Data Science @ SIO

October 15, 2017

Peter Gerstoft, Bruce Cornuelle, Steve Diggs, Scott Sellars, George Sugihara, Guy Masters

Summary

The Scripps Institution of Oceanography (SIO) welcomes the creation of the Halicioglu Data Science Institute (HDSI), an inclusive campus-wide data science institute at UCSD.
SIO is a world leader in Earth environmental sensing, modeling, and data analysis, with emphasis on collection and interpretation of large, real-world data sets for advancing basic science as well as informing environmental policy and improving public health. With growing awareness of the interdisciplinary breadth required for this mission, SIO welcomes the chance to connect across various UCSD departments participating in the HDSI. Further, we believe that acquiring additional expertise in data science for our researchers and students is essential to maintain and accelerate SIO's preeminent role in basic research, and education in environmental and earth science. In turn, SIO can provide high profile accomplishments to help establish a UCSD brand in data science.

Science in the 21st century is being driven and enabled by the explosion of data and data-driven analysis techniques in all fields of science and engineering. These developments allow researchers to address ever more complex and ambitious problems, and to validate results by real world prediction. However, the solutions to such problems require holistic, systematic frameworks and interdisciplinary collaboration. While the last century made great strides with reductionist and specialized discipline-specific approaches (e.g., significant progress in biology gained from idealized simple experimental systems, and in atmospheric science from parameterized first principle physics models), the emerging hallmark of the current century involves addressing complex problems that cannot be completely abstracted to first principles and that require interdisciplinary convergence. Thus, as a conduit for collaboration, a UCSD data science institute has the potential to be institutionally transformative.

In summary, we wish to present ourselves as solving critical problems in the field of data science while at the same time harnessing emerging data science methodologies to address pressing global environmental, health and social needs. Thus, we have identified some specific sets of problems and data with potentially signature applications that we see as best communicating our unique potential. We believe that these problems and applications would help UC San Diego create a brand in data science.

0) Background

The HDSI will provide support for the incorporation of data science practices with SIO's mission. Based on current information, HDSI will have a large endowment, and UCSD will further provide about 12 research FTE and 10 endowed chairs with main offices in Atkinson Hall. It is envisioned that HDSI scientists will be at three UCSD locations, with one satellite location at SIO. It is expected that HDSI will be a catalyst for data science carried out at SIO. HDSI will be an independent institute reporting directly to the Chancellor with no academic affiliation, however, teaching will be a significant component of HDSI.

First, we summarize SIO's current data-science status (Sections I—III), and then give our vision for HDSI in Section IV.

I) Education

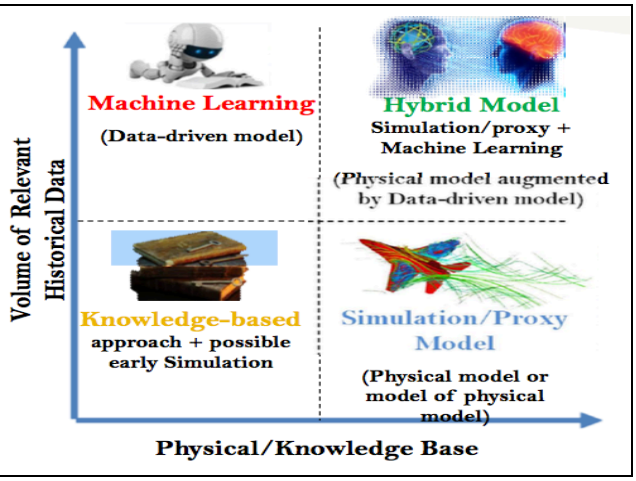
SIO is focused on educating researchers in many fields of earth sciences. The current curriculum needs additional offerings to train students to use data science tools. New undergraduate and graduate data science courses, with discipline-specific data and emphasis on statistical and computational approaches in data science, must be developed. While nearly all courses at SIO are focused on data, there are currently few dedicated data science classes.

Many graduate students from SIO have obtained jobs in the data science industry. Thus, both for SIO's institutional mission and to prepare graduates who choose to enter the data science industry, it is important to collaborate with the HDSI, to develop our SIO curriculum for the institutional mission. The graduate students will then be better prepared for these jobs.

II) Data science analysis efforts at SIO

In the first parts of this section we focus on machine learning, which explores the study and construction of algorithms that can learn from and make predictions from data. Such algorithms are solidly based on statistical theory (Bayesian and non-Bayesian), and combine statistics with optimization to deliver algorithms for learning from data.

Fig 1: Machine learning is data-driven, and obtains models from potentially large data sets. The physical modeling depends on our understanding of the domain knowledge. Numerical weather predictions are an example of a simulation/proxy model, currently built on physical principles. We would like to see a hybrid model where a physical model is augmented by a data-driven model.



Physical (and chemical or biological/ecological) models and hypotheses are central to most work at SIO. Earth science data sets tend to be poorly sampled, noisy, and incomplete, and are often difficult to use in standard machine learning algorithms. It would be transformative if we could develop a modeling framework that combines machine learning and physical modeling. This would provide a means of specifying a physical model as a component of the machine learning or a means of using machine learning to train better physical parameterizations in earth system/biological/chemical models. Transparency of the learned algorithms would enable human learning and allow validation by testing for physical consistency, just as algorithms used in social problems (e.g. parole) should be tested for ethical consistency. Determining how

machine learning best can be used in the earth sciences requires multi-disciplinary collaboration and could be a major priority for the HDSI.

(1) **Specific example: Numerical weather prediction**

Several groups at SIO are studying atmospheric dynamics. Specifically, the Center for Western Weather and Water Extremes (CW3E) is working towards prediction of atmospheric state from days to months as a tool to understand the precipitation impacts of Atmospheric Rivers (ARs). These highly-intermittent events supply more than half of California's rain, and show strong inter-annual variability. California's wet and dry years can be traced to the frequency and duration of ARs, and improvements in their prediction would be hugely beneficial to California and the other Western States.

Accurate long-range (one month lead time or more) precipitation forecasts for California have continued to be elusive, with little advancement or increase in accuracy since the 1920's. These predictions are critical for planning water allocations, reservoir operation, and preparation for possible floods and drought. These forecasts are an area where collaboration with the HDSI could provide real-world impact. Currently, machine-learning projects in weather forecasting are underway at SIO for precipitation to soil moisture, run-off, and river flow predictions.

(2) **Data assimilation, state estimation** (MAE and ECE Collaboration)

Data assimilation as used at SIO means fitting models to observations, both as a way of doing a physics-based reanalysis (mapping) of multivariate observations and as a test of the model as a hypothesis. Machine learning has tended to emphasize the supervised learning popularized by artificial neural networks but unsupervised and semi-supervised learning algorithms could be more appropriate. Though it may be a challenge to incorporate machine learning with highly-structured physical models, successfully doing so would prove tremendously rewarding.

Models for all components of the Earth system are in rapid development, including coupling between components. Ocean models have been coupled to atmospheric models, ice models, biogeochemical and ecosystem models, land models, and even human impacts, such as agriculture. Physics-based models might have advantages compared to purely statistical models as they are constructed from basic principles, independent of observations. They contain important limits on possible system behaviors, and their testing improves our understanding of the physical processes. Such models might not be able to predict data not included in the fitting set (e.g. the failure of climate models to predict on decadal and subdecadal time scales). This underscores that despite the established physical mechanisms, such models are at some level simply hypotheses (e.g. scaling assumptions, lack of connectivity etc.). Evolution of the underlying system, e.g., climate change, might erode a stationary statistical model built on old observations. However, physical models are hugely complicated, costly to run, and testing them with observations is a challenge.

A large volume and variety of observations are critical for model testing and improvement, but are challenging to use. Modern data assimilation and state estimation methods address this problem, making the most efficient use of the observations by taking full advantage of the

dynamical constraints that models supply while also keeping track of errors. Present state estimation techniques match the model evolution to multivariate observations by adjusting model control parameters, such as initial conditions, boundary conditions, forcing, and other modeled processes, e.g. turbulence. Both the control adjustments necessary to match the observations and the residual observational misfits are examined for consistency with expectations and as a tool for discovering poorly-modeled processes.

(3) **Machine learning in physics** (ECE Collaboration)

Given the vast scale of the Earth, oceans, and challenges of obtaining direct measurements of their properties such as rock structures or temperatures, earth scientists and oceanographers rely upon acoustic and seismic signals to infer such properties. Acoustic and seismic signals are used to infer properties of the Earth and oceans. These signals, which are generated by natural sources as earthquakes and storms (or man-made sources), travel great distances, acquiring environmental information in the process. When these waves are recorded by arrays of sensors, an inverse problem can estimate source location, and ocean/earth structures and properties. These inverse problems are critical for advancing our understanding of the structure of the Earth and oceans, and climate dynamics.

The focus in this topic is to work more broadly with machine leaning techniques, adapt and appreciate these so that they can be used in physical problems. It thus includes geophysical inverse problems. Many inverse problems include sequential estimation, which is a generalization of item 1 and 2.

Efforts have been undertaken at SIO to develop machine learning tools for generating data-driven wave propagation models, with the aim of providing enhanced prediction and inference capabilities over more conventional inversion. These methods show great promise in enhancing our ability to extract information from these useful signals.

Examples where such methods have succeeded in ocean acoustics include: source localization in an ocean waveguide using supervised machine learning, use of compressive sensing in array processing, and dictionary learning of ocean sound speed profiles. In seismology developments include hierarchical clustering for earthquake relocation, supervised learning to improve data-driven ground motion and hazard analysis, and graph-based signal processing for weak source localization within seismic networks. Infrasound data is used to detect bolides and other atmospheric events, on very long period data to detect atmospheric gravity waves at the Earth's surface, and on seismic data to detect fracking and other small events. In oceanography, improved mapping with random forest of sparsely sampled biogeochemical variables from the SIO Argo floats, covering the world ocean.

(4) Biology**, complex systems in nature** (Economics and Psychology Collaborations)

Many complex natural systems such as those found in biology are driven, not by a few factors acting independently, but by complex interactions between many components acting together in time –nonlinear dynamic systems. Physical science notwithstanding, nonlinearity in living systems means that its parts are interdependent. They interact, and as a consequence associations

(correlations) between them will change as the overall system context (state) changes. This sets up a 300-year-old problem regarding how to identify causal variables. The typical assumption that correlation is necessary can be shown to mislead. Causal variables that are uncorrelated with their effects are invisible to current methods used to construct causal networks (such as bioinfomatic analysis). This problem is cuts across many scientific disciplines and is fueled by observational data.

There is an active effort at SIO to address this problem by exploiting interdependence – the fact that nature is complex – that is based on the concept that if states of the driving variable can be recovered from the time series of the affected variable, causation is established. Efforts such as this are being fueled by the growing availability of large amounts data in all fields, that are enabling studies of systems in their natural complexity – an enterprise requiring a more inclusive view to produce models that predict. SIO is collaborating with groups on this in areas as diverse as gene expression, chemical limnology, astrophysics, and forecasting dengue epidemics.

(5) Data Curation and Rescue (Library Collaboration)

SIO maintains several regional, national and international-scale data facilities, and has diverse experience with data from biological, physical, chemical, and geological sensors and models. Data are at the heart of UCSD's research outputs and discoverable, reusable data are a critical component of both reproducible science and scholarly communications. Geoscience has seen an explosion of data from *in situ* sensors on autonomous platforms in addition to petabytes of model data. Acquiring earth science data often comes with a high deployment and acquisition price, especially in remote and/or polar regions. Unfortunately, many datasets have large spatial and temporal gaps rendering them unfit for use in certain types of analysis.

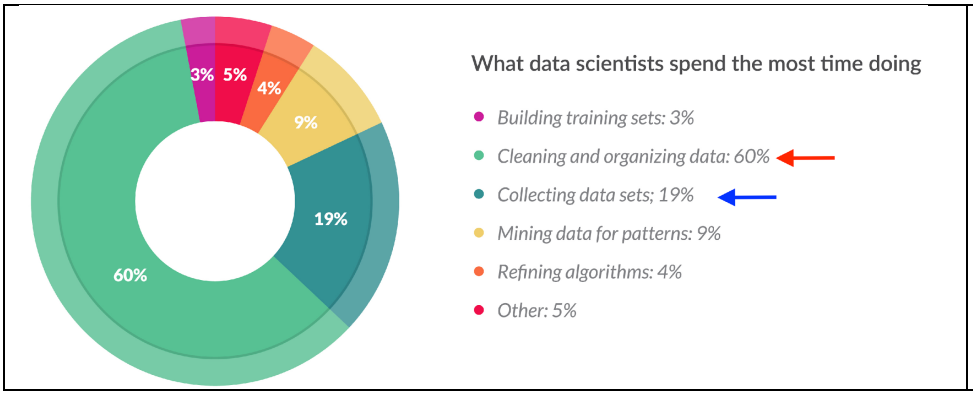


Fig 2: A recent study found that data scientists spent almost 80% of their time collecting, cleaning and organizing data. (source: FORBES 2016-03-23)

Recent surveys have brought attention to the fact that many researchers spend up to 80% of their time discovering, cleaning and organizing their data, long before they can begin extracting knowledge from the numbers. In addition, the relative paucity of uniform data standards often impedes the scientific workflow, consuming untold man-hours in the process of exploring certain types of pressing environmental problems. Students and researchers who could better use their time to creatively understand and analyze data, spend most of their time and energy simply trying to locate and reformat relevant data sets.

UCSD's HDSI should play a central role in the creation of domain specific as well as interdisciplinary full-stack data curation infrastructure that:

- protects the University's investment in the creation of these data
- makes these data **FAIR** (**F**indable, **A**ccessible, **I**nteroperable and **R**eusable)
- is responsive to private, state and federal funding agency data-related regulations
- builds and disseminates expertise in rescuing and recovering both grey and dark data

State-of-the-art data curation must reduce the "time to science" and increase publication, and should therefore be a component of the HDSI. There are related initiatives within the UCSD Libraries, thus data curation through the HDSI can leverage existing funds, personnel and expertise, thereby reducing the resources needed. Most successful data curation programs combine deep digital curation expertise, such as the library provides, with a deep knowledge of the sources, characteristics, and uses of the data, which faculty and staff from the SIO data facilities can bring. This is a strength of the proposed HDMI cross-department framework.

III Suggestions for HDSI

We agree with the overall vision set forth by the UC San Diego Data Analytics Initiative Working Group Statement of Principles (Nov 28, 2016) and the Division of Physical Sciences white paper (April 6, 2017), in addition:

Teaching and research are primary missions at UCSD, leaving an enduring impact. Thus, contributing to student training would be important for HDSI. Most students are rooted in a specific domain in the physical sciences. The HDSI should develop an innovative education framework that allows for students to learn data science, computer science, machine learning, as well as their chosen domain. Recommendations include:

- Encourage communication and collaboration across UCSD.
- Short courses for UCSD faculty, staff, and students.
- Boot camps in data science applied to the physical sciences, including machine learning packages
- Secure and offer online material for education on data science.

We are enthusiastic about including a data-science emphasis in the SIO PhD degrees.

We encourage collaboration between domain scientists and data-scientists. Basic collaboration between scientists at SIO and HDSI should naturally be encouraged. This would lead to improved research output and strengthen research proposals. Many of the basic problems are similar within diverse fields. Thus, there should be a wider collaboration than just between two PIs, preferably across several departments at UCSD.

- Have faculty with split appointments in a domain science institute as well as at the HDSI.
- Post Doc or similar shared personnel between with multiple PI is a good way to get a collaboration started.
- Giving seed funding for graduate support for graduate students based on collaboration between three PIs on a data science problem. The UCSD Graduate Division's "Interdisciplinary Collaboratories" could be a model for this.