1 Cover

Faculty Rank of Principal Contact:
Last Name:
First Name:
University:
Department:
Funding Request: $
Is this a Joint Proposal? ☑Yes ☐ No
If Yes, please fill in information for co-proposers (add separate sheets as necessary):
Name_____Rank_____Department_____
Name_____Rank_____Department_____
Is this a Continuation Project? Yes ☑No If Yes, you must complete Appendix A.2
E-mail of Principal Contact: therese.smith@ccsu.edu Phone Number of Principal Contact: 2-2718
Campus Address of Principal Contact: Maria Sanford - 30303

Please select one disciplinary group category in which this project best fits:
☐ Fine Arts and Humanities ☐ Social Sciences, Buisiness and Education ☑Life and Physical Sciences, Mathematics, Computer Science, Engineering and Technology
Please select one research focus area in which this project best fits:
☐ Creation of new knowledge ☑Application of disciplinary/multidisciplinary knowledge, methodologies and/or insights ☐ Production of creative works ☐ Research in student learning

Project Title: Distributed Processing Environment for Computational Medicine

2 Abstract

Abstract

We wish to install a distributed programming environment based upon Hadoop, which will serve several purposes. First, we can teach students with hands-on experience of distributed computing. Second, we can support large datasets, and parallelization of suitable algorithms; this includes algorithms for computational medicine. We can provide more extensive support for our statistical calculations (which support biomarker discovery), and carry them out faster, in the distributed programming framework, taking advantage of multiple inexpensive machines. Biomarkers are helpful in medical diagnosis. Kits that test for biomarkers are a possible product suitable for manufacture.

IRB/IACUC Statement

(If "yes" to either question please see Section 5, p. 3 of the program guidelines)

YES	NO	
☑	☐	Does your research involve human beings as research subjects?
☐	☑	Does your research involve vertebrate animals?

Sign-Off Statement (To be signed individually be each faculty applicant. Please add separate sheets as needed)
I hereby acknowledge my understand that the lack of compliance with the porposal format and other requirements spelled out in the CSU - AAUP Faculty Research Grant Guidelines for the Spring 2018 Competition may result in the proposal being disqualified without review.

Signature of Permanent, Full-Time Faculty	Date
Signature of Permanent, Full-Time Faculty	Date
Signature of Permanent, Full-Time Faculty	Date

3 Narrative

3.1 Significance

We have a three-fold purpose:

- Provide distributed processing infrastructure for our own research in computational medicine
- Provide distributed processing infrastructure for teaching students about distributed algorithms
- Conduct research on biomarkers that is intended to lead to manufacturable test kits, likely with consumables, analogous to home pregnancy tests, or blood sugar testing strips

This activity will develop infrastructure for distributed processing, that complements our department's existing support for cloud computing. This infrastructure will enable us to educate the students with hands-on experience on algorithms that exploit parallel processing in a distributed mode. Such algorithms include Hadoop's Map Reduce applied to a gene analysis toolkit [MHB+10, Tay10], and work with FASTA and FASTQ files [FPRCG17] and others [CGP+16].

We would like to extend our research in the learning of students of computer science about mathematical proofs [SM13, SM14, Smi16] to mathematical proofs about software executing in distributed systems, as are given in Lynch's Distributed Algorithms [Lyn96].

By providing a facility for distributed processing in our computer science (CS) department, and making use of it available to any member of the Biomolecular Sciences department, we hope to increase the visibility of CCSU's CS department. Conceivably in the future we might extend the availability of this facility more broadly.

We hope to use this capability to support our research in computational medicine[AS13, SK99]. We have reached out to the Biomolecular Sciences department, in connection with a planned graduate specialization in computational support for biomedicine. We hope that this facility will support cooperation between our faculty members. Moreover, we hope to support existing projects at the larger scale available with distributed processing, including the search for biomarkers in the relatively restricted contexts of blood tests, urine tests and breath tests. We hope to develop opportunities for commercial activity in the development of these tests, which correspond to manufacturable test kits. Because of the anticipated support for commercial activity, we hope to develop external funding.

3.1.1 Outline of Related Research

Hadoop and software created to work with it has been found very useful in distributed programming generally and as applied to research in bioinformatics [Tay10, FPRCG17] and computational medicine [WLL+11].

Hadoop can aid our objective of finding biomarkers, using survival analysis[Rod15]. We use R in our single processor work, and there is an interface between Hadoop and R[RHa], named RHIPE. We expect the combination of R-based survival analysis and Hadoop to remain active, because it has been shown to be useful in marketing on the Internet[dV13].

Use of survival analysis and Hadoop has provided insight useful to both insurance and health care [DSSS15, BGK+16, htt].

Researchers have used survival analysis in development of a test refining a diagnosis of myocardial infarction[APC+07]. Odom et al. [KSS+15] developed a breath test for malaria.

3.2 Work Plan

The software infrastructure, a modification of Hadoop, has been developed by Roland DiPratti. The plan is to identify machines onto which we can install this software, to make an operating facility. We plan to use machines obtained separately from the grant. T. Smith has some Raspberry Pi machines to offer, on a temporary to-be-returned basis, to the project. Hadoop has been shown by [pih] to be suitable for Raspberry Pi clusters. Raspberry Pis are small, which implies our need for space to be only a few square feet. Then we plan to install and verify the installation of the modified Hadoop. Then we plan to install software used by the application, including Python and R, which are generally useful packages. We plan to investigate a programmatic interface to Mathematica, as it is said to be installed on the Raspberry Pis. In particular, we would like to be able to interoperate Mathematica and R with the modified Hadoop. Then we plan to install software used by the application, a survival analysis package of R, which is more focussed in purpose, though of interest to both insurance and medical applications. Then we plan to load and execute our first application, which is expected to be the biomarker software, which uses

Python and R. This software might be converted into strictly R, if that seems to be desirable. We plan to apply the biomarker software to larger datasets than those to which it has been applied so far; this will be facilitated by the modified Hadoop software. We plan to coordinate this facility's capabilities with Biomolecular Sciences, in order to support any interest they may have in a specialization of software engineering for computational medicine. We plan to coordinate this facility's capabilities with Biomolecular Sciences, in order to support any computational facility they might find useful.

3.2.1 Joint proposal individual contributions and level of participation

T. Smith has contributed, on a to-be-returned basis, the initial hardware.

T. Smith intends to devote 50% time in the summer, and over the break between fall and spring semesters, to this activity.

R. DiPratti intends to contribute the modified Hadoop software, and to install it on any machines used in the platform.

3.3 Outcomes and Reporting

It is certainly our intention to submit the results of our biomarker research to a journal such as BMC Bioinformatics (`https://link.springer.com/journal/12859`). This research is related to our previous research [AS13].

We intend to submit the results of our research in approaches to teaching distributed programming to ICER and Koli Calling, which have accepted our work previously.

4 Budget proposal

4.1 Budget

a separate form Do we need to estimate the impact on the electric bill? Air conditioning will probably not be affected.

4.2 Budget Justification

Table 1: **2018 - 2019 CSU - AAUP Faculty Research Grant**

Budget Item	Amount (No Cents)	Brief Justification
Faculty Stipend		
Support Services *		
Supplies and Equipment		
Travel		
Total		**N/A**

* For definition see Section 9.4 of the "Collective Bargaining Agreement between Connecticut State University, American Association of University Professors and Board of Regents for Connecticut State Colleges & Universities System, August 26, 2016 August 26, 2021", Section 9.4, pp. 56-57.

5 CVs

5.1 CV-TS

Thérèse Smith, PhD

- Spring 2018 full time instructor, Central Connecticut State University and Adjunct Associate Professor at University of Maryland University College, and Adjunct Assistant Professor at University of Connecticut/Hartford
- Fall 2017 – full time instructor, Central Connecticut State University and Adjunct Associate Professor at University of Maryland University College
- Spring 2017 – full time instructor, Central Connecticut State University and Adjunct Associate Professor at University of Maryland University College
- Fall 2016 – part time instructor, University of Rhode Island, and Adjunct Assistant Professor at University of Connecticut/Hartford and Adjunct Associate Professor at University of Maryland University College
- 2009 - 2016 PhD student in Computer Science at University of Connecticut, Storrs
- 2001 – 2010 subcontracted to Federal Aviation Administration (FAA)
- 2000 employee NavCanada
- 1998 founded Air Traffic Software Architecture, Inc.
- 1979 - 1998 Member of the Technical Staff (Full staff), MIT/Lincoln Laboratory

References

Professor James Robertson, EdD
james.robertson@umuc.edu
cell: 443-889-5850.
Computer and Information Science Program Chair
University of Maryland University College
S. K. Bhaskar, PhD, Vice Dean
University of Maryland University College
3501 University Blvd. East, Adelphi, MD 20783 - USA
Phone: 240-684-2840

Professor Reda Ammar
reda.ammar@uconn.edu
University of Connecticut
371 Fairfield Way U-4155
Storrs, CT 06269-4155
860 486-5285

Professor Donald Sheehy
University of Connecticut
371 Fairfield Way U-4155
Storrs, CT 06269-4155
860 486-0006

Professor Robert McCartney
University of Connecticut
371 Fairfield Way U-4155
Storrs, CT 06269-4155
860 486-5232

5.2 CV-RD

5.3 CS-FA

5.4 CV-SC

5.5 CV-KM

References

[APC+07] Ronny Alcalai, David Planer, Afsin Culhaoglu, Aydin Osman, Arthur Pollak, and Chaim Lotan. Acute coronary syndrome vs non-specific troponin elevation: clinical predictors and survival analysis. *Archives of internal medicine*, 167(3):276–281, 2007.

[AS13] Reda Ammar and Therese Smith. Developing time constraints in petri net models of biochemical processes via computation struc-

ture modeling. In *Signal Processing and Information Technology (ISSPIT), 2013 IEEE International Symposium on*, pages 000034–000039. IEEE, 2013.

[BGK+16] Mukesh Borana, Manish Giri, Sarang Kamble, Kiran Deshpande, and Shubhangi Edake. Healthcare data analysis using hadoop. *International Journal of Engineering Science*, 4598, 2016.

[CGP+16] Giuseppe Cattaneo, Raffaele Giancarlo, Stefano Piotto, Umberto Ferraro Petrillo, Gianluca Roscigno, and Luigi Di Biasi. Mapreduce in computational biology-a synopsis. In *Italian Workshop on Artificial Life and Evolutionary Computation*, pages 53–64. Springer, 2016.

[DSSS15] Prashant Dhotre, Sayali Shimpi, Pooja Suryawanshi, and Maya Sanghati. Health care analysis using hadoop. *International Journal of Scientific & Technology Research*, 4(8):279–281, 2015.

[dV13] Andrie de Vries. Using survival analysis for marketing attribution (with a big data case study). In *The R User Conference, useR! 2013 July 10-12 2013 University of Castilla-La Mancha, Albacete, Spain*, volume 10, page 75, 2013.

[FPRCG17] Umberto Ferraro Petrillo, Gianluca Roscigno, Giuseppe Cattaneo, and Raffaele Giancarlo. Fastdoop: a versatile and efficient library for the input of fasta and fastq files for mapreduce hadoop bioinformatics applications. *Bioinformatics*, 33(10):1575–1577, 2017.

[htt] http://www.actuaries.org. Health actuaries and big data.

[KSS+15] Megan Kelly, Chih-Ying Su, Chad Schaber, Jan R Crowley, Fong-Fu Hsu, John R Carlson, and Audrey R Odom. Malaria parasites produce volatile mosquito attractants. *MBio*, 6(2):e00235–15, 2015.

[Lyn96] Nancy A Lynch. *Distributed algorithms*. Morgan Kaufmann, 1996.

[MHB+10] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[pih] http://www.widriksson.com/raspberry-pi-hadoop-cluster/.

[RHa] http://cran.us.r-project.org/web/views/HighPerformanceComputing.html.

[Rod15] James A Rodger. Discovery of medical big data analytics: improving the prediction of traumatic brain injury survival rates by data mining patient informatics processing software hybrid hadoop hive. *Informatics in Medicine Unlocked*, 1:17–26, 2015.

[SK99] Therese M Smith and Patrick A Kelly. Random sets technique for information fusion applied to estimation of brain functional images. In *Proceedings of SPIE-The International Society for Optical Engineering*, volume 3661, pages 1158–1169, 1999.

[SM13] Therese Smith and Robert McCartney. Mathematization in teaching pumping lemmas. In *Frontiers in Education Conference, 2013 IEEE*, pages 1671–1677. IEEE, 2013.

[SM14] Thérèse Smith and Robert McCartney. Computer science students' concepts of proof by induction. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research*, pages 51–60. ACM, 2014.

[Smi16] Thérèse Smith. Categories of conceptions of proofs by students of computer science. 2016.

[Tay10] Ronald C Taylor. An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. *BMC bioinformatics*, 11(12):S1, 2010.

[WLL+11] Fusheng Wang, Rubao Lee, Qiaoling Liu, Abulimiti Aji, Xiaodong Zhang, and Joel Saltz. Hadoop-gis: A high performance query system for analytical medical imaging with mapreduce. *Atlanta–USA: Technical report, Emory University*, pages 1–13, 2011.

Cattaneo G. et al. (2016b). MapReduce in Computational Biology A Synopsis. In Proceedings of the 11th Italian Workshop on Artificial Life and Evolutionary Computation. Springer.

5.6 Optional Appendices

6 proposal review criteria

- coversheet abstract
- signoff statement
- proposal narrative

7 Priorities

7.1 Educational mission

7.2 Visibility

7.3 Research stature