

Student Training in the Era of Big Data Physics Research

Amit Chakrabarti, William and Joan Porter Professor and Head, Department of Physics, Kansas State University

This article summarizes how the availability of Big Data is changing the research landscape in Physics. Some thoughts on student training in this new era are also presented here. Although these topics are discussed in the context of the physics department at Kansas State University, the implications go beyond the borders of one physics department or one University.

Introduction

Availability of Big Data is having a major impact on research and student training in all sub disciplines of physics. These topics are discussed in this paper in the context of the physics department at Kansas State (K-State) University. Thus, some background on this department will be useful to the readers. This is a research intensive department in a land grant university. With only 27 permanent faculty members, the department receives competitive external funding of over \$7 million each year which places K-State at the top of the Regents-designated peer institutions. The department routinely ranks in the top 50 universities in the United States in National Science Foundation listings of external funding for research [1]. In the recently released National Research Council data [2], K-State physics faculty is placed above the median in research productivity measures, including publications per faculty member and percentage of faculty members with external support. K-State Physics faculty conducts internationally recognized research in atomic, molecular and optical physics, soft and biological

matter physics, high-energy physics, cosmology, and physics education.

Next, a brief summary is presented on how Big Data is shaping physics research in some of the sub-disciplines where Kansas State physics department has strong record of achievements.

High Energy Physics and Cosmology

High Energy Physics and Cosmology are at the forefront of Big Data Physics [3, 4]. Glenn Horton-Smith, the current director of our High-Energy Physics program has presented some details of Big Data projects in High Energy Physics and Cosmology at Kansas State University. As Glenn has pointed out, there are three major aspects of High-energy physics and Cosmology research today:

- 1) **The Energy Frontier:** This includes the Large Hadron Collider at CERN with its goals of discovering new particles (e.g., Higgs boson) and new fundamental phenomena;
- 2) **The Intensity Frontier:** This includes multiple neutrino experiments with its goals of understanding the nature of mass and matter/antimatter asymmetries;

3) **Cosmology:** This includes Dark Energy modeling and testing of alternative theories based on large amount of data from various Astronomical surveys.

Atomic, Molecular, and Optical Physics (AMO)

The K-State Atomic, Molecular, and Optical Physics (AMO) program housed in the nationally renowned James R. Macdonald (JRM) Laboratory [5], is one of the largest in the country and ranked 13th nationally by the latest U.S. News and World Report [6]. JRM lab has recently added several ultrafast, intense laser facilities. As a result, the direction of the program has shifted toward intense-laser-matter interactions. Besides doing their research at the JRM lab, some of our AMO faculty members use X-Ray Lasers (Linac Coherent Light Source or LCLS) in the Stanford National Accelerator Laboratory [7]. LCLS stores data at a rate and scale comparable to experiments at the Large Hadron Collider in CERN. The LCLS data team manages about 10 petabytes (1 petabyte = 1000^5 bytes) of data from experiments [8]. This is already an amazing number – about three times more than the total data library for movie-streaming and rental company Netflix. An experiment at one LCLS instrument produces about 10 million X-ray images in about 48 hours, on average, with the largest experiments generating 150-200 terabytes (about 154,000 to 205,000 gigabytes) of data. However, this is just the beginning in data production and data management. Amedeo Perazzo, who leads the Photon Controls and Data Systems Department at SLAC said that "Pretty soon we will be taking a factor of

20 more data than we are taking today" [8].

Soft Matter and Biological Physics

The Soft Matter and Biological physics group at K-State has traditionally depended on large amounts of small data [9]. That, however, is changing with the more recent focus on nanomaterials research and computations in biological physics [10-11]. For a detailed theoretical understanding of such complex systems one often needs to resort to hierarchical or multi-scale computing. One usually starts from atomistic studies of relatively small systems and then uses the results as input for tackling larger "coarse-grained" systems. This way, one can make headways toward understanding and predicting macroscopic materials properties of interest. For a quick introduction to what sort of large-scale computing is cutting edge these days, here is a list of few topics from the last couple of years' Gordon Bell prize [12] (given for outstanding achievement in high-performance computing applications) winners:

- 2012 --- 4.45 Pflops Astrophysical N-Body Simulation on K computer - The Gravitational Trillion-Body Problem
- 2011 --- First-principles calculations of electron states of a silicon nanowire with 100,000 atoms on the K computer
- 2011 --- Peta-scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer

How do we train undergraduate physics majors and graduate students in this era of Big Data physics research? Here are some specific thoughts on that.

Student Training:

Physics as both a fundamental and foundational science: All physics students must be encouraged to view physics as both a fundamental and foundational science that provides an effective background for a diversity of career paths. Many of the problems that will need to be solved in the coming decades will occur on the interface between physics and related areas. These include understanding and controlling new forms of energy, developing new materials for the next generation of computers and improving methods of medical imaging.

Introduction to Theoretical Models:

Of foremost importance is to train students in the physical models that have been so successful in explaining Nature. This is essential to provide the students with Big Data interpretation skill. This training cannot however be achieved by a traditional lecture format. Early involvement in research is a must. Research experience lets students put to use theories they learn in class and acquaint themselves with the faculty, post-docs and other students. These experiences help students make good career decisions. Involvement in research is also fun.

Specialized Computation Skills:

Another essential component of student training in this new era is the introduction of specialized computational skills early in their career. On one hand, this will teach them to apply tailor-made computational algorithms based on understanding the specific physics of the problem at hand. On the other hand, introduction to Open Source and Visual programming skills will help them with their career decisions after they graduate.

Communication skills: Physicists are trained to formulate their understanding of a problem or phenomenon in precise terms and to communicate these ideas to others. Training in both oral and written technical communication skills and the ability to translate from Techie language to English will be critical for success in a wide variety of situations.

Entrepreneurial skills: The first step in this direction will be to increase commercialization efforts in the area of K-State strength. Once new opportunities for the physics faculty are identified, their research programs can be broadened by systematically engaging companies in the research work. This will bring industrial support to research and create a culture of solving practical problems. Such experience in “producing products” will have a profound impact on professors and students equally. K-State will be a powerful economic driver for growth and development by generating new knowledge and producing graduates who will impact Kansas, the nation and the world.

Assessment of student achievements in the Big Data era

Finally, a brief discussion of assessment of student achievements in the Big Data era is warranted. The major problem in assessment is that Faculty and departments are not familiar in interpreting research-based assessments, or comparing their students to a national group of similar students. To achieve these goals, one would want to analyze large datasets of nationally representative data, but these datasets do not exist. Such a work, however, is in progress and K-State Physics Education Research Group [13] is in the forefront of creating such a database

[14] with support from the American Association of Physics Teachers and the National Science Foundation. Once the database is created, faculty will be able to visualize and compare their students' performance to huge national database of results from 50+ research-based assessment instruments.

Conclusions

In conclusion, Big Data is changing the physics research landscape big time. Curriculum development and student training must be undertaken in view of these recent developments. Topics on student training and Big Data Physics projects discussed here are in the context of the physics department at Kansas State University. Their implications, however, go beyond the borders of one physics department or one University.

References

1. <http://www.nsf.gov/statistics/nsf13325/pdf/tab26.pdf>
2. <http://www.nap.edu/rdp/>
3. <http://home.web.cern.ch/>
4. http://sites.nationalacademies.org/bpa/BPA_049810
5. <http://jrm.phys.ksu.edu/>
6. <http://grad-schools.usnews.rankings-andreviews.com/best-graduate-schools/top-science-schools/atomic-science-rankings>
7. <http://www.slac.stanford.edu/>
8. <http://www6.slac.stanford.edu/news/2013-06-13-X-ray-Laser-Explores-Big-Data-Frontier.aspx>
9. See for example, <http://www.pvos.org/?p=9>
10. <http://on-demand.gputechconf.com/gtc-express/2012/presentations/scaling-soft-matter-physics-to-a-thousand-gpus-and-beyond.pdf>
11. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5147
12. <http://awards.acm.org/bell/year.cfm>
13. <http://perg.phys.ksu.edu/>
14. <http://perusersguide.org/>