Doing Big Data Projects:
What's the Best Team Process Methology?

October 2015

Executive Summary

What's the Best Team Process Methology?

Executive Summary

What's the Best Team Process Methology?

Nobody Knows

Executive Summary

What's the Best Team Process Methology?

Nobody Knows

- *We don't even know a good process methodology*

Data Science Process Methodology

A Process for Big Data Teams?

Questions to Explore:

- What has been done to date?
- Is there a need for a 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?

Could it improve big data science project effectiveness and efficiency?

Previous Research

Examples of Previous Process Descriptions

Data Science as a "Task Focused" activity

Jagadish:

- Acquisition
- Information extraction & cleaning
- Data integration
- Modeling
- Analysis
- Interpretation and deployment

Guo:

- Preparation
- Analysis
- Reflection
- Dissemination

Previous Research

Examples of Previous Process Descriptions

Not much has changed in the past 20 years!!!

Jagadish:	CRISP-DM*:
• Acquisition	• Business Understanding
• Information Extraction	• Data Understanding
• Data Integration	• Data Preparation
• Modeling	• Modeling
• Analysis	• Evaluation
• Interpretation & Deployment	• Deployment

*Cross Industry Standard Process for Data Mining, defined in the 1990s, used within the KDD community

Previous Research

An Example of Current (Lack of) Research

Very little has been written about methodologies that could enable teams to more effectively and efficiently "do" big data projects.

Analysis of papers/posters from IEEE Big Data 2014

	# of Papers / Posters
Total in IEEE Big Data	295
Papers Focused on Process Methodology	0 (0%)
Papers Mentioning Socio-Technical issues*	23 (8%)

*Typical Examples: Privacy, Enabling humans to help "steer the analysis"

A Data Science Process Methodology

Is there a Need to Define a Process for Big Data Teams?

Current Big Data Focus:

- Improve the algorithm!

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.

Why Does a Team Need to Use a Methodology?

Impact of No Accepted Data Science Process Methodology

Existing methodologies are not being used

- Fewer people are using CRISP-DM [Piatetsky]
  → and more people are using their own methodology

Things could be better:

- Data analysis teams work in an ad hoc fashion [Bhardwaj]
  → In other words – there is a low level of process maturity

- A survey of 300 companies [Kelly & Kaskade] noted that:
  - → 55% of Big Data projects don't get completed
  - → Many others fall short of their objectives

Why Does a Team Need to Use a Methodology?

Impact of a Data Science Process Methodology

With a robust team-based process methodology, one would expect improved project outcomes:

- Many issues to be identified prior to the start of the project
- Issues to be mitigated via the project coordination

Without a well-defined methodology, the team might:

- Forget a step (tasks might be missed)
- Not follow best practices.
- Have lack of coordination across team members

Why Does a Team Need to Use a Methodology?

Why Should My Team Spend Time Improving our Methodology?

Expected benefits when using a well-defined process

- Improve coordination With Others
- Ensure/Improve Quality
- Provide Data Ownership, Security and Privacy
- Prioritize Requirements
- Robust Requirements Analysis
- Deployment (i.e., be able to use in “production”)

Can We Define a Process for Doing Data Science?

Isn't Data Science often "Exploratory"?

Open-ended does not mean "no process"

- Agile software projects can be "open ended" – but it's still a process.
- Benefits of having a robust methodology include ensuring that:
  - There is an appropriate data architecture
  - There is a prioritized set of goals for doing the data analysis.

Doesn't one need a well-defined process
when one organization hires another organization to do data analysis.

Exploring Related Domains

Couldn't Teams Use a Methodology From a Related Domain?

Example domains that could be leveraged

- Software development methodology
- "Small data" quantitative research
- Business Intelligence
- Operations Research

But which one?
Perhaps a combination across several?

Exploring Related Domains

Comparing Software and Data Projects

Similarities	Differences
Understanding the requirements	Data: Not always known, exploratory in nature, capturing / cleaning data
Doing Analysis & Design	QA: Timeliness / accuracy of data, validity & accuracy of analytics
Large number of people that need to be coordinated	Challenges in validating / describing the results.

Agile? Waterfall?

Exploring Related Domains

Comparing to Quantitative Research for “Small” Datasets

Similarities	Differences
Examining the distribution of data to identify possible bias	IT requirements needed to do the analysis
Testing for validity	Large number of people that need to be coordinated
Carefully documenting the steps taken and results generated so that work can be replicated	Challenges in validating / describing the results.

Some possible points of leverage, but some differences

How to Assess Data Team Performance

How Can One Evaluate Which Teams Work “Better”

Starting points

- Critical Success Factors (CSFs)
- General Models of Team Effectiveness
- Information Systems Success Models
- Common Maturity Model for Data Teams

How to Assess Data Team Performance

Does the Team Follow Critical Success Factors?

Critical Success Factors – CSFs [Gao]

- 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)

How to Assess Data Team Performance

General Models?

General Model of Team Effectiveness [Hackman]

- It's not just the team leader:
  - If members behave cooperatively and competently
    → leaders tend to operate more participatively
  - Five conditions enhance team performance & effectiveness
    → real team, compelling direction, enabling structure, supportive context, and competent coaching
- How to describe team behavior and performance:
  - Input factors
  - Process & moderating factors
  - Outputs

How to Assess Data Team Performance

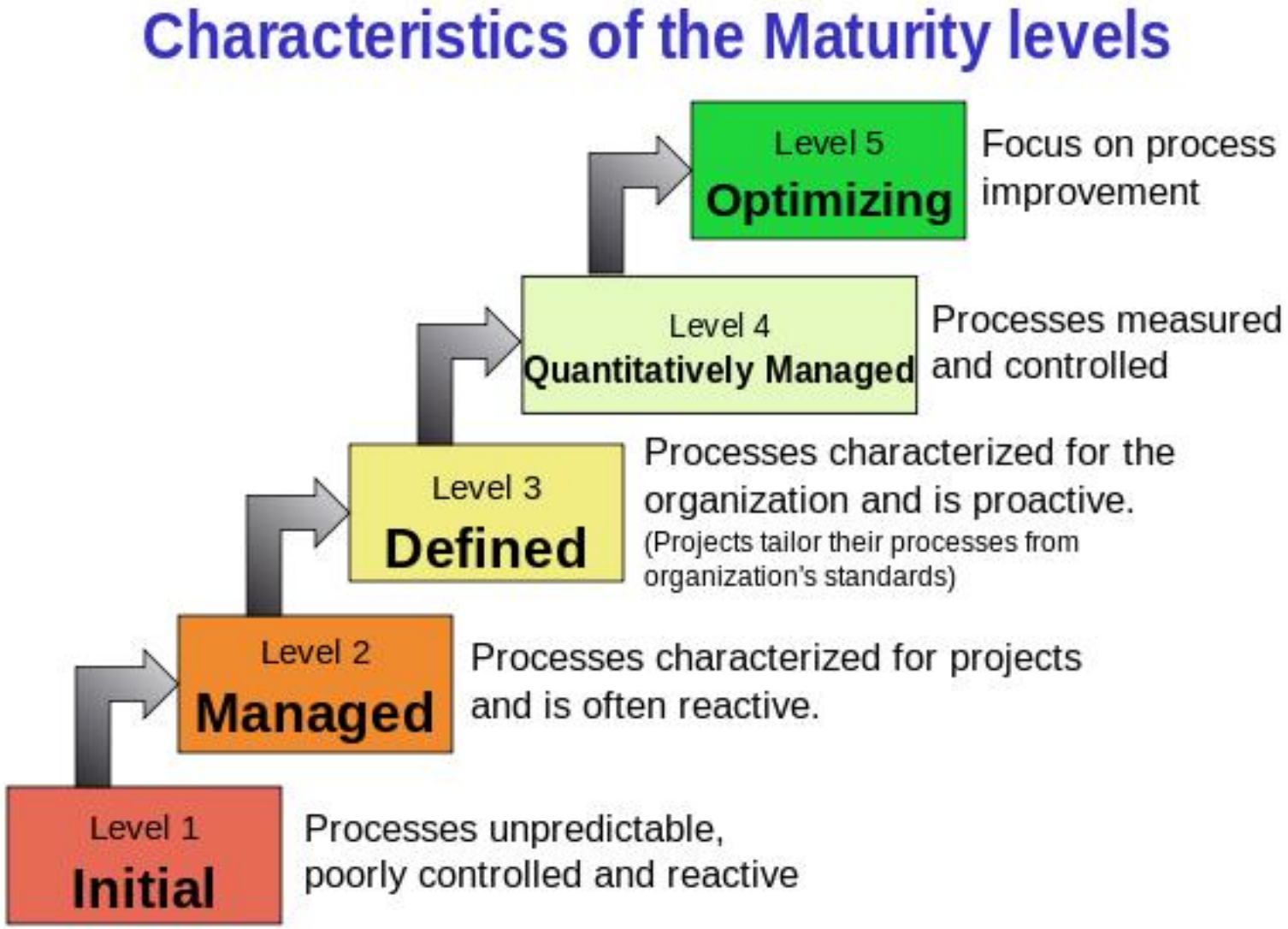
How Successful was the Team?

Information Systems Success Models [DeLone & McLean]

- Creation of a System
- Use of the system
- Impact of the system

How to Assess Data Team Performance

Create a Common Maturity Model to Compare Teams



http://istqbexamcertification.com/what-is-cmm-capability-maturity-model-what-are-cmm-levels/

How Do Teams Operate Today?

Depends on their Frame of Reference

Initial findings after surveying 10 organizations

- Data Science is a new field – so where did the leader come from?
  - Software Development
  - Research
  - ?
- Many teams have given little thought to this topic…

Next Steps

How Do We Start?

Some ideas

- Case studies
- Survey existing big data science papers for any mention of process / methodologies used
- Form a community

People

Cited Articles

- 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.

- 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

- 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

- 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).

- 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/

- 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.

- 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.

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

People

Thanks To the Following:

Providing valuable insight/discussion/help:

- Jason Dedrick
- Kevin Crowston

Ph.D. Students actively involved:

- Ivan Shamshurin
- Ehsan Sabaghian