# 

# 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 sup- port 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.

# Narrative

## Significance

The methods to complete a medical diagnosis through non-invasive methods have steadily increased over the last 4 decades. The promise of making an accurate diagnosis without surgery and its associated risks, especially in early stages of the process prior to any symptoms, drives researchers to discover new and less-invasive methods. Examples of this approach include using ultrasound [1] and breathalyzer diagnostics [2].

Incorporating data mining algorithms to large amounts of data can be used by researchers to find patterns that can lead to new non-invasive diagnostics. Using the Hadoop platform has made it possible to incorporate more data into this data mining and decrease the time needed to do this analysis.

Instructing students in recent theory and techniques is an important role that we take on as educators. Companies looking for employees with skills in Big Data technologies exist across all industries ranging from business to health care. An article in *Science Magazine*, published June 2014, states that ‘Big data is everywhere, and its influence and practical omnipresence across multiple industries will just continue to grow. For life scientists with expertise and an interest in bioinformatics, computer science, statistics, and related skill sets, the job outlook couldn’t be rosier” [3]. Across all industries, demand for new hires with Big Data skills has an upward trajectory. According to an article published in *Forbes* *Magazine* in June 2015, “Demand for Computer Analysts with big data expertise increased by 89.9% in the last twelve months” [4].

Our research proposal has four goals:

* + - Determine the best alternative for a reproducible, distributed processing research platform, examining both local and cloud choices;
    - Design and provide distributed processing infrastructure for our own research in computational medicine;
    - 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.
    - Provide distributed processing infrastructure for teaching students about distributed algorithms;

2.1.1 A Reproducible, Distributed Processing Research Platform

Choosing a computer environment for scientific research in a small to medium sized university is often a matter of availability. Teaching universities do not have shared computer resources designated for research or the IT support capabilities that a research university would provide. In a research area that involves big data or large processing needs, this limits which type of research can be undertaken.

With the advent of inexpensive computers (RPi, Arduino) and cloud-computing services (AWS, Google, Azure), researchers have choices that allow them to build research platforms inexpensively. In the case of cloud computing, the researchers are charged only for its use and the cloud provider is responsible for many management issues (hardware upgrades, OS software patches, etc.). Many universities have begun exploring the cloud as collaborative research platform that is often referred to as a science gateway [3] [4]. Cloud platforms are also be used to do distributed public health research [5]. Researchers are also beginning to use the Raspberry Pi as a viable research environment. Although much of the research has focused on using the Raspberry Pi as a teaching tool, researchers have explored its use as a Hadoop environment [4][5] and one team has used it to do biometric analysis for neonatal identification [7].

Our intent is to research which of these choices, Raspberry Pi versus the cloud, will provide the best big data platform to develop a collaborative research environment that can be used for our biomarker research, as well as for future research. By providing a facility for a science gateway, and making it available to any member of the Computer Science and Biomolecular Sciences departments, 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.

The designed platform must have the ability to capture runtime characteristics, including configuration and execution logs, that could be provided for reproducibility.

2.1.2 Use the platform to conduct research on cancer biomarkers

We hope to use this capability to support our research in computational [medicine [AS13,](#_bookmark1) [SK99].](#_bookmark14) We have reached out to the Biomolecular Sciences department, regarding 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.

Hadoop has been an important tool in bioinformatics research [12] [13]. Other researchers have used cloud providers to develop platforms for doing proteomic analysis without using Hadoop 14]. The field of Big Data, although nascent, is a fast-moving area. MapReduce, the original processing engine for Hadoop, has been superseded by a new processing engine, Apache Spark, which due to its more efficient use of memory, has delivered faster turnaround times. Researchers are beginning to apply this new technology to bioinformatic problems [14]. It is also driving the development of Bioinformatic frameworks in Spark to solve genomic processing problems [15]. Our goal is to use Apache Spark to develop solutions that support non-invasive cancer diagnostics.

2.1.3 Support Student Educational Opportunities

This activity will develop infrastructure for distributed processing, that augments 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,](#_bookmark10) [Tay10],](#_bookmark18) and work with FASTA and FASTQ files [[FPRCG17]](#_bookmark6) and others [[CGP+](#_bookmark3)16].

We would like to extend our research in the learning of students of computer science about mathematical proofs [[SM13,](#_bookmark15) [SM14,](#_bookmark16) [Smi16]](#_bookmark17) to mathematical proofs about software executing in distributed systems, as are given in Lynch’s Distributed Algorithms [[Lyn](#_bookmark9)96].

### 2.2 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,](#_bookmark18) [FPRCG17]](#_bookmark6) and computational medicine [[WLL+](#_bookmark19)11].

Hadoop can aid our objective of finding biomarkers, using survival analysis[Ro[d15].](#_bookmark12)

We use R in our single processor work, and there is an interface between Hadoop and [R[RHa],](#_bookmark13) 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 In[ternet[dV13].](#_bookmark5)

Use of survival analysis and Hadoop has provided insight useful to both insurance and health care [[DSSS15,](#_bookmark4) [BGK+16,](#_bookmark2) [h](#_bookmark7)tt].

Researchers have used survival analysis in development of a test refining a diagnosis of myocardial infarction[APC[+](#_bookmark0)07]. Odom et al. [[KSS+15]](#_bookmark8) developed a breath test for malaria.

## 2.3 Work Plan

The software infrastructure, a modification of Hadoop, has been developed by Roland DePratti. 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]](#_bookmark11) 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. 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 focused 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, 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, 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.

1. DePratti 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/](https://link.springer.com/journal/12859) [12859](https://link.springer.com/journal/12859)). This research is related to our previous research [[AS13].](#_bookmark1)

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.

# CVs

## CV-TS

Th´er`ese 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 Lab- oratory

References

Professor James Robertson, EdD [james.robertson@umuc.edu](mailto:james.robertson@umuc.edu)

cell: 443-889-5850.

Computer and Information Science Program Chair University of Maryland University College

1. 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](mailto: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

## CV-RD

* 1. **CS-FA**
  2. **CV-SC**
  3. **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 structure 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.](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/.](http://www.widriksson.com/raspberry-pi-hadoop-cluster/)

[RHa] [http://cran.us.r-project.org/web/views/HighPerformanceComputing.html.](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 im- ages. In *Proceedings of SPIE-The International Society for Optical Engineering*, volume 3661, pages 1158–1169, 1999.

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

[SM14] Th´er`ese Smith and Robert McCartney. Computer science students’ concepts of proof by induction. In *Proceedings of the 14th Koli Call- ing International Conference on Computing Education Research*, pages 51–60. ACM, 2014.

[Smi16] Th´er`ese 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.

[1] Siegel, R., Bar-Cohen, Y, Grandia. 1997. Noninvasive Medical Diagnostics & Treatment Using Ultrasonics. in Proceedings of ASNT '98 Spring Conference.

#### [2] Morad K. Nakhleh et al. 2017. Diagnosis and Classification of 17 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules. *ACS Nano* **2017** *11* (1), 112-125

[3] Rajesh Kalyanam, Rob Campbell, Derrick Kearney, Leif Delgass, Larry Biehl, Lan Zhao, Carolyn Ellis, and Carol Song. 2017. Cloud-enabling a Collaborative Research Platform: The GABBs Story. In Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact (PEARC17). ACM, New York, NY, USA, Article 23, 8 pages.

[4] Jason C. Christopher. 2017. Analytics Environments on Demand: Providing Interactive and Scalable Research Computing with Windows. In Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact (PEARC17). ACM, New York, NY, USA, Article 4, 5 pages.

[5] Kenneth Fox, William Mongan, and Jeffrey Popyack. 2015. Raspberry HadooPI: A Low-Cost, Hands-On Laboratory in Big Data and Analytics (Abstract Only). In Proceedings of the 46th ACM Technical Symposium on Computer Science Education (SIGCSE '15). ACM, New York, NY, USA, 687-687.

[6] Jimmy Lin. 2015. Scaling Down Distributed Infrastructure on Wimpy Machines for Personal Web Archiving. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 1351-1355.

[7] S. Sumathi, R. Poornima, and T. Haripriya. 2016. Implementation of multimodal neonatal identification using Raspberry Pi 2. In Proceedings of the 2nd International Conference on Communication and Information Processing (ICCIP '16). ACM, New York, NY, USA, 88-92.

[8] Jia-kun Li, Le Zhang, and Ming Xiao. 2017. The High Performance Computing Applications for Bioinformatics Research. In Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science (ICBBS '17). ACM, New York, NY, USA, 70-75.

[9] Hamid Mushtaq, Frank Liu, Carlos Costa, Gang Liu, Peter Hofstee, and Zaid Al-Ars. 2017. SparkGA: A Spark Framework for Cost Effective, Fast and Accurate DNA Analysis at Scale. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics (ACM-BCB '17). ACM, New York, NY, USA, 148-157

[10] Urmi Bhayani and John A. Springer. 2015. Toward an analytical framework for proteomics software. In Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '15). ACM, New York, NY, USA, 582-588.

[11] Slagel, J., Mendoza, L., Shteynberg, D., Deutsch, E. W., & Moritz, R. L. (2015). Processing Shotgun Proteomics Data on the Amazon Cloud with the Trans-Proteomic Pipeline. *Molecular & Cellular Proteomics : MCP*, *14*(2), 399–404. http://doi.org/10.1074/mcp.O114.043380

[12] Taylor, R. C. (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*, *11*(Suppl 12), S1.

[13] Pratt, B., Howbert, J. J., Tasman, N. I., & Nilsson, E. J. (2012). MR-Tandem: parallel X!Tandem using Hadoop MapReduce on Amazon Web Services. *Bioinformatics*, *28*(1), 136–137.

[14] Mahidhar Tatineni, Xiaoyi Lu, Dongju Choi, Amit Majumdar, and Dhabaleswar K. (DK) Panda. 2016. Experiences and Benefits of Running RDMA Hadoop and Spark on SDSC Comet. In Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale (XSEDE16). ACM, New York, NY, USA, Article 23, 5 pages.

[15] Frank Austin Nothaft, Matt Massie, Timothy Danford, Zhao Zhang, Uri Laserson, Carl Yeksigian, Jey Kottalam, Arun Ahuja, Jeff Hammerbacher, Michael Linderman, Michael J. Franklin, Anthony D. Joseph, and David A. Patterson. 2015. Rethinking Data-Intensive Science Using Scalable Analytics Systems. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15). ACM, New York, NY, USA, 631-646.

## Optional Appendices

1. **proposal review criteria**

* coversheet abstract
* signoff statement
* proposal narrative

# Priorities

## Educational mission

* 1. **Visibility**
  2. **Research stature**