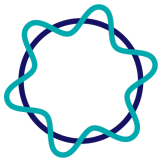


Moore-Sloan Data Science Environments Summit - Program Abstracts

November 2019
Santa Fe, New Mexico

How Universities can Attract and Retain Data Scientists

Nick Adams

Data Scientists can add incredible value to your research community. But how do you evaluate, attract, and retain them? This lightning talk describes the six most common types of academic data scientists, what they can offer to your community, how they're motivated, and what they need to thrive and foster data science throughout your campus. With insights distilled from 5 years as a sociologist, methodological innovator, software architect, and institution builder at the UC Berkeley Institute for Data Science, Nick Adams offers recommendations that will help you define your recruiting & retention strategy and prepare your campus to attract the sort of catalysts already upgrading scientific practice at the world's leading universities.

Public Editor: The Collaborative Way to Credible News

Nick Adams

Public Editor is a new collective intelligence system guiding the public to label and score news articles' credibility. The system is ready to deploy at a national scale to protect the public discourse during our upcoming Presidential elections.

TagWorks: Best Ever Textual Training Data by the Crowd

Nick Adams

Natural Language Processing has seen slow gains over the decades. But just as labeled training data revolutionized image classification, textual data is now tractable data for machine learning and AI thanks to TagWorks.

Exploring the future of the hackweek

Anthony Arendt

Hackweeks provide opportunities for community building, peer learning, networking and collaborative project work within a welcoming and inclusive environment. The hackweek model

evolved from collaborations within the MSDSE astronomy community and has since been applied to neuroscience, geospatial science and many other sub-domains. This session invites conversation around lessons learned from previous events, and explores our collective vision for the future of the hackweek.

The challenge of analyzing not-your-own-data: documentation lessons and opportunities from high-energy physics to homelessness support

Matthew Bellis

The last 20 years has seen an explosion of machine learning tools, algorithms, and platforms all geared toward analyzing both big and small datasets and quantifying the uncertainties on the results. With more and more data being made "open" to anyone who wants to analyze it, these tools are immensely useful as it means that data scientists don't need to continuously need to reinvent the wheel. However, just because you can apply an algorithm to some dataset doesn't mean you fully understand the details of how that data was collected and how that might affect any conclusions you draw from it. I will discuss efforts made by physicists at CERN to make available to anyone interested subsets of the very same data in the very same format that CERN physicists themselves analyze and our efforts to educate users on how to use the data properly. I will also discuss the other side of the coin in my and my colleagues' struggles in analyzing data on the homeless population in eastern-NY state using a federally-mandated data schema.

Talking With the Public About Data Science

Meredith Broussard (moderator), Joshua Tucker, Andrea Jones-Rooy, Sara Stoudt

Communicating with the public about data science is different than communicating with colleagues. In this session, we will talk about best practices for talking to a lay audience about data science, from blogging to journalism to writing op-eds for major publications.

Making open datasets more accessible with Gigantum

Dav Clark

Even with all of the progress that's been made on open tools and data, there's often a lot of assumed knowledge even for the most reproducible papers and results. We start from the perspective of a paper illustrating the potential of the Healthy Brain Network, discuss barriers for

newcomers wanting to extend or interrogate this work, and finally demonstrate one approach using tools like FlyWheel and Gigantum.

Sustainability and profitability

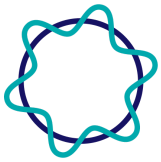
Dav Clark

In my own experience, transitions between academe and industry can be unnecessarily rocky and stressful, but many such transitions have worked out pretty well - both with novel companies (e.g., Anaconda) and efforts within established companies (e.g., Nvidia). We can actively support models that have worked to sustainably fund critical data science infrastructure via commercial models. I'll ground this by talking a little about my experience inside Gigantum - where we're building a highly automated and streamlined approach to portable, collaborative data projects.

Communities of practice for Jupyter deployments on shared infrastructure

Jim Colliander

Projects like Pangeo and education-driven deployments like the UC Berkeley DataHub have demonstrated the power of assembling open tools, deploying them on the cloud and making them available to researchers and students. Similarly, in Canada, the Syzygy (https://syzygy.ca) and Callysto (https://callysto.ca) projects, which deploy JupyterHubs for researchers and grades 5-12 classrooms across the country, are victims of “catastrophic success,” where use of the hubs and interest for expansion of services (e.g. custom Pangeo-like hubs for research projects or workshops) is increasing much more rapidly than anticipated. In the Canadian context, this rapid growth in demand is driving ideas for the creation of a nonprofit organization whose mission is to support people who are engaged in interactive computing. The organization, International Interactive Computing Collaboration (2i2c), would be seeded in Canada with the intention of fostering a global network of people invested in advancing interactive computing, openly sharing software and best practices, and advocating for the continued development of cloud agnostic infrastructure. In this session, we will give a brief overview of the Syzygy project, introduce the vision of 2i2c and discuss with participants how such a nonprofit organization could benefit the broader community, and what principles it should follow to successfully grow into an international network where individual nodes can fulfill national/regional missions.



Data Science Consulting as a University Service

Kyle Cranmer

This BOF aims to discuss approaches to building new university services resulting from the MSDSEs, from setting career paths for RSEs to strategizing around sustainability of services. A motivation for this BOF is the ongoing efforts at New York University to establish Data Science and Software Services (DS3). DS3 aims to assist faculty and research staff by bringing bring software engineering expertise as well as data science and statistical methods, tools, and thinking to academic projects across campus. DS3 will be a central service for faculty and research staff to access data science and statistical methodology expertise and labor for projects and grants.

Enabling science through better data management

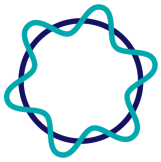
Diya Das

We have observed an exponential increase in the amount of health-related data amassed in the past few years. In order for scientific and technological advances in health care to scale at an equivalent rate, we need to implement smarter and more strategic methods of acquiring, standardizing, and storing data. Our data management team at Genentech is tasked with ensuring that data is available, actionable, and reusable for analysis. Through the implementation of cutting edge data management strategies, the team is enabling scientists across the company to use old data to answer new questions, while also ensuring that prospectively-generated data is Findable, Accessible, Interoperable, and Reusable (FAIR). Facilitating better data management processes will enable us to realize our scientific mission by making every data point count.

Making the Most of Your Data Science Institute

Diya Das, Daniela Huppenkothen

So you've decided to join a data science institute. How can you make the most of it? We'll provide some guiding principles for founding members of data science institutes, with examples drawn from our experiences at all three Moore-Sloan Data Science Environments. This talk is aimed at leaders (at all levels), but we hope will be informative for all.



GEOME Evolution Blockchain

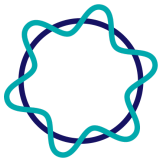
Neil Davies

Blockchain technology promises to digitally integrate the value chain of scientific samples, seamlessly maintaining the chain of custody of specimens and data. Blockchains can elegantly address the technological challenge of registering digital assets with cryptographic identifiers, linking all exchanges of assets in an immutable decentralized ledger, and enabling programmable ‘smart’ contracts to enforce terms of use. The Genomic Observatories Meta-Database (GEOME) is a web-based database capturing the who, what, where, and when of biological samples and associated genetic sequences. GEOME helps users ensure the metadata from biological samples are findable, accessible, interoperable, and reusable, comply with global data standards, integrate with R, and ease publication to DNA sequence archives. Here we present a proof of concept GEOME Evolution Blockchain that will manage storage, querying and retrieval of GEOME sample data, and be able to provide a provenance history of any sample. The goal of the proof of concept is to demonstrate the use of a blockchain for managing data provenance.

Goodhart’s Law: Are Academic Metrics Being Gamed?

Michael Fire

The academic publishing world is changing significantly, with ever-growing numbers of publications each year and shifting publishing patterns. However, the metrics used to measure academic success, such as the number of publications, citation number, and impact factor, have not changed for decades. Moreover, recent studies indicate that these metrics have become targets and follow Goodhart's Law, according to which" when a measure becomes a target, it ceases to be a good measure." In this study, we analyzed over 120 million papers to examine how the academic publishing world has evolved over the last century. Our study shows that the validity of citation-based measures is being compromised and their usefulness is lessening. In particular, the number of publications has ceased to be a good metric as a result of longer author lists, shorter papers, and surging publication numbers. Citation-based metrics, such citation number and h-index, are likewise affected by the flood of papers, self-citations, and lengthy reference lists. Measures such as a journal's impact factor have also ceased to be good metrics due to the soaring numbers of papers that are published in top journals, particularly from the same pool of authors. Moreover, by analyzing properties of over 2600 research fields, we observed that citation-based metrics are not beneficial for comparing researchers in different fields, or even in the same department. Academic publishing has changed considerably; now we need to reconsider how we measure success.



Best Practices for Best Practices

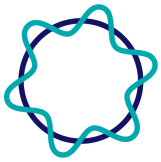
Stuart Geiger

This session will begin with a brief overview of BIDS's Best Practices in Data Science Series, which has been a pilot initiative over the past academic year. This has involved twice a month lunchtime discussions about a particular issue or problem in doing data science, in which notes are taken and written up as whitepaper reports (see tinyurl.com/bidsbp). Some of the participants in the series will share their experiences and lessons learned in both BIDS's specific effort and the broader idea of "best practices" or even just "good enough practices." We will then spend most of the time in an open discussion about both this specific effort and the idea of best practices in general. Possible topics of discussion may include: to what extent this effort could and should be scaled up to the other MSDSEs and beyond; related efforts and initiatives in this space (e.g. PLoS's 10 Simple Rules Series; The Turing Way; discipline-specific initiatives); potential future topics for sessions/papers; to what extent are there or should be generalizable best or good enough practices in data science; and how these efforts relate to other efforts in education and training, reproducibility and open science, and more.

Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening

Krzysztof Geras

We present a deep convolutional neural network for breast cancer screening exam classification, trained and evaluated on over 200,000 exams (over 1,000,000 images). Our network achieves an AUC of 0.895 in predicting whether there is a cancer in the breast, when tested on the screening population. We attribute the high accuracy of our model to a two-stage training procedure, which allows us to use a very high-capacity patch-level network to learn from pixel-level labels alongside a network learning from macroscopic breast-level labels. To validate our model, we conducted a reader study with 14 readers, each reading 720 screening mammogram exams, and find our model to be as accurate as experienced radiologists when presented with the same data. Finally, we show that a hybrid model, averaging probability of malignancy predicted by a radiologist with a prediction of our neural network, is more accurate than either of the two separately. To better understand our results, we conduct a thorough analysis of our network's performance on different subpopulations of the screening population, model design, training procedure, errors, and properties of its internal representations.



Collecting resources on best practices for scientific software development

Lindsey Heagy, Fernando Pérez

Software engineering skills are becoming a critical competency in many scientific fields, as more and more researchers are looking to develop open source software projects to advance research in their field. There are many good resources available for researchers to learn basic programming skills, but resources for going to the next level of developing a software package and growing a community are relatively scarce and scattered. In this BoF, we plan to draw upon the experience of participants to identify and prioritize topics for which educational resources should be developed and collect material which has been effective in workshops, HackWeeks, university courses, etc.. Finally, we will discuss how, collectively, we can assemble missing pieces, curate content, and develop shared resources that will serve the scientific community.

Special Interest Group in Quantitative Cell Biology and Communities (QCBC)

Joseph Hellerstein

Rapid progress in the quality and quantity of laboratory data on cellular processes (including low cost DNA sequencing and high throughput laboratory techniques) enable the development of quantitative models of cells and cell communities that can usher in a new era of materials, personal medicine, environmental remediation and more. We refer to this emerging area as Quantitative Cell Biology and Communities (QCBC). The QCBC special interest group at the University of Washington (UW) eScience Institute focuses on data management, processing pipelines, and modeling, and machine learning methodologies for understanding the biology of cells, viruses, and their communities.

What Systems Biology Can Learn From Software Engineering

Joe Hellerstein

Systems Biology is a sub-field of quantitative biology that focuses on reaction-based mechanistic models of biological systems. A grand challenge of Systems Biology to create a whole cell model (with experimental validations). For humans with 20K to 30K coding genes, ~10 pathways per gene, and 10 to 30 reactions per pathway, this means that a whole cell model contains around 4M reactions. In contrast, today's System Biology models range in size from tens to a few thousand reactions.

Systems biology is where software was in the 1950s. At that time, programs were tens of statements. Since then, computer scientists developed best practices and tools so that today's open source software is often tens of millions of statements.

The software experience has much to inform Systems Biology. For example, a core part of the exponential growth in software complexity is the ability to reuse software components written by others. Systems Biology has yet to develop techniques and practices that facilitate reuse.

Mapping Marine Picophytoplankton Biogeography using Statistical Learning

Corinne Jones

Phytoplankton, small photosynthetic organisms living near the surface of the ocean, form the base of the oceanic food chain and account for approximately half of the photosynthesis that occurs on Earth. They play a key role in the transfer of carbon from the atmosphere to the deep ocean. As such, changes in the abundance of phytoplankton species both affect and are affected by global warming. In order to better understand the impact of climate change it is therefore important to study the community structure of phytoplankton over large areas and how this structure varies over time. Existing methods for examining community structure typically use small-scale datasets and rely on manually identifying the phytoplankton species. Unfortunately, such methods are insufficient for mapping the community structure over large regions of the ocean. In this work we use data from a flow cytometer called SeaFlow, which continuously measures the size and composition of individual picophytoplankton particles during research cruises. This data possesses a novel structure, and we propose a statistical method that leverages this structure to effectively detect community shifts along individual cruises without labeling the particles. We examine the nature of the community shifts detected and analyze the relationship between the locations of biological shifts and similarly detected shifts in the physical environment.

Using Data Science to Understand the Film Industry's Gender Gap

Dima Kagan, Michael Fire

Data science can offer answers to a wide range of social science questions. Here we turn attention to the portrayal of women in movies, an industry that has a significant influence on society, impacting such aspects of life as self-esteem and career choice. To this end, we fused data from the online movie database IMDb with a dataset of movie dialogue subtitles to create the largest available corpus of movie social networks (16,303 networks). Analyzing this data, we investigated gender bias in on-screen female characters over the past century. We find a trend

of improvement in all aspects of womens roles in movies, including a constant rise in the centrality of female characters. There has also been an increase in the number of movies that pass the well-known Bechdel test, a popular---albeit flawed---measure of women in fiction. Here we propose a new and better alternative to this test for evaluating female roles in movies. Our study introduces fresh data, an open-code framework, and novel techniques that present new opportunities in the research and analysis of movies.

Intro to Julia

Stefan Karpinski

Julia combines the ease-of-use and productivity of Python with the speed of C or C++. It also offers 21st century programming language features, including:

1. multiple dispatch—like classes but better!
2. modern Unicode support that never throws errors on bad data
3. fast, easy multithreading with the same programming model as Go

If you've been curious about Julia and want to find out more, here's your chance! Stefan Karpinski, one of Julia's co-creators will give a gentle introduction to the language.

Unifying deep learning with item response theory: interval measurement, debiasing, efficiency, and explainability

Chris Kennedy

We propose a novel method for measuring complex social phenomena through the unification of the Constructing Measures approach to Rasch item response theory (IRT) with deep natural language processing (BERT, XLNet, etc.). Our approach establishes a new way of viewing supervised machine learning as an extension of psychometric measurement theory, in which algorithms learn to rate training observations on a labeling instrument by emulating human comment reviewers. This IRT methodology fits naturally into a multi-output, weight-sharing deep learning architecture in which our theorized components of hate speech provide a supervised shortcut structure for the neural network’s internal representation learning, improving sample efficiency and promoting generalizability. Built-in interpretability is an inherent advantage of our method, because the final prediction can be directly explained by the predictions on the 9 constituent components. We further show that the traditional reliance on the inter-rater reliability of a training dataset as a quality indicator is flawed and can be superseded through faceted partial credit modeling, which estimates and corrects for arbitrary rater bias in the labeling process - to our knowledge an adjustment never before implemented in computational text

analysis but critical for algorithmic fairness. We demonstrate our method on a new dataset of 50,000 online comments sourced from YouTube, Twitter, and Reddit.

Reflections on machine learning for (bio)acoustics

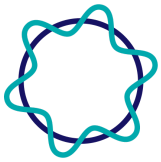
Justin Kitzes

As I've discussed at several previous DSE summits, our group has been hard at work on collecting and classifying bird songs using audio recorders. Here I present some reflections on our attempts to date, focused specifically on current challenges surrounding the use of machine learning models in applied domain science. I will (very briefly) touch on four key challenges: the difficulty in choosing between many different NN architectures, the practice of data augmentation, classification of simple but repetitious events, and communication of uncertain results.

Gradient Group Lasso Identifies Sparse Functional Basis for Molecular Manifolds

Samson Koelle

We present a method for analyzing low-energy paths between molecular conformations by combining techniques in both manifold learning, which identifies such paths, and functional regression, which can parameterize them by explanatory non-linear functions. Unsupervised manifold learning approaches are useful for understanding molecular dynamics simulations since they disregard small-scale information such as peripheral hydrogen vibrations that can nevertheless drastically affect the observed energy. However, understanding the role of covariates such as bond rotation in determining the energy landscape is made difficult by non-trivial data topology and geometry. In order to deal with these difficulties, we regress gradients of embedding coordinates on functional covariate gradients, and use a group-lasso inspired penalty for inducing sparsity. Differentiation of functional covariates is done automatically, while embedding gradients are estimated. This method replaces visual inspection for determining which bonds describe the slow dynamical modes of small molecules.



Career Paths in Data Science

Michael Laver

This is a discussion draft of a "white paper" on lessons learned in the partner institutions from five years of the MSDSE about careers in data science. The white paper will be an important "deliverable" for funders, and a full and frank discussion of this draft by as many participants as possible will surely make it stronger.

Managing a Productive Data Science Team

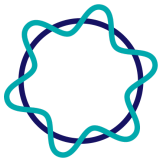
Ciera Martinez

Handling and managing data on your own is challenging enough, some might argue data science is never performed alone, so what are the effective strategies to overcome the challenges of performing data science with others? Data Science teams in academia are diverse, ranging from a research group to open source projects and everything in between. And every team is composed of a diverse group of individuals with diverse skills, needs, and personalities. We will spend this session discussing the overlapping strategies of what works and what doesn't work in team management with these panelists which represent a spectrum of team sizes, types, and experiences. We will be talking about challenges, workflow strategies, establishing and fostering community, data standards, and effective tools. Come with questions and your own unique experiences.

The side project you love and ignore

Ciera Martinez

Everyone has them, the project you daydream about blowing off Saturday night plans for, but that Saturday never comes. Maybe you even figured out the perfect name for this project and started the Github repo for it, only for the project to haunt you with no commits for months or even years. Let's dust off all these projects and let them shine. Come with a brief (5 min max) description of the project, the Github repo link (optional), and why you just can't seem to get it out of your mind. All project types welcome even if it has nothing to do with your main research focus or expertise - the more distant the better! Who knows, maybe you will find a kindred spirit that has insight into valuable next steps or even a new collaborator.



Reproducible Open Benchmarks for Data Analysis

Heiko Mueller, Irina Espejo, Kyle Cranmer

We demonstrate the Reproducible Open Benchmarks for Data Analysis Platform (ROB). ROB is an experimental prototype for enabling community benchmarks of data analysis algorithms. The idea is to allow participants to evaluate the performance of their algorithms in a controlled competition-style format. In ROB, the benchmark coordinator defines a workflow template along with input data. The template contains placeholders for workflow steps that are implemented by the benchmark participants (e.g., by providing Docker containers that satisfy the workflow steps). The ROB backend processes workflows on submission. Execution results are maintained in a database. The ROB user interface allows participants to submit new benchmark runs and to view the current leader board for the benchmark.

Scoping Research Engineer work

Andreas Mueller

The role of "Research Engineer" is still not very established and even though the MSDSE did some pioneering work,I don't think a clear definition emerged yet. I will briefly discuss questions of funding, scoping, evaluating and incentives for these positions.

Is academic data science diverse? Data from 42 data science centers, institutes and think tanks

Laura Norén

Using data from 42 academic data science centers and non-profit think tanks, we see that academic data science has shown a strong commitment to some kinds of diversity, but not others.

Virtually all data science centers talk about the importance of disciplinary diversity and many have a bimodal disciplinary distribution with a cluster around computer science, math, engineering, stats and physics and a secondary cluster either around the natural sciences or the social sciences. Most centers and institutes do not have all three disciplinary clusters; some have clusters in medicine, business, or law. There is vanishingly little inclusion of traditional humanities scholars beyond a handful of technical artists.

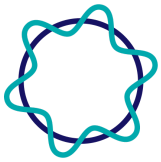
In terms of gender and race, data science is looking rather pale (72% white) and male (74% men) even when it is drawing from disciplines that are themselves better represented by women and racial minorities.

As an emerging professional field, the choices that are made now may become path determinant for future choices. Taking a look at which disciplines are represented and which people have a seat at the data science table now may help us make informed choices as we continue to admit students, make hiring decisions, and form collaborative research teams.

Wiki-Atlas: Notable Outdoor Site Recognition using Deep Learning on Mobile

Anastasios Noulas

Wiki Atlas, is a web platform that enables the exploration of Wikipedia content in a manner that explicitly links geography and knowledge. To explore content, users can click and interact with three dimensional polygon structures, each corresponding to a wikipedia article embedded in geographic space. As a use case example, let's imagine a user who is eager to explore museums in the city of London. By setting 'London' as the city of interest in the location search container available on the map and typing the 'Museum' search keyword, they can view all museums in the city that are featured in Wikipedia. Using Mapbox as the primary cartography platform for development, the Atlas offers global coverage with users being able to explore geo-tagged Wikipedia content wherever it is present around the world. Our first prototype of the altas features almost 1.2 million articles of the English Wikipedia, as well as five other popular languages and expanding. Moreover, we are in the process of integrating new tools for personalisation that, for example, will allow users to save lists of articles (“knowledge lists”) in order to view them at a later time or share them with peers. Besides being a general knowledge exploration tool, we aspire that Wiki Atlas also evolves to a knowledge discovery and learning platform that will help educators and students alike to systematically acquire knowledge through a fun and interactive medium. In this presentation, we will walk through a demo of the tool and the technologies that were used to build it, aiming at receiving feedback from the community in terms of improvements as well as possible future directions and use cases that could be enabled through the platform. Moreover, we will take some time towards the end of the presentation to showcase a mobile version of the Atlas that is currently under development at New York University and involves the geographic discovery of knowledge through Augmented Reality and Computer Vision technologies. We will discuss the possibilities and opportunities for novel way not only to acquire, but also contribute media other content, through new immersive mediums.



Docker for Research and Pedagogy in Data Science

Sang-Yun Oh

Managing software installations can be a complex and time-consuming task. Docker software can simplify the process of installing, documenting, and deploying your computational environments. This tutorial will guide you through customizing and deploying your computational research environment with Jupyter notebook and R studio.

Data-driven de-identification of clinical notes

Vikas Pejaver

Unstructured data in clinical notes are estimated to represent about 80% of the clinically and scientifically relevant information present in an electronic health record (EHR). Despite their high value to the research community, the sharing of clinical notes is non-trivial due to the presence of highly private and sensitive patient information. While several approaches have been proposed for the automated de-identification of clinical notes, existing methods largely focus on the removal of 18 identifiers defined by Health Insurance Portability and Accountability Act (HIPAA), such as name, social security number, contact information, among others. However, it is well known that patients can be re-identified from other information not covered by HIPAA identifiers, such as the occurrence of a particular disorder or combination of symptoms. Here, we develop a method for the redaction of potentially identifiable phrases from clinical notes based on their frequencies stored in a centralized repository of phrases. Preliminary results indicate that this simple approach yields reasonable sensitivity in redacting HIPAA identifiers while leaving two-thirds of the original note unredacted. We also propose the design and implementation of a cloud-based API service to support the secure sharing of phrases to enable the construction of the centralized repository of phrase frequencies across different institutions. We anticipate that such a frequency-based redaction platform will serve as a baseline method upon which more sophisticated de-identification methods can be built and provide an objective metric for data governance.

Transitioning Open Source projects to a self-sustaining organizational model

Matti Picus

Over the past year, and for the first time since its creation, NumPy has been operating with dedicated funding. This move has allowed the project to address long standing technical debt,

improve engineering infrastructure, and implement significant new features that benefit the entire ecosystem.

We view funded contributions as an inevitable developmental stage in the natural progression of a mature Open Source project.

We will lead a breakout session to share our experience and explore with others how they view the evolution of their own projects.

A Nonconvex Approach for Exact and Efficient Multichannel Sparse Blind Deconvolution

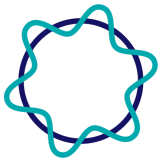
Qing Qu

We study the multi-channel sparse blind deconvolution (MCS-BD) problem, whose task is to simultaneously recover a kernel and multiple sparse inputs from their circulant convolution. We formulate the task as a nonconvex optimization problem over the sphere. Under mild statistical assumptions of the data, we prove that the vanilla Riemannian gradient descent (RGD) method, with random initializations, provably recovers both the kernel and the signals up to a signed shift ambiguity. In comparison with state-of-the-art results, our work shows significant improvements in terms of sample complexity and computational efficiency. Our theoretical results are corroborated by numerical experiments, which demonstrate superior performance of the proposed approach over the previous methods on both synthetic and real datasets.

Nonconvex approaches for sparse deconvolution problems

Qing Qu

In this work, I will present our advances in recovery guarantees for solving sparse (blind) deconvolution problems, and introduce efficient nonconvex optimization methods that solving the problem to global optimality. We demonstrate the practicality of our nonconvex methods on several computational imaging problems, such as super-resolution microscopy imaging and calcium imaging, etc.



Complex ideas for data science and data science for all

Andy Rominger

The goal of complexity science is to dive into the gaps between traditional disciplines and seek new knowledge that reflects the fluid and dynamic world and cosmos we live in. Data, at new levels of heterogeneity and size, have been critical to allowing complexity science to flourish. The Santa Fe Institute and its researchers have found their home in the data rich gaps between disciplines and I will share vignettes of their work, from understanding opinion formation through combining statistical physics and online discussion boards, to peering into the origin of life with chemistry and biology, to studying biological diversity and its fate in the Anthropocene through the lens of information theory and ecology. Along the way we will highlight the data science tools that have enabled this research and the current limitations that present opportunities for development. Key among the current limitations, I will argue that limited inclusion in data science holds back the movement, and I will present ideas for dialogue about development on this front.

apricot: Submodular selection for summarization of large data sets

Jacob Schreiber

Recently, machine learning practitioners face the problem of having too much data. The problem typically manifests in long training times for machine learning models, or an infeasible number of points to visualize in a single panel. These problems can pose serious problems for the development cycle of new approaches.

Submodular selection offers a solution to the problem of having too much data by reducing it down a representative subset. This subset can be used to train machine learning models with higher efficiencies than using the full data set and can result in much higher accuracy than using a random subset.

apricot is a Python package that implements submodular selection using the API of a sklearn transformer. In this talk, we will demonstrate how to use apricot to select a subset of points to train machine learning models. Additionally, we will talk about the different types of submodular functions and what types of datasets they each work best on.

Academic Data Science Alliance

3D Organ Shape Reconstruction from Topogram Images

Elena Sizikova

Automatic delineation and measurement of main organs such as liver is one of the critical steps for assessment of hepatic diseases, planning and postoperative or treatment follow-up. However, addressing this problem typically requires performing computed tomography (CT) scanning and complicated postprocessing of the resulting scans using slice-by-slice techniques. In this paper, we show that 3D organ shape can be automatically predicted directly from topogram images, which are easier to acquire and have limited exposure to radiation during acquisition, compared to CT scans. We evaluate our approach on the challenging task of predicting liver shape using a generative model. We also demonstrate that our method can be combined with user annotations, such as a 2D mask, for improved prediction accuracy. We show compelling results on 3D liver shape reconstruction and volume estimation on 2129 CT scans.

Learning on the job at a hypergrowth unicorn

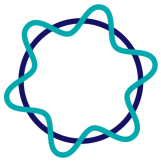
Allison Smith

I have worked at a rapidly growing startup for over a year, and I'm still learning every day. However, everyone at a startup is learning because new challenges are constantly developing. I'll focus my talk on the following questions: What is data science at a startup? What skills are needed for data science at a startup? How does data science at a startup grow? How to get a job in data science at a startup?

Building Local Research Communities through Special Interest Research Groups

Valentina Staneva

In an attempt to expand the influence of MSDSEs across campus various Special Interest Research Groups have been created which unify researchers from different departments around common research interests, data types, programming tools, etc. A few examples at University of Washington are Neuroinformatics, Satellite Image Analysis, Text as Data, Computational Demography, Python for Geosciences, Data Science Studies Groups. We believe they exist in all sorts of forms and shapes across the institutions.



Echopype: Enhancing the Interoperability and Scalability of Ocean Sonar Data Processing for Biological Information

Valentina Staneva

Instrumentation advancements in the last decade have produced a deluge of ocean sonar data. These data provide opportunities to study the marine ecosystems at unprecedented spatial and temporal scales. However, to date, these data remain significantly under-utilized, mainly due to the lack of data interoperability and scalable analysis workflow. We address these challenges by developing an open-source Python package, echopype, which defines an interoperable netCDF file format and leverages the power of Xarray and Dask for explicit and efficient sonar data analysis in the Jupyter environment.

Investigating and Archiving the Scholarly Git Experience

Vicky Steeves

A requirement of the MSDSE is that any programming done under its auspices must be open source. For most if not all, that means putting a source code repository on a Git hosting platform like GitHub, GitLab, or Bitbucket. But these platforms make no guarantee to keep your work around forever, and the toolkits to get folks to use any sort of version control is challenging at best. As such, NYU Libraries is working on *Investigating & Archiving the Scholarly Git Experience*, an Alfred P. Sloan funded project that addresses the need to investigate how academics are using Git hosting platforms, how they can be adapted to academic needs, and how the scholarship hosted on them can be archived. The results of this project aim to inform the way code and annotations on Git hosting platforms move from a phase where they are highly active and collaborative, to a state where they are stable, permanently citable, and under active, professional preservation, along with their contextualizing information (e.g. pull requests threads, issue discussions, wikis, etc.). Vicky Steeves, Project Lead, will debrief about the projects' preliminary findings and efforts.

Diversity and Inclusion in Data Science

Sara Stoudt

This session will be focused on sharing the BIDS Diversity and Inclusion working group’s lessons learned from organizing "Fostering diverse and inclusive data science at Berkeley: a series for underrepresented undergraduate students” and providing an opportunity for others to share information from their own diversity and inclusion initiatives. The session will be

discussion based with the goal of brainstorming ideas for new events and programming that promote explicit support of diversity and inclusion at our respective institutions and beyond.

Data Science Capstone Research Programs

Anthony Suen

Discussion of various data science for good or capstone programs around the country. It will look at the lessons learned and strategies for scaling these student research programs.

An artificial neural network controller for insect flight

Callin Switzer

Insect flight is a highly non-linear dynamical system. As such, strategies for understanding its control have typically relied on either simulation methods (e.g., Model Predictive Control (MPC), genetic algorithms) or linearization of the dynamical system. Here we develop a new framework that combines MPC and deep learning to create an efficient method for solving the inverse problem of flight control. First, we used a feedforward, fully connected artificial neural network to answer the question, “What is the temporal pattern of forces required to follow a complex trajectory?” Combining neural networks with simulations based on dynamical systems models yields a data-driven controller where the data are derived from a non-linear physical model. We trained a network (4 hidden layers, with hundreds of nodes) on ~9 million simulated 2D insect trajectories. Our network accurately predicted the force, force angle, torque, and tangential and angular velocities (7 outputs), when it was provided with initial conditions and a goal location (10 inputs). The coefficient of determination (r^2) for all predictions was > 0.98 on a test dataset (~1 million additional trajectories). Second, we pruned the network by removing weak connections. Only ~20% of the original weights were necessary to produce comparable results to the fully connected network. Overall, this work shows that sparsely connected artificial neural networks may be an efficient approach for controlling nonlinear dynamical systems.

The Pangeo Project

Amanda Tan

The Pangeo project is a community that fosters open, reproducible workflows by linking scientists with software and computing infrastructure. While started by Earth scientists, researchers in other disciplines have started to use this infrastructure for a variety of investigations. The Pangeo community is now well established, with recent successes in deploying supporting infrastructure on Cloud and HPC systems and in demonstrating Pangeo’s

capabilities to support a diversity of collaborative research applications. These include individual research efforts, as well as deployments for use in classes and hackweeks. In this session we invite feedback from the community to learn more about potential use cases and applications. We are especially interested in identifying key architectural components to the system (for example, Kubernetes, JupyterHub, Binder), what currently works well and what could be improved. We are also interested in contrasting use by individual researchers compared to educational group settings.

Open Long-Tailed Recognition in the Real World

Stella Yu

Real world data often have a long-tailed and open-ended distribution. A practical recognition system must classify among majority and minority classes, generalize from a few known instances, and acknowledge novelty upon a never seen instance. The Open Long-Tailed Recognition task is to learn from such naturally distributed data and optimize the classification accuracy over a balanced test set which include head, tail, and open classes. The key challenges are how to share visual knowledge between head and tail classes and how to reduce confusion between tail and open classes. This session introduces some of the latest advances on this topic in computer vision, and invites data scientists to share their problems, techniques, observations, and insights.