6-2017

Mental Models: What do you mean by Data Science?

Mari W. Buche
mwbuche@mtu.edu

Mental Models: What do you mean by Data Science?

Mari W. Buche
Michigan Technological University
mwbuche@mtu.edu

ABSTRACT

This research-in-progress investigates the diverse comprehension and definition of the Data Science. This field is rapidly growing in popularity around the world. There are numerous degree programs being rolled out at both the graduate and undergraduate levels. However, there is notable inconsistency in the approach to this highly regarded area of study. The individual components and the focus of data science have produced misconceptions in academia as well as in industry. As the build out approaches broad saturation, this seems like an ideal time to define what is considered Data Science. Mental models will be created from focus group interactions. Each session will represent participants from homogenous backgrounds (e.g. computer science or management information systems). We will compare the models, analyzing the variations and commonalities from the data. Results of this study will increase general understanding of data science. Implications for universities and industry practitioners will be discussed.

Keywords

Data science, mental models, degree program development.

INTRODUCTION

Data science is a phrase that is being employed in many diverse disciplines, creating a buzz in the industry and producing numerous lucrative positions for graduates majoring in the field. These are excellent outcomes for students and researchers focusing their energies on this current “hot topic”. However, when put to the test, few people share a common view of this phenomenon, particularly when the comparisons are made across established disciplinary boundaries. There is even disagreement on where a degree program in Data Science should be located within an academic institution. Specifically, should it be in a business school or computer science? Or, is another unit more appropriate?

Although the field is still in its nascent phase, it will be critical for Data Science to develop a firm identity going forward. Definitions and explanations will need to coalesce around theoretical foundations and practical implications. The roots of this field are borrowed from a number of existing disciplines. Although it might be provocative to espouse an imprecise identity, this ambiguity causes discomfort and suspicion if prolonged. For example, recruiters from businesses search for applicants that meet a subset of skills selected to fulfill a predetermined corporate function. Defined knowledge and abilities are listed as required and desirable. A broad palette of unique skills is difficult to align with prescribed job openings. Recruiters just aren’t sure what to do with those applicants.

The purpose of this study is to explicate key attributes of data science from focus group participants. The data will be used to develop mental maps, identifying concepts (nodes) and associations that link the concepts that produce a shared understanding of data science within a particular discipline. The discipline-specific composite maps will be compared across various disciplines, thereby observing commonalities and differences. Inconsistencies between mental maps often indicate areas prone to misunderstandings. In corporate settings, collaboration and communication can be improved by developing greater uniformity across disciplines.

Variety, Variance, Volume, and Veracity are the common V’s associated with general explanations of the Data Science. These terms present an underlying foundation for discussion, but there is a belief that everyone holds a common mental representation of these components. However, if we investigate the conceptual dimensions more closely, we will quickly find that perceptions are not crystal clear. For instance, the simplest concept of the four might be Volume. The general understanding is that, by definition, big data connotes a large amount of data. But, what do we mean by “large amount”? Accountants have collected, stored, analyzed, and reported out on large data sets for decades. The simple observation might be that analysis breaks down under such large volume when we attempt to use traditional, standard analytical tools such as Microsoft Excel. However, that definition is not quite accurate, either.

Fundamental Skills

The basic skills considered fundamental to success as a data scientist might vary, but, in general, it includes a combination of both technical and applied skills. The transferable (soft) skills include interpersonal communication, complex problem solving, persistence, inquisitive nature, and systems (or big picture) thinking. Technical skills include some combination of data mining, machine learning, data visualization, data modeling techniques, time series and statistical analysis, and cloud computing. In short, data scientists are tasked with solving problems using computing technologies, interpreting results, and communicating their findings in the form of "actionable" information. The combination of skills causes many people to refer to data scientists as unicorns – nearly impossible to find because the perfect combination of skills doesn't exist. This perception fuels increased salaries afforded to data scientists, which in turn leads to a growing attraction and interest in degree programs.

Outcomes and Actionable Information

In some academic disciplines, outcomes are valued for their contribution to theoretical foundations. In other disciplines, application of results to measurable performance is viewed as most important. The value proposition is generally context specific. Therefore, the foundations of data science are intricately woven into the fabric of the context and the purpose driving the initiative.

Academic Placement

Educational programs are increasing rapidly. The actual components that constitute a data science degree vary greatly. Likewise, there is disagreement on the appropriate physical location or domain for such programs. Degree programs that focus on technical programming skills (e.g. data mining) tend to favor computer science for as the parent discipline. Programs that emphasize business problem solving (e.g. business analytics) opt for a business school as the natural choice. Others focus on the mathematical underpinnings of the course content, and the appropriate parental unit would be sciences and arts (e.g. statistics). Several programs recognize the interdisciplinary strength of data science, taken advantage of synergies between requisite skills. Those multi-disciplinary programs often encounter friction due to the inherent design of not having an official home. This negativity appears inevitable, at least within institutions fragmented by traditional academic disciplines.

Industry Placement

Data scientists are in high demand (e.g. Press, 2015) across industries. Companies are aware that they are sitting on mountains for data. The data have potential value, leading either to decreased costs of doing business or increasing revenues (BLS 2017), or both. Few examples exist that successfully analyze and interpret big data with resultant increases in productivity and corporate profit. Data scientists are skilled technicians, but require domain expertise in order to meet employer expectations. The question often becomes, what do we do with our data scientist? Or, how do we get the most value out of our data scientists? Corporate executives must determine the most effective role and location for these experts. In some businesses, data scientists are co-located with business units. One benefit is that the experts rapidly develop domain expertise by being embedded with their internal clients. Data scientists would likely be motivated by solving relevant problems plaguing the unit, or by attaining unexpected benefits through the pursuit of previously unexplored opportunities.

In contrast, data scientists could be located in a separate unit where their technical expertise would develop quickly through spontaneous interchanges of ideas. This resembles the "think tank" model, where the data science experts tackle obscure problems while building on each other's strengths. Intrinsic motivation accrues via intriguing challenges and professional competition among the participants.

DATA COLLECTION

Focus groups will be convened to gather perceptions and definitions of Data Science/Big Data from knowledgeable experts in a variety of industries. Participants will be grouped by common educational background (e.g., computer science, mathematical sciences/statistics, business/management information systems, and computer engineering). The size of each group will be capped at no more than five persons per session in order to encourage greater participation and meaningful contributions. The researchers will follow a standardized question guide and they will actively probe for details based on individual responses. The unit of interest is discipline representatives, so individual responses will be aggregated when developing composite mental maps.

METHODOLOGY

Mental models have been studied in many settings and numerous research disciplines. The degree of overlap using network analysis techniques provides an indication of the shared understanding between groups or individuals with regard to a specific topic. Mapping the cognitive conceptual models of individuals from various academic disciplines and comparing the resultant composite maps will highlight aspects of commonality and divergence. Revealed causal mapping technique will be used to analyze the transcripts, coding the raw statements for verbal identification of causal linkages between relevant concepts (Armstrong, 2005). Identification of specific primary nodes and associations will illustrate the breadth of the domain of interest. In other words, by comparing maps across disciplines, we will highlight the inconsistencies and boundary-spanning view of data science.

ANTICIPATED RESULTS AND LIMITATIONS

Diagrams representing the discipline-specific mental maps will be presented for visual comparison. The researchers will analyze the various maps using rigorous network analysis techniques (Axelrod, 1976). Findings of variation will specify the degree of divergence and convergence. Commonalities of nodes and associations will be considered shared understanding.

The methodology used to analyze and produce mental models is a content analysis. Even though the research team will be trained in the RCM methodology, there may be differences in the coding of certain statements. Discussion and communication is used to arrive at consensus (Armstrong, 2005). We will also be relying on a convenience sample, which raises the concern of generalizability. The results of this exploratory study will provide deeper understanding of the topic. This, in turn, may lead to further research that builds sound theoretical foundations (Lee and Baskerville, 2003).

DISCUSSION AND CONTRIBUTIONS

We are currently in the growth stages of data science/big data. This build out consists of increasing number of academic degree programs at both the undergraduate and graduate levels. Although universities use identical terminology, the actual implementation at the programmatic level can and does exhibit profound diversity. On an arbitrary continuum, academic programs can be described as highly managerial (e.g. business) or highly technical (e.g. computer programming) in nature. This study will not impose any artificial or subjective qualifying assessments on overall value to society. The purpose of this study is to provide an initial comparison, in order to increase awareness of the immense variation in this important multi-disciplinary field.

The interview guide is currently in the development stage. Existing questionnaires from the literature review will be modified for this study, employing previously validated instruments whenever possible. Data collection will begin immediately following pilot testing of the semi-structured guide. Questions will be revised based on feedback from a panel of academic researchers familiar with focus groups and interviewing techniques.

The comments from focus groups will be coded by the research team. Discrepancies in the coding will be resolved based on common practices from extent research (e.g., Armstrong, 2005). Software will be used to analyze the findings from the coded data, and maps will be generated (e.g., Stout, Cannon-Bowers and Salas, 1996). The discipline-specific maps will be compared through visual inspection and mathematical analysis (e.g., Johnson, and O'Connor, 2008; Klimoski and Mohammed, 1994; Lim and Klein, 2006; Mohammed and Dumville, 2001).

The findings from this study will increase awareness of the variation in perceptions regarding data science. In addition to theoretical contributions, these results will have implications for industry practitioners, academic programs under development, recruitment of potential students, and informed communication of admissions personnel. Assuming that all agencies share a common perception of data science can, and often does, lead to misunderstandings. This study will promote a shared view of the topic, in spite of diverse disciplinary-myopic conceptual lenses.

REFERENCES

1. Armstrong, D. J. (2005) Causal Mapping: A Discussion and Demonstration, in *Causal Mapping For Research in Information Technology*, V.K. Narayanan, and D.J. Armstrong (eds.), Idea Group Publishing, Hershey, PA, 20-45.
2. Axelrod, R. (1976) *The Structure of Decisions*, Princeton University Press, Princeton, NJ.
3. Bureau of Labor Statistics (2017) www.bls.gov (Last accessed 27 February 2017).
4. Johnson, T.E. and O'Connor, D.L. (2008) Measuring Team Shared Understanding Using the Analysis-Constructed Shared Mental Model Methodology, *Performance Improvement Quarterly*, 21:3, 113-134.
5. Klimoski, R. and Mohammed, S. (1994) Team mental model; construct or metaphor, *Journal of Management*, 20:2, 403-437.
6. Lee, A.S. and Baskerville, R.L. (2003) Generalizing Generalizability in Information Systems Research, *Information Systems Research*, 14:3, 221-243.

7. Lim, B.C. and Klein, K.J. (2006) Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy, *Journal of Organizational Behavior*, 27, 403–18.

8. Mohammed, S. and Dumville, B.C. (2001) Team mental models in a team knowledge framework: expanding theory and measurement across disciplinary boundaries, *Journal of Organizational Behavior*, 22:2, 89-106.

9. Press, G. (2015) The Supply and Demand of Data Scientists: What the Surveys Say, *Forbes*, URL https://www.forbes.com/sites/gilpress/2015/04/30/the-supply-and-demand-of-data-scientists-what-the-surveys-say/#58945ccd6b90 (Last accessed 27 February 2017).

10. Stout, R.J., Cannon-Bowers, J.A. and Salas, E. (1996) The role of shared mental models in developing team situational awareness: implications for training, *Training Research Journal*, 2, 86-116.