Attributes of Successful Computational Tools in Big Data Science

Grace Kim, Megan Mogannam and Kerk Kee
COM-491: Spring 2014, Chapman University; Orange, California

Introduction

In the rapidly changing landscape of big data, there has been an emergence of new computational tools and virtual organizations (VO) that enable the tools. Certain attributes of the tool make it well adopted and diffused within the scientific community. However, there are other uses of the tool that go beyond its original development and intended use.

Literature Review

Computational tools are used in a number of interdisciplinary ways that advance research for new discoveries in domains such as biotechnology, astronomy, computer science, chemistry, physics, etc. (Atkins, 2003). For instance, GLEaMviz is a publically available software that has become increasingly important in the assessment and control of public health by predicting global health crises scenarios (Van den Broeck et al., 2011).

A widely used framework for understanding innovation adoption is the diffusion of innovations (DOI) theory by Everett Rogers in 1962. The theory suggests five main factors that influence the adoption of an innovation: (1) *Relative Advantage*, (2) *Compatibility*, (3) *Complexity*, (4) *Triability*, (5) *Observability* (Rogers, 2003). Rogers' theory was used mostly to examine technologies that are pre-designed, mass-produced, bought off-the-shelf, and used-as-instructed after the product is fully developed. But the big data science context is unique because the tools are experimental, developed by lead users, and often in dispersed teams (Kee et al, 2011).

Hence, this study examines the characteristics of experimental computational tools in prototype and incubator phase. Certain characteristics make them well adopted and diffused beyond the incubators to the larger scientific community. Therefore, this study seeks to answer the research question: *What attributes of a computational tool (as a dynamic prototype) impact the tool's adoption & diffusion from one virtual organization to another within the scientific community?*

Discussion

Our findings cross-reference and have similarities with Rogers' theoretical framework. Similar to the *relative advantage* and *complexity*, the greater the *perceived usefulness* of the tool, the faster its rate of adoption. *Triability* is the extent in which uncertainty can be reduced by experimenting with the innovation first before the user commits to adopt. The *triability* of the product can be measured by the open source nature and amount of documentation available to the users. The same exact finding of *compatibility* with existing practices is matched. However, the outlier is the factor of *observability*, which represents the provision of intangible and tangible results that can be seen everywhere. This makes sense because the computational movement is still in its infancy.

Methodology

This poster employed the grounded theory approach (Corbin & Strauss, 1990) and analyzed 25 interviews conducted with domain scientists (in bioinformatics, computational chemistry, theoretical physics, etc.) and computational technologists. Interview participants came from across the US (including CA, IL, IN, SC, MI, TX, etc.) and three from the UK (specifically Scotland). Interviews range from 16 minutes to 2:25 hours, with 10 conducted in person at the Supercomputing 2013 conference in Denver, and 15 over the phone, between Nov 2013 and April 2014. Guided by the stated research question, the co-authors performed multiple iterations of data analysis and literature integration, yielding preliminary findings presented in this poster.

Findings

Usefulness

Tools that are easy to use have greater chances of adoption and diffusion because the user's ability to understand the tool determines its adoptability.

- "… [to be a successful tool]… it's a tool that if somebody asks me, 'Do you know a tool to do X with?', I will say, 'Yes, actually that's a perfect thing to do'. So it's a tool that people want to recommend for… its usefulness." (Statistician, UK, November 18, 2013)
- "I'd say that in order for a tool to be adopted by a virtual organization, it [must] … provide a useful function because otherwise there is no point in using the tool." (Bioinformatics Researcher, CA, March 19, 2014)

Compatibility with Bigger Systems and Existing Practices

Successfully developed tools are built to last because it is made to fit well into a larger technological environment. An innovation that is incompatible within the context of larger systems and preexisting practices will not be adopted as quickly as an innovation that is compatible.

- "Most people focus on some core parts…they've got a new idea, a way to solve a problem, and they implement a tool to do that. But the reality is that almost all these tools have to be put into a bigger system, and if it is not easy to connect and use the tool nobody will use it." (Computer Scientist, IL, November 20, 2013)
- "So sometimes the tools make the existing thing not work well. [Because if] the existing application requires a lot of change, then most of the people are not going to use your tool. [B]ut if your tool can work with the existing tool implementation, then there will be much more option in the process. So nobody really wants to scratch all the work that they have done just to use one tool. But if the tool can work with your current stuff, then it will be adopted." (Computer Science, Graduate Student, IL, November 20, 2013)

Domain Flexibility

By not being too domain specific, a tool has the ability to be used in different disciplines and in other ways beyond the purpose of what it was original designed for because it has the characteristic of domain flexibility.

- "Don't let the domain penetrate into the tool itself. When the VO is maturing, they should really figure out what is specific to the VO and what is their common problem. Not to be too domain specific. The VO has to be domain specific but keep the layer separate…Once you extract out all the domain out of it. Keep the layer separate." (Administrator, IN, November 18, 2013)
- "I think first and foremost, they are for the main scientists… [B]ut in the process of implementing it, they created a new parallelization algorithm so that in itself can become…a good tool for the computer scientists later on to use in some other projects, not just in that one." (Biostatistician, WA, March 26, 2014)
- "So MatLab… it's meant for doing all sorts of maths… in engineering and sciences… [A] lot of people use it. An engineer might use it, a scientist, a [biologist] might use it. Another one called R, just R, which is a statistical package actually is an example of an open source [software]. So R was started by a group in New Zealand, of all things. And it started on like wildfire, and R is the popular open source data packages that are in use. If you just search for R you will find it all over the place. It's a statistical analysis…[My] wife is an economist, and she's learning R and she loves it. She can just use it and do all sorts of things. So that's an example of where R is again used by everybody, scientists on one side, all the way to ecologists [and] people who do not think they are very sophisticated, all the way to very advanced data analytics minded and all that. So those kind of core software define diffusion. The other examples would be GIS software. Google Maps actually is an example. Everybody uses Google Maps." (Data Computer Scientist, CA, November 21, 2013)

Open Source and Free

Well adopted and diffused tools are not "owned" by a specific individual or organization, but emerge as a result of collective efforts where everyone can freely access, use, and modify the tools for themselves.

- "So one of the primary reasons why we adopted AutoDock Vina and use it on a large scale using XSEDE was that it is open source and free. So there are different versions of the program but they are very expensive. We have a fairly small budget. So…it is out there, it's been published. Other people have used it and verified it. So that was really the reasoning…that it had been tried and true and free." (Biologist, MA, April 15, 2014)
- "…To get diffusion into lots of different virtual organizations…it should be open sourced. I think that an open sourced tool is far more likely to move from one domain to another. Cause it's adaptable… [T]hat means it has to have a license, an open sourced license. [I]f you wanted to change from one domain to another, it's got to be accessible to developers so somebody needs to be able to look at it, understand what it does and adapt its function." (Institute Administrator, UK, November 18, 2013)

Documentation

Proper documentation (e.g. manual guides, etc.) for multiple audiences increases the adoption because it aids in the ease of use by new users who are not familiar with the history of the tool.

- "…[I]t's got to be clearly described so there there's clearly something that sort of, some form of documentation that explains what the tool does [and] how it does it." (Institute Administrator, UK, November 18, 2013)
- "The amount of documentation… that is aimed specifically at the people that you have tried to persuade…So if you've persuaded them that they should put effort in from their integration team or their development team, it should be documentation for developers or administrators. If you're talking about usage…by the domain scientists, it should be documentation for the domain scientists." (Computational Physicist & Institute Administrator, UK, November 18, 2013)

Conclusion

Big data science has revolutionized the way we process and analyze immense amounts of data across different disciplines. Hence, computational tools are vital to the advancement of research and knowledge across domains such as computer science, biology, astronomy, chemistry, physics, social sciences, humanities, and others. A diffusible tool is characterized as flexible, adaptable, open source, well documented, usefulness and/or compatibility in the current network of existing systems. Innovations in technology can be the driving force of change in human behavior, civilization, business, and scientific breakthroughs. Understanding why certain innovations spread more quickly or fail determines the future trajectory of how corporations, federal government, and non-profit businesses decide to fund money to institutions or groups of people that can elicit change and challenge the status quo of existing practices and former knowledge.

References

Atkins, D. E. (2003). Transformation through cyberinfrastructure-based knowledge environments. In W.H. Dutton, B. Kahin, R. O'Callaghan, A. W. Wyckoff (Eds.), *Transforming Enterprise: The Economic and Social Implications of Information Technology* (155-175). MA: Massachusetts Institute of Technology.

Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*(1), 3-21.

Kee, K. F., Cradduck, L., Blodgett, B., & Olwan, R. (2011). Cyberinfrastructure inside out: Definitions and influencing forces shaping its emergence, development, and implementation. In D. Araya, Y. Breindl & T. Houghton (Eds.), *Nexus: New intersections in Internet research (pp. 157-189)*. New York: Peter Lang.

Rogers, E. M. (2003). *Diffusion of innovations* (5th Ed.). NY: Free Press.

Van den Broeck, W., Gioannini, C., Gonçalves, B., Quaggiotto, M., Colizza· V., & Vespignani, A. (2011) The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infectious Diseases*, *11*(37). doi:10.1186/1471-2334-11-37