Big Data for Health

Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong,
and Guang-Zhong Yang, *Fellow, IEEE*

Abstract—This paper provides an overview of recent developments in big data in the context of biomedical and health informatics. It outlines the key characteristics of big data and how medical and health informatics, translational bioinformatics, sensor informatics, and imaging informatics will benefit from an integrated approach of piecing together different aspects of personalized information from a diverse range of data sources, both structured and unstructured, covering genomics, proteomics, metabolomics, as well as imaging, clinical diagnosis, and long-term continuous physiological sensing of an individual. It is expected that recent advances in big data will expand our knowledge for testing new hypotheses about disease management from diagnosis to prevention to personalized treatment. The rise of big data, however, also raises challenges in terms of privacy, security, data ownership, data stewardship, and governance. This paper discusses some of the existing activities and future opportunities related to big data for health, outlining some of the key underlying issues that need to be tackled.

Index Terms—Big data, bioinformatics, health informatics, medical imaging, medical informatics, precision medicine, sensor informatics, social health.

I. INTRODUCTION

THE term "big data" has become a buzzword in recent years, with its usage frequency having doubled each year in the last few years according to common search engines. Fig. 1 illustrates the fast increase in the number of publications referring to "big data," regardless of disciplines, as well as those in the healthcare domain. Although the popularity of big data is recent, the underlying challenges have existed long before and have been actively pursued in health research. Big data in health is concerned with meaningful datasets that are too big, too fast, and too complex for healthcare providers to process and interpret with existing tools. It is driven by continuing effort in making health services more efficient and sustainable given the demands of a constantly expanding population with an inverted age pyramid, as well as the paradigm shift of
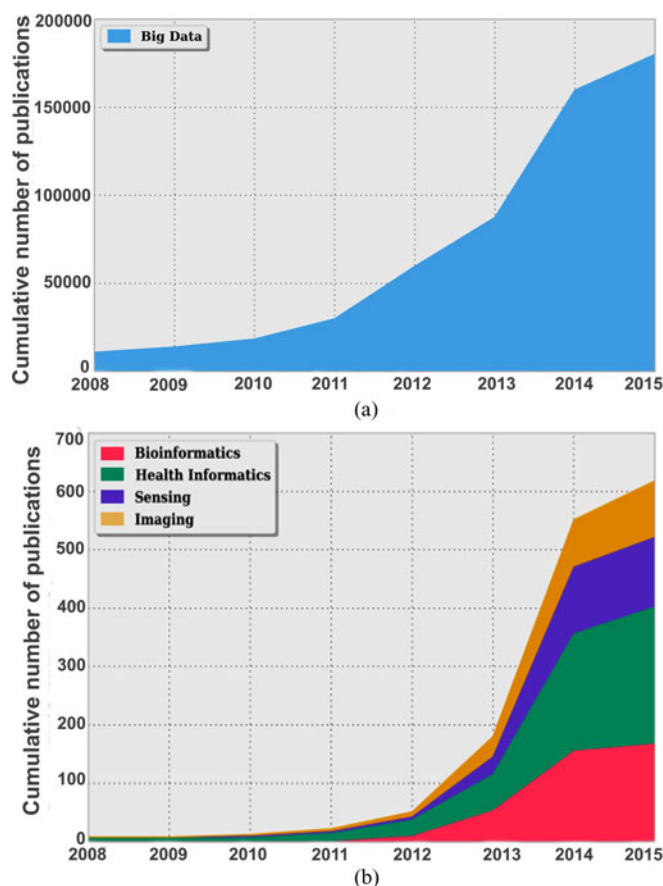
J. Andreu-Perez, R. D. Merrifield, and G.-Z. Yang are with the Hamlyn Centre, Imperial College London, London SW7 2AZ, U.K. e-mails: javier.andreu@imperial.ac.uk; rdm99@imperial.ac.uk; g.z.yang@imperial.ac.uk.

C. C. Y. Poon is with the Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: cpoon@surgery.cuhk.edu.hk).

S. T. C. Wong is with the Houston Methodist Research Institute, Weill Cornell Medical College, Houston, TX, 77030, USA (e-mail: stwong@houstonmethodist.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2450362

Fig. 1. (a) Cumulative number of publications referring to "big data" indexed by Google Scholar. (b) Cumulative number of publications per health research area referring to "big data," as indexed in IEEE Xplore, ACM Digital library, PubMed (National Library of Medicine, Bethesda, MD), Web of Science, and Scopus.

delivering health services toward *prevention, early intervention, and optimal management*.

In this paper, several ways of defining big data exist as a broad term to encapsulate the challenges related to the processing of a *massive amount of structured and unstructured data*. Clearly, the size (or volume) of data is an important factor of big data. Indeed, the US healthcare system alone already reached 150 exabytes (10^{18}) five years ago [1]. Before long, we will be dealing with zettabyte (10^{21}) and yottabyte (10^{24}) data for countries with large populations including emerging economies, such as China and India. This trend is due to the fact that multiscale data generated from individuals is continuously increasing, particularly with the new high-throughput sequencing platforms, real-time imaging, and point of care devices, as well as wearable computing and mobile health technologies. They provide genomics, proteomics and metabolomics, as well as long-term

Fig. 2. Six V's of big data (value, volume, velocity, variety, veracity, and variability), which also apply to health data.

continuous physiological features of an individual. In parallel, environmental factors present yet another set of variables that can be captured by continuous sensing that are important to population health.

However, size itself does not qualify big data. Other challenges include speed, heterogeneity, and variety of data in health. With the versatility, diversity, and connectivity of data capturing devices, additional data is generated at increasingly high speed, and decision support must be made available near real time in order to keep up with the constant evolution of technologies. In managing an influenza pandemic, for example, heterogeneous information from managed and unmanaged (e.g., social media, air travels) sources can be processed, mined, and turned into decisive actions to control the outbreak.

In healthcare, *data heterogeneity and variety* arise as a result of linking the diverse range of biomedical data sources available. Sources can be either quantitative (e.g., sensor data, images, gene arrays, laboratory tests) or qualitative (e.g., free text, demographics). The objectives underlying this data challenge are to support the basis for observational evidence to answer clinical questions, which would not otherwise been solved via studies based on randomized trials alone. In addition, the issue of generalizing results based on a narrow spectrum of participants may be solved by taking advantage of the potential of big data for deploying longitudinal studies.

Volume, *Velocity*, and *Variety* are the three Vs in the original definition of the key characteristics of big data in the research report published by META Group, Inc. (now Gartner, Inc.) [2]. Since then, other factors have also been considered, including *Variability* (consistency of data over time), *Veracity* (trustworthiness of the data obtained), and *Value*. These characteristics are summarized in Fig. 2 along with the key features that each captures.

Veracity is important for big data as, for example, personal health records may contain typographical errors, abbreviations, and cryptic notes. Ambulatory measurements are sometimes taken within less reliable, uncontrolled environments compared to clinical data, which are collected by trained practitioners. The use of spontaneous unmanaged data, such as those

from social media, can lead to wrong predictions as the data context is not always known. Furthermore, sources are often biased toward those young, internet savvy, and expressive online.

Last but not least, real *value* to both patients and healthcare systems can only be realized if challenges to analyze big data can be addressed in a coherent fashion. It should be noted that many of the underlying principles of big data have been explored by the research community for years in other domains. Nevertheless, new theories and approaches are needed for analyzing big health data. The total projected healthcare spending in the UK by 2021 will make 6.4% of the gross domestic product (GDP) [3] whilst the total projected healthcare share of the GDP in the United States is expected to reach 19.9% by 2022 [4]. In these respects, if used properly, big data can be a valuable resource that can provide significant insights toward improving contemporary health services and reducing healthcare costs. However, it also raises major social and legal challenges in terms of privacy, reidentification, data ownership, data stewardship, and governance.

In this paper, we will discuss some of the existing activities and future opportunities related to big data for health. More specifically, we will discuss its value for *Medical and Health Informatics*, *Translational Bioinformatics*, *Sensor Informatics*, and *Imaging Informatics*.

II. MEDICAL AND HEALTH INFORMATICS

With the ability to deal with large volumes of both structured and unstructured data from different sources, big data analytical tools hold the promise to study outcomes of large-scale population-based longitudinal studies, as well as to capture trends and propose predictive models for data generated from electronic medical and health records. A unique opportunity lies in the integration of traditional medical informatics with mobile health and social health, addressing both acute and chronic diseases in a way that we have never seen before.

A. Electronic Health Records (EHRs)

EHRs describing patient treatments and outcomes are rich but underused information. Traditional health data centres capture and store an enormous amount of structured data concerning a wide range of information including diagnostics, laboratory tests, medication, and ancillary clinical data. For individual patient reports, the use of natural language processing plays an essential role for systematic analysis and indexing of the underlying semantic contents. Mining EHRs is a valuable tool for improving clinical knowledge and supporting clinical research, for example, in discovering phenotype information [5]. Mining local information included in EHR data has already been proven to be effective for a wide range of healthcare challenges, such as disease management support [6], [7], pharmacovigilance [8], building models for predicting health risk assessment [9], [10], enhancing knowledge about survival rates [11], [12], therapeutic recommendation [11], [13], discovering comorbidities, and building support systems for the recruitment of
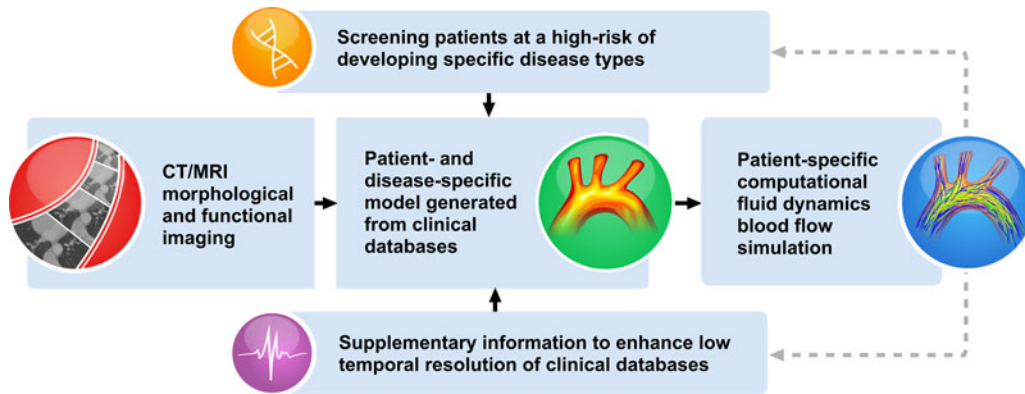
Fig. 3. Integration of imaging, modeling, and real-time sensing for the management of disease progression and planning of intervention procedures. This example of thoracic aortic dissection illustrates how risk stratification and subject-specific haemodynamic modeling substantiated with long-term continuous monitoring are used to guide the clinical decision process.

patients for new clinical trials [14]. Most of this work focused on the analysis of very large multidimensional longitudinal patient data collected over many years. However, most clinical databases provide low temporal resolution information due to the difficulty in collecting rich long-term time-series data. To bridge this gap, current clinical databases can be enhanced by connecting with mobile health platforms, community centres, or elderly homes such that other information can be incorporated into the system to facilitate clinical descision-making and address unanswered clinical questions. One interesting direction will be to build patient-specific models using data already available in existing clinical databases, and, then, update the model with data that can be collected outside the hospitals. In particular, some chronic diseases are manifested with acute events that are unlikely to be predictable solely by sporadic measurements made within hospitals.

Taking thoracic aortic dissection, a relatively rare disease (3–4 per 100 000 people per year), as an example, the disease is typically manifested as a tear in the intimal layer of the aorta, which can later on develop into either type A (involving both ascending and descending aorta) or type B dissection (involving descending aorta only). Type-A patients would require immediate surgical intervention, whereas for type B dissection, it is generally considered as a chronic condition requiring careful long-term control of blood pressure (BP).

Individuals with connective tissue disorders such as Marfan syndrome (MFS) are often more susceptible to aortic aneurysms or tears. Large-scale population screening for this rare disease will, therefore, be useful in identifying people who are at higher risk of developing aortic dissection. For a tear to develop into type A dissection, while others into type B dissection, one hypothesis would be that it is due to different flow patterns generated close to the tear location and across the aorta. Although an initial model built from imaging can give good insights into the problem, this does not take into account progressive hemodynamic variation over time and the impact of lifestyle and daily activities. By incorporating ambulatory BP profiles, it is possible to create simulation results as a lon-

gitudinal model spanning over a longer period of time for a better understanding of disease progression as summarized in Fig. 3.

B. Social Health

One of the primary tasks of telemedicine involves connecting patients and doctors beyond the clinic. However, this communication has been expanded, with the involvement of social networks, to new levels of social interaction. This new feature has opened up new possibilities of patient-to-patient communication regarding health beyond the traditional doctor-to-patient paradigm. One-fourth of patients with chronic diseases, such as diabetes, cancer, and heart conditions, are now using social network to share experiences with other patients with similar conditions, thereby providing another potential source of big data [15]. In addition to biological information, geolocation and social apps provide an additional feature to understand the behaviors and social demographics of patients, while avoiding resource intensive and expensive studies of large statistical sampling. This advantage has already been exploited by several epidemiological studies in areas, such as influenza outbreaks [16], [17], collective dynamics of smoking [18], and the misuse of antibiotics [19]. Text messages and posts on online social networks are also a valuable source of health information, e.g., for the better management of mental health. Compared to traditional methods, such as surveys, fluctuations and regulation of emotions, thoughts and behaviors analyzed over social network platforms, such as Twitter, offer new opportunities for the real-time analysis of expressed mood and its context [20]. For example, when validating against known patterns of variation in mood, the 2.73×10^9 emotional tweets collected over a 12-week period in a study reported by Larsen *et al.* [20] claimed to find some correlation between emotion tweets and global health estimates from the World Health Organization on anxiety and suicide rates.

Social media and internet searches can also be combined with environmental data, such as air quality data, to predict the sudden increase of asthma-related emergency visits [21].

Similar models are anticipated to help other areas of public health surveillance.

C. Lifestyle Environmental Factors, and Public Health

Climatological data, such as heat-stress and cold-related mortality, present another dimension to predict personal health [22], [23]. Recent remote sensing technologies and geographic information systems allow climate data for global land areas to be interpolated at a spatial resolution of 500 m to 1 km [24], [25]. Achieving high-resolution measurements are necessary so as to be able to monitor the real impact of pollution on human wellbeing in urban environments. With this aim the dense grid of wireless sensor networks facilitates the capture of spatiotemporal variability in toxic air pollutants [26]. Such technologies will become increasingly important for connecting epidemic intelligence with infectious disease surveillance and launching effective heat response plans [27]–[29]. Similarly, patterns of social factors influencing unhealthy habits such as smoking can be studied using the collective dynamics of social networks [18]. As an example of this, Christakis and Fowler found that smokers mostly belonged to the periphery of social networks, and by the time of quitting, they behaved collectively [18]. In addition, smokers with higher education tended to have a greater influence on their peers toward smoking behavior, compared to less educated smokers. As regards psychological states, emotional levels denoting hostility and stress, expressed in social media such as Twitter tweets, can serve as predictors of heart disease mortality per geographical area [30].

A mobile phone is an excellent platform to deliver personal messages to individuals to engage them in behavioral changes to improve health. Although at present, there is limited evidence that mobile messaging-based interventions support preventive health care for improving health status and health behavior outcomes [31], a better understanding of how this platform can be used is an interesting area to explore. For example, type-2 diabetes is generally thought to be preventable by lifestyle modification; however, successful lifestyle intervention programs are often labor intensive. It has been shown that mobile phone messaging can be used as an alternative to deliver motivational and educational advices for changing population lifestyles [32].

III. TRANSLATIONAL BIOINFORMATICS

Translational bioinformatics, a field that emerged after the first human genome mapping, focuses on bridging molecular biology, biostatistics, and statistical genetics with clinical informatics. The field is evolving at a tremendously fast pace, and many related areas have been proposed. Amongst them, pharmacogenomics is a branch of genomics concerned with individuals' variations to drug response due to genetic differences. The area is important for designing precision medicine in future.

New discoveries, resulting from the Human Genome Project, are now frequently applied to develop improved diagnostics, prognostics, and therapies for complex diseases, which is known as "translational genomics". In particular, the sequencing cost per genome has markedly reduced over the last decade, according to the data presented by the National Institutes of Health
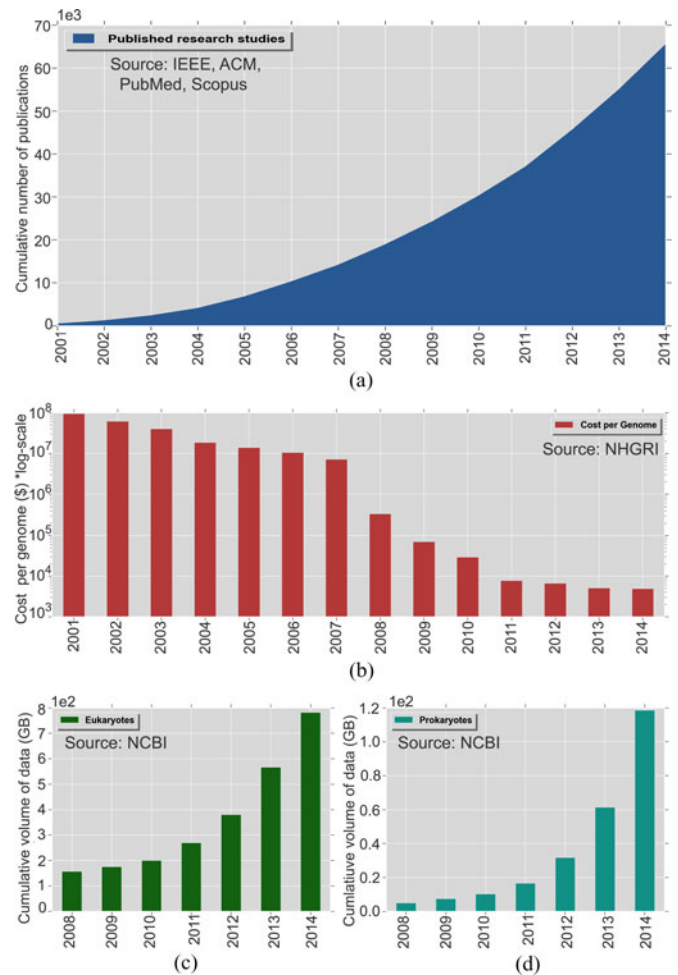

Fig. 4. a) Number of research studies sequencing DNA or genomes (source: PubMed, Web of Science, Scopus, IEEE, ACM). b) Sequencing cost per human-sized genome (source: National Human Genome Research Institute, NHGRI). Total volume of genomic data per year reported by completed studies for c) eukaryotes and d) prokaryotes in 1e2 GB (source: National Center for Biotechnology Information).

(NIH) Human Genome Research Institute as shown in Fig. 4. This further gives rise to new opportunities for personalized treatment and risk stratification.

On the other hand, research in bioinformatics has broadened from solely sequencing the genome of an individual to also measuring epigenomic data (i.e., above the genome), which include processes that alter gene expression other than changes of primary DNA sequences, such as DNA methylation and histone modifications. Information technologies for acquiring and analyzing biological molecules other than the genome, for example, transcriptome (the total mRNA in a cell or organism), proteome (the set of all expressed proteins in a cell, tissue, or organism), and metabolome (the total quantitative collection of low molecular weight compounds, metabolites, present in a cell or organism that participate in metabolic reactions) are also needed for future advances in the field. To summarize, OMICS aims at collectively characterizing and quantifying groups of biological molecules that translate into the structure, function, and dynamics of an organism. The OMICS profile of each

Fig. 5. Outline of the "OMICS" approach for studying disease mechanisms. OMICS aims at collectively characterizing and quantifying groups of biological molecules that translate into the structure, function, and dynamics of an orga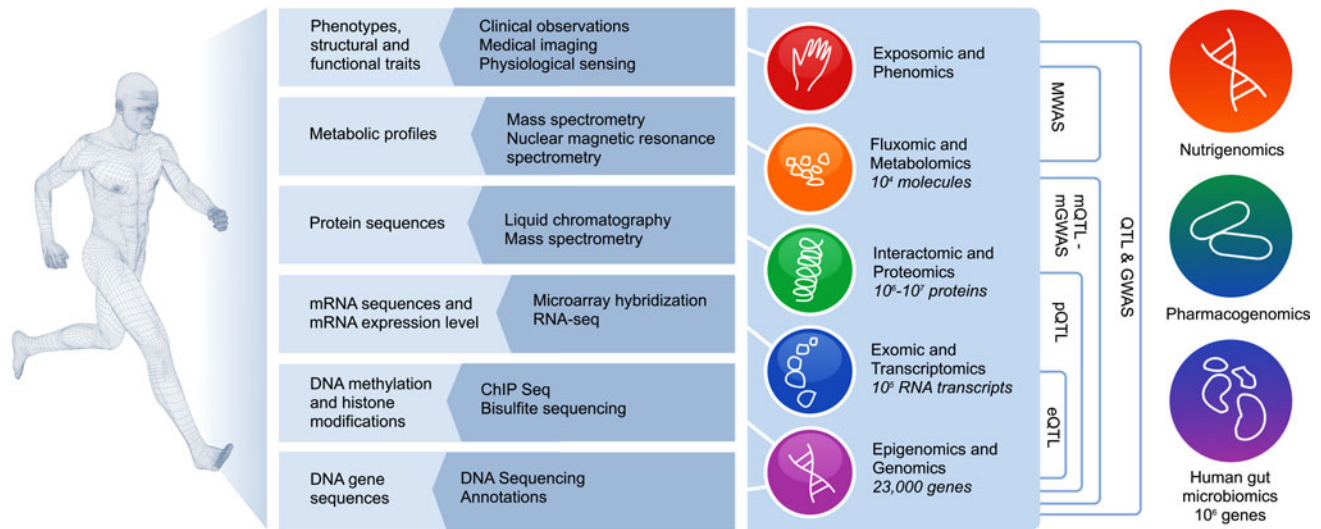nism. The OMICS profile of each individual, including the genome, transcriptome, proteome, and metabolome, should be eventually linked up with phenotypes obtained from clinical observations, medical images, and physiological signals. Different acquisition technologies are required to collect data at each biological level. Interaction within each level and across different levels as well as with the environment, including nutrition, food, drugs, traditional Chinese medicine, and gut microbiome presents grand challenges in future bioinformatics research.

individual should eventually be linked up with phenotypes obtained from clinical observations, medical images, and physiological signals (see Fig. 5).

A. Pharmacogenomics

A single whole human genome obtained by the next-generation sequencing (NGS) is typically 3 GB. Depending on the average depth of coverage, this can vary up to 200 GB, making it a clear source of big data for health. Nevertheless, only about 0.1% of the genome is different amongst individuals, which accounts for roughly 3 million variants. From a signal processing point of view, the data can be considered as highly compressible; however, in practice, compressed genotyping is not widely adopted at present.

Whole genome sequencing by NGS is important to the study of complex diseases such as cancer. It has been a long-standing problem in cancer treatment that drugs often have heterogeneous treatment responses even for the same type of cancer, and some drugs only show profound sensitivity in a small number of patients [33], [34]. Currently, large-scale personal genomics and pharmacogenomics datasets have been generated to uncover unique signalling patterns of individual patients and discover drugs that target these unique patterns. These include cancer cell line databases of nonspecific cancer cell types [35], [36] or a specific cancer cell type such as breast cancer [37]. The Cancer Genome Atlas Project of the NIH has tested the personal genomic profiles of over 10 000 individuals across over 20 types of cancer [38], and uncovered new cancer subtypes based on those profiles [39]. Patients with distinct genomics aberrations are believed to be responsible for the variability of drug response [40]. Large-scale datasets as such can be used to enable drug repositioning [41], [42], predict drug combinations [43], [44], and delineate mechanisms of action [45]. They are

becoming an important component in drug development [46], [47]. It is, therefore, possible to design precision medicine for individual patients based on their genomics profiles.

Pharmacogenomics has gone beyond studying individuals' drug response based on genome characteristics (e.g., copy number variations and somatic mutations) and now incorporates additional transcriptomic and metabolic features such as gene expression, considering factors that influence the concentration of a drug reaching its targets and factors associated with the drug targets. Since the gene expression profiles of cell lines are known to vary considerably in the process of prolonged culture under different culture conditions and techniques, the use of gene expression from cell lines for prediction of drug response in the patient is currently controversial. A recent algorithm for predicting *in vivo* drug response with the patient's baseline gene expression profile achieved 60%–80% predictive accuracy for different cases [48]. Other research [49], [50] studied drug response using immunodeficient mice xenografted with human tumors, which have the advantage of potentially studying both genetic and nongenetic factors that affect cancer growth and therapy tolerance [51].

Similar pharmacogenomics studies are also important to vascular diseases. Although antiplatelet agents such as clopidogrel are widely prescribed for diseases such as acute coronary syndrome (ACS), their responses vary greatly from person to person and approximately 30% of the patients may exhibit resistance to clopidogrel [52], 53]. Since clopidogrel is activated by the citocromo P450 (CYP) enzyme system to active metabolite, CYP2C19 loss-of-function (LOF) allele(s) affects the responsiveness of clopidogrel, but not the new antiplatelet agents (prasugrel and ticagrelor). Therefore, it is cost effective to use the genotype-guided method to screen carrier of CYP2C19 LOF allele(s) when using antiplatelets in high-risk ACS patients [54].

B. Translational Genomics

Although comprehensive genotyping is still relatively recent, it has a high potential for genetic stratification in patient screening, for instance, in the case of factors arising from genotyping, such as high-risk DNA mutations [55], milk and gluten intolerance, and muscovisciosis. Genetics combined with phenotypic information provided by EHR may help to provide greater insights into low penetrant alleles [56]. For example, it is well known that mutations of fibrillin 1 (FBN1) cause MFS. Nevertheless, the aetiology of the disease leads to marked clinical variability of MFS patients of the same family as well as different families [57]. Combining genetic tests of FBN1 and a series of related genes (TGFBR1, TGFBR2, TGFB2, MYH11, MYLK1, SMAD3, and ACTA2) will help to screen out patients who are more likely to develop aortic aneurysms that lead to dissections [58]. Further studies on these high-risk patients based on morphological images of the aorta may provide insight into the rate of disease development.

Another potential area for translational genomics is to study the gene networks of different syndromes of the same person in order to better understand how these syndromes are interrelated. For example, this has been used to study different genes on chromosome 21 (HSA21) and their role in Down's Syndrome (DS), as well as to understand the underlying reason why nearly half of DS patients exhibit an overprotection against cardiac abnormalities related to the connective tissue [59]. One hypothesis is based on the recent evidence that there is an overall upregulation of FBN1 in DS (which is normally down regulated in MFS) [59]. The construction of genetic networks will, therefore, provide a clearer picture of how these syndromes are related. By understanding the gene networks of the related syndromes, it may be possible to provide specific gene therapy for the related diseases.

C. OMICS and Large-Scale Databases

In addition to the Human Genome Project, several large-scale biological databases launched recently will further facilitate the study of disease mechanisms and progressions, particularly at the system level as outlined in Fig. 5. The Research Collaboratory for Structural Bioinformatics Protein Data Bank [60], [61] is a worldwide archive of structural data of biological macromolecules, providing access to the 3-D structures of biological macromolecules, as well as integration with external biological resources, such as gene and drug databases [62]. ProteomicsDB [63] is another example, encompassing mass spectrometry of the human proteome acquired from human tissues, cell lines, and body fluid to facilitate the identification of organ-specific proteins and translated long intergenic noncoding RNAs, with due consideration of time-dependent expression patterns of proteins [63].

Parallel to these developments, the Human Metabolome Database [64] consists of more than 40 000 annotated metabolites entries in the latest version released in 2013. It provides both experimental metabolite concentration data and analyses through mass spectrometry and Nuclear Magnetic Resonance (NMR) spectrometry [64]. Databases as such are believed to greatly facilitate the translation of information into knowledge for transforming clinical practice, particularly for metabolic-
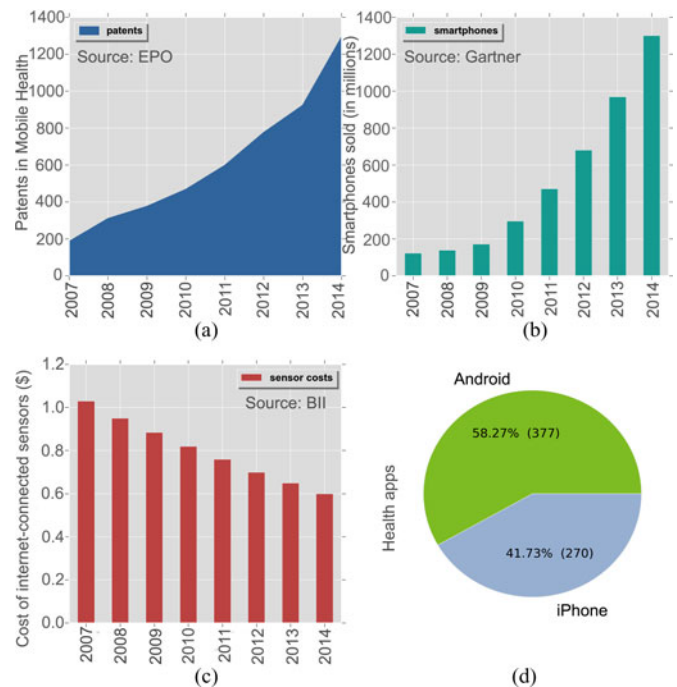

Fig. 6.   a) Evolution of the number of patents published in the area of mobile health (source: European Patent Office); b) evolution of the number of smartphones sold per year in million units (source: Gartner); c) evolution of the cost of Internet-enabled sensors in dollars (source: Business Intelligence International); d) number of mobile health apps published in Google play and iTunes as of May 2015.

related diseases, such as diabetes and coronary artery diseases [65]. In fact, metabolomics has emerged as an important research area that does not only include endogenous metabolites of the human body but also chemical and biochemical molecules that can interact with the human body [66]. Specifically, ongoing efforts have been placed for fingerprinting metabolites from food and nutrition products [67], drugs [68], and traditional Chinese medicine [69], as well as molecules produced by the gut bacterial microbiota [67], [70]. These will eventually help us to better understand the interaction between the host, pathogen and environment.

The availability of the genomic, proteomic, and metabolic databases allows a better understanding of the development of complex diseases such as cancer. They also allow the search of new biomarkers using different pattern mining and clustering techniques [68]–[71]. The clusters can be either partitional (hard) or hierarchical (tree-like nested structure). These methods can be further accelerated by using multicore CPU, GPU, and field-programmable gate arrays with parallel processing techniques.

IV. SENSOR INFORMATICS

Advances in sensing hardware have been accelerating in recent years and this trend shows no signs of slowing down [72]. According to the analysis in the BI Intelligence report (Garner) published at the end of 2014, the price of one MEMS sensor has decreased by half from US\$ 1.30 to US\$ 0.60 during the last decade as shown in Fig. 6. This has partly driven a paradigm shift of future internet applications toward what
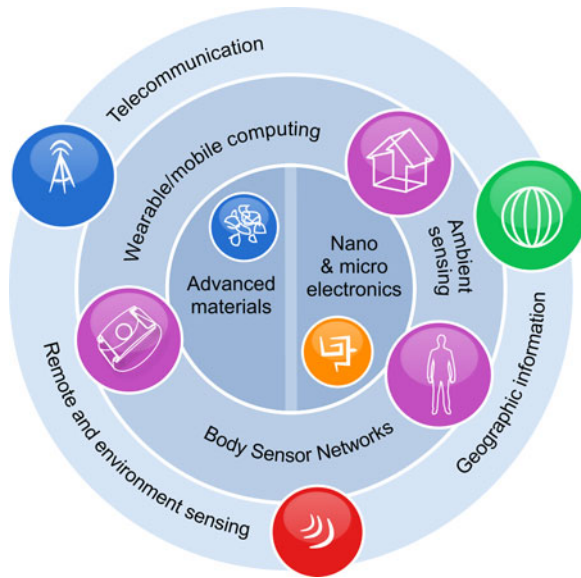
Fig. 7. Big sensing data in health are all around us, enabled by technologies ranging from nano- and microelectronics, advanced materials, wearable/mobile computing, and telecommunication systems as well as remote sensing and geographic information systems. The inner loop presents technologies for sensor components, while the middle loop presents devices and systems potentially own by each individual or household. The outer loop presents sensing technologies required at the community and public health level.

is termed "the Internet of Things" (IoT). Moreover, enabling technologies ranging from nano- and microelectronics, advanced materials, wearable/mobile computing, and telecommunication systems, as well as remote sensing and geographic information systems have made it possible for sensing health information to be collected pervasively and unobtrusively [73] as illustrated in Fig. 7.

A. Wearable, Implantable, and Ambient Sensors

As outlined in a recent review article [15], three factors, in particular, have contributed to the rapid uptake of wearable devices. These include increased data processing power, faster wireless communications with higher bandwidth, and improved design of microelectronics and sensor devices [15]. Example platforms include earlier systems with limited connectivity and single sensing elements developed solely for use in research laboratories to more recent ambient sensors as well as easy-to-wear wearable/implantable devices equipped with *continuous* multimodal sensing capabilities and support for data fusion deployed in a wide range of clinical applications [74]–[76]. Furthermore, parallel developments in miniaturized sensor embodiment, microelectronics and fabrication processes, and the availability of wireless power delivery have made *miniaturized* implantable sensors increasingly versatile [73].

Implantable sensors address the challenges of both acute and chronic disease monitoring by providing a means of capturing critical events and continuous streamlining of health information. Recent advances in microelectronics and nanotechnology have greatly improved the sensitivity of different sensors. For example, based on metal nanoparticle arrays and single

nanoparticles, the sensitivity of localized surface plasmon resonance optical sensors can be pushed toward the detection limit of a single molecule [77]. This has enabled the development of the next generation of high-throughput sequencing technologies, as well as the detection of biomolecules, such as glucose, lactate, nitric oxide, and sodium ions [78]. For diabetic patients, a myriad of new sensors for both wearable and implantable applications have been developed, which provide continuous monitoring and corresponding response to the time-varying glucose level, which is well known to be diet dependent [79]–[81].

There is a clear trend of moving from the scenario where a centralized large computing infrastructure is shared between multiple users toward one where each individual possesses multiple smart devices, most of which are sufficiently small to be wearable or implantable such that the use of these sensing devices will not affect normal daily activities. These sensor systems have the potential to generate datasets which are currently beyond our capabilities to easily organize and interpret [82]. Meanwhile, healthcare services delivered via ambient intelligence consisting of *ambient sensors* and objects interconnected into an integrated IoT represent a promising and supportive solution for the ageing society. It is important that such systems should take into account the sensor, service, and system integration architecture [83]. Such distributed systems require decentralized inference algorithms, which are frequency explored, either in the framework of parametric models, in which the statistics of phenomena under observation are assumed to be known by the system designer, or nonparametric models, when the underlying data is sparse and prior knowledge is limited [84], [85].

B. From Sensor Data to Stratified Patient Management

Physiological sensing by these smart devices can be long term and continuous, imposing new challenges for interpreting their clinical relevance. For example, the current clinical practice defines hypertension based on measurements taken during infrequent hospital visits. Although automated oscillometric BP measurement devices are now available, studies in these areas are often limited to taking BP once every hour over a 24-hr period. With the newly emerging ambulatory devices [75], a comprehensive BP-related profile of an individual can be made available. Nevertheless, the interpretation of these data is nontrivial, since in many situations, they may not be equivalent to the clinical BP readings that are currently being used by practitioners [86], [87]. The signals, however, carry underlying physiological meanings that, if properly processed and managed, can be used as additional information for understanding uncontrolled hypertension or to enhance the current hypertension management schemes. In addition to vital sign monitoring, smart implantable sensors provide a promising technology to monitor postoperative complications, such as slow tissue healing and infections. Moreover, smart implants can also have a reactive role by delivering drugs for chronic pain [88] and acting as brain stimulators for neurological diseases including refractory epilepsy [89] and Parkinson's disease [90]. This makes
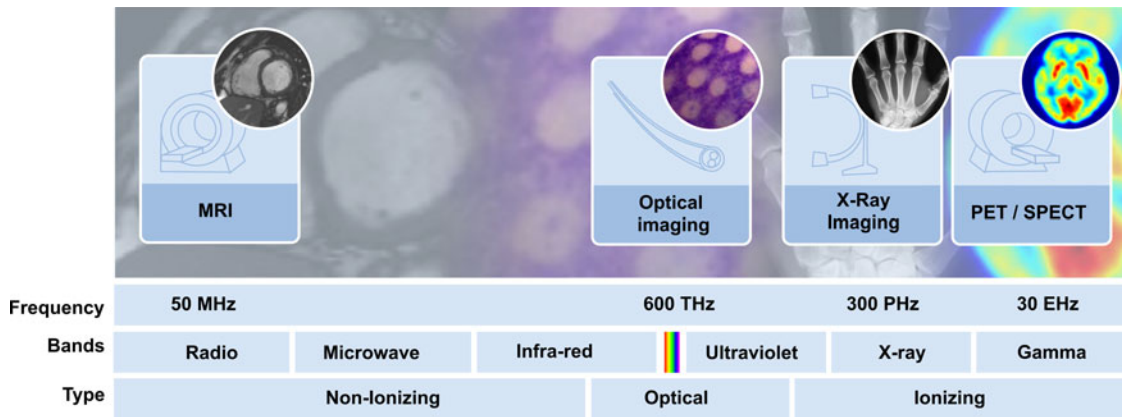
Fig. 8.   Different imaging modalities across the electromagnetic spectrum. They are playing an increasingly important role in early diagnosis, treatment planning, and deploying direct therapeutic measures.

smart implants not just another resource for data collection but also an integral part of early intervention.

With increased volume and acquisition speed of data from both wearable and implantable sources, new automated algorithms are needed to reduce false alarms such that they are sufficiently robust to support large-scale deployment, particularly for free-living environments [91]. Automatic classifications are necessary since the dataset sizes are beyond the capability of manual interpretation within a reasonable time period. New compression-based measures are, therefore, proposed as high-quality cloud computing services to reduce the computation time for the automated classification of different types of cardiac arrhythmia [92]. In many situations, measurements must be interpreted together with the context under which the data is collected. For example, many physiological parameters, such as BP or episodes of gastroesophageal reflux disease are posture dependent [93], [94], which can be captured by inertial sensors. Therefore, multimodal integration and context awareness are essential to the analysis of pervasive sensing data.

C. Mobile Health

Nowadays, smart phones have become an inseparable companion for more than 1.75 billion users. The data generated by the use of smart phones provides highly descriptive and continuous information anytime and anywhere. The penetration of smartphones, which has reached over 200% of the total population in some cities such as Hong Kong, makes it logical to use it as a personal logging device of health information. The new generation of smartphones has a wide range of health apps with standardized protocol to connect to sensors provided by different companies. They can potentially serve as a platform to centralize health data, from which additional new information that was previously untraceable by individual sensors can now be mined. In fact, earlier versions of mobile phones consist of only simple motion sensors, while newer models are packed with sophisticated sensors that facilitate the extraction of different types of vital signs, even without the need for external devices. These sensors, when properly used, can provide valuable health information for the management of many long-term

illnesses. For example, the video cameras of mobile phones can be used to collect heart rate and heart rate variability [95], embedded accelerometers and gyroscopes to track energy expenditure [96]. Furthermore, the pulse transmission time as measured by time delays between electrocardiographic and photoplethysmographic sensors can be used as a surrogate measure for BP [75], [97]. This information can be calculated from two devices that connect with a mobile phone independently, one with an electrocardiographic sensor and the other one with a photoplethysmographic sensor. When connected to health providers, a closer level of interaction in healthcare can be maintained toward greater personalization and responsiveness [98].

V. IMAGING INFORMATICS

The ever-increasing amount of annotated and real-time medical imaging data has raised the question of organizing, mining, and knowledge harvesting from large-scale medical imaging datasets. While established imaging modalities are getting pervasive, new imaging modalities are also emerging. These modalities are rapidly filling up the entire EM spectrum as shown in Fig. 8. Many of these imaging techniques are now geared toward real-time *in situ* or *in vivo* applications, making multimodality imaging an exciting yet challenging big data management problem.

Recent developments in imaging are progressing in multiple frontiers. First, there is relentless effort in making existing imaging modalities faster, higher resolution, and more versatile. Take cardiovascular magnetic resonance imaging (MRI) as an example, imaging sequences are no longer limited to morphological and simple tissue characterization (e.g., via T1, T2/T2* relaxation times). Details concerning vessel walls, myocardial perfusion and diffusion, and complex flow patterns *in vivo* can all be captured. When facilitated with new minimally invasive interventional techniques, novel drugs and other forms of treatment, MRI now serves as a therapeutic and interventional aid, rather than solely a diagnostic modality. Similar advances can also be appreciated for ultrasound, computed tomography (CT), and other imaging modalities. Moreover, extensive efforts in combining different imaging modalities, not by postprocessing,

but at the hardware level, e.g., MRI/PET and PET/CT, open up a range of new opportunities, particularly for oncological imaging and targeted therapy.

A. Imaging Across Scales

There have been extensive research efforts for developing new technologies that probe deeper into the biological system, from tissue (up to micrometer) to the protein level (micronanometer). In particular, recent advances in stimulated emission depletion fluorescence microscopy allow the generation of 3-D superresolution images of living biological specimens [99]. It overcomes the classical optical resolution limit of light microscopy and pushes the spatial resolution of optical microscope toward the nanoscale [100]–[102]. This opens up the possibility of imaging not only the fine morphological structure of many organ systems (e.g., microfibrils that form blood vessels), but also subcellular behavior and molecular signaling. The use of quantum dots or *qdots* also pushes the boundary of imaging resolution, allowing the study of intracellular processes at molecular levels (20–40 nm) [103]. Another class of fluorescent labels is made by conjugating qdots with biorecognition molecules, which emission wavelength can be tuned by changing the particle size such that a single light source can be used for simultaneous excitation of all different-sized dots [104]. These technologies have already been used for immunofluorescence labeling of tissues, fixed cells, and membrane proteins, such as cancer markers [105], the hybridization of chromosomes [106], the labeling of DNA [107], and contrast-enhanced image-guided resection of tumors [108].

B. From Morphology to Function

The understanding of many biological processes requires the identification and representation of structure–function relationships. This expands across different spatial scales, namely proteins, cells, tissues, and organs. For instance, haemodynamic analysis combined with contractile analysis, substantiated with myocardial perfusion data, can be used to elucidate the underlying factors associated with cardiac abnormalities. Starting with modeling, the tissue and scaling up toward a more specific description of organ behaviors has made it possible to create integrative models of heart function [109], [110]. These architectural models fuse information such as fibrous-sheet geometrical models of tissue and membrane currents from ion channels at the subcellular level [111].

Amongst all organs that have been studied to define their function from its morphology, the brain is the one that has received the most attention recently. This is motivated by the fact that brain structure and function are keys to understand cognitive processes, hence the need for unveiling neuronal behavior from the molecular level up to the functioning of neural circuits. Super-resolution fluorescence microscopy has been applied to study neural morphology and their subcellular structures. These techniques may enable us to achieve a resolution as high as 20 nm [112]. Needless to say, the myriad of markers necessary for each single type of cell and synapse would result in an enormous database.

Methods, such as functional MRI (fMRI) and functional diffusion tension imaging provide flexible information in the form of macrostructural, microstructural, and dense connectivity matrices. Improved fMRI sampling methods produce time-series data of multiple blood oxygenation-level-dependent volumes of the brain [113]. In addition, there is an increasing trend in making neuroimaging multimodal. In some studies, several modalities are used to compensate the benefits and tradeoff of one another. Furthermore, information from lower cost and rapid noninvasive methods, such as wearable electroencephalography and functional near-infrared spectroscopy allows gathering brain functional data for examining cortical responses due to more complex tasks.

An indirect way of inferring function consists of a combination of imaging modalities as well as medical records, demographics, and lab test results. In order to maximize the information contained in these heterogonous sources, linking different metadata with features extracted from image modalities is key to characterize the structure, function, and progression of diseases. Solving this challenge presents a unique opportunity for bridging the semantic gap between images and more effective prediction, diagnosis, and treatment of diseases. However, this issue entails many independent yet interrelated tasks, such as generating, segmenting, and extracting enormous amounts of quantifiable spatial objects and features (nuclei, tissue regions, blood vessels, etc.). This requires the implementation of effective and optimized querying systems [114] in order to reduce the computational complexity of handling these data. Fig. 9 represents a schema of what big data means for imaging, as defined by both structural and functional data.

Existing efforts in improving the spatiotemporal constraints of brain imaging are significantly increasing the computational resources needed for neuroimaging studies. RAM memory is an important resource for neuroimaging analysis. For instance, to perform subject-, voxel- and trial-level analysis, a significant amount of fMRI images needs to be loaded into memory. Fig. 10 illustrates the evolution of the required amount of RAM reported by neuroimaging-related studies in pubmed.org. From 2013 onwards, there has been a fast increase in the amount of RAM reported (from 8 to 60 GB). If this trend is confirmed, the amount of memory used in a study could reach values of around 260 GB by 2020.

C. Research Initiatives to Understand the Human Brain

Another active topic in imaging is to study the functional connectivity of the human brain, which is fundamental to both basic and applied neurobiological research [115]. Both U.S. and European Union (EU) have launched large-scale Human Brain projects in recent years with an aim to unravel the organ's complexity. The NIH-funded Human Connectome Project (HCP), with a funding scale of 30 million US$, aims at leveraging the latest advances in DTI to study brain areas in relation to their functional, structural, and electrophysiological connectivity [116]. The idea behind the HCP is that neural connectivity is as unique as the fingerprint to each individual. Genetics, environmental influences, and life experience are factors contributing to the formation of each individual's neural circuitry [117]. This is
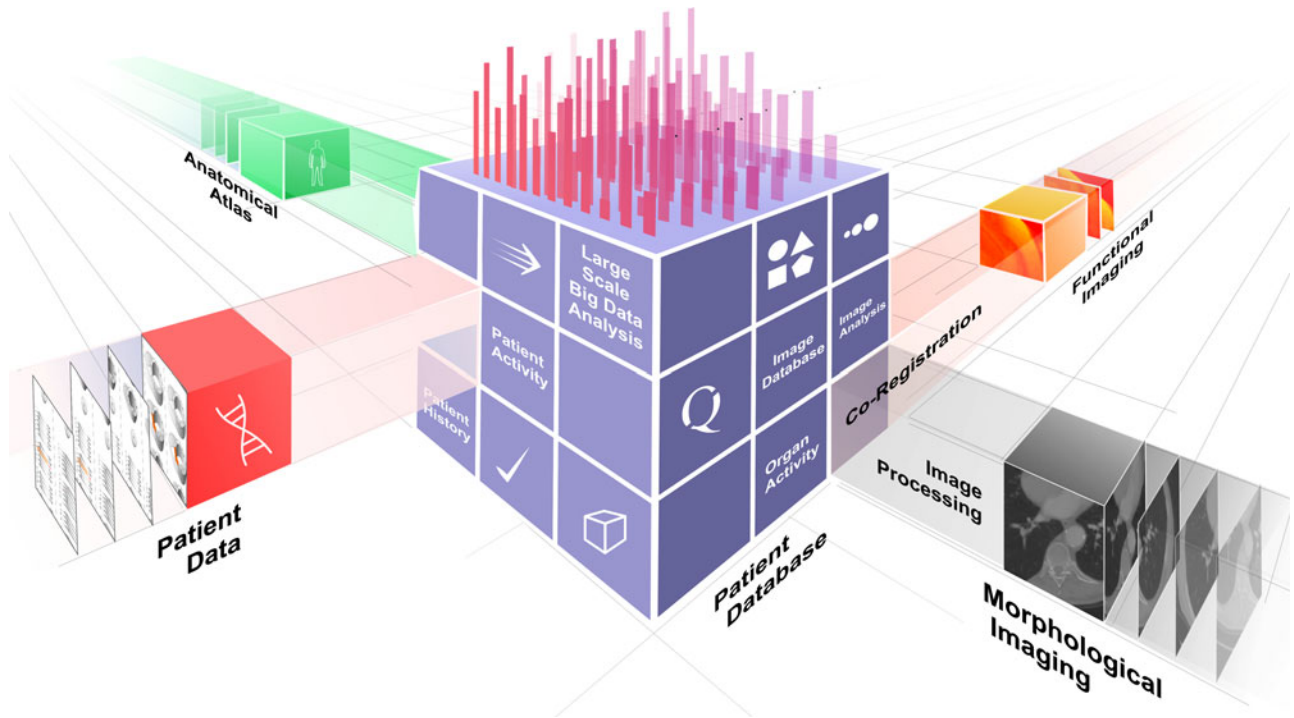
Fig. 9. Processing schema of imaging toward big data. Nonfunctional medical imaging is acquired and processed to serve as a model to register organ activity in the resulting functional imaging. Results from image processing and functional imaging are stored in databases with specific metadata protocols. Large-scale big data analysis is performed in these databases linking then the features extracted through medical imaging processing and functional imaging.
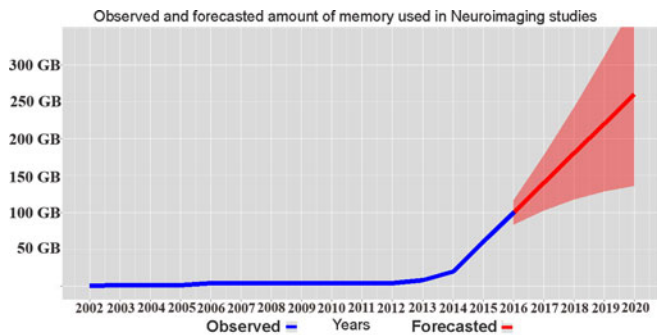

Fig. 10. Amount of RAM needed and forecasted to be used in neuroimaging studies.

supported by genome-wide association studies that link genetic variants with neurological and psychiatric disorders that have abnormal brain connectivity, e.g., variants at human clusterin (CLU) on chromosome 8 and complement receptor 1 on chromosome 1 are associated with Alzheimer's disease [118] as well as those specific markers associated with schizophrenia [119] and dementia [120].

Recently, the EU commission is providing €1.1 billion for the human brain project, which aims to develop a biological model of the brain that simulates different aspects of the nervous system, including point neuron models, neural circuitry, and cellular models at different scales. The main idea is to provide a simulation platform for theoretical neuroscientists to study how the brain processes information. For this purpose, it would require simulating all functions, architecture, and chemical properties for the 86 billion neurons and trillions of synapses of the human brain as estimated by Azevedo *et al*. [121]. There-

fore, the aim of the project is both ambitious and controversial. A panel review disclosed earlier this year, after the project had been launched for 18 months, urges the project team to adjust its governance and scientific direction [122]. Specifically, the report emphasizes that it is overambitious for whole-brain simulation, and that the project should consider the perspective of other sciences involved in the study of how the brain works, such as neuropsychology or neuroimaging. It is hope that the adjusted aim could complement the U.S. BRAIN initiative [122].

VI. DISCUSSION

According to International Data Corporation, worldwide spending on information and communication technology will reach 5 trillion US$ by 2020, and at least 80% of the growth will be driven by platform technologies, which encompass mobile technology, cloud services, social technologies, and big data analytics. Table I shows a selection of studies illustrating the potential of applying big data to health and the considerable increase in data complexity and heterogeneity in the field.

Applying big data to health is not only important to biological and physical sciences, but equivalently important to what has traditionally been considered as "soft" sciences, such as behavior and social sciences [123]. It is well known that human behavior is a significant driver for environmental problems, such as climate change, air pollution, and medical issues. Nevertheless, there are few studies that actually study these issues systematically and quantitatively. With the advanced technologies reviewed in this paper, it is now possible to study human behavior, including their physical actions, observable emotions, personality, temperament, and social interaction patterns, all of

Area	Sample	Methods	Data type	Ref
B	2708 subjects	Biostatistics	Gene expression data	[125]
HI (EHR)	2974 patients	Machine Learning (NLP)	Patient records and laboratory results	[126]
HI (EHR)	42 160 control 8549 patients	Statistics	Categorical database of patient records	[127]
B	876* subjects	Genomics	Gene expression data	[128]
S	200* patients	Machine Learning	Wearable sensor and annotation data	[129]
HI	3000 animal sample	Statistics	Veterinary records of health assessment	[130]
HI	745 053 patients	Machine Learning	Preoperative risk data and patient records	[131]
IMG	1414 subjects	Network Analysis	Resting state of neural fMRI data	[132]
IMG EHR	228* patients	Machine Learning	PET scans and patient records	[133]
HI (SN and ENV)	465 million records	Machine Learning	Social network and air quality data	[20]
HI (SN)	686 003 Social network users	Machine Learning (NLP)	Emotions in users' news feeds during 20 years	[134]

Acronyms: B (Bioinformatics), HI (Health Informatics), S (Sensing), IMG. (Imaging), EHR (Electronics Health Records), ENV (Environmental data), SN (Social Network), NLP (Natural Language Processing).

*Although these samples do not make more than 1000 instances, they can be considered large for the particular area of study.

which are conventionally difficult to measure and quantify. This will further help us to understand the mechanisms of disease development, and how these diseases spread and affect one another at the community level.

Health informatics applications are known to generate datasets that are complicated to store, untangle, organize, process, and, above all, interpret. From a scientific perspective, studies with a limited cohort of patients and controls can only serve as a proof-of-concept for future treatments and diagnoses.

Close to 3000 scientific studies indexed in pubmed.org since 2005 state that their conclusions should be "interpreted with caution" due to issues relative to statistical sampling. Large longitudinal and multimodal studies are necessary to discover the causes, risk, and improvement factors of several health diseases, such as cancer, Parkinson's, Alzheimer, and arthritis.

It must be emphasized that the interpretation of big data should be handled with care in all situations. In particular, proven cases show large discrepancies between the predicted and actual values. After all, predicting the future is always difficult. Despite its early success, Google Flu Trend (GFT) in 2013 was predicting more than twice the proportion of doctor visits for influenza-like illness than that of the Centers for Disease Control and Prevention [124]. There were a number of attributes to this problem which should be avoided in future studies in this area. First, the quality of the data collected should not be comprised with the quantity of the desired data. In many problems that researchers are dealing with, the number of parameters considered in a model may be exemplarily overfitting. Thus, the trained model was unable to predict future trends in this example because it put too much focus on the idiosyncrasies of the data at hand. Moreover, specific datapoints (outliers) may dominate in the trained model and those may have no predicting values. For the case of GFT, the nonseasonal 2009 influenza A–H1N1 pandemic was also incorporated in the model, which makes it partly a flu detector and partly a winter detector. Second, algorithm dynamics can induce errors in the prediction, particularly for analyzing big data. Often, both the data collected and the algorithms are changing at different paces. Capturing a specific instance can, therefore, be difficult due to the enormous amount of variations.

A. Processing Big Data

A bottleneck in analyzing big data is to obtain fast inference in real time from large and high-dimensional observations. For instance, high-dimensional spaces may arise from an extensive set of biomarkers [135], health attributes, and sensor fusion [136]. From a software point of view, processing big data is usually linked to parallel programming paradigms such as MapReduce [137]. Several open-source frameworks such as Hadoop have been considered to store distributed databases in a scalable architecture, as a basis for tools (e.g., Cascading, Pig, Hive) that allow developing applications to process vast amounts of data on commodity clusters. However, when combined with the continuous streams of pervasive heath monitoring data, this also requires capacities for iterative and low-latency computations, which depends on sophisticated models of data caching and in-memory computation.

In addition to the processing architecture, machine-learning-based data analysis also requires specific tuning to learn a classifier or repressor over large-scale datasets. Dimensionality reduction and feature selection can help us to cope with the curse of dimensionality. Nevertheless, whether supervised or unsupervised, these algorithms also require the regular implementation of a learning process to obtain a mapping or a set of maximally informative dimensions. Some machine learning methods, such as deep learning, involve learning several layered transformations of the data in order to find the best high-level abstraction for the problem at hand, mimicking the way neuroscience explains learning [138], [139]. Most machine learning techniques involve learning a set of model parameters that need to be found by means of optimization. The complexity of this learning process typically increases when dealing with big data. When the number of observations grows sample-by-sample iterative parameter learning methods can be a solution [140]. Another interesting option for scalable learning is to incrementally generate the set of required parameters or update the model structure whilst new data is being added [141], [142]. Online methods of variable selection and regularization are recommended to deactivate spurious variables in order to ease this scalability to large dimensions during learning [143], [144].

B. Data Privacy and Security

The emergence of big data for health raises additional challenges in relation to privacy, security, data ownership, stewardship, and governance. Personal data, which is regarded as the "New Oil" of the 21st century as coined during the 2011 World Economic Forum [145], are being generated at a tremendously fast speed due to the launch of many new intelligent devices, sensors, networks, and software applications. While these datasets often used to be generated and stored at a centralized location, today they are often distributed over various servers and networks. In the healthcare domain, data privacy is of utmost importance as regulated by laws in countries with large population. Closely related to the privacy issue is that data must be linked to the right person to ensure correct diagnosis and treatment. Therefore, the collected data about an individual must be uniquely tagged with an identifier. Furthermore, data security should be ensured at all levels of the healthcare system, including at the sensor level at which the data is collected [146].

VII. Conclusion

Big data can serve to boost the applicability of clinical research studies into real-world scenarios, where population heterogeneity is an obstacle. It equally provides the opportunity to enable effective and precision medicine by performing patient stratification. This is indeed a key task toward personalized healthcare. A better use of medical resources by means of personalization can lead to well-managed health services that can overcome the challenges of a rapidly increasing and aging population. Thus, advances in big data processing for health informatics, bioinformatics, sensing, and imaging will have a great impact on future clinical research. Another important factor to consider is rapid and seamless health data acquisition, which will contribute to the success of big data in medicine. Specifically, sensing provides a solid set of solutions to fill this gap. Frequencies of health data acquisition still involve a slow and complex process requiring the involvement of special health personal and laboratories. In this context, faster and unobtrusive health data can be provided by means of pervasive sensing. The use of sensors means the capacity of covering large periods of continuous monitoring without the need for performing sporadic screening, which may only represent a narrow picture of the development of a disease. However, the fact of deploying continuous sensing over a large population will result in a large amount of information that requires both on-node data abstraction and distributed inference. From a population level, one's unfortunate past can provide significant insight into forecasting and preventing the same incident from occurring in others. Last but not the least, the governmental policy and regulation are required to ensure privacy during data transmission and storage, as well as during subsequent data analysis tasks.

References

[1] M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. W. Treister, *Transforming Health Care Through Big Data*, Institute for Health Technology Transformation, Washington DC, USA, 2013.

[2] D. Laney, "3D data management: Controlling data volume, velocity and variety," Meta Group Inc., Stamford, CT, USA, Tech. Rep. 949, 2011.

[3] *Public Expenditure Statistical Analyses*, HM Treasury, London, U.K., 2012.

[4] G. A. Cuckler, A. M. Sisko, S. P. Keehan, S. D. Smith, A. J. Madison, J. A. Poisal, C. J. Wolfe, J. M. Lizonitz, and D. A. Stone, "National health expenditure projections, 2012–22: slow growth until coverage expands and economy improves," *Health Affairs*, vol. 32, pp. 1820–1831, 2013.

[5] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak, "Automated encoding of clinical documents based on natural language processing," *J. Amer. Med. Informat. Assoc.*, vol. 11, pp. 392–402, 2004.

[6] J. Sun, C. D. McNaughton, P. Zhang, A. Perer, A. Gkoulalas-Divanis, J. C. Denny, J. Kirby, T. Lasko, A. Saip, and B. A. Malin, "Predicting changes in hypertension control using electronic health records from a chronic disease management program," *J. Amer. Med. Informat. Assoc.*, vol. 21, pp. 337–344, 2014.

[7] G. N. Forrest, T. C. Van Schooneveld, R. Kullar, L. T. Schulz, P. Duong, and M. Postelnick, "Use of electronic health records and clinical decision support systems for antimicrobial stewardship," *Clin. Infectious Dis.*, vol. 59, pp. 122–133, 2014.

[8] R. Eriksson, T. Werge, L. J. Jensen, and S. Brunak, "Dose-specific adverse drug reaction identification in electronic patient records: Temporal data mining in an inpatient psychiatric population," *Drug Saf.*, vol. 37, pp. 237–247, 2014.

[9] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, pp. 1123–1131, 2014.

[10] M. R. Boland, G. Hripcsak, D. J. Albers, Y. Wei, A. B. Wilcox, J. Wei, J. Li, S. Lin, M. Breene, and R. Myers, "Discovering medical conditions associated with periodontitis using linked electronic health records," *J. Clin. Periodontol.*, vol. 40, pp. 474–482, 2013.

[11] H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, M. Levy, A. Shah, X. Han, and X. Ruan, "Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality," *J. Amer. Med. Informat. Assoc.*, pp. 1–10, 2014.

[12] Y. Hagar, D. Albers, R. Pivovarov, H. Chase, V. Dukic, and N. Elhadad, "Survival analysis with electronic health record data: Experiments with chronic kidney disease," *Statist. Anal. Data Mining, ASA Data Sci. J.*, vol. 7, pp. 385–403, 2014.

[13] T. Cars, B. Wettermark, R. E. Malmström, G. Ekeving, B. Vikström, U. Bergman, M. Neovius, B. Ringertz, and L. L. Gustafsson, "Extraction of electronic health record data in a hospital setting: Comparison of automatic and semi automatic methods using anti TNF therapy as model," *Basic Clin. Pharmacol. Toxicol.*, vol. 112, pp. 392–400, 2013.

[14] M. Marcos, J. A. Maldonado, B. Martínez-Salvador, D. Boscá, and M. Robles, "Interoperability of clinical decision-support systems and electronic health records using archetypes: A case study in clinical trial eligibility," *J. Biomed. Informat.*, vol. 46, pp. 676–689, 2013.

[15] J. Andreu-Perez, D. Leff, H. M. D. IP, and G.-Z. Yang, "From wearable sensors to smart implants—Towards pervasive and personalised healthcare," *IEEE Trans. Biomed. Eng.*, pp. 1–13, 2015, submitted for publication.

[16] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PLoS One*, vol. 6, pp. 1–10, 2011.

[17] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proc. 1st Workshop Soc. Media Analytics*, 2010, pp. 115–122.

[18] N. A. Christakis and J. H. Fowler, "The collective dynamics of smoking in a large social network," *N. Engl. J. Med.*, vol. 358, pp. 2249–2258, 2008.

[19] D. Scanfeld, V. Scanfeld, and E. L. Larson, "Dissemination of health information through social networks: Twitter and antibiotics," *Amer. J. Infection Control*, vol. 38, pp. 182–188, 2010.

[20] M. Larsen, T. Boonstra, P. Batterham, B. O'Dea, C. Paris, and H. Christensen, "We feel: Mapping emotion on Twitter," *IEEE J. Biomed. Health Informat.*, pp. 1–7, 2015, submitted for publication.

[21] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, "Predicting asthma-related emergency department visits using big data," *IEEE J. Biomed. Health Informat.*, pp. 1–8, 2015, submitted for publication.

[22] R. S. Kovats and S. Hajat, "Heat stress and public health: A critical review," in *Annual Review of Public Health*, vol. 29. Palo Alto, CA, USA: Annual Reviews, 2008, pp. 41–57.

[23] W. R. Keatinge, G. C. Donaldson, K. Bucher, G. Jendritsky, E. Cordioli, M. Martinelli, L. Dardanoni, K. Katsouyanni, A. E. Kunst, J. P. Mackenbach, C. McDonald, S. Nayha, and I. Vuori, "Cold exposure and winter mortality from Ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe," *Lancet*, vol. 349, pp. 1341–1346, May 1997.

[24] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, "Very high resolution interpolated climate surfaces for global land areas," *Int. J. Climatol.*, vol. 25, pp. 1965–1978, Dec. 2005.

[25] M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. M. Huang, "MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets," *Remote Sens. Environ.*, vol. 114, pp. 168–182, Jan. 2010.

[26] S. Moltchanov *et al.*, "On the feasibility of measuring urban air pollution by wireless distributed sensor networks," *Sci. Total Environ.*, vol. 502, pp. 537–547, 2015.

[27] B. Lobitz, L. Beck, A. Huq, B. Wood, G. Fuchs, A. S. G. Faruque, and R. Colwell, "Climate and infectious disease: Use of remote sensing for detection of Vibrio cholerae by indirect measurement," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 97, pp. 1438–1443, Feb. 2000.

[28] G. Luber and M. McGeehin, "Climate change and extreme heat events," *Amer. J. Prev. Med.*, vol. 35, pp. 429–435, Nov. 2008.

[29] J. C. Semenza and B. Menne, "Climate change and infectious diseases in Europe," *Lancet Infectious Dis.*, vol. 9, pp. 365–375, Jun. 2009.

[30] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, C. Weeg, E. E. Larson, L. H. Ungar, and M. E. Seligman, "Psychological language on Twitter predicts county-level heart disease mortality," *Psychol. Sci.*, vol. 26, pp. 159–69, Feb. 2015.

[31] V. Vodopivec-Jamsek, T. de Jongh, I. Gurol-Urganci, R. Atun, and J. Car, "Mobile phone messaging for preventive health care," *Cochrane Database Syst. Rev.*, vol. 12, pp. 1–44, 2012.

[32] A. Ramachandran, C. Snehalatha, J. Ram, S. Selvam, M. Simon, A. Nanditha, A. S. Shetty, I. F. Godsland, N. Chaturvedi, A. Majeed, N. Oliver, C. Toumazou, K. G. Alberti, and D. G. Johnston, "Effectiveness of mobile phone messaging in prevention of type 2 diabetes by lifestyle modification in men in India: A prospective, parallel-group, randomised controlled trial," *Lancet Diab. Endocrinol.*, vol. 1, pp. 191–198, 2013.

[33] A. R. Zlotta, "Words of wisdom: Re: Genome sequencing identifies a basis for everolimus sensitivity," *Eur. Urol.*, vol. 64, p. 516, Sep. 2013.

[34] G. Iyer, A. J. Hanrahan, M. I. Milowsky, H. Al-Ahmadie, S. N. Scott, M. Janakiraman, M. Pirun, C. Sander, N. D. Socci, I. Ostrovnaya, A. Viale, A. Heguy, L. Peng, T. A. Chan, B. Bochner, D. F. Bajorin, M. F. Berger, B. S. Taylor, and D. B. Solit, "Genome sequencing identifies a basis for everolimus sensitivity," *Science*, vol. 338, pp. 221–223, Oct. 2012.

[35] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. S. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. J. Zhou, F. Jewitt, T. H. Zhang, P. O'Brien, J. L. Boisvert, S. Price, W. Hur, W. J. Yang, X. M. Deng, A. Butler, H. G. Choi, J. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott, and C. H. Benes, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, pp. 570–575, Mar. 2012.

[36] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. W. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. Y. K. Yu, J. J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. X. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, pp. 603–607, Mar. 2012.

[37] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J Wang, M. Bansal, M. Ammad-ud-din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, NCI DREAM Community, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, and G. Stolovitzky, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnol.*, vol. 32, pp. 1202–1215, Dec. 2014.

[38] TCGA-web. (2015). [Online]. Available: https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp

[39] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, Oct. 2012.

[40] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The Cancer genome atlas pan-cancer analysis project," *Nature Genetic*, vol. 45, pp. 1113–1120, Oct. 2013.

[41] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte, "Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease," *Sci. Transl. Med.*, vol. 3, pp. 1–8, Aug. 2011.

[42] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Sci. Transl. Med.*, vol. 3, pp. 1–10, Aug. 2011.

[43] L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, and S. T. C. Wong, "DrugComboRanker: drug combination discovery based on target network analysis," *Bioinformatics*, vol. 30, pp. 228–236, Jun. 2014.

[44] J. Lee, D. G. Kim, T. J. Bae, K. Rho, J.-T. Kim, J.-J. Lee, Y. Jang, B. C. Kim, K. M. Park, and S. Kim, "CDA: Combinatorial drug discovery using transcriptional response modules," *PLoS One* vol. 7, no. 8, pp. 1–11, 2012.

[45] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. di Bernardo, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 14621–14626, Aug. 2010.

[46] V. v. van Noort, S. Scholch, M. Iskar, G. Zeller, K. Ostertag, C. Schweitzer, K. Werner, J. Weitz, M. Koch, and P. Bork, "Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene expression profiling," *Cancer Res.*, vol. 74, no. 20, pp. 5690–5699, 2014.

[47] N. S. Jahchan, J. T. Dudley, P. K. Mazur, N. Flores, D. Yang, A. Palmerton, A.-F. Zmoos, D. Vaka, K. Q. T. Tran, M. Zhou, K. Krasinska, J. W. Riess, J. W. Neal, P. Khatri, K. S. Park, A. J. Butte, and J. Sage, "A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors," *Cancer Discovery*, vol. 3, no. 12, pp. 1364–1377, Sep. 2013.

[48] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," *Genome Biol.*, vol. 15, no. 3, pp. 1–12, 2014.

[49] A. Prahallad, C. Sun, S. D. Huang, F. Di Nicolantonio, R. Salazar, D. Zecchin, R. L. Beijersbergen, A. Bardelli, and R. Bernards, "Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR," *Nature*, vol. 483, pp. 100–103, Mar. 2012.

[50] M. Nunes, P. Vrignaud, S. Vacher, S. Richon, A. Lievre, W. Cacheux, L. B. Weiswald, G. Massonnet, S. Chateau-Joubert, A. Nicolas, C. Dib, W. D. Zhang, J. Watters, D. Bergstrom, S. Roman-Roman, I. Bieche, and V. Dangles-Marie, "Evaluating patient-derived colorectal cancer xenografts as preclinical models by comparison with patient clinical data," *Cancer Res.*, vol. 75, pp. 1560–1566, Apr. 2015.

[51] A. Kreso, C. A. O'Brien, P. van Galen, O. I. Gan, F. Notta, A. M. K. Brown, K. Ng, J. Ma, E. Wienholds, C. Dunant, A. Pollett, S. Gallinger, J. McPherson, C. G. Mulligham, D. Shibata, and J. E. Dick, "Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer," *Science*, vol. 339, pp. 543–548, Feb. 2013.

[52] K. A. Kim, P. W. Park, S. J. Hong, and J. Y. Park, "The effect of CYP2C19 polymorphism on the pharmacokinetics and pharmacodynamics of clopidogrel: A possible mechanism for clopidogrel resistance," *Clin. Pharmacol. Therapeutics*, vol. 84, pp. 236–242, Aug. 2008.

[53] H. G. Xie, J. J. Zou, Z. Y. Hu, J. J. Zhang, F. Ye, and S. L. Chen, "Individual variability in the disposition of and response to clopidogrel: Pharmacogenomics and beyond," *Pharmacol. Therapeutics*, vol. 129, pp. 267–289, Mar. 2011.

[54] M. H. Jiang and J. H. S. You, "Review of pharmacoeconomic evaluation of genotype-guided antiplatelet therapy," *Expert Opinion Pharmacotherapy*, vol. 16, pp. 771–779, Apr. 2015.

[55] Y. A. Lussier and Y. Liu, "Computational approaches to phenotyping: High-throughput phenomics," *Proc. Amer. Thoracic Soc.*, vol. 4, pp. 18–25, 2007.

[56] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genetic*, vol. 13, pp. 395–405, 2012.

[57] S. Hutchinson, A. Furger, D. Halliday, D. P. Judge, A. Jefferson, H. C. Dietz, H. Firth, and P. A. Handford, "Allelic variation in normal human FBN1 expression in a family with Marfan syndrome: A potential modifier of phenotype?" *Hum. Mol. Genetics*, vol. 12, pp. 2269–2276, Sep. 2003.

[58] A. W. den Hartog, R. Franken, A. H. Zwinderman, J. Timmermans, A. J. Scholte, M. P. van den Berg, V. de Waard, G. Pals, B. J. Mulder, and M. Groenink, "The risk for type B aortic dissection in Marfan syndrome," *J. Amer. College Cardiol.*, vol. 65, pp. 246–254, 2015.

[59] M. Vilardell, S. Civit, and R. Herwig, "An integrative computational analysis provides evidence for FBN1-associated network deregulation in trisomy 21," *Biol. Open*, vol. 2, pp. 771–778, Aug. 2013.

[60] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res.*, vol. 28, pp. 235–242, Jan. 2000.

[61] P. W. Rose, A. Prlic, C. X. Bi, W. F. Bluhm, C. H. Christie, S. Dutta, R. K. Green, D. S. Goodsell, J. D. Westbrook, J. Woo, J. Young, C. Zardecki, H. M. Berman, P. E. Bourne, and S. K. Burley, "The RCSB protein data bank: Views of structural biology for basic and applied research and education," *Nucleic Acids Res.*, vol. 43, pp. 345–356, Jan. 2015.

[62] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlić, M. Quesada, G. B. Quinn, and J. D. Westbrook, "The RCSB protein data bank: Redesigned web site and web services," *Nucleic Acids Res.*, vol. 39, pp. 392–401, 2011.

[63] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J. H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster, "Mass-spectrometry-based draft of the human proteome," *Nature*, vol. 509, pp. 582–587, May 2014.

[64] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. F. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. G. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert, "HMDB 3.0-The human metabolome database in 2013," *Nucleic Acids Res.*, vol. 41, pp. 801–807, Jan. 2013.

[65] J. R. Bain, R. D. Stevens, B. R. Wenner, O. Ilkayeva, D. M. Muoio, and C. B. Newgard, "Metabolomics applied to diabetes research moving from information to knowledge," *Diabetes*, vol. 58, pp. 2429–2443, Nov. 2009.

[66] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: Acquiring and understanding global metabolite data," *Trends Biotechnol.*, vol. 22, pp. 245–252, May 2004.

[67] H. Tilg and A. R. Moschen, "Food, immunity, and the microbiome," *Gastroenterology*, vol. 148, pp. 1107–1119, 2015.

[68] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, pp. 264–323, Sep. 1999.

[69] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[70] S. K. Mazmanian, J. L. Round, and D. L. Kasper, "A microbial symbiosis factor prevents intestinal inflammatory disease," *Nature*, vol. 453, pp. 620–625, 2008.

[71] D. Z. Wang, R. C. C. Cheung, and H. Yan, "Design exploration of geometric biclustering for microarray data analysis in data mining," *IEEE Trans. Parallel Distributed Syst.*, vol. 25, no. 10, pp. 2540–2550, Oct. 2014.

[72] C. C. Poon and Y.-T. Zhang, "Perspectives on high technologies for low-cost healthcare," *IEEE Eng. Med. Biology Mag.*, vol. 27, no. 5, pp. 42–47, Sep/Oct. 2008.

[73] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 3, pp. 579–590, May 2013.

[74] Y. Zheng, X. Ding, C. C. Y. Poon, B. Lo, H. Zhang, X. Zhou, G. Yang, N. Zhao, and Y. Zhang, "Unobtrusive sensing and wearable devices for health informatics," *IEEE Trans. Biomed. Informat.*, vol. 61, no. 5, pp. 1538–1554, May 2014.

[75] Y.-L. Zheng, B. P. Yan, Y.-T. Zhang, and C. C. Y. Poon, "An Armband wearable device for overnight and cuff-less blood pressure measurement," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 7, pp. 2179–2186, Jul. 2014.

[76] G. Z. Yang, *Body Sensor Networks*, 2nd ed. New York, NY, USA: Springer, 2014.

[77] J. N. Anker, W. P. Hall, O. Lyandres, N. C. Shah, J. Zhao, and R. P. Van Duyne, "Biosensing with plasmonic nanosensors," *Nature Mater.*, vol. 7, pp. 442–453, Jun. 2008.

[78] G. S. Wilson and R. Gifford, "Biosensors for real-time in vivo measurements," *Biosens. Bioelectron.*, vol. 20, pp. 2388–2403, Jun. 2005.

[79] C. R. Yonzon, C. L. Haynes, X. Y. Zhang, J. T. Walsh, and R. P. Van Duyne, "A glucose biosensor based on surface-enhanced Raman scattering: Improved partition layer, temporal stability, reversibility, and resistance to serum protein interference," *Analytical Chem.*, vol. 76, pp. 78–85, Jan. 2004.

[80] R. Hovorka, "Continuous glucose monitoring and closed-loop systems," *Diab. Med.*, vol. 23, pp. 1–12, Jan. 2006.

[81] M. Breton, A. Farret, D. Bruttomesso, S. Anderson, L. Magni, S. Patek, C. D. Man, J. Place, S. Demartini, S. Del Favero, C. Toffanin, C. Hughes-Karvetski, E. Dassau, H. Zisser, F. J. Doyle, G. De Nicolao, A. Avogaro, C. Cobelli, E. Renard, and B. Kovatchev, "Fully integrated artificial pancreas in type 1 diabetes: Modular closed-loop glucose control maintains near normoglycemia," *Diabetes*, vol. 61, pp. 2230–2237, Sep. 2012.

[82] S. Redmond, N. Lovell, G. Yang, A. Horsch, P. Lukowicz, L. Murrugarra, and M. Marschollek, "What does big data mean for wearable sensor systems?: Contribution of the IMIA wearable sensors in healthcare WG," *Yearbook Med. Informat.*, vol. 9, pp. 135–142, 2014.

[83] Z. Pang, L. Zheng, J. Tian, S. Kao-Walter, E. Dubrova, and Q. Chen, "Design of a terminal solution for integration of in-home health care devices and services towards the Internet-of-things," *Enterprise Inf. Syst.*, vol. 9, pp. 86–116, 2015.

[84] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks," in *Proc. Annu. Allerton Conf. Commun., Control Comput.*, 2005, pp. 1–10.

[85] J.-J. Xiao, A. Ribeiro, Z.-Q. Luo, and G. B. Giannakis, "Distributed compression-estimation using wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 27–41, Jul. 2006.

[86] Q. Liu, B. P. Yan, C.-M. Yu, Y.-T. Zhang, and C. C. Y. Poon, "Attenuation of systolic blood pressure and pulse transit time hysteresis during exercise and recovery in cardiovascular patients," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 346–352, Feb. 2014.

[87] Y.-L. Zheng, B. P. Yan, Y.-T. Zhang, and C. C. Y. Poon, "Noninvasive characterization of vascular tone by model-based system identification in healthy and heart failure patients," *Ann. Biomed. Eng.*, 2015, to be published.

[88] D. C. Turk, "Clinical effectiveness and cost-effectiveness of treatments for patients with chronic pain," *Clin. J. Pain*, vol. 18, pp. 355–365, 2002.

[89] T. Loddenkemper, A. Pan, S. Neme, K. B. Baker, A. R. Rezai, D. S. Dinner, E. B. Montgomery, and H. O. Lüders, "Deep brain stimulation in epilepsy," *J. Clin. Neurophysiol.*, vol. 18, pp. 514–532, 2001.

[90] Deep-Brain Stimulation for Parkinson's Disease Study Group, "Deep-brain stimulation of the subthalamic nucleus or the pars interna of the globus pallidus in Parkinson's disease," *N. Engl. J. Med.*, vol. 345, pp. 956–963, 2001.

[91] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, Jan. 2013.

[92] J. M. Lillo-Castellano, I. Mora-Jimenez, R. Santiago-Mozos, F. Chavarria-Asso, "Symmetrical compression distance for arrhythmia discrimination in cloud-based big-data services," *IEEE J. Biomed. Health Inform.*, pp. 1–11, 2015, to be published.

[93] J. H. Wang, J. Y. Luo, L. Dong, J. Gong, and M. Tong, "Epidemiology of gastroesophageal reflux disease: A general population-based study in Xi'an of Northwest China," *World J. Gastroenterol.*, vol. 10, pp. 1647–1651, Jun. 2004.

[94] F. I. Caird, G. R. Andrews, and R. D. Kennedy, "Effect of posture on blood-pressure in elderly," *Brit. Heart J.*, vol. 35, pp. 527–530, 1973.

[95] M. Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Express*, vol. 18, pp. 10762–10774, May 2010.

[96] L. Atallah, B. Lo, R. King, and G.-Z. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 4, pp. 320–329, Aug. 2011.

[97] C. C. Y. Poon and Y. T. Zhang, "Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time," in *Proc. IEEE Eng. Med. Biol. Soc. Int. Conf.*, 2005, pp. 5877–5880.

[98] R. Atun, S. Jaffar, S. Nishtar, F. M. Knaul, M. L. Barreto, M. Nyirenda, N. Banatvala, and P. Piot, "Improving responsiveness of health systems to non-communicable diseases," *Lancet*, vol. 381, pp. 690–697, 2013.

[99] S. W. Hell and J. Wichmann, "Breaking the diffraction resolution limit by stimulated emission depletion fluorescence microscopy," *Opt. Lett.*, vol. 19, pp. 780–782, Jun. 1994.

[100] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging intracellular fluorescent proteins at nanometer resolution," *Science*, vol. 313, pp. 1642–1645, Sep. 2006.

[101] A. Sundaramurthy, P. J. Schuck, N. R. Conley, D. P. Fromm, G. S. Kino, and W. E. Moerner, "Toward nanometer-scale optical photolithography: Utilizing the near-field of bowtie optical nanoantennas," *Nano Lett.*, vol. 6, pp. 355–360, Mar. 2006.

[102] S. W. Hell, "Far-field optical nanoscopy," *Science*, vol. 316, pp. 1153–1158, May 2007.

[103] C. Dais, G. Mussler, T. Fromherz, E. Muller, H. H. Solak, and D. Grutzmacher, "SiGe quantum dot crystals with periods down to 35nm," *Nanotechnology*, vol. 26, no. 25, pp. 1–6, Jun 2015.

[104] W. C. Chan, D. J. Maxwell, X. Gao, R. E. Bailey, M. Han, and S. Nie, "Luminescent quantum dots for multiplexed biological detection and imaging," *Curr. Opinion Biotechnol.*, vol. 13, pp. 40–46, 2002.

[105] X. Wu, H. Liu, J. Liu, K. N. Haley, J. A. Treadway, J. P. Larson, N. Ge, F. Peale, and M. P. Bruchez, "Immunofluorescent labeling of cancer marker Her2 and other cellular targets with semiconductor quantum dots," *Nature Biotechnol.*, vol. 21, pp. 41–46, 2003.

[106] S. Pathak, S.-K. Choi, N. Arnheim, and M. E. Thompson, "Hydroxylated quantum dots as luminescent probes for in situ hybridization," *J. Amer. Chem. Soc.*, vol. 123, pp. 4103–4104, 2001.

[107] J. Farlow, D. Seo, K. E. Broaders, M. J. Taylor, Z. J. Gartner, and Y.-w. Jun, "Formation of targeted monovalent quantum dots by steric exclusion," *Nature Methods*, vol. 10, pp. 1203–1205, 2013.

[108] A. Mohs, M. Mancini, J. Provenzale, C. Saba, K. Cornell, E. Howerth, and S. Nie, "An integrated widefield imaging and spectroscopy system for contrast-enhanced, image-guided resection of tumors," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 5, pp. 1416–1424 May 2015.

[109] I. LeGrice, P. Hunter, and B. Smaill, "Laminar structure of the heart: A mathematical model," *Amer. J. Physiol.-Heart Circul. Physiol.*, vol. 272, no. 5, pp. 2466–2476, 1997.

[110] I. LeGrice, P. Hunter, A. Young, and B. Smaill, "The architecture of the heart: A data–based model," *Philosoph. Trans. Roy. Soc. London. Ser. A, Math., Phys. Eng. Sci.*, vol. 359, pp. 1217–1232, 2001.

[111] C.-h. Luo and Y. Rudy, "A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes," *Circulation Res.*, vol. 74, pp. 1071–1096, 1994.

[112] A. Dani and B. Huang, "New resolving power for light microscopy: applications to neurobiology," *Curr. Opinion Neurobiol.*, vol. 20, pp. 648–652, 2010.

[113] D. A. Feinberg, S. Moeller, S. M. Smith, E. Auerbach, S. Ramanna, M. F. Glasser, K. L. Miller, K. Ugurbil, and E. Yacoub, "Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging," *PLoS One*, vol. 5, no. 12, p. e15710, 2010.

[114] A. Aji, F. Wang, and J. H. Saltz, "Towards building a high performance spatial query system for large scale medical imaging data," in *Proc. 20th Int. Conf. Adv. Geograph. Inf. Syst.*, 2012, pp. 309–318.

[115] O. Sporns, G. Tononi, and R. Kotter, "The human connectome: A structural description of the human brain," *PLoS Comput. Biol.*, vol. 1, pp. 245–251, Sep. 2005.

[116] P. Kochunov, N. Jahanshad, D. Marcus, A. Winkler, E. Sprooten, T. E. Nichols, S. N. Wright, L. E. Hong, B. Patel, T. Behrens, S. Jbabdi, J. Andersson, C. Lenglet, E. Yacoub, S. Moeller, E. Auerbach, K. Ugurbil, S. N. Sotiropoulos, R. M. Brouwer, B. Landman, H. Lemaitre, A. den Braber, M. P. Zwiers, S. Ritchie, K. van Hulzen, L. Almasy, J. Curran, G. I. Dezubicaray, R. Duggirala, P. Fox, N. G. Martin, K. L. McMahon, B. Mitchell, R. L. Olvera, C. Peterson, J. Starr, J. Sussmann, J. Wardlaw, M. Wright, D. I. Boomsma, R. Kahn, E. J. C. de Geus, D. E. Williamson, A. Hariri, D. van 't Ent, M. E. Bastin, A. McIntosh, I. J. Deary, H. E. Hulshoffpol, J. Blangero, P. M. Thompson, D. C. Glahn, and D. C. van Essen, "Heritability of fractional anisotropy in human white matter: A comparison of human connectome project and ENIGMA-DTI data," *NeuroImage*, vol. 111, pp. 300–311, May 2015.

[117] J. M. Perkel, "Life Science Technologies: This is your brain: Mapping the connectome," *Science*, vol. 339, pp. 350–352, 2013.

[118] J. C. Lambert, S. Heath, G. Even, D. Campion, K. Sleegers, M. Hiltunen, O. Combarros, D. Zelenika, M. J. Bullido, B. Tavernier, L. Letenneur, K. Bettens, C. Berr, F. Pasquier, N. Fievet, P. Barberger-Gateau, S. Engelborghs, P. De Deyn, I. Mateo, A. Franck, S. Helisalmi, E. Porcellini, O. Hanon, M. M. de Pancorbo, C. Lendon, C. Dufouil, C. Jaillard, T. Leveillard, V. Alvarez, P. Bosco, M. Mancuso, F. Panza, B. Nacmias, P. Bossu, P. Piccardi, G. Annoni, D. Seripa, D. Galimberti, D. Hannequin, F. Licastro, H. Soininen, K. Ritchie, H. Blanche, J. F. Dartigues, C. Tzourio, I. Gut, C. Van Broeckhoven, A. Alperovitch, M. Lathrop, P. Amouyel, "Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease," *Nature Genetics*, vol. 41, pp. 1094–199, Oct. 2009.

[119] H. Stefansson, R. A. Ophoff, S. Steinberg, O. A. Andreassen, S. Cichon, D. Rujescu, T. Werge, O. P. H. Pietilainen, O. Mors, P. B. Mortensen, E. Sigurdsson, O. Gustafsson, M. Nyegaard, A. Tuulio-Henriksson, A. Ingason, T. Hansen, J. Suvisaari, J. Lonnqvist, T. Paunio, A. D. Borglum, A. Hartmann, A. Fink-Jensen, M. Nordentoft, D. Hougaard, B. Norgaard-Pedersen, Y. Bottcher, J. Olesen, R. Breuer, H. J. Moller, I. Giegling, H. B. Rasmussen, S. Timm, M. Mattheisen, I. Bitter, J. M. Rethelyi, B. B. Magnusdottir, T. Sigmundsson, P. Olason, G. Mason, J. R. Gulcher, M. Haraldsson, R. Fossdal, T. E. Thorgeirsson, U. Thorsteinsdottir, M. Ruggeri, S. Tosato, B. Franke, E. Strengman, L. A. Kiemeney, I. Melle, S. Djurovic, L. Abramova, V. Kaleda, J. Sanjuan, R. de Frutos, E. Bramon, E. Vassos, G. Fraser, U. Ettinger, M. Picchioni, N. Walker, T. Toulopoulou, A. C. Need, D. Ge, J. L. Yoon, K. V. Shianna, N. B. Freimer, R. M. Cantor, R. Murray, A. Kong, V. Golimbet, A. Carracedo, C. Arango, J. Costas, E. G. Jonsson, L. Terenius, I. Agartz, H. Petursson, M. M. Nothen, M. Rietschel, P. M. Matthews, P. Muglia, L. Peltonen, D. St Clair, D. B. Goldstein, K. Stefansson, and D. A. Collier, "Common variants conferring risk of schizophrenia," *Nature*, vol. 460, no. 7256, pp. 744–747, Aug. 2009.

[120] N. Jahanshad, P. Rajagopalan, X. Hua, D. P. Hibar, T. M. Nir, A. W. Toga, C. R. Jack, A. J. Saykin, R. C. Green, M. W. Weiner, S. E. Medland, G. W. Montgomery, N. K. Hansell, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, M. J. Wright, P. M. Thompson, "Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 110, pp. 4768–4773, Mar. 2013.

[121] Editorial, "Rethinking the brain," *Nature*, vol. 519, pp. 389–389, Mar. 2015.

[122] F. A. C. Azevedo, L. R. B. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. L. Ferretti, R. E. P. Leite, W. J. Filho, R. Lent, and S. Herculano-Houzel, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled–up primate brain," *J. Compara. Neurolog.*, vol. 513, no. 5, pp. 532–541, 2009.

[123] Editorial. "In praise of soft science," *Nature*, vol. 435, p. 1003, Jun 2005.

[124] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The Parable of Google Flu: Traps in big data analysis," *Science*, vol. 343, pp. 1203–1205, Mar 2014.

[125] K. Schramm, C. Marzi, C. Schurmann, M. Carstensen, E. Reinmaa, R. Biffar, G. Eckstein, C. Gieger, H.-J. Grabe, and G. Homuth, "Mapping the genetic architecture of gene regulation in whole blood," *PLoS One*, vol. 9, pp. 1–13, 2014.

[126] H. J. Murff, F. FitzHenry, M. E. Matheny, N. Gentry, K. L. Kotter, K. Crimin, R. S. Dittus, A. K. Rosen, P. L. Elkin, and S. H. Brown, "Automated identification of postoperative complications within an electronic medical record using natural language processing," *JAMA*, vol. 306, pp. 848–855, 2011.

[127] Á. Skow, I. Douglas, and L. Smeeth, "The association between Parkinson's disease and anti epilepsy drug carbamazepine: A case—Control study using the UK general practice research database," *Brit. J. Clin. Pharmacol.*, vol. 76, pp. 816–822, 2013.

[128] G. P. Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[129] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 3, pp. 722–730, May 2014.

[130] J. L. Nielson, C. F. Guandique, A. W. Liu, D. A. Burke, A. T. Lash, R. Moseanko, S. Hawbecker, S. C. Strand, S. Zdunowski, K. A. Irvine, J. H. Brock, Y. S. Nout-Lomas, J. C. Gensel, K. D. Anderson, M. R. Segal, E. S. Rosenzweig, D. S. Magnuson, S. R. Whittemore, D. M. McTigue, P. G. Popovich, A. G. Rabchevsky, S. W. Scheff, O. Steward, G. Courtine, V. R. Edgerton, M. H. Tuszynski, M. S. Beattie, J. C. Bresnahan, and A. R. Ferguson, "Development of a database for translational spinal cord injury research," *J. Neurotrauma*, vol. 31, pp. 1789–1799, Nov. 2014.

[131] J. E. Anderson and D. C. Chang, "Using electronic health records for surgical quality improvement in the era of big data," *JAMA Surg.*, vol. 150, pp. 24–29, Jan. 2015.

[132] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, and S. Colcombe, "Toward discovery science of human brain function," *Proc. Nat. Acad. Sci.*, vol. 107, pp. 4734–4739, 2010.

[133] A. Mikhno, F. Zanderigo, R. T. Ogden, J. J. Mann, E. D. Angelini, A. F. Laine, and R. V. Parsey, "Toward non-invasive quantification of brain radioligand binding by combining electronic health records and dynamic PET imaging data," *IEEE J. Biomed. Health Informat.*, 2015, to be published.

[134] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proc. Nat. Acad. Sci.*, vol. 111, pp. 8788–8790, 2014.

[135] M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," *Briefings Bioinformat.*, vol. 9, pp. 102–118, 2008.

[136] G.-Z. Yang, J. Andreu-Perez, X. Hu, and S. Thiemjarus, "Multi-sensor fusion," in *Body Sensor Networks*. New York, NY, USA: Springer, 2014, pp. 301–354.

[137] E. A. Mohammed, B. H. Far, and C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends," *BioData Mining*, vol. 7, no. 22, pp. 1–23, 2014.

[138] G. E. Hinton, "Learning multiple layers of representation," *Trends Cogn. Sci.*, vol. 11, pp. 428–434, 2007.

[139] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning—A new frontier in artificial intelligence research [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010.

[140] O. Bousquet and L. Bottou, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing System*. Cambridge, MA, USA: MIT Press, 2008, pp. 161–168.

[141] C. C. Aggarwal, *Data Streams: Models and Algorithms*, vol. 31. New York, NY, USA: Springer, 2007.

[142] J. Andreu and P. Angelov, "Real-time human activity recognition from wireless sensors using evolving fuzzy systems," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2010, pp. 1–8.

[143] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B, Methodological*, vol. 58, pp. 267–288, 1996.

[144] V. R. Carvalho and W. W. Cohen, "Single-pass online learning: Performance, voting schemes and online feature selection," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 548–553.

[145] R. Wainwright, F. Donck, M. Fertik, M. Rake, S. C. Savage, and J. H. Cloppinger, "Personal Data: The 'new oil' of the 21st century," presented at the World Economic Forum Europe Central Asia, Vienna, Austria, 2011.

[146] C. C. Y. Poon, Y. T. Zhang, and S. D. Bao, "A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health," *IEEE Commun. Mag.*, vol. 44, no. 4, pp. 73–81, Apr. 2006.

Authors' photographs and biographies not available at the time of publication.