Scientific Computing, High-Performance Computing and Data Science in Higher Education

Lessons learned from the training program in a super-computer center in Canada

Marcelo Ponce
Erik Spence
Ramses van Zon
Daniel Gruner
mponce@scinet.utoronto.ca
ejspence@scinet.utoronto.ca
rzon@scinet.utoronto.ca
dgruner@scinet.utoronto.ca
SciNet HPC Consortium, University of Toronto
Toronto, ON, Canada

ABSTRACT

We present an overview of current academic curricula for Scientific Computing, High-Performance Computing and Data Science. After a survey of current academic and non-academic programs across the globe, we focus on Canadian programs and specifically on the education program of the SciNet HPC Consortium, using its detailed enrollment and course statistics for the past six to seven years. Not only do these data display a steady and rapid increase in the demand for research-computing instruction, they also show a clear shift from traditional (high performance) computing to data-oriented methods. It is argued that this growing demand warrants specialized research computing degrees.

CCS CONCEPTS

• **Social and professional topics → Model curricula**; **Computing education programs**; *Accreditation*;

KEYWORDS

Training and Education, Scientific Computing, High-Performance Computing, Data Science, Master's Program

1 INTRODUCTION

The computational resources available to scientists and engineers have never been greater. The ability to conduct simulations and analyses on thousands of low-latency-connected computer processors has opened up a world of computational research which was previously inaccessible. Researchers using these resources rely on scientific-computing and high-performance-computing techniques; a good understanding of computational science is no longer optional for researchers in a variety of fields, ranging from bioinformatics to astrophysics.

Similarly, the advent of the internet has resulted in a paradigm where information can be more easily captured, transmitted, stored, and accessed than ever before. Researchers, both in academia and industry [19], have been actively developing technologies and approaches for dealing with data of previously-unimaginable scale. Researchers' ability to analyze data has never been greater, and many branches of science are actively using these newly-developed techniques.

Unfortunately, the skills needed to harness these computational and data-empowered resources are not systematically taught in university courses [20]. Some researchers, postdocs and students may find non-academic programs to fill this void, but others either do not have access to these courses or cannot commit the time to follow them. These researchers typically end up learning by trial and error, or by self-teaching, which is rarely optimal.

A number of academic programs that aim to address this issue have emerged at universities across the world (a few examples are [11, 28]). Some of these grew out of the training efforts of High Performance Computing (HPC) centres and organizations (*e.g.* [27]). Recognizing the need for additional skills in their users, computing centres such as those in the XSEDE partnership [29] in the U.S., PRACE in Europe, and Compute/Calcul Canada have been providing local and online HPC training as part of their user support. Universities have also developed graduate programs in both Scientific and High-Performance Computing, to train scientists and engineers in the use of these computational resources.

A more-recent complement to these graduate programs is the development of the degree in Data Science (DS), that is, degrees which focus on the analysis of data, especially at scale. These degrees come in a variety of forms, from multi-year academic graduate programs to specialized private-sector training. These programs are in strong demand at present, as large companies have discovered the value in thoroughly analysing the vast quantities of customer data which they collect. It is expected that this field will continue to grow, and academic programs will continue to be introduced to meet this demand.

https://doi.org/10.22369/issn.2153-4136/10/1/5

The SciNet HPC Consortium [15, 26] is the high-performance-computing center at the University of Toronto. SciNet provides both computational resources and specialized user support for Canadian academic researchers, and as members of its support team, we are responsible for training researchers, postdocs and graduate students at the University of Toronto in HPC techniques. In this paper we provide a review of the current state of graduate-level Scientific Computing, High-Performance Computing and Data Science academic programs, and endeavour to share our training and teaching experiences as they might result useful for other super-computer centers. The paper is organized as follows. In Sec. 2 we discuss how computation has become an essential ingredient in many academic research endeavours; in Sec. 3 we review the current status of education in the areas of High-Performance and Scientific Computing. In Sec. 4 we present the Data Science education efforts at the academic and non-academic level. We conclude with final remarks and perspectives for the future in Sec. 5 and 6.

2 THE ROLE OF HPC IN CURRENT RESEARCH

It is essentially impossible to give an overview of *all* the uses of computational methods in current scientific and academic research. We will nonetheless attempt a review of at least some computational scientific research, since the way computers are used in research (and other realms of inquiry) influences what should be taught to students.

Astrophysical computational research inherently involves large scale computing, such as the simulation of gravitational systems with many particles, magnetohydrodynamic systems, and bodies involving general relativity. Atmospheric physics requires large weather and climate models with many components to be simulated in a variety of scenarios. High-energy particle physics projects, such as the ATLAS project at CERN, require the analysis of many recorded events from large experiments, while other high-energy physics projects have a need for large scale simulations (*e.g.* lattice QCD investigations). Condensed matter physics, quantum chemistry and materials science projects must often numerically solve quantum mechanical problems in one approximation or another; the approximations make the calculations feasible but still rely on large computing resources. Soft condensed matter and chemical biophysics research often involve molecular dynamics or Monte Carlo simulations, and frequently require sampling a large parameter space. Engineering projects can involve optimizing or analyzing complex airflow or combustion, leading to large fluid dynamics calculations. Bioinformatics often involves vast quantities of genomic input data to be compared or assembled, requiring many small computations. Research in other data-driven fields such as social science, humanities, health care and biomedical science [13], is also starting to outgrow the capacity of individual workstations and standard tools.

Examining these cases in more detail, one can distinguish different ways in which research relies on computational resources:

(1) Research that is inherently computational, *i.e.* it cannot reasonably be done without a computer, but which requires relatively minor resources (*e.g.* a single workstation).
(2) Research that investigates problems that do not fit on a single computer, and therefore rely on multiple computing nodes attached through a low-latency network.
(3) Research that requires many relatively small computations.
(4) Research that requires access to a large amount of storage, but not necessarily a lot of other resources.
(5) Research that requires access to a lot of storage, on which many relatively small calculations are performed.

The distinction between the various types of research determines the appropriate systems and tools to use. Graduate students that are just starting their research often do not have enough knowledge to make the distinction (as nobody has taught them about this), let alone select and ask for the resources that they will need [20].

Note that all five categories fall under "Advanced Research Computing" (ARC). The categories are not mutually exclusive, but research of the second and third kind are usually associated with HPC, while the fourth and fifth, and sometimes the first, are associated with Data Science (DS). Although there is a lot of overlap between HPC and DS, these fields require somewhat different techniques. For that reason, we will consider separate programs for HPC and DS.

3 PROGRAMS IN HIGH-PERFORMANCE COMPUTING

Much of the research presented in the previous section falls in the category of *Scientific Computing* (SC). The growth in the computational approach to research, both academic and industrial, has prompted some institutions to develop graduate-level programs crafted to teach the skills needed to design, program, debug and run such calculations. These programs, having been in development for more than two decades, are now fairly widespread and mature, and are known by the names of "Scientific Computing" or "Computational Science and Engineering". Scientific Computing graduate degrees are offered internationally in several graduate education hubs around the world (U.S., England, Germany, Switzerland, *etc.*, — lists of which can be found at the SIAM and HPC University [14] websites). Canada is no exception, with at least eight universities offering graduate-level programs in Computational Science. These programs include one-year and two-year Master's programs, as well as Ph.D. programs. Most of these programs (*e.g.* the ones shown in Tables 1 and 4) require a final thesis. The projects and theses are faculty-guided research projects and are usually one-term long, though, as with all research, these projects sometimes take longer.

A typical curriculum for a two-year Master's program in Scientific Computing (in this case from San Diego State University) is presented in Table 1. It clearly shows that Scientific Computing has its roots in research in the physical sciences; the program heavily emphasizes numerical analysis and scientific modelling. In some ways this is not surprising: computers are very apt at solving such problems, and the formalism of the physical sciences often lends itself easily to computer programming. Other topics of study which are also often encountered in these programs include finite element analysis, matrix computations, optimization, stochastic methods, differential equations and stability.

In contrast to Scientific Computing, HPC requires somewhat wider knowledge; its practitioners need to understand more than

Course Name	Type
Introduction to Computational Science	required
Computational Methods for Scientists	required
Computational Modelling for Scientists	required
Computational Imaging	required
Scientific Computing	required
Applied Mathematics for Computational Scientists	required
Seminar Problems in Computational Science	required
Computational and Applied Statistics	elective
Computational Database Fundamentals	elective
Research	required
Thesis	required

Table 1: The curriculum for the two-year Master's program at the Computational Science Research Center at San Diego State University [9]; this forms a good example of a typical Scientific Computing graduate program.

just the theoretical and numerical principles. They require skills such as serial and parallel programming (often in several languages, and on different platforms) and scripting, as well familiarity with numerics, data handling, statistics, and supercomputers and their technical bottlenecks. In addition, these practitioners are usually not computer science students, so they must cope without that background. This is somewhat unavoidable as they need to have sufficient domain knowledge as well. Much of the same holds true for Data Science.

3.1 Academic HPC Programs

There are not many academic programs that focus on HPC. Part of the reason may be that such programs require access to a high-performance-computing machine so that students can develop their skills on real hardware, in a real supercomputing environment. These machines require multiple computing nodes which are connected by a low-latency network. Fortunately, such systems do not need to be local: as long as the machine is accessible through the internet the machine could be used for teaching. Nonetheless, having the hardware local to the students lends advantages, since most of the administrators and analysts of the system are typically available to assist students with optimizing their codes and developing good computational strategies. Not surprisingly, the majority of the currently offered HPC graduate programs seem to have been developed by or in conjunction with supercomputer centres. Availability of HPC resources at the local institution or department level varies significantly by country or region. While most top-ranked institutions in the US have HPC facilities, the funding model in Canada [6] is such that all the HPC resources are concentrated in a few national centers (like SciNet).

As examples of High Performance Computing programs, the University of Edinburgh (UK) offers an MSc in High Performance Computing, the Universitat Politècnica de Catalunya/Barcelona Tech (Spain) offers a Master in High Performance Computing and a Master program in Data Mining and Business Intelligence, SISSA/ICTP in Italy offers a Master in High Performance Computing, while a collaboration between the University ITMO (Russia) and the University of Amsterdam (Netherlands) offers a Double-Degree Master Programs in Applied Mathematics and Informatics (Computational Science). Note that many of these programs emerged from locations with a very strong tradition and consolidated background in HPC.

3.2 SciNet's HPC Programs

Many HPC centers provide training for their users to fill the computational-skills gap for the wider scientific community, such as, SDSC, PSC, TACC, NCSA, BSC, EPCC, CSCS, SHARCNET, AceNet, Calcul Québec, among many others. In its capacity as an HPC and ARC centre based at the University of Toronto, SciNet has developed several education and training programs [24] aimed at helping students and users obtain the skills and knowledge required to get the most out of advanced-research-computing resources. SciNet's training events and courses are currently taken by researchers, postdocs, and graduate students across many different departments and even from outside of the University of Toronto (UofT). Some of these courses are considered part of the official curricula and count as graduate level courses within the Ph.D. programs at UofT.

Initially SciNet provided training specifically oriented toward Scientific Computing, with the purpose of maximizing user productivity. These early classes focused on parallel programming (MPI and OpenMP), best coding practices, debugging, and other scientific computing needs. Over the years the breadth of courses has grown, with classes offered in Linux shell programming, parallel input/output, advanced C++ and Fortran coding, accelerator programming, and visualization. This is in addition to the annual HPC Summer School which SciNet runs in collaboration with two other HPC centres within Compute Ontario[4], CAC[1] and SHARCNET[2]. The summer school is a week-long intensive workshop[1] on HPC topics, and more recently, also Data Science and Medical/Bio-Informatics topics. Table 2 shows the curriculum of our summer school for last year.

Table 3 shows the training events and courses that SciNet has already been teaching in the areas of HPC and Data Science. The number and types of classes which SciNet teaches have grown significantly [25]. This can be seen in Figure 1, which presents the total student class-hours taught by SciNet over the last six years. This remarkable growth is a testament to the latent need for this material to be taught. The need for this training is supported by the enrolment statistics: our students constitute 35% of SciNet's total users, clearly showing that even in a specialized audience this kind of training is still needed.

For several years the four-week graduate-style classes offered by SciNet have been accepted for *graduate* class credit by the departments of Physics, Chemistry and Astrophysics at UofT. This was possible by accepting the classes as "modular" (or "mini") courses, one-third semester long, and bundling three such classes into a full-semester course. This arrangement has been so popular with students and faculty that the Physics Department recently listed SciNet's winter twelve-week HPC class in the course calendar [5], allowing graduate students from other departments in the university to take the class for university credit. Following exactly the same path, a six-week module designed to teach students from Biological and Medical Sciences, basics on data analysis with emphasis

[1]Similar initiatives and trends are being carried on by the International HPC Summer School [3] within the theme of HPC Challenges in Computational Sciences.

HPC Stream	Data Science Stream	BioInformatics/Medical Stream
Introduction to HPC & SciNet	Introduction to Linux Shell	PLINK
Shared Memory Programming with OpenMP	Introduction to R	Next Generation Sequencing
Programming Clusters with MPI	Data Science with Python	RNASeq Analysis
Programming GPUs with CUDA	Parallel R for Data Science	Python for MRI Analysis
Debugging and Profiling	Python for HPC (Parallel Python)	Image Analysis at Scale
Bring your own code Labs	Visualization with Python	Machine Learning for NeuroImaging
	Scientific Visualization Suites	R for MRI Analysis
		Public Datasets for Neuroimaging
		HCP with HPC: Surface Based Neuroimaging Analysis

Table 2: Curriculum for the 2017 SciNet's summer school, showing three parallel streams: the traditional High-Performance Computing, Data Science and adding a new parallel stream on BioInformatics/Medical applications. This type of events not only benefit the students and participants of the summer school, but also enables collaborations between departments and consortia, as it was this particular case, where part of the training was delivered in partnership with colleagues from SHARCNET and the Center for Addition and Mental Health (CAMH).

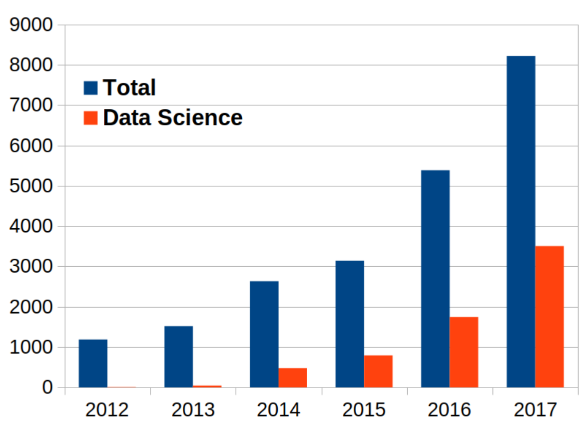

Figure 1: Attendance hours at SciNet training and education events, per year, for all SciNet classes and Data Science specific classes.

Course Name	Certificate	Credits
Quantitative Applications for Data Analysis‡	DS/SC	36
Introduction to Computational BioStatistics with R‡2	DS/SC	36
Introduction to Neural Network Programming	DS/SC	4
Neural Network Programming	DS/SC	16
Advanced Neural Networks	DS/SC	4
Intro to Apache Spark	DS	3
Machine Learning Workshop	DS/SC	6
Hadoop Workshop	DS	3
Scalable Data Analysis Workshop	DS	12
Relational Database Basics	DS/SC	6
Storage and Input/Output in Large Scale Scientific Projects	DS/SC	6
Workflow Optimization for Large Scale Bioinformatics	DS/HPC/CS	6
Python for High Performance Computing	DS/HPC/SC	12
Parallel R	DS/HPC/SC	3
Python GUIs with Tkinter	DS/SC	2
Scientific Visualization	DS/SC	6
Visualizing Data with Paraview	DS/SC	6
Scientific Computing for Physicists‡3	HPC/SC	36
Intro to Programming with Python‡	SC/DS	12
Intro to Research Computing with Python‡	SC	8
Intro to High Performance Computing	HPC	3
Advanced Parallel Scientific Computing	HPC	12
Intro to Scientific C++	HPC/SC	6
Intro to Scientific Programming with Modern FORTRAN	HPC/SC	7
Intro to Parallel Programming	HPC/SC	7
Programming Clusters with Message Passing Interface	HPC/SC	12
Programming Shared Memory Systems with OpenMP	HPC/SC	6
Practical Parallel Programming Intensive	HPC/SC	32
Intro to GPGPU with CUDA	HPC/SC	9
Programming GPUs with CUDA	HPC/SC	12
SciNet/CITA CUDA GPU Minicourse	HPC/SC	12
Coarray Fortran	HPC/SC	2
Parallel I/O	HPC/SC	6
Debugging, Optimization, Best Practices	HPC/SC	6
HPC Best Practices and Optimization	HPC/SC	3
HPC Debugging	HPC/SC	3
Intro to the Linux Shell	HPC/SC	2
Advanced Shell Programming	HPC/SC	3
Seminars in High Performance Computing	HPC/SC	4
Seminars in Scientific Computing	HPC/SC	4

Table 3: Courses taught by SciNet on *Data Science* (DS), *High-Performance Computing* (HPC), and *Scientific Computing* (SC). ‡ denotes courses already recognized at the University of Toronto in several departments, such as, Physics, Astrophysics, Chemistry, Ecology and Evolutionary Biology, Institute of Medical Science, Physical and Environmental Sciences, Engineering, as graduate level credits. We should also mention that students from other universities in the province of Ontario –e.g. Ryerson University– were allowed to enroll in some of these graduate courses for credit via a provincial academic transfer program.

in statistical analysis using the R Statistical Language [18], has now become one of the most popular courses at the Institute of Medical Science at UofT [7].

The skills that SciNet aims to transfer are rare and sought-after, and complement and enhance the skills students learn in regular curricula. That is why SciNet has developed a set of *Certificate Programs* [23], that users and students can pursue in *Scientific Computing*, *High Performance Computing*, and/or *Data Science*, once they have completed enough credit-hours. As a document that proves the holder has highly competitive skills, and in lieu of graduate credit for most SciNet courses, the certificates are in high demand. In a resounding endorsement of our teaching, thus far students have completed a total of 161 certificates (116 in Scientific Computing, 23 in High-Performance Computing and 22 in Data Science[4]). According to the current registration and trends, we are projecting to have over 250 certificates completed by mid-2018. Moreover, anecdotal feedback from some of our students suggests that the courses and their SciNet certificates were instrumental in successful job applications, in industry and in the financial sector.

4 PROGRAMS IN DATA SCIENCE

The wide adoption of the internet in the professional and the personal sphere ushered in the age of "Big Data". The ease of recording of people's online behaviour, and the ability to rapidly move data, lead to a large, diffuse, complex amount of data waiting to be mined for useful information. Because of the typically large size of the data special hardware and training are often needed. In contrast to Scientific Computing and HPC, there are many applications of Data Science in the private sector, e.g. in the medical, banking, retail, insurance, and internet industries.

Bioinformatics also has a large component in the academic world. Though a more-recent addition to the HPC disciplines, the bioinformatics field is well-populated with graduate programs, a testament to its rapid growth and latent demand. Its emergence as a major user of HPC systems has resulted in the development of "Master's of Bioinformatics", and related degrees. A typical Master's program is outlined in Table 4, this one from the Indiana-Purdue University at Indianapolis. While having many features in common with a more-standard SC Master's program, such as the study of programming and algorithms, it exhibits the particular needs of the bioinformatics community, stressing the importance of genetics and biological processes, and a lesser emphasis on mathematics and programming theory.

Degrees in Data Science are relatively new, with the first Master's program only being introduced in the U.S. (by North Carolina State University) in 2007. A sample of some of the classes offered in one such program is given in Table 5. As can be seen, these programs have a strong focus on data, with statistics, machine learning, and databases being their standard focus. Analyzing data that are too big to fit on a standard desktop computer requires specialized equipment; such training is also part of these graduate-level programs, as indicated by the presence of the "Cloud Computing" and "Distributed Systems" classes. Like typical graduate-level programs, these degrees usually require a final project or thesis to be presented by the student.

[4]This number has triplicated since the launch of the program in 2016.

Course Name	Type
Introduction to Bioinformatics	required
Seminar in Bioinformatics	required
Biological Database Management	required
Programming for Life Science	required
High Throughput Data in Biology	required
Machine Learning in Bioinformatics	elective
Computational System Biology	elective
Structural Bioinformatics	elective
Transitional Bioinformatics Applications	elective
Algorithms in Bioinformatics	elective
Statistical Methods in Bioinformatics	elective
Computational Methods for Bioinformatics	elective
Next Generation Genomic Data Analytics	elective
Next Generation Sequencing	elective
Bioinformatics Project	required

Table 4: The curriculum for the "Project Track" two-year Master's of Science in Bioinformatics at the Indiana University-Purdue University in Indianapolis; this forms a good example of a typical Bioinformatics graduate program.

One could argue that the novelty of methods in Data Science is due to its roots in Business Analytics (BA), where the objective is to make a decision. The field has certainly grown beyond that, and BA is now considered a sub-field of Data Science. Another more-recently developed sub-field is in the realm of health care ("Health Informatics"). Because these sub-fields are directly applicable to the private sector (and the associated revenue streams these present) these have become the most-commonly implemented post-graduate programs. The Business Analytics programs focus on using data to refine business administration, as well as develop marketing strategies. Health Informatics programs concentrate on using clinical data to optimize health care processes.

The practical focus of Data Science is reflected in the presence of an internship in the Data Science curriculum listed in Table 5.

Course Name	Type
Analysis of Algorithms	required
Machine Learning	required
Advanced Database Concepts	required
Distributed Systems	elective
Advanced Database Concepts	elective
Cloud Computing	elective
Information Retrieval	elective
Data Mining	elective
Web Mining	elective
Applied Machine Learning	elective
Complex Networks and Their Applications	elective
Relational Probabilistic Models	elective
Internship in Data Science	elective

Table 5: A selection of the courses available for the Master's of Data Science at the Indiana University.

Internships in such programs are similar to other co-op-type arrangements: the student works with an employer for a semester, allowing the student to gain hands-on experience applying the skills learnt during such period.

4.1 Academic Data Science Programs

Graduate level programs in Data Science are not difficult to find. For instance, programs in bioinformatics (a data-driven field), can be found on the web site of the International Society for Computational Biology. It speaks to the rapid rise of the field bioinformatics, that there are more bioinformatics programs available than Scientific Computing programs. Examples lists of other Data Science programs can be found at the NCSU analytics web site, the online business analytics programs site of `predictiveanalyticstoday.com` and at `online.coursereport.com`. Initially these programs were not as common as Scientific Computing, due to the fact that Data Science was a relatively new field of study. This situation has changed dramatically in the last few years. For instance in the United States alone there are hundred of Data Science degrees and certificates [8]. Among those programs about half are offered in the fields of Business Analytics and Health Informatics, with the other half being proper Data Science programs. There has also been a surge in Machine Learning and Artificial Intelligence programs, which span data science and scientific computing.

4.2 Non-academic Data Science training

The demand for Data Science skills (or "Data Analytics" skills as they are often called in the private sector) is so high [19] that the private sector has developed programs to meet the growing demand. See, for instance, on skilledup.com, which contains a list of data science boot-camps. The format of these classes is varied, though they are all oriented toward a "boot-camp" format: some are in person, some online; some are one-week long, others twelve weeks. These programs are very applied, often with one-on-one mentorship with a seasoned Data Analytics expert. They also include direct contact with possible future employers.

Moreover, a great number of these training programs are not focused on developing analytical thinking or problem-solving skills, [12] but rather are aimed at Ph.D.s and postdocs, whose problem-solving skills are assumed to have already developed. This allows them to focus on the technical training relevant to the job market. Some of these programs are free, some of them offer fellowships, and many of them charge on the order of 10-30 thousand US-dollars for a training period of, typically, three months. These programs have acquired such a level of popularity among young and recent graduates that the companies offering these programs have started to perform evaluation tests in order to assess which candidates are more suitable to be accepted to their programs. Perhaps the most appealing part for trainees is the networking platform offered by these programs, as in most cases they provide the opportunity to interact with actual companies looking for new talent and avoid recruitment layers.

Institutions in the non-profit arena are also starting to offer programs on Data Science. For instance the Fields Institute, a traditional institution for mathematical research, has offered several workshops and courses, and developed a thematic program on Big Data. Other examples include the International Centre for Theoretical Physics (ICTP) and the International School for Advanced Studies (SISSA), prestigious institutions with a well known tradition in theoretical physics, now offering training in "Research Data Science".

HPC centers are also venturing into Data Science training, offering workshops on R, Hadoop, machine learning, *etc.* SciNet started offering classes with greater data-oriented content (*cf.* Table 3) in 2013, with a four-week class in scientific analysis using Python. Having now finished its third year, the class remains popular, with about twenty students taking the class each year. The 2015 fall semester also inaugurated SciNet's first "Data Science with R" class, focusing on data analysis techniques using the R language. This class was very popular with over twenty-five students finishing the course, and most students requesting a second installment with more advanced material. Continuing its growth in the Data Science area, in the last year SciNet has held workshops in machine learning, scalable data analysis, and Apache Spark.

Comparing the student- and taught-hours per year shown in Fig. 2, one sees that the Data Science classes have been growing consistently, both in absolute as well as relative numbers (Data Science related courses roughly constituted less than 2% (2012), 4% (2013), 21% (2014), 22% (2015), 27% (2016) and 28% (2017) of the total classes taught respectively in each year. While attendance to Data Science courses, follows even a more significant trend: 1% (2012), 23% (2013 and 2014), 29% (2015), 40% (2016) and 41% (2017). We project for the current year a growth of at least 10%, positioning at equal levels the traditional Scientific Computing and Data Science trainings.

5 DISCUSSION

As mentioned above, scientific computing is used by scientists and engineers as never before, and graduate-level programs in Scientific Computing are numerous in Canada and around the world. In contrast, the development of HPC and Data Science programs is in its early stages, both in academia and the private sector. These programs are being developed to meet the continued shortfall in skill in these areas, with the McKinsey Global Institute estimating that the United States will be short 140,000 to 190,000 data analytics professionals by 2018 [16].

One may wonder whether online learning could not satisfy this need. A few examples of MOOCs (Massively Open Online Courses) in HPC and Data Science do exist[5]. However, seeing the growth in enrolment in SciNet's in-person courses and the summer school over the years (*cf.* Figs. 1, 2 and 3) shows that many students still prefer the face-to-face format.

Similarly, one may wonder why certificate programs do not suffice for HPC and DS education. As successful as these programs are, they have a few disadvantages. Firstly, they are mostly collections of fairly specific technical training: this leaves no room for more fundamental material. Secondly, it is also hard to encorporate an internship or thesis into such a certificate. Finally, certificates tend to carry less weight than degrees, and, in line with this, the demand for for-credit courses is larger than that for not-for-credit courses, as

[5] E.g. https://www.citutor.org/, http://www.hpc-training.org, https://www.futurelearn.com.

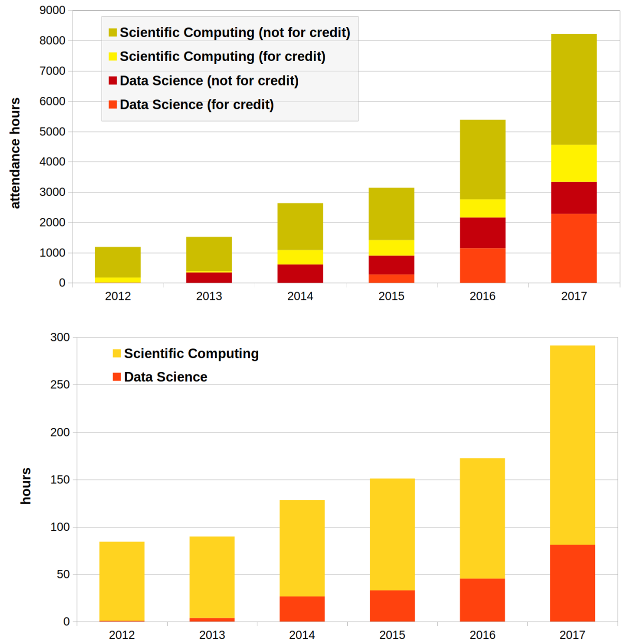

Figure 2: Total student hours (top) and taught hours (bottom) per year, for SciNet's Data Science and Scientific Computing related courses. The trends clearly show the need for training courses in both fronts, at the level of university recognized courses.

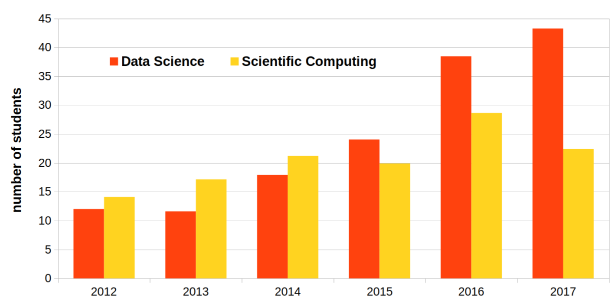

Figure 3: Average class size trend for Scientific Computing and Data Science courses at SciNet.

our experience with SciNet's Scientific Computing graduate course has shown.

A degree program in HPC or DS could offer more academic and fundamental education, which would leave the student with the analytical skills and high-level knowledge to stay on top of their field regardless of changes in computational technology.

6 CONCLUSION

We believe we have offered quantitative evidence demonstrating the need for programs in higher education in High-Performance Computing and Data Science, in particular in our own institution. If the qualitative evidence of this seems somewhat limited, it should be understood that existing HPC and DS programs (academic and non-academic) are still relatively new. While some such programs are already in existence, in many cases students must use non-academic options, or teach the material to themselves. Academic programs would offer the benefit of not just teaching specific technical skills, but an education in the fundamentals of HPC and DS and instilling the analytical skills needed to adopt to an ever-changing technological landscape.

We have reviewed existing academic and non-academic education programs, in both HPC and DS. In light of this review, we invite eager readers to look at a longer and complementary version of this manuscript [17] where we present the design for a tentative Master's programs in HPC and DS, based on the examples discussed here and drawing from the experience and enrollment statistics in not-for-credit training in HPC and DS by the SciNet HPC Consortium at the University of Toronto.

Getting well-founded graduate programs off the ground will not be without challenges. It will likely involve partnerships and discussions with other departments and institutes in order to offer a stronger and multi-disciplinary program. Existing HPC Centers, which already operate across multiple disciplines, can play a fundamental role in bringing together such programs. Thus, we have described SciNet's path in developing and transgressing the usual role of training events for users, into full credited graduate courses recognized at the university level for masters and doctorate degrees. Moreover, the creation of these graduate courses allowed us to leverage two key elements:

(1) Several of our analysts involved in the educational efforts got recognized by obtaining graduate teaching affiliations with the corresponding institutes/departments that were sponsoring the courses.
(2) Interact with students at a more professional level, collaborating and participating in research projects, allowing us to launch research initiatives [22] for which we already have successful cases [10, 21] and many more in the process of being developed. Perhaps the most important point here is to note that these collaborations were catalysed by the direct interaction with our students and researchers.

We hope this might help other super-computer centers transition a similar path and develop strategies to play a fundamental role in the multidisciplinary fronts that are HPC, SC and DS.

REFERENCES

[1] Centre for Advanced Computing (CAC) at Queen's University, Kingston, Ontario, Canada. https://cac.queens.ca/. (????). Accessed: 2016-06-14.
[2] 2001. Shared Hierarchical Academic Research Computing Network (SHARCNET). https://www.sharcnet.ca. (2001). Accessed: 2016-04-08.
[3] 2010. International HPC Summer School. http://www.ihpcss.org/. (2010). Accessed: 2016-04-08.
[4] 2014. Compute Ontario. http://computeontario.ca/. (2014). Accessed: 2016-04-08.
[5] 2016. Scientific Computing for Physicists (PHY1610H). http://www.physics.utoronto.ca/students/graduate-courses/current/phy1610h-f. (2016). Accessed: 2016-04-06.
[6] 2017. Canada's Fundamental Science Review. http://www.sciencereview.ca. (2017). Accessed: 2018-06-02.
[7] 2017. Introduction to Computational Biostatistics with R (MSC1090H). https://ims.utoronto.ca/wp-content/uploads/2014/04/IMS-syllabus-2017-REV-MSC1090H.pdf. (2017). Accessed: 2018-04-11.
[8] 2018. Masters's in Data Science. http://www.mastersindatasience.org/. (2018). Accessed: 2018-06-02.

[9] Computational Science Research Center at San Diego State University. 2011. Master's programs. http://www.csrc.sdsu.edu/masters.html. (2011). Accessed: 2016-04-06.

[10] Philippe Berger and George Stein. 2018. A volumetric deep Convolutional Neural Network for simulation of dark matter halo catalogues. (2018). arXiv:astro-ph.CO/1805.04537

[11] Helmar Burkhart, Danilo Guerrera, and Antonio Maffia. 2014. Trusted High-performance Computing in the Classroom. In *Proceedings of the Workshop on Education for High-Performance Computing (EduHPC '14)*. IEEE Press, Piscataway, NJ, USA, 27–33. https://doi.org/10.1109/EduHPC.2014.13

[12] Ed Lazowska. November 2015. Dear GeekWire: A coding bootcamp is not a replacement for a computer science degree. http://www.geekwire.com/2015/dear-geekwire-a-coding-bootcamp-is-not-a-replacement-for-a-computer-science-degree. (November 2015). Accessed: 2016-04-06.

[13] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H. Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, Miguel Ángel Rubio, Judy Sheard, Bronius Skupas, Jaime Spacco, Claudia Szabo, and Daniel Toll. 2015. Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports (ITICSE-WGR '15)*. ACM, New York, NY, USA, 41–63. https://doi.org/10.1145/2858796.2858798

[14] Scott Lathrop, Ange Mason, Steven I. Gordon, and Marcio Faerman. 2013. HPC University: Getting Information About Computational Science Professional and Educational Resources and Opportunities for Engagement. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery (XSEDE '13)*. ACM, New York, NY, USA, Article 67, 4 pages. https://doi.org/10.1145/2484762.2484771

[15] Chris Loken, Daniel Gruner, Leslie Groer, Richard Peltier, Neil Bunn, Michael Craig, Teresa Henriques, Jillian Dempsey, Ching-Hsing Yu, Joseph Chen, L Jonathan Dursi, Jason Chong, Scott Northrup, Jaime Pinto, Neil Knecht, and Ramses Van Zon. 2010. SciNet: Lessons Learned from Building a Power-efficient Top-20 System and Data Centre. *Journal of Physics: Conference Series* 256, 1 (2010), 012026. http://stacks.iop.org/1742-6596/256/i=1/a=012026

[16] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. *Big data: The next frontier for innovation, competition, and productivity*. Technical Report. McKinsey Global Institute.

[17] Marcelo Ponce, Erik Spence, Daniel Gruner, and Ramses van Zon. 2016. Designing Advanced Research Computing Academic Programs. *CoRR* abs/1604.05676 (2016). arXiv:1604.05676v2 http://arxiv.org/abs/1604.05676v2

[18] R Core Team. 2017. R: A Language and Environment for Statistical Computing. (2017). https://www.R-project.org/

[19] Alex Radermacher and Gursimran Walia. 2013. Gaps Between Industry Expectations and the Abilities of Graduates. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education (SIGCSE '13)*. ACM, New York, NY, USA, 525–530. https://doi.org/10.1145/2445196.2445351

[20] Mark Richards and Scott Lathrop. 2011. A Training Roadmap for New HPC Users. In *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery (TG '11)*. ACM, New York, NY, USA, Article 56, 7 pages. https://doi.org/10.1145/2016741.2016801

[21] Alejandro Saettone, Jyoti Garg, Jean-Philippe Lambert, Syed Nabeel-Shah, Marcelo Ponce, Alyson Burtch, Cristina Thuppu Mudalige, Anne-Claude Gingras, Ronald E. Pearlman, and Jeffrey Fillingham. 2018. The bromodomain-containing protein Ibd1 links multiple chromatin-related protein complexes to highly expressed genes in Tetrahymena thermophila. *Epigenetics & Chromatin* 11, 1 (09 Mar 2018), 10. https://doi.org/10.1186/s13072-018-0180-6

[22] SciNet HPC Consortium. Research SciNet. https://www.scinethpc.ca/research-scinet/. (????). Accessed: 2018-04-13.

[23] SciNet HPC Consortium. SciNet's Certificate Program. http://www.scinethpc.ca/scinet-certificate-program. (????). Accessed: 2016-04-06.

[24] SciNet HPC Consortium. SciNet's Education website. https://support.scinet.utoronto.ca/education. (????). Accessed: 2016-04-06.

[25] SciNet HPC Consortium. SciNet's Training, Outreach and Education. http://www.scinethpc.ca/training-outreach-and-education. (????). Accessed: 2016-04-06.

[26] SciNet HPC Consortium. SciNet's website. http://www.scinet.utoronto.ca. (????). Accessed: 2016-04-06.

[27] Kris Stewart. 1995. HPC Undergraduate Curriculum Development at SDSU Using SDSC Resources. In *Proceedings of the 1995 ACM/IEEE Conference on Supercomputing (Supercomputing '95)*. ACM, New York, NY, USA, Article 19. https://doi.org/10.1145/224170.224209

[28] Guy Tel-Zur. 2014. PDC Education in the BGU ECE Department. In *Proceedings of the Workshop on Education for High-Performance Computing (EduHPC '14)*. IEEE Press, Piscataway, NJ, USA, 15–20. https://doi.org/10.1109/EduHPC.2014.9

[29] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering* 16, 5 (Sept 2014), 62–74. https://doi.org/10.1109/MCSE.2014.80