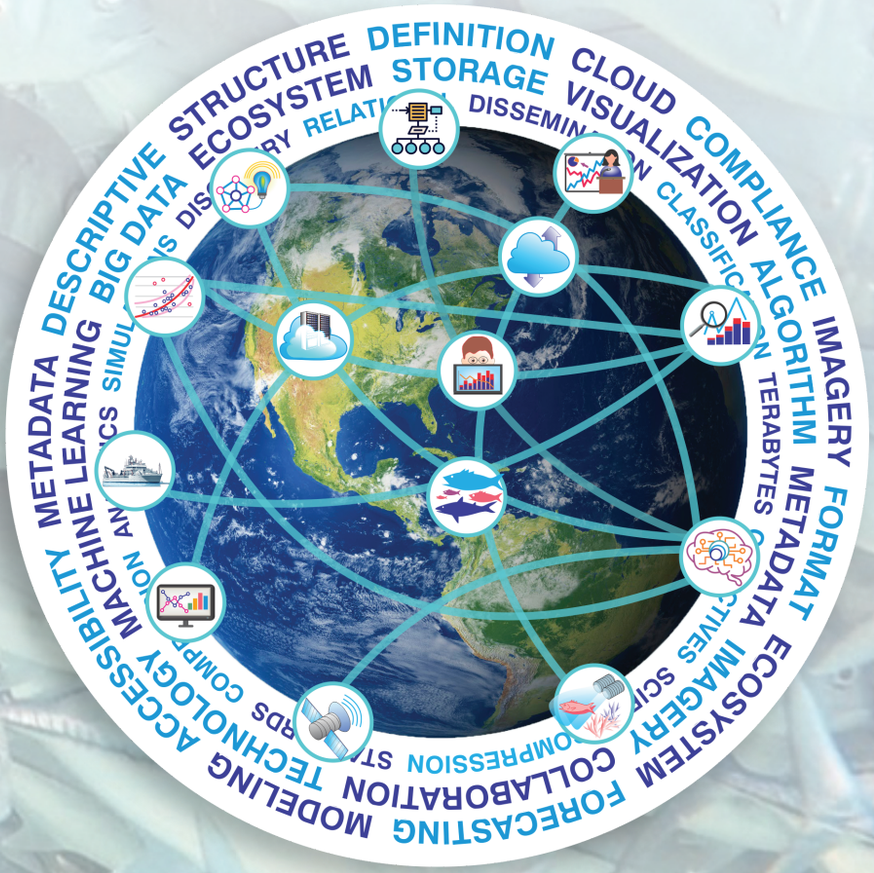
NOAA FISHERIES

Accessibility of Big Data Imagery for Next Generation Machine Learning Applications



NOAA Technical Memorandum NMFS-F/SPO-194
April 2019

Accessibility of Big Data Imagery for Next Generation Machine Learning Applications

Sarah Margolis, William L. Michaels, Brett Alger, Christopher Beaverson, Matthew D. Campbell, Edward J. Kearns, Mashkoor Malik, Charles H. Thompson, Benjamin L. Richards, Carrie C. Wall, Farron Wallace

May 2019
NOAA Technical Memorandum NMFS-F/SPO-194

U.S. Department Of Commerce

Wilbur L. Ross, Jr.
Secretary of Commerce

National Oceanic and Atmospheric Administration

Neil A. Jacobs, PhD
Acting Under Secretary of Commerce for Oceans and Atmosphere and NOAA Administrator

National Marine Fisheries Service

Christopher W. Oliver
Assistant Administrator for Fisheries

Recommended citation:

Margolis, S., W.L. Michaels, B. Alger, C. Beaverson, M.D. Campbell, E.J. Kearns, M. Malik, C.H. Thompson, B.L. Richards, C.C. Wall, F. Wallace. 2019. Accessibility of Big Data Imagery for Next Generation Machine Learning Applications. NOAA Tech. Memo. NMFS-F/SPO-194, 64 p.

Copy of this report may be obtained from:

https://spo.nmfs.noaa.gov/tech-memos

Contents

Acknowledgements

The comprehensive overview of the storage and accessibility of imagery and its data in the report was possible with the collective input from various intra-agency and inter-agency experts who understand the increasing importance that machine learning will have on the quality and timeliness of scientific products. In addition to the editors, several contributors provided insight on the necessary components of big data accessibility to optimize its use for long-term research and discovery by the broader scientific community. Matt Dornback (NESDIS) provided insight on the requirements for annotation and archival pertinent to the Big Earth Data Initiative deep sea coral using SeaTube. Megan Cromwell (NESDIS), Kirsten Larsen (NESDIS), Luke Thompson (NMFS) and Nathan Wilson (NMFS) provided insight on the Inport requirements and standards for the NOAA Fisheries' metadata repository. Caitlin Ruby (NMFS) was instrumental in executing the Video Data Management Pilot project for archiving imagery from NOAA's Southeast Fisheries Science Center to the National Center for Environmental Information. Dvora Hart (NMFS) and Kresimir Williams (NMFS) provided an overview of how optical data projects collect and process data. Anthony Hoogs and Matt Dawkins from Kitware Inc. provided technical expertise from a training workshop that utilized machine learning to streamline image processing with image recognition. Euan Harvey provided insight on machine learning for underwater imagery processing in Australia. Todd Ruston and Bill Kirkwood summarized information about MBARI's data collections, processing, and future of machine learning. Ken Casey (NESDIS) from The National Centers of Environmental Information provided expertise on archival practices, and NOAA's enterprise architect, Dave Layton, assisted with information related to NOAA's cloud strategy. Frank Schwing also provided input and guidance on the document.

List of Acronyms in Report

AIASI: Automated Image Analysis Strategic Initiative

AI: Artificial Intelligence

AFSC: Alaska Fisheries Science Center

AVED: Automated Video Event Detection

BEDI: Big Earth Data Initiative

CIMS: Cruise Information Management System

CLASS: Comprehensive Large Array-Data Stewardship System

CNN: Convolutional Neural Network

CRADA: Cooperative Research and Development Agreement

EBS: East Bering Sea

EM: Electronic Monitoring

ER: Electronic Reporting

FAR: The Federal Acquisition Regulation

FISMA: The Federal Information Security Management Act

FITARA: The Federal Information Technology Acquisition Reform Act

GUI: Graphical User Interface

ICES: International Council for the Exploration of the Sea

ISO: International Standard Organization

JSON: JavaScript Object Notation

ML: Machine Learning

MSA: Magnuson-Stevens Act

MBARI: Monterey Bay Aquarium Research Institute

NARA: National Archives and Records Administration

NASEM: National Academies of Sciences, Engineering, and Medicine

NIST: National Institute of Standards and Technology

NRC: National Research Council

NCEI: National Centers for Environmental Information

NEFSC: Northeast Fisheries Science Center

NFWF: National Fish and Wildlife Foundation

NWFSC: Northwest Fisheries Science Center

NMFS: National Marine Fisheries Service

NSF: National Science Foundation

OAIS: Open Archival Information System

OER: Office of Exploration and Research

PARR: Public Access to Research Results

PIFSC: Pacific Islands Fisheries Science Center

ROI: Region of Interest

SEFSC: Southeast Fisheries Science Center

SRI: Stanford Research Institute

SVM: Support Vector Machines

SWFSC: Southwest Fisheries Science Center

UHSI: Untrawlable Habitat Strategic Initiative

UCSD: University of California San Diego

UTC: Coordinated Universal Time

VARS: Video Annotation and Reference System

VDMMI: Video Data Management Modernization Initiative

VDMP: Video Data Management Pilot

VIAME: Video and Image Analytics for a Marine Environment

Executive Summary

NOAA generates tens of terabytes of data a day, also known as “big data” from satellites, radars, ships, weather models, optical technologies, and other sources. This unprecedented growth of data collection in recent years has resulted from enhanced sampling technologies and faster computer processing. While these data are publicly available, there is not yet sufficient access to the data by next generation processing technologies, such as machine learning (ML) algorithms that are able to improve processing efficiencies. Accessibility is the key component for utilizing analytical tools and ensuring our processing meets 21st century data needs. This report focuses on the challenges of accessibility of imagery (defined as still images and video) from the marine environment. Vast amounts of imagery are collected from optical technologies used in marine ecosystem monitoring and ocean observation programs. While technologies have dramatically increased the spatial and temporal resolution of data and increased our understanding of marine ecosystems, the drastic increase in big data, specifically imagery, presents numerous challenges. Case studies discussed in this report highlight that big data imagery are readily being collected and stored, yet the foundation for the long term storage and accessibility of big data must be based on the necessary guidance for its architecture, infrastructure, and applications to enhance the accessibility and use of these data to help fulfill NOAA’s cross-functional missions. Additionally, the report highlights key considerations and recommendations for NOAA’s data modernization efforts that align with mandates such as Public Access to Research Results, the Evidence-Based Policy Making Act, Department of Commerce Strategic Plan, the President’s Management Agenda, and White House Executive Order on Artificial Intelligence (AI).

As big data and analytical tools become more commonplace for NOAA’s research and scientific operations, there is an increasing need to create end-to-end data management practices that improve data accessibility for analytical tools that utilize AI, computer vision (AI applied to the visual world), and ML. The development and application of AI and ML analytics will progress as long as there is accessibility of big data with enriched metadata; however, accessibility appears to be the primary challenge to fully utilize ML analytics. Rapid, optimal access to entire imagery and data collections is critical to create annotated imagery libraries for supervised analysis using ML algorithms. This report highlights the common need to implement accessibility solutions to facilitate efficient imagery processing using available analytical tools.

Other critical requirements to enable AI include the necessary metadata for discovery, long term data archive and access, and economical multi-tier storage. As big data imagery are made more readily available to open source tools such as ML analytics, significant cost reductions in data processing will be realized by reducing the labor-intensive efforts currently needed. ML tools accelerate processing of imagery with automated detection and classification resulting in more timely and precise scientific

products for management decisions. Furthermore, as the broader scientific community expands its research and discovery from increased accessibility of big data imagery, the ML applications will increase the number of insightful science-based products beyond the scope of the original operational objectives, thereby increasing the value of the agency's scientific products.

1. Introduction

1.1 Background

NOAA's expansive collection of high-quality environmental data and expertise are publicly available, although sufficient accessibility to next generation processing technologies, such as machine learning (ML) algorithms, is lacking. Obtaining access to raw imagery (still images and video) files is an inefficient process i.e. shipping a hard drive in the mail, and third party ML applications have limited access to imagery whose processing would benefit from greater accessibility. There has been exponential growth in the volume of imagery collected and information produced in recent years as a direct result of enhanced sampling technologies and faster computer processing. This innovation and growth in technologies is largely influenced by several driving documents and mandates highlighting modernizing technology, ML, big data, and accessibility such as, Public Access to Research Results, the Evidence-Based Policy Making Act, U.S. Department of Commerce 2018-2022 Strategic Plan, the President's Management Agenda, and the February 2019 Executive Order (EO) on Maintaining American Leadership in Artificial Intelligence.

The marine ecosystem is Earth's largest resource, and the demand to collect big data using ocean technologies requires attention to data management and computing infrastructure efforts that increase the scalability and accessibility of these data. For the purpose of this report, the reference to big data is in regard to the increasing rate of data collection which traditional approaches struggle to process and analyze in a timely manner. In particular, data from the deployment of ocean sensor technologies such as underwater acoustic and optical technologies dramatically increased the spatial and temporal resolution of information collected from environmental and resource monitoring programs. Such innovations in sampling technologies have helped resolve problems from data-limited assessments from habitats that were previously difficult to sample. While scientists are continuously making advances in the post-processing and analytical tools to address big data, the main bottleneck in fully utilizing these data is the need to further develop the supporting data enterprise with the necessary infrastructure, architecture, and accessibility.

Exponential growth in imagery collection requires adaption from the current data enterprise to new analytical tools that expedite processing.

It is critical for the future of science and management to integrate artificial intelligence, specifically ML, into the workflow of imagery processing so that machines and minds can work together (Fortunato et al., 2018). This will enhance the use of big data to enable discovery among the wider scientific community using open source analytical

tools, thereby increasing the value of scientific data. In addition, accessibility of collected imagery and data aligns with one of the key drivers for improving data accountability and transparency in work done by federal agencies, as stated in the President's Management Agenda (United States, 2018).

Currently, much of the imagery and data (information gathered from imagery) from marine research and survey monitoring programs are stored at the regional level. This might be acceptable for regionally focused objectives, but the scope of this report is to highlight considerations in making these big data imagery more accessible to the broader community for research and discovery. For the purposes of this report, still images and video collected from optical technologies will be referred to as *big data imagery* and quantitative information resulting from processed imagery will be referred to as *imagery data*. The storage of big data imagery databases must be founded on sound principles for its architecture, infrastructure, and applications with analytical tools that also impose their own infrastructure requirements. Developing this data storage framework requires enriched metadata for long term accessibility. This requires an understanding of not only the original objectives of the data collection, but an understanding of the new scientific products that can be derived from increased accessibility for data mining (the process of discovering new information or patterns within the data). Furthermore, tools for processing and analyzing imagery have advanced with ML algorithms. These analytics with automated image recognition dramatically reduce the processing time of imagery data, provide more precise measures of length and abundance estimates of living marine resources, and increase opportunities to analyze big data, including interdisciplinary analysis between imagery and other environmental information. Data integration and visualization are key components of the big data platform to increase the use of data by the wider scientific community and citizen science. Therefore, we must strive to increase the accessibility of big data imagery for the analytical tools that are becoming more readily available (Figure 1).

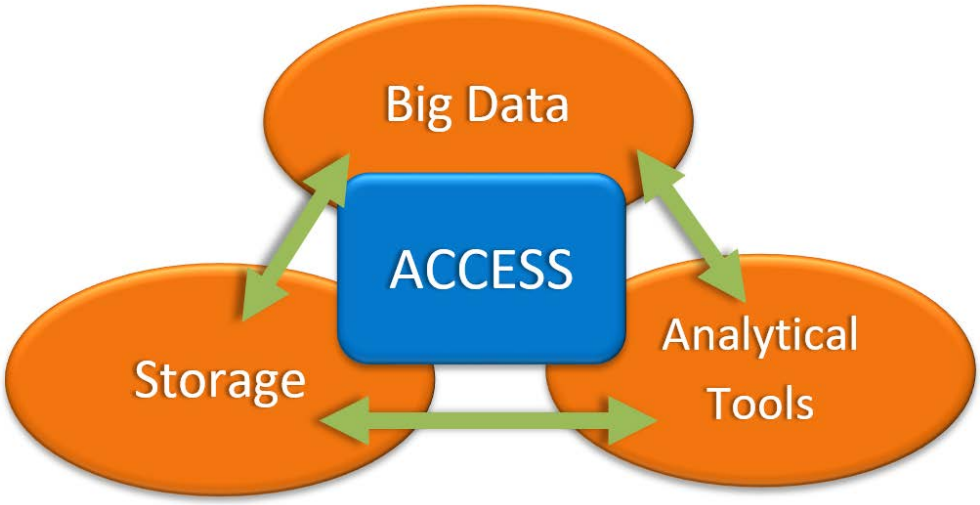


Figure 1. The framework for managing big data must optimize the accessibility of data sources for utilizing state of the art analytical tools, including access and development of training datasets and annotation libraries for machine learning applications.

1.2 Goal

This report is intended to provide guidance and recommendations to improve the accessibility of big data imagery for machine learning applications for NOAA's Science and Technology Enterprise. NOAA provides science-based products in support of its mission, and strives to make its data accessible to the scientific community that also contributes to the sustainability and health of our living oceans. Case studies are provided to highlight specific guidelines and recommendations on the metadata, storage, and accessibility of underwater imagery. Traditional human processing is unable to keep up with the growing volume of imagery data collected from the marine environment and should be augmented by available tools. While sampling technologies in marine monitoring programs are frequently deployed, only recently have ML analytics become available to reduce processing time. Accessibility also remains a key consideration for next generation collections and processing of NOAA's big data imagery. As the big data enterprise examines how best to increase the accessibility of big data imagery, the following analytical questions must be considered to understand the end to end process of data collection and information management:

Big data imagery management should consider the analytical questions relevant to the end to end process for deriving scientific products with analytics during research and discovery.

Descriptive analytics: What was collected and why was it observed?

Diagnostic analytics: What patterns can be derived, and why did they occur?

Predictive analytics: What are the metrics, and what is the forecast?

Decisional analytics: How is it used, and what improvements can be made?

The preceding questions address both the short and long term accessibility requirements of the big data enterprise, and should be used to determine whether certain imagery and data should remain at the regional level or be centralized for accessibility to the broader community. The methods and tools used to answer these analytical questions may impose requirements on the access methods, level of service, and characteristics of big data imagery. They also may address accessibility of big data imagery to ML applications that likely will provide scientific products beyond the scope of the original operational objectives. Therefore, big data imagery requires the necessary metadata to organize the relationships among databases to optimize the use of analytical tools for research and discovery.

Overarching components of managing and access to large sets of imagery in this report will address the following topics.

Importance of Metadata: Explain hierarchical structure and types of metadata that are critical requirements for accessibility and discoverability of the source data in the long term (Section 2).

Storage and Archives: Address considerations for storage and archival practices and highlight the importance of organized, reliable storage as a gateway to accessibility (Sections 3 and 4).

Significance of Accessibility: Provide clarification on the significance and interconnectivity of big data storage, accessibility, and readily available analytical tools (Section 5).

Analytical Tools: Describe how analytical tools, such as ML, can accelerate the processing of imagery and yield more timely results to inform management decisions (Section 6).

Case studies presented in this report will demonstrate how NOAA and other institutions have addressed big data challenges to advance efficient imagery accessibility and processing (Sections 7-10), yet there is more to be done as analytics evolve and become more readily used. The overarching goal to enhance the accessibility of big data imagery collected from the marine environment to ML analytics will provide immediate benefits

in streamlined processing, more precise measures of length, count and species identification, and timely products for marine ecosystem science.

2. Metadata Standards for Accessibility

Metadata involves establishing policies and processes that ensure information can be integrated, accessed, shared, connected, analyzed and maintained across the organization. The importance of metadata is well established to automate administrative processes, improve data quality and extraction of information from data. Recently, metadata-driven accessibility for data-mining and analytical tasks has become a focus for developing big data enterprise (Froeschl, 1997; Vaduva and Vetterli, 2001; Hay, 2010; Vnuk, 2014; Blokdyk, 2018). Metadata has various categories of information (Figure 2) used to build relationships across databases that make research data discoverable and reusable in the long term. Furthermore, metadata can be enriched to achieve a number of objectives that result from increased accessibility, use of analytics, and discovery. The importance of metadata has dramatically increased during the past two decades with the exponential increase in data storage, and most certainly with the recent use of ML tools.

Enriched metadata improves data accessibility for new analytics like machine learning, enabling more timely research and discovery.

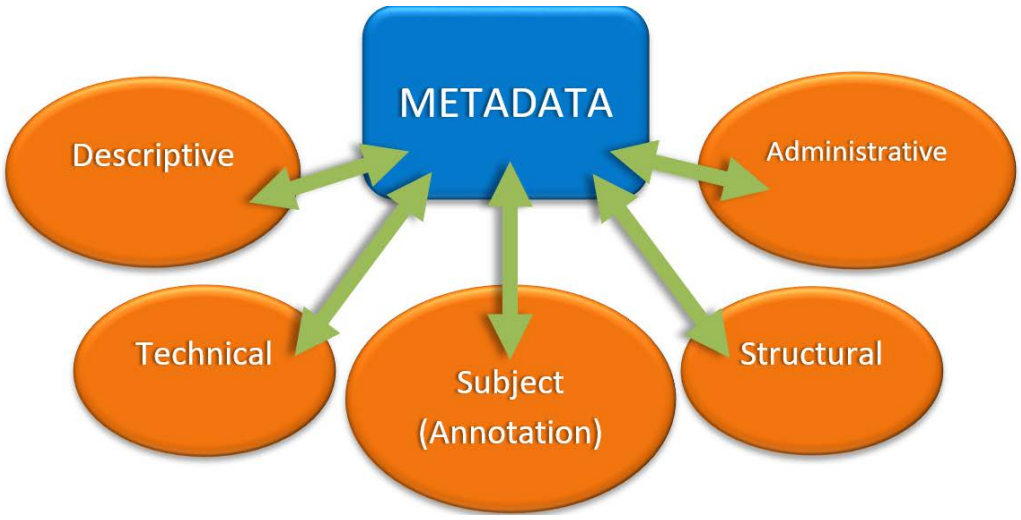


Figure 2. The different categories of metadata that govern the structure of imagery collection, organization, storage and therefore accessibility.

The value of metadata for accessibility and discovery is clear, yet there are ongoing discussions within metadata communities on how it might evolve with the applications

of new tools that utilize ML algorithms. The standard classification scheme for metadata include the following categories:

Descriptive metadata: Describes the data resource (type of data, keywords, subject title, project description, learning objectives, catalog information, pertinent publications), and who created the data (creator or investigator, affiliation, date, time, location). The descriptive metadata describes the resources for purposes such as discovery and identification.

Technical metadata: Provides further descriptive information about the resource, such as digital object management (creation date and time, compression requirements, format and size, time-stamp, and other interoperability variables). UTC timestamps (Coordinated Universal Time) of imagery are highly recommended for accurate time reference independent of time zone and for synchronization of other relational data. Furthermore, geographical coordinates make imagery accessible through geo-based map portals.

Subject metadata: Referred to as annotation and textual representations describing what is inside the imagery, such as fish species, length or number. The ability to annotate during the post-processing and analyses are critical for a user interface to properly search through and accurately filter imagery for subjects of interest, such as a specific organism or habitat feature.

Structural metadata: Describes the hierarchical organization of the data, structural definitions, and relationships with other data.

Administrative metadata: Provides information on the origins of the resources, file format size and software, preservation location and retention period, regulatory compliance, and restrictions on access and usage.

The analytical questions previously discussed (Section 1) are important to enrich metadata for operational efficiency in the dissemination of datasets and to fit the requirements for better discovery. This brings big data imagery into a more accessible and inter-operative data structure that provides immediate benefits from streamlined processing and discovery using analytics and visualization.

3. Data Management Regulations and Compliance

Data management governance and relevant administrative metadata establish the creation, retention, security, data integrity and accessibility rights of the data enterprise.

Data management standards with enriched metadata structure establish the administrative compliance, archival retention and security, data integrity and accessibility rights of the data enterprise.

When assigning metadata to imagery, it is important to consider existing guidelines and standards. The International Organization for Standardization (ISO) has put forth standards that focus on metadata and have been adopted by international organizations. NOAA Fisheries optical imagery adheres to ISO metadata standards. These standards do not provide naming classification for imagery annotations, although established standards exist, such as the Darwin Core and Coastal and Marine Ecological Classification Standards, and are widely used in the marine community (Wieczorek et al., 2012; Madden et al., 2009). Clear, standardized metadata make imagery data more useful and discoverable. In particular, subject metadata (annotations) are critical for a user interface to properly search through and accurately filter imagery and data for the user's needs. The NOAA Data Documentation Directive[1] provides guidance on metadata to ensure all environmental data are accompanied by machine-readable metadata to enable discovery, access, and use of online and offline data holdings. Standards for formatting and minimum information required for discovery, access and use are outlined and referenced to the ISO; however, NOAA does not require any particular wording for subject metadata. Existing classification guidelines are widely used (e.g. Carollo et al., 2013; Shumchenia and King 2010) and may provide guidance for future agency tf cstandardization.

Data management policies for the NOAA data enterprise are discussed in Section 5.1. With regard to data integrity of the enterprise, the format and relational standards with appropriate time stamps and coordinates are a priority. The use of timestamps, especially Coordinated Universal Time (UTC), are critical when annotating data. They vastly improve the reusability of the data, as the data become searchable by the time they were collected or annotated. UTC timestamps are standard and not time-zone dependent. In addition, geographical coordinates attached to images and video are exceptionally important for referencing imagery data components to one another and its collection location (Section 5.2). Coordinate information allows discoverability of imagery data through map-based portals.

[1] NOAA Data Documentation Directive:
NOAA Data Documentation Directive

Data management compliance is a critical requirement in the determination of data access. For this reason, the scope of this report will be primarily focused on accessibility of big data imagery to the broader scientific community that will utilize machine applications. There is recognition of the far-reaching benefits of increasing accessibility of certain data for research and discovery by the broader community, while other data would be subject to accessibility restrictions due to privacy and confidentiality regulations. For example, imagery data collected from electronic monitoring (EM) fishing vessel operations would have certain confidentiality requirements (Section 8). Although this report acknowledges the increasing volume of imagery data collected from EM programs, this report is limited to the community's use of open big data imagery collected from fishery-independent monitoring programs (Section 7). This report focuses on how to optimize the scientific products from ML analytics for big data imagery that can be made available to the broader scientific community.

Time stamps and coordinates are a priority to imagery annotations, allowing for geographic discovery and cross referencing of the imagery with relational databases.

4. Storage

Storage Organization

As more imagery are generated for near real time analysis, solutions will be needed to handle and store more imagery than ever. The old model of block storage is a system in which metadata are not directly connected to imagery and have limitations in scalability. Although block storage may have faster performance when the application and storage are local, it is no longer effective for large processing demands of the big data enterprise. Object-based storage provides greater accessibility benefits as it has unlimited scalability and enhances search capabilities. This type of storage is best for unstructured data (data that are not organized in a pre-defined manner).

Interconnectivity between storage and archival resources should be optimized for data accessibility and workflow to utilize analytical tools such as machine learning.

When addressing imagery storage and archives, it is important to understand the definitions, misconceptions and objectives associated with each term so that an appropriate repository, or more likely a combination of storage solutions, can be identified for big data imagery collections. For this paper, the term "storage" is understood as a temporary solution to host imagery for processing needs and requires quick access to fulfill operation missions. Archival storage tends to be for long-term

historical storage preserved for future use. The challenge is how the data enterprise should optimize storage and accessibility of big data imagery for research and discovery using analytical tools. Archiving is a more time-consuming form of storage as it follows the formal archive process determined by NOAA and the National Archives and Records Administration (NARA); however well-archived data should be discoverable via metadata, and timeliness of access depends on the type of media used to preserve the imagery (e.g., spinning disks versus tapes). A popular use of the word "archive" often implies slow and difficult access to big data, but this may not be the case if data are archived properly and accessible under the tenets of the Open Archival Information System (OAIS) Reference Model (OAIS Reference Model, ISO 14721). Rapid accessibility for computational purposes by analytical software, such as ML, may not be best suited on a preserved copy of archived data. In recent years, commercial cloud computing (storing and accessing imagery and programs over the internet instead of a local computer's hard drive) has increased in popularity.

Cloud Services

Enterprises are moving toward cloud storage due to the lower cost associated with fewer dedicated storage arrays, and also for their dynamic storage software that can be more readily integrated with deep learning algorithms (Avram, 2014). For example, industry has developed cloud storage capability that utilizes artificial intelligence algorithms to improve data security and accessibility, and these storage services are scalable to store and retrieve large amounts of data with high levels of reliability and redundancy. Informed decisions on the services of commercial cloud providers must be based on cost, security and performance. The NOAA Big Data Project led by the Office of the Chief Information Officer (Section 5.2) provides an example of evaluating the costs and benefits of services provided by potential industry cloud partnerships.

Although cloud computing offers an array of benefits, it can be quite complex; decisions on constructing and changing the internal environment to support the cloud are just as much about the business model as the technology. The cloud environment requires a strong foundation of best practices in software development and architecture along with well-defined service and security management foundations. Transitioning to the cloud can be challenging; however, it is becoming less so as the various cloud platforms develop tools to ease the transition. The idea of merging multiple cloud models, below, may be well suited for a data enterprise in terms of security, privacy, scalability, and cost.

Multiple Cloud Models

NOAA generally subscribes to standards and definitions put forth by the National Institute of Standards and Technology[2] (NIST). Terminology surrounding cloud service models and deployment models will reference NIST 800.145, The Definition of Cloud Computing (Mell and Grance 2011). There are four deployment models outlined in the literature that each offer benefits and challenges to be addressed by the organization based on storage objectives, need for access, and considerations of privacy (Aryotejo et al., 2018; Fox et al., 2009; Goyal 2014; Mell and Grance, 2011).

Multiple cloud models exist with varying services toward ownership, security, scalability and cost to enhance the accessibility of big data imagery.

Deployment Models

Public: The cloud is available to the general public in a pay-as-you-go system.

Private: Cloud computing and storage through internal data centers and not made available to the general public.

Community: A cloud infrastructure shared between several organizations.

Hybrid: Combined private and public cloud services.

Brief details on each model's services for ownership, security, scalability, and cost are presented in Table 1. There is presently a debate on the functionality of private versus public cloud storage, particularly in regard to research with analytical tools such as ML algorithms. Although it is common to manage data and imagery as one pool, regardless of the type of storage (Islam et al., 2017), this presents challenges to accessibility for discovery by the broader community and multiple cloud models may need to be integrated into data storage solutions. Enterprises need to be flexible in adapting to new storage technologies and models of cloud-based data management.

[2] NIST Definition of Cloud Computing:
The NIST Definition of Cloud

Table 1. Cloud Deployment Models. Credit: Aryotejo et al., 2018.

Deployment Models	Holder	Security	Scalability	Cost
Private Cloud	Single private organization	Higher than other deployment models	Limited	High
Community Cloud	Two or more private organizations with identical requirements	Lower than Private Cloud and higher than Public Cloud	Limited	Medium
Public Cloud	Cloud Service Provider (CSP)	Lower than other deployment models	Very High	Pay-per-use
Hybrid Cloud	CSP and private organizations	Lower than Private and Community Cloud and higher than Public Cloud	High	Pay-per-use

Literature shows the hybrid cloud approach of combined public and private cloud as an ideal solution that reaps benefits while dealing with the limitations of the two models (Aryotejo et al., 2018, Goyal, 2014; Mirajker et al., 2012; 2015; Venkat et al., 2015). Hybrid approaches will optimize storage and accessibility requirements with third party IT vendors (e.g., Amazon Web Services, Google Cloud, IBM Cloud, Microsoft Azure, and other private cloud platforms), and the big data enterprise may find the hybrid cloud has more flexibility to meet compliance requirements for its data requirement regulations.

There are several cloud service models that can guide the integration of the cloud into storage solutions (Huwitz et al., 2010). Each is complex, and this report merely presents a brief overview of the following service models.

Service Models

Infrastructure-as-service: This service rents computer resources instead of buying and installing at the vendor's data center that includes servers, network technology and storage. It may also include operating systems and virtualization technology to manage resources. Infrastructure can be automatically scaled up and down.

Platform-as-service: The provider of this service delivers infrastructure and integrated software to build applications for software development. The consumer has the capability to manage all software development stages from planning to deployment and maintenance with the flexibility to test new software.

Software-as-a-service: This service hosts applications and makes them available to customers over the internet. These are purpose-built business applications.

Data-as-a-service: This service allows users to access data the same way they are used to accessing applications and infrastructure: as a service, available instantly, anywhere, on demand.

When deciding upon a cloud strategy, it is important to consider the strategy of the agency's overall mission and IT organization. It is particularly important that IT departments have the necessary control to manage all components of the service they receive and provide (Huwitz et al., 2010).

Although cloud services offer significant benefits, it is imperative to understand the storage and accessibility needs of the imagery and analysis methods. It is possible that a combined approach, where different IT environments of the cloud and servers are integrated, may deliver the required performance.

Combined Storage Technologies

The optimal imagery storage solution may be a mixture of locations and platforms, as cloud computing may not be an ultimate replacement for a data storage center. For example, cloud models and hard drive storage may be utilized for processing local imagery, while servers may be used to archive historical images and video. Imagery accessibility is directly linked to storage and network decisions; therefore, the data enterprise should inventory storage requirements and the frequency of data access to understand if and how the cloud should be integrated into storage and archiving solutions. For example, a public cloud is less expensive than maintaining region-based storage arrays, yet using public cloud for storage on historical data that is rarely accessed would result in outsized storage costs over time. However, at present, cloud providers are designing storage tiers with different pricing to address storage of infrequently used data. Another management possibility may be to keep processed video on the cloud for further discovery, and archive the raw images used to make that video. Images and video that have been processed and analyzed may contain no further information as we currently have the ability to analyze it, however, and may not require frequent accessibility.

Imagery storage can be optimized by utilizing an integrated technology system of servers and commercial cloud providers.

To manage different degrees of accessibility for big data imagery, the hybrid approach of using cloud models and physical servers might help to optimize storage and accessibility needs, including scalability, as these needs evolve over time with the development of analytics. It is critical to comprehensively analyze the phases of imagery storage that maintains the data integrity and optimizes accessibility in a cost efficient and timely manner.

Phases of Storage

Scientific imagery collected to fulfill operational objectives may have several storage requirement phases, as described below.

Computational-based storage: *easy access for post-processing; not long-term storage.*

The initial phase of storage involves collected imagery data with structured formats, filenames, and metadata. Typically, access is needed for up to one year during the imagery post-processing procedures. On-premise storage is most often used because routine access is required during post-processing, annotation, auditing, and analysis to generate routine products (e.g., standardized biomass indices for stock assessments). Note that these computational procedures could easily be accomplished in the cloud as well, and are becoming more common. After the imagery is fully analyzed for the purposes of fulfilling original research objective requirements, it is archived.

Archival storage: *Federal requirement for data preservation.*

Once the imagery has been audited and meets the data quality requirements, it is archived amongst historical data. Archiving imagery requires a rigorous auditing process attached to extensive resource costs, mostly directed to archive access. If archived data evolves, such as modification of metadata formats, the cost and effort to amend archive records can be demanding. While archives focus on preservation, more resources are most often dedicated to the accessibility of preserved data. The OAIS reference model[3] outlines accessibility as a key entity and data are not archived unless they are accessible (OAIS Reference Model, ISO 14721). The instantaneous accessibility needs for computational purposes and processing by ML software however may be met by access to preserved copies of archived data. Historical imagery and data are infrequently accessed for the purpose of obtaining performance metrics or improving the time series indices when new analytical methods become available. Generally, the purpose of an archive is to preserve the data for a defined, extended period of time as required for federal agencies. The National Archives and Records Administration has the authority under Chapters 21, 29, and 33 (3302, 3303a) of the Federal Records Act, 44 United States Code, to determine whether federal records have archival value. NOAA, and all federal agencies, must comply through definition and approval of retention schedules for federal records. NOAA's imagery data are subject to the Federal Records Act, and therefore must adhere to NARA's and NOAA's Appraisal and Archive Policies[4]. Commercial cloud storage may or may not be adopted by NOAA to fulfill the storage

[3] International Standards OAIS model 14721:
OAIS Reference Model (ISO 14721)

[4] NARA Appraisal Policy:
Appraisal Policy of the National Archives and Records Administration

requirements of the official NOAA archive. Further information on NOAA data archival can be found below in Section 5.3, for Environmental Information. The National Centers of Environmental Information (NCEI) is a centralized NOAA facility that archives a wide variety of environmental data using several different archival storage mechanisms.

Storage for discoverable imagery: *easy, open access, discovered through metadata portal or catalog.*

As big data imagery continues to grow and require more storage, the data enterprise will be challenged to optimize data accessibility for AI and ML analytics. This will open a new suite of discovery objectives that extend beyond the original research mission. Once imagery have been initially processed and archived, they should be stored with considerations for future discovery. Imagery metadata that are well annotated improve the frequency of specific data queries and will maximize requests for historical or large imagery retrievals. Sets of large imagery must be easily accessible to train AI and ML algorithms. To optimize cost efficiency, the mechanism for search and discovery that will provide access to metadata and storage information should be functionally separate than systems storing the imagery. Storage systems that provide access to ML must be scalable and affordable, and these two competing factors make for challenging decisions.

Data accessibility for machine learning and discovery by the broader community will likely evolve with enhanced interconnectivity between on-premise storage and cloud storage services.

Imagery accessibility for AI and ML applications will most likely require a mixture of technologies pertinent to the hybrid storage approach. While some imagery can remain on the cloud, managers should be mindful of privacy restrictions. For example, imagery and any data or metadata collected by NMFS EM programs that contains identifiable information of fishing operations is confidential (Section 8). Another example is data from sensitive protected marine areas, such as critical fish spawning and aggregation habitat, which may not be available for public access. For the imagery that can be publicly accessible, complete access for research and discovery depends on comprehensive metadata including time stamps, GPS coordinates, standardized annotations and optimal file formats. Consider what kind of access is needed for your imagery and what file formats would be best supported.

Recent advancements with cloud storage solutions, combined with traditional agency servers, will likely provide benefits in programmability, automation, scalability and efficiency to the big data enterprise. When deciding on appropriate storage options, an evaluation of the imagery storage/archive objectives must be done for each stage of the imagery lifecycle, and should consider the combined benefits of using commercial cloud providers and on-premise storage already within the domain of the NOAA Enterprise (Figure 3).

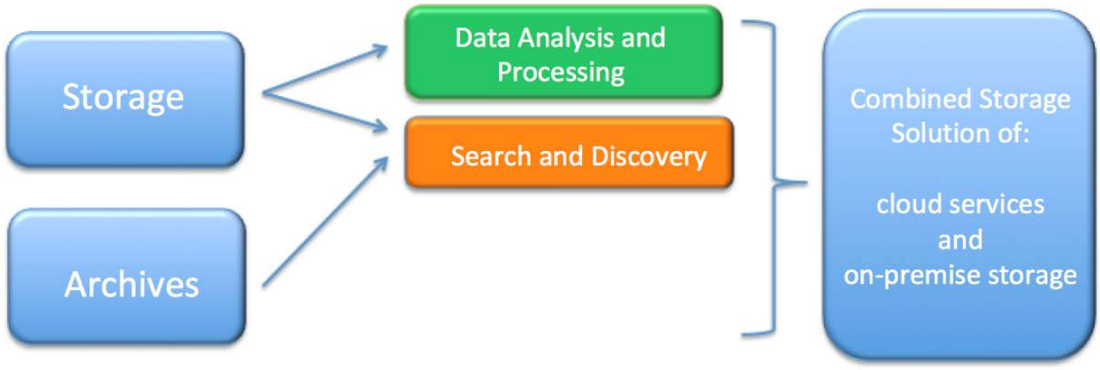


Figure 3. An idealized solution for the storage and archival processes of big data imagery will likely utilize cloud services provided by vendors that are integrated into on-premise storage.

Relevant Storage Questions

When deciding on a combination storage approach, there are several considerations that should drive the analysis of storage phases. The following components have several associated questions that are important when assessing the data enterprise: amount of imagery storage volume, frequency of access, centralization, and cost.

Volume: How much storage is required for historical, current and anticipated imagery?

Access: How often are the imagery accessed or will it be accessed?

Egress: Will the imagery be processed on the storage system, or moved via network to another location for processing? This is generally the largest cost-driver.

Location: Should imagery be kept and managed locally or is a central location needed to promote standards and easy, global accessibility? What are the challenges of transporting imagery from the current location to centralized storage?

Cost: How much will it cost per gigabyte to transport, upload, store, archive, and access the imagery of interest?

Business: Who will bear the responsibilities and costs?

The decision on the optimal storage system will likely be determined from a comprehensive analysis of analytical questions (section 1.2), data infrastructure (section 2) and accessibility requirements. The data enterprise must strive for an integrated approach that effectively interconnects processes between raw imagery, analysis, and timely scientific products. Easily accessible storage will allow scientists and the wider

community to extract training datasets and develop pooled annotated libraries of imagery for supervised training of ML algorithms.

5. NOAA Data Enterprise

NOAA Fisheries has a dichotomy of fishery-independent imagery that must be publicly accessible, while fishery-dependent imagery has privacy compliance requirements under the Magnuson Stevens Act.

The NOAA data enterprise is continuing to execute successful efforts toward imagery accessibility and processing that align with the core NOAA mission of sharing knowledge with others, and conserving and managing coastal and marine ecosystems and resources. In 2018, NOAA had a total of 200 petabytes[5] of information residing on IT systems at any given time, coming from around 70,000 datasets. To address future access and use of existing and increasing data, the NOAA Big Data Project is using Cooperative Research and Development Agreements (CRADAs) to explore industry partnerships to understand best business practices for integrating cloud computing into storage solutions (Section 5.2). In addition, NCEI (NOAA's established environmental data archive) is working toward effective imagery transfer and preservation methodologies (Section 5.3). The following case studies for NOAA Fisheries and NOAA Office of Ocean of Exploration and Research (OER) offer challenges and lessons learned for imagery accessibility and integration of ML applications for timely and effective processing. NOAA Fisheries has improved image data collection and processing from fishery-independent surveys (Section 7) and fishery-dependent monitoring (Section 8). OER has made advances in storing and annotating imagery from the *Okeanos Explorer* operations (Section 9). The case studies below are presented in the context of the agency's mission and the NOAA Enterprise requirements to improve accessibility of big data imagery.

5.1 NOAA Enterprise Data Management Requirements

As a science-based government agency, NOAA has several requirements in place to aid in the preservation and availability of data, as well as the safety of information and assets to support the NOAA mission. NOAA data and imagery must adhere to NOAA's Public Access to Research Results (PARR) policies, a plan for public access to publicly funded research that was drafted by NOAA in 2013 as a response to a White House Office of Science and Technology policy memorandum (Holdren, 2013). Imagery archived by NOAA must observe the appraisal and archive policies adopted by NOAA in alignment with NARA requirements. There are also several data management

[5] 250 bytes or a million gigabytes

regulations and acts that must be considered for imagery storage and vendor partnerships and investments: Paperwork Reduction Act, Privacy Act, The Federal Acquisition Regulation (FAR 2010), The Federal Information Technology Acquisition Reform Act (FITARA), and The Federal Information Security Management Act (FISMA). NOAA's Procedural Directives[6] deliver guidelines for data management planning, data access, documentation, citation, scientific records appraisal, and data sharing. Additionally, NOAA's Environmental Data Management Framework[7] defines and categorizes the policies, requirements, activities, and technical considerations relevant to the management of observational data and derived products (Davis 2010).

5.2 NOAA Big Data Project

The NOAA Big Data Project is exploring CRADA partnerships with industry to utilize the cloud's scalable infrastructure to host and increase accessibility of NOAA data.

The Office of the Chief Information Officer is exploring the use of cloud storage and government/industry partnerships through Cooperative Research and Development Agreements (CRADA). To develop CRADA partnerships[8], NOAA has teamed up with five cloud providers to provide free public access to NOAA's data, while the companies create their own data products to monetize services based on free data and access to NOAA expertise. The cloud providers are: Amazon Web Services, Google Cloud Platform, IBM, Microsoft, and Open Commons Consortium. This is an ongoing study to explore the business case of utilizing industry partnerships for cloud models, and is part of a larger agency initiative to continue investigating the resources commercial cloud vendors can provide for NOAA's data storage, dissemination, and operational solutions.

Benefits: The NOAA Big Data Project will be able to utilize the scalability and existing online infrastructure to host big data while simultaneously increasing the usage of NOAA data (Figure 4) and cyber security (Figure 5).

[6] NOAA Procedural Directives:
NOAA Procedural Directives
[7] NOAA Environmental Data Management Framework:
NOAA Environmental Data Management Framework
[8] NOAA Cooperative Research and Development Agreements:
NOAA's Technology Partnerships Office's Cooperative Research and Development Agreements (CRADAs)
Contact: Ed Kearns, NOAA Chief Data Officer

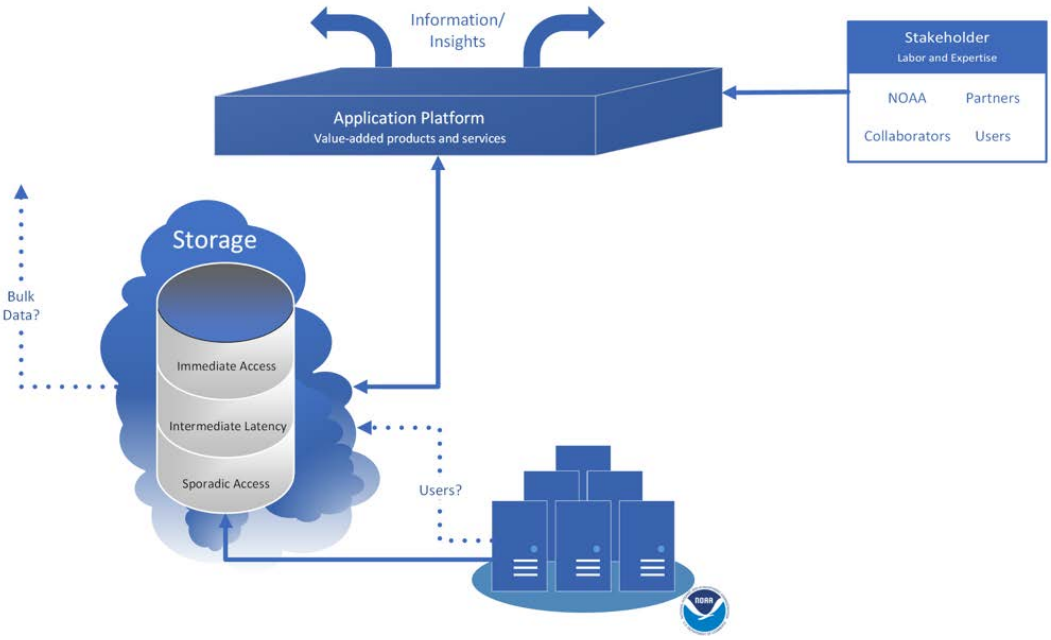


Figure 4. Federally verified and trusted third party commercial cloud providers can broker NOAA data that allows the agency to reallocate security assets and staffing to focus on mission-critical systems. *Credit: Big Data Project, NOAA Data Management Integration Team.*

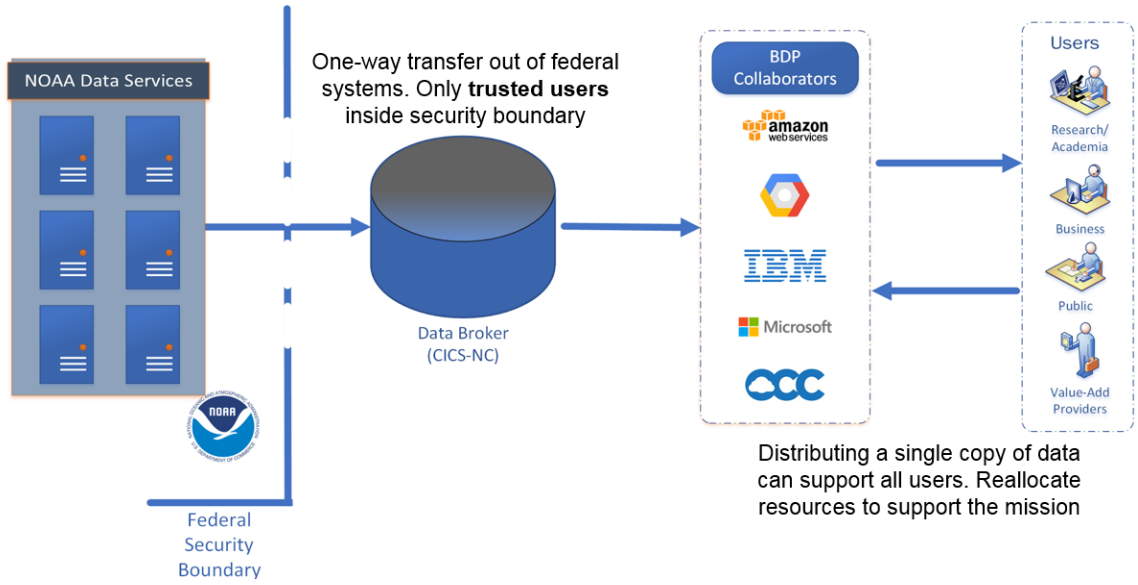


Figure 5. Federally verified and trusted third party commercial cloud providers have scalable platforms with quick data dissemination that allows for greater usage of NOAA data. *Credit: Big Data Project, NOAA Data Management Integration Team.*

5.3 National Centers for Environmental Information

NCEI hosts and provides public access to one of the most significant archives for environmental data, encompassing coastal, oceanic, atmospheric, and geophysical data on Earth[9]. NCEI has made great improvements to NOAA's data accessibility; for example, NCEI has centralized much of NOAA's underwater acoustic data, and currently has archived over 100 terabyte of water-column sonar data, of which more than 56 terabytes is NOAA Fisheries survey data. Presently, there are challenges for NCEI to centralize and host NOAA's big data imagery. The NCEI archives have petabytes of storage capacity, but an established, efficient and long-term pipeline to transfer imagery from NOAA to NCEI is still in development. Preliminary work was completed through NCEI's Video Data Management Pilot (VDMP) for NOAA Fisheries' underwater imagery data collected during 2014-2015 Untrawlable Habitat Strategic Initiative (UHSI) (Section 7.2). Challenges were identified with the UHSI metadata quality and efficiency in the transfer of the data to NCEI. The sheer volume of NOAA's video and digital still imagery data that reside at regional laboratories on external media (e.g., hard drives or data arrays) with varying levels of metadata completeness and format will need to be carefully addressed. Close collaboration between NCEI staff and NOAA Fisheries regional staffs will be necessary to ensure the video collections are properly and efficiently described, archived, and discoverable in future efforts to centralize and steward NOAA Fisheries big data imagery.

Evaluation of archival and accessibility requirements for big data imagery is underway by NOAA's National Center for Environmental Information, which hosts much of NOAA's environmental data.

One possible solution to improve imagery throughput is NOAA's Enterprise network, N-Wave (Figure 6).

[9] Current NCEI metadata records: *Current NCEI Data Collections*

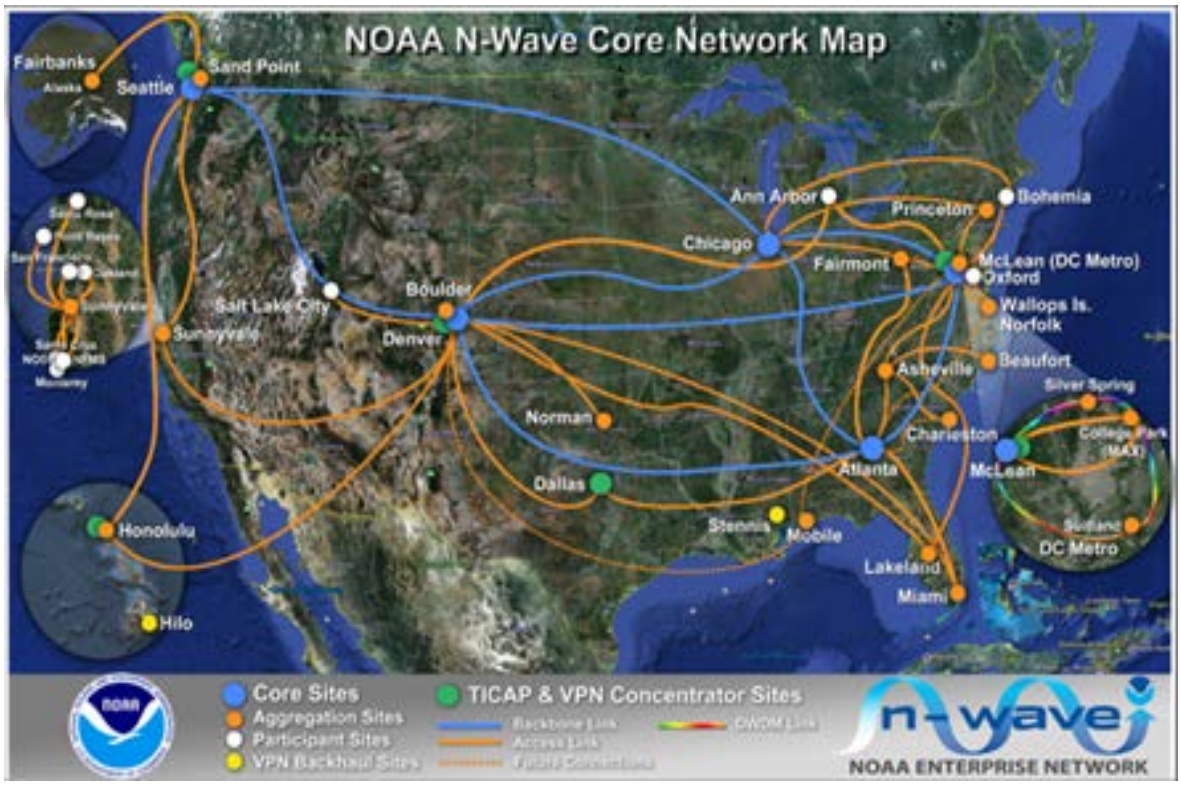


Figure 6. NOAA N-Wave Core Network map.

This scalable, secure network quickly transfers data using 10 gigabytes or 100 gigabytes per second Wave Division Multiplexed fiber-optic links. Methods to efficiently transfer large data from NOAA facilities to NCEI are still in development as the N-Wave connections are established and speed tested.

Most of NCEI's archived data are discoverable and accessible through dedicated web portals. These portals have a map interface, filter functions driven by the underlying metadata, and access capabilities. These tools were built with community input and enable users to discover and extract the datasets of interest using fields specific to each data type. Refined searches ensure that just the data and data files in which a user is interested are requested. This selectivity creates a more efficient system for the archive and the user. The NOAA Fisheries Video portal is an excellent example of NCEI data discoverability and accessibility. There, users can view the location on a map where data were collected, search based on the associated annotation, and view and download desired video clips (Section 7.2). The underlying metadata fields drive the enhanced search capabilities in these portals. Therefore, high quality, complete, and consistent metadata are vital to the discoverability and accessibility of high volume, complex datasets such as big data imagery through dedicated data portals.

NCEI has recently developed a generalized discovery tool called NOAA OneStop[10] in addition to the NCEI Geoportal Server[11] web interface (website links provided below). NCEI's metadata can also be searched via the following interoperable machine services: OneStop Search API, Catalog Service for the Web, OpenSearch, and Really Simple Syndication. In the future, the NOAA OneStop interface will replace the NOAA Data Catalog, enabling discovery of all of NOAA's cataloged data through a single interface.

6. Machine Learning

6.1 Benefits and Challenges of Machine Learning

Machine learning is a subfield of computer science that gives computers the ability to learn without explicitly being programmed (Samuel, 1959). In ML, computer programs learn from experience with respect to some class of task, and the performance measure improves from this experience (Mitchell, 1998). The field began to flourish in the 1990's as statistical methods and models were applied from probability theory (Langley, 2011). Machine learning has been increasingly integrated into a variety of fields that collect and process imagery. Industry, academic, and government entities across the globe have been utilizing ML to augment imagery processing during recent decades. Only recently has the development of ML become advanced enough to successfully evolve into a priority in environmental fields such as biological and habitat classification and resource assessments (Diesing et al., 2001; Puser et al., 2009; Lüdtke et al., 2012; Nian et al., 2014).

Machine learning software with trained algorithms decreases processing cost and effort for more timely and precise results.

Benefits of Machine Learning

Machine learning software can decrease the cost and effort associated with human-based processing of big data imagery. Well-trained algorithms have the capability to process imagery at a greater rate with higher precision than humans. In addition to improved quality and timeliness of scientific products, ML will likely increase public,

[10] NOAA's OneStop:
NOAA OneStop Portal Homepage
[11] NOAA's Geoportal:
NCEI Archives Search Page
NCEI Geoportal
Contact: Carrie Wall Bell, NCEI

scientific and other stakeholder use of the imagery and data beyond the original operational purpose. As ML continues to advance, its performance and capabilities will outperform those of humans. The resulting growth in research and discovery from ML will likely increase the value of scientific data. A summary of ML benefits are listed below.

Cost effective: Algorithms will take less time to process big data and therefore require fewer resources and human effort to complete annotations.

Complete processing: Machine learning can compute faster than humans, and therefore, will likely resolve the big data imagery backlog and process future collections of imagery and data in a timely manner.

Resolve difficult annotations: Detectors can be trained to rapidly identify object features or conduct pixel-level characterization beyond human capabilities at a faster rate.

Increased accuracy: Algorithms can identify difficult objects and address object occlusion by tracking individuals to increase the precision of identification, measurements, and other quantitative estimates.

Machine Learning Cycle

Big data accessibility must allow users to readily extract imagery to train classifier and detector algorithms (Figure 7).

Figure 7. The machine learning cycle uses annotated imagery to train machine learning classifier and detector algorithms. Analytics can be done with these new classifiers that enable annotated storage of imagery and can be run in a graphical user interface (GUI) through an iterative process.

Workflow using machine learning: Machine learning should be integrated with the analysis and annotation of imagery to greatly increase processing efficiencies (Figure 8).

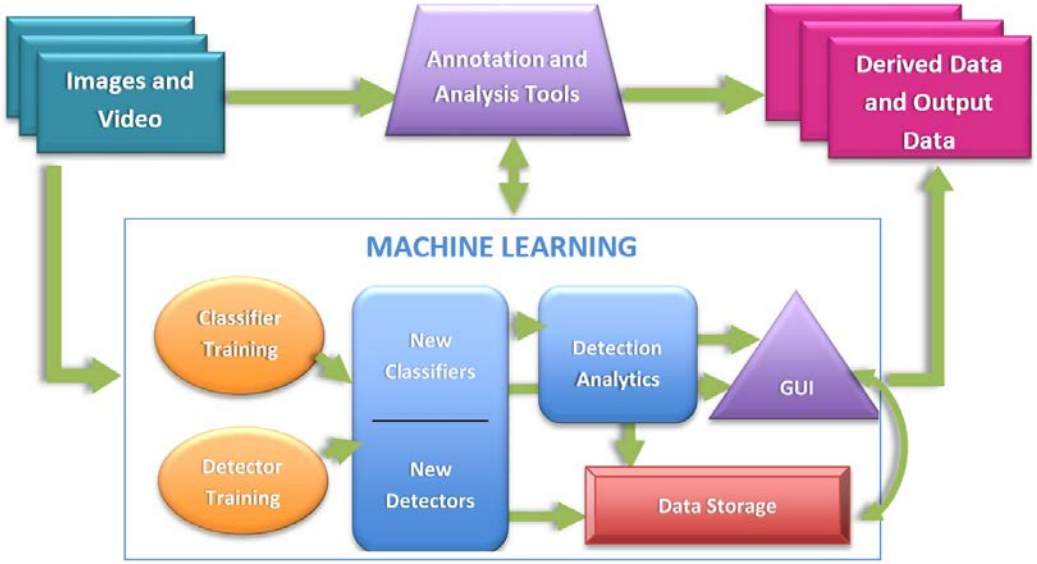


Figure 8. Imagery processing workflow with machine learning.

Challenges of Machine Learning

The topic of ML has become increasingly popular as analytical tools become more readily available and are necessary to keep up with 21st century big data processing. The wide-spread use and integration of AI and computer vision programs into the workflow yields great challenges, especially when applied to Big Data and Big Data imagery:

Data accessibility: Training sets must be optimal stored and accessible with enriched metadata.

Defining the Question: What discovery objectives can be applied to the data beyond the original research objective?

Unclear Representation: Without enriched metadata, the representation of data is unclear.

Computational Resources: Keeping up to date with fast advances with the cloud and GPUS's.

Selection of the Algorithm: Machine learning algorithms must be selected based on assumptions and rules; this is linked the statistical validation of the results.

6.2 Machine Learning in the Marine Environment

Image recognition has recently advanced by using deep learning models and convolution networks (Krizhevsky et al., 2012; Goodfellow et al., 2014; Simonyan and Zisserman 2014; Zeiler and Fergus 2014; Ren et al., 2015; Szegedy et al., 2015; He et al., 2016), which have the ability to label recognized objects (Karpathy and Fei-Fei 2014). Machine learning techniques in the wider world have advanced considerably and are routinely used for facial or thumbprint recognition, for example to unlock smartphones (Chen et al., 2014). When models to analyze large sets of information, such as ML algorithms applied to imagery, it is critical to conduct statistical analyses and validate the models to produce accurate scientific products (NASEM 2016; NRC 2013).

In applying ML to marine environments, there are challenges in training the algorithms to detect and classify living marine resources such as the turbidity of the water, direction of the marine organism or object, lighting variability from day/night to cloudy conditions, and other environmental factors. Despite challenges, scientists have explored the utility of ML algorithms to process imagery collected from the marine environment. Investigators have demonstrated the feasibly of utilizing ML algorithms in the field of marine science, including for topics such as fisheries acoustics, fisheries imagery, environmental conditions to predict fishery status, classifying plankton, and ageing fish (e.g. Benfield et al., 2007; Fernandes et al., 2010; Rodríguez et al., 2012; Uusitalo et al., 2016; Thomas et al., 2018). Machine learning software can be most useful when applied to imagery collected for estimating abundance and other measures of marine organisms and their habitat classification. Pertinent results have been demonstrated for underwater image processing and automated image recognition using deep neural networks in the workflow (Shortis et al., 2013, Salman et al., 2016, Shafait et al., 2016, Shafait et al., 2017, Siddiqui et al., 2017, Villon et al., 2018). Training sets and annotated imagery are crucial to increase the accuracy of ML, enhance the learning ability of the algorithm, and optimize its use for imagery processing. NOAA Fisheries' recently released Video and Image Analytics for the Marine Environment (VIAME) open source software is a good example of the processing efficiencies and precision that can be achieved with deep learning algorithms (Section 7.3). As more users utilize the VIAME software for various marine environment applications, it has become apparent that making big data imagery more accessible is urgent in order to create pooled annotated imagery collections and to more efficiently advance the capabilities of ML.

To provide timely scientific information for ocean policy decisions, there is an urgency to make big data imagery accessible to machine learning to resolve the bottleneck of costly processing time.

This urgency to advance ML applications is also recognized by the International Council for the Exploration of the Sea (ICES) which recently formed the ICES ML working group

to assess current international priorities in this field in relevance to the marine environment. Given that the ICES mission is focused on the sustainability of marine resources, this working group is tasked with providing technical guidance on improving the quality and timeliness of scientific products using ML. One priority is undoubtedly focused on increasing the use of ML algorithms, which will clearly be dependent on the accessibility of big data collected from the marine environment.

6.3 Video and Image Analytics for the Marine Environment

NOAA Fisheries recently completed its 2013-2018 Automated Image Analysis Strategic Initiative (AIASI) to resolve the big data imagery bottleneck by collaborating with computer vision experts to develop more efficient image processing software. The foundation for this initiative was constructed from a NMFS-funded workshop focusing on methods used to analyze imagery for stock assessments (NRC 2015). After 5 years, the AIASI delivered Video and Image Analytics for a Marine Environment[12] (VIAME): an open source framework for underwater image processing that utilizes ML. VIAME provides a crucial platform for the development of ML for various applications in the marine environment ranging from underwater visual fish surveys to aerial surveys of marine mammals. VIAME exemplifies ML, as it utilizes deep learning algorithms and computer vision technology to streamline the processing of imagery data, and provides more precise quantitative measures from automated detection, tracking, classification, and performance evaluation (Figure 9).

The open source VIAME software utilizes machine learning analytics to automatically detect and classify objects in images collected from the marine environment.

[12] VIAME Kitware Inc.Website*:*
VIAME Toolkit Homepage
VIAME GitHub:
VIAME Github Page
VIAME Instruction Document:
VIAME Instruction Document
Contacts: Ben Richards, PIFSC
Matt Dawkins, Kitware Inc.
Anthony Hoogs, Kitware Inc.

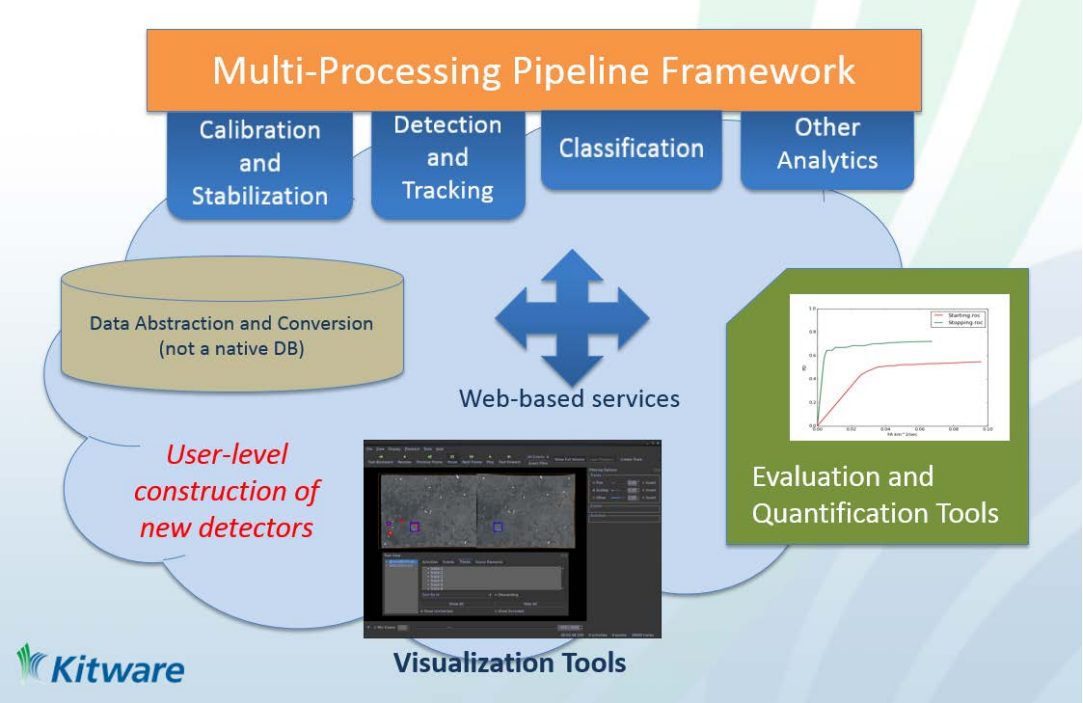


Figure 9. VIAME multi-processing pipeline framework allows users to enhance imagery processing by using detection and tracking, classification, visualization, and evaluation tools. *Credit: Kitware Inc.*

Optical training datasets are crucial for VIAME and its ML capabilities, as they enable the computer algorithms to learn and increase annotation accuracy. The VIAME toolbox applies ML algorithms to training sets from imagery data collected by NOAA's Fisheries Science Centers. Refer to Section 7 for further details on the NOAA Fisheries' imagery data used to develop and train the VIAME toolbox.

The Northeast Fisheries Science Center (NEFSC) has incorporated VIAME into its video analysis since early 2018. The HabCam data collected for the NEFSC Optical Scallop Survey is processed by VIAME to identify the presence of skates in the images. VIAME processes the datasets and produces a probability that an identified object is a skate. NOAA Fisheries' VIAME Toolbox will continue using imagery data sets to train the software and increase its accuracy of object identification, tracking, measurements, and more (Figures 10 and 11). As VIAME becomes more proficient, it will be able to increase efficiency in data processing by automatically annotating the optical data.

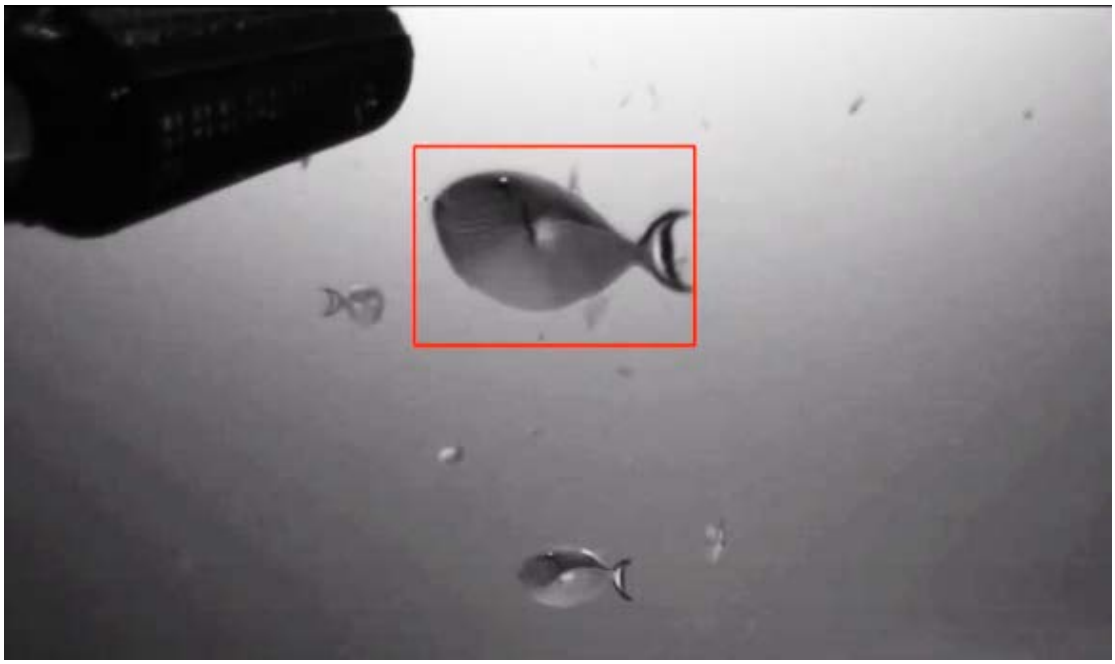


Figure 10. The VIAME object identification feature is able to detect objects of interest and specific targets once algorithms are trained through training imagery collections. *Credit: Kitware Inc.*

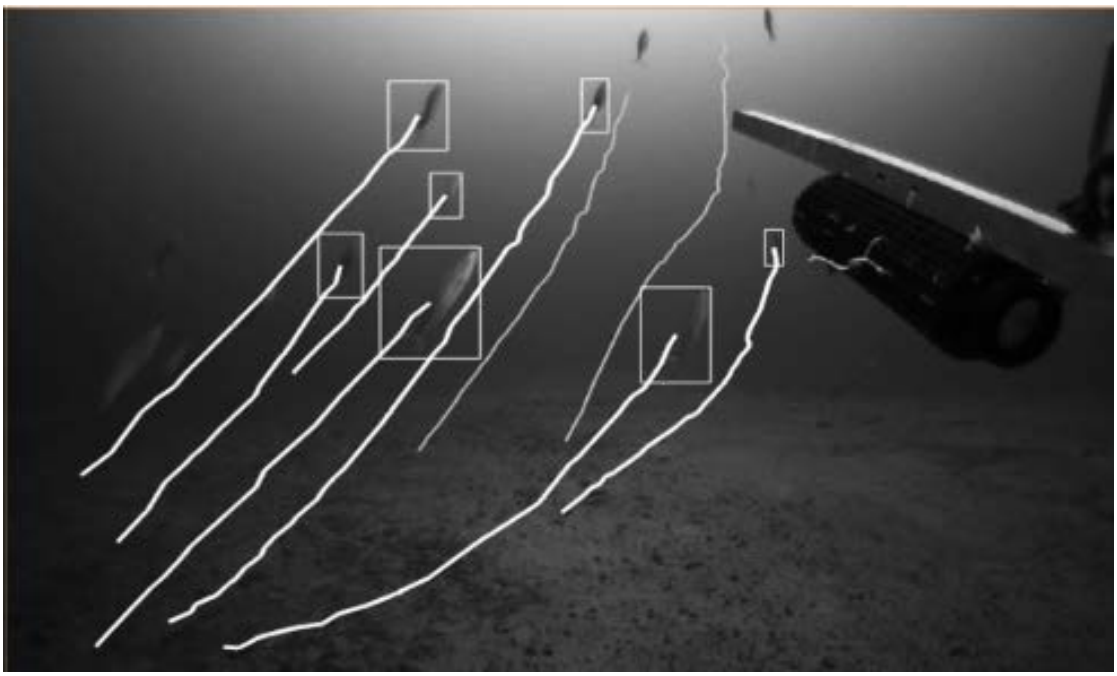


Figure 11. The VIAME track annotation feature creates tracks for individual objects in a frame. This will greatly aid in the processing of imagery, and helps determines if the same fish is swimming in and out of the frame thereby reducing double-counting uncertainties in abundance estimates. *Credit: Kitware Inc.*

The number of VIAME users has increased dramatically after training in its application was provided in 2018 at each regional Fisheries Science Center. The next step of this initiative during 2019 is to solicit recommendations on upgrades to VIAME to expand its applications for marine science and to determine how best to improve accessibility of data to ML analytics.

Refer to Section 10.1 CoralNet and Section 10.2 FLASK for two other NOAA Fisheries' AIASI-funded projects that have developed ML analytics for streamlining imagery processing for marine environment applications.

7. NOAA Fishery-Independent Surveys

7.1 Imagery Data from Regional Fishery-Independent Surveys

NOAA Fisheries is a science-based agency with mandates to provide the best scientific information available for policy decisions on the conservation and management of our nation's living marine resources. This drives NOAA's priorities to enhance its scientific monitoring programs with innovative sampling and analytical technologies (Lubchenco, 2012). With recent advancement of optical technologies for fisheries science (Mallet and Pelletier, 2014), NOAA Fisheries has dramatically increased its collection of digital images and video from its fishery-independent surveys. Although these visual surveys have helped to address data-limited stock assessments in untrawlable habitats, a bottleneck has developed in processing these big data imagery. The case studies in this section provide examples of NOAA Fisheries' progress with utilizing optical technologies to enhance fisheries survey operations. These examples emphasize the urgency of making big data imagery more accessible to new analytical tools such as ML in order to accelerate the processing of imagery to yield more precise and timely scientific information for management decisions.

NOAA Fisheries' ongoing efforts with optical imagery storage and localized management of the raw files and data is focused on accessibility of big data for research and discovery by the broader community. Projects collecting routine imagery for optical surveys are beginning to integrate ML into their workflows to significantly reduce the laborious manual processing.

Optical Scallop Survey: The NEFSC upgraded its dredge scallop survey[13] to an optical survey in 2012. This optical scallop survey has collected about 40 million image pairs, accounting for 200-250 terabytes of disk space stored on NEFSC servers. Copies of most

[13] NEFSC Optical Scallop Survey:
HabCam Homepage
Contact: Dvora Hart, NEFSC

of the images reside on additional servers at the Woods Hole Oceanographic Institution. Images are accessed within the year of collection for processing, but imagery collected from previous years may be accessed for additional purposes. Due to security concerns, the accessibility of these images has devolved from open-access via a web-based annotation platform, to local, NEFSC-only access via an annotation system within the center's firewall. The NEFSC is currently working on automated detectors for scallops and fish, such as skates, and improvements to processing and camera calibration procedures. The imagery from the optical scallop survey resides on regional storage servers, and its training datasets were used to develop and test the performance of the VIAME ML algorithms.

Bering Sea Pollock Survey: The Alaska Fisheries Science Center (AFSC) biennial Eastern Bering Sea (EBS) Pollock Survey[14] has collected roughly 2.02 terabytes (6 million images) of stereo camera imagery since 2012. An automated image analysis system is used from the cod-end of the survey trawl to extract fish species and length, and the output data is quality controlled before analysis. These data are loaded into the AFSC server database and used in EBS pollock stock assessments. In addition to the automated process, NMFS survey staff review and annotate the images using custom-developed software for rapid viewing of stereo still imagery. The images and the annotations are copied to a network drive at the AFSC and accessed occasionally by survey analysts to associate fish species and sizes with acoustic backscatter on an echogram. In general, historic data that are more than one year old and not actively being analyzed are rarely accessed. Prior to open-source automated image processing software, the AFSC used an automated fish length estimation processing routine written in the Matlab programming language, which has evolved over the years to include a species classification module. The stereo length estimation Matlab routine was transcoded into Python and included as an example with the VIAME software system.

The imagery data from the EBS pollock survey were used to train and test the performance of the VIAME toolbox; results of the VIAME process are comparable to the AFSC's Matlab output. The AFSC has not seen an advantage to changing their image analysis protocol to the VIAME toolbox at this time. Further analysis of the VIAME toolbox will be performed. This testing will require access to other imagery data within NOAA to fully evaluate the performance and quantitative output of VIAME. Current automated image-based size and species estimation processes have proven to be adequately precise for the intended use in the survey, so immediate research efforts into performance improvements are not high priority at this time. However, AFSC scientists are currently investigating the possibility of starting a camera-based survey of untrawlable rock habitat to complement existing trawl-based surveys of groundfish in

[14] AFSC Bering Sea Pollock Survey:
NOAA's Walleye "Alaska" Pollock Survey Information
Contact: Kresimir Williams, AFSC

the Gulf of Alaska, using similar stereo camera platforms. When this survey becomes active, it will rely heavily on automated methods to enable efficient data processing of large volumes of image data. The VIAME toolbox will be strongly considered for this task.

Hawaiian Bottomfish Survey: The underwater video survey of Hawaiian bottomfish[15] has 50 terabytes of data stored to date. As the data is only stored on individual hard drives and local servers at the Pacific Islands Fisheries Science Center (PIFSC), only PIFSC staff have access to the imagery. About 20 terabytes of imagery is accessed per year, usually within a 3-month time period of active processing after survey collections. These data were used for the development of the VIAME software, and the PIFSC is in the beta-testing stage of using ML for processing and is testing newly-trained detectors on human-annotated video. Processing will become more automated as ML detectors/classifiers are trained and tested (Section 6). Imagery collections will continue to grow at a rate of no less than 20 terabytes per year, and these data will likely continue to be stored on local hard drives until an alternative process to improve accessibility for ML can be identified and is available.

The PIFSC has also begun using VIAME to aid in annotation of modular optical underwater survey system (MOUSS) stereo-camera data from the Bottomfish Fishery-Independent Survey in Hawaii (BFISH). To assist in tuning VIAME detection and classification modules for the Hawaii Deep7 bottomfish complex (six species of deepwater snapper and one deepwater grouper), bounding boxes and track lines have been made for all species using the WAMI-Viewer semi-automated annotation module. Annotations with track lines should assist the software to identify fish moving over complex backgrounds where they may be difficult to distinguish from the substrate in still images. These training annotations were used to tune a species-specific VIAME convolutional neural network (CNN), which is currently being tested.

Gulf of Mexico Reef Fish Video Survey: The Southeast Fisheries Science Center (SEFSC) has long relied on its reef fish video survey[16] in the Gulf of Mexico for stock assessments, and to date, imagery from this survey has stored 148 terabytes in various formats. It is accessed by personnel in the SEFSC Laboratory in Pascagoula, Mississippi, and is occasionally requested by outside users. Data sets from the SEFSC reef fish video survey were used to train staff with the VIAME toolbox, and ML will be used to streamline their imagery processing. This includes a priority of building the imagery

---

[15] Pacific Islands Fisheries Science Hawaiian Bottomfish Survey:
Pacific Islands Fisheries Science Center Bottomfish Survey Webpage
Contact: Ben Richards, PIFSC

[16] Southeast Fisheries Science Center Gulf of Mexico Reef Fish Video Survey:
Southeast Fisheries Science Center Reef Fish Survey Page
Contact: Matt Campbell, SEFSC

library with pooled annotations to train the ML detection algorithms. The goal is to create separate detectors for color imagery, black and white imagery, and one for a mixture of the two. A subset of the imagery has been archived at NCEI and made accessible through a Fisheries Video Portal (Section 7.2). Making this video reef fish survey data more accessible to ML would benefit many institutions and government agencies working in the Gulf of Mexico and Caribbean regions.

The four survey examples presented above demonstrate the urgency for NOAA's Big Data Enterprise to prepare for improving storage and accessibility of its big data imagery for ML analytics. These case studies constitute a large portion of NOAA Fisheries' collected imagery data, although NOAA's six Fisheries Science Centers collect imagery from a number of projects, including surveys used for fisheries stock assessments. These projects vary by instrumentation, platform, ecosystem, sampling regime, data type, and volume of data collected (Appendix A).

Currently, imagery and imagery data are stored at the Science Centers. Many of these imagery and data sets are stored on local hard drives, if a third party would like access to the imagery, a primary contact at the Science Center needs to be identified, and then a hard drive with the data will be shipped to the potential user. Others archive their imagery data sets on the cloud or local servers. This variety of storage locations hinders the accessibility of the data beyond the scope of the project, and many times the data are only discoverable by those who know how and where to find it.

In addition to the four survey examples presented above, the NWFSC and SWFSC also collect and store imagery (Figure 12). The future collection and storage requirements for NOAA Fisheries will increase exponentially in upcoming years; therefore, the projected estimates in Figure 12 are likely underestimates. The key point is NOAA Fisheries' imagery data collections and processing demands will out-pace the agency's ability to produce scientific products from the big data imagery without access to analytics such as ML. A single survey using optical technology may have hundreds of hours of video or millions of images which can be costly and labor intensive to process manually. It is also apparent that the management of the imagery metadata at the regional Science Centers varies widely among its data managers, IT staff, or the scientists themselves. Consensus on the metadata standards, annotations and formats may be the first step for those promoting that big data imagery be centralized in more efficient and integrated storage systems that optimize accessibility. Improving the accessibility of the NOAA Fisheries' big data imagery would enable ML applications that produce more precise, cost-effective and timely scientific information. Additionally, increased accessibility to ML

Enhanced accessibility is required for large imagery collections in order to utilize analytics allowing for more timely scientific information for policy decisions on the management and conservation of living marine resources.

and larger datasets provides added value to NOAA's scientific products from the research and discovery by the broader scientific community.

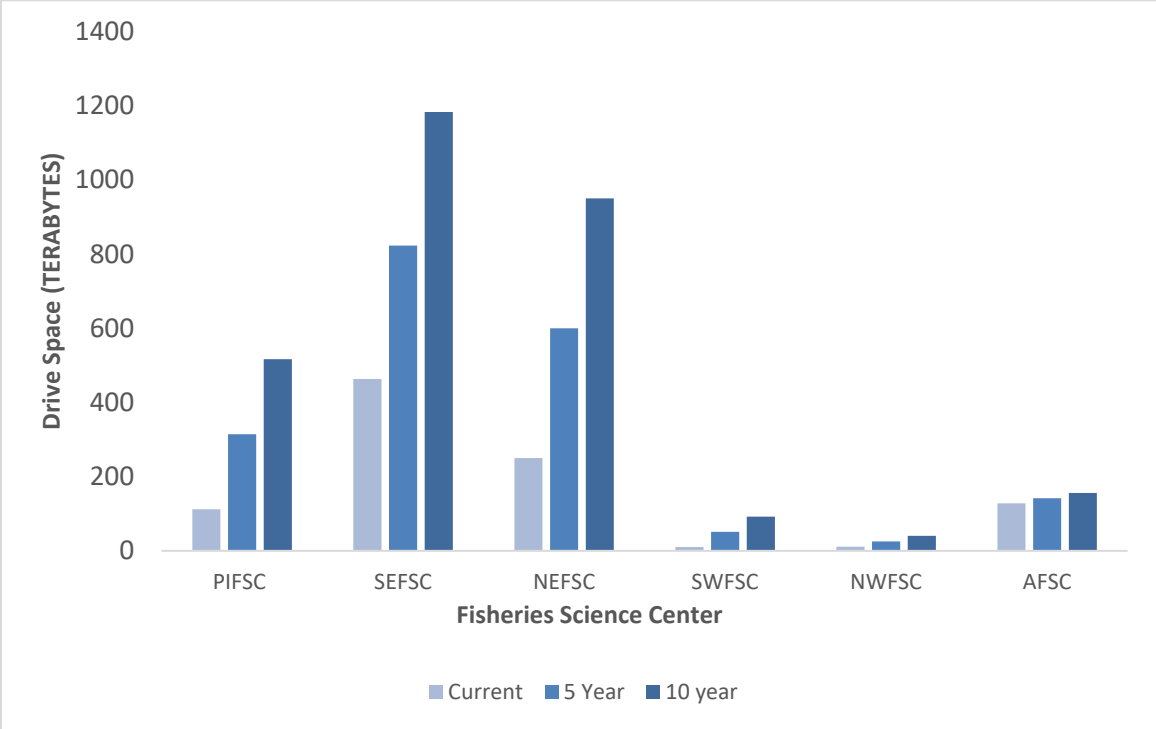


Figure 12. Current and projected (next 5 and 10 years) drive storage space are estimated required for archived imagery data from fishery-independent surveys for each NOAA Fisheries' regional Science Center (PIFSC: Pacific Islands Fisheries Science Center, SEFSC: Southeast Fisheries Science Center, NEFSC: Northeast Fisheries Science Center, SWFSC: Southwest Fisheries Science Center, NWFSC: Northwest Fisheries Science Center, AFSC: Alaska Fisheries Science Center). The recent increase in imagery data collected from the EM of commercial fishing operations is not included. For more details please see Appendix A.

7.2 Video Data from Untrawlable Habitat Strategic Initiative

The Video Data Management Pilot (VDMP) was established between NOAA Fisheries and NCEI to establish the framework for archiving big imagery data at NCEI using the 2014-2015 NOAA Fisheries Untrawlable Habitat Strategic Initiative (UHSI) video and imagery data collections. During the UHSI initiative, underwater video and still images were collected during roughly two months of tested studies to evaluate the sampling performance of various platforms and camera systems to obtain absolute abundance

estimates of marine fish. Approximately 13 terabytes of imagery data (encompassing over 5.5 million files) were collected, and these imagery data were made available to the NCEI VDMP project staff to evaluate their storage, archival and accessibility requirements. The NCEI VDMP project staff examined NOAA's OER Video Data Management Modernization Initiative (Section 9.2) when developing the metadata schema, associated database, and prototype NOAA Fisheries Video Portal for the UHSI imagery data. Due to the volume of data, a hybrid approach was chosen to archive and access the data collections at NCEI. The millions of still images and large, high resolution videos were archived on cheaper, less accessible tape storage. While the smaller volume of manageable short video clips was archived on tape as well as stored on spinning disk for live viewing and immediate download from the video portal. This practice matches that of the OER video data management (Section 9.1) and offers a cost effective and efficient compromise to long-term storage and easy access. The video portal[17] provides geospatially-based access to annotated video data (Figures 13 and 14).

For imagery data collected from the marine environment, standardized data collection protocols with metadata guidance must be followed to optimize data processing and accessibility for analytics.

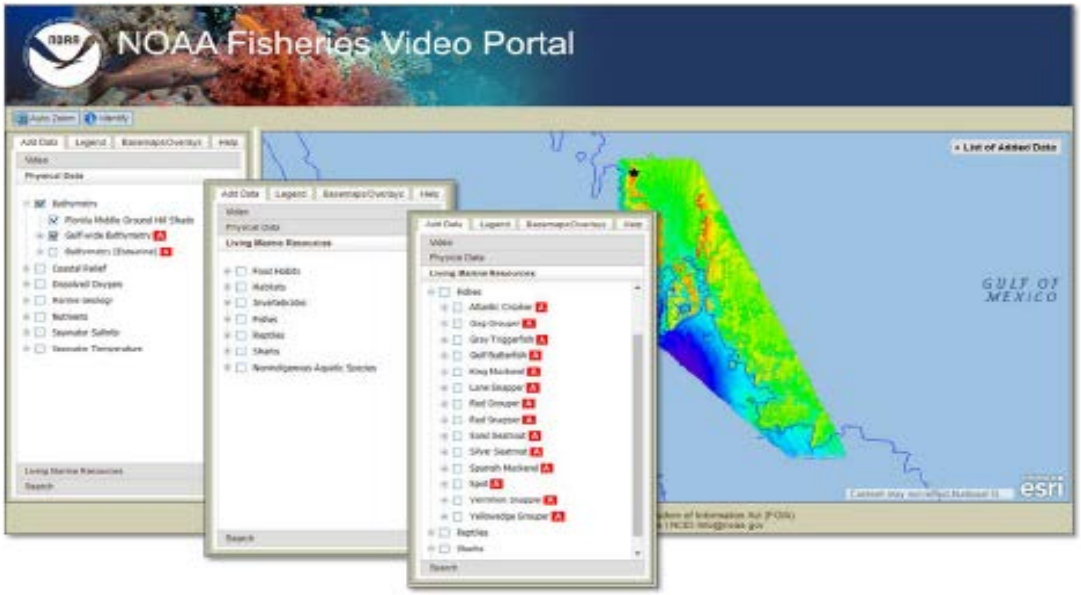


Figure 13. The above screenshots depict the geospatial services included within the NOAA Fisheries Video Portal from the Gulf of Mexico Digital Data Atlas. *Credit: NOAA Fisheries Video Data Management Project.*

[17] NOAA Fisheries Video portal prototype:
NOAA Fisheries Imagery Portal
Contact: Carrie Wall Bell, NCEI

Figure 14. The above screenshot shows the tabular information presented with annotated video in the pop-up window. A video and download button appear when the user hovers the mouse over the video pop-up window. *Credit: NOAA Fisheries Video Data Management Projec*t.

Additional geospatial services are embedded in the video portal to provide a holistic understanding of the environment, surrounding species and habitat when examining the video.

Upon completion of the project, the NCEI VDMP project staff provided recommendations for imagery archival and georeferenced accessibility, which are summarized below. Many challenges identified in this project are similar to those faced by other managers of big data imagery. The NCEI VDMP project staff found inconsistencies with NOAA Fisheries' data management structure, formats and metadata for the imagery data collections within this one UHSI project. This lack of consistency suggests that consensus is needed on the implementation of a standardized sampling protocol and metadata framework when collecting underwater imagery data across NOAA and potentially the scientific community.

Annotations are another important aspect of the imagery data management framework that are critical for data accessibility, research, and discovery. For example, time-stamped annotations help reference specific events within a video and allows

synchronization with other relational databases. Annotated video segments with lost coordinate information cannot be supported by the map-based portal. Critical locational information was lost due to camera coordinate information absent from the data spreadsheet. The VDMP found not all video formats and codecs are supported by modern web browsers. The pilot dataset contained a mixture of MPG, AVI, and MP4 video file formats. MP4 (H.264 codec) is the only format natively read by all modern web browsers. MPG and AVI videos are not widely supported by most modern web browsers (e.g., Firefox, Internet Explorer, Safari). Challenges arose with larger files (e.g., 25 gigabytes and above). To overcome this issue the imagery files were split into more manageable sizes and data collections with large file counts were aggregated. Overall, the VDMP found the OER video data management model (Section 9) too specialized for OER's exploratory surveys to easily support highly variable optical data consistent with the big data imagery collections throughout NOAA Fisheries. The methodologies and findings from the VDMP are important to consider when moving forward with centralizing imagery storage and accessibility. The Fisheries Video Portal is an exemplary initial effort of a user-friendly portal that allows access to geo-referenced optical data.

Challenges

Inconsistencies within a data collection and file formats: It is difficult and may not be possible to store and access imagery if they are not stored in the same formats as one another.

Costly to prepare historic imagery: It will take more effort and resources to alter historic imagery than to create systematic standards moving forward.

High-volume datasets: Most imagery data are high-volume and need substantial effort to initially store, process and make accessible unless the data collection follows standardized data structure, formatting, and appropriate metadata, enabling analytical tools to streamline and automate processing and analysis.

Lost coordinate information: Imagery cannot be made available through a map-based portal without coordinates.

Lessons Learned

Consistent files and names: File and directory names should be consistent.

Time stamps: Annotations need to be UTC time-stamped (Coordinated Universal Time).

Coordinates: Latitudinal and longitudinal information in data records allows discovery by map-based portals. It also provides a reference to the local time when data were collected.

Video formats: Not all video formats and codecs are supported by modern web browsers.

Chunking: For storage at NCEI, individual files larger than 25 gigabytes data must be split into smaller, more manageable sizes.

Other management models: In this instance, the OER video data management model is too specialized to be used with other imagery.

8. Fishery-Dependent Electronic Monitoring

Commercial fisheries are rapidly moving to incorporate EM into existing fishery-dependent data collection programs to augment or replace human observers in a number of regions, and collection of imagery is dramatically increasing. In 2018, Congress appropriated ~ $7 million to implement EM and electronic reporting (ER) aboard fishery-dependent surveys (e.g., commercial fishing operations). The funds are split about evenly to the National Fish and Wildlife Foundation[18] (NFWF) and NOAA[19] to fund external and internal projects to further implement EM and ER in commercial and recreational fisheries in the U.S. Funded projects address big data imagery collections and processing with ML algorithms.

Challenges moving forward with the implementation of EM and the reason for Congressional appropriations to further implementation of EM and ER beginning in 2016 include the high costs of video transmission and storage, the current need for labor intensive (and thus costly) human video review, as well as the need to improve the timeliness of data management associated with new EM data streams. Any observer[20] data collected by the agency is confidential under the Magnuson-Stevens Act (MSA); therefore a barrier remains between the imagery data and public access. Challenges also arise when providing access to NOAA Fisheries partners to develop ML algorithms to expedite imagery processing. Accessibility of imagery data will be

Recent surges in funding and imagery collections for fishery electronic monitoring require a balance between accessibility for processing with analytics and privacy constraints.

[18] NFWF's grant for electronic monitoring and reporting: The National Fish and Wildlife Foundation Grant to Support Electronic Technologies

[19] NOAA Fisheries Electronic Monitoring:
NOAA Office of Science and Technology Electronic Monitoring and Reporting Webpage
Contacts: Farron Wallace, AFSC, Brett Alger, OST

[20] NOAA Fisheries National Observer Program:
NOAA Fisheries Fishery Observer Program

important to promote the development of ML applications that will decrease processing cost and effort and ensure timely fishery data (Figure 15).

Figure 15. Data and imagery collected from fishing boats have confidentiality agreements and therefore public access should be provided with caution. It is critical that EM imagery be available for processing for ML, as fisheries observers are costly and their work is often dangerous.

Challenges of Electronic Monitoring

Video Costs: High costs are associated with video transmission and storage.

Labor: Processing video is labor intensive, costly and requires improvements for efficiencies and timeliness of information.

Confidentiality: Observer data are confidential under the Magnuson-Stevens Act.

Video formatting: There are no national standards for EM. As some vendors use proprietary software, lack of national standardization could lead to issues related to aggregating data for long-term storage and access. NOAA Fisheries is developing national standards for EM data.

NOAA Fisheries' science centers have increased and will continue to increase the collection of imagery data from fishery EM operations, and access to ML is necessary to address the potential processing backlog of these big data imagery. Researchers at the AFSC are developing machine vision systems for a chute and stereo camera tool that incorporates ML to automate image processing. The chute system is moving steadily toward a mature technology, for future implementation. Automated image processing currently provides real time image analyses for monitoring and estimating halibut discard on trawl vessels, providing immediate feedback to the vessel operator after each haul via a simple text file summarizing catch statistics. Machine learning algorithms also provide onboard image analyses to evaluate image quality (good image quality = precise length), identify catch events, and provide count and weight (inferred from length) of discarded fish (Figure 16). This tool can also provide species identification with high

accuracy for 45 species commonly caught along the West coast and Alaska groundfish trawl fisheries.

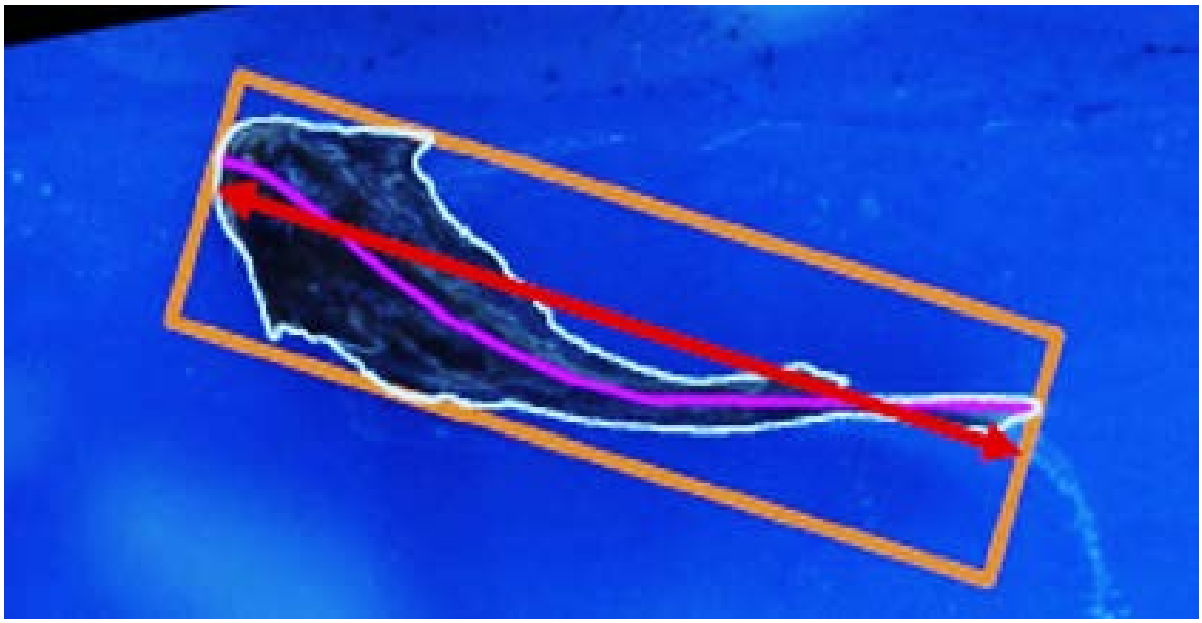


Figure 16. Screen capture dorsal view of the automated midline measurement of a fish in the camera chute system.

Machine learning algorithms are being used in stereo imagery, which can provide length measurements while fish are being hauled on longline vessels (Figure 17). Recent advances in stereo camera ML include accurate species identification for the three most commonly caught species (Pacific halibut, sablefish and dogfish), determination of disposition (discarded or retained) and catch event detection. Machine learning algorithms to estimate catch volume of trawl nets and sorting tables (pot vessels) are being developed to reduce costs and the workload associated with fishery observers.

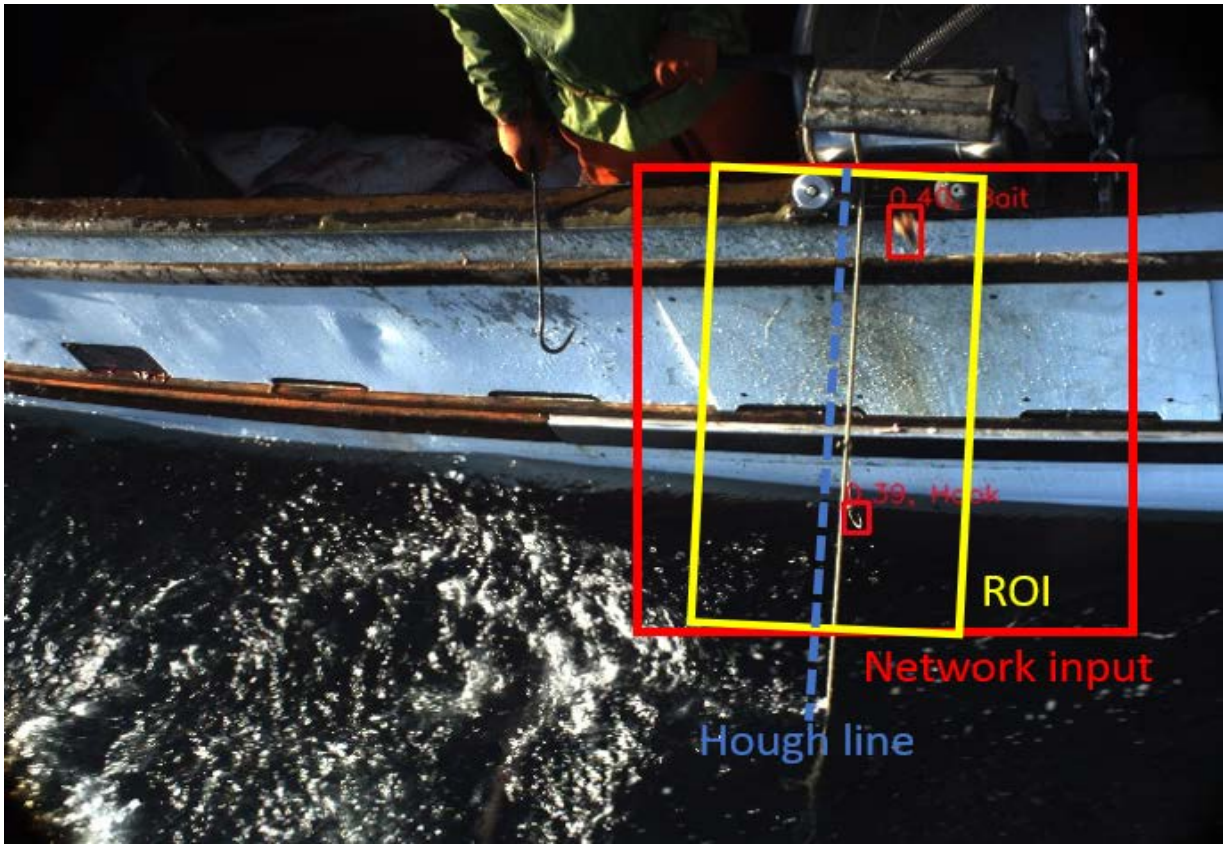


Figure 17. Example of region of interest (ROI) and annotation of hooks (small red box) during processing of longline stereo camera imagery.

A challenge faced by EM researchers is that few have experience with annotation techniques or developing ML algorithms. An advantage of ML algorithms is that they can be trained on other image datasets for any other fisheries, providing a cost-effective transfer of technology especially where imagery data are similar.

Currently, approximately 760 terabytes of imagery have already been collected by existing EM programs (Table 2). As these programs continue to collect imagery, and implement new programs in the future, the storage and accessibility needs must be addressed.

Table 2. Total Electronic Monitoring Video Storage through 2018.

#	Region	Fishery	Total Storage of all Video (in TB)
1	AK	Bering Sea and Aleutian Island (BSAI) Non-Pollock Trawl Catcher/Processor (C/P)	1
2	AK	Bering Sea Pollock trawl Catcher/Processors and motherships	1
3	AK	Central Gulf of Alaska Rockfish Trawl C/P	1
4	AK	BSAI Pacific Cod Longline C/P	1
5	AK	BSAI Halibut/Sablefish Longline	40
6	AK	Small boat pot	N/A *program is in first year of implementation*
7	WC	Whiting at-sea mid-water trawl	130
8	WC	Shore based whiting and non-whiting mid-water trawl	
9	WC	Fixed gear IFQ	14
10	WC	Groundfish bottom trawl	30
11	NE	Groundfish sectors - Audit Model Project	47
12	NE	Mid-water trawl	45
13	HMS	Pelagic longline	360
14	SE	Shrimp Trawl	2
15	PI	Pacific Longline	90
		Total:	759 TB

Big data imagery will increase dramatically when NOAA Fisheries implements the following new regional EM programs:

West Coast: Whiting midwater trawl and fixed gear (2019)

West Coast: Bottom trawl and non-whiting midwater trawl (2019)

Northeast: Herring mid-water trawl (2020)

Northeast: Groundfish fishery implementation (2021)

9. NOAA Ocean Exploration Research Program

9.1 Imagery Storage and Retrieval from the NOAA Ship *Okeanos Explorer*

NOAA's OER program focuses on exploration and discovery to understand the world's oceans. It uses cutting edge technologies and methodologies to enhance research, policy, and management decisions to develop new lines of scientific inquiry. The premier vessel used in exploration missions, the NOAA Ship *Okeanos Explorer*, uses optical technologies to collect imagery that supports OER's mission to increase the pace, scope and efficiency of ocean exploration. Its imagery data is also used for education. The ship has been collecting imagery data for NOAA missions since 2010. By 2016, 120 terabytes of imagery and associated products have been collected. This amount of imagery data and need for accessibility and dissemination motivated a Video Data Management Modernization Initiative[21] (VDMMI) with the goal of creating management methods[22] for the data.

Comprehensive storage systems and user-friendly portals are critical for widespread access to imagery and related data.

Video Data Management Modernization Initiative

In 2016, OER completed a VDMMI project to investigate modern methods of video data management. The high volume of video collected by the *Okeanos Explorer* is in demand by a wide audience of scientists, broadcast journalism and academia, while at the same time faced challenges with accessibility as the old video data management systems were based on a physical media model. The VDMMI has led to an ongoing partnership and collaboration with NCEI with a dedicated pipeline, archive, and access mechanism specific to the *Okeanos Explorer* imagery data that uses both the NCEI spinning disk system as well as NOAA's Comprehensive Large Array-Data Stewardship System (CLASS) infrastructure (Figure 18).

[21] NOAA OER's VDMMI report:
Video Data Management Modrenization Initiative Report
[22] NOAA OER's best video data management practices:
OER Video Data Management Best Practices Document

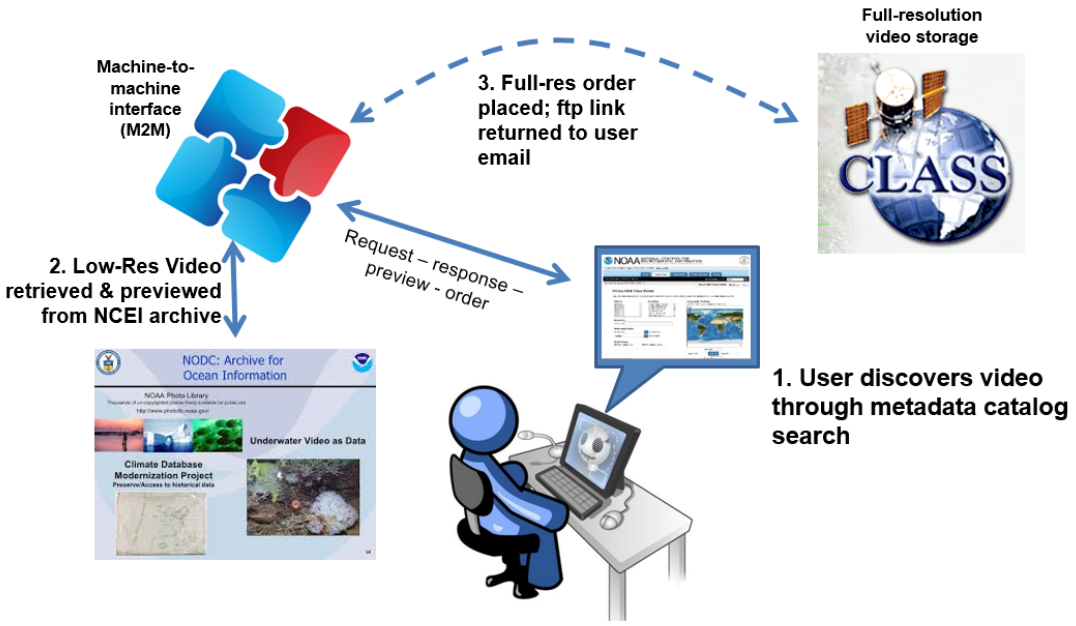


Figure 18. Dataflow of OER video data. The CLASS storage and retrieval system was chosen as the video repository. *Credit: NOAA Office of Exploration and Research.*

Another focus of VDMMI is to make video and its data easily accessible to the public. Public accessibility maximizes the imagery and data's value as it can be leveraged by advances in technology and is open to social uses of scientific data. The key for discovering the *Okeanos Explorer* video data is detailed metadata coupled with a metadata search portal that efficiently filters for criteria designated by the user. The cruise metadata is gathered in planning stages and saved in a database called the Cruise Information Management System (CIMS). CIMS outputs ISO collection-level metadata record. To date, OER has stored over 91 terabytes of imagery, comprising about 72,406 video segments discoverable through the OER Video Portal[23]. In 2018, over 20 terabytes of the video and products were downloaded for use. There are currently no computer vision applications automating annotations or processing.

Challenges

Physical media: There is limited access to imagery data on physical media and it is at risk for deterioration.

[23] *NOAA OER video data portal:* *NOAA NCEI OER Video Portal*

High-volume datasets: Most imagery data are high-volume and need substantial effort to initially store and make accessible.

High-demand to wide audience: *Okeanos Explorer* imagery and data are in high demand by scientists and the general public interested in ocean exploration.

User Portal: Web-based portals are crucial for discovery and access of imagery and may need to be created if not already available.

Lessons learned

Metadata: Metadata should be reliable and complete for best storage and access.

Video quality: Video should be collected at the highest quality and lowest levels of compression.

Pilot projects: Smaller pilot projects should be completed first before large scale implementation.

9.2 Imagery Annotations from the NOAA Ship *Okeanos Explorer*

SeaScribe Annotations and SeaTube Accessibility

NOAA's OER has also made advancements with the annotations of the *Okeanos Explorer* imagery. Within the last several years, the OER has worked closely with the NOAA Fisheries Deep Sea Coral Program and other partners to understand imagery data requirements and improve real-time annotations. The relationship between video collection, annotation, and management are highlighted in the conceptual diagram below.

Effective real-time annotation systems utilized during imagery collections enhance post-processing and analyses for research and discovery.

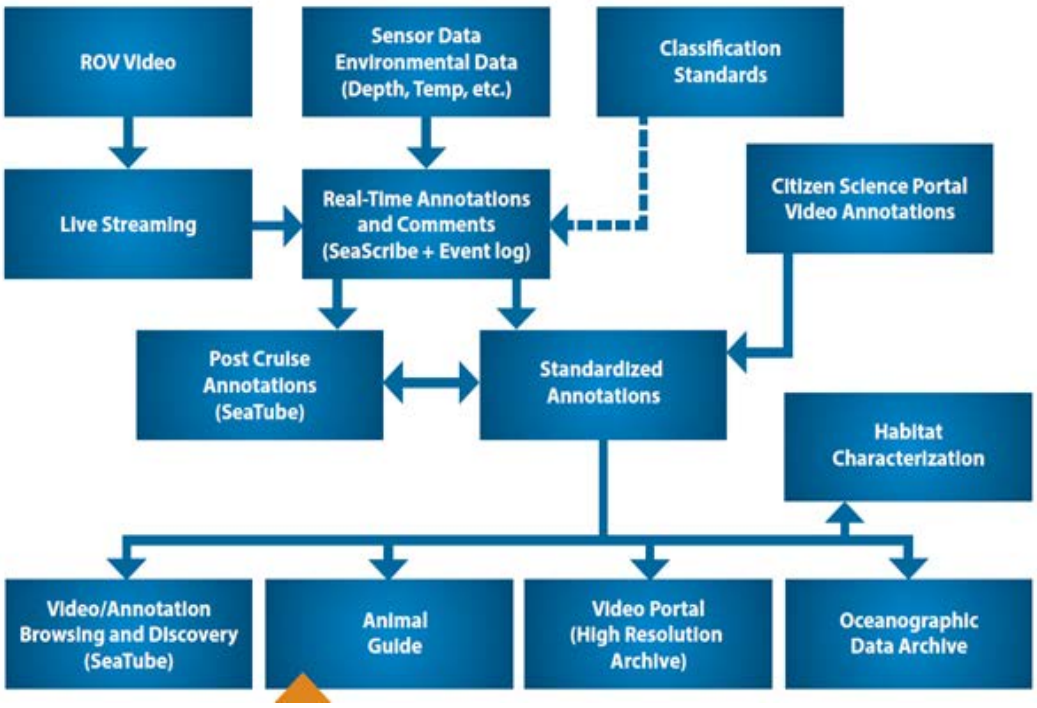


Figure 19. Conceptual diagram of connections between video data collection, annotation, and management. *Credit: NOAA Office of Ocean Exploration and Research.*

Since 2010, OER invited experts to participate remotely in the remotely operated vehicle (ROV) dives to provide these video annotations. In 2017, OER partnered with Ocean Networks Canada to implement annotation software tools that provide an end-to-end annotation workflow for the scientists to create, review, and validate video observations. SeaScribe[24] is an online annotation application that shore-based and shipboard participants can access concurrently during an ROV dive. SeaTube[25] is a cloud-based video archive and browsing interface that enables playback of previously recorded videos and entry of new annotations (Figure 20).

[24] SeaScribe Annotation Software:
OER SeaScribe Overview
[25] SeaTube Software*:*
Ocean Networks Canada SeaTube Software

Figure 20. Example screen from the annotation software SeaTube showing the user interface that enables video playback, position, and time-stamped observations. *Credit: NOAA Office of Exploration and Research and Ocean Networks Canada.*

These tools reduce post-cruise annotation time and improve the utility of OER video for enhanced insight into explored areas. Annotations thus provide a rich data set that can be "mined" by many different scientists for different purposes.

Through a NOAA Big Earth Data Initiative (BEDI) project, additional work with SeaTube was completed in July 2018 by Mississippi State University (MSU) and NCEI. The project created an in-depth interface with collaborative team-based annotations, and implemented standardized vocabulary and a standardized CSV table output. This project determined the feasibility of using SeaTube as an annotation for deep sea coral data, using SeaTube outputs to populate databases and archive metadata, and potentially for the future use of training ML for automated image analysis annotations.

10. Extramural Efforts to Improve Imagery Processing

10.1 CoralNet

Scientists at the University of California San Diego (UCSD) developed the CoralNet[26] software as a web-based image classification system for the world's coral reef habitats.

[26] CoralNet Website:
CoralNet Website

NOAA Fisheries' AIASI provided partial support to the CoralNet program to improve the use of deep learning for coral habitat classification. The CoralNet system was transferred from a single host to a scalable distributed system on Amazon Web Services to handle large datasets, and workflows were improved to process images 10 times faster using ML. For example, the CoralNet software can reduce the 10-week manual processing of 1200 images from a diver coral reef survey to only 1 week of processing time using deep learning algorithms that automate classification.

CoralNet preserves many desirable characteristics of Coral Point Count with Excel extensions (CPCe), including a familiar interface, the ability for users to create a unique set of target descriptor codes, a function to overlay points randomly, and no acquisition or usage fees. The flat data structure used by CoralNet removes the inherent file structure problem in CPCe. Image metadata and annotations can be downloaded and archived and images can be randomly assigned to different analysts, a desired feature that was not possible using CPCe. The web-based deployment of CoralNet also makes it possible to easily collaborate with remote analysts.

CoralNet Alpha allowed users to upload image datasets, randomly distributed annotation points across those images, manually annotate a subset of the images using a web interface with study-specific labels (e.g., functional groups), and use those manual annotations as training data. It then automatically proposed labels for annotation points across the rest of the images and allowed users to verify and correct the proposals. In estimation of coral cover at the functional group level, CoralNet Alpha achieved a level of accuracy commensurate with human analysts (Beijbom et al., 2015), but challenges remained in identifying algal classes and many coral species. CoralNet Alpha characterized intra- and inter-expert variation, and found significant variation, particularly amongst algal classes.

Analytical tools, such as machine learning, are readily available and being used to classify benthic images from coral reef surveys.

The transition from CoralNet Alpha to CoralNet Beta resulted in significant improvements. Cloud-based processing significantly increased throughput, decreased latency, and reduced model training time. The transition from support-vector machines (SVM), a supervised learning model, to Deep Learning (deep CNN), layers of models, increased classification accuracy and reduced end-to-end human analyst effort.

CoralNet VIMEO channel:
CoralNet VIMEO Channel
Contacts: Ben Richards, PIFSC
Oscar Beijbom, UCSD
David Kriegman, UCSD

Through 2018, 550,000 images have been uploaded to CoralNet Beta from 629 sources from around the globe, comprising over 20 million annotations. Currently, CoralNet supports nearly 1,000 registered users with over 1,000 images uploaded and analyzed every day. Of the 550,000 images, over 100,000 images are from NOAA.

PIFSC has implemented the CoralNet tool for operational annotation of benthic photoquadrat imagery from Reef Assessment and Monitoring Program surveys in the U.S. Pacific Island region. In trials using manually annotated imagery from the Main Hawaiian Islands and American Samoa, a trained CoralNet model was able to accurately assess site-level coral cover (Figure 21), and its performance showed highly comparable results to those generated by human analysts. CoralNet was also effective at estimating cover of common coral genera, while performance was mixed for other groups including algal categories.

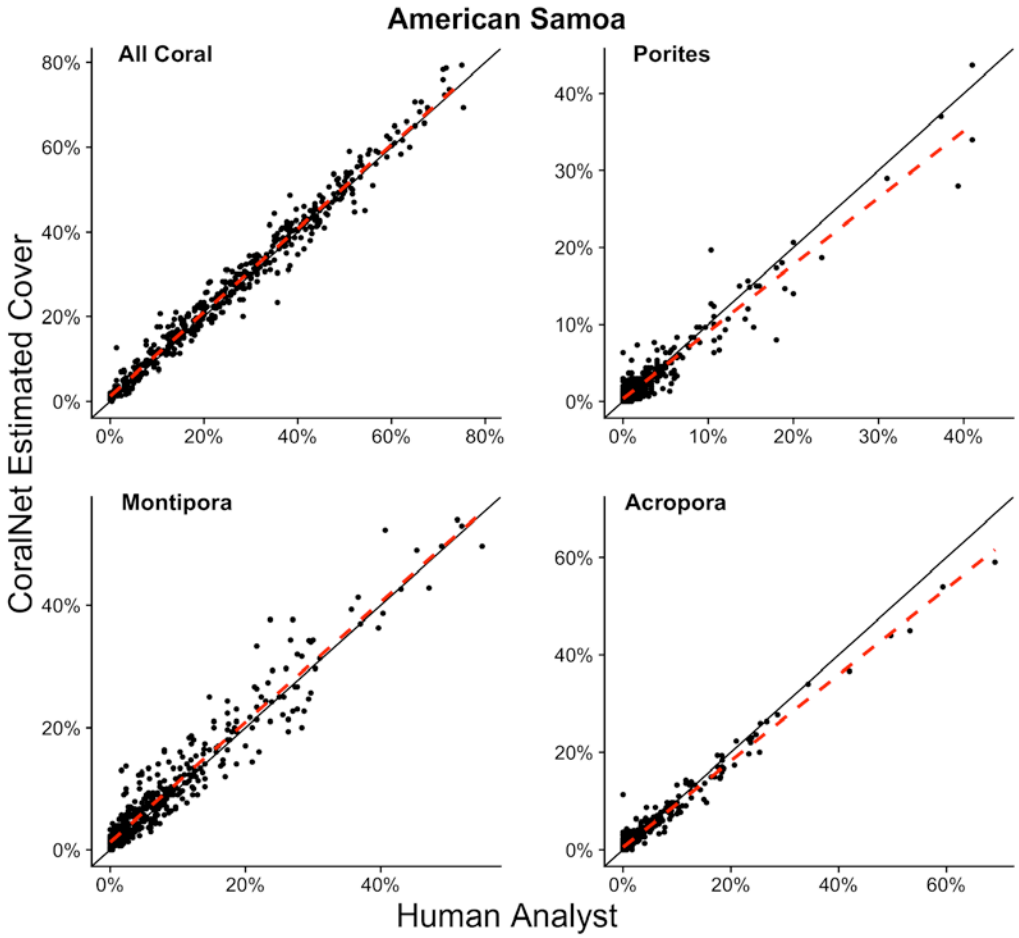


Figure 21. Site-level coral cover computed via manual (human) and automated (CoralNet) analysis for all coral and for common coral genera. Data comes from sites in American Samoa, surveyed by NOAA PIFSC in 2015. The solid black line is the 1:1 line, the dashed red line is a linear fit of the point data.

10.2 Flask

The NOAA Fisheries' AIASI initiative provided partial funding in the development of the Flask program. FLASK[27] was designed as a means for scientists to rapidly count and classify fish observed in optical surveys using remote camera systems. Initial design and development determined that recently available neural network capabilities for automated image processing to automate fish counting and classification. It became clear that many groups possessed large image datasets but without the bounding-box-level and specific ontology-driven fish annotations required to provide usable training data for algorithm development. To meet this need, the Stanford Research Institute (SRI) developed a semi-automated rapid annotation tool that allows analysts to rapidly ingest and annotate their video as they train a novel neural network. The neural network is then used for performing rapid classification of other raw video sources and produces JavaScript Object Notation (JSON) or hierarchical data format 5 (HDF5) output files containing all region of interest (ROI) bounding boxes with specific fish types. These data can then be parsed to provide fish types and counts. FLASK's neural network framework provides preprocessing functionality to identify key fish features and segment fish images to allow for interactive annotation and training. The tool framework is able to rapidly pool similar features into clusters, which can then be viewed through a Graphical User Interface (GUI) that allows for rapid annotation of hundreds to thousands of segmented objects simultaneously (Figure 22).

Artificial intelligence integrated into Flask software, rapidly annotates, counts and classifies fish in underwater optical surveys.

[27] SRI International's Flask website: *Flask Software*
Contacts: Michael Piacentino, SRI
David Zhang, SRI
Ben Richards, PIFSC

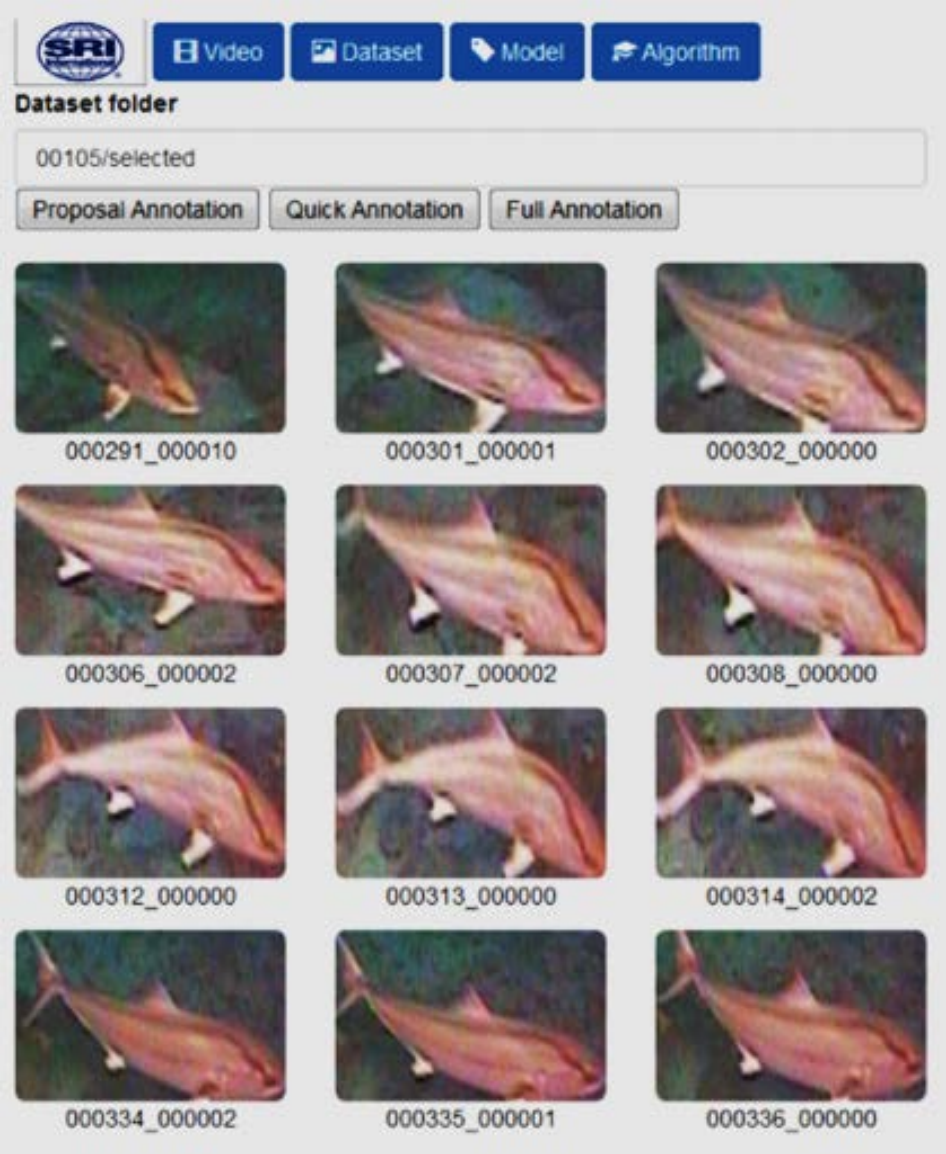


Figure 22. Flask's rapid annotation of objects clustered together with similar features.

In an iterative manner with just a few passes, as large clusters of fish are annotated, retraining of Flask's neural network occurs. After a few iterations of training cycles the network can be used to automatically provide labeled bounding boxes for any video with fish or other objects similar to the trained data. The HDF5 output can be loaded to directly overlay the fish classifications (text with bounding box) and bounding boxes directly on the source videos for reviewing results.

As researchers use the Flask tool they are quickly realizing its benefits and coming up with new features they would like to add to the tool framework. Recent extensions

include the ability to train on different video collections for a more diverse training data set and adding neural network temporal object tracking to improve fish counting performance.

Figure 23 shows three detected fish of two distinct species provided with the ROI coordinates. Each distinct ROI allows independent fish tracking even when one fish is partly occluded by the other as shown above.

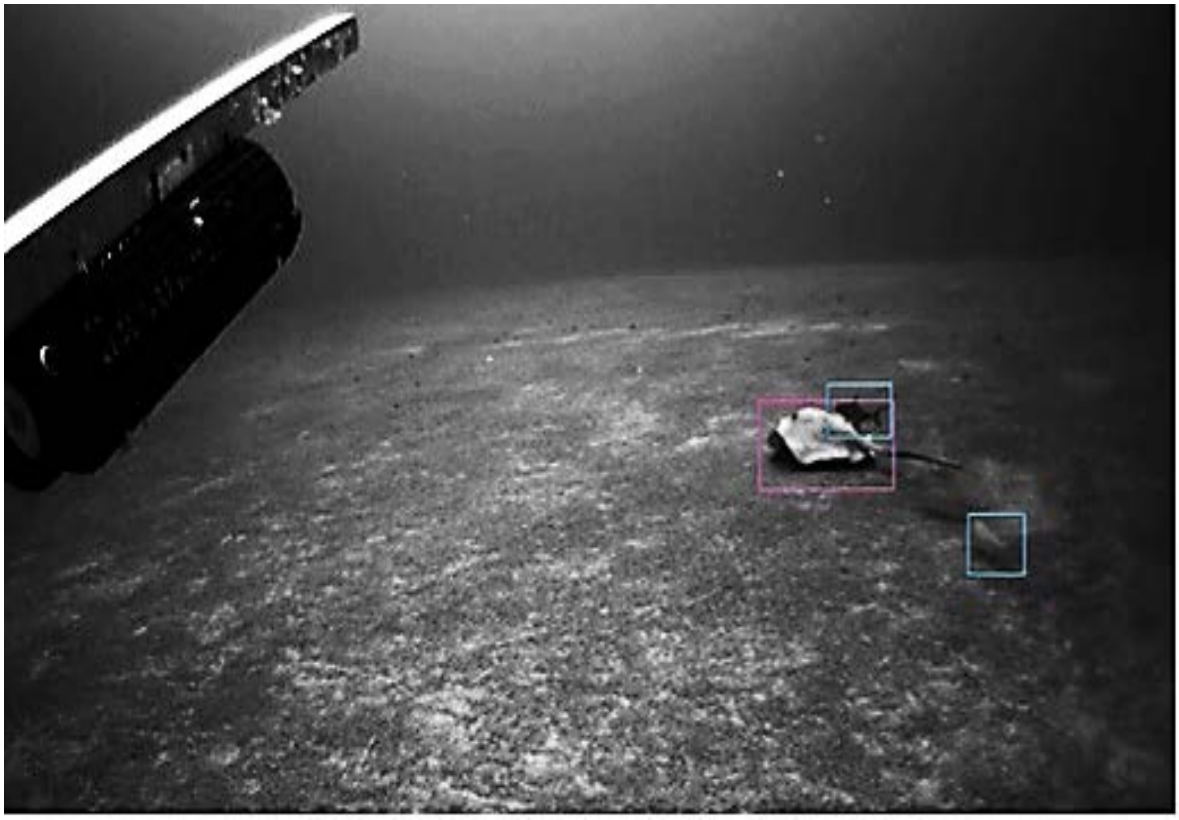


Figure 23. Example fish segmentation and tracking using deep learning for classification.

10.3 Monterey Bay Aquarium Research Institute (MBARI)

The Monterey Bay Aquarium Research Institute (MBARI) is an oceanographic research center that develops new tools and methods to study the oceans. MBARI developed an advanced knowledge-based annotation system, referred

Annotation systems are being used by imagery processors and are critical to rapidly classify and query complex objects observed on deep-sea video.

to as Video Annotation and Reference System[28] (VARS). It uses effective taxonomic classification and records the environmental information of marine organisms observed during deep-sea video operations. The VARS annotation system is interfaced to its database to enable complex querying to retrieve MBARI's deep-sea video observations. Users, primarily the video lab staff, access the visual, descriptive and quantitative data associated with the video archive (Figure 24).

Figure 24. A flowchart of the Monterey Bay Aquarium Research Institute video collection and Video Annotation and Reference System. *Credit: Monterey Bay Aquarium Research Institute.*

Roughly 28,000 hours of video have been archived and manually annotated in the past three decades. Researchers take frame grabs and enter observations and annotations of

[28] MBARI's Video Annotation and Reference System (VARS): *MBARI's Video Annotation and Reference System*

ROV footage via the VARS platform. Video are annotated in more detail once the ship comes back to shore, and all annotations are approved by the knowledge administrator. Ancillary data such as coordinates, temperature, oxygen, depth, and salinity are merged with observations and all combined information goes into the VARS database. In 2017, MBARI switched to all-digital recording with almost 1000 hours of HD digital file recordings (roughly 170 terabytes) from its primary video platforms (ROVs). There are plans to convert some of the videotape archive to digital.

The video lab staff accesses the video for annotation purposes and also to fulfill requests from researchers and media. They access all the video at least once for initial annotation review, and then an indeterminate number of times for the research and media requests. The entire library of digital video is technically accessible to all MBARI staff.

To streamline video processing, MBARI developed the Automated Video Event Detection[29] (AVED) which utilizes neuromorphic vision algorithms to detect and track objects in the video. Efforts are underway to integrate its AVED and VARS systems.

In the future, MBARI expects to utilize more platforms with higher resolution sensors and higher sampling rates (pixel dimensions, color depth, and frame rate in the case of video). These next generation technologies will produce vastly larger imagery datasets. Therefore, MBARI plans to utilize ML to address the increased imagery data, and they are in the process of implementing a new storage and archive system (IBM Spectrum Scale and Spectrum Archive) that will provide enhanced accessibility to its big data imagery.

10.4 Ocean Networks Canada

There are many international efforts taking place to address imagery accessibility. Ocean Networks Canada (ONC) is presented as a case study having an online data management system for scientific information collected from the west and east coasts of Canada and the Arctic Ocean. The ONC mission is to improve global accessibility of near real-time and historical data from Canada's oceans including the Arctic region to help communities, governments and industry make informed decisions on ocean-related policies. ONC enables observatories across Canada to have the technical and scientific capabilities that allow researchers to operate remotely and receive data at their home laboratories in

Increased accessibility is necessary to promote informed decisions on ocean policies.

[29] MBARI's Automated Video Event Detection (AVED): *MBARI's Automated Video Event Detection Webpage*

real time.

The University of Victoria hosts the ONC data management system which presently includes storage of about 170 terabytes of imagery data. It provides web quality video streamed (~5 Mbps) via satellite from the vessel to the shore, which is archived in five minute segments. HD imagery are made available to SeaTube (Section 9.2) and the Oceans 2.0 data portal[30] (Figure 25).

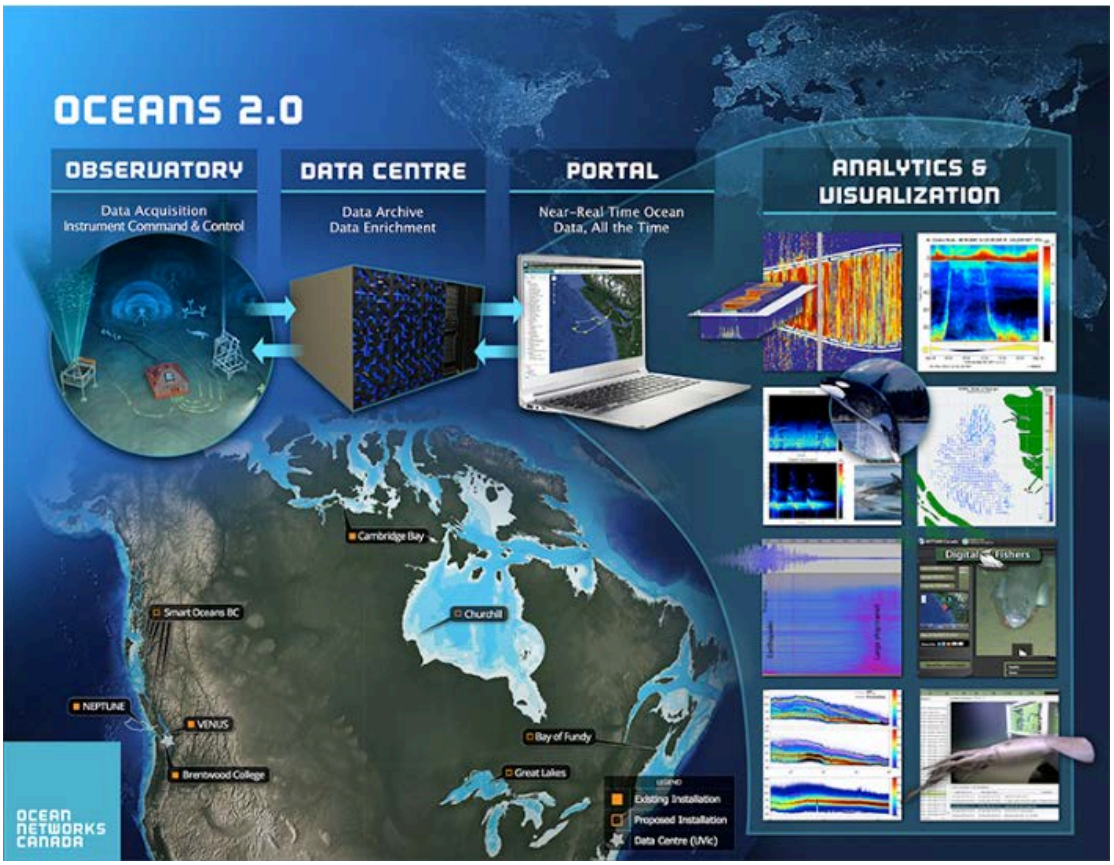


Figure 25. A flowchart of Ocean Networks Canada video collection to the Oceans 2.0 data portal. *Credit: Ocean Networks Canada.*

This data portal supports video acquisitions, annotation, and archiving with ship- and shore-based loggers. Videos archived by ONC are accessed by about 50-200 users per week, and ML is used to process imagery from stationary cameras through collaboration with the University's Computer Science Department. In the short-term, ONC is focusing on validating the quality of imagery annotations and improving scientific products from these annotated data sets; however, long-term goals involve processing of ROV footage using ML.

[30] Ocean Networks Canada Oceans 2.0 Data Portal*:*
Ocean Networks Canada Oceans 2.0 Data Portal

11. Conclusion

In summary, efficient utilization of readily available analytical AI tools, such as ML, will be dependent on the accessibility of big data imagery. This report emphasizes that the collection of imagery has grown and will continue to increase rapidly from monitoring programs in the marine environment. ML analytical software has recently become more readily available for the marine science community; however the most immediate challenge is to improve the accessibility of big data imagery to train the analytical tools that will provide more precise and timely scientific products. This will not only provide significant cost savings in processing, but also added public value to the scientific archives and products to the broader scientific community for research and discovery. The future of big data imagery processing will rely on the coupling of both humans and machines working together with increased data accessibility to enhance the performance of ML analytics for specific applications at various scales. This effort will include coding ML for specific applications and statistical validation of results to confirm the quality and reliability of the scientific products used in ocean policy decisions.

This report highlights the efforts within NOAA and its extramural partners to address the data enterprise requirements and analytical tools for big data imagery collected from the marine environment. Big data management should consider the analytical questions relevant to the end-to-end process for deriving scientific products with research analytics during search and discovery. While this report provides a foundation for important considerations with regard to improving the accessibility to analytics, this section outlines key priorities highlighted by selected experts involved with the collection, processing and analysis of imagery data that should be addressed in the near term as imagery collections exponentially increase.

As NOAA continues to enhance its national data management enterprise, the following considerations were highlighted.

- The recent exponential growth of big data imagery collected from the marine environment requires adaption from the current data enterprise to utilize analytical tools such as ML algorithms to expedite processing and analysis.

- The accessibility of big data imagery must be improved to utilize the analytical tools that deploy advances in ML, and this requires:

    - Improving data queries and accessibility through consistent, standardized, and enriched metadata to represent imagery and facilitate research and discovery for a range of analytical questions. Case studies in this report suggest metadata and annotation of underwater image data collections vary between regional programs.

- Improving interconnectivity between data storage and archival systems to optimize accessibility and workflow using analytical tools. The integration of on-premise storage and cloud services should be evaluated as a potential hybrid storage solution.

- Accessibility of big data imagery to scientists and the public must be improved to utilize the analytical tools that deploy advances in ML and computer vision.

- Statistical validation of ML computing is critical to ensure continued output of quality enterprise data and data products, in addition to increased accuracy beyond human processing capabilities.

- The NOAA Big Data Enterprise should promote partnerships among key NOAA scientists and extramural experts with diverse hands-on experience in the collection, processing and analysis of big data imagery by:

  - Increasing communication among line offices to build consensus and share expertise on the requirements and applications of big data imagery for NOAA's scientific products to support the NOAA mission.

  - Identifying analytics and accessibility requirements for enhancing NOAA's data enterprise to support a range of applications.

  - Promoting and strengthening academic and industry partnerships to provide insight on improving the data enterprise utilization of analytics and advance ML for research and discovery; therefore, reducing imagery processing costs and increasing efficiencies.

Overall, the NOAA Big Data Enterprise is transitioning into a new era where AI tools, such as ML will be used routinely by NOAA scientists and the broader, international scientific community, as well as the public. Accessibility of big data imagery and other relevant data is one of the largest challenges, and remains a bottleneck with applying ML tools to NOAA's big data imagery from the marine environment. This report serves as an overview of key considerations for NOAA's data modernization initiative relevant to enhancing data accessibility for ML analytical tools, and this initiative has recently been elevated as a NOAA cross-functional mission priority, pursuant to the White House Executive Order on Artificial Intelligence. Prioritizing modernization of the agency's data enterprise with cloud platforms and user-friendly data accessibility for ML analytics will promote scientific exchange and collaborations across various sectors of the broader scientific community, thereby providing added value to NOAA's missions as a science based agency.

12. References

Aryotejo, G., Kristiyanto, D. Y and Mufadhol. May. Hybrid cloud: bridging of private and public cloud computing. In *Journal of Physics: Conference Series,* 1025(1): 012091). IOP Publishing.

Avram, M.G. 2014. Advantages and challenges of adopting cloud computing from an enterprise. *Perspective Procedia Technology, 12*: 529-534.

Beijbom, O., Edmunds, P. J., Roelfsema C., Smith, J., Kline, D. I., Neal, B. P., Matthew J., Dunlap, M. J., Moriarty, V., Fan, T., Tan, C., Chan, S., Treibitz, T., Gamst, A. and Mitchell, B. G., and D. Kriegman. 2015. Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation." *PloS one,* 10(7): e0130312.

Benfield, M. C., Grosjean, P., Culverhouse, P. F., Irigoien, X., Sieracki, M. E., Lopez-Urrutia, A., Dam, H. G., Qiao, H., Davis, C. S., Hansen, A. Pilskaln, C. H., Riseman, E. M., Schultz, H., Utgoff, P. E. and Gorsky, G. 2007. RAPID: research on automated plankton identification. *Oceanography*, *20*(2): 172-187.

Blokdyk, G. 2018. Enterprise metadata management standard requirements. Emereo Pty Limited. 118 pp. ISBN 0655167498.

Chen, S., Pande, A., and Mohapatra, P. 2014. Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services* (pp. 109-122). ACM.

Davis, D. S. 2010. NOAA Environmental Data Management Framework.

Diesing, M., Green, S. L., Stephens, D., Lark, R. M., Stewart, H. A. and Dove, D. 2014. Mapping seabed sediments: comparison of manual, geostatistical, object-based image analysis and machine learning approaches. *Continental Shelf Research*, *84*:107-119.

Federal acquisition regulation. 2010. *Federal Employees' Compensation Act Federal Employees' Compensation Act*.

Fernandes, J. A., Irigoien, X., Goikoetxea, N., Lozano, J. A., Inza, I., Pérez, A., and Bode, A. 2010. Fish recruitment prediction, using robust supervised classification methods. *Ecological Modelling*, *221*(2): 338-352.

Fortunato, S., Bergstrom, C. T., Boerner, K., Evans, J.A., Helbing, D., Milojevie, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D. and A. Barabasi. 2018. Science of science. *Science:* 359, eaao0185. DOI: 10.1126/science.aao0185.

Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G. and I. Stoica. 2009. Above the clouds: A Berkeley view of cloud computing, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Rep. UCB/EECS, 28, 2009.

Froeschl, K.A. 1997. Metadata management in statistical information processing. Springer-Verlag Wien, New York. 538pp. ISBN 978-3-211-82987-5. DOI 10.1007/978-3-7091-6856-1.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*: 2672-2680.

Goyal, S. 2014. Public vs private vs hybrid vs community - Cloud computing: A critical review. *International Journal of Computer Network and Information Security*, *6*(3): 20.

Hay, D.C. 2010. Data model patterns: A metadata Map. The Morgan Kaufmann Series in Data Management systems, Elsevier Inc., 432 pp. ISBN 0080477038.

He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 770-778.

Holdren, J. P. 2013. Memorandum for the heads of executive departments and agencies: Increasing access to the results of federally funded scientific research.

Howland, J., Gallager, S., Singh, H., Girard, A., Abrams, L., Griner, C., Richard, T. and Vine, N. 2006. *Development of a towed survey system for deployment by the fishing industry*. Woods Hole Oceanographic Institution, MA.

Hurwitz, J., Bloor, R., Kaufman, M. and F. Halper. 2010. Cloud Computing for Dummies. Wiley Publishing, Inc., Indianapolis, Indiana.

Islam, S., Husain, A. and Zaki, H.M. 2017. Pooling of computing resources in private Cloud deployment. In *Journal of Physics: Conference Series,* 1025(1): 012091. IOP Publishing.

International Standard Organization. 2003. Organización Internacional de Normalización. *ISO 14721: 2003: Space data and information transfer systems: open archival information system: reference model*.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE conference on computer vision and pattern recognition*: 3128-3137.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*: 1097-1105.

Langley, P. 2011. The changing science of machine learning. *Machine Learning*, *82*(3): 275-279.

Lubchenco, J. 2012. NOAA Annual Guidance Memorandum. *NOAA Tech Memorandum*. http://www.performance.noaa.gov/agm/

Lüdtke, A., Jerosch, K., Herzog, O. and Schlüter, M. 2012. Development of a machine learning technique for automatic analysis of seafloor image data: Case example, Pogonophora coverage at mud volcanoes. *Computers & Geosciences*, *39*: 120-128.

Madden, C. J., Goodin, K., Allee, R. J., Cicchetti, G., Moses, C., Finkbeiner, M. and Bamford, D. 2009. Coastal and marine ecological classification standard. *National Oceanic and Atmospheric Administration and NatureServe*.

Mallet, D. and Pelletier, D. 2014. Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). *Fisheries Research*, *154*: 44-62.

Mell, P. and T. Grance. 2011. The NIST definition of cloud computing. National Institute of Standards and Technology, *Department of Commerce.*

Mirajkar, N., Barde, M., Kamble, H., Rahul, A., and S. Kumud. 2012. Implementation of private cloud using eucalyptus and an open source operating system. *International Journal of Computer Science Issues*. 9: 360-364.

Mitchell, Tom M. *Machine Learning*. McGraw Hill, 1998.

National Academies of Sciences, Engineering, and Medicine. 2016. Refining the Concept of Scientific Inference When Working with Big Data: Proceedings of a Workshop—in Brief. Washington, DC: The National Academies Press. https://doi.org/10.17226/23616. National Research Council. 2013. Frontiers in Massive Data Analysis. Washington, DC: The National Academies Press. https://doi.org/10.17226/18374.

National Research Council. 2015. Robust Methods for the Analysis of Images and Videos for Fisheries Stock Assessment: Summary of a Workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/18986.

Nian, R., He, B., Zheng, B., Van Heeswijk, M., Yu, Q., Miche, Y. and Lendasse, A. 2014. Extreme learning machine towards dynamic model hypothesis in fish ethology research. *Neurocomputing*, *128*: 273-284.

Purser, A., Bergmann, M., Lundälv, T., Ontrup, J. and Nattkemper, T. W. 2009. Use of machine-learning algorithms for the automated detection of cold-water coral habitats: a pilot study. *Marine Ecology Progress Series*, *397*: 241-251.

Ren, S., He, K., Girshick, R., and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*: 91-99.

Riley, J. 2016. Understanding Metadata. A Primer Publication of the National Information Standards Organization, Baltimore, MD. ISBN: 978-1-937522-72-8.

Rodríguez J.G., Fernandes J.A., Garmendia J.M., Muxica I., and Borja A. 2012. Development of a Bayesian Networks based method for assessing the status of hard bottom substrata biota. XVII Simposio Ibérico de Estudios de Biología Marina. 11-14 September. Donostia-San Sebastián (Spain).

Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M., Seager, J., and E. Harvey. 2016. Fish species classification in unconstrained underwater environments based on deep learning. *Limnology and Oceanography: Methods*, *14*(9): 570-585.

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3): 210-229.

Shafait, F., Mian, A., Shortis, M., Ghanem, B., Culverhouse, P. F., Edgington, D., Cline, D., Ravanbakhsh, M., Seager, J. and E. S. Harvey. 2016. Fish identification from videos captured in uncontrolled underwater environments. *ICES Journal of Marine Science*, *73*(10): 2737-2746.

Shafait, F., Harvey, E. S., Shortis, M. R., Mian, A., Ravanbakhsh, M., Seager, J. W., Culverhouse, P. F., Cline, D. E. and D. R. Edgington. 2017. Towards automating underwater measurement of fish length: a comparison of semi-automatic and manual stereo–video measurements. *ICES Journal of Marine Science*, *74*(6): 1690-1701.

Shortis, M. R., Ravanbakskh, M., Shaifat, F., Harvey, E. S., Mian, A., Seager, J. W., Culverhouse, P. F., Cline, D. E. and D. R. Edgington. 2013. A review of techniques for the identification and measurement of fish in underwater stereo-video image sequences. *Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection,* 8791: 87910G. International Society for Optics and Photonics.

Shumchenia, E. J., and King, J. W. 2010. Comparison of methods for integrating biological and physical data for marine habitat mapping and classification. *Continental Shelf Research*, *30*(16): 1717-1729.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Strasser, C. 2015. Research Data Management. A Primer Publication of the National Information Standards Organization, Baltimore, MD. ISBN: 978-1-937522-65-0.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 1-9.

Thomas, M.K., Fontana, S., Reyes, M., Kehoe, M., and Pomati, F. 2018. The predictability of a lake phytoplankton community, over time-scales of hours to years. *Ecology Letters, 21*: 619-628.

United States. (2018). *The President's management agenda*.

Uusitalo, L., Fernandes, J. A., Bachiller, E.,Tasala, S., and Lehtiniemi, M. 2016. Semi-automated classification method addressing marine strategy framework directive (MSFD) zooplankton indicators. *Ecological Indicators*, *71*: 398-405.

Vaduva, A., and T. Vetterli. 2001. Metadata management for data warehousing: An overview. International Journal of Cooperative Information Systems 10.03.

Venkat, T., Rao, N., Naveena, K., David, R. and Narayana, M. 2015. A new computing environment using hybrid cloud. *Journal of Information Sciences and Computing Technologies*, *3*(1): 180-185.

Villon, S., Mouillot, D., Chaumont, M., Darling E. S., Subsol, G., Claverie, T. and S. Villeger. 2018. A deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecological Informatics*: doi: 10.1016/j.ecoinf.2018.09.007

Vnuk, L. 2014. A success factor-based framework for enterprise metadata management. University of South Australia, School of Information Technology and Mathematical Sciences. 400 pp.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R. and Vieglais, D. 2012. Darwin Core: an evolving community-developed biodiversity data standard. *PloS one*, *7*(1): e29715.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. *European conference on computer vision*: 818-833. Springer, Cham.

Appendix A. NOAA Fisheries Imagery Data from Independent Surveys

Current and planned NOAA Fisheries' optical data sets, broken down by regional Science Center, data description and current and planned space requirements.

Science Center	Dataset Description	Current drive space required (TB)	Yearly drive space required (TB)
PIFSC			
	BotCam Stereo-Video of MHI Bottomfish	15.0	2.0
	SeaBed AUV Stereo-Video of MHI Bottomfish	0.7	1.0
	SeaBed AUV Benthic Imagery	0.5	-
	MOUSS Stereo-Imagery of MHI Bottomfish	50.0	15.0
	CRED Benthic Photoquadrats	10.0	1.0
	BRUV Reef Fish Stereo-Video	21.7	-
	CRP Cetacean Photos	4.0	0.5
SEFSC			
	MOUSS UHSI Test Project Gulf of Mexico (Stereo Stills)	22.0	0.0
	MSLabs Reef fish survey (Stereo Stills Plus Video)	115.5	11.0
	MSLabs Full Spherical Camera (Future SEAMAP)	1.8	13.5
	SEFIS Reef Fish Survey (High Resolution Video)	300.0	40.0
	PCLab Reef Fish Survey (Stereo Stills Plus Video)	9.0	3.0
	PCLab Reef Fish Survey (Video)	14.0	4.2
	Acropora Palmata Demographic Monitoring	0.3	0.3
	Sea Turtle Photo ID	0.0	0.1
	Sea Turtle Bone Section Images	0.4	0.1
NEFSC			
	HabCam (Stereo stills)	200.0	30.0
	APH-22 (Aerial Stills)	0.4	0.4

SWFSC			
	ROV Still Images	0.2	0.2
	ROV Video	5.0	2.0
	Other HD Video	3.0	2.0
	MOUSS UHSI Test Project California (Stereo Stills)	1.9	4.0
NWFSC			
	AUV	8.0	2.0
	Camera on Trawl Images	0.1	0.2
	ROV Images (ROPOS)	0.2	0.0
	ROV Images	2.0	0.0
	Hook and Line Video	0.5	0.5
	Groundfish Video	0.2	0.2
AFSC			
	Camtrawl Stereo-Images of Fish in Trawl	3.0	0.5
	Drop Stereo Camera	7.2	1.5
	ROV Images: Rockfish and Habitat Video and Photos	1.0	0.1
	Shark Morphology Photos	0.5	0.1
	Benthic Habitat Dive and ROV Video and Photos	1.0	0.1
	Coral/Sponge/Benthic Habitat ROV, Sub, Dive Video and Photos	3.0	0.5
	TACOS Towed Video and Mosaics	0.4	-
	SEABOSS Video and Stills	0.0	-
	LRSSS transmissometer (Water Quality)	0.0	-
	Age and Growth Groundfish Aging Structure Images	0.8	0.1
	GOAIERP Near shore Habitat Work	0.5	0.0
	Camera Chute Survey Trawl	0.5	-
	Stereo Rail Camera Alaska	10.0	-
	Total:	**814.2**	**175.8**

U.S. Secretary of Commerce
Secretary of Commerce
Wilbur L. Ross, Jr.

Acting Under Secretary of Commerce
for Oceans and Atmosphere
and NOAA Administrator
Neil A. Jacobs, PhD

National Marine Fisheries Service
Assistant Administrator for Fisheries
Christopher W. Oliver

May 2019

www.fisheries.noaa.gov

OFFICIAL BUSINESS

National Marine Fisheries Service
1315 East West Highway
SSMC3, F/ST
Silver Spring, MD 20910