Aligning Nuclear Physics Computing Techniques with Non-Research Physics Careers

Wouter Deconinck

Supported by the National Science Foundation under Grant Nos.
PHY-1156997, PHY-1206053, PHY-1359364, PHY-1405857, DUE-1625872.

Future Trends in Nuclear Physics Computing — May 2017

William & Mary
CHARTERED 1693

Introduction: Who Am I? Why Do I Care?

Who am I?

- Experimental nuclear physicist (parity-violating electron scattering)
- Faculty at research-intensive PhD-granting liberal arts university which requires undergraduate research of all undergraduate seniors

Why do I care?

- Frustration with duplication of efforts in software development
- Worked with many graduate and undergraduate students
  - lack of sufficient background for software development
  - often frustration with ramp-up and delay before first results
- Explored agile development and alternative approaches to analysis

Conclusion

Centering the experiences of students, who are the main users of nuclear physics data analysis software, naturally drives us towards considering current general industry-standard data tools in an increasingly modular data analysis environment with a low barrier of entry to first results and algorithm modifications.

User-Centered Design: Who Are Software Users? I

Nuclear physics analysis software users

- Long-term researchers
  - faculty, staff scientists
- Short-term researchers
  - undergraduate, graduate researchers, post-doctoral researchers

User-Centered Design: Who Are Software Users? II

Software use by short- vs. long-term researchers

- Long-term researchers (in particular faculty) are largely disconnected from direct low-level software tasks, instead looking at physics outputs
  - common complaint: students don't see the physics, only the code!
  - suggested solution: maybe they shouldn't be writing as much code then?
- Short-term researchers are trained or in training, and not (typically) proficient in software development until later in their student career (learning by doing, but without training wheels)
- Yet, at least 50% of their time is spent on writing code instead of thinking about physics (e.g. graduate students at LHC)
- Nuclear physics data analysis has gotten a reputation of being coding intensive at the expense of generating physics insights

User-Centered Design: Who Are Software Users? III

User-centered design of nuclear physics analysis software

- The users (and often developers) of nuclear physics analysis software are predominantly the short-term researchers, i.e. graduate students and postdoctoral researchers
- User-centered design means we should align software development with their goals

What Do Physics PhDs Do After Graduation?

Job titles

- Postdoctoral researcher
  - Postdoctoral researcher
    - Postdoctoral researcher
- Faculty position
- Industry position
  - Data scientist, data analyst

What Do Physics PhDs Do After Graduation?

Job titles

- Postdoctoral researcher
  - Postdoctoral researcher
    - Postdoctoral researcher
- Faculty position
- Industry position
  - Data scientist, data analyst

The myth of the traditional physicist

- A physicist is NOT someone like you or me, i.e. a researcher
- NOT someone following the linear career path from undergrad to grad school to postdoc to faculty or staff position
- Deviations are referred to as “leaving physics”, value judgments implied...

Physicists Find Careers Primarily Outside Academic Research

Bachelor degrees in physics

- Only 1 out of 6 physicists continues to PhD degree (AIP SRC)
- All other physicists not included in “traditional physicists” interpretation

PhD degrees in physics

- Majority of the permanent jobs is outside of academic research
- About 1700 physics PhDs per year, but significantly fewer jobs
- All other physicists not included in “traditional physicists” interpretation

Mismatch between curriculum and reality of physics teaching

- How can we prepare our undergraduate and graduate students better for their most likely career?
- How can we align the training of undergraduate and graduate students to the skills they will benefit from?

Physicists Find Careers Primarily Outside Academic Research

Type of Employment of Physics PhDs by Employment Sector One Year After Degree, Classes of 2013 & 2014 Combined

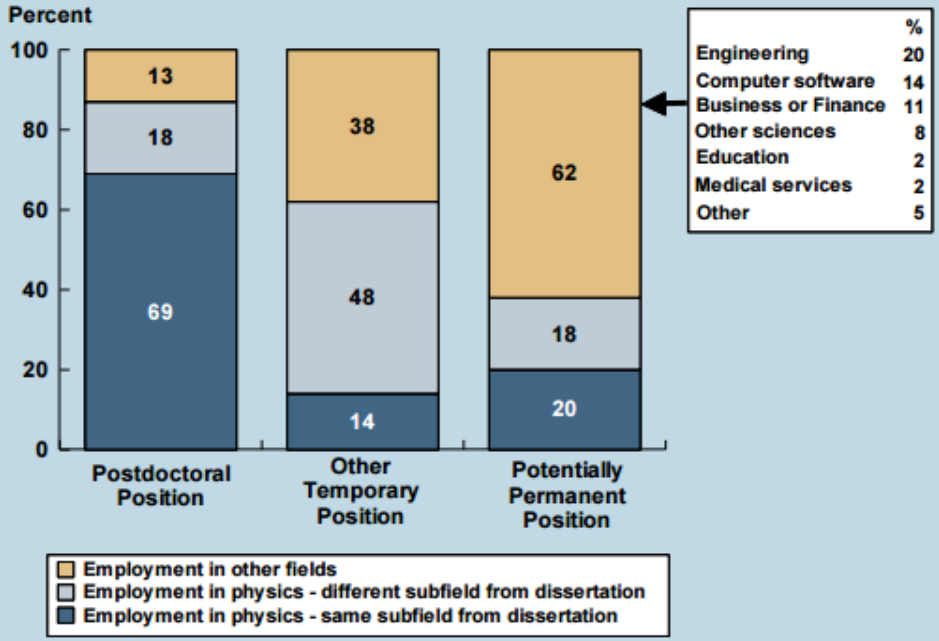
Sector of Employment	Initial Employment Type: Postdoc %	Potentially Permanent %	Other Temporary %	Overall %
Academic*	75	20	71	52
Private	1	70	18	31
Government	21	8	3	14
Other	3	2	8	3
	100%	100%	100%	100%

Note: Data only include US-educated physics PhDs who remained in the US after earning their degrees. Data are based on the responses of 655 postdocs, 523 individuals working in potentially permanent positions and 126 individuals working in "other temporary positions."

*The academic sector includes two- and four-year colleges, universities, and university affiliated research institutes.

http://www.aip.org/statistics

Employment Field of Physics PhDs One Year After Degree, Classes of 2013 & 2014 Combined



Note: Employment in physics means an individual's primary or secondary employment field was in physics or astronomy. Data only include US-educated PhDs who remained in the US after earning their degrees. Data are based on the responses of 419 postdocs, 297 individuals working in potentially permanent positions and 87 individuals working in "other temporary positions".

http://www.aip.org/statistics

Physicists Find Careers Primarily Outside Academic Research

What skills do physicists have?

- Comfortable with computer modeling and simulation
- Fluent with math
- Strong problem solvers
- Experience with coding
- Technical software skills

Physicists Find Careers Primarily Outside Academic Research

What skills are physicists missing?[1]

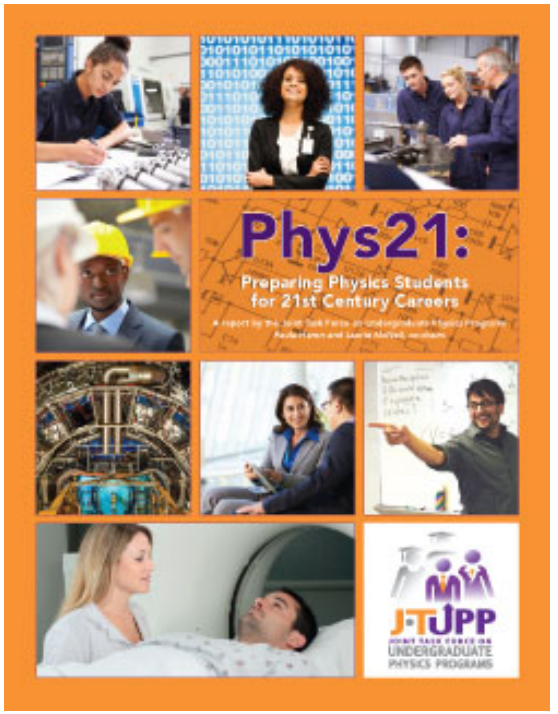
- Ability to design a system, component or process to meet a specific need
- Ability to function on multi-disciplinary teams
- Ability to recognize value of diverse relationships (customers, supervisors, etc)
- Leadership skills
- Familiarity with basic business concepts (i.e. cost-benefit analysis, funding sources, IP, project management)
- Communication skills (oral and written), esp. how to tailor message to audience
- Real-world experience in companies before graduation
- Awareness of career paths outside of academia

[1] *Sources: ABET Survey of Applied and Engineering Physics Graduates, Kettering University; APS Workshop on National Issues in Industrial Physics, Industrial Physics Lunches.*

Joint Task Force on Undergraduate Physics Programs

Findings

- "The overwhelming majority of physics bachelor's recipients are employed outside academia for all or part of their careers."
- "Since only about one-third of physics Ph.D. recipients end up in academic careers, even students who plan to obtain graduate degrees will benefit from developing skills and knowledge that are valued outside the academic community."



Promote career readiness: Scientific and technical skills

- "Competencies in instrumentation, software, coding, and data analytics."
- "Introduce students to industry-standard tools and software packages."

Physicists Are Not The Preferred Data Scientists Anymore... I

Over the previous two decades...

- Rapid expansion of data collection capabilities outside of research enterprise, surpassing what high energy and nuclear physics have routinely dealt with for years
- This expansion and presence of physicists with large-scale data analytics training resulted in employment opportunities for many physicists

Physicists Are Not The Preferred Data Scientists Anymore... II

But the educational environment has adapted rapidly

- Nearly every college or university has now implemented a data science concentration, minor, or major, with relevant content based on requirements in industry: relevant algorithms, languages and platforms
- Nuclear and high energy physics have remained largely the same: no major changes in languages, no major adoption of new methodologies (notable exception: GPUs)
- Physicists have lost their edge in the large-scale data analytics job market

Intermediate Conclusion

Physics undergraduate and graduate students participating in nuclear physics data analysis research are likely to find permanent employment outside of academic research. If we are to continue attracting outstanding students, we should provide students with the scientific and technical skills for future careers inside and outside academia, in particular in data science.

Three Ways Of Adapting To This New Reality

Allowing students to get up and running quickly through modularity

Easy access to data with the flexibility to modify independent parts of the data analysis chain will allow students to develop the relevant data analytics skills and portfolio, rather than getting stuck in a morass of sloppy code

Using industry-standard data analytics tools

Experience with tools and languages that are recognized as important will help students get a foot in the door on the job market, while externalizing the development of software frameworks that do not require data analysis skills

Increased focus on data analytics results

By reducing the amount of time coding, students can increase the amount of time spent applying physics knowledge to their data analysis

Monolithic Nuclear Physics Frameworks Limit Flexibility I

"One size fits all" software solutions are overwhelming

- From raw data to final histograms in one process (if multiple threads)
- Advantages: centralized framework allows for code reuse
- Disadvantages: modifications have steep learning curve, no intermediate branch points to start partial analysis

Examples of monolithic "modular" frameworks

- At Jefferson Lab: Podd, Jana, Clara,...
- Exquisitely designed class hierarchies, but limited to single language and workflow
- Modularity by implementing different classes in a strict syntactically enforced environment

Monolithic Nuclear Physics Frameworks Limit Flexibility II

Towards independent processes with minimal operations

- Formatted input and output streams (JSON, XML)
- Each process is responsible for one logical task (decoding, correcting, fitting, calibrating, regression,...), modifications can be more easily overseen
- Scalability in number of cores and across connected sites
- Compartmentalization of functionality
- Full container approach (e.g. Docker)

Importance is in data types over algorithms

- With clearly defined interfaces between individual operations, individual algorithms can be swapped out easily

Can We Get Back To The Unix philosophy?

Quote by Doug McIlroy, inventor of pipes (the Unix kind)

- Write programs that do one thing and do it well.
- Write programs to work together.
- Write programs to handle text streams, because that is a universal interface.

Also by Doug McIlroy

- Design and build software, even operating systems, to be tried early, ideally within weeks.
- Don't hesitate to throw away the clumsy parts and rebuild them.

When did we move away from these principles?

Using Industry-Standard Tools Benefits Research Too

Reuse of existing frameworks

- Take advantage of available architectural frameworks reduces development time and overhead
- Examples: Apache Hadoop and Spark for map-reduce-like functionality around atomic operations (arguably performs many of the same functions that Jana and Clara perform)

Application of industry-standard analytics and visualization

- Ability to use R and Python for analysis or prototyping at any stage, for generation of figures
- Use of business analytics software, such as interactive visualizations in Tableau

Early Focus On Speed Ignores Life Cycle Management

Interpreted versus compiled languages

- Many physicists dismiss interpreted languages such as Java or Python out of hand based on (at best) outdated assumptions or anecdotes
- This presents a higher barrier of entry for new users, and a mismatch with currently taught programming languages (primarily python)

Use of data storage formats with poor interchangeability

- Many options: ROOT, SQL, csv, hdf5,...
- This often prevents new students from getting up and running with data analysis quickly

Increased focus on user-centered design

- Increase productivity of short-term researchers

Let Physicists Think Primarily About Physics Again

Conclusion

Centering the experiences of students, who are the main users of nuclear physics data analysis software, naturally drives us towards considering current general industry-standard data tools in an increasingly modular data analysis environment with a low barrier of entry to first results and algorithm modifications.