

Introduction to the Minitrack on Big Data and Analytics: Pathways to Maturity

Stephen H. Kaisler,
D.Sc SHK & Associates
Laurel, MD 20723
Skaisler1@comcast.net

Frank Armour, Ph.D. Kogod
School of Business
American University
Washington, DC
farmour@american.edu

Alberto Espinosa, Ph.D. Kogod
School of business American
University Washington, DC
alberto@american.edu

This minitrack is focused on exploring theory, techniques, applications, and understanding of the maturing field of Big Data and Analytics. We have selected papers that demonstrate innovative approaches to analytics and introduce new concepts.

Session 1

The first paper of Session 1 by Michael Werner is entitled *Materiality Maps – Process Mining Data Visualization for Financial Audits*. As the financial industry has evolved with new tools and techniques, new challenges await financial auditors. Since many transactions involve large and complex data sets, new tools to facilitate the auditing process are necessary. Process Mining is a novel technique that uses transactional event data to create process models that represent the ebb and flow of business transactions in an organization. The basic idea is to create process and heat maps that allow auditors to understand the business processes of an organization and to detect interactions between different processes. The heat maps show where the most activity within a set of processes is occurring. By detecting hotspots according to the heat maps, analysts can drill down to examine the causes for the increased activity. Since process mining leads to perhaps hundreds of different variants, process and heat maps can help auditors focus on the most important elements of their tasks.

Paper 2 in Session 1 by Jeremy Glassman, Robert St.Louis, and Benjamin Shao is entitled *Don't Get the Cart before the Horse: There are no Shortcuts to Prescriptive Analytics*. This paper suggests that talent is not as much of a problem as the granularity, accuracy and integration of data in corporate business analytic efforts. Chief Purchasing Officers (CPOs) were “intimidated” by the concept of prescriptive analytics, which constrained them to focus largely on descriptive analytics. Thus, they are unable to see how to automate some of the predictive and prescriptive

processes which could lead to automating some of the decision-making processes. This paper addresses some of the roadblocks that CPOs identified in getting to advanced analysis of the data they have (e.g., beyond descriptive and diagnostic analytics). In particular, it validated the idea that 80% of the effort is in wrangling the data – including collecting, organizing, and integrating – which leaves little time and resources to do the analysis. The authors noted that five out of 15 firms had advanced to predictive analytics, but only one beyond the level of prescriptive analytics. The authors concluded that getting the data and preparing it for analysis is more important than talent and other resources. To sum up, they noted that “there are no shortcuts to prescriptive analytics”. This requires closer coordination between management expectations and needs and the activities of the analytics team.

Paper 3 of Session 1 by J. Albert Espinosa, Stephen Kaisler, Frank Armour, and William H. Money is entitled *Big Data Redux: New Issues and Challenges Moving Forward*. The authors published the predecessor to this paper in 2013 at HICSS-47. Since then, it has received over 600 citations across a broad swath of technical papers on Big Data (although 530 were noted when the paper was written). The authors determined that significant evolution in Big Data activities had occurred in the intervening five years and decided to revisit the issues identified in the original paper to see if they were still valid and to identify any new issues that had arisen. Of the 530 papers citing the original paper at the time of their analysis, the authors selected 110 of those for review based on selected criteria such as papers with 50 or more Google citations. After reading all these papers, they reduced them to 54 papers after those that merely mentioned the author’s paper without contributing anything new. They mapped those papers to the original categories and identified new categories. They

found that some of the old topics continued to dominate the technical literature, but noted the emergence of many new tools and techniques the evolution of predictive analytics, and the emergence of prescriptive analytics.

Session 2

The first paper in Session 2 by Johannes Weibl and Thomas Hess is entitled *Finding the Next Unicorn: When Big Data meets Venture Capital*. As the world economy improved in recent years, the amount of venture capital available has greatly increased. This means more opportunities but substantially more data to sift through to evaluate an opportunity. The authors reviewed 13 venture capital firms that synthesize and use novel data sources in making decisions about investment opportunities. A data driven approach to analysis of these data source leads to transactional and information benefits that leads to lower operating costs. Such data gives VC firms deeper insight into a businesses' operations that allow VC firms to provide better mentoring and investments and, thereby, increase their ROI. Their contributions give greater insight into how VC firms can use big data to improve their selection of opportunities and continued oversight as the business evolves.

The second paper of Session 2 by Lucas Baier, Niklas Kuhl, and Gerhard Satzger is entitled *How to cope with Change? Preserving Validity of Predictive Services Over Time*. Predictive analytics are becoming readily available to assist firms in monitoring and analyzing data streams for better service. But, sudden changes in the content, format, or other characteristics can affect the validity and functionality of the data stream. The authors have developed a framework to describe and assess predictive for their validity. The authors also propose a research agenda for developing tools and techniques to improve the long-term validity of predictive service. They note the use of supervised machine learning, such as used in service analytics to improve or extend service offerings. Their framework allows the differentiation between predictive services based on their characteristics. Their research agenda identifies two topics: the availability of a labeling operation to support supervised learning and the integration of domain experts to provide knowledge for assessment. They also note that validity is only one aspect of predictive services and that others need to be identified, evaluated and tested for use in assessing predicting services.

The final paper in Session 2 by Jeffrey Saltz, Robert Heckman, Kevin Crowston, Sangseok You, and Yatish Hegde. It is entitled *Helping Data Science Students Develop Task Modularity*. Building the skills of data science students can improve their ability to work effectively within a data science team. The authors report on a mixed method study to evaluate different approaches to task decomposition in order to facilitate task modularity. The authors assessed different aspects of the Kanban methodology with different aspects of the methodology. The authors found that small tasks improved understanding of the overall project and that task modularity improved individual task tracking. Conversely, they found that it was hard to divide complex tasks into chunks of appropriate size. The authors observed that task decomposition leading to task modularity was a major challenge for the students. They concluded that data science students do not have the same advantage as traditional computer science students and that this needs to be factored into data science programs.