The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness

Jeffrey S. Saltz
Syracuse University
Syracuse, NY, USA
jsaltz@syr.edu

Abstract— **As data continues to be produced in massive amounts, with increasing volume, velocity and variety, big data projects are growing in frequency and importance. However, the growth in the use of big data has outstripped the knowledge of how to support teams that need to do big data projects. In fact, while much has been written in terms of the use of algorithms that can help generate insightful analysis, much less has been written about methodologies, tools and frameworks that could enable teams to more effectively and efficiently "do" big data projects. Hence, this paper discusses the key research questions relating methodologies, tools and frameworks to improve big data team effectiveness as well as the potential goals for a big data process methodology. Finally, the paper also discusses related domains, such as software development, operations research and business intelligence, since these fields might provide insight into how to define a big data process methodology.**

Keywords-Big Data, Data Science, Process Methodology

I. INTRODUCTION

With the increasing ability to collect, store and analyze an ever-growing diversity of data generated with ever-increasing frequency, the field of data science is growing rapidly. Likewise, there has been a rapid increase in the number of organizations doing analysis on large amounts of data they are able to collect. This analysis is increasingly becoming a team activity, as opposed to being done by a single data scientist. For example, when the data is fairly static and not too large, one person might be able to do the analysis. In this case, the methodology / process used is not as critical as the actual analytics performed. However, as the amount of data expands, and one person becomes a team of people, then there is a risk that the team's working effectiveness is not optimal. At a high level, the question to be addressed is how to ensure a data team is working efficiently and effectively. There are many foundational questions that might be addressed to understand and potentially improve a data team's performance. This includes questions such as:

- *Do we need a big data team process methodology?*
- *What might be some goals for such a methodology?*
- *How could one approach defining a methodology?*
- *How could one assess a data team's performance?*
- *What tools might be helpful to improve big data team effectiveness?*

This rest of this paper discusses these foundational questions with a goal of laying a framework for future research with respect to defining methodologies and tools to improve team processes within data science teams.

II. RELATED WORK

Much has been written about the use of data science and algorithms that can generate useful results. In fact, many in the field believe that big data research needs to continue to focus on analytics [1]. However, less has been written about how to effectively "do" Big Data projects, that is, how processes can be institutionalized within organizations to improve the efficiency and effectiveness of Big Data projects. To gain a better understanding on the current focus within the Big Data community, we reviewed the proceedings from the 2014 IEEE Big Data conference. The goal of this inspection was to determine which papers focused on methodologies and tools for improved team effectiveness in doing big data projects. In all, 296 articles and posters were examined. Of those, none focused on what process methodology the team did (or should) use. In fact, only 8% were found to discuss any aspect of the socio-technical challenge in doing a big data project, most often that focus was with respect to information security or how an algorithm should interact with the user to better understand the data and refine the results.

More broadly, current descriptions on how to do a data science project generally adopt a task-focused approach, conveying the techniques required to analyze data. For example, Jagadish described a process that includes acquisition, information extraction and cleaning, data integration, modeling, analysis, interpretation and deployment [2]. O'Neil approached this challenge similarly [3]. Guo approached the problem from a slightly different perspective and provided a Data Science Workflow framework [4]. Guo's workflow defined several high-level phases such as preparation, analysis, reflection, and dissemination, with each phase having a specific series of steps that can be repeated within that phase in an iterative analysis.

These views on how to do data science have not materially evolved in the past 20 years. For example, they are similar to the KDD (Knowledge Discovery in Databases) process described nearly twenty years ago [5]. In another example, *CRISP-DM* (*Cross Industry Standard Process for Data Mining)* might also be viewed as a possible first step towards defining a data science methodology. CRISP-DM was established in the 1990s, and is a data mining process model for data mining experts [6]. The model mentions six high-level phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. However, no model seems to have achieved wide acceptance, and in fact, it has been reported that there is a decrease within the KDD community of people using CRISP-DM, and an increase in people using their own methodology [7].

Since there is no currently accepted process methodology for doing big data projects, one might naturally want to explore the need for such a process – perhaps there is no common process methodology because one is not needed!

III. Methodology

This section first addresses the need for a process methodology, and the related question of defining a methodology for an often open-ended set of questions that are being explored. Then, the importance of assessing team performance, as well as some possible initial areas of exploration, is discussed.

A. Is there be a need to define a process methodology?

First, while the use of a step-by-step description of data science provides some understanding of the tasks involved, it does not provide much guidance about the tasks a data science team should be carrying out. Hence, not surprisingly, Bhardwaj [8] noted that teams doing data analysis and data science work in an *ad hoc* fashion, using trial and error to identify the right tools, that is, at a low level of process maturity [9]. In fact, currently data analysis is much like the early days of software development, when organizations had little ability to predict whether a project would be successful, on time or on budget, and projects were overly reliant on the heroic efforts of particular individuals (i.e., a low level of process maturity). Just as then, organizations that start a data science project have little assurance of its success and that success is overly dependent on the competencies and efforts of the individuals involved. Demonstrating this lack of success, Kelly and Kaskade [10] surveyed 300 companies, and reported that "55% of Big Data projects don't get completed, and many others fall short of their objectives". While there are many reasons a project might not get completed, with a robust team-based process methodology, one would expect many of those reasons to be identified prior to the start of the project, or to be mitigated via some aspect of the project execution and/or coordination methodology. It is interesting to note that many larger software projects used to have similar challenges.

Having a well-defined repeatable methodology for carrying out Big Data analyses will help projects address a range of challenges, such as understanding what roles will be played and, hence, who needs to be included in the broader extended team, determining the desired goals for the analysis, selecting an appropriate data architecture or technical infrastructure, and determining what techniques might be appropriate for analyzing and validating the results. Without a well-defined methodology, necessary tasks might get addressed, but the team might forget a step or not follow best practices. Similarly, without well-defined team roles, tasks might be missed or not coordinated across team members.

B. How can one define a process for something so exploratory in nature?

Some people might feel that one cannot define a process for doing a Big Data project, since the analysis of the data is often open ended and iterative. However, the fact that the process is open ended does not contradict the value of a process. In fact, one can view the agile software development methodology as supporting an open-ended process to develop/refine/improve software (such as improving a user interface). While that is very different from the far more open ended question of trying to find "value in the data", it nonetheless suggests that there is value in using a methodology within a more open-ended process to achieve goals such as group coordination / communication and the prioritization of "what to try next". Within a data context, benefits of having a robust methodology range from helping to ensure that there is an appropriate data architecture to ensuring that there is a prioritized set of goals for doing the data analysis. In addition, within a business context, there are often consultants hired to do data analysis, and clearly, one would want a well-defined process methodology when one organization is hiring another person (or organization) to do the data analysis.

C. Exploring Related Domains

There are many domains where process methodologies have been defined and are actively used. While data projects have their unique characteristics, one approach to defining a methodology is to explore other domains to see what aspects of a methodology might be leveraged for big data projects. One such example is the software development life cycle (SDLC) process used for software development. However, other process methodologies might also be applicable – such as those used by operations research teams (such as the optimization of a business process) or a quantitative research methodology (such as exploring a smaller dataset using scientific methods and statistical analysis). Finally some of the KDD (Knowledge Discovery in Databases) and data mining process descriptions might also be applicable, such as the previously noted *CRISP-DM*.

However, the Big Data project challenges involved might or might not be similar to software, BI or OR project challenges. Hence, a process methodology, or even a process methodology task such as requirements analysis, might not be appropriate or that methodology might need to be adapted to support data projects. Some of the most relevant domains with the potential to help shape the definition of a Big Data process methodology are discussed below.

1) Software development methodology: At a high level, there are similarities between software and data projects. Both have similar objectives such as trying to understand the project requirements and doing analysis and design, which can be mapped to a software development process. However, data projects and software projects also have many differences. First, with respect to data, there is a much greater focus on what data might be needed, as compared to software projects, where there is typically less ambiguity of the data needed for a project. Furthermore, since there is often a large amount of data, storing, retrieving, cleaning and validating the cleansed data is non-trivial. While part of this can be thought of as the data architecture for a software project, there is often the challenge of identifying appropriate data sources, evaluating those sources (e.g., quality, timeliness), capturing those data sources and trying to determine a methodology to ensure the data has no issues, such as mismatching of timing or data issues after cleaning. This stage has no parallel in a software project but is critical for the success of a data project. Quality assurance is another example of how software and data projects could be different. For example, the acceptable level of data quality is often dependent on the use of the data, as described by Kaisler [11] who noted that trend analysis may not require the precision that traditional DB systems provide. Finally, even if one could leverage a software development methodology, one would still need to determine which methodology was most appropriate, since there are multiple software development process alternatives, such as agile and waterfall methodologies. Some teams only use agile, some only use waterfall, and some use a combination of both methodologies.

2) Quantitative research for "small" datasets: Many of the challenges encountered by Big Data teams are not the typical challenges encountered by more traditional quantitative research efforts. These differences include the IT requirements needed to do the analysis, the larger number of people that need to be coordinated, and challenges in validating / describing the results. Nevertheless, there are some key concepts from quantitative research that could prove useful in a data science methodology. These include examining the distribution of data to identify possible bias, testing for validity, and carefully documenting the steps taken and results generated so that work can be replicated, which is especially important in scholarly big data teams where results have to stand up to expert peer review and examination by others in the field. But the risk and potential cost of drawing erroneous conclusions from big data is equally serious in other fields such as finance, health care, education, or public safety.

3) Other domains such as operations research and business intelligence: Operations Research (OR) is a discipline that is focused on the application and use of advanced analytical methods, with a goal of making better decisions. Hence, just as software development has some commonality with data science, operations research also has many similar characteristics. One example description of the OR process includes steps such as recognizing & formulating the problem, constructing a model, finding solution, establishing a procedure and implementing the solution [12]. So, for example, the focus on establishing a procedure (since an optimization is often ongoing rather than a unique one-time task) is similar in spirit to many big data projects, such as generating a customer score that is updated on a regular basis. Another field, Business intelligence (BI), focuses on using a set of techniques and tools to transform data into useful information for business insight. BI has been well researched [13], and in fact, many leverage concepts from both software development and CRISP-DM [14]. However, it has been found that there has been a need to adapt traditional requirements engineering process for BI due to the fact that the requirements analysis for BI systems differ substantially from requirements analysis for conventional information systems [14]. In a related finding, the need to develop Business Intelligence systems that are able to react to unforeseen or volatile requirements [15] has also been suggested. Taken together, the processes used within OR and BI projects are both possible points of leverage for Big Data projects, but expecting one of these methodologies to work unchanged might not be realistic.

D. How to Assess Data team Performance

Since data teams are often tasked with "exploring the data", the work is often viewed as open ended. As such, standard metrics to evaluate team performance have often been seen as not appropriate for evaluating data team performance. However, to enable the identification of best practices with respect to how teams should operate, there needs to be a way to evaluate team performance. In other words, if one wants to compare different methodologies and tools to improve Big Data team performance, then it becomes critical to be able to evaluate and compare the effectiveness of the different teams. There are several areas of exploration to address this challenge, some of which are outlined below.

1) General Models of Team Effectivness: Researchers in social and organizational psychology have studied teams and their performance for decades and have developed a plethora of models describing and explaining team behavior and performance. One of the most widely used normative models was proposed by Hackman [16]. Hackman's model seems appropriate due to its intended purpose of identifying factors related to team effectiveness, broadly defined, and its inclusion of team process factors. In brief, this model focuses on the inputs factors (such as organizational context and group design), process and moderating factors and outputs (including task output, the team's continued capability to work together and the satisfaction of individual team members).

2) Information Systems Success Models: While Hackman includes some forward-looking criteria (e.g., continued capability to work together), there is no doubt that a very important factor in understanding team performance is measuring if the team executed a data project successfully, which is the focus of how people have modeled Information Systems success. A commonly cited model for IS Success is from DeLone and McLean [17], which is based on the system and information quality that drives use and user satisfaction, which drives individual impact and leads to organizational impact. DeLone and McLean's model has three components: the creation of a system, the use of the system, and the consequences of this system use. If applied to big data teams, items such as data quality would still very important, but would be just one aspect of analysis usage and consequences of such usage (i.e., poor data quality will lead to less usage or negative consequences when using analytics which were based on bad data). However, impact, by itself might not be the best measure, since it's possible that another data science team might have identified knowledge within the dataset with would have resulted in significantly more impact to the organization.

3) Establishing Critical Success Factors: Another way to try and understand how effective a data science team is working is by exploring how they are doing with respect to the critical success factors (CSFs) for doing a data project. Towards this end, Gao [18] recently led an effort that explored what these CSFs might be on data teams. Specifically, after reviewing 60 Big Data Case Studies, as well as 14 published surveys, Gao identified the CSFs for Big Data projects. Based on Gao's work, below are some possible processes/people CSFs that might be helpful to evaluate team readiness and performance:

- Identifiable business value of the project
- Clear project scope
- High data quality and appropriate data security
- Clear project goals with appropriate deadlines
- Measurable outcomes
- Iterative Process Model
- Multidisciplinary Teams (ex. IT, Business)

4) Exploring A Maturity Model for Data Teams: Finally, by defining and then using a set of data science maturity levels, one could then easily compare process maturity across organizations. The common maturity model (CMM) framework provides a standard definition of process maturity, which includes five levels of process maturity [9] and has been used extensively within the software industry to understand and improve the level of maturity of an organizations' software development processes. While CMM has been focused on software development, a step towards using this framework for data science was suggested by Crowston & Quin [19], where the CMM framework was used to create a common maturity model for scientific data management. Note that a data science process maturity model would be different from maturity models that have been proposed for Big Data 'organizational readiness'. For example IBM's maturity model [20] provides broad-based organizational readiness dimensions, but does not focus on the process maturity of the actual data science effort.

IV. GOALS

Beyond the basic goal of ensuring a high quality and timely data science project, a set of high level objectives that one would hope to achieve via the use of a data science methodology might include that the team knows *what* is the desired output of the analysis, *how* this output could be achieved within the constraints of available resources and time, *who* are the people that will be involved in the effort and *why* they are doing the analysis. If one person is doing everything within the data science process, this knowledge might be implicit. However, when larger teams start to do data science projects, this knowledge needs to be explicitly known by all team members.

In addition to these high level objectives, and the foundational requirement of being able to do advanced analysis and visualizations, the following challenges also need to be addressed within the methodology.

A. Coordination with Others

A cornerstone of any methodology should be to ensure that there is a process to effectively communication across all the interested parties, both within the core "data science team" as well as others interested in the results of the work. This includes helping others in the organization have insight into the team's progress and status as well as in terms of understanding the context of how the insights could be valuable. This coordination and communication might be via sprints, phases or some other set of well defined processes.

B. Ensuring Quality

The current step-by-step process often has an implicit assumption that the results will have no errors. Certainly, current data scientists do see quality as important. However, in reality, there has been minimal focus on ensuring quality results via a defined process, especially compared to software development methodologies. Quality within a data focused project includes ensuring no coding bugs (ex. an error in parsing data), no misuse of statistics, nor misleading communication of the results. While this has always been important, as the field grows, there needs to be a visible effort to ensure valid results. A specific focus with respect to quality should be data quality, which is an important subset of the overall quality of the data science project. Issues of data quality can include aligning the timing of data from different sources, the frequency of data updates (to ensure the data is up to date), the meaning of data attributes, and the cleaning of data (e.g., what assumptions are acceptable, how is missing data treated). Just as with overall quality, while data quality has always been important, as the field grows, there should be a well defined process to ensure valid data is being used to drive the results. While this is also true for "small data" efforts, big data efforts can swamp the usual methods used in smaller data projects, such as identifying outliers and determing if those outliers are the result of an error in data entry or repreent actual readings.

C. *Data Ownership, Security and Privacy*

For many datasets, due to data privacy, not all team members should be able to access all of the data. Hence, ensuring that the data is appropriately protected needs to be explicitly evaluated and if necessary, appropriately monitored. This can range from ensuring the correct permissions have been assigned to the appropriate team members to ensuring personally identifying information has been correctly masked.

D. *Priorization of Requirements*

Often times, there are many possible directions one might focus, with respect to analyzing the data. As the data science team grows, it is important to ensure that the focus is on answering useful questions and providing valuable insight to the organization. In other words, there needs to be a process to make sure the team is focused on trying to answer the most relevant / interesting questions.

E. *Requirements Analysis*

Once the requirements are understood, the process should ensure that there is an analysis of how to achieve the desired goals. This analysis should cover a range of questions, such as expected time required, costs involved and exploring potential challenges in doing the data investigation (ranging from data quality to determining if a new modeling algorithm needs to be developed). Based on the results of this project analysis, the team might need to refine the goals or expected timeline for the project. This can also be used as a way to ensure external support teams (such as the "IT department") understand their respective roles in the effort.

F. *Deployment*

Within many organizations, both research and commercially focused, one must think about being able to share the results in a "productized" manner, such as a real-time update of a computed score. This includes ensuring the system is reliable, stable and scalable as well as ensuring that the software can handle appropriate error conditions. While some might view this at an IT problem that data scientists do not need to worry about, often times it is helpful to think through possible deployment challenges prior to the analysis being complete. In addition, data scientists might also need to consider the timeliness of the results and how long an algorithm can to run, ranging from milliseconds to days, prior to the results being deployed.

V. POTENTIAL NEXT STEPS

One area of future work is to perform and document case studies (focusing on the process methodology used by data teams) across a variety of settings (large team / small team, for profit / research, open ended exploration / enhancement to current analytic). From these case studies, we can collectively determine what are the biggest challenges within big data projects. In case studies where improved methodologies where introduced, we should be able to better understand the sociotechnical challenges and benefits of an improved methodology. One might also be able to refine the critical success factors as well as understanding the value having a standardized maturity model (e.g., to help an organization understand the value of increasing its process maturity).

In addition to case studies, another way to gain insight into current practices is via more broad-based surveys, which could add additional context to the overall discussion. However, work needs to be done to determine the appropriate questions to be addressed. For example, one question that could be analyzed is how often a data science project is a one-off analysis, where a team is brought together for a limited time as compared to an ongoing flow of data that needs to be analyzed. If the latter, does that process need to be automated (e.g., reporting energy use daily for customers to track)? In other words, a broad-based survey could be used to define the key characteristics within a data science team, with different types of data science teams needing different process methodologies.

With respect to maturity models, work could be done to refine the broadly defined maturity model to specific methodologies. Another avenue to explore is process simulation, which could be used to explore different process methodologies. While a simulation is clearly a simplified model of the sociotechnical challenge, analyzing simulation results might be useful to better frame the problem and understand the strengths and weaknesses of possible solutions.

Beyond which process methodology the team should use, a related question is if there are tools that can support the big data team. Researchers have just begun to address the need for tools to help team-based Big Data projects. For example, in a recent project, Bhardwaj [8] introduced the notion of version control for datasets. This concept is similar in spirit to version control for software development, but rather than tracking software code, the tool tracks changes to a dataset. Another step forward in helping teams do data science is a project that provides a framework to help teams determine the most appropriate technology infrastructure for doing data science [21]. However, overall there has been little thought given as yet to broader team issues, e.g., the collaboration tools to support a Big Data team process or methodology. Open questions with respect to tools include what types of tools are most useful and team acceptance of such tools.

VI. CONCLUSION

A recent Gartner Consulting report advocates for more careful management of analysis processes, though a specific methodology is not identified [22]. With this in mind, this paper outlines several key questions with respect to a big data team's process methodology. It is interesting to note that the evolution envisioned for how to do data science projects is similar to the evolution that has occurred for software development. At first, programming was thought to be a solitary task, and the defined work process was focused on the key steps required to create a software solution. There was an implicit assumption that the process for working across groups was not an important challenge. For example, when the classic waterfall software development model was defined, the process was described as a series of tasks. However, as demonstrated by the growing use of the agile methodology, it has become clear that it is useful to establish a methodology that ensures good group communication and acknowledges that the process is iterative.

VII. References

[1] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, *36*(4), 1165-1188.

[2] Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J., Ramakrishnan, R., Shahabi, C. 2014. "Big data and its technical challenges," *Commun. ACM* 57, 7, pp 86-94.

[3] O'Neil, M. (2014). As Data Proliferate, So Do Data-Related Graduate Programs, *The Chronicle of Higher Education*. DOI= http://m.chronicle.com/article/As-Data-Proliferate-So-Do/144363

[4] Guo, P. (2013). Data Science Workflow: Overview and Challenges, *Commun. ACM Blog*. DOI= http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

[5] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases," *AI Magazine, Volume 17 Number 3*

[6] Shearer, C. 2000, "The CRISP-DM model: The new blueprint for data mining," *Journal of Data Warehousing*, 5(4).

[7] Piatetsky, G. 2014. "CRISP-DM, still the top methodology for analytics, data mining, or data science projects," *KDD News*. DOI= http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[8] Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A., Madden, S., and Parameswaran, A. (2015). "DataHub: Collaborative Data Science & Dataset Version Management at Scale," *Biennial Conference on Innovative Data Systems Research (CIDR)*.

[9] Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. (1993). "Capability maturity model Version 1.1," *IEEE Software*, 10(4), pp 18-27.

[10] Kelly, J., & Kaskade, J. (2013). CIOS & BIG DATA What Your IT Team Wants You to Know. DOI= http://blog.infochimps.com/2013/01/24/cios-big-data/

[11] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 995-1004). IEEE.

[12] Jensen, P. (2004). What is Operations Research? DOI= http://www.me.utexas.edu/~jensen/ORMM/models/unit/or_method/process.html

[13] Jourdan, Z., Rainer, R. K., & Marshall, T. E. (2008). Business intelligence: An analysis of the literature 1. *Information Systems Management*, *25*(2), 121-131.

[14] Britos, P., Dieste, O., & García-Martínez, R. (2008). Requirements elicitation in data mining for business intelligence projects. In *Advances in Information Systems Research, Education and Practice* (pp. 139-150). Springer US.

[15] Krawatzeck, R., Dinter, B., & Thi, D. A. P. (2015, January). How to make business intelligence agile: The Agile BI actions catalog. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 4762-4771). IEEE.

[16] Hackman, J. R. (1987). The design of work teams. In J. W. Lorsch (Ed.), *The Handbook of Organizational Behavior* (pp. 315–342). Englewood Cliffs, NJ: Prentice-Hall.

[17] Delone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: a ten-year update. *Journal of management information systems*, *19*(4), 9-30.

[18] Gao, J., Koronios, A., & Selle, S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. *AMCIS 2015 Proceedings*.

[19] Crowston, K., & Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1-9

[20] Nott, C. (2014). Big Data & Analytics Maturity Model. DOI= http://www.ibmbigdatahub.com/blog/big-data-analytics-maturity-model

[21] Ebner, K., Bühnen, T. and Urbach, N. (2014). "Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments", *Proceedings of the 47th Hawaii International Conference on Systems Sciences (HICSS-47)*.

[22] Chandler, N., & Oestreich, T. W. (2015). Use analytic business processes to drive business performance. DOI= https://www.gartner.com/doc/2994617