Big Holes In Big Data

29.02.2016

Prof. Tim Menzies (Ph.D.)
tim.menzies@gmail.com
Computer Science
Via the NC State Engineering Foundation
c/o NC State University Department of Computer Science
Campus Box 8206, Raleigh, NC 27695-8206
919-513-4292 (Office) / 919-513-1684 (FAX)

About the NC STATE ENGINEERING FOUNDATION

The NC State Engineering Foundation, Inc. (NCSEF) was chartered in 1944 to secure private financial support critical to the improvement of the College. The Foundation continues to raise funds and engage with alumni to support the important work of the people who define the College. All gifts to the College, its departments and programs, make a difference in the ability of engineering students and faculty to fulfill their own dreams and make a positive impact on the world.

Table of contents

1 Summary

Prof. Tim Menzies of NC State University, USA, wishes to collaborate with IBM on Big Data applications, and specifically seeks an “unrestricted” award through the NC State Engineering Foundation in support of the College of Engineering and the Department of Computer Science to support his teaching and research efforts in this space.

Of particular and mutual interest to IBM are the “big holes” in “big data” research; i.e. whata are the next generation issues that that not being explored by any research.

To this end, Prof Menzies will “watch over” multiple data mining projects on IBM data looking for “anti-patterns”; i.e. actions by analysts that confuse, rather than clarify how conclusions are generated from data mining.

It is recommended that the work be funded as a “gift” not a “contract” since the latter incurs a 48% overhead at NC State while the former does not.

- All work conducted in a private Github repository, shared with the client (so client can look over our week to week work)
- All code written under an MIT open source license so IBM will be free to use any to all of the code developed by NC State.

Project Funding Request: $40K - The requested funding for this project is as an unrestricted research grant made payable to the NC State Engineering Foundation, and as such involves no express or implied deliverables. The gift check should be accompanied by either a letter from the donor acknowledging there are no associated deliverables or expectations.

Schedule:

- Invoicing + hiring + initial investigations: now to June 2016
- Phase1 analysis : June to December 2016
- Phase2 planning: December 2016
- Phase2 analysis : January to May 2017

2 Facilities, Equipment, Other Resources

For2016/2017, this work will be conducted at NC State Computer Science and NCSU has extensive resources to address the above issues.

NCSU completed the Engineering Building II on its Centennial Campus in January 2006. The new building now houses the Computer Science Department with its laboratories, which are accessible to the proposed project. In particular, one laboratory will be devoted to the research tasks of this proposal.

2.1 Facilities within NCS CS Department:

The department has a 108-node compute cluster named ARC with about 2,000 cores (AMD Mangy-Cores), Infiniband QDR interconnect, per node power monitoring, GPUs and SSDs and parallel file system support, which was funded by an NSF CRI that he is the main PI of together with 5 co-PIs. The ARC facility is providing local and remote researchers with administrator/root privileges for Computer Science experiments at medium scale. This allows any of the software layers, including the operating system and Infiniband switch network routing tables, to be modified for experimental purposes, e.g., to experiment with different network topologies. For large-scale demonstrations, remote facilities will be utilized (see below).

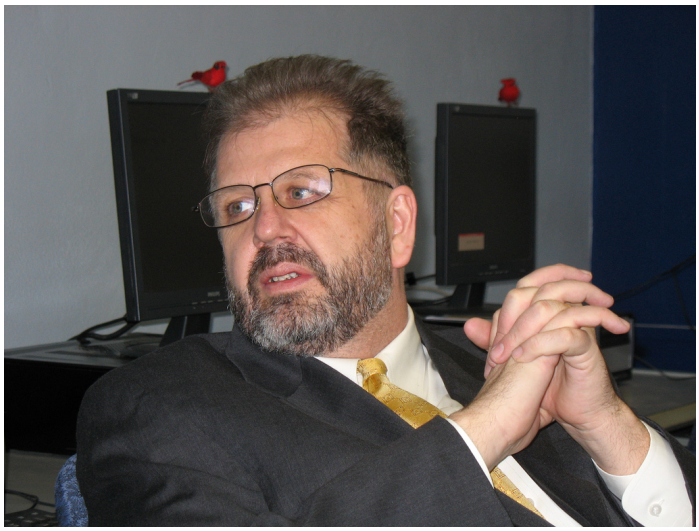
2.2 Computing Resources

The College of Engineering at North Carolina State University has built a distributed computing environment named “Eos” for engineering education. The Eos environment consists of more than 1,000 public and private workstations and supports more than 12,000 users campus wide. The success of Eos in the College of Engineering has spawned similar projects in other colleges and in the campus computing center. Recently, these projects have merged into a single distributed computing system supporting more than 30,000 faculty, staff, and students. This computing environment not only serves the academic computing needs of the campus, but is also becoming the primary means of communication between the students and the faculty. A large number of software packages are available on the Eos system.

2.3 Other Resources

The departments provide the space and basic networking services to carry out the experiments, secretarial and administrative support as well as general-purpose office equipment (e.g., fax, photocopiers, etc.).

3 Brief Bio of Tim Menzies



Tim Menzies (Ph.D., UNSW, 1995) is a full Professor in CS at North Carolina State University where he teaches software engineering and automated software engineering. His research relates to synergies between human and artificial intelligence, with particular application to data mining for software engineering.

In his career, he has earned over $8 million in competitive research funding. He has graduated 7 Ph.D. and 27 masters-by-research students.

He is the author of over 230 referred publications; and is one of the 100 most cited authors in software engineering out of over 80,000 researchers (http://goo.gl/BnFJs). In his career, he has been a lead researcher on projects for NSF, NIJ, DoD, NASA, USDA, as well as joint research work with private companies.

Prof. Menzies is the co-founder of the PROMISE conference series devoted to reproducible experiments in software engineering (http://openscience.us/repo). He is also an associate editor of:

- IEEE Transactions on Software Engineering,
- Journal of Big Data Research
- Empirical Software Engineering,
- Information Software Technology,
- The Automated Software Engineering Journal
- The Software Quality Journal.

In 2015, he served as co-chair for the ICSE'15 NIER track. In 2016, he serves as co-general chair of ICMSE'16. In 2017 he will serve as co-PC chair for the Symposium on Search-Based SE.

For more, see his vita (http://goo.gl/8eNhYM) or his list of publications https://goo.gl/qNQAIq) or his home page http://menzies.us.

4 Organization

- NDAs:
    - All team members sign non-disclosure agreements with IBM.

- MEETINGS:
    - All team members will attend a bi-weekly teleconferences with the client.

- MONITORED:
    - All work conducted in a private Github repository, shared with the client (so client can look over our week to week work)

- RESULTS ALWAYS AVAILABLE:
    - Goal: IBM is free to use any to all of the code developed by NC State.
        - Method: All code written under an MIT open source license so the
    - Goal: IBM has full access to reports on all our interim results).
        - Method: All reports and their associated graphics for the bi-weekly meeting will be made as Github issues

- PAPERS:
    - As to protecting proprietary IBM information, all research papers written as part of this work will be shown to IBM prior to publication. IBM has 30 days to offer either (a) rewrites or (b) require that certain sections be culled (and up to 100% of the paper may be culled in that way).
    - Optionally (but not required) it is hoped that IBM will share authorship with NC State personnel for papers arising from this work.