Where is the Science in Big Data Visual Analytics? From Pretty Pictures to Transformative Biomedical Discoveries

Suresh K. Bhavnani, PhD[1], Shyam Visweswaran, MD PhD[2], Rohit D. Divekar, MBBS PhD[3], Gowtham Bellala, PhD[4]

[1]Inst. for Translational Sciences, Univ. of Texas Medical Branch, Galveston, TX; [2]Dept. of Biomedical Informatics, Univ. of Pittsburgh, Pittsburgh, PA; [3]Div. of Allergic Diseases, Mayo Clinic, Rochester MN; [4]Analytics Lab, Hewlett Packard Laboratories, Palo Alto, CA

Abstract

Numerous visual analytical representations (e.g., networks visualizations) have been developed to analyze big biomedical data obtained from sources such as electronic medical records and whole genome assays. The goal of using these methods is to make unplanned discoveries leading to transformative changes in healthcare. However, such discovery-based approaches are often referred to as "fishing expeditions" that capture mere correlations, and therefore inferior to gold standard methods such as hypothesis testing and machine learning. Given the vast opportunities offered by big data visual analytics, there is an urgent need for the informatics community to address this analytical debate and define a path forward. Towards that goal, this panel will assemble analysts from informatics, clinical practice, and industry to debate the role of visual analytics in biomedical informatics research. Two proponents of visual analytics will stress why visual analytics is indispensable in the age of big biomedical data, and two critics will argue that visual analytics does not adequately address the limitations of correlation, multiple testing, and human bias. This debate format is designed to engage the audience in a spirited discussion about the future role of big data visual analytics in biomedical informatics.

Introduction

The Open Science movement[1] (e.g., data from NIH-funded studies being made publicly available), coupled with rapid advances in the development of inexpensive high throughput technologies (e.g., multiplex assays), has resulted in vast digital resources accessible by both scientists and the lay public. However, the sheer magnitude of such resources far exceeds our cognitive abilities to exploit them for the prevention, diagnosis, and treatment of diseases.

One approach that appears promising to help comprehend such vast quantities of information is the use of *visual analytics*[2]. This emerging field, defined as the "science of analytical reasoning facilitated by visual interactive interfaces", has been shown to aid researchers in the rapid comprehension of complex and big biomedical data[3,4,5]. For example, networks have been used to analyze Medicare claims from more than 30 million patients, which enabled researchers to infer patterns in the progression of different diseases[5]. However, such discovery-based approaches are often referred to as "fishing expeditions" that produce correlations that are vulnerable to false positives and human bias. At best, visual analytical methods are relegated to the category of hypothesis generation, whose results require considerable verification and validation.

Given the transformative opportunities offered by big data visual analytics, there is an urgent need to address this analytical debate head on. Towards that goal, this panel brings together analysts from informatics, clinical practice, and industry to debate and explore the role of big data visual analytics in biomedical informatics. Two proponents will argue that visual analytics is indispensable for the analysis and comprehension of big biomedical data, and will present concrete examples of how such approaches have been successfully used. In contrast, two critics will argue that big data visual analytics has not adequately addressed the limitations related to correlation, multiple testing, and human bias and therefore destined to remain an approach for hypothesis generation. The goal of this debate format is to polarize the views about this critical topic so that the audience can comprehend the positions clearly, and engage in the issues by voicing their opinions.

Position Statements of Panel Members

1. Why Visual Analytics is Indispensable for the Analysis of Big Biomedical Data (Suresh K. Bhavnani, PhD, Moderator)

Dr. Bhavnani, is associate professor of biomedical informatics at the Institute of Translational Science in the University of Texas Medical Branch, and holds an adjunct appointment at the School of Biomedical Informatics at the University of Texas in Houston. As PI of the Discovery and Innovation through Visual Analytics lab, he specializes in (1) the discovery of complex patterns in big biomedical data such as heterogeneity in diseases and their respective pathways, and (2) the innovation of novel visual analytical methods to analyze and comprehend large and complex datasets. His research in the use of networks has received several awards including a

distinguished paper award[6] at the AMIA fall symposium, and a distinguished paper award[7] at the AMIA Summit of Translational Bioinformatics.

Dr. Bhavnani will briefly introduce the cognitive and computational foundations of visual analytics, and discuss three goals that have proved useful[8] in the analysis of medical data: (1) **discovery of complex relationships** through network visualizations and transformations; (2) **verification and validation of patterns** through the use of graph-based and biostatical measures; and (3) **inference of mechanisms** underlying the emergent patterns. Each of these goals will be illustrated through examples of completed network analysis projects involving biomedical data. He will conclude with why visual analytics is indispensable in the comprehension of complex phenomena including heterogeneities in complex diseases such as Alzheimer's disease and asthma.

2. How Visual Analytics Accelerates the Discovery of Disease Mechanisms in Big Biomedical Data (Shyam Visweswaran, MD PhD)

Dr. Visweswaran is assistant professor of biomedical informatics at the Department of Biomedical Informatics and the Intelligent Systems Program, University of Pittsburgh. He specializes in the application of artificial intelligence and machine learning to problems in clinical medicine and translational bioinformatics. As PI of the Vis lab, his research includes (1) computer-aided diagnosis and prediction[9], (2) discovery and prediction from high-dimensional genomic data[10], and (3) patient-specific predictive modeling for personalized medicine[11].

Dr. Visweswaran will define biomarker discovery as a search for informative models of biomarkers and subjects. He will argue that since the number of such models in big data tends to be very large, exhaustive search for informative models is impractical. Therefore visual analytics can play a key role by enabling domain experts to use domain knowledge to direct the search for models in big biomedical data that are both statistically and biologically significant. Using examples from an existing dataset of Alzheimer's SNP-subject dataset, he will demonstrate how domain experts have made key biological discoveries using visual analytics that were overlooked using conventional approaches. He will conclude with proposals for addressing the technical challenges of integrating visual analytics with machine learning for high-dimensional data.

3. Why Inferences from Visual Analytical Representations Can Lead to Bad Science (Rohit Divekar, MBBS PhD)

Dr. Divekar is assistant professor and practicing clinician in the Division of Allergic Diseases at Mayo Clinic. His combined training as a biologist and immunologist enables him to analyze immune pathways in complex molecular data, and translate those results into evidence-based treatments for patients. He has helped in making inferences of disease mechanisms from visual analytical representations in complex diseases such as asthma[8].

Dr. Divekar will discuss the perils of making inferences from visual analytical representations such as networks. He will argue that because biological mechanisms activated in many diseases are complex and redundant, almost any pattern in data can be interpreted as having biological meaning. He will present examples from a network analysis of Alzheimer's disease where there could be many different interpretations for the same pattern in the data. Given that many of these inferences can be inaccurate, he will propose several strategies to reduce incorrect inferences including (1) the use of inter-rater reliability where several independent experts are asked to interpret relevant patterns, and the degree of their agreement quantitatively measured and reported, and (2) enhanced tools that automatically gather and present biological findings for specific patterns that appear in visual analytical representations.

4. Why Visual Analytics Needs to Transcend the Limitations of Correlation and Multiple Testing (Gowtham Bellala, PhD)

Dr. Bellala is senior research scientist in the Analytics lab at HP Labs. He specializes in conducting inter-disciplinary research on the applications of machine learning and data mining techniques to a wide range of domains including biomedical[7], software-defined networking, and cyber-physical systems[12].

Dr. Bellala will argue that discovery-based methods like visual analytics have yet to address the limitations of correlations which appear compelling to domain experts, but could be the result of mere coincidences in the data. Furthermore, he will argue that the mere act of visually looking for patterns in a visual analytical representation is equivalent to the well-known problem of multiple testing, and therefore highly suspect as a credible scientific method. He will conclude by stating that while inferences made through visual analytical methods might be appropriate in domains such as social networking where the cost of an incorrect hypothesis is easily absorbed, it is

not appropriate or ready for biomedical research where an incorrect hypothesis can lead to a costly drain in limited funding and intellectual resources.

Discussion and Engagement of Panel Attendees

Before the panel position statements are presented, two microphones marked "Agree" and "Disagree" will be set up in the seating aisles. After each of the above position statements, the attendees will be first asked (with a show of hands) to determine how many agree or disagree with the position statement just presented. Those who wish to speak can voice their opinions using the "Agree" or "Disagree" microphones. The panel members will engage with these opinions, and the moderator will determine when the next position statement will begin. After all the positions statements are delivered and responded to, the attendees will be asked through a final show of hands to determine if the proponents or the critics of big data visual analytics were more compelling in their arguments. The moderator will then attempt to integrate the ideas in order to determine the next steps in defining the role of big data visual analytics in biomedical research.

We expect this debate to enable the attendees to understand the issues related to making transformative discoveries using big data visual analytics, while avoiding the pitfalls of arriving at incorrect inferences. We also hope that new ideas will emerge about how visual analytics can transcend its current predominant role of hypothesis generation. The panel discussion should be of benefit to: (1) **biomedical informaticians** and **domain experts** who have no prior background in visual analytics but need to acquire an intuition about current and future methods and their strengths and limitations; (2) **network researchers** who wish to understand the hurdles of using visual analytics to analyze big biomedical data, (3) **program managers** from funding agencies who wish to understand how best to support research informed by visual analytics that have the potential of wide impact in making medical discoveries.

Acknowledgements

This research is funded by NIH grants 1U54RR02614 UTMB CTSA(ARB).

References

1. Molloy JC. The Open Knowledge Foundation: Open Data Means Better Science. 2011 PLoS Biology 9.
2. Thomas JJ, Cook KA (2005) Illuminating the path: the R&D agenda for visual analytics. National Visualization and Analytics Center.
3. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci U S A. 2007 May 22;104(21):8685-90.
4. Christakis, NA, Fowler, JH. The Spread of Obesity in a Large Social Network Over 32 Years. New England Journal of Medicine 2007, 357 (4): 370–379.
5. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA. A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol, 2009, 5(4).
6. Bhavnani, S.K., Abraham, A., Demeniuk, C., Gebrekristos, M., Gong, A., Nainwal, S., Vallabha, G.K., and Richardson, R.J. Network Analysis of Toxic Chemicals and Symptoms: Implications for Designing First-Responder Systems. AMIA Annu Symp Proc. 2007:51-55.
7. Bhavnani S.K., Drake, J.A., Bellala, G., Dang, B., Peng, B., Oteo, J.A., Santibañez-Saenz, P., Visweswaran, S., Olano, J.P. How Cytokines Co-occur across Rickettsioses Patients: From Bipartite Visual Analytics to Mechanistic Inferences of a Cytokine Storm. Proceedings of AMIA Summit on Translational Bioinformatics (2013).
8. Bhavnani S.K., Drake, J.A., Divekar, R. The role of visual analytics in asthma phenotyping and biomarker discovery. In Heterogeneity in Asthma. (ed. A. Brasier), Springer, (2014) 289-305.
9. Visweswaran, S, Mezger, J, Clermont, G, Hauskrecht, M, Cooper, GF. Identifying deviations from usual medical care using a statistical approach. AMIA Annu Symp Proc., 2010.
10. Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. Journal of the American Medical Informatics Association 2011:18(4):370-5
11. Visweswaran, S, Angus, DC, Hsieh, M, Weissfeld, L, Yealy, D, Cooper, GF. Learning patient-specific predictive models from clinical data. Journal of Biomedical Informatics 2010:43(5):669-85.
12. G. Bellala, M. Marwah, A. Shah, M. Arlitt, and C. E. Bash, "A Finite State Machine-based Characterization of Building Entities for Monitoring and Control," ACM BuildSys, 2012.

The first author affirms that all panel members have agreed to participate, and have contributed to the preparation of this document.