WINTER 2023 VOL. 2

APS Topical Group on Data Science GDS

DATA SCIENCE

APS Topical Group on Data Science Newsletter

Letter from the Newsletter Editor

Dear GDS community,

I'm slowly but surely finding my footing for this newsletter. I have enjoyed getting to know the GDS community and have been impressed with how welcoming they have been. My goals are to cultivate a newsletter that includes a wide range of perspectives and to have articles that will pique the interests of the GDS community and beyond. This newsletter features some exciting pieces that go into important aspects of data science, the future, and all sorts of career things. I'm thankful and appreciative that so many people are willing to contribute these articles- without them, this newsletter just wouldn't exist.

There are also some announcements related to the activities GDS engages in. The DSECOP project has its save the date, and there are terrific events happening at the March and April Meetings. I encourage you to check these out.

I hope you enjoy the newsletter, and consider sharing it with others. A friend mentioned that they shared the last newsletter with some students, who were excited to learn about what is going on in the data science and physics world.

If you have any ideas regarding potential articles, do feel free to reach out.

Best wishes,

Alexis V. Knaub
Editor

IN THIS ISSUE

HOW TO CONNECT

There are a variety of ways that you can stay up to date, and informed about GDS. Follow GDS on social media if you haven't done so already!

Email: gds@aps.org
Website: aps.org/units/gds/index.cfm
LinkedIn: linkedin.com/company/apsdatascience
Twitter: twitter.com/apsdatascience
Facebook: facebook.com/APSDataScience

Message from the Chair

Dear GDS members,

It has been an honor to serve as Chair of the GDS and contribute to the vibrant community of data scientists.

Data science is experiencing rapid growth and making a significant impact in various fields. From the successful application of machine learning algorithms in detecting gravitational waves to the analysis of particle colliders and simulation of complex systems, the role of data science in physics research is becoming increasingly apparent. This is reflected in the growing number of physics research initiatives worldwide that utilize data science techniques.

It's also important to highlight the interdisciplinary nature of data science and its connection to the industry. Data science has the potential to bridge gaps between different fields and drive innovation, as evidenced by the successful collaborations between industry and academia that resulted in cutting-edge advancements in fields such as energy and healthcare.

I would like to take this opportunity to highlight some of the GDS's achievements, including the growth of our community and the establishment of the new GDS IMPACT Award. The award recognizes and celebrates students who have made a significant impact in the field of data science, and the first recipients will be honored at the 2023 APS meetings. Additionally, the GDS will host exciting sessions at the March and April Meetings 2023, providing a platform for researchers to showcase their work and engage with their peers.

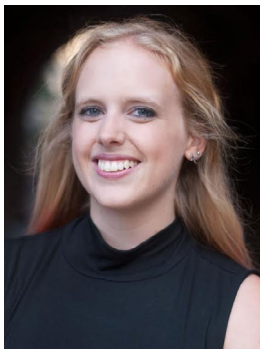
I am grateful for the unwavering support of the Executive Committee during my term as Chair. Their commitment to the GDS and the advancement of data science has been an inspiration, and it has been a privilege to work alongside them. As my term comes to an end in March 2023, I would like to extend my best wishes to the incoming Chair, William Ratcliff, who will be taking the helm. I have confidence in William's passion and commitment to leading the GDS.

Finally, I would like to express my gratitude to all GDS members for their continued support and participation. I am eager to see the future of data science in physics continue to grow and evolve and for our GDS community to thrive!

Best regards,

Maria Longobardi
GDS Chair

Ethics, Machine Learning, and Physics: An Interview with Savannah Thais



SAVANNAH THAIS

By Sanha Cheong

Sanha Cheong is a physics PhD candidate at Stanford University, working on the ATLAS and the MAGIS Experiments. Outside of research, he is passionate about statistics and machine learning education. Savannah Thais is a Research Scientist at the Columbia University Data Science Institute where she focuses on machine learning (ML). She is interested in complex system modeling and in understanding what types of information is measurable or modelable and what impacts designing and performing measurements have on systems and societies. She is passionate about the impacts of science and technology on society and is a strong advocate for improving access to scientific education and literacy, community centered technology development, and equitable data practices. She was the ML Knowledge Convener for the CMS Experiment from 2020-2022, currently serves on the Executive Board of Women in Machine Learning and the Executive Committee of the APS Group on Data Science, and is a Founding Editor of the Springer AI Ethics journal. She received her PhD in Physics from Yale University in 2019 and was a postdoctoral researcher at Princeton University from 2019-2022.

Sanha Cheong: Can you tell us about your physics background? What brought you to a physics PhD in particle physics, especially?

Savannah Thais: I became interested in particle physics during high school. My AP physics teacher had us do an 'independent research project' and report on a topic in modern physics. As part of that assignment I read "Warped Passages" by Lisa Randall and ended up focusing my report on the (then still undiscovered!) Higgs Boson. From that point on, I was hooked; understanding the mathematical foundations of the universe and characterizing the smallest, most fundamental building blocks of nature seemed like the coolest thing anyone could possibly work on! I decided I wanted to pursue a physics degree and dreamed of working at CERN. From there, I went to the University of Chicago for college and pursued a double major in mathematics and physics, and it was there that I got my first research position working for Dr. Young-kee Kim, first on the MICE experiment and then for my senior thesis on the ATLAS experiment at the Large Hadron Collider (LHC). I went on to pursue my PhD in high-energy physics (HEP) at Yale, where I continued working on the ATLAS experiment with Dr. Sarah Demers. My research focused on developing software for identifying electrons in the detector and looking for an as-of-yet unmeasured production and decay channel of the Higgs Boson.

Cheong: What then motivated you to expand your research interest to artificial intelligence (AI), more broadly?

Thais: I started learning about and using machine learning (ML) methods during grad school; in particular, I worked on applying computer vision techniques to the electron identification task. During this time I started attending ML/AI conferences and became deeply interested in ML research more broadly. After finishing my PhD, I went on to a post-doc position at the Princeton Institute for Computational Science and Engineering with IRIS-HEP, a multi-institution research institute that focuses on software and computing in HEP. During my post-doc, I focused on geometric ML for particle trajectory reconstruction at the LHC. However, I also pursued side projects in computational social science/public health and AI ethics. I had been following these areas of research, and particularly many issues of bias, reliability, interpretability, and social impact that were plaguing the broader ML/AI community. I felt like this kind of work was the perfect way to combine my long standing interests in math, coding, model building, policy, community building, and social good. I started my dream position in the Columbia University Data Science Institute this fall where I conduct highly interdisciplinary research combining physics principles

and ML methods to study complex systems including cities, policy development, and public health ecosystems.

Cheong: You work a lot in AI ethics and societal applications of AI tech. Please give us an overview, especially for the more traditional physicist audience, of your current research in this direction.

Thais: I would say I have three main throughlines in my research. The first is focused on methods to better precisely understand the behavior, robustness, expressivity, and limitations of ML models. Put another way, I'm interested in research that moves us towards a more scientific theory of ML. This includes things like characterizing the space of possible performant models and identifying the optimal solution in that space, thinking about how we encode known information efficiently into ML models, and understanding what mathematical guarantees we can place on model performance. Right now, specifically, I'm looking at ways of requiring ML models to respect symmetry conservation that is present in different kinds of physical science data and understanding in which cases this leads to more efficient models.

My second area of research is focused on what I would refer to as computational social science. My work in this space applies techniques from model building in physics and ML to other types of systems. The goal of this work is to provide reliable mathematical insight into complex, real world situations. This work is highly interdisciplinary and deals with questions like what type of information is directly measurable, how we can build reliable models with incomplete or incorrect data, and how we conceptualize causality in models. My current project in this area is trying to precisely quantify the impact that sex ed. policies have on the outcomes of interest for Planned Parenthood (like teen pregnancy, STI rates, and inter-partner violence rates). This work would allow Planned Parenthood and other similar organizations to effectively lobby for more beneficial sex ed. policies.

Finally, I have some research projects that are less computational and look more holistically at the socio-technical ecosystem surrounding AI. These include questions around who benefits from and who is harmed by real-world AI systems, how we can effectively regulate emerging technology, when AI systems should actually be deployed, and more. My current work in this area is focused on the issue of AI hype and the real-world research and social consequences of the ways we talk about AI capabilities. I am also focused on developing public technical literacy, particularly amongst community groups like mutual aid networks in Brooklyn, and understanding how this knowledge can be used to shift power and democratize technology.

Cheong: What kind of research that is currently underexplored would you most like to see happen?

Thais: I would definitely like to see more work focused on precisely understanding the abilities and limitations of ML, both in a technical mathematical sense (for example, in what cases can we truly trust model predictions) and in broader societal sense (for example, in what cases do humans need or want an ML system to be deployed, or how can we understand the real impact these systems will have on humanity from a very holistic, humanistic perspective). There are absolutely some amazing researchers working in these areas, but I think unfortunately a lot of AI research right now is focused on artificial general intelligence (AGI) and a rush to build larger models that can quickly be profitably utilized by large corporations.

Cheong: How did your training in physics help with this transition/expansion to AI research with a focus on ethics and social sciences?

Thais: To me, physics is the science of reliable model building, and therefore so much of what we learn translates directly to working in the ML/AI space. I think, in fact, this gives us a really unique and useful approach to this area of research. This has certainly contributed a lot to how I approach my social science work, and I think a lot about the robustness and certainty in the type of models I build and measurements I make.

Cheong: What was the most difficult for you, during this transition/expansion?

Thais: I will give two answers here. The first is unlearning what I'll call 'physics exceptionalism', a common belief among physicists that physics is the purest or most useful way to understand the

universe. While of course physics can tell us a lot about the physical world, understanding any kind of societal systems necessarily requires interdisciplinary collaboration and putting other disciplines of knowledge on equal footing with physics. The second challenge—the bigger challenge for me personally—was finding an intellectual home for this work. Academia is still extremely siloed, and my work doesn't fit neatly into a specific discipline. I feel really lucky to have landed in the Data Science Institute at Columbia because I am able to collaborate with people from across the entire university, and I think data science is an apt descriptor for the kind of work I find myself doing.

Cheong: Do you have any tips or advice for young physicists who might be interested in similar research topics and expansion into AI ethics and AI in social sciences?

Thais: Get outside of the physics department! Pursue anything that really sparks your passion and is related to the kind of work you eventually want to do. Some of the most impactful things I did in grad school were participating in student government, taking an operations research class in the Yale business school, and attending AI ethics workshops at ML conferences. Being at a university is a truly unique and incredible opportunity to learn about many different topics, and I encourage students to take full advantage of that. At the end of the day, your career and your research trajectory are yours, and they should serve you. One thing about me is that I can be very stubborn about certain things, and I have always refused to compromise on my values and interests to be a better physicist in a traditional sense. It has worked out so far, and as long as it keeps working out I'll stay in academia, and as soon as it stops working out, I'll leave and do something else. The social impact of my work and the way I am able to live out my personal value system through my research is my north star.

Cheong: What actions could scientists, as individuals, take to make our AI-related technology and tools more ethical?

Thais: I think documentation is a big and relatively easy step. Make sure that you are carefully documenting what data you're using, where it comes from, how you processed it, and then what iterations you went through to build your model, what kind of evaluation metrics you used, and what use cases this model can be employed for. There is some great research in the AI ethics space around datasheets and model cards that pioneered these ideas within the ML community. This is a great practice to get into and a great standard to set for any AI work.

Cheong: How could the physics academia better prepare and train the new generation of physicists for the new era with ethical AI tech? Any systematic changes you would like to see?

Thais: First, I think all physics students should be offered (or required to take even) a computational course that combines data science, data visualization, statistics, and software development principles. Then, part of this course should absolutely be focused on the ethical implications and considerations around this technology. Additionally, I think we need to be more willing to talk about these issues in physics circles. I bring it up in nearly all of my talks and would love to see more folks do the same. We are physicists, but we're also community members, and these issues affect us and everyone around us.

Expanding the Space of Machine Learning for Physics

By Jesse Thaler, Center for Theoretical Physics, Massachusetts Institute of Technolog, NSF Institute for Artificial Intelligence and Fundamental Interactions, jthaler@mit.edu

Jesse Thaler is a professor of physics at MIT. In his research, he fuses techniques from quantum field theory and machine learning to address outstanding questions in fundamental physics. Thaler was elected as a 2022 APS Fellow by the GDS.

"Can a Computer Devise a Theory of Everything?" This provocative question, posed in the title of a New York Times article from 2020, anticipates a (dystopian?) future where theoretical physicists might be replaced by their robotic counterparts. Whether or not such a future will ever come to pass is debatable, but already today, questions like these reflect the immense promise of machine learning (ML) to advance the frontiers of the physical sciences. On the one hand, this kind of question is maddening, since the present-day algorithms driving ML innovations are far removed from those employed by human physicists. On the other hand, this kind of question is inspiring, since it forces us to reflect on what aspects of the scientific process could be systematized and automated.

Robot futures aside, ML has already had an irreversible impact on my field of high-energy physics (HEP), as represented by the growing catalog of papers on the topic. A few years ago, one could legitimately question the benefits of modern ML techniques given the strong performance of traditional methods. Now, one can legitimately question the benefits of solely relying on traditional methods given the dramatic gains seen from ML applications. This is particularly true when one can infuse "physics intelligence" into artificial intelligence, such that decisions made by machines are underpinned by principles, best practices, and domain knowledge from the physical sciences.

As Director of the NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI), I have a front-row seat to the many innovative ways that physicists are incorporating ML into their research. From calculating the properties of nuclei from first principles, to inferring the nature of dark matter from astrophysical observations, to improving the operations of large-scale physics experiments, ML is having a transformational impact across theoretical, experimental, and computational physics. Equally exciting, physics techniques from statistical mechanics and quantum field theory are being used to understand the inner workings of ML systems.

This fruitful dialogue between ML and physics is unlikely to slow down anytime soon. When I was an undergraduate student, multivariate calculus, differential equations, and linear algebra were viewed as the basic mathematical prerequisites for physics research. Today, given the ubiquity of rich data sets in physics, it is equally important for physics students to learn the foundations of probability, statistics, computational methods, and data analysis. In recognition of the growing importance of data science for physics research, we recently launched an Interdisciplinary Ph.D. in Physics, Statistics, and Data Science at MIT. The first recipients of this degree have written Ph.D. theses in fields ranging from particle physics, to plasma science, to gravitational waves, all enabled by ML in some form.

To maintain this exciting momentum at the intersection of physics and ML, I believe we need to expand the space of ML in three complementary directions.

Expanding the space of ML methods. Much of the buzz in ML is around "deep learning", specifically supervised learning with multi-layer feed-forward neural networks. Such techniques have shown transformational potential in physics, but we have a larger opportunity to leverage analysis strategies from various areas of mathematics, statistics, and

computer science. My recent HEP research leverages techniques from optimal transport (OT), by using the "earth mover's distance" to determine whether two collider events are similar or dissimilar. I heard about OT from an MIT colleague working on computational geometry, which is usually not placed under the banner of ML. Instead of limiting the scope of ML to specific tools, though, I think it is more fruitful to view ML as a big umbrella that includes any algorithm that helps us make sense of scientific data sets. This framing inspires us to look beyond neural networks and find the best method to solve the problem at hand.

Expanding the space of ML applications. Much of the recent progress in ML for physics has targeted low-hanging fruit, where off-the-shelf ML algorithms have replaced existing elements of the analysis pipeline. For HEP applications in particular, deep learning is supplanting traditional strategies for object reconstruction and identification. As the community gains more expertise in ML methods, we have an opportunity to translate more aspects of physics into a computational language and thereby enhance the way we manipulate theoretical and experimental data. Two key tools used in physics are symbolic computation (e.g. for theoretical calculations) and numerical simulation (e.g. for experimental modeling). To incorporate these tools into an ML-based pipeline, physicists are exploring the use of symbolic regression and differentiable programming, respectively. These developments force us to go beyond off-the-shelf ML tools and develop custom solutions to integrate ML into physics frameworks.

Expanding the space of ML career pathways. There is a vibrant community in ML for physics, especially among early-career researchers. The members of this community straddle the traditional boundaries between theoretical physicist, experimental physicist, statistician, and data scientist. Through the IAIFI Fellowship program, a group of talented postdocs has assembled in the Boston area to collaborate at the multi-disciplinary intersection between physics and ML. To continue to recruit and retain the most promising talent in this area, we have an opportunity to create new career pathways at the physics/ML interface, both in academia and in industry. Researchers in physics are united in the quest to understand nature. Researchers with dual training in physics and ML have the skills necessary to tackle a broader universe of questions in and beyond the physical sciences.

By expanding the space of ML in physics, we can increase the discovery potential of physics experiments, invigorate theoretical physics research, and demonstrate the value of physics training to address broader societal challenges. We are likely far away from a future where a computer could devise a theory of everything without human input. We are likely closer to a (utopian?) future where a group of humans with interdisciplinary training could devise a theory of everything with help from ML-enabled computation.

An Interview with Jennifer Hobbs: A "Lapsed" Physicist, on a Career in Data Science

By Margaret Zientek

Dr. Jennifer Hobbs, is the VP of Data and Analytics, Lead Data Scientist at Zurich North America, an international commercial insurance provider. She previously worked as the Director of Machine Learning at Intelinair. Prior to that, she worked on computer vision projects at Stats Perform and as a Risk Insights Analyst at Zurich. Hobbs' did her PhD in physics at Northwestern University where she studied how the brain encodes touch which was where she was first exposed to machine learning. She generously gave her time for an interview on how she got started in data science, her advice to physicists, and her thoughts on the future of the field .

Margaret Zientek: What made you interested in switching from academic physics to data science?

Jennifer Hobbs: I had always anticipated staying in academia; it was only in my very final years of my PhD that I realized how challenging the academic job market was and the types of sacrifices that had to be made. I decided going into industry was my best route. I applied to over 200 jobs; only three companies even contacted me for an initial screening. At that time, switching from academia to industry was not the status quo. The university career services could not tell me what someone with a PhD in Physics did for a living outside of being a physicist. I knew another physicist who had gone into analytics in the insurance industry and was doing very well, so that gave me encouragement.

Zientek: Do you have advice for any early career physicists now who may be thinking about switching fields and going into industry as a data scientist?

Hobbs: You don't have to switch out of your physics program just because you want to pursue a career in machine learning (ML). The use of ML within physics is becoming increasingly common, so there could be opportunities even within your core research. If you are specifically interested in the business-data science side, see if your university has courses on entrepreneurship, management, or "business for engineers". If you are interested in pursuing a more technology-focused role, brush up on your coding and learn the fundamentals of test driven development.

Look at summer or academic year internships as an opportunity to broaden your skills and experience what working in industry is like. It can't be overemphasized enough, but networking is key. Even if you consider yourself an introvert, push yourself a bit outside your comfort zone to meet people and build up your network.

Zientek: This is all great advice for getting your foot in the door and gaining some experience. What do you think makes a successful data scientist?

Hobbs: Determine what "type" of data scientist you want to be. "Data science" is an overloaded term and could include anything from a data analyst, to a data engineer, to machine learning researcher, to various things in between. Understand if you like solving business problems, building technology, working with data, or building software. In some of these roles, coding skills may be the most important thing. Others require more business savvy and people skills. Finding that right fit of what you're both good at and enjoy sets you up for success.

Zientek: What keeps you interested in your work as a data scientist now?

Hobbs: I learned that my passion was about solving interesting, challenging problems, which is what attracted me to physics in the first place. I enjoy the pace and variability of problems that working in this space enables. Adoption of data science and machine learning is still a major challenge, but I see this as a form of teaching, which I very much enjoyed when I was in academia. My

recent roles have also enabled me to stay close to the research community through the ability to publish papers, attend conferences, and organize workshops.

Zientek: It's great to hear that you are able to stay connected with the research community. What is your relationship with physics now that you are in industry?

Hobbs: As far as my day-to-day work, I don't "do physics" anymore, but certainly my training influences the way I approach problem solving more broadly. I have worked with a number of "lapsed" (former) physicists in my current and past roles - we're truly everywhere in this domain.

Zientek: What do you think physicists bring to the field of data science?

Hobbs: Physicists are grounded in reality, inherently skeptical about easy results, and committed to reproducibility in ways others aren't. They approach data science as a science: a series of hypotheses, experiments, and analysis. They anticipate what reasonable performance is before conducting an experiment (i.e. building a model), and if it's unexpectedly high, they usually assume they've done something wrong (i.e. there's a leak in the data), not that they've built the greatest model ever. Physicists are also unintimidated by, and usually welcome, difficult, complex, problems. They know how to break something down into smaller parts and solve each one in turn.

Zientek: These do sound like good skills for both physicists and data scientists. What could academic physicists learn from data scientists and vice versa?

Hobbs: If there was something I wish the research and industry sides would learn from one another is that many of their processes and needs are the same, but the terminology and tools might be slightly different.

For example, reproducibility is critical for science. However, sometimes our tools and discipline around accomplishing this are lacking. This is usually what is referred to as "research code": paths are hardcoded, plots are made from GUIs, etc. In industry, code needs to work at scale. The best engineering teams understand the value of testing: unit tests, regression tests, and end-to-end tests. The scientific community, which understands the importance of reproducibility, would benefit from adopting more of this formalized testing rigor and tooling in their work to ensure results are in fact fully reproducible.

Zientek: To switch gears, I want to talk about some of the ways the field of data science is changing and moving in new directions. One is through reducing the barriers to entry for budding data scientists and the democratization of AI. What are the ways you see these movements are advancing?

Hobbs: The open source movement has been critical to the democratization of AI. Imagine how different things would look if PyTorch and Tensorflow were as expensive as Matlab. Having these platforms freely available, plus thousands of open code repositories, has enabled almost anyone interested in learning the field to pick up the necessary tools and start exploring. Beyond this, the public release of large real-world datasets is huge for the field. Educational platforms like Coursera and Udacity offer access to free training, code, and collaboration from highly qualified individuals. Kaggle, hackathons, and other competitive venues give newcomers the opportunity to demonstrate their skills.

Democratization is also used to refer to a number of tools out there (often which cost money) which abstract away the data science process with little or no code. This was driven out of the desire to enable less technical, more business focused roles to generate the same insights to what data scientists had worked to generate (and spent weeks/months building). However, making tools available for everyone to get quick results isn't the solution. They need to understand how the approaches work. They need to understand the implications of their results. They need to be informed citizens of the "data democracy". And that mindset- that quest for knowledge and understanding over pure "results" - is what the scientific community should bring to the broader data science community.

Zientek: I'm curious to hear where you think data science is going in the future. What are some of the

biggest opportunities and challenges in the field today?

Hobbs: Building off the discussion on democratization, I think the barrier to entry for delivering a model will continue to decrease. With the increased tooling (and competition) in the data annotation and auto-ML space, it is easier to get to proof-of-concept with minimal modeling effort. Being able to automatically pipe in more and more training data in an active-learning fashion often delivers more improvement (and therefore more business value) than building a fancy, custom model.

As a result of that, on the career front I think the need for specialized engineers to run and train models at a scale, implement monitoring systems, and measure model impact will only continue to grow. At the same time, the democratization of modeling has dramatically reduced the effort needed to implement and train models. Together, I think this will continue to drive the demand for ML Engineers, MLOps engineers, and "full stack" data scientists, while reducing the demand for data scientists who are only involved in the model development process.

At the other end of the spectrum, the need for technical individuals who can communicate and serve as a conduit between the business side and engineers is not going away. So for someone entering the field, my encouragement is to make sure you stay sharp on the engineering and communication fronts and not focus just on the modeling, even if that is the exciting part. If you hate the engineering elements, you may want to reassess what it is about "data science" that interests you and develop your strengths more on the business side.

I think one of the challenges the field faces right now is actually driven by the disconnect between the ease of doing *something* in ML and the challenge of doing something impactful. Especially over recent years, it is easy to get support, sales, funding, and publications by incorporating machine learning into an existing product or workflow. This has led to an explosion of ML-based technology and features like improved chatbots for customer support. But have these meaningfully changed our lives in a positive manner? Certainly some have been helpful, but are these the big problems of our time? Identifying important, impactful problems in both science and technology, which were previously out of reach prior to machine learning, and committing to solving those problems may actually be the biggest challenge.

The Rise of the Data Physicist

By Benjamin Nachman

Benjamin Nachman is a Staff Scientist at Berkeley Lab leading the Machine Learning for Fundamental Physics Group, Research Affiliate at the UC Berkeley Institute for Data Scientist, and Secretary of GDS.

The rapid advances in machine learning for physics has been fueled by data. Our data are complex and are often complemented with equally complex ab initio simulations to connect our experiments with underlying physics principles. On the other hand, researchers developing and deploying machine learning to physics are often not the ones who directly collected and curated the data. No one knows the data as well as the people who designed, built, and executed the experiments, but the reality is that specialization has created researchers with different skill sets. How can we respect the work of experimentalists while also ensuring that state-of-the-art tools are used to make the most of our precious data?

I would argue that we need a new[1] type of physicist that is neither an "experimentalist" or a "theorist" - they are a "Data Physicist". These scientists have the core skills to understand and interrogate data as well as the computational and theoretical background to relate these data to underlying physical properties. Unlike a traditional "experimentalist", a Data Physicist will likely not have extensive (or any) hands-on instrumentation experience and unlike a traditional "theorist", they may not have extensive (or any) experience with first-principles calculations aside from coursework. It is not enough to make data easily accessible and well-documented - we also need to train a cohort of scientists who are well-equipped to use these data for science. In particular, I expect the Data Physicist to have a strong physics background, but extensive training in statistics/data science/machine learning and scientific computing. Software and computing are becoming commensurate with instrumentation and it is important and necessary physics research to work on these topics for maximizing the science from our data.

How can we create this cohort? There needs to be career paths (starting in graduate school) for Data Physicists. This includes degree programs as well as long term funding prospects (that are not all short grants as is typical for computing). New initiatives like the NSF Artificial Intelligence (AI) Institutes (e.g. IAFAI and A3D3) are great examples of interdisciplinary machine learning for physics research and I strongly support continuing and expanding these initiatives. However, smaller scale funding for individual research groups that is not tied to a particular experimental effort will also be important for a sustained effort.

It is fantastic that many people are excited about sharing and exploring data. The reason I became an "experimentalist" in graduate school is precisely because I wanted to interact directly with data. On the other hand, I have found through the course of my career so far that the labels "experimentalist" and "theorist" can be unnecessarily restrictive. With a growing need for cross-cutting methodology and the growing availability of publicly available datasets, we need to train and support Data Physicists to make the most of our precious data. With open minds and the right skill set, we will be ready to make the discoveries of tomorrow.

This article is adapted from *Computing and Software for Big Science* 6 (2022) 17. See also *The Future of High Energy Physics Software and Computing*, arXiv: 2210.05822. The author acknowledges Professor David Shih for coining the term "Data Physicist" at the recent Particle Physics Community Planning Exercise.

REFERENCES

1. Some fields of physics have a structure similar to what I'm describing, where "instrumentationlists" are separate from people who do data analysis. I am writing from the perspective of particle/nuclear physics, where this is not the case.

March Meeting 2023

SUNDAY, MARCH 5

T5. Graph Neural Networks
Chair: William Ratcliff, National Institute of Standards and Technology
Room: Room 129

MONDAY, MARCH 6

A53. AI and Materials I
Sponsoring Units: GDS DMP
Chair: Rama Vasudevan, Oak Ridge National Lab
Room: Room 307

B53. Autonomous Control
Sponsoring Units: GDS DQI
Chair: Maria Longobardi, University of Basel, Switzerland
Room: Room 307

D02. Statistical Physics Meets Machine Learning I
Sponsoring Units: GSNP DSOFT DBIO GDS
Chair: Yuhai Tu, IBM T. J. Watson Research Center
Room: Room 125

D53. Machine Learning for Spectroscopy
Sponsoring Units: GDS
Chair: Nina Andrejevic, Argonne National Laboratory; Davis Unruh, Argonne National Laboratory
Room: Room 307

TUESDAY, MARCH 7

F02. Statistical Physics Meets Machine Learning II
Sponsoring Units: GSNP DSOFT DBIO GDS
Chair: David Schwab, The Graduate Center, CUNY
Room: Room 125

F53. AI and Materials II
Sponsoring Units: GDS
Chair: Carolina Adamo, Northrop Grumman
Room: Room 307

K53. Designing Neural Networks for the Structure of Physics Data
Sponsoring Units: GDS DCOMP
Chair: William Ratcliff, National Institute of Standards and Technology
Room: Room 307

WEDNESDAY, MARCH 8

M53. Computer Vision
Sponsoring Units: GDS
Chair: William Ratcliff, National Institute of Standards and Technology
Room: Room 307

N53. AI and Materials III
Sponsoring Units: GDS DMP
Chair: Trevor Rhone; Ayana Ghosh, Oak Ridge National Lab
Room: Room 307

N60. Emerging Trends in Molecular Dynamics Simulations and Machine Learning IV
Sponsoring Units: DCOMP DSOFT GDS DPOLY
Chair: Sabrina Wahler, LMU Munich, RoseExplosive UG
Room: Room 419

Q53. AI and Statistical/Thermal Physics
Sponsoring Units: GDS
Chair: Caroline Desgranges, University of Massachusetts Lowell
Room: Room 307

Q60. Emerging Trends in Molecular Dynamics Simulations and Machine Learning V
Sponsoring Units: DCOMP DSOFT GDS DPOLY
Chair: Aravind Krishnamoorthy, University of Southern California
Room: Room 419

Q52. Advanced Technologies for Medical Physics
Sponsoring Units: GMED GDS
Chair: Wojciech Zbijewski, Johns Hopkins University; Alejandro Sisniega
Room: Room 308

THURSDAY, MARCH 9

S53. Machine Learning
Sponsoring Units: GDS DFD DMP
Chair: Jennifer Hobbs, Zurich North America
Room: Room 307

T12. Invited Undergrad Friendly Energy Research and Data Science
Sponsoring Units: GERA GDS
Chair: Davis Unruh, Argonne National Laboratory
Room: Room 235

T51. Invited Undergrad Friendly Data Science for Industry
Sponsoring Units: FIAP GDS
Chair: Maria Longobardi, University of Basel, Switzerland
Room: Room 321

T53. Focus Data Science, ML and Active Matter
Sponsoring Units: GDS DBIO
Chair: Jerome Delhommelle, University of Massachusetts, Lowell
Room: Room 307

W53. Focus Data Science for Climate
Sponsoring Units: GDS GPC DFD
Chair: William Ratcliff, National Institute of Standards and Technology
Room: Room 307

FRIDAY, MARCH 10

Y53. Open Science and Data Sets
Sponsoring Units: GDS FIAP
Chair: Savannah Thais, Columbia University
Room: Room 307

MONDAY, MARCH 20

EE02. V: Machine Learning in Physics
Sponsoring Units: GSNP DSOFT DBIO GDS
Chair: Arpan Biswas, Oak Ridge National Lab
Room: Virtual Room 2

TUESDAY, MARCH 21

LL10. V: Data Science I
Sponsoring Units: GDS
Chair: Yisheng Chai, Chongqing University
Room: Virtual Room 10

WEDNESDAY, MARCH 22

UU10. V: Data Science II
Sponsoring Units: GDS
Chair: Saravana Prakash Thirumuruganandham, Centro de Investigación de Ciencias Humanas y de la Educació
Room: Virtual Room 10

Save the Date

The Data Science Education Community of Practice (DSECOP) is planning on having its second workshop from June 26 through 28 in College Park, MD! Curious about DSECOP?

Please visit our site: dsecop.org

GDS Executive Committee

Chair
Maria Longobardi (03/22–03/23)
University of Basel

Chair-Elect
William Ratcliff (03/22–03/23)
National Institute of Standards and Technology

Vice Chair
Jerome P Delhommelle (03/22–03/23)
University of Massachusetts Lowell

Past Chair
Wolfgang Losert (03/22–03/23)
University of Maryland, College Park

Treasurer
Savannah J Thais (03/22–03/25)
Columbia University

Secretary
Benjamin Nachman (03/21–03/23)
Lawrence Berkeley National Laboratory

Assigned Council Representative
James Knox Freericks (01/20–12/23)
Georgetown University

Member-at-Large
Emine Kucukbenli (03/21–03/23)
Harvard University

Member-at-Large
Nima Dehmamy (03/21–03/23)
Northwestern University

Member-at-Large
Jennifer Hobbs (03/22–03/25)
Intelinair, Inc.

Member-at-Large
Trevor David Rhone (03/22–03/25)
Rensselaer Polytechnic Institute

Early Career Member-at-Large
Stefano Roberto Soleti (03/22–03/24)
Lawrence Berkeley National Laboratory

Student Member
Kyle Peter William Hall (03/21–03/23)
Memo Univ of Newfoundland