses biological system variants at the molecular scale. With prevailing customised medicine patterns, there is a growing necessity generate, store, and evaluate these huge datasets in a controllable timespan. Next-generation building the next generation allows for the rapid collection of genetic information [31, 32]. Big data techniques play a role in bioinformatics applications by offering datasets, computing facilities, and effective data manipulation techniques for researchers to assemble and analyse biological information.

Taylor mentions how Hadoop and MapReduce are now widely used in the biomedical field [33].

This section categorises big data tools and technologies into 4 categories: (1) data storage and information extraction, (2) error recognition, (3) data processing, and (4) platform integration and implementation. These classifications are linked and may coincide; for example, most data processing applications may endorse basic data analysis and conversely. Nevertheless, in this study, we only classify methods that focus on their main purposes.

6.2.1 Data storage and retrieval

A sequencing machine can now generate millions of short DNA sequencing data in a single run. To be used for further research, such as genotyping and expression variation analyses, the sequencing data must be mapped to specified reference genomes. The CloudBurst [34] parallel computing model makes the genome mapping procedure easier. To enhance the scalability of reading massive sequencing data, CloudBurst parallelizes the short-read mapping procedure. A 25-core cluster was used to test the CloudBurst model, and the findings show that it processed seven million short reads about 24 times faster than a single core system. In order to promote biomedical research, the CloudBurst team has created new tools based on CloudBurst, such as Crossbow [36] and Contrail [35] for assembling huge genomes and single nucleotide polymorphisms (SNP) identification from sequencing data.

On a Hadoop cluster, *DistMap* [37] is a toolbox for distributed short-read mapping. With DistMap, more types of mappers will be supported, allowing it to address a wider range of sequencing applications. BWA, Bowtie, Bowtie2, GSNAP, SOAP, STAR, Bismark, BSMAP, and TopHat are among the nine supported mapper types. DistMap incorporates a procedure for mapping that may be used with straightforward commands. It works well for mapping short-read data, as demonstrated by an evaluation test utilising a 13-node cluster. DistMap enables the BWA mapper to complete 500 million read pairs (247 GB) at a rate that is 13 times faster than a single-node mapper.

To facilitate access to massive whole-genome datasets for bioinformatics researchers, *SeqWare* [38] is a query engine built on the Apache HBase [39] database. An interactive interface was developed by the SeqWare team to incorporate genome browsers and tools. The U87MG and 1102GBM tumour datasets were loaded for a prototype study, and the researchers utilised this engine to compare the capacities of the Berkeley DB and HBase back ends for loading and exporting variant data. According to the findings, the HBase solution is quicker when reading more than 6M variants, whereas the Berkeley DB solution is quicker when reading that many variants.

The DNA Data Bank of Japan (DDBJ) created the Read Annotation Pipeline [40], a cloud-based pipeline for high-throughput analysis of data from next-generation sequencing. This cloud computing system was set developed by DDBJ to assist in sequencing analysis. It provides a user-friendly interface for processing sequencing data sets and supports two levels of analysis: (1) basic-level tools accept FASTQ format data and pre-process them to remove low-quality bases; and (2) second-level tools map the data to genome references or assemble them on supercomputers. For sophisticated analysis, including SNP detection, RNA-sequencing (RNA-seq), and ChIP-seq analysis, this pipeline makes use of the Galaxy interface. In 6.5 hours, DDBJ successfully mapped 34.7 million sequencing reads to a 383 MB reference genome in a benchmark test.

Scalable Hadoop-based proteomic search engine *Hydra* [41] leverages distributed computing. A distributed computer architecture that facilitates the scalable searching of enormous volumes of spectrometry data is implemented by the software package Hydra, which is used to process vast peptide and spectra databases. Hydra divides the proteome search process into two stages: creating a database of peptides and scoring the spectra and obtaining the data. On a Hadoop cluster with 43 nodes, the system can score 27 billion peptides in roughly 40 minutes.

6.2.2 Error identification

A number of tools have been developed to identify errors in sequencing data; *SAMQA* [42] identifies such errors and ensures that large-scale genomic data meet the minimum quality standards. Originally built for the National Institutes of Health Cancer Genome Atlas to automatically identify and report errors, SAMQA includes a set of technical tests to find data abnormalities (eg, sequence alignment/map [SAM] format error, invalid CIGAR value) that contain empty reads. For biological tests, researchers can set a threshold to filter reads that could be erroneous (empty reads) and report them to experts for manual evaluation. Hadoop, which was evaluated on a cluster, processed a 23 GB sample about 80 times faster than SAMQA, which was tested on a single-core server, according to a comparison (18.25 hours).

For three major sequencing platforms—454 SequencingTM, Illumina, and SOLiD—ART [43] offers simulation data for sequencing analysis. Base substitutions, insertions, and deletions are the three forms of sequencing errors that ART can detect thanks to built-in profiles of read error and read length.

An error-correction approach for high-throughput sequencing data that is built on a parallel, scalable framework is called *CloudRS* [44]. This technique was created using the RS algorithm. [45] The GAGE benchmarks were used by the CloudRS team to analyse the system on six distinct datasets. The results indicate that CloudRS has a greater precision rate than the Reptile [47] technique.

6.2.3 Data analysis

The Genome Analysis Toolkit (GATK), [48] is a MapReduce-based programming framework created to support large-scale DNA sequence analysis in addition to the mentioned frameworks and toolkits for sequencing data analysis. Many data types are supported by GATK, such as SAM files, binary alignment/map (BAM), HapMap, and dbSNP. Via the use of "traversal" modules, GATK prepares and reads sequencing data into the system, providing associated references to the data, such as data organisation by loci. Data is consumed by the "walker" module, which then produces analytics results. The Cancer Genome Atlas and 1000 Genomes Projects both made use of GATK.

An international partnership has created the *ArrayExpress Archive of Functional Genomics* data repository [49,50] for combining high-throughput genomics data. More than a million assays and 30,000 experiments can be found in the repository. Around 80% of the data were taken straight from the GEO data repository, and the remaining 20% came from user submissions to ArrayExpress. More than 1,000 unique users access the platform each day, and more than 50 GB of data are downloaded. To assist with data movement and analysis, the platform also has connections with R and GenomeSpace.

The R package, *BlueSNP* [51] for GWAS analysis focuses on statistical tests (e.g., P-value) to uncover strong connections between significant genotype-phenotype datasets. BlueSNP uses the Hadoop platform, which lowers obstacles and boosts the effectiveness of GWAS analysis performed on clustered computers. BlueSNP examined 1,000 phenotypes on 106 SNPs in 104 individuals on a 40-node cluster in 34 minutes.

A cloud-based computing pipeline called *Myrna* [52] computes the variations in gene expression found in huge RNA-seq datasets. The m-sequencing reads used to generate RNA-seq data come from mRNA molecules. Reads alignment, normalisation, and statistical modelling are only a few of the features that Myrna enables for RNA-seq analysis in an integrated pipeline. Gene differential expression is reported by Myrna as a P-value and q-value. The Amazon Elastic Compute Cloud (Amazon EC2) was used to test this system using 1.1 billion RNA-seq reads, and the results reveal that Myrna can process data in under two hours for a cost of about $66.

Data imports from sequencer reads, data mapping to reference genomes, alignment filters, transcription expression calculations, expression normalisations using edgeR, and differential expression detection are all part of the pipeline for analysing the differential transcript expressions that was implanted by the *Eoulsan* package [53]. Three different configurations of Eoulsan are available: standalone, local cluster, and cloud using Amazon Elastic MapReduce. Eight mouse samples totaling 188 million readings were used in the Eoulsan test on Amazon EC2. The processing of the data cost between $18 and $66 and took between 109 and 822 minutes.

A quick, scalable, cloud-ready software suite for interactive genomic data processing with nucleotide precision is called *SparkSeq* [54]. For RNA/DNA investigations, SparkSeq offers interactive queries. The project is built on Apache Spark and uses the Hadoop-BAM library to analyse bioinformatics files.

6.2.4 Platform integration deployment

The usage of big data platforms typically necessitates a solid understanding of networking and distributed computing. New approaches are required to combine existing big data technologies with user-friendly operations so that biomedical researchers may adopt big data technology. The systems listed below have been created to assist in achieving this objective.

Bioinformaticians no longer need to learn the specialised technical knowledge required to use MapReduce thanks to *SeqPig* [55]. The SeqPig project adds functionality for feature-rich sequence processing to the Apache Pig scripts. SeqPig resolves the issue of reading big BAM files to feed analytic programmes with the use of Hadoop-BAM [56]. Common sequencing formats supported by SeqPig include FASTQ, SAM, BAM, and QSeq. Additionally, it supports widely used processing methods like distribution, read coverage, read frequency count, and pileup.

Virtual machines are also incorporated into the current bioinformatics platform. A sequencing analysis programme called CloVR [57] is delivered via virtual machines. CloVR supports both local desktop and cloud platforms to allow high-throughput data processing by lowering the technological obstacles for evaluating huge sequencing datasets. The virtual machine incorporates a number of automated bioinformatics workflows and pipelines, such as those for whole-genome, metagenome, and 16S rRNA-sequencing analysis. The CloVR team tested the system's portability on a local workstation (4 CPU, 8 GB RAM) and on the Amazon EC2 cloud platform (80 CPU). The results demonstrate that CloVR works on both platforms, however the EC2 instance performs around five times quicker. In a similar vein, CloudBio-Linux [54] is a virtual machine solution that offers over 135 bioinformatics packages for sequencing analysis, including preconfigured tools (such as GATK, Bowtie, Velvet, and FASTX) and programming libraries (eg, BioJava, R, Bioconductor).

Deploying the Hadoop cloud platform could be quite challenging for researchers without a background in computer science. CloudDOE is a software programme that gives a simple interface for building the Hadoop cloud because the Hadoop platform is frequently too difficult for scientists without computer science background and/or comparable technical skills. A user-friendly tool called CloudDOE [59] makes it easier for bioinformatics researchers to set up the Hadoop cloud and utilise MapReduce to analyse high-throughput sequencing data. A number of other programmes are integrated into the CloudDOE package (CloudBurst, CloudBrush, and CloudRS), and wizards and graphical user interfaces further simplify use.

7. Application of Big Data in Biology: Multicellular Datasets

Big data research in Life Sciences typically focuses on big molecular datasets of protein structures, DNA sequences, gene expression, proteomics and metabolomics. Now, however, new developments in three-dimensional imaging and microscopy have started to deliver big datasets of cell behaviours during embryonic development including cell trajectories and shapes and patterns of gene activity from every position in the embryo. This surge of multicellular and multi-scale biological data poses exciting new challenges for the application of ICT and applied mathematics in this field.

7.1 Advancements in the field

Technological developments in microscopy and image analysis are now producing a flood of new data that excites me much more. With this data, it is now possible to track the movements and behaviours of any cell, in an early embryo, organ, or tumour. With this capability we will now be able to identify what makes cells take a wrong turn in children with birth defects or how tumour cells can change their metabolism and movement to out compete their well-behaved neighbours and disrupt the structure and function of an organ. Such mechanistic insights will eventually make it possible to interfere with developmental mechanisms with a greater specificity than currently possible.

Conventional light microscopy can already follow the migration of a subset of individual cells (labelled with fluorescent markers) in organs but techniques are getting better. Two-photon microscopy techniques, used in conjunction with advanced image analysis, allow researchers to routinely generate all-cell datasets of developing embryos or organs. Applying this approach, the BioEmergences platform at CNRS (Gif-sur-Yvette, France) recently produced a gene expression atlas featuring cellular resolution of developing zebrafish [60]. Soon we will be able to follow every cell in developing organisms and tissues and concurrently identify what genes they are expressing and what metabolites they are producing.

Ongoing initiatives in the field of information sciences are laying the foundations for similar data standards and domain-specific languages in the multicellular biology community. New versions of SBML will allow users to describe the distribution of molecules in fixed geometries and coupled cells. However, in a recent paper that proposed a Cell Behaviour Ontology (CBO) [61], it was argued that SBML is not the most efficient or insightful way to annotate embryological data. The multicellular organism is a collection of thousands to trillions of individual cells. Individually describing the gene expression levels and biophysical properties of each cell will create huge datasets but not necessarily yield useful insights. Even the most detailed three-dimensional movies or sets of cell trajectories are merely pretty pictures unless we can identify and label their components meaningfully. A useful comparison is thinking about the difference between providing a list of pixels in an image versus the list of things in that image. CBO focuses on describing the behaviour of cells and the dependency of those behaviours on the cell's internal machinery. This includes its gene expression pattern and local environment. This declarative approach allows the CBO to categorise each cell in a developing embryo using a manageable set of cell types which range from the tens to hundreds in number. Each cell type is characterised by the same class of behaviours, thus, cells belonging to the same cell type share the same behaviours. Each cell follows a set of logical input and output rules that guide these behaviours and its transition from one cell type to another (i.e., differentiation). Many cell types in multicellular organisms are 'sub-types' whose behaviour varies in subtle ways around a general 'base' cell type. For example, the endothelial cells in a developing blood vessel are made up of two sub-types: 'tip' cells at the end of a sprouting blood vessel which are usually spikier and more motile and 'stalk' cells which occur to the back of the sprout. This approach allows the CBO to develop a hierarchical classification of cell types and cell behaviours.

Besides compressing the data, the classification of cell behaviours will also enable quantitative biologists to understand biological development to a point that, with the aid of applied mathematicians, they can then reconstruct it using agent-based computer simulations. This will then enable them to unravel how subtle changes in cell behaviour, driven by factors such as inherited disease or cancer, can affect the outcome of development and why. Thus, the resulting datasets become more meaningful descriptions of the observations as well as sets of rules to construct agent-based computer simulations of those observations. In this way, CBO takes a 'cell-based approach' [62], which views embryogenesis as the collective behaviour of a 'colony' of individual cells.

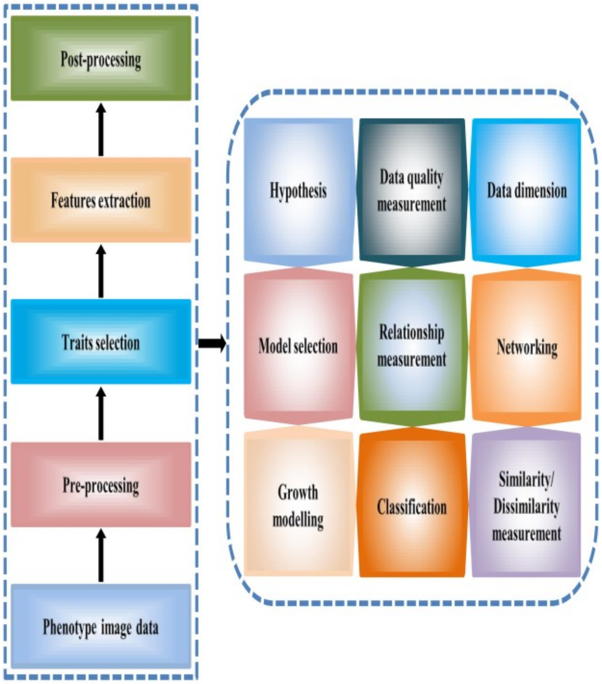

Figure 7 Application of Big data in Biology

8. Big-data medicine by dynamical network biomarkers

It is commonly recognized that a complicated living organism cannot be completely appreciated by merely analyzing individual components. Phenotypes and functions of an organism are ultimately determined by interactions between these components or networks in terms of structures and dynamics [2]. Network and dynamics are two key aspects in computational systems biology [6,7–13]. However, majority of traditional research focuses on the static and statistic properties (e.g., GWAS) of big data, rather than the essential dynamics and networks of life in living organisms. Generally, a disease is a problem resulting not from malfunction of individual molecules but from failure of the relevant system or network, which can be considered as a set of interactions among molecules.

Thus, rather than single molecules, the networks are stable forms as biomarkers to reliably characterize complex diseases. The era of big data [14,15] provides great opportunities for predictive, preventive, personalized and participatory (P4) medicine, which is expected to lead to big-data medicine. The study of network and interactions of biological elements rather than biological elements themselves, can capture the previously-unobserved features at the levels of both network (or edges) and dynamics. Therefore, with the demand from both theoretical and clinic aspects, biomarkers are evolving from single molecules (e.g., individual genes) to multiple molecules (e.g., gene set), associated molecules (e.g., molecule network) and dynamical interactive molecules (e.g., dynamical molecule network) due to the availability of big data, in particular, high-dimensional data, which can be categorized as node biomarkers [14,15], network-based biomarkers [16–18], network biomarkers [19,20] and dynamical network biomarkers (DNBs) [21,22], respectively. By exploiting the network information from big data, recent studies on Edge Marker [14,20] demonstrate that non-differentially expressed genes, which -are usually ignored by traditional methods, can be as informative as differentially expressed genes in terms of classifying different biological conditions or phenotypes of samples. By exploiting the dynamical information from big data, a novel biomarker, DNB, was recently developed [22]. In contrast to the disease state detected by traditional biomarkers, DNB is able to identify the pre-disease state before the occurrence or serious deterioration of diseases, which can actually be used to prevent from further disease progression before deteriorating into their irreversible states [21–24]. In other words, by high-dimensional data (such as gene expression, RNA-seq, protein expression, and imaging data), this new type of biomarkers can achieve the early diagnosis of ‘‘pre-disease’’ state or ‘‘un-occurring disease’’ state, which is a concept raised in ‘‘Yellow Emperor’s Canon of Internal Medicine’’ (one of the earliest books for Traditional Chinese Medicine) [14].

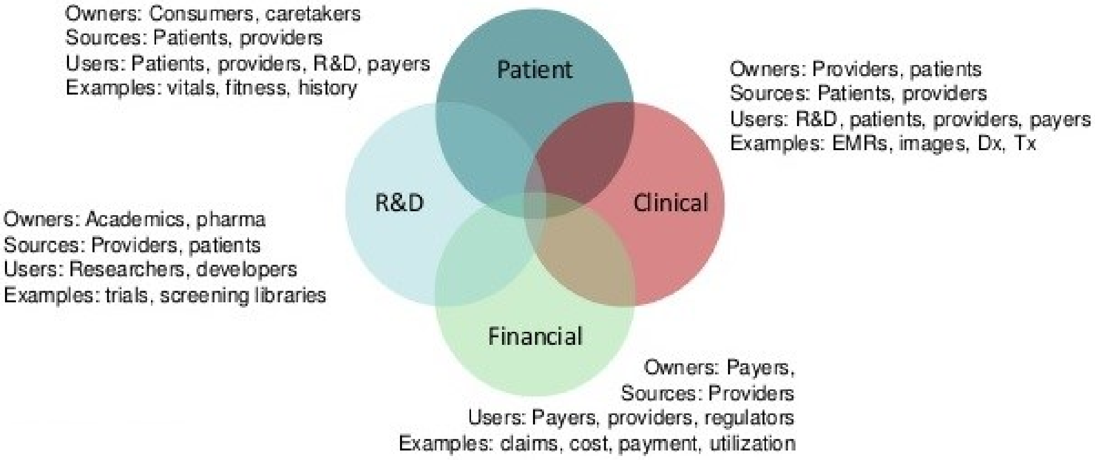
Owners: Consumers, caretakers
Sources: Patients, providers
Users: Patients, providers, R&D, payers
Examples: vitals, fitness, history
Patient
Owners: Providers, patients
Sources: Patients, providers
Users: R&D, patients, providers, payers
Examples: EMRs, images, Dx, Tx
R&D
Clinical
Owners: Academics, pharma
Sources: Providers, patients
Users: Researchers, developers
Examples: trials, screening libraries
Financial
Owners: Payers,
Sources: Providers
Users: Payers, providers, regulators
Examples: claims, cost, payment, utilization

Figure 8 Big data applications in medicine

8. Conclusion

Big Data arise from many frontiers of scientific research and technological developments. They hold great promise for the discovery of heterogeneity and the search for personalized treatments. They also allow us to find weak patterns in presence of large individual variations. Salient features of Big Data which include experimental variations, computational cost, noise accumulation, spurious correlations, incidental endogeneity, and measurement errors should be seriously considered in Big Data analysis and in the development of statistical procedures.

The use of advanced technologies by healthcare providers to gather knowledge from their clinical and other data repositories and make wise judgements has the potential to change as a result of big data analytics. Big data analytics will soon be quickly and extensively implemented throughout the healthcare system and the wider healthcare sector. To that goal, it is necessary to overcome the several difficulties mentioned above. As big data analytics gains popularity, concerns like ensuring privacy, securing data, establishing standards and governance, and upgrading tools and technology will come to light. Although big data analytics and applications in healthcare are still in their infancy, tremendous advancements in platforms and technologies could hasten their maturation.

It must be emphasized that the health care industry remains well within its infancy of leveraging big data for business and clinical use. Although there have been some successes, many are unproven at the outcome level and much work remains to determine whether those strategies and systems that work best at one facility (eg, The Rhode Island Beacon Community Program, see page 9) can work equally well at another due to cultural, technological, and other types of variables.

Big data could change what scientists know and how they do science. Rather than analyzing data to answer a particular question, creative data mining may allow data to inspire questions—opening the door for hypothesis- generating as well as hypothesis-driven science.

The incremental expansion of health information across different contexts has compelled computational specialists to devise novel techniques for analysing and interpreting such massive amounts of data in a short period of time. All these scientists and practising healthcare practitioners are increasingly integrating computational systems for signal analysis. As a result, the next big goal could be to build an in-depth representation of the human body by incorporating physiological parameters and "omics" methodologies. This novel concept has the potential to improve our understanding of disease circumstances and contribute to the creation of innovative diagnostic equipment. The increasing availability of genomic data, which include inherent hidden errors from experimentation and analytical practises, necessitates additional investigation. Nevertheless, there's many possibilities to incorporate continuous improvement in healthcare analytics at every stage of this rigorous process. The healthcare industry is obviously transitioning from a broad volume base to a personalised or participant field. As a result, understanding the emerging scenario is critical for scientists and engineers and professionals. Big data analytics is expected to progress forward towards a predictive system in the future. This would imply forecasting future consequences in an individual 's wellbeing according to present or historical data (such as EHR-based and omics-based). Correspondingly, structured information gained from a specific geography may result in the creation of population health records. Big data will help healthcare by having to introduce epidemic forecasting (in connection with population health), supplying advance warning of disease states, and assisting in the development of new genetic markers and intellectual individual therapy techniques for an enhanced standard of living.

References

[1] New Horizons for a Data-Driven Economy - Springer. doi:10.1007/978-3-319-21569-3

[2] Hellerstein, J. Parallel Programming in the Age of Big Data. Gigaom Blog, 2008.

[3] Segaran, T. & Hammerbacher, J. Beautiful Data: The Stories behind Elegant Data Solutions. O'Reilly Media, 2009, 257, ISBN 978-0-596-15711-1.

[4] Snijders, C., Matzat, U. & Reips, U.D. ‘Big Data': Big gaps of knowledge in the field of Internet. International Journal of Internet Science, 7, 2012, 1–5.

[5] Kevin N. Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. September 12, 2013. http://insidebigdata.com.

[6] Khalili, A. and Chen, J. Variable selection in finite mixture of regression models. Journal of the American Statistical Association, 2007, 102, 1025-1038.

[7] Fan, J., Samworth, R., & Wu, Y. Ultrahigh dimensional feature selection: Beyond the linear model. The Journal of Machine Learning Research, 2009, 10, 2013-2038.

[8] Cai, T. & Jiang, T. Phase transition in limiting distributions of coherence of high-dimensional random matrices. Journal of Multivariate Analysis, 2012, 107, 24-39.

[9] Tibshirani, R.J. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 1996, 58, 267-288.

[10] Fan, J. & Fan, Y. High-dimensional classification using features annealed independence rules. The Annals of Statistics, 2008, 36, 2605.

[11] Cottle, Mike, et al. Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry. Institute for Health Technology Transformation, 2013. http://ihealthtran.com/big-data-in-healthcare.

[12] LaValle, S., et al. Big Data, Analytics and the Path from Insights to Value'Sloan Management Review, (Winter 2011Research Feature), 21 December 2010. 2011.

[13] Burghard, C. Big data and analytics key to accountable care success. IDC Health Insights, Sponsored by: IBM, 2012, 3-4.

[14] Raghupathi, W. & Viju R. Big data analytics in healthcare: promise and potential. Health Information Science and Systems 2.1, 2014, 1.

[15] Knabel, M. K., K. D. & Dennis S. F. Intellectual Property Protection for Synthetic Biology, Including Bioinformatics and Computational Intelligence. Big Data Analytics in Bioinformatics and Healthcare, 2014, 380.

[16] Stratton M.R., Campbell P.J. & Futreal P.A. The cancer genome. Nature. 2009; 458(7239), 719–724.

[17] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008, 26(10), 1135–1145.

[18] Metzker M. L. Sequencing technologies – the next generation. Nat Rev Genet. 2010, 11(1), 31–46.

[19] Nielsen R, Paul JS, Albrechtsen A, et al. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet, 2011, 12(6), 443–451.

[20] Zhang J, Chiodini R, Badr A, et al. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011, 38(3), 95–109.

[21] Murdoch T.B. Detsky AS. The inevitable application of big data to healthcare. JAMA. 2013, 309(13),1351–1352.

[22] Lynch C. Big data: how do your data grow? Nature, 2008, 455(7209, 28–29.

[23] Wilhelm M, Schlegl J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. Nature, 2014, 509(7502), 582–587.

[24] Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010, 363(6), 501–504.

[25] Botsis T., Hartvigsen G., Chen F., et al. Secondary use of EHR: data quality issues and informatics opportunities. In AMIA Summits on Translational Science Proceedings, AMIA, San Francisco, California, 2010, pp.1.

[26] Rea S., Pathak J., Savova G., et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform, 2012, 45(4), 763–771.

[27] Margolis R., Derr L., Dunn M., et al. The National Institutes of Health's Big Data to Knowledge (BD2 K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc., 2014, 21(6), 957–958.

[28] White T. Hadoop: The Definitive Guide. Sebastopol. O'Reilly Media, Inc., CA, 2012.

[29] Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM, 2008, 51(1), 107–113.

[30] Armbrust M., Fox A., Griffith R., et al. A view of cloud computing. Commun ACM, 2010, 53(4), 50–58.

[31] Schuster SC. Next-generation sequencing transforms today's biology. Nature, 2007, 200(8), 16–18.

[32] Morozova, O., Marra, M.A. Applications of next-generation sequencing technologies in functional genomics. Genomics, 2008, 92(5), 255–264.

[33] Taylor R. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics, 2010, 11(suppl 12), S1.

[34] Schatz M.C. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics, 2009, 25(11), 1363–1369.

[35] Schatz M., Sommer D., Kelley D., et al. Contrail: assembly of large genomes using cloud computing. CSHL Biology of Genomes Conference, Cold Spring Harbor, New York, CSHL, 2010.

[36] Gurtowski J., Schatz M.C., Langmead B. Genotyping in the cloud with crossbow. Curr Protoc Bioinformatics, 2012, Chapter 15, Unit15.3.

[37] Pandey R.V., Schlötterer C. DistMap: a toolkit for distributed short read mapping on a Hadoop cluster. PLoS One, 2013, 8(8), e72614.

[38] O'Connor B.D., Merriman B., Nelson S.F. SeqWare query engine: storing and searching sequence data in the cloud. BMC Bioinformatics, 2010, 11(suppl 12), S2.

[39] George L. HBase: The Definitive Guide. Sebastopol, O'Reilly Media Inc, CA, 2011.

[40] Nagasaki H., Mochizuki T., Kodama Y., et al. DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. DNA Res., 2013, 20(4), 383–90.

[41] Lewis S., Csordas A., Killcoyne S., et al. Hydra: a scalable proteomic search engine which utilizes

the Hadoop distributed computing framework. BMC Bioinformatics. 2012, 13(1), 324.

[42] Robinson T., Killcoyne S., Bressler R., et al. SAMQA: error classification and validation of high-throughput sequenced read data. BMC Genomics, 2011, 12(1), 419.

[43] Huang W., Li L., Myers J.R., et al. ART: a next-generation sequencing read simulator. Bioinformatics, 2012, 28(4), 593–594.

[44] Chen C. C., Chang Y. J., Chung W. C., et al. CloudRS: an error correction algorithm of high-throughput sequencing data based on scalable framework. 2013 IEEE International Conference on Big Data; Santa Clara, California: IEEE, 2013, pp.717–722.

[45] Gnerre S., MacCallum I., Przybylski D., et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A, 2011, 108(4), 1513–1518.

[46] [46] Salzberg S.L., Phillippy A.M., Zimin A., et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res, 2012, 22(3), 557–567.

[47] Yang X., Dorman K.S. & Aluru S. Reptile: representative tiling for short read error correction. Bioinformatics, 2010, 26(20), 2526–2533.

[48] Van der Auwera G.A., Carneiro M.O., Hartl C., et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013. 11(1110). 11.10.1–11.10.33.

[49] Rustici G., Kolesnikov N., Brandizi M., et al. Array Express update – trends in database growth and links to data analysis tools. Nucleic Acids Res., 2013, 41(D1), D987–990.

[50] Brazma A., Parkinson H., Sarkans U., et al. ArrayExpress – a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2003, 31(1), 68–71.

[51] Huang H., Tata S., Prill R.J. BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters. Bioinformatics, 2013, 29(1),135–136.

[52] Langmead B., Hansen K.D., Leek J.T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol., 2010, 11(8), R83.

[53] Jourdren L., Bernard M., Dillies M.A., et al. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics, 2012, 28(11), 1542–1543.

[54] Wiewiórka M.S., Messina A., Pacholewska A., et al. SparkSeq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision. Bioinformatics, 2014, 30(18), 2652–2653.

[55] Schumacher A, Pireddu L, Niemenmaa M, et al. SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. Bioinformatics, 2014, 30(1), 119–120.

[56] Niemenmaa M., Kallio A., Schumacher A., et al. Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics, 2012, 28(6), 876–877.

[57] Angiuoli S.V., Matalka M., Gussman A., et al. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics, 2011, 12(1), 356.

[58] Krampis K., Booth T., Chapman B., et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics, 2012, 13(1), 42.

[59] Chung W.C., Chen C.C., Ho J.M., et al. CloudDOE: a user-friendly tool for deploying Hadoop clouds and analyzing high-throughput sequencing data with MapReduce. PLoS One, 2014, 9(6), e98146.

[60] Castro-González, C. et al. A Digital Framework to Build, Visualize and Analyze a Gene Expression Atlas with Cellular Resolution in Zebrafish Early Embryogenesis, PLoS Comp. Biol., 2014, 10 (6), e1003670.

[61] Sluka, J. P. et al . The Cell Behavior Ontology: Describing the Intrinsic Biological Behaviors of Real and Model Cells Seen as Active Agents, Bioinformatics, 2014, 30(16), 2367-2374

[62] Merks, R.M.H. and Glazier, J.A. A Cell-Centered Approach to Developmental Biology, Physica A, 2005, 352(1), 113–130