Topical Group on Data Science

Winter 2020 Newsletter

Message from the Editor

Kyle Wm. Hall, GDS Newsletter Editor

Hello GDS Members!
It is my great pleasure to provide you with the first newsletter for the Topical Group on Data Science (GDS). My vision for the newsletters is that they will provide GDS members with original content (e.g., interviews and perspective pieces) while also highlighting physics-relevant data science news and events. I want the newsletter to grow while addressing the needs and interests of the GDS membership, so I encourage you to reach out to me by email at gds@aps.org. If you publish some exciting work at the intersection of physics and data science, write to me. Perhaps, we can feature it in an upcoming newsletter. Alternatively, you might have some news or views that you would like to share with your fellow GDS members. I am eager to have members contribute to the newsletter, and look forward to hearing from you.

This issue contains a number of exciting elements starting with a welcome message from GDS Chair, Mohammad Soltanieh-ha. For this issue, I had the privilege to interview Nathan Sanders (a physicist who made the leap to industrial data science) on topics ranging from data science tools and trends to data science job hunting. There are contributed pieces by Dimitri Bourilkov and Sergei Kalinin. This issue also contains an overview of GDS events at the Marching Meeting in Denver, and pointers to additional items. I hope that you enjoy this issue.

Kind Regards,

Kyle Wm. Hall

Kyle Wm. Hall is currently a Postdoctoral Fellow with Michael L. Klein at Temple University. Kyle obtained an interdisciplinary PhD from the University of Calgary under the supervision of Drs. Peter G. Kusalik and Sheelagh Carpendale. His research lies at the intersection of molecular simulations and data science, and spans data visualization, human-computer interaction, polymer physics, and crystallization.

In this Issue

How to Connect

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

Email
gds@aps.org

Website
https://www.aps.org/units/gds/index.cfm

LinkedIn
https://www.linkedin.com/company/apsdatascience

Twitter
https://twitter.com/apsdatascience

Facebook
https://www.facebook.com/APSDataScience

The information and views provided herein are those of the individuals involved, and do not necessarily correspond to those of APS, the GDS, or the Newsletter Editor.

Message from the Chair

Mohammad Soltanieh-ha, GDS Chair

Dear GDS members,

It is my pleasure to welcome you to the first issue of the GDS Newsletter, as the founding chair of the Topical Group on Data Science (GDS).

In under a year, GDS has attracted over 700 members and is one of the fastest-growing units within APS. The mission of this unit is to facilitate data science education amongst physicists as well as to promote the research in the field at this crucial time. This mission has been heavily supported by the APS community which demonstrates the necessity of such a group.

Data science approaches are becoming standard in many branches of science and have opened a new window to scientific discovery. Physics is not an exception and this can be easily observed by looking at the trend of machine learning and data-science-related papers as well as talks at the APS meetings in the past few years. Both physics and data science fields can benefit tremendously from knowledge sharing as there are many similar use-cases and common problems. We are hoping that GDS can connect people across multiple physics disciplines leveraging the same methods. I believe the creation of this group was inevitable and I am extremely lucky to have had the opportunity to work with the bright executive committee of GDS to form this unit in April 2019.

Data science needs within industry are on the rise and it has been the case that physicists make the best candidates for many of those roles. As someone who once made this transition, before returning to academia, I am well aware of the lack of resources within the physics departments and within APS for such a career shift. In physics, data science education is lacking in graduate school curricula and even at the post-doc and faculty level, because the techniques are so new. It has been one of our main priorities to bring in more training sessions to APS meetings to support faculty, post-docs, graduate and undergraduate students, not only to enable them in their research, but also to make them marketable, should they want to transition to an industry career. The technology is evolving very rapidly and the techniques that are emerging every day can benefit our community. For the March Meeting 2020, we have arranged many invited and focus sessions that I hope you will attend. We are also offering a short course, two tutorials, as well as a workshop in partnership with Google Cloud. Please find the details of our program later in this newsletter.

I want to close with a special thanks to the executive committee and, in particular, to the GDS Chair-Elect Jie Ren, for their efforts during this first year.

I hope that you will support the activities of GDS by attending our events at the March and April Meetings or considering to present your work in future events. There are many other ways that you can get involved; please feel free to reach out to us if you'd like to learn more about the opportunities: gds@aps.org . To stay updated follow us on LinkedIn, Facebook, and Twitter: @APSDataScience.

Kind Regards,

Mohammad Soltanieh-ha

Mohammad Soltanieh-ha is a Clinical Assistant Professor at the Information Systems department at Boston University. Mohammad obtained his Ph.D. in computational physics in 2015 from Northeastern University where he studied strongly correlated electronic systems in low dimensions. Upon graduation, he accepted a role as a data scientist at Infor. In 2018, Mohammad returned to academia and joined the faculty of Boston University to teach data science. Mohammad is also a Faculty Expert at Google Cloud. His current research interest revolves around computer vision applications in automating cancer diagnosis as well as large scale computing and HPC. Mohammad has been an active member of the Ameican Physical Society; he's been a founding member of Boston Local Links (2015) and FECS (2016). He has also served on the Committee on Membership (2015-2018).

Practitioner Perspectives on Data Science

An Interview with Nathan Sanders (Chief Scientist at Warner Media Applied Analytics)

Nathan Sanders, pictured to the right, holds a BS in Astrophysics and Physics from Michigan State University, and a PhD in Astronomy and Astrophysics from Harvard University, which he completed under the supervision of Alicia Soderberg. While completing his PhD, Dr. Sanders served as a Legislative Fellow for the Massachusetts State Legislature. Dr. Sanders was Vice President of Quantitative Analytics at Legendary Entertainment before assuming his current position as the Chief Scientist of Warner Media's Applied Analytics division.

KH: How did you move from astronomy into data science for media applications?

NS: As part of my astronomy PhD, I took courses in statistics and computational science to support my thesis research activity. I found this work so compelling that I decided that I wanted to emphasize it in my career. At the same time, "big data" was an emerging topic in business and "data science" was being built as a discipline at many companies, so I realized that I could apply my skills far beyond just astronomy. When I made that decision, I set out to find industries and organizations that were staged to take a really fresh, dramatically new perspective on their business using data. I wanted to pursue transformational rather than incremental research, which meant finding organizations that didn't have long, well-established quantitative research teams. It turned out to be the right time and place to do that kind of work in the entertainment domain.

KH: Describe some of your current data science projects.

NS: Professionally, my team has long worked on predictive modeling for targeted marketing, natural language processing for social media data, market modeling and box office projection, inference on behavioral influence and advertising effectiveness, and other topics. More recently, we have had a focus on applying multi-modal deep learning models for audiovisual data to understand and characterize media content. Much of this work has been led by my close colleagues Ben Lawson and Jonathan Foster. Jonathan is a fellow astronomer by training.

Personally, I continue to do a lot of work applying data science to public policy problems. I have studied how water pollution effects different communities at different rates, and particularly how sewage discharged into rivers tends to overburden minority and low-income communities. This has, I believe, profound environmental justice implications. I have also been collaborating with criminologists to infer the trends in incidence and severity of mass public shootings. Because these events happen (fortunately) only a few or a handful of times per year, this is a very challenging small-data statistical inference problem. Because the rates have been (unfortunately) growing, it is critical to understand what practical effect policy decisions can have.

KH: What data science techniques and tools do you use in your line of work?

NS: Our team has intentionally pursued a very broad suite of techniques ranging from Bayesian inference (my frequent area of focus) to natural language processing to computer vision to predictive modeling to active learning and more. The probabilistic programming language Stan is a personal favorite of mine, and invaluable for Bayesian models. Our shop has been mostly oriented around python and R, but also pulling packages from other environments as needed.

KH: What is your favorite data science resource and why?

NS: Two things. First, the Harvard Data Science Review (HDSR) is a really exciting new journal that

serves as a "kaleidoscopic" forum for data scientists working across academia, industry, and government to come together to exchange ideas. I think this intersection of domains is exactly what makes it exciting to be a data scientist. For disclosure, I serve as an Associate Editor for the HDSR. Second, Andrew Gelman's blog cannot be beat for always interesting commentary and perspective on statistical inference.

KH: What data science trends are you watching in 2020?

NS: Influenced in large part by friends like Alex D'Amour (Google) and Victor Lei (TripAdvisor) doing fascinating work in this area, I am interested to watch the expanding focus on causality and causal
inference in applied data science. How can companies move beyond uninterpretable black-box predictive models and descriptive exploration of correlations to understand the true underlying causal effects that drive experimental and business outcomes?

KH: Where do you see the greatest differences between data science in academia and in industry?

NS: This depends sensitively on the organization, but I think the biggest hurdle for many scientists making the transition is project timescale. In academia, researchers often largely control their project timelines, and work on timescales of many months or even a year or more. In applied work in industry, project timelines are often dictated by external stakeholders and organizational requirements, often down to timescales of a week or even days or hours. An important strategy for data scientists in industry is to integrate a thoughtful development process with these externally-oriented timelines, so that ideas and models can be developed persistently across multiple project milestones, or even across projects.

KH: What is your advice for physicists who want to pursue data science roles in industry?

NS: The best way to explore career options in any other field is to get in the routine practice of informational interviewing. Read about jobs you are interested in, find people doing that work, and send them a polite email asking for a few minutes of their time to chat by phone to learn more. When you've connected, ask them who else you should talk to. If you do one or a few of these per week, you'll quickly build up a strong understanding of how data science is applied in different fields and where you would like to work. Equally important, you will have a network of people who know about and care about you, who will be ready to assist when your next job search comes to pass.

One outside the box idea for the job search is to indulge your personal interests, and be an active member of your community. Volunteer, do science outreach, get involved in advocacy for policy issues you care about. The contacts and friends you form through those activities will be people who share your passions and interests, and may be your greatest advocates when you go on the market.

In Review: Deep Learning & Particle Physics

Dimitri Bourilkov (Associate Scientist, Department of Physics, University of Florida)

Flashback ten years - the Large Hadron Collider in Geneva, Switzerland started taking data, and unprecedented volumes were soon being collected. More or less traditional methods for analysis and extraction of the underlying physics were the norm of the day. This was about to change soon: both the CMS and ATLAS collaborations used machine learning methods based on decision trees to discover the Higgs boson by extracting the small signals from a huge background, like needles in a haystack.

A recent review by D. Bourilkov, "Machine and deep learning applications in particle physics" – just published in the International Journal of Modern Physics A, Vol. 34, No. 35, 1930019 (2019) – tracks the many amazing ways in which artificial intelligence techniques are transforming the analysis and simulation of data in particle physics. The traditional way to do such work is to first develop algorithms based on domain knowledge, implement them in software, and then use the resulting programs. This process is labor intensive, and analyzing complex datasets with many input variables becomes increasingly difficult and sometimes intractable. Artificial intelligence and the subfield of machine learning attack these problems in a different way: instead of humans developing highly specialized algorithms, computers learn from data how to analyze complex data and produce the desired results. There is no need to explicitly program the computers. Artificial neural networks try to imitate in a simplified way biological brains. The neurons and synapses are replaced with connected layers of nodes and links. Nodes take inputs from their connections, and perform non-linear transformations to extract the relevant information present in the input data. Once trained, a neural network can analyze large amounts of new data much faster compared to traditional methods.

Advances in academic research paired with the needs of large companies like Google, IBM, Amazon, Facebook and Netflix, just to name a few, are producing a fundamental paradigm shift, especially with the recent successes of deep learning, using neural networks with several layers of neurons. The review introduces novel analysis techniques based on boosted decision trees and various types of neural networks, highlights cutting-edge applications in the experimental and theoretical/phenomenological domains, and describes the challenges in their application. The interaction between physics and machine learning is a two-way street, enriching both disciplines and helping to meet the present and future challenges of data-intensive science at the energy and intensity frontiers.

For more information and to access the full review, visit https://doi.org/10.1142/S0217751X19300199

Dr. Dimitri Bourilkov was born in Sofia, Bulgaria, where he obtained his PhD in particle physics at the Institute for Nuclear Research and Nuclear Energy. He has conducted research at the largest accelerators in the world with the BIS-2experiment at Serpukhov, Russia, and the L3 and CMS experiments at LEP and LHC in Geneva, Switzerland. After starting his career in Sofia, Dr. Bourilkov has worked at the Joint Institute for Nuclear Research in Dubna, Russia, at the Radboud University in Nijmegen, the Netherlands, at ETH Zurich, Switzerland, and is now a Scientist in CMS with the University of Florida,Gainesville.

Parsimony Is Physics

Sergei V. Kalinin (Distinguished Research Staff Member , CNMS ORNL)

Over the last decade, machine learning and artificial intelligence catapulted from the relatively obscure subfield of computer science (or SciFi domain) to virtually all aspects of everyday life - from front pages of newspapers to job announcements. Yet the adoption of machine learning tools by the physics community has been relatively slow and highly non-even. Only the last year, APS formed the Data Sciences Group, recognizing the role that machine learning is starting to play in physics. In this newsletter, I and my colleagues aim to share some thoughts and ideas on what can be the future for ML in physical sciences. Indeed, it is worth remembering that the vast majority of machine learning methods are correlative in nature. In other words, ML serves as an interpolator between the large dimensional input and output spaces, for example establishing correlation between the image of the cat and the class of animals. However, the well-known maxim is that correlation is not causation. Finding the relationship between strongly correlated variables is not equivalent and is generally insufficient to establish functional, and even more so causal, relationship. For example, see http://tylervigen.com/spurious-correlations for multiple examples of curious correlations. Hence, the question becomes how can the correlative ML methods be connected to hypothesis-driven nature of modern physics. One such pathway is through simplicity, or parsimony. The immeasurable multitude of crystallographic structures are built from a small number of simple building blocks. The collective dynamics of ferroics can be described via small number of macroscopic order parameters. Similarly, dynamics of macromolecular systems can be well described by a small number of collective variables. Finding these compressed representations from the real space modelling or imaging data is an ideal task for ML methods, as exemplified by Variational Autoencoders and similar methods. Here, the ML algorithms directly matches the reductionist understanding of physics – finding simplest representation of complex systems. Similar approach can be extended to function rather than structure. For example, the complexity of the Mandelbrot set is described by several lines of code. Similarly, the range of exotic phases and behaviors observed in real materials can often be described through simple physical laws defining the interactions between the individual blocks, were it the exchange integrals in lattice Hamiltonians or more complex force fields and equations of motion. Finding these functional descriptors from observations is one task for ML in physics going forward. It is very curious how much physics can be learnt from observations, even before doing experiment. Can AI figure out Newton mechanics from observations of solar system (yes). Can it postulate existence of Pluto from anomalies in Neptune orbit? What about special relativity from observations of Mercury orbit? Now, if this is possible, what if we observe atoms, molecules, or ferroelectric domain, and not planets - and deal with non-differentiable trajectories, images rather then coordinates, lost and missing data, etc? Some of the example of such recent work is graph networks (https://arxiv.org/abs/1909.05862), statistical distance minimization (https://arxiv.org/abs/1907.05531), pattern formation [Physical Review Letters 124, 060201 (2020)]. However, this field has just started!

Sergei Kalinin is the distinguished staff member at the Center for Nanophase Materials Sciences at Oak Ridge National Laboratory. He received his MS degree from Moscow State University in 1998 and Ph.D. from the University of Pennsylvania (with Dawn Bonnell) in 2002. His research presently focuses on the applications of big data and artificial intelligence methods in atomically resolved imaging by scanning transmission electron microscopy and scanning probes for atom by atom fabrication, extraction of relevant physics and chemical behaviors on the single-atom levels, as well as mesoscopic studies of electromechanical and transport phenomena via scanning probe microscopy. Sergei has co-authored >600 publications, with a total citation of >30,000 and an h-index of >85. He is a fellow of MRS, APS, IoP, IEEE, Foresight Institute, and AVS; a recipient of the RMS medal for Scanning Probe Microscopy (2015); Blavatnik Award for Physical Sciences (2018), Presidential Early Career Award for Scientists and Engineers (PECASE) (2009); Burton medal of Microscopy Society of America (2010); 4 R&D100 Awards; and a number of other distinctions.

March Meeting 2020

Here are some of the exciting GDS events and sessions taking place at this year's Marching Meeting. Some additional events related to data science are also included.

Kavli Foundation Special Symposium
Frontiers of Computation: Machine Learning and Quantum Computing
Wednesday, March 4
2:30 pm – 5:30 pm
Room: Bellco Theatre

GDS Business Meeting
Session Q15
Wednesday, March 4th
5:45 pm – 6:45 pm
Room: 706

FIAP Industry Day Reception
Co-sponsored by GDS; Open to GDS Members
Thursday, March 5th
5:30 pm – 7:30 pm
Stout Street Lobby (Outside Room 407)

GDS Short Course
Deep Learning for Image Processing Applications
Registration Required
Sunday, March 1st
8:00 am – 5:30 pm

Workshop
Google Cloud Hero
Registration Required
Tuesday, March 3rd
6:00 pm – 9:00 pm

Tutorial
Active Learning and AI for Computational and Autonomous Experiments
Tutorial 2
Registration Required
Sunday, March 1st
8:30 am – 12:30 pm

Invited Symposia
New Ways of Seeing with Data Science
Session J28
GDS/FIAP
Tuesday, March 3rd
2:30 pm – 5:30 pm
Room: 405-407

Data Science in the Physics Curriculum
Session M36
GDS/FED
Wednesday, March 4th
11:15 am – 2:15 pm
Room: 601/603

GDS Focus Sessions
Data Science I: Big Data & ML
Session F20
Tuesday, March 3rd
8:00 am – 11:00 am
Room 301

Data Science II: Machine Learning
Session G20
Tuesday, March 3
11:15 am – 2:15 pm
Room 301

Data Science III: Deep Learning
Session R20
Thursday, March 5th
8:00 am – 11:00 am
Room 301

Joint Sessions
Machine Learning for Quantum Matter
Sessions L39, M39, R39, S39, U39, W39
DCOMP/GDS/ DMP

Emerging Trends in Molecular Dynamics Simulations and Machine Learning
Sessions J45, L45, M45, P45
DCOMP/GDS/DSOFT/DPOLY

Statistical Physics Meets Machine Learning
Session U24
GSNP/GDS

Additional Items of Interest

Recent Special Issues

Machine Learning for Physical Systems
Journal of Computational Physics

Machine Learning and Statistical Physics: Theory, Inspiration, Application
Journal of Physics A: Mathematical and Theoretical

Upcoming Events

The Confluence of Science-Based and Machine Learning Approaches in Energetic Materials Research
Energetic Materials Gordon Research Conference
Location: Newry, ME, USA
May 31st – June 5th, 2020

Emerging Imaging Techniques at the Intersection of Physics and Data Science
Image Science Gordon Research Conference
Location: Easton, MA, USA
Dates: June 7th – June 12th, 2020

Machine Learning and Informatics for Chemistry and Materials
Telluride Science Research Center Workshop
Location: Telluride, CO, USA
Dates: July 27th – July 31st, 2020

GDS Executive Committee

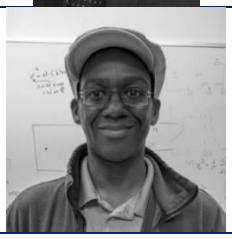
Chair
Mohammad Soltanieh-ha
Boston University

Chair-Elect
Jie Ren
Merck & Co.

Vice Chair
Wolfgang Losert
University of Maryland

Treasurer
Skanda Vivek
Georgia Gwinnett College

Secretary
William Ratcliff
National Institute of Standards and Technology



Member-at-Large
Brian Barnes
U.S. Army Research Laboratory

Member-at-Large
Sergei V. Kalinin
Oak Ridge National Lab

Member-at-Large
Dimitri Bourilkov
University of Florida

Member-at-Large
Cheng-Chien Chen
University of Alabama - Birmingham

Early Career Member-at-Large
Rachel J Henderson
Michigan State University

Student Member
Alexandra M Courtis
University of California, Berkeley

Newsletter
Kyle Wm. Hall
Temple University

Webmaster
Emine Kucukbenli
Harvard University