Fulbright - Charles University Distinguished Chair at the Faculty of Mathematics and Physics

Project Statement

Bringing Statistics to High Performance Computing

George Ostrouchov
Oak Ridge National Laboratory and University of Tennessee

This document is a proposal narrative submitted to the Fulbright U.S Scholar Program for the Charles University Distinguished Chair at the Faculty of Mathematics and Physics 2021/2022 competition. The proposal was selected and approved for funding. Unfortunately, unexpected events prevented me from traveling in early 2022 and led me to decline the grant.

The virtual lectures planned for the Summer semester titled "Adventures in Supercomputing with R" are an outgrowth of this proposal. However, the lectures are not affiliated with the Fulbright program or its sponsors in any way.

As Charles University Distinguished Chair at the Faculty of Mathematics and Physics in the Department of Probability and Mathematical Statistics (DPMS), my hope is to teach, to create collaborative conditions, and to write a book that collectively promote more participation by statisticians and other data scientists in high performance computing (HPC). I will connect the department with the Czech supercomputing center (IT4I) [2] to enable practical components of teaching and research collaborations to take place on IT4I HPC systems. The resulting book will be used in further US teaching opportunities and to bring more international visibility to Oak Ridge National Laboratory in the emerging area of data driven computational sciences.

I am a computational statistician, with a background in mathematical statistics and numerical mathematics, whose career so far has been performing research, consulting, and software development at what has become the world's largest supercomputing center. I view this Fulbright opportunity as my "return" to a statistics environment with the purpose of teaching statisticians and other data scientists to utilize HPC resources and refreshing my perceptions of this environment. Enabling this community on HPC also brings modern statistical methods to HPC, where they are still largely absent. While an HPC scientist may make a glib remark that statisticians only deal with small data sets and are not needed if lots of data is available, a statistician may have a glib reply that same results can be

accomplished without brute force big data approaches if one just thinks about a problem more deeply and uses statistical methods! There is some truth to both of these statements and more interaction between these communities can be beneficial in the US, in Czechia, and around the world.

The project I propose for this Fulbright position is compelling for reasons that include teaching and technical components, community building in Czechia and in the US, potential interactions between the host and home institutions, and a lasting product that continues to build new community interactions. My personal connections to Czechia provide additional motivation and capability. I address each of these in what follows.

Teaching and Technical Components: I propose to teach a class and a seminar series in the Spring semester, 2022, both with practical components at the IT4I Czech Supercomputing center.

The class, titled "Statistical Computing on Parallel Systems with R," will contain a large dose of components from the *programming with big data in R* project (see pbdR.org) that I started eight years ago and continue to lead. But it will be broader and include strategies for speeding up serial code, profiling to find intensive sections, and other parallel approaches already in the base R system. It will also include an overview of current parallel hardware and software that make up today's clusters and supercomputers. Algorithms that we study will be drawn from modern statistical methods. The class will assume upper undergraduate or graduate standing in statistics or a similar background that includes some matrix computation.

R has been described as the *lingua franca* of statistics [4]. This continues today in statistics as well as in other domains that make heavy use of statistics, such as biology, economics, and finance. Most statisticians, including faculty at DPMS, use the R programming language.

The seminar series will have some overlap with the class but be more descriptive of how statistics and HPC can collaborate. Both the use of statistics for HPC and the use of HPC for statistics will be discussed. My hope is that this will generate interactions with PhD students and faculty to discuss their own projects to potentially scale on IT4I systems or propose the use of statistics to improve the state of HPC. A goal from these interactions is to have a paper submission or a poster presentation result. Potentially, other speakers can be invited to this seminar series.

Community in Czechia: The practical component of the class and possibly other collaborations will take place remotely on IT4I resources, which include supercomputers in partnership with EuroHPC Joint Undertaking [5]. The proposed course and seminar series build upon my previous visits to the Czech supercomputing center IT4I, both to present multi-day tutorials at IT4I, one funded by the Partnership for Advanced Computing in Europe [3] and the second visit funded by IT4I itself. It was during my second visit that I also gave a presentation at Charles University at the invitation of Prof. Jaromir Antoch, DPMS, whom I met earlier at an International Statistical Institute biennial conference. My

presentation and tutorials filled classrooms to capacity at each institution, indicating strong interest in the topic.

In preparing this proposal, I reconnected with Dr. Branislav Jasnik, IT4I Supercomputing Services Director, to enquire about the use of IT4I systems. He responded that class use of the systems is highly encouraged and directed me to the application process that will be required about 3 months prior to the project start. Ideally, this class and collaborations can be a catalyst for future inclusion of IT4I resources in DPMS curricula and research as well as collaborations that benefit IT4I operations. There are other potential connections around the premise of bringing statistics to HPC, for example at the Czech Technical University in Prague (CTU).

Host and Home Institution Interactions: ORNL is the largest open science laboratory in the US. Charles University is the largest and arguably the top Czech university. Its Mathematics and Physics Faculty, colloquially known as "Matfyz" has a remarkably large discipline overlap with ORNL and especially ORNL's Computer Science and Mathematics Division where I am located. Overall, ORNL is more applied with large user facilities and Matfyz is relatively more theory focused.

My 35-year career at ORNL as a consulting and research statistician has enabled me to have a broad familiarity with ORNL facilities and research programs as well as with many people who run them. A ready "Research and User Facilities at ORNL" presentation that I can deploy for various Charles University departments or even any Czech organization can be a catalyst to starting conversations. A more targeted presentation "Bringing Statistics to Supercomputing" can be useful for disciplines closer to applied statistics and other data sciences.

As the current COVID-19 situation is changing the landscape of scientific interactions to online meetings and improving online meeting capablities, it is easy to start conversations between the institutions. I also hope to arrange a small number of in-depth meetings and talks that would require travel funding.

An interesting connection between Charles University and ORNL from a long time ago is that Albert Einstein was briefly one of Charles University's faculty members and he was also instrumental in the creation of the US National Laboratory system, which includes ORNL.

Community in the US and Lasting Product: The Oak Ridge National Laboratory, the largest open science US Department of Energy (USDOE) Laboratory, has almost no involvement in Fulbright programs at this time, either as a host institution or home institution. On the other hand, the University of Tennessee (UT), which is a partner institution in UT Battelle LLC, the corporation that operates ORNL, already participates at a high level. I am communicating with ORNL's Office of Research Excellence to consider joining the UT effort for joint participation. I also hold a Joint Faculty appointment at the Univerity of Tennessee that may help facilitate this process.

The proposed project fulfills a personal goal to write a book for statisticians about statistical computing on HPC platforms, a planned artifact from my course lecture notes. The

pbdR project currently has a book-length set of tutorial notes that can serve as base material to be broadened and updated into the proposed book.

As it is my mission to connect more statisticians in the United States to its abundant supercomputing centers, I plan to use the book to apply for a traveling course program sponsored by the American Statistical Association Council of Chapters [1]. The book will also bring more international visibility to Oak Ridge National Laboratory in the emerging area of data driven computational sciences.

Personal Connections: I have some family background and a last name that is ethnically Russian (so I am often perceived as an American with Russian heritage) but I was born in Prague, I have some Czech heritage, and I completed my elementary education there. As far as languages, English is my preferred language in both personal and technical communication. My Czech is native although lacking in technical vocabulary. I am hoping to improve this through my visit. I also speak fluent Russian but not as well as I speak Czech or English.

These personal connections make the Fulbright opportunity more compelling beyond the already strong technical and community building aspects I propose. My knowledge of the Czech language should make the teaching and ambassadorial aspects of this Fulbright visit easier and more effective.

Teaching Experience and Philosophy: In teaching, I will draw primarily on my experience with multi-day tutorials (IT4I: Ostrava, Czechia; GIAN: Mangalore, India; ICSA Conference Course: Kerala, India; ISM: Tokyo, Japan; UseR! Conference, Spain; JSM Conference, USA; Several ORNL tutorials)[1] and on my experience with mentoring summer student projects (US Department of Energy Office of Science Outstanding Mentor Award)[2].

A multi-day tutorial is much more concentrated than a class so considerable adaptation will be needed. Aside from decomposition into digestible chunks, practical components and intuitive explanations will be added.

In teaching mathematical, statistical, and computer concepts, It is important to describe intuition, not just the concept, and connect it to broader aspects. The broader context may be different for every student so that interactive feedback is an important component of teaching. The ultimate goal is understanding and that is possible only if it is connected to what the students already know.

The perceived distance between student and professor is greater in most countries than it is in the US and being a "Distinguished Chair" can only make this distance greater! Although I tend to be informal in my interactions with students, additional sensitivity in this respect is needed. Getting interactive feedback about understanding and giving positive reinforcement are important. Matfyz students are expected to function with English language instruction and many are international students who are not Czech, nevertheless my knowledge of Czech should help in that interaction.

[1]

[2]See CV for details and acronyms

Research: I will mostly concentrate on writing the book. However, I expect that my interaction with the students and faculty, and particularly with the PhD students, will enable me to bring more HPC into their research and encourage them to use statistics for enabling HPC. My background in statistics and experience in numerical mathematics on parallel resources in HPC can serve as a catalyst to scale their research ideas.

References

[1] ASA. Council of chapters traveling course. `https://community.amstat.org/coc/chapterresources/travelingcourse`, 2020. [Online; accessed 15-September-2020].

[2] IT4I. It4innovations National Supercomputing Center. Link. [Online; accessed 15-September-2020].

[3] PRACE. Partnership for Advanced Computing in Europe (PRACE). `https://prace-ri.eu/`, 2020. [Online; accessed 15-September-2020].

[4] David Smith. The language of Statistics. Technical report, 2010.

[5] Wikipedia contributors. European High-Performance Computing Joint Undertaking — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=European_High-Performance_Computing_Joint_Undertaking&oldid=941177922`, 2020. [Online; accessed 15-September-2020].