The Convergence of Computational and Social Approaches for Unveiling Meaningful and Valuable Data

Angela P. Murillo
Indianapolis University-Purdue University Indianapolis
Indianapolis, U.S.A.
apmurill@iu.edu

Renata G. Curty
Universidade Estadual de Londrina
Londrina, Brazil
renatacurty@uel.br

Wei Jeng
National Taiwan University
Taipei, Taiwan
wjeng@ntu.edu.tw

Daqing He
University of Pittsburgh
Pittsburgh, U.S.A.
dah44@pitt.edu

ABSTRACT

The current data paradigm is seeking a more integrated and comprehensive framework to make sense of data and its derived issues. From the perspective of the data life cycle, we argue that computational and social approaches complement each other to confront data challenges. Computational approaches consist of ETL (extract, transform, and load), modeling, and machine learning techniques; social approaches include policy and regulations, data sharing and reuse behavior, reproducibility, ethical and privacy issues. In this panel, we frame these two approaches as data acumen and data stewardship. The merging of these two perspectives allows data not only to become discoverable, accessible, and interoperable, but also to further the value of revealing meaningful patterns and become supportive evidence for important decision making. In this panel, the opening facilitator and three panelists will report on their recent studies in terms of this convergence of both data acumen and stewardship while sharing their recent research insights on case studies in three disciplines: agriculture, biomedicine, and archeology.

KEYWORDS

Data acumen, data stewardship, agricultural data, biomedical data, archaeological data.

ASIS&T THESAURUS

Data Science, Cross Disciplinary Fertilization, Reproducibility

INTRODUCTION

As the world becomes increasingly more data-intensive, there is a need to examine how data-related work is impacting academia, the workforce, and society at large. The growing demand for data experts has been documented. In a Linked Workforce Report (2018), more than 151,000 data science jobs went unfilled across the United States. Additionally, there is an estimated shortage of 200,000 workers with deep analytics skills and 1.5 million employees who lack the knowledge to make data-driven decisions (Manyika et al., 2011). Despite the fact that Curty and Serafim (2017) observed a highly computational and business orientation in 93 data science/data analytics degrees offered by U.S institutions, and that these emphases were also found by Verma, Yurov, Lane, and Yurova (2019) in the majority of the job posts for data analysts, some research has shown that data knowledge is needed across all disciplines and domains (Börner et al., 2018).

Liberal arts and humanities students who pair their major with a data-related component qualify for additional jobs and increased salaries (Burning Glass Technologies, 2013). Additionally, studies have shown the need to incorporate "soft" social skills into the data fields, indicating the need for both social and computational approaches to data work (Börner et al., 2018). As a result, data-related job titles which were initially limited to data scientists or data analysts have been branching out. Lyon, Mattern, Acker, and Langmead (2015); Lyon and Mattern (2016) mapped iSchool curriculum to real-world job descriptions to examine the variety of data-related employer requirements and the current state of data science-related educational curriculum. In the series, six data science roles were compared in two articles: Data Archivist, Data Librarian, and Data Steward/Curator (Lyon et al., 2015), as well as Data Analyst, Data Engineer, and Data Journalist (Lyon & Mattern, 2016).

Grounded in the computational and social data approaches, for this panel, we consider that data acumen and stewardship are interconnected concepts that should be instilled in data practices and research. While the former relates to the capability to draw more informed and better decisions based on data, the latter is the capability which envisions the management of data among and between different data lifecycle stages (Baker, 2009), which allows decisions from data acumen to be readily available and accessible, respecting high-quality requirements.

This convergence of data acumen and data stewardship is impacting jobs, education, and how practitioners and researchers are responding to these computational and social knowledge domains. Both practitioners and academics are seeking a more integrated and comprehensive framework to deal with data issues and challenges, by combining highly computational approaches (i.e., data gathering, extraction, modeling, etc.) for acumen analytics, with contributions from social perspectives, (i.e., data policy and regulations, data sharing and reuse behaviors, data stewardship, ethics and privacy, and so on). The merging of these two

This is the author's manuscript of the article published in final edited form as:

Murillo, A. P., Curty, R. G., Jeng, W., & He, D. (2019). The convergence of computational and social approaches for unveiling meaningful and valuable data. Proceedings of the Association for Information Science and Technology, 56(1), 554–557. https://doi.org/10.1002/pra2.89

perspectives allows full compliance with the FAIR principles (Wilkinson et al., 2016) so that data becomes not only findable, accessible, and interoperable, but also reusable. Such integration facilitates the discovery of meaningful patterns and rules from data, as well as the investigation of questions which could not have always been anticipated by primary researchers. The panelists will provide examples of how data acumen and data stewardship are bound together while addressing data issues from case studies in clinical sciences, agriculture, and archeology.

OBJECTIVES OF THE PANEL

The objectives of this panel include:

- To introduce the framing perspectives of data acumen and data stewardship for the computational and social behavioral approaches, respectively;
- To examine how these two perspectives complement each other to solve data issues in different knowledge domains;
- To explore, as a result of the group interaction, some alter-natives to strengthen the ties between data acumen and data stewardship.

STRUCTURE OF THE PANEL

The structure of the panel is as follows:

- The opening facilitator (He) will provide a framing intro-duction to the session and introduce the dynamics of the panel.
- Each one of the panelists (Murillo, Jeng, and Curty) will present their recent research findings on how both computational approaches and social and behavior-wise foci interplay in different knowledge domains.
- The facilitator and the panelists will host an interactive session which invites attendees to share their opinions and thoughts through the use of gamification and data visualization strategies.

PANELISTS AND TOPICS

Facilitator

Daqing He is a professor at the School of Computing and Information as well as has a joint appointment at the Intelligent Systems Program, an interdisciplinary academic unit focusing on research and education on Artificial Intelligence, both of which are at the University of Pittsburgh. He earned his PhD degree in artificial intelligence from the University of Edinburgh, Scotland. He has been the Principal Investigator (PI) and Co-PI for more than 10 research projects, funded by the National Science Foundation (NSF), United States Defense Advanced Research Projects Agency (DARPA), University of Pittsburgh, and other agencies. His current research projects include open corpus personalized learning, models for sentence simplification, and information access.

In this panel, He will serve as an opening facilitator and will provide an introduction to the panel theme and the basic structures of the panel. Equipped with a computer science background and years of teaching and service experience in the iSchool community, He is able to provide unique insights and identify promising research opportunities that can be achieved only through close integration of data acumen and stewardship.

Topic: Computational and Social Research in Agriculture

Angela P. Murillo is an Assistant Professor and the Program Director of the Applied Data and Information Science Program at the School of Informatics and Computing at Indiana University-Purdue University Indianapolis (IUPUI). Her dissertation investigated earth and environmental data sharing and reuse. More recently, her research has focused on the impacts of data science on LIS education (Murillo & Jones, 2018), and the impact of data science on earth and environmental science data management. Prior to joining IUPUI, she worked as a Research and Development Information Scientist at Novozymes, where she assisted with research and analytics in relation to Novozymes R&D, product development, and competitors.

For this panel, Murillo will share her recent research and perspectives of an ongoing research project related to agricultural data and new tools that are being used by farmers to gather data in the fields related to weather conditions, soil conditions, seed treatments, and other field data. This research is a collaboration between Murillo and data science faculty to produce predictive analysis to assist with the under-standing of the return of investment of various treatments, field, and weather conditions in regards to agriculture and an understanding of farmers' use of new tools to gather agricultural data. In this study, the computational approaches will include ETL and modeling of field data provided by farmers and industry partners, and social approaches will include an analysis of farmers' data gathering processes, use of techno-logical tools, and reproducibility considerations. Murillo will report on early research insights of this ongoing research project of data acumen and stewardship in agricultural.

Topic: Preserving the Research Workflow in Biology and Medicine

Wei Jeng is an Assistant Professor at the Department of LIS, as well as the Center for Research in Econometric Theory and Applications in the National Taiwan University, Taiwan. She completed her PhD studies at the School of Computing and Information, University of Pittsburgh. Her recent research interests involved research data sharing and reproducibility. While focusing on the social aspects of data research, Jeng frequently collaborates with computer science and bio- medicine researchers practicing data acumen, and thus has first-hand experience in resolving tensions and identifying challenges when bridging the two worlds.

In this panel, Jeng will share her recent results and early research insights on an ongoing research project which explores better solutions for preserving the research workflows for PIs. Jeng will discuss the context of reproducibility crisis recently raised in these fields, and review the most recent information services toward ensuring reproducibility and a potentially better solution--a decentralized, secure, and auditable infrastructure. Jeng will also report some early results about data gathering from a participatory design workshop which is expected to be conducted in Summer 2019. Inviting a group of university faculty and researchers in biology and medical science, the design workshop aims to reveal PIs' strategies or techniques for data acumen (either standardized or "homegrown"), their data management practices for ensuring the reproducibility of their research around the data lifecycle, as well as potential challenges arose from the entanglement between data acumen and stewardship.

Topic: The Role of Paradata in Digital Archaeology

Renata G. Curty is an Assistant Professor in the Information Science Department at the Universidade Estadual de Londrina, in Southern Brazil. Her research relates to sociological aspects of science, digital scholarly communication and metrics, data curation, and research data sharing and reuse.

In this panel, Curty will present her current findings of an ongoing study about paradata and its role in promoting more informed, responsible, and trustworthy data (re)use. She will first define paradata in relation to data documentation and its integration with contextual and descriptive metadata despite their conceptual and practical boundaries. Later, she will discuss different types of paradata sources in the sciences, considering the massive quantity of data which are gathered as a by-product of computer-mediated collection processes, such as web surveys. Taking a social approach, she will cover the need for paradata to be documented and easily accessible along with the structured dataset and its metadata in a consistent manner, and describe a few strategies for doing so. From a computational perspective, she will talk about how paradata can assist data interpretation and resignification in the context of digital archeology to increase scientificity, transparency, and reproducibility for digitally re-created environments and artifacts.

PANEL STRUCTURE

The structure of the proposed panel is as follows:

Introduction (5 minutes)

Dr. He will introduce the panel theme and objectives and talk briefly about how computational and social approaches are important for allowing excellence in data acumen and stewardship.

Presentations (30 minutes-10 minutes each)

The panelists (Murillo, Jeng, and Curty) will present their individual research projects and findings.

Q&A session (5-10 minutes)

A brief Q&A and discussion session regarding the content of panel presentations will be hosted.

Interactive Session (45 minutes)

The interactive session will be moderated by the panelists who will use a combination of gamification, visualization strategies, and team collaboration tools to better engage the audience. A set of situational-based data scenarios related to the fields panelists will cover on their presentations will be provided to attendees, and they will be asked to respond them both from a computational and social perspective.

We plan to use the Role Playing Game (RPG) as a methodological technique to elicit and stimulate conversations, and promote retention of acquired knowledge and group discussion outcomes. Predefined archetypes will be provided to each participant in to guide discussions. The participants will be invited to "change hats" during the group discussion to address how they would approach data issues differently according to computational or social perspectives. Participant responses to questions elaborated by the panelists will be registered through live poll tools, e.g., Sli.do (www.sli.do), and the results will be displayed to the audience to allow participants to reflect about the responses provided. The responses will be discussed collectively, and the most salient remarks will be shared with the audience through a real-time online whiteboard. We will then, debrief and wrap up the session with all attendees.

EXPECTED CONTRIBUTION

This panel aims to raise attention to data issues and cultivate constructive conversations among scholars and practitioners from information science and data science communities, by integrating both computational and social perspectives to address data challenges in different knowledge domains. Through the interactive session, the panel expects to assist researchers and practitioners from both perspectives to consider the importance of understanding computational and social considerations of data, as well as potentially find areas of opportunity to expand their research frameworks, strengthening the ties of researchers and practitioners from both perspectives.

We anticipate that this awareness will positively reflect on data-related initiatives and research endeavors that will not consider data acumen and data stewardship as antagonistic, but rather an opportunity to unify them to look at the data phenomenon. Through the sharing of experiences in the inter-active session, this panel expects to learn of others techniques in confronting data challenges both from a social and computational perspective, to assist the panelists and participants in learning about practices, techniques, and tools when working with data. Ultimately, this panel will assist in examining how data acumen and data stewardship are complementary for unveiling meaningful and valuable data, as well as in allowing the collective development of more robust interconnected practices for both concepts.

ACKNOWLEDGMENTS

The panel of "Preserving the Research Workflow in Biology and Medicine" by Jeng was supported by Grant MOST108-2636-H-002-002 and MOST107-3017-F-002 -004, as well as the Grant (#107L900204) by the Ministry of Science and Technology and Ministry of Education (MOE) in Taiwan.

REFERENCES

Baker, K. S. (2009). Data stewardship: Environmental data curation and a web of repositories. *Digital Discourse: The E-volution of Scholarly Communication, 1*(1).

Börner, K., Scrivner, O., Gallant, M., Ma, S., Liu, X., Chewning, K., … Evans, J. A. (2018). Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences*, 115(50), 12630–12637. https://doi.org/10.1073/pnas.1804247115

Burning Glass Technologies. (2013). *The art of employment: How liberal arts graduates can improve their labor markets prospects*. Retrieved from http://www.burning-glass.com/wp-content/uploads/BGTReportLiberalArts.pdf

Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective, Computer Science and Telecommunications Board, Board on Mathematical Sciences and Analytics, Committee on Applied and Theoretical Statistics, Division on Engineering and Physical Sciences, Board on Science Education, … National Academies of Sciences, Engineering, and Medicine. (2018). *Data Science for Undergraduates: Opportunities and Options*. https://doi.org/10.17226/25104

Curty, R. G., & Silva, J. S. da (2017). A formação em ciência de dados: uma análise preliminar do panorama estadunidense. *Informação & Informação, 21*(2), 307-331.

LinkedIn. (2018). LinkedIn Workforce Report | United States | August 2018. Retrieved from https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

Lyon, L., & Mattern, E. (2016). Education for Real-World Data Science Roles (Part 2): A Translational Approach to Curriculum Development. I*nternational Journal of Digital Curation, 11*(2), 13–26. https://doi.org/10.2218/ijdc.v11i2.417

Lyon, L., Mattern, E., Acker, A., & Langmead, A. (2015). *Applying Translational Principles to Data Science Curriculum Development.* IPRES 2015 Proceedings.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers Hung, A. (2011). *Big data: The next frontier for innovation, competition, and productivity | McKinsey.* Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Murillo, A. P., & Jones, K. M. L. (2018). The Development of an Undergraduate Data Curriculum: A Model for Maximizing Curricular Partnerships and Opportunities. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *Transforming Digital Worlds* (Vol. 10766, pp. 282–291). https://doi.org/10.1177/0165551517748149

Ortiz-Repiso, V., Greenberg, J., & Calzada-Prado, J. (2018). A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools. *Journal of Information Science, 44*(6), 768–784.

Varvel Jr., V. E., Bammerlin, E. J., & Palmer, C. L. (2012). Education for data professionals: A study of current courses and programs. *ACM International Conference Proceeding Series*, 527–529. https://doi.org/10.1145/2132176.2132275

Verma, A., Yurov, K. M., Lane, P. L., & Yurova, Y. V. (2019). An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. *Journal of Education for Business, 94*(4), 243-250. https://doi.org/10.1080/08832323.2018.152068