POGO 2019 Workshop on Machine Learning and Artificial Intelligence in Biological Observations

Tutorial leaders

Danelle Cline, Monterey Bay Aquarium Research Institute, ***Acoustics***
Tristan Cordier, University of Geneva, ***Genomics***
Anders Lanzén, AZTI-Tecnalia, ***Genomics***
Eric Orenstein, UC San Diego/Scripps Institution of Oceanography, ***Imaging***
John Ryan, Monterey Bay Aquarium Research Institute, ***Acoustics***
Simon-Martin Schröder, University of Kiel, ***Imaging***

Introduction

The past two decades have seen rapid advances in technology available to oceanographers seeking to study and manage marine ecosystems. Relatively cheap, compact computers and digital storage have allowed scientists to collect big, complex datasets. Cruises now regularly return to port with terabytes of data, high temporal resolution coastal time series contain billions of measurements, and water samples are parsed into millions of DNA sequences. These information rich datasets have grown so large that analysis with traditional methods has become untenable.

Oceanographers have begun exploring high-throughput, automated methods to make sense of their big datasets. Recent developments in machine learning and artificial intelligence (ML/AI) offer the means to analyze a variety of types of data; acoustic recordings, digital images and video, and eDNA samples to name a few. ML/AI techniques, when properly applied, could be used to expedite analysis of existing oceanographic data and enable novel experimental designs.

The POGO Biological Observations Working Group recognizes an immediate need for the implementation of automated workflows for biological oceanographic observations. In an effort to kick start development and community-wide discussions, the Working Group proposed hosting a workshop on ML/AI. After a year of organization, the 2019 POGO Workshop on Machine Learning and Artificial Intelligence in Biological Observations took place from May 20-22, 2019 at the Flanders Marine Institute in Ostend Belgium.

Six tutorial leaders from four countries developed content to teach oceanographers how to develop and deploy ML/AI algorithms. 41 participants representing 31 institutions from around the globe spent three days working with cutting edge techniques. The attendees used computing resources courtesy of Amazon Web Services (AWS) to get hands-on experience with applying new ML/AI methods to ocean specific data.

This white paper summarizes the proceedings of the workshop, discusses improvements to the format, and makes recommendations for how POGO might engage with and facilitate ML/AI

in biological oceanographic observations. The report consists of three sections in the main body and several appendices:

Section I outlines the workshop structure, provides details of the topics covered, and discusses successes and short-comings.

Section II identifies critical areas for ML/AI capacity building for biological observations based on the experiences of the workshop participants.

Section III contains concluding remarks and discusses the long-term potential of ML/AI enabled observations.

Appendix I includes information and references to guide future discussions and workshop organization.

Appendix II contains the workshop documents: a detailed copy of the schedule and a list of participants.

SECTION I

The Workshop brought together 55 scientists from all over the world for a two-and-a-half day, hands-on training session with the latest ML/AI techniques. The participants applied to one of three sections: acoustics, genomics, and imaging. Each section was led by a pair of domain experts who developed instructional material hosted on AWS. Attendees were encouraged to load their own data onto AWS in advance of the workshop for experimenting with the techniques.

The Workshop schedule built in group discussions at the beginning and end of the workshop to solicit the opinions of the participants on ML/AI. The rest of the time was allocated for hands-on work in the domain specific groups.

TOPICS

Most of the Workshop was dedicated to hands-on material relevant to each domain area. Each tutorial section included example data to illustrate the methods with the coded examples.

Acoustics – The focus on this session was on acoustic preprocessing, and classification of cetacean sounds, specifically Blue whale A and D calls. Blue whale calls are stereotypical and relatively abundant, making them an excellent exemplar for the session. Although participants were working on a broad range of sounds, the methods outlined were generally applicable, and some reported immediate improvement using PCEN in their workflows. Due to the duration of the workshop, more in-depth topics like acoustic event detection were not covered. Topics included:

- Efficient decimation filtering
- Spectrogram generation: choosing the right color map and normalizing using per-channel energy normalization (PCEN)
- Classification using transfer learning with deep neural networks
- Bias versus variance and the importance of learning curves

Genomics – The use of ML techniques in genomics has so far focused mainly on technical tasks such as genome annotation, assembly or protein structure prediction. Here, we instead pioneered two tasks that are more directly relevant to marine ecology: 1) ecosystem impact assessment or monitoring using environmental DNA (eDNA) metabarcoding and 2) prediction of traits based on metagenome data. Metabarcoding targets specific gene markers to identify or group sequences taxonomically and can be applied to a wide range of life forms and environments, including ancient eDNA. It is relatively cost-effective and standardized, requiring less advanced treatment of generated sequence data, compared to metagenomics. Several participants applied the demonstrated methods to their own datasets during the workshop. The second part of the tutorial instead utilized the publicly available Tara Oceans dataset.

- Error sources and protocols in environmental genomics
- Supervised and unsupervised ML methods commonly applied in genomics

- Metabarcoding data treatment, clustering and taxonomic assignment
- Ecosystem impact assessment using a Random Forest Classifier trained on eDNA metabarcoding data
- Assembly and binning, basics of metagenomics data analysis
- Trait prediction using metagenome-assembled genomes (MAGs) based on predicted protein families (Traitar and GenePhene).

Imaging – The section focused on working with plankton image data. *In situ* plankton data is relatively easy to work with from a programmatic perspective as the images contain light foreground pixels on a dark background (or vice versa); ideal for illustrating basic image processing routines. All the techniques can be applied to other data types with modifications. During the workshop, several participants began porting these methods for use with benthic images and data from pelagic fish surveys.

- Image manipulation
- Region finding
- Hand-engineered feature extraction
- Ensemble and margin classifiers
- Feature extraction with deep neural networks
- Fine-tuning deep neural networks
- Data augmentation to boost neural network performance
- Cross domain classification with neural networks

AMAZON WEB SERVICES

The tutorial leaders used AWS cloud computing to streamline the learning experience for attendees. With AWS, the leaders set up computing environments that all attendees could access with minimal set up on their personal computers. AWS generously donated $50 credits for every participant, covering sufficient data storage and computing costs for the duration of the workshop. Each tutorial leader was given a $100 credit to provide extra development and testing time.

The Workshop primarily made use of three AWS resources: S3 Buckets for data storage, the SageMaker interface for computation, and CloudFormation for scripting the computing environment set. Each group of tutorial leaders uploaded data to S3, wrote instructional material on SageMaker, and used CloudFormation to give participants appropriate access to the materials. Together, these services allowed the attendees to quickly get started on running ML/AI tests.

INSTRUCTIONAL MATERIALS

During the workshop

The tutorial leaders independently developed instructional materials to guide the attendees through processing data with ML/AI. The teaching units were written in Jupyter Notebooks, a code interface that allows for in-line text formatting and interactive execution. With Jupyter, the participants were able to see step-by-step output from the code along with explanatory text written by the tutorial leaders. They were also able to manipulate the code to get a feel for how it operates and do exercises to solidify understanding.

The domain area leaders broke their material into several specific topics. Each unit included text explaining concepts and code that illustrated mechanics. Activities at the end of each unit were designed to solidify understanding by expanding upon the concepts and code laid out in the hands-on portion.

Access after the proceedings

All the tutorial materials produced for the workshop are currently available on GitHub[1]. Interested individuals can download all the code to their personal computer. Once setting up the appropriate computing environment, they can independently step through the lessons. The data provided by the tutorial leaders and the associated presentations will be available through the OceanTeacher platform. Further development of the materials will be necessary to make the materials into a fully cohesive course without the need for direct instruction.

SUCCESSES, FAILURES, AND IMPROVEMENTS

AWS

The tutorial leaders felt that AWS was generally a good platform for a hands-on workshop. The set-up scripts run in CloudFormation effectively insulated tutorial attendees from dealing with the minutia of setting up the appropriate computational work environment. It also gave the attendees easy access to more computing power than might have otherwise been available.

Future workshop organizers might consider directly contracting technical support from AWS or an equivalent service provider to ease development. The tutorial leaders benefited from the expertise of UC-San Diego's Research IT department, but still had a lot to figure out on their own. Having on-site support would expedite some of the inevitable troubleshooting that will arise during the workshop.

After the workshop, several attendees incurred extra charges from AWS related to active computing instances and storage. AWS waived the charges after the individuals contacted customer service. While the tutorial leaders reminded participants to shut down all their AWS resources at the end of the workshop, more needs to be done to ensure such unexpected

[1] https://github.com/eor314/pogo_bioobs

charges do not occur. This could be done by centrally administering participant accounts or including an explicit CloudFormation shutdown script.

Materials

The hands-on format was generally met with enthusiasm from the workshop participants. Getting started with machine learning requires overcoming a few conceptual and programmatic hurdles. The guided tutorials jump started the participants' facility with these tools by laying the techniques out step-by-step with prescribed data. This approach effectively illustrated how to design and implement a bespoke pipeline for one's data.

The tutorial leaders felt rushed on the two-day schedule. Indeed, all domain areas skipped prepared material to allow time for the participants to work with their own data. Future iterations of the proceedings should budget two full days for instruction and at least one full day for participants to experiment with their own data.

SECTION II

Several discussions were hosted outside of the tutorial sections to identify areas for growth and coordination. Three broad themes emerged that the attendees agreed would facilitate development at their home institutions and the broader community: storage and computation resources, community data standards, and enhanced coordination with computer scientists and industry players.

STORAGE AND COMPUTATION

Collaborative data repositories

Data pooling – There is ample evidence that pooling data of similar types enhances the discriminatory ability of a machine classifier. This is particularly true for acoustic and image data where procedures such as fine tuning can repurpose trained algorithms for new, similar data types. In this context, sharing data to the largest extent possible would speed progress in the community. Given sufficient resources, the development of a global, publicly available data repository would be ideal. A meaningful intermediate step would be producing accepted data standards to facilitate data sharing between organizations.

Storage resources – As observational methods become more efficient, the amount of data collected grows, necessitating expanded storage capacity. Developing long term storage options would assist researchers with expanding, long term data sets.

Storage standards – The community as a whole will soon need to grapple with questions regarding data access and longevity. Guidelines for how long data must be rapidly accessible versus in "cold storage" should be considered and issued as appropriate.

Computation resources

Access – Many researchers, especially those outside of Europe and the United States, are limited by access to computational power. Modern machine learning techniques require computations that might be untenable on a personnel computer. Securing access to computing resources will be beneficial to many projects seeking to monitor ocean ecosystems. A grant program supplying access to computation time on remote resources would help practitioners with lots of data jump start analysis. This could be best accomplished by establishing a relationship with an existing super computer or exploring developing an international center for oceanographic data processing.

There is precedent for such repositories and standards as laid out in the FAIR Guiding Principles[2]. Applying this approach to the oceanographic realm would facilitate interoperability between datasets in each domain and across domains. There are efforts underway to this end

[2] https://www.go-fair.org/fair-principles/

Shared code repositories – A central location to share algorithms and trained models would benefit new ML/AI users. There is often overlapping development effort as many models, particularly deep neural networks, are essentially the same from an architectural stand point. Sharing parameter files and architectures will allow researchers to constructively use and build off of existing work. The concept of a "model zoo" has been used in the computer science community, notably by the developers of the Caffe software package at UC Berkeley. This resource could be a stand-alone website or a moderated page on GitHub or a similar service.

Curated information – A regularly maintained Wiki or formal review that lists relevant ML/AI approaches and code bases would be of broad utility for oceanographers. For scientists unfamiliar with ML/AI, selecting the right approach, both from a theoretical and practical perspective, can be challenging. A centralized resource outlining how different techniques are used, domains where they are applied, and how to access the necessary code will expedite development.

Data and metadata standards

Workshop participants and organizers alike expressed a desire for coordination on standards for data storage and annotation. The adoption of a consistent storage structure will aid scientists beginning new studies, allow researchers to easily compare data geographically or temporally, and enable engineers wishing to use outside resources to develop learning systems. Each domain area has specific needs to facilitate such standards:

- Acoustics
    - Sound files with standard metadata. Standard sound file formats exist, but no standards for acoustic environmental metadata exist.
    - Some species have well-documented sounds; others are more subjective such as "Humpback song." Anthropogenic sounds can also vary widely. Establishing standard names for sounds and their associated characteristics like duration, frequency and other relevant spectral measurement would aid in sharing and comparing techniques.
    - Calibrated sound levels. For environmental monitoring of soundscapes, acoustic calibration methods need to be well-documented and shared to ensure baseline measurements are unbiased.
    - Sound acquisition details, e.g. geographic location (latitude, longitude, depth), device metadata, and sound reception envelope according to acoustic propagation models are needed for data provenance and general sharing.
- Genomics
    - Metadata associated with sample preparation and processing is virtually nonexistent.

    - The lack of such standards limits the sharing of computing techniques from lab-to-lab and project-to-project, thus limiting the utility of such techniques for environmental monitoring.
    - Several existing, but non-coordinated, databases already exist. Still need unification in terms of formatting to ease collaboration.
- Imaging
    - Well-defined and cohesive classes with detailed information regarding the organizational structure. Often classes are defined haphazardly for an individual project. Guidelines for organization and class definition would enable better coordination among monitoring efforts. This could include taxonomic information such as that codified by the World Register of Marine Species[3].
    - A standard format for attaching semantic descriptors (e.g. "lateral view," "puffy," or "mottled") to images. Such annotations are useful for richer model development and could be extended for cross-dataset analysis.
    - Datasets need metadata associated with the collection method, illumination type, and, for microscopy, magnification.

Coordination with computer science community and industry

The data and scientific goals of many members of the oceanographic community may require development of new techniques rather than application of existing ones. Such undertakings require expertise outside the purview of most domain science laboratories. To approach these projects, many workshop participants expressed interest in developing ties with their counterparts in the computer engineering community.

Establishing such collaborations can be a challenge between groups that often speak different scientific languages. Initiatives at the university level to encourage dialogue between relevant departments could pay dividends. Some more general programs are already in development, such as the Massachusetts Institute of Technology's newly endowed Schwarzman College of Computing[4], that focus on applied ML/AI in a broad range of academic disciplines.

Industry players are often eager to get involved with researchers seeking to use new ML/AI techniques. Several examples of this sort of collaboration are already in existence: AWS has a special program for non-profits seeking to use their services[5]; Google researchers have used their technology to classify whale calls from a NOAA data repository[6]. Much like encouraging relationships between academic departments, developing such collaborations could be jump started at the institutional level.

---

[3] http://marinespecies.org/

[4] http://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015

[5] https://aws.amazon.com/government-education/nonprofits/

[6] https://ai.googleblog.com/2018/10/acoustic-detection-of-humpback-whales.html

SECTION III

The 2019 Workshop on ML/AI in Biological Ocean Observations introduced practitioners in the field to new analysis techniques. It is our hope that the participants will be able to take what they learned back to their home institutions to teach their colleagues and speed analysis of data from their high-throughput observational systems. We also hope that the experiences and insights from the researchers on the needs of the community will aid POGO's efforts to engender a positive environment for global collaborations in this area.

POGO has the ability to be a guiding force toward adopting ML/AI at the scale necessary to facilitate longer term biological monitoring projects. ML/AI resources – intellectual, computing, and storage – will be necessary to ensure that conducting large scale studies remains viable as increasingly data intensive technologies come online. Early investments toward codifying data standards and promoting consistent analysis will pay particularly high dividends toward speeding scientific efforts.

One can readily envision a future where ML/AI methods are an integral part of marine studies. Acoustic arrays will become more common on buoys and moorings; genomic assays will be regular parts of water sampling campaigns; imaging systems will increasingly be deployed on autonomous vehicles and shore stations. The future is digital and data intensive. The community can be prepared and coordinated to make the efforts as enlightening as possible.

APPENDIX I

ACADEMIC REFERENCES

Bollier, D., and C. M. Firestone, 2010: *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC, USA.

Howe, D., and Coauthors, 2008: Big data: The future of biocuration. *Nature*, **455**, 47–50.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, https://doi.org/10.1038/nature14539.

MacLeod, N., M. Benfield, and P. Culverhouse, 2010: Time to automate identification. *Nature*, 467, 154–155, https://doi.org/10.1038/467154a.

Marx, V., 2013: Biology: The big challenges of big data. *Nature*, 498, 255–260.

Pesant, S., and Coauthors, 2015: Open science resources for the discovery and analysis of Tara Oceans data. *Scientific data*, **2**, 1–16.

Wilkinson, M. D., and Coauthors, 2016: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018, https://doi.org/10.1038/sdata.2016.18.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson, 2014: How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems 27*, 3320–3328.

TECHNICAL LINKS

- Pytorch: https://pytorch.org/
  - Deep learning library for rapid development
- Jupyter: https://jupyter.org/
  - Graphical interface for python development
- GluonCV: https://gluon-cv.mxnet.io/
  - Set of python wrappers for MXnet deep learning library
- Scikit-learn: https://scikit-learn.org/stable/index.html
  - Standard python machine learning toolbox
- GenePhene: https://www.geno2pheno.org/
  - Protein database
- Traitar: https://github.com/hzi-bifo/traitar
  - Protein database

APPENDIX II

Partnership for Observation of the Global Ocean

Machine Learning/Artificial Intelligence for Biological Observations Workshop

MONDAY, MAY 20, 2019

0830	*Registration, coffee, networking*
0900	Introductory remarks (Margaret Leinen)
0910	Overview of workshop format (Eric Orenstein)
0920	Brief ML/AI overview. What it is and what it is not (Eric Orenstein)
0945	State of the field for each of the 3 tutorial subjects (Tutorial Leaders)
1030	*Group photo and Coffee Break*
1100	Break into domain-specific groups for "Lightning talk style" introductions (All Participants) Participants will have no more than 2 minutes and a single slide to give a (very) brief overview of their work. Focus: data you work with, the challenges associated with analyzing it, and what you hope to get from the workshop
1230	*Lunch*
1400	Tutorial session 1
1530	*Coffee break*
1545	Tutorial session 1 continued
1800	*Reception at VLIZ*

TUESDAY, MAY 21, 2019

0900	Tutorial session 2
1030	*Coffee Break*
1045	Tutorial session 2 continued
1230	*Lunch*
1330	"Homework" session
1500	*Coffee break*
1515	"Homework" session continued
1730	*Adjourn*

WEDNESDAY, MAY 22, 2019

0900	Discussion and wrap-up in domain-specific groups
1030	*Coffee Break*
1045	All attendees come together for final discussion and wrap-up
1200	*Workshop concludes*

POGO ML/AI Biological Observations Workshop
List of attendees

Imaging

- Baburaj, Reshma; Central Marine Fisheries Research Institute (CMFRI)
- Button, Rio; University of Cape Town, South Africa
- Cornils, Astrid; Alfred Wegener Institute
- Currie, Jock; South African National Biodiversity Institute & Nelson Mandela University
- Deneudt, Klaas; VLIZ
- Giering, Sari; National Oceanography Centre, Southampton
- Gomes, Alessandra; University of Sao Paulo
- Irisson, Jean-Olivier; Sorbonne Université
- Jansen, Jan; IMAS, University of Tasmania
- Johnson, Craig Richard; IMAS, University of Tasmania
- Kitahashi, Tomo; JAMSTEC
- Lindley, Anthony JW; University of Southampton
- Mendes Lopes, Rubens; University of Sao Paulo
- Ollevier, Anouk; VLIZ
- Pérez, Nerea Valcárcel; Spanish Institute of Oceanography
- Scoulding, Ben; CSIRO
- Simon, Julien; IFREMER

Genomics

- Arce, Paola; SAMS
- Bagi, Andrea; NORCE Norwegian Research Centre
- Dillen, Nick; VLIZ
- Dully, Verena; University of Kaiserslautern, Germany
- Evans, Susan; National Oceanography Center, UK
- Hempel, Christopher Alexander; University of Guelph, Canada
- Hernández de Rojas, Alma; Spanish Institute of Oceanography
- Keeley, Nigel; Institute of Marine Research, Norway
- Knapik, Kamila; NORCE Norwegian Research Centre
- Krolicka, Adriana; NORCE Norwegian Research Centre
- Macher, Till-Hendrik; University of Duisburg-Essen, Germany
- Ohnesorge, Alica; HIFMB, Germany
- Organo Quintana, Cintia; University of Southern Denmark
- Salles Bernal, Soluna; Spanish Institute of Oceanography
- Samuel, Robyn; National Oceanography Centre, UK

Acoustics

- Ariza, Alejandro; British Antarctic Survey
- Chapuis, Lucille; University of Exeter, UK
- de Pina Júnior, Luís Filipe; Lurio University, Mozambique
- Develter, Roeland; VLIZ
- Djokic, Divna; Universidade Federal do Rio Grande do Norte (UFRN)
- Gaughan, Paul; Marine Institute Ireland
- Laute, Amelie; Whale Wise (Iceland)
- Maslov, Dmytro; University of Minho
- Wisniewska, Danuta Maria; CNRS, France

Guests and Observers

- Alexander, Britt; European Marine Board
- Bangle, Brandi; Scripps Institution of Oceanography, UCSD
- Chavez, Francisco; MBARI
- Leinen, Margaret; Scripps Institution of Oceanography, UCSD
- Michaels, William; Director--NOAA Fisheries' Ocean Technology Program

Tutorial Leaders

- Cline, Danelle; MBARI
- Cordier, Tristan; University of Geneva
- Lanzen, Anders; AZTI
- Orenstein, Eric; Scripps Institution of Oceanography, UCSD
- Ryan, John; MBARI
- Schröder, Simon-Martin; Kiel University