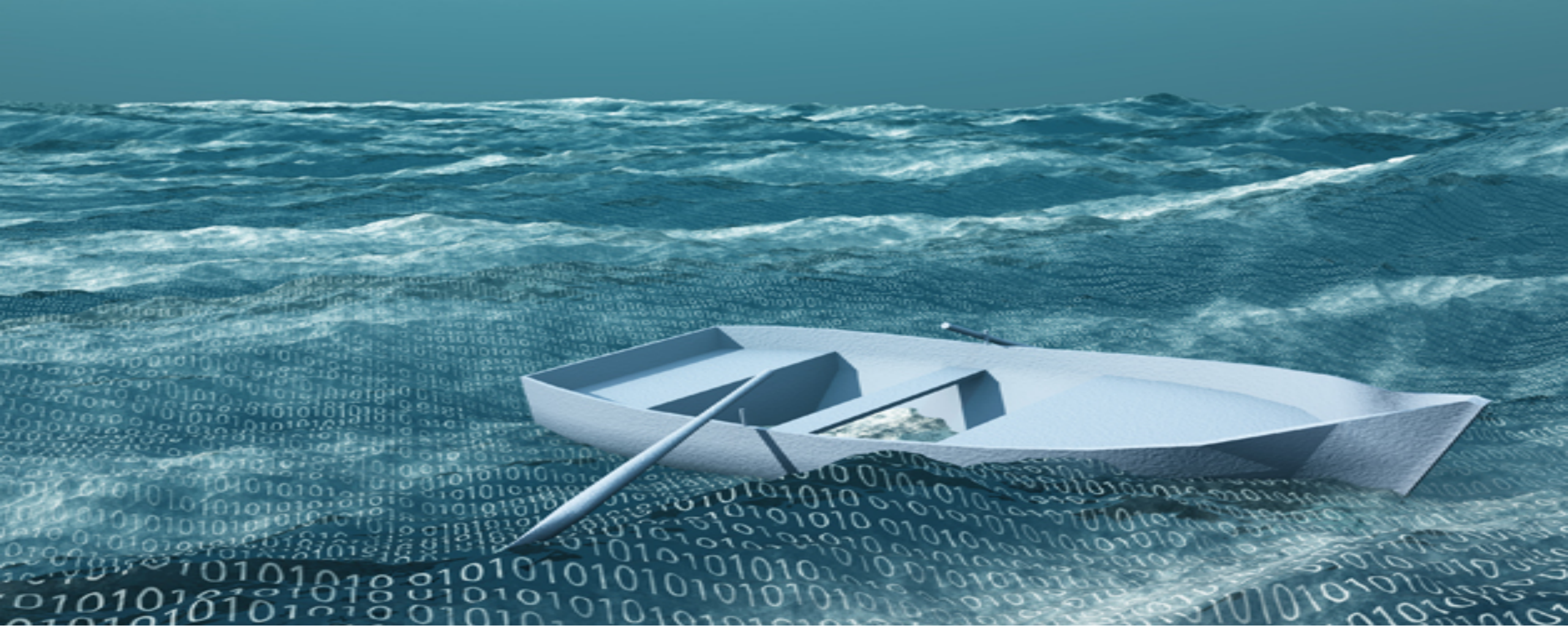


BIG DATA TEAM PROCESS METHODOLOGY: A LITERATURE REVIEW AND THE IDENTIFICATION OF CRITICAL FACTORS FOR A PROJECT'S SUCCESS

School of Information Studies
SYRACUSE UNIVERSITY

J.SALTZ, I.SHAMSHURIN

2016 IEEE INTERNATIONAL CONFERENCE ON BIG DATA

5 DECEMBER, WASHINGTON DC, USA

OUTLINE

Background

Research Questions

Methodology

Data Collection

Findings

Conclusion

Next Steps

INTRODUCTION

- Lack of focus on the process teams should use to actually do a data science project [Saltz, 2015]
- 55% of Big Data projects don't get completed, and many others fall short of their objectives [Kelly&Kaskade]

- Data science teams do not have an explicit data science team-based process methodology:
  - What steps should be done first?
  - How long each phase of a project should take?
  - Which people with what skills should be involved in the project?

- There is no coherent review and analysis of the work that has explored big data project methodologies

RESEARCH QUESTIONS

RQ1. What are the current approaches to plan, organize and perform big data projects?

RQ2. What are the issues and actionable insights that we can synthesize from our literature analysis?

SLR, SOURCES

- Systematic Literature Review (SLR) [Kitchenham&Charters, 2007]
- Sources:
  - Manual Search

    top conferences and journals in 2014 and 2015 according to SRJ ranking and scope;
  - Online Search

    search terms for Google Scholar: “big data” + “process methodology”, “data science” + “process methodology”, “data science” + “team coordination”, “data science” + “project management”; papers from 2011-2015

SELECTION CRITERIA

- if a paper discussed the process (or methodology) the team used to execute the project (was there a discussion on how the project team worked together and coordinated their tasks);
- identified if there was a description of a phase, perhaps pointing out a best practice (such as planning or time spent in a specific phase of the effort);
- if a paper was focused on the maturity and/or the key factors that help drive the success of a big data project.

SELECTION AND CODING PROCESS

	Total
Step 1: automatic search	1645
Step 2: manual search	8383
Step 3: coding, 96% agreement	92
Step 4: detailed analysis	42

TYPES OF IDENTIFIED PAPERS

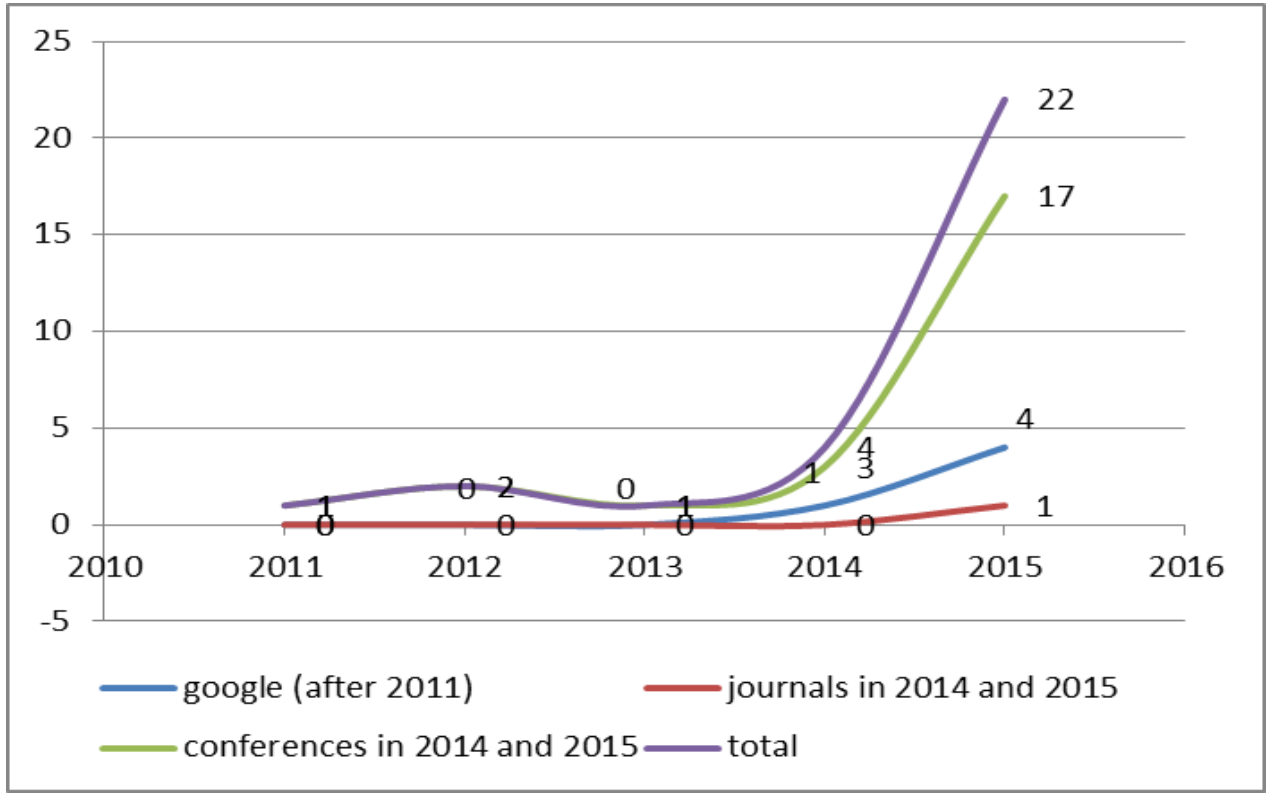
- 42 papers reviewed in step 4

Source	Process	Phase	Maturity
Conferences	12	16	4
Journals	0	0	1
Google	6	1	2
	4journ+ 1book+ 1conf	1conf	1journ+ 1conf
Total	18	17	7

FINDINGS

- Recent Growth
- Team Process Insights
  - The need for a process;
  - Process attributes;
  - The need for effective communication;
- Project execution insights
  - Suggestions across the entire project;
  - Phase specific insights;
- Maturity Levels and Success Factors

FINDINGS: RECENT GROWTH



School of Information Studies
SYRACUSE UNIVERSITY

FINDINGS: TEAM PROCESS INSIGHTS

- The need for a process
  - well-known Data Mining Models CRISP-DM and SEMMA might not be suitable for Big Data projects due its for V's [Duta&Bose, 2015];
  - agile-like methodologies have several advantages over traditional waterfall-like methodologies [Larson&Chang, 2016] such as early time-to market, good collaboration, early risk revealing through iterative development [Dharmapal&Sikamani, 2016];
  - some agile methodologies might be better than other methodologies [Saltz, et al.];

FINDINGS: TEAM PROCESS INSIGHTS

- Process attributes
  - iterative process model is a success factor [Gao et al., 2015];
  - more waterfall like data collection phase followed by a more iterative data analysis phase [Saltz and Shamshurin, 2015] [Vanauer et. al., 2015];
  - multiple models needed based on clarity and complexity of a project [Ahangama&Poo, 2015], type of project (data- or question-driven project) [Vanauer et al., 2015], organization capacity assessment and capacity building [Kee, 2015].
- The need for effective communication
  - including multidisciplinary teams is a CSF [Gao et al., 2015], cross functional team [Dutta&Bose, 2015]

FINDINGS: PROCESS EXECUTION INSIGHTS

- Suggestions across the entire project
  - guidelines for each phase of CRISP-DM [Asamoah&Sharda, 2015] that might be generalized to other domains;
  - big data competence can be measured in each phase of Big Data lifecycle rather than just focus on the analytics phase [Kung et al., 2015];
  - there is a need for mechanisms to evaluate the validity of the results and ensure the analytical results are presented in an appropriate format for the end-user [El-Gayer&Timsina, 2014];

FINDINGS: PROCESS EXECUTION INSIGHTS

- Phase specific insights

  - first step should be getting clarity about the purpose [32] [Das et al, 2015] and understanding the business objectives [Niño et al, 2015];

  - most studies focus on business requirements phase, ex:
    - an integrated methodology that links business concepts with data via a business information model to structure and formalize business requirements [Priebe&Markus, 2015];
    - CRISP-DM based business requirement model: determining business objectives, assessing the situation, determining the data mining goals and producing a project plan [Niño et al, 2015]

  - data storage requirement phase that includes defining strategic goals, decision goals (to answer how strategic goals can be satisfied), and information goals (to answer decision goals) [Di Tria et al, 2014];

  - methodology for model formulation and retrieval, including a Model Management Warehouse that supports the methodology [Corral et al, 2015];

FINDINGS: MATURITY LEVELS

- Organizations with a high level of maturity run their projects holistically, with checks, feedback loops and mechanisms for improvement [Booth, 2015];
- Studer and Leimstoll [Studer&Leimstoll, 2015] provide a step-by-step analytics process that takes into account the organization's analytics maturity across capabilities, culture and technology;

FINDINGS: SUCCESS FACTORS

Groups of CSF synthesized from multiple studies:

- Data (ability to store and access appropriate data), ex. Data & data quality management, Data Integration & Security, Unstructured/structured data, Representativeness of data, Document collection/access to sources
- Governance (well defined roles and responsibilities), ex. Big Data strategy alignment, Project management process defined, Well defined organizational structure, Performance management, Data protection and privacy by design
- Process (using a formal methodology such as Agile), ex. collaboration between IT and business, Communication about the data and initiatives, Flexibility and agility, Project difficulty explored and communicated, Clarity of project deliverables
- Objectives (with measurable KPIs), ex. Focus on small projects and known questions, Specified business case, Feasibility study, Skill gap analysis, well defined scope, measurable project outcome
- Team (skills in data-driven decision-making), ex. Development of skills, People skills & ability to self-organize, Data science, technology, business & management skills, Multidisciplinary team, Stakeholder coordination / shared understanding
- Tools (to enable data-derived insights), ex. Investment in IT infrastructure, technology & tools, Investment in data sources & data storage, Reporting and visualization technology, Discovery technology

CONCLUSION

- identified the most valuable papers with respect to how to execute Big Data projects using a three-criterion based framework to identify Big Data methodology related papers;
- synthesized the knowledge from these studies;
  - there is currently no agreed upon standard for executing these projects and that an improved process methodology would be useful.
- an integrated set of success factors to help practitioners execute Big Data projects was proposed: 33 CSFs were defined, grouped by the six key project characteristics.

POSSIBLE NEXT STEPS

- Comparing the effectiveness of how teams operate using different methodologies;
- Additional case studies;
- Prioritizing, refining, and validating list of identified success factors;
- All CSFs can be ranked according to their importance for the project's success.

THANK YOU

APPENDIX I

Data

- Data & data quality management / ownership
- Data Integration & Security
- Unstructured/structured data
- Representativeness of data
- Document collection/access to sources

Governance

- Management priority / sponsorship / support
- Big Data strategy alignment (with organization's vision)
- Project management process defined
- Well defined organizational structure
- Performance management
- Data protection and privacy by design
- Culture of being Data-driven

APPENDIX II

Process

- Close collaboration between IT and business
- Communication about the data and initiatives
- Flexibility and agility, with freedom for experimentation
- Focus on change management
- Project difficulty explored and communicated
- Clarity of project deliverables (clear or ambiguous)

Objectives

- Focus on small projects and known questions
- Specified business case
- Feasibility study
- Skill gap analysis
- Well defined scope – that understood by the team
- Measurable project outcome

APPENDIX III

Team

- Development of skills / training
- People skills & ability to self-organize when needed
- Data science, technology, business & management skills
- Multidisciplinary team (i.e., across different departments)
- Stakeholder coordination / shared understanding

Tools

- Investment in IT infrastructure, technology & tools
- Investment in data sources & data storage
- Reporting and visualization technology
- Discovery technology