Harnessing the Heart of Big Data

Sarah B. Scruggs, Karol Watson, Andrew I. Su, Henning Hermjakob, John R. Yates III,
Merry L. Lindsey, Peipei Ping

The exponential increase in Big Data generation combined with limited capitalization on the wealth of information embedded within Big Data has prompted us to revisit our scientific discovery paradigms. A successful transition into this digital era of medicine holds great promise for advancing fundamental knowledge in biology, innovating human health, and driving personalized medicine; however, this will require a drastic shift of research culture in how we conceptualize science and use data. An e-transformation will require global adoption and synergism among computational science, biomedical research, and clinical domains.

Overview of Big Data Science Research

A scarce number of scientific investigations have innovated clinical diagnosis, prognosis, and therapeutics, despite decades of research and enormities of National Institutes of Health (NIH)–funded research dollars.[1,2] This situation requires a global reassessment of whether linear thought processes and reductionistic approaches alone can describe biological processes in a way that translates to valuable information on human systems. Information gleaned from population science using large Big Data data sets has perpetuated a shift in the paradigm of how we define and investigate health and disease in the individual patient.[3] We are recognizing the profound value in unorthodox data types and in the integration of diverse data to describe individuals to sufficient depths for discerning clinical outcomes. Biomedicine, along with other fields, has been awakened and awed by the digital wave of major corporations such as Google and Amazon, who have revolutionized the Internet roadmap through developing and refining sophisticated data analytics platforms to accurately describe individual human behavior.[4] The reality in biomedical science is that there are zettabytes of high-quality data sitting idly on servers and in cloud infrastructures, and an abundance of biomedical knowledge lies hidden within, yet only a small fraction of this wealth has been harvested. There is an immediate need for data science to penetrate every area of biology, and the future of biomedicine rests on our collective ability to transform Big Data into intelligible scientific facts and knowledge.

The inception of the Big Data to Knowledge (BD2K) Initiative is a testament to the foresight of the NIH and our community (http://bd2k.nih.gov/). Revolutionary changes are occurring in every area of biology, including cardiovascular medicine, on how diverse data types are accessed, extracted, organized, integrated, and modeled, and how they affect basic science investigation and clinical care alike. It has become increasingly apparent that Big Data are everywhere and affect the global population in everyday life, through activities as ordinary as Internet shopping or as advanced as retail genome sequencing. Enthusiasm extends from the White House and major scientific organizations to laypersons and social media. Federal resources have been allocated to support national efforts in harnessing the enormous power embedded within Big Data and to advance biomedicine. NIH Centers of Excellence (COE) have been established to drive a transformation in the research culture, addressing data science challenges in an array of disciplines including cardiovascular medicine (http://bd2k.nih.gov/FY14/COE/COE.html). A significant effort is committed to shift the paradigm of scientific progress from the duplication and fragmentation of efforts across many competing groups to a synergistic accumulation and integration of unified community efforts in Big Data science. This reframing requires innovations aimed toward increasing the interactivity of and communication with Big Data data sets, as well as bridging the gap between layperson/patient and professional domains.

Data Science Promise for Supporting Cardiovascular Investigations

What is data science? Data science can be defined as the process of extracting, inferring, and validating knowledge from data sets that are acquired in a broad, minimally user-biased fashion. Data science builds tools and enhances access of datasets for investigators. Our vision of Big Data science is for it to support and to benefit the cardiovascular community at large. We do not see it as taking the place of fundamental research; on the contrary, we see it as synergizing with fundamental research. Many of the data science tools are being built to support individual investigators that conduct hypothesis-driven research. These include Omics data analysis tools, as well as text mining tools, and annotation pathway tools. Data science is data-driven, tool-driven and user-driven, rather than hypothesis-directed (Figure 1).

From the Departments of Physiology, Medicine, and Bioinformatics (S.B.S., P.P.) and Department of Medicine (K.W.), University of California, Los Angeles School of Medicine; Department of Molecular and Experimental Medicine (A.I.S.) and Departments of Chemical Physiology and Molecular and Cellular Neurobiology (J.R.Y.), The Scripps Research Institute, La Jolla, CA; Proteomics Services, European Molecular Biology Laboratories, European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom (H.H.); and Departments of Physiology and Medicine, University of Mississippi Medical Center, Jackson (M.L.L.).

Correspondence to Peipei Ping, PhD, Departments of Physiology, Medicine, and Bioinformatics, University of California, Los Angeles School of Medicine, 675 CE Young Dr S, MRL 1609, Los Angeles, CA 90095. E-mail peipeiping@earthlink.net

(*Circ Res*. 2015;116:1115-1119.
DOI: 10.1161/CIRCRESAHA.115.306013.)

***Circulation Research* is available at http://circres.ahajournals.org**
DOI: 10.1161/CIRCRESAHA.115.306013

Nonstandard Abbreviations and Acronyms

BD2K	Big Data to Knowledge
COE	Center of Excellence
NIH	National Institutes of Health

Data

Data are the currency of data science. The Big in Big Data describes not only the size or volume but also the potential of the data to (1) be useful and reused, (2) accumulate value over time, and (3) innovate a multidimensional, systems-level understanding. Importantly, these features are inversely proportional to user bias. Omics datasets, for example, are great examples of Big Data, in that global profiles of biomolecular features (eg, metabolites and proteins) are acquired using unbiased methods of detection (eg, mass spectrometry). There are of course physiochemical constraints of acquisition technologies that introduce instrument bias, but in general, they are unbiased in that they discern features of biomolecules based on a least common denominator—molecular mass. Although this type of data set may initially be collected for biological inquiries of narrow focus, Big Data datasets are amenable to repurposing and reuse to answer a myriad of other biological questions.

Data exist in innumerable, noncommensurate formats prohibiting interoperability. Some data exist as unstructured or unlinked data (eg, gene, disease, or drug data) that are not in a format readily amenable to computational analyses. For example, >1 million new articles are indexed in PubMed every year (1 every 30 s) and the knowledge is almost completely unstructured, making information access overly time-consuming, incomplete, and void of learning/memory. Big Data are thus in large part inaccessible, which can be because of this unstructured nature or other issues such as inadequate data descriptors (metadata) or data privacy ethics. A notable example is patient electronic health records,[5] which contain a wealth of largely unstructured clinical information. Accessing these data requires substantial changes in the clinical healthcare systems, and in how healthcare professionals are managing unstructured knowledge. Clinical data are not the only data that are inaccessible; most basic science investigators are hesitant to practice open data science for reasons such as the risk of data misuse by other parties and lack of data sharing incentives. Top-tiered journals, such as *Nature* have aimed to rectify the situation by creating journals like *Scientific Data*, a peer-reviewed, open-access publication for detailed data descriptors aimed at enhancing data set reuse (http://www.nature.com/sdata/about). However, widespread change requires a paradigm shift in research culture at all levels. To this end, the Biomedical and healthCAre Data Discovery and Indexing Engine Center led by Lucila Ohno-Machado at the University of California at San Diego has been awarded the NIH BD2K Data Discovery Index Coordination Consortium, which has been tasked with developing incentives, policies, and tools for data sharing and data discovery. Moreover, the NIH BD2K COE at Stanford University led by Mark A. Musen is developing innovative computational strategies to standardize metadata across all areas of biomedical science. For data science to be successful in the biomedical field, data and descriptive metadata must be carefully procured and transformed into an open and common currency; essential to this process is systematic security measures (eg, proper deidentifications) for protecting patient privacy.

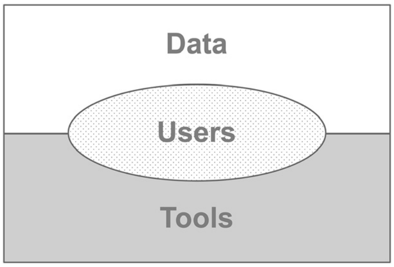

Figure 1. Central theme of data science—data, tools, and users. These are 3 essential components of data science architectures. Data refer to datasets that are reusable, accumulate value over time, and provide a multidimensional, systems-level understanding. Tools enable organization of and knowledge inference from data, in areas such as on-cloud data processing, multi-scale data integration, machine learning, crowdsourcing and text mining, data visualization, and mechanistic modeling. Users are anyone who has access to a digital device and an Internet connection. Individuals such as healthcare professionals, biomedical investigators, and layperson/patient populations are users.

In this regard, cardiovascular medicine has been highly fortunate to receive support and leadership from the NIH (eg, National Heart Lung and Blood Institute and National Institute of General Medical Sciences; both are global leaders in data science). The National Heart Lung and Blood Institute has supported many large cohort studies for decades (http://www.nhlbi.nih.gov/research/resources/obesity/population), including, for example, the Jackson Heart Study and Multi-Ethnic Study of Atherosclerosis. The National Institute of General Medical Sciences has supported the development of novel tools for use in data science (http://www.nigms.nih.gov/Research/Pages/ResearchResources.aspx), including the Human Genetic Cell Repository, Lipidomics Gateway, and Protein Data Bank. These high-quality data and tools have provided virtually inexhaustible resources for future data science-driven discoveries.

Tools

The technological platform of data science is driven by innovations in software tools and computational models; these new tools and models comprise a second integral component of data science. They represent the computational translators of data that enable communication with and knowledge translation from datasets. Many types of tools with diverse functionalities are required to adapt to user needs. We will briefly discuss here types of tools that have received high priority for overcoming the bottleneck of data to knowledge translation. These include innovations in (1) on-cloud data processing, (2) crowdsourcing and text mining, (3) multi-scale data integration, (4) data mining and machine learning, (5) mechanistic modeling, and (6) Big Data visualization.

Cloud computing infrastructure has been a springboard for the Big Data science revolution by enabling scientists to access and use shared pools of high-powered computational resources for data processing, which would otherwise exceed the capabilities of most desktop laboratory computers. This is a key innovation in that Big Data processing tools can be refined and maintained by experts in computational infrastructure and data science, and subsequently be made readily available to the global scientific community. The emphasis on crowd and community resources eliminates the requirement for each individual research group and organization to purchase, maintain, and update the latest hardware. *Crowdsourcing*, generically defined, is the process of engaging large communities of individuals to collectively accomplish a shared mission. Our BD2K COE at the University of California, Los Angeles (UCLA), leverages crowdsourcing of genomic knowledge to improve and expedite the gene annotation process. These efforts aim to systematically define relationships among key biomedical entities (eg, genes, proteins, diseases, and drugs) from the biomedical literature, through a combination of text mining, professional biocuration, and crowdsourcing. This strategy enlists both professional and patient/layperson crowds, the latter proving to be an enthusiastic and powerful resource. Although they may lack the formal training to fully appreciate the scientific context, it is increasingly clear that citizen scientists have both the motivation and ability to contribute to efforts to organize biomedical knowledge.[6] We envision a virtuous cycle that synergistically combines the efforts of scientific professionals, citizen scientists, and computational text mining. *Multi-scale data integration* tools are being developed to integrate and define relationships among distinct data entities (eg, molecular, drug, and disease information). The heterogeneous formats of biomedical data currently hinder knowledge aggregation, which prevents researchers from interpreting datasets using all relevant knowledge. *Data mining and machine learning* innovations are being applied to Big Data datasets to unveil biological patterns and emergent properties of data to make valuable and reliable inferences. Notably, investigators in the BD2K COE at the University of Wisconsin led by Mark W. Craven are using this strategy to take unstructured, heterogeneous clinical data and extract definitive, measurable and, importantly, predictable clinical phenotypes that are otherwise ill-defined. *Mechanistic modeling* innovations are being developed to enable scientists and clinicians to conduct more systematic investigations. These include strategies using Bayesian networks to connect molecular data with mechanistic information, such as correlating individual phenotypes, health histories, and multi-scale molecular profiles to examine disease mechanisms. Investigators in the BD2K COE at Stanford University led by Scott L. Delp are taking the heterogeneous pool of mobility Big Data and using novel strategies to innovate biomechanical modeling and behavioral and social modeling of physical activity data to transform diagnosis and treatment of limited mobility-associated disorders. Finally, significant efforts are being put forth to advance strategies in *Big Data visualization*. This includes creating visual analytics platforms for displaying multi-scale interaction network and pathway models of different data types (eg, genes, proteins, and metabolites) in a way that is customizable to different user inquiries and adaptable to the inherent complexities of the data.

One example of an innovative data science architecture showcasing certain types of tools described above is shown in Figure 2. This illustrates how data science can support cardiovascular investigations at-large by offering computational solutions for common inquiries, such as integrating diverse data (eg, genomics and proteomics) to predict disease phenotypes and support personalized medicine. Noteworthy is the modular structure of the workflow, making it integrable and adaptable to evolving user needs. Moreover the workflow is intuitive and generalizable; it is user-friendly, yet powerful enough for a broad range of biomedical applications. The vast utility and potential of data science tools are best exemplified in scientific investigations that have successfully harnessed Big Data and have gleaned valuable insights to advance science and medicine. A study by Denny et al[5] used a phenome-wide association study using electronic medical record–linked genetic data to examine associations between 3144 single nucleotide polymorphisms known from genome-wide association studies analysis to mediate human traits, and 1385 electronic medical record phenotypes in 13 835 patients. The phenome-wide association study analysis successfully replicated 66% of genome-wide association studies associations and discovered 63 novel associations; worthy of note, the strongest of these associations were validated using an independent cohort. This study highlights the tremendous potential of electronic medical record-linked genetic data to advance our understanding of disease phenotypes and human diversity. An additional study published this year by Shah et al[7] sought to improve the classification of heart failure with preserved ejection fraction, a heterogeneous clinical syndrome with no known treatment, to pave the way for more tailored therapeutic strategies. Dense phenotyping data from patients (n=397) clinically diagnosed with heart failure with preserved ejection fraction included 46 distinct measurements from clinical, laboratory, ECG, and echocardiographic analysis. Unbiased phenotype mapping, termed phenomapping, was performed using unsupervised machine learning algorithms to cluster patients into 3 groups that differed in clinical characteristics, cardiac structure/function, invasive hemodynamics, and outcomes. Importantly, results were validated in a prospective cohort. This study underscores the value of data science approaches for embracing the complexities of heterogeneous clinical phenotypes, thus innovating clinical decision-making and targeted treatment strategies.

Users

Users are the final integral component of data science, including virtually anyone with access to a digital device and Internet connection. Data science tools are most effective when they are user-centric, achieved by interactive development between data scientists and users. This process should ideally harness efforts by a diverse membership of biomedical professionals, or domain experts, and nonprofessionals alike, realizing that laypersons are both the source of data and ultimate consumers of insights gleaned from data science. The user base will be a self-propagating system; the premier quality of datasets, organization, software, and analytic tools contained within will attract users, and from that proximal community of users, new data contributors and users will emerge.

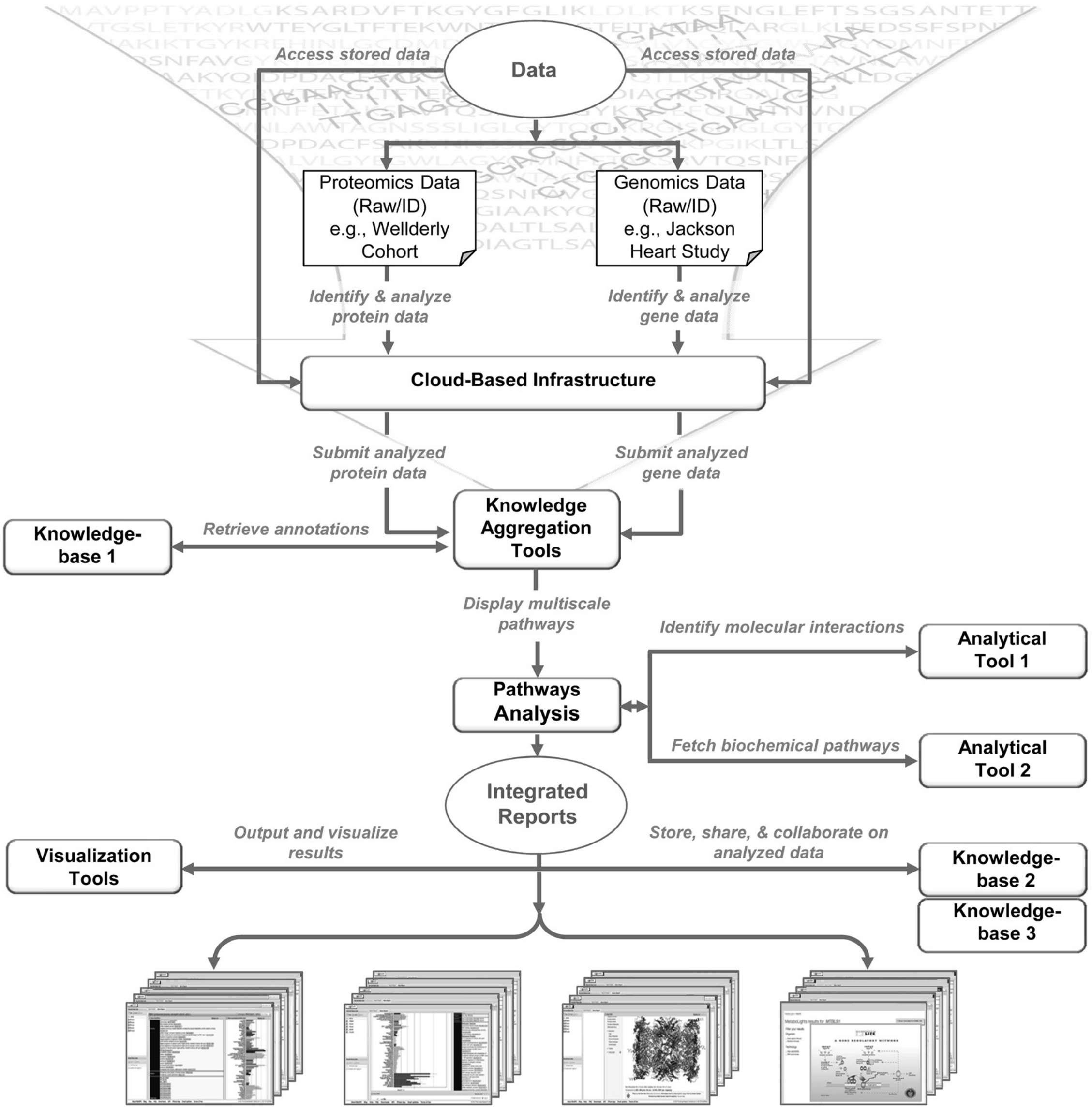
Data
Access stored data
Access stored data
Proteomics Data (Raw/ID) e.g., Wellderly Cohort
Genomics Data (Raw/ID) e.g., Jackson Heart Study
Identify & analyze protein data
Identify & analyze gene data
Cloud-Based Infrastructure
Submit analyzed protein data
Submit analyzed gene data
Knowledge-base 1
Retrieve annotations
Knowledge Aggregation Tools
Display multiscale pathways
Identify molecular interactions
Analytical Tool 1
Pathways Analysis
Fetch biochemical pathways
Analytical Tool 2
Integrated Reports
Output and visualize results
Store, share, & collaborate on analyzed data
Visualization Tools
Knowledge-base 2
Knowledge-base 3

Figure 2. Example of a modular data science architecture for supporting cardiovascular investigations. The workflow above provides an example to illustrate data science platforms correlating multi-scale molecular expression and phenotypic data from different experiments and the literature. The workflow begins with users uploading their own genomics or proteomics data, or data shared on and retrieved from a cloud-based infrastructure. Subsequently, with their submitted protein/gene data, they access the knowledge aggregation tools that enable location and access of both knowledgebase and analytic tools for processing and analysis. Data types are automatically annotated using community intelligence knowledgebase 1 (eg, Gene Wiki[9]). Multi-scale pathway information is integrated into a cohesive model via a pathways analysis tool, which retrieves molecular interaction and biochemical pathway information from analytical tools 1 and 2 (eg, PSICQUIC[10] and Reactome,[11] respectively). Results are output to visualization tools (eg, BioJS[12]) for tailored, multi-faceted visualization. Processed data can be stored and reaccessed via knowledgebases 2 and 3 (eg, COPaKB-Data[13] or Sage Synapse [http://www.sagebase.org/]).

However, because Big Data concepts are currently only in the common vocabulary of a select few communities, the key to rapidly overcoming this barrier of unfamiliarity is to implement a multi-faceted Big Data assimilation and education plan. This plan must target 3 general user domains in unique ways. The first is the biomedical researcher/users, including for example, physicians, and basic science investigators. The goal here will be to empower their ability to manage and interpret Big Data using data science software tools, and to capitalize on their highly specialized domain expertise to give meaning to the data. This can be accomplished through virtual classrooms, where tool dissemination and development occur interactively. The second population is Big Data science researchers, specifically targeting the new generation of scientists to grow the population of developers with transdisciplinary expertise in both computational biology and biomedical informatics. The final population is the general public/laypersons, which include diverse age groups/backgrounds, patient populations, government employees, and clinical personnel, in an effort to heighten public awareness and enthusiasm for the opportunities couched in Big Data. Social media, gaming tools, and crowdsourcing tactics will be highly effective here in showcasing and teaching bioinformatics concepts to laypersons.

Challenges and Opportunities

Despite the overwhelming promise of data science to innovate science and medicine, a few notable challenges require our attention. Perhaps the most formidable barrier for transitioning into this new era involves rigid ways of thinking within the research culture. There are ample opportunities to advance biomedicine by expanding our views and our laboratories to broader, systems-level ideas and approaches, and by positioning ourselves within scientific teams of complementary expertise. Academic departments in the biological realm will benefit from a balanced representation of data scientists, clinicians, and biologists. We will learn to be comfortable with data-driven, in parallel to hypothesis-driven, strategies from which unpredicted biological phenomena emerge.[8] It cannot be overstated how critical fundamental domain scientists and the knowledge gleaned from targeted science are to the Big Data science research paradigm. The supremely sophisticated information achieved from decades of hypothesis-driven research has provided a wealth of structural and functional information for the scientific community. Data science-born knowledge is not a competitor, but rather a synergistic elevator and integrator of targeted knowledge in that it provides multidimensional tools and dissemination channels for fully capitalizing on these focused efforts.

A second Big Data challenge comes in understanding the absolute requirement for validation of computational models with copious amounts of independent data. The emergent, open-ended nature of data science-driven research is a strength in that it lessens user bias and incorporates complexities of the data that are often excluded. However, it is paramount to understand that a derived model—although appropriate for the experimental datasets—may not be universally generalizable. Overstating results can lead to false positives and false confidence. This underscores the principal importance of open science, so that findings may be replicated and interrogated to ensure high fidelity.

This notion leads into a third major Big Data challenge, data ownership. A small percentage of scientific investigators in biomedicine currently share data openly; the majority of investigators remain relatively reluctant to making their data available for reuse and repurposing. The success of the Big Data era requires a global adoption of open science and the community working together as dutiful citizens of science about the manner in which data are collected, stored, accessed, and reused. NIH has established the aforementioned Biomedical and healthCAre Data Discovery and Indexing Engine Center to spearhead efforts toward creating a beneficial and safe environment for open science and data sharing. This will involve formulating policies for NIH-funded research that ensure optimal data curation, privacy, and quality. It is important to recognize that responsible open science and data sharing will breed science of superior integrity and higher value, which is in itself a most noble objective.

We are at an exciting and critical juncture in medicine and scientific investigation; a time when funding mechanisms are available for accessing the vast complexity of human health and redefining personalized medicine. BD2K is not a trendy, fleeting movement; rather, it is an essential advancement in and progression of science and medicine that has been birthed by the complexity of the questions we are asking. This effort is entirely dependent on the community working together, as polarized science will likely result in a failed BD2K effort. A unified community effort for translating Big Data to knowledge will achieve virtually endless returns on investments initially put forth for the acquisition of Big Data, producing a sum that is much greater than its parts.

Acknowledgments

We thank Dr Ding Wang at UCLA for help in preparing the figures and Dr Edward Lau at UCLA for his critical input on the content.

Sources of Funding

This work was supported in part by the National Institutes of Health U54 GM114833 (to Dr Ping, Dr Watson, Dr Lindsey, Dr Su, H. Hermjakob, Dr Yates) and R37 HL063901 (to Dr Ping); and by the T.C. Laubisch endowment at University of California, Los Angeles (to Dr Ping).

Disclosures

None.

References

1. Collins FS. Reengineering translational science: the time is right. *Sci Transl Med*. 2011;3:90cm17. doi: 10.1126/scitranslmed.3002747.
2. Kell DB. Finding novel pharmaceuticals in the systems biology era using multiple effective drug targets, phenotypic screening and knowledge of transporters: where drug discovery went wrong and how to fix it. *FEBS J*. 2013;280:5957–5980. doi: 10.1111/febs.12268.
3. Hayes DF, Markus HS, Leslie RD, Topol EJ. Personalized medicine: risk prediction, targeted therapies and mobile health technology. *BMC Med*. 2014;12:37. doi: 10.1186/1741-7015-12-37.
4. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff (Millwood)*. 2014;33:1163–1170. doi: 10.1377/hlthaff.2014.0053.
5. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31:1102–1110. doi: 10.1038/nbt.2749.
6. Good BM, Nanis M, Wu C, Su AI. Microtask crowdsourcing for disease mention annotation in pubmed abstracts. *Pac Symp Biocomput*. 2015;20:282–293.
7. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, Bonow RO, Huang CC, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*. 2015;131:269–279. doi: 10.1161/CIRCULATIONAHA.114.010637.
8. Friend SH, Schadt EE. Translational genomics. Clues from the resilient. *Science*. 2014;344:970–972. doi: 10.1126/science.1255648.
9. Good BM, Clarke EL, de Alfaro L, Su AI. The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic Acids Res*. 2012;40:D1255–D1261. doi: 10.1093/nar/gkr925.
10. Aranda B, Blankenburg H, Kerrien S, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*. 2011;8:528–529. doi: 10.1038/nmeth.1637.
11. Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011;39:D691–D697. doi: 10.1093/nar/gkq1018.
12. Gómez J, García LJ, Salazar GA, et al. BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics*. 2013;29:1103–1104. doi: 10.1093/bioinformatics/btt100.
13. Zong NC, Li H, Li H, et al. Integration of cardiac proteome biology and medicine by a specialized knowledgebase. *Circ Res*. 2013;113:1043–1053. doi: 10.1161/CIRCRESAHA.113.301151.

Key Words: crowdsourcing ■ database ■ heart diseases ■ information storage and retrieval ■ metabolomics ■ proteomics ■ user-computer interface