Future Science Brief

Nº 6 April 2020

Big Data in Marine Science

European MARINE BOARD
Advancing Seas & Ocean Science

www.marineboard.eu

European Marine Board IVZW Future Science Brief 6

The European Marine Board is an independent and self-sustaining science policy interface organisation that currently represents 34 Member Organizations from 18 European countries. It was established in 1995 to facilitate enhanced cooperation between European marine science organizations towards the development of a common vision on the strategic research priorities for marine science in Europe. The EMB promotes and supports knowledge transfer for improved leadership in European marine research. Its membership includes major national marine or oceanographic institutes, research funding agencies and national consortia of universities with a strong marine research focus. Adopting a strategic role, the European Marine Board serves its member organizations by providing a forum within which marine research policy advice is developed and conveyed to national agencies and to the European Commission, with the objective of promoting the need for, and quality of, European marine research.

www.marineboard.eu

European Marine Board Member Organizations

European Marine Board IVZW Future Science Brief 6

This Future Science Brief is the result of the work of the European Marine Board Expert Working Group on Big Data in Marine Science. See Annex 1 for the list and affiliations of the Working Group members.

Working Group Chairs
Lionel Guidi, Antonio Fernandez Guerra

Contributing Authors
Dorothee C. E. Bakker, Carlos Canchaya, Edward Curry, Federica Foglini, Jean-Olivier Irisson, Ketil Malde, C. Tara Marshall, Matthias Obst, Rita P. Ribeiro, Jerry Tjiputra

Series Editor
Sheila J. J. Heymans

Publication Editors
Britt Alexander, Ángel Muñiz Piniella, Paula Kellett, Joke Coopman

External Reviewers
Björn Backeberg, Ghada El Serafy, Frank Muller-Karger, David Schoeman

Additional Contribution
Dick M. A. Schaap

Internal review process
The content of this document has been subject to internal review, editorial support and approval by the European Marine Board Member Organizations.

Suggested reference
Guidi, L., Fernandez Guerra, A., Canchaya, C., Curry, E., Foglini, F., Irisson, J.-O., Malde, K., Marshall, C. T., Obst, M., Ribeiro, R. P., Tjiputra, J., Bakker, D. C. E. (2020) Big Data in Marine Science. Alexander, B., Heymans, J. J., Muñiz Piniella, A., Kellett, P., Coopman, J. [Eds.] Future Science Brief 6 of the European Marine Board, Ostend, Belgium. ISSN: 2593-5232. ISBN: 9789492043931. DOI: 10.5281/zenodo.3755793

www.marineboard.eu
info@marineboard.eu

Design and cover image
Zoeck

First edition, April 2020

Foreword

Whether you are familiar with the term or not, big data are part of everyday life for the majority of citizens. We routinely use navigation apps on our smartphones, are inundated with targeted advertisements on social media, and recently, tracking using big data is a strategy in some countries to manage the spread of the novel coronavirus. However, various projects, business professionals and vendors often use the term 'big data' quite differently making the definition difficult to establish. The concept of big data gained momentum in the early 2000s and is now becoming critical to provide policy-makers with the tools they need to make well informed, evidence based decisions in real-time. It is important now more than ever to effectively understand and manage the multitude of threats the ocean faces in a timely manner, given that the window of opportunity to take effective action is diminishing. The UN Decade of Ocean Science for Sustainable Development (2021-2030) provides an opportunity to improve the uptake of ocean science in sustainable development. Preserving the ocean health requires open access to data to ensure efficiency of activities and, importantly, transparency. Big data will play an essential role during the Ocean Decade in participating in a digital revolution in ocean data, providing new insight into the complexities of the ocean, and increasing our knowledge and tools for the sustainable management of human impacts on marine resources.

Given the rapid advancements and digitalization of ocean technologies and data collection, in 2017 the European Marine Board (EMB) identified big data as an area of interest. In May 2019 the EMB working group on 'Big Data in Marine Science' kicked-off with a meeting in Ostend. This Future Science Brief is the primary output of the working group. It aims to raise awareness of big data, give some examples of its potential applications in marine science, and identifies actionable recommendations needed to fully bring marine science into the world of big data.

On behalf of the EMB membership, I would like to thank the members of the EMB working group on Big Data in Marine Science (Annex I) for their hard work and expertize in contributing to this Future Science Brief. I would also like to thank the external reviewers (Annex II) for their constructive comments, and Dick Schaap for providing additional comments and input. My thanks go to the EMB Secretariat for their work in coordinating the working group and the synthesis and publication of this document, namely Britt Alexander, Sheila Heymans, Ángel Muñiz Piniella, Paula Kellett and Joke Coopman. Last but not least, I would like to thank the students from the Arteveldehogeschool in Ghent, Belgium - Jonas Willems, Maud Chiau, Cristian Moraru and Paul Dekeyser - for their work to design and produce the infographics that are included in this document.

Gilles Lericolais
Chair, European Marine Board
April 2020

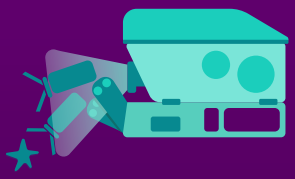
Table of Contents

Executive summary

This document explores the potential of big data, i.e. large volumes of high variety data collected at high velocity, to advance marine science. Marine science is rapidly entering the digital age, as introduced in **Chapter 1**. Expansions in the scope and scale of ocean observations, as well as automated sampling and 'smart sensors', are leading to a continuous flood of data. This provides opportunities to transform the way we study and understand the ocean through more complex and interdisciplinary analyses, and offers novel approaches for the management of marine resources. However, more data do not necessarily mean that we have the right data to answer many critical scientific questions and to make well-informed, data-driven management decisions on the sustainable use of ocean resources. To increase the value of the wealth of marine big data, it must be openly shared, interoperable and integrated into complex transdisciplinary analyses, which can be based on artificial intelligence.

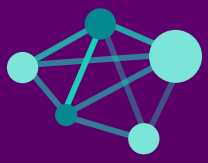
The marine science community has not yet reached the big data revolution and the 'data deluge' introduces a unique set of challenges that are new to many marine scientists. This document identifies bottlenecks and opportunities related to data acquisition, data handling and management, computing infrastructures and interoperability, data sharing, big data analytics, data validation, and training and collaboration. Specific challenges should be overcome to ensure the maximum value of marine big data can be reaped. We present topics and case studies of some recent advances in the application of big data to support marine science that demonstrate these challenges and recommendations. **Chapter 2** covers climate science and marine biogeochemistry, with particular focus on European and global initiatives to integrate carbon and other biogeochemical data that are used to inform global climate negotiations. **Chapter 3** discusses how big data could be used to create high-resolution, multidisciplinary habitat maps for planning new marine protected areas. **Chapter 4** looks at marine biological observations including genetic sequences, imagery and hydro-acoustic data and calls for a globally connected network of long-term biological observations for more complex interdisciplinary analyses using big data. **Chapter 5** addresses food provision from the ocean and seas with a focus on aquaculture and the management of sea-lice outbreaks and escaped, farmed salmon using artificial intelligence.

RECOMMENDATIONS

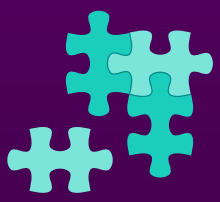


DATA ACQUISITION
Continued development of smart sensors for automated sampling and data processing as well as more efficient data transfer, so more ocean data can be collected by machines rather than humans.

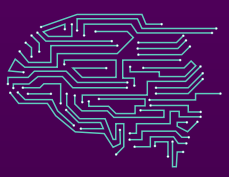
DATA HANDLING AND MANAGEMENT
Develop community standards and well-designed data management plans ensuring Findable, Accessible, Interoperable and Reusable (FAIR) data.



COMPUTING INFRASTRUCTURE AND INTEROPERABILITY
Marine data services need to be interoperable and incorporate cloud-computing, cloud-storage, and analytical tools.



DATA SHARING
Data need to be open and data sharing should be incentivized between scientists, industry and governments.



BIG DATA ANALYTICS AND DATA VALIDATION
Develop standardized algorithms and community maintained data sets that can be used for model training and calibration.

TRAINING AND COLLABORATION
Develop specialized training for marine scientists to adopt the use of artificial intelligence. Collaborations are needed between marine scientists and computer scientists.

To pave the way towards making marine science a big data-driven discipline, in **Chapter 6** we recommend to:

- **Enhance data acquisition** through the continued development of 'smart sensors' for automated sampling and data processing so that more marine data can be collected by machines. We also propose to increase the efficiency of data transfer to allow more real-time, or near real-time analyses and decision making;
- **Enhance data handling and management** through more widespread adoption of community data standards and well-designed data management plans based on Findable, Accessible, Interoperable and Reusable (FAIR) principles so that data are machine-readable. We also recommend the increased use of existing marine data management infrastructures;
- **Increase data interoperability and accessibility** by upgrading European marine data management infrastructures to handle and exchange more multidisciplinary and real-time data. These infrastructures should include more integrated cloud computing, data storage and big data analytical tools. We recommend increased participation of the European marine science community in development of Virtual Research Environments (VREs) and European Open Science Cloud (EOSC) initiatives. We also recommend that these infrastructures should be sustained in the long term and that there should be more cross-disciplinary fertilization of computing technology from multimedia and digital sectors;
- **Improve data sharing** between scientists, industry and governments through new incentives and protocols such as social networks or data impact factors;
- **Increase the use of big data analytics and ensure data validation** by developing close collaborations between data scientists and marine scientists, developing standardized models, and well-curated community data sets to train algorithms;
- **Develop specialized training** on artificial intelligence by establishing new regional and global marine science networks and consolidating already existing networks. We recommend training data curators to maintain the quality of data feeding into artificial intelligence algorithms; and
- **Increase collaborations between marine scientists, computer scientists, data scientists and data managers** in the form of working groups and the involvement of data scientists in the design of marine research.

Credit: Oliver Beardon, SailBritain

1 Introduction

> "We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely."
>
> (Wilson, 1998)

We have more ocean data than ever before, yet we do not have enough data to answer many critical scientific questions. Big data (see Box 1) offers the potential to revolutionize marine science, allowing the use of data in novel and innovative ways to enhance our understanding of the ocean and the impact of human activities.

Box 1: What are big data?

The term 'big data' has been used to label data with different attributes and several definitions of big data have been proposed over the last decade (Curry, 2016). In this document, we use the most widespread definition, which describes big data as a three-dimensional approach: ***"Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization*** (Laney 2001)[1].

Big data are commonly described using the three 'Vs', which each introduce unique challenges for data processing:

Volume large quantities of data;

Velocity high-frequency of incoming real-time data (e.g. the Internet of Things); and

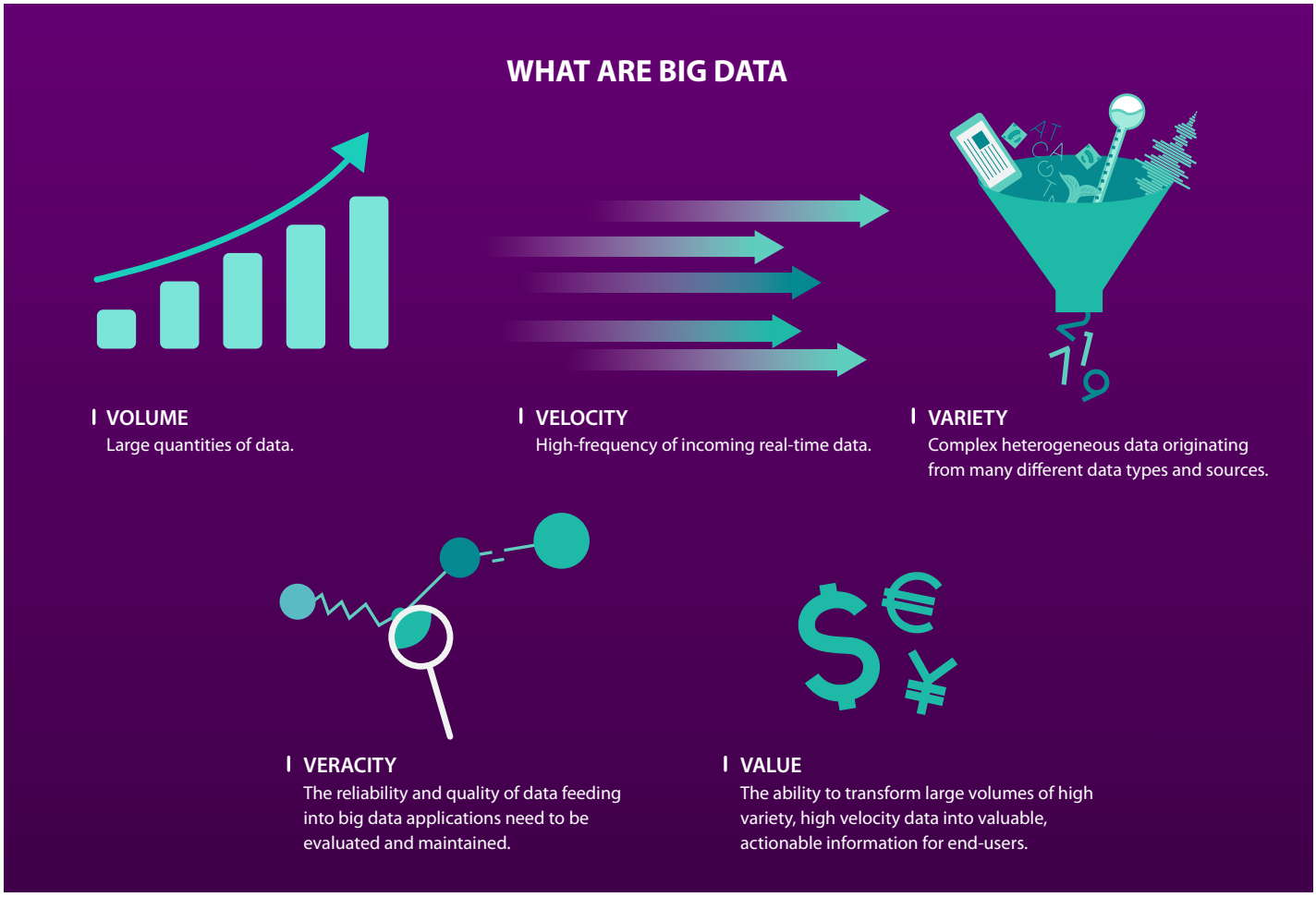
Variety complex heterogeneous data originating from a wide range of data types and sources with different syntactic formats and schemas, usually requiring standardization before its use.

In this document we adopt two additional dimensions to describe big data:

Veracity trustworthiness of data and potential uncertainty due to ambiguity, inconsistency (e.g. sensor reading failures), incompleteness, data drift (i.e. change in input data), statistical inconsistency, and approximations (e.g. from data compression algorithms); and

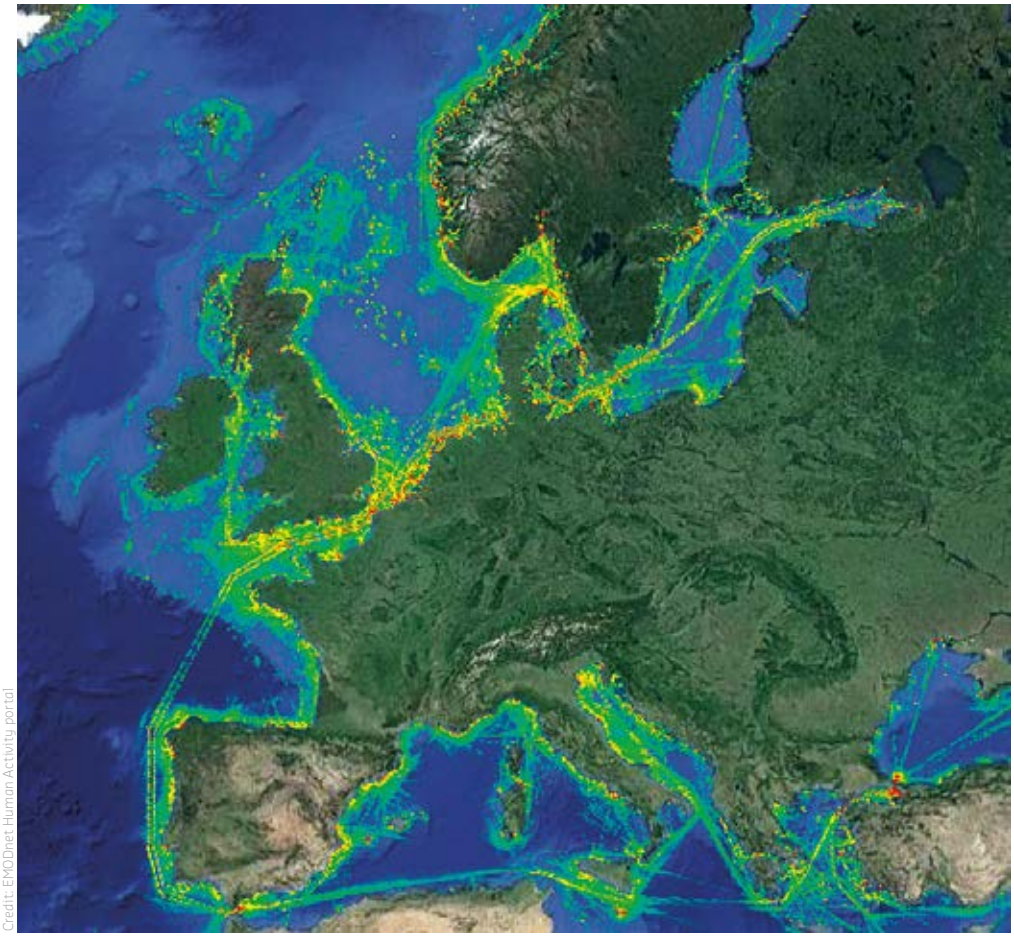
Value understanding the costs and benefits of collecting and analyzing data to ensure that its value can be reaped.

[1] https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf



In recent years, the observatory-based approach to marine science has increased in scale from regional to global. It has expanded in scope from traditional physical and biogeochemical observations to a more interdisciplinary approach that includes more levels of complexity from genes, individuals, populations, communities, ecosystems to the biosphere. This leads to high volume and high velocity data, with high variability. Several international observational initiatives like Argo[2], Global Ocean Observing System[3] (GOOS) or the Ocean Observatories Initiative (OOI[4]) are using a myriad of infrastructures including satellites, seafloor electro-optic cables, drifting buoys, floats, moorings, and gliders, alongside classical ship-based observations. These infrastructures have sensors installed that measure physical, chemical, biological, geological and geophysical parameters. This is leading to a deluge of heterogeneous data originating from monitoring and observations of the marine environment, from a spectrum of multidisciplinary projects and programs. Effectively using this data for management will require a big data approach. Next-generation ocean observing technologies, e.g. autonomous underwater vehicles (AUVs), will offer completely new avenues for studying marine ecosystems due to their high levels of automation, autonomy and precision, enabling them to sample larger areas of the ocean. These will have the capability to measure biological parameters, including -omics to measure DNA, RNA and proteins, and to acquire and preserve environmental DNA (eDNA), e.g. DNA collected directly from seawater, which will further increase the pace of data acquisition (Benedetti-Cecchi *et al.,* 2018). Continual advances in these technologies will make spatio-temporal data richer and more ubiquitous, opening up opportunities for performing larger and more complex analyses that combine disparate scientific data across disciplines. One example is combining acoustic survey data, trawl survey data, fish catch data, and automatic identification system (AIS) data for vessel tracking (see Figure 1.1) to understand the impact of shipping noise on commercial fish landings. Artificial intelligence, machine learning, and data mining (see Box 2) are needed to identify patterns in complex data for a deeper understanding of processes that is not possible with traditional methods. This can provide critical insights to help manage human impacts on the marine environment and its living resources, as well as increasing automation and efficiency for researchers, industry and data users at the science-policy interface. Artificial intelligence algorithms are beginning to be installed on-board Earth observation satellites to speed up data processing and transmission using 'Edge AI', meaning that algorithms are run locally on hardware devices where the data are collected. These advances, coupled with developments in ocean observation technologies, offer enormous potential to collect larger volumes of heterogeneous marine data in real-time, or near real-time.

[2] http://www.argo.net/
[3] https://www.goosocean.org/
[4] https://oceanobservatories.org/



Credit: EMODnet Human Activity portal

Figure 1.1 Vessel density map created using automatic identification system (AIS) data.

The United Nations has proclaimed a Decade of Ocean Science for Sustainable Development[5,6], (2021 – 2030) to ensure that ocean science guides and supports the 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals (SDGs). A global ocean data portal is a key priority for the Ocean Decade, to enable improved transfer of data and data products to end-users, thereby addressing some of the most critical societal challenges (Ryabinin *et al.,* 2019). In addition, in Navigating the Future V the European Marine Board proposed a digital ocean twin where all historical and current data about the ocean could be uploaded, accessed, and updated in real-time and used in decision-making (European Marine Board, 2019). This digital ocean twin will require simulation modelling to address gaps in data, and hence flag associated uncertainties. It would integrate next-generation ocean observing technologies into a network of *Ocean Internet of Things*, within which data are made available and processed in real-time using artificial intelligence and cloud computing.

The digital ocean twin could then be used by managers to make informed data-driven decisions regarding, *inter alia*, a sustainable ocean economy, targets for greenhouse gas emissions, fisheries and marine resource management options, and marine spatial planning. In the USA, the National Oceanographic and Atmospheric Administration (NOAA) has recently called for increased investments in artificial intelligence, unmanned observing systems, -omics, and cloud computing in order to make transformative advancements in their science and end-user products[7].

Big data are transforming our world and have compelled a paradigm-shift in how we face major societal challenges from climate change to public health. Currently, decisions and solutions are increasingly based on data-driven models with big data driving digital transformation and automation across all sectors in what has been termed the fourth industrial revolution. Real-time data processing allows rapid reactions to events, which is critical for many situations. From financial fraud to public security and environmental policy, big data will contribute to establishing a framework that enables a safe and secure digital economy (Zillner *et al.,* 2017). Big data is transforming scientific research in fields such as astrophysics (e.g. the Large Hadron Collider), biomedicine, bioinformatics, climate science, and material sciences. Some fields such as astrophysics and biomedicine have more advanced unified frameworks for integrating heterogeneous data, cloud computing, and analysis tools and marine science can learn from these. For example open-source software frameworks, such as Rucio[8] (Barisits *et al.,* 2019), are currently used in high-energy physics and to support Large Hadron Collider experiments to manage widely distributed heterogeneous data from different data centres.

[5] https://en.unesco.org/ocean-decade

[6] https://oceandecade.org/

[7] https://www.noaa.gov/media-release/noaa-releases-new-strategies-to-apply-emerging-science-and-technology

[8] https://rucio.cern.ch/

Box 2: Artificial intelligence, machine learning, deep learning and data mining

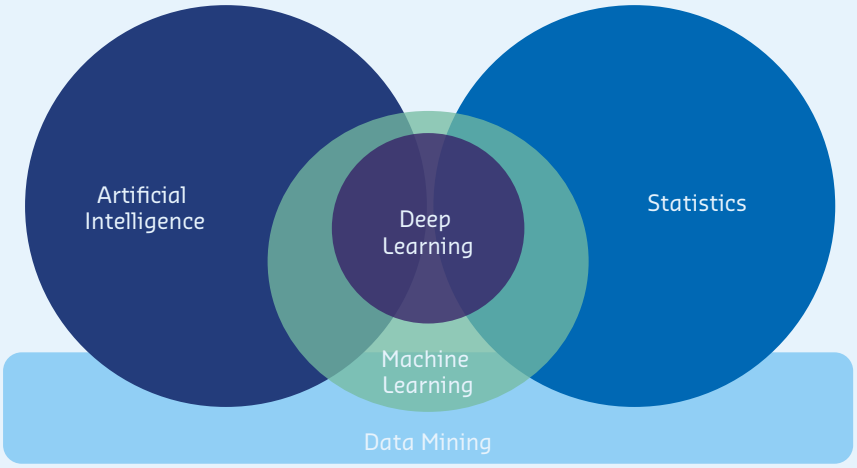
In order to extract information and knowledge from data, analysis is required, often through statistical methods. When dealing with big data, classic inferential statistics are not ideal because large volumes of data tend to render statistically significant false positive outcomes, and high uncertainty means more insight on causality is needed. As a result, such tests have limitations and drawbacks in the decision-making process. Dedicated methods that can cope with, and process, large and heterogeneous data sets and extract valuable information are therefore required. These include:

Artificial intelligence the theory and development of computer systems that are able to perform tasks or exhibit behaviour normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. Artificial intelligence is an umbrella term that includes machine learning, machine reasoning and robotics;

Machine learning considered a subfield of artificial intelligence and statistics. It refers to algorithms that automatically learn to recognize complex patterns in new data sets, improve their performance from experience and produce models that have predictive power. Machine learning models are usually divided into supervised and unsupervised models, where the former is given known examples to learn from (e.g. images with a label) while the latter attempts to find structure from the data directly (e.g. group images by colour);

Deep learning a subfield of supervised machine learning that refers to powerful machine learning algorithms able to learn a model with complex raw data as its input. This method has been very successful with highly non-linear data such as images. Classic image classification requires the operator to manually extract features in the form of summary characteristics from each image (e.g. dimensions, luminance, colour), which the algorithm can be trained on, while deep learning can learn directly from the pixels in the image;

Data mining the process of discovering information from data through pattern recognition. It is built on several fields including machine learning, artificial intelligence, statistics, mathematical modelling and database activities. Data mining can extract useful insights that can both summarize the available data and provide actionable information to make crucial decisions.

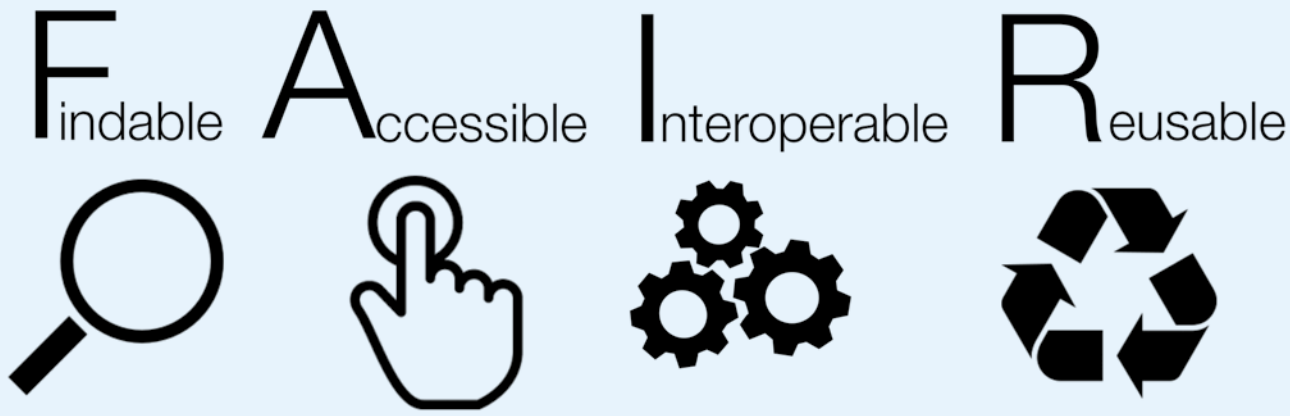


Simplified diagram showing the relationship between artificial intelligence, machine learning, deep learning, statistics and data mining.

Ocean data are collected by many different stakeholders including scientists, government organizations and private industries such as fisheries, aquaculture, and oil and gas. These include historical (e.g. time-series, publications) and real-time data. Once data has been acquired, it is important to ensure its maximum benefit can be derived and the principles of 'capture once, use many times' and the FAIR principles (Findable, Accessible, Interoperable, Reusable, see Box 3), are major targets to make ocean data sets available to user communities. Discovery of and access to aggregated data sets are important as well as analytical e-infrastructures that support complex transdisciplinary analyses. Several European initiatives for the management of ocean data have made significant progress in marine data accessibility and interoperability by developing standards and dedicated infrastructures (see Box 4).

However, there is currently no unifying framework to realize the full potential of big data in marine science. Many data are scattered, fragmented, not publicly available, and sometimes discarded after a few years. Widening big data applications in marine science depends on increasing and improving strategies to search for and link distributed data (including historical data and real-time data). Google's Data Search Tool™, [9], is promising and allows searching, but not access to, distributed data sets. Access to data and the use of best practices across the data value chain (see Box 5) will also be important for the uptake of big data in marine science. This will allow increased use of machine learning in marine science, for which some examples are given in this document. The International Council for the Exploration of the Sea (ICES) working group 'Machine Learning in Marine Science'[10] is focusing on reviewing current applications, new developments in machine learning of interest for marine science, and improved collaboration between data scientists and marine scientists to advance machine learning use and applications. These aspects are therefore not discussed in detail in this document.

Box 3: FAIR data



FAIR data principles describe standards of *Findability, Accessibility, Interoperability and Reusability*[11]. FAIR data are essential for large-scale, machine-driven, multidisciplinary analyses to realize the full scientific and societal value of data. The widespread adoption of FAIR data principles requires a shift in research culture and technology. Metrics are also being developed to define and evaluate the FAIRness of data submitted to data networks, e.g. the FAIR metrics group[12]. For more recommendations on how to achieve the FAIR principles, see the 'Final Report and Action Plan from the European Commission Expert Group on FAIR data'[13].

Image credit: Sangya Pundir, European Bioinformatics Institute, CC BY 4.0.

Working with data of increasing volume and complexity introduces a set of data management challenges that require the widespread deployment of solutions. New skills are needed to cope with the 'data deluge' since many scientists are spending a large proportion of their time on discovery and re-use of complex, heterogeneous data. This document provides advice on capacity building and initiatives to enhance the use and uptake of big data in marine science for Europe. We demonstrate some examples of current state of the art, challenges and potential future paths using societally relevant marine science topics and case studies. Data volume, velocity, variety, veracity and value (i.e. the 'V's' we use to define big data) are discussed to varying degrees in each topic and case study.

[9] https://datasetsearch.research.google.com/

[10] https://www.ices.dk/community/groups/Pages/WGMLEARN.aspx

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

[12] https://github.com/FAIRMetrics/Metrics

[13] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf

Box 4: Examples of European marine data management infrastructures

Several European initiatives funded and/or supported by the European Commission provide infrastructure for collecting and managing marine *in situ* and remotely sensed data for increased discovery, access and long-term data stewardship. These include **SeaDataNet**[14] (marine environment), **Euro-Argo**[16] (ocean physics and marine biogeochemistry), **EMODnet**[17] (bathymetry, chemistry, geology, physics, biology, seabed habitats and human activities), **ELIXIR-ENA**[18] (biogenomics), **Copernicus Marine Environmental Monitoring Service**[19] (CMEMS, ocean analysis and forecasting), **ICOS-Ocean**[20] (carbon) and **LifeWatch**[21] (biodiversity). Significant progress has been made during the past three decades to develop these infrastructures, which function in connection with national data centres and develop data standards. These infrastructures are developed and operated by research, governmental, and industry organizations from European states, and in close interaction with international initiatives on data management led by the Intergovernmental Oceanographic Commission (IOC), the World Meteorological Organization (WMO), the Food and Agricultural Organization (FAO), the Group on Earth Observations (GEO), the International Council for the Exploration of the Sea (ICES), and others. Each of these data infrastructures has established links to data originators and their data collections, facilitating data collection, validation, storage and distribution. Several also create data products and models, which are made available as services for external users including research, government and industry. These infrastructures are mostly complementary to each other, dealing with different data originators and/or different stages in the data value chain (see Box 5) from data acquisition to data products. They collaborate in EU projects such as ENVRI-FAIR[22], which aims to analyze and improve FAIRness of data services for environmental research infrastructures, including marine science, in order to align with the requirements of the European Open Science Cloud[23] (EOSC), as well as the Blue-Cloud Project[24].

Box 5: The Data Value Chain

The Data Value Chain is a series of steps and information flows needed to generate value and useful insights from data (Curry, 2016). In addition to these steps, it is important to make well thought out decisions on the types of data to collect and to have good sampling design, which influences the end-use and value of data. The data value chain is the centre of the future knowledge economy, bringing opportunities of digital developments to traditional sectors as part of the Big Data Value Public Private Partnership (BDV PPP) (Zillner *et al.,* 2017, European Commission 2013) as part of the Big Data Value Public Private Partnership (BDV PPP) (Zillner *et al.,* 2017, European Commission 2013)[25]. The same value chain applies for exploiting the value from data within marine science, leveraging large volumes of complex and heterogeneous data originating from multiple data sources. The key steps in the data value chain are:

Data acquisition gathering, filtering and cleaning raw data before it can be put in data repositories, analysed and stored;

Data analysis making raw data available for decision-making. It involves exploring, standardizing, transforming and modelling data. This includes the use of machine learning and data mining approaches;

Data curation the management of data over its lifecycle to ensure it meets quality requirements. It is performed by data curators that are responsible for ensuring trustworthiness, discoverability, accessibility and reusability of data;

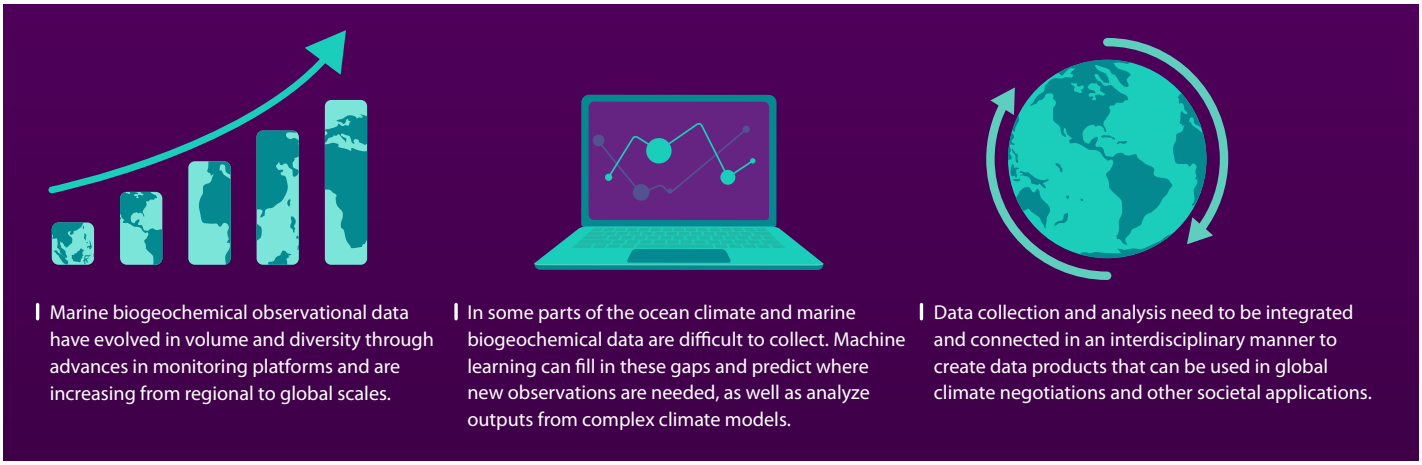
Data storage the persistent and scalable management of data to enable rapid access and end-user applications; and

Data usage the activities and applications that use data e.g. increased automation for data analysis or decision-making based on data products.

[14] https://www.seadatanet.org
[16] https://www.euro-argo.eu/
[17] https://www.emodnet.eu
[18] https://www.ebi.ac.uk/ena
[19] http://marine.copernicus.eu/
[20] https://otc.icos-cp.eu/
[21] https://www.lifewatch.eu/home
[22] https://envri.eu/home-envri-fair/
[23] https://ec.europa.eu/digital-single-market/en/european-open-science-cloud
[24] https://www.blue-cloud.org/
[25] https://ec.europa.eu/digital-single-market/en/news/elements-data-value-chain-strategy

2 Climate and marine biogeochemistry

The ocean supports life on Earth and is essential for regulating climate and absorbing excess heat and carbon originating from human activities (IPCC, 2019). Understanding how physical, biogeochemical and biological processes in the oceans will respond to and affect future climate change is therefore one of the most pressing grand challenges facing our society. Interest in emerging climate knowledge has expanded into the policymaking landscape. Impacts of climate change on the ocean include warming, stratification, sea-level rise, marine heatwaves, melting of sea ice, as well as deoxygenation, uptake of carbon dioxide (CO_2) by the ocean, the associated ocean acidification and consequences on marine ecosystem services. The adoption of a big data approach has the potential to revolutionize our ability to predict climate change trends and their impacts on the ocean and society.



Marine biogeochemical observational data have evolved in volume and diversity through advances in monitoring platforms and are increasing from regional to global scales.

In some parts of the ocean climate and marine biogeochemical data are difficult to collect. Machine learning can fill in these gaps and predict where new observations are needed, as well as analyze outputs from complex climate models.

Data collection and analysis need to be integrated and connected in an interdisciplinary manner to create data products that can be used in global climate negotiations and other societal applications.

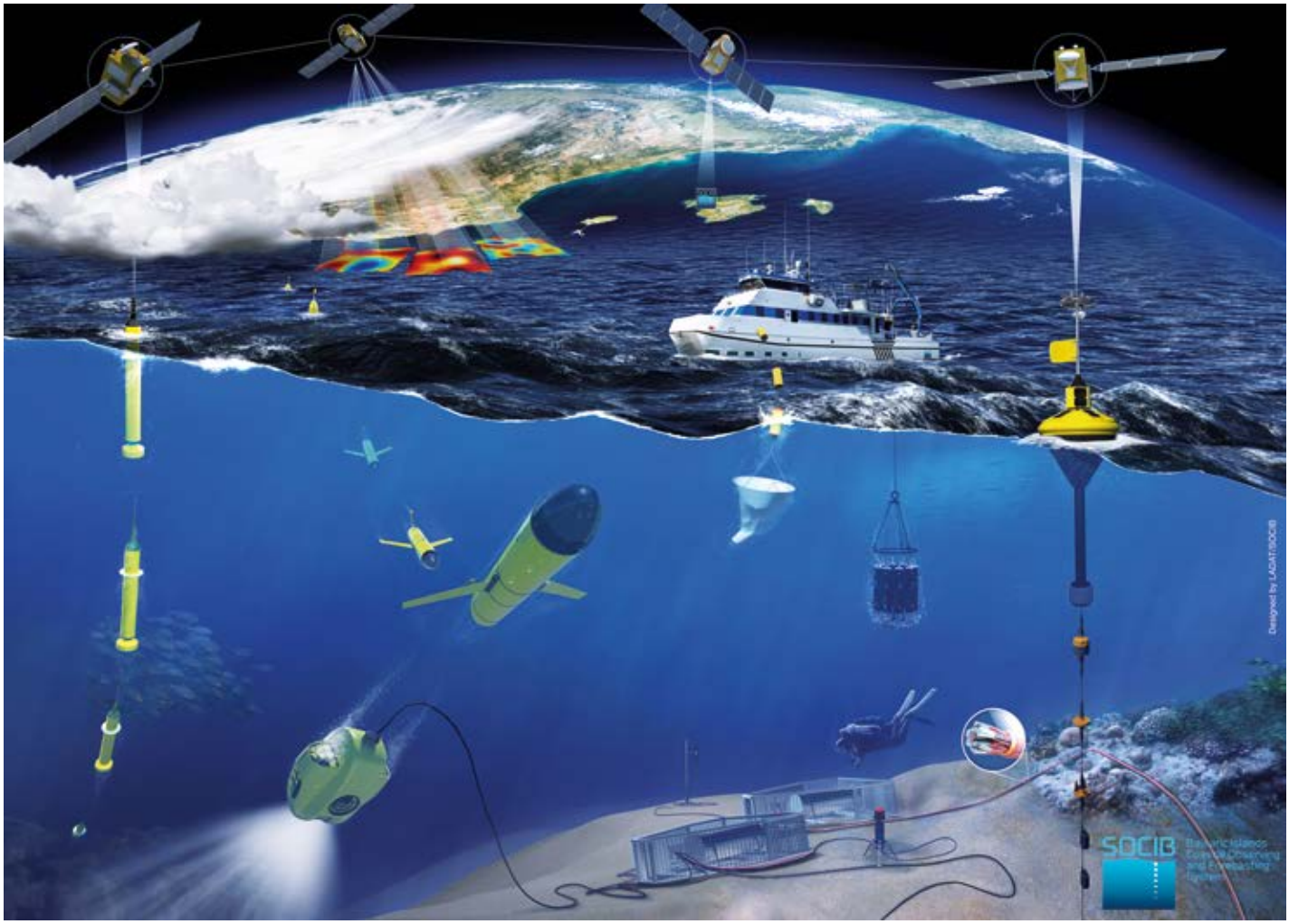
Marine data routinely collected and synthesized from observations and models are used to deepen our understanding of the climate system to improve our predictive capability. Non-linear interactions and feedback processes operate within the climate system. Big data approaches are needed to connect disjointed data collection and analysis in an interdisciplinary manner, in what today remains the topic of different sub-disciplines (physical, chemical, biological, geological, social, etc.). Marine data have grown rapidly in volume and variety through the advancement of observing infrastructures (examples shown in Figure 2.1), all of which serve different aspects of value creation from monitoring physical, biogeochemical and biological states, to the understanding of processes and enhancing forecasting ability. Part of these data are collected to satisfy the requirement of Essential Climate Variables[26] (ECV), which are needed to systematically observe Earth's changing climate, and which increasingly incorporate biological and biogeochemical parameters defined by the Global Climate Observing System[27] (GCOS). ECVs are informed by Essential Ocean Variables[28] (EOVs) defined by the Global Ocean Observing System[29] (GOOS) of the Intergovernmental Oceanographic Commission (IOC). These observational data, together with increasing outputs from model simulations and reanalysis products, cover broad disciplines of marine science (e.g. physical, biogeochemical and biological oceanography, and fisheries). To properly assess both direct and indirect impacts of climate change on the ocean, a fully integrated interdisciplinary approach is required. This section focuses on current initiatives working towards this vision via regional and global integration of marine climate and biogeochemical data, synthesis into data products, application of big data analytics, and use in global climate negotiations and other societal applications.

[26] https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables

[27] https://gcos.wmo.int/en/home

[28] www.goosocean.org/eov

[29] https://goosocean.org/



Credit: Balearic Island Coastal Observing and Forecasting System (SOCIB).

Figure 2.1. Examples of marine climate and biogeochemical observing infrastructures that enable data collection that passes though the data value chain to provide information used by policymakers, including Remotely Operated Vehicles (ROVs), Autonomous Underwater Vehicles (AUVs), buoys, remote sensing and ships.

Collection of marine biogeochemical observational data has evolved considerably through the advancement of sensor development, autonomous platforms (e.g. moorings, gliders, surface drifting buoys, coordinated Argo profiling floats), remote sensing, etc., and increasing use of ships of opportunity that are fitted with sensors. This has further accelerated our understanding of the physical dynamics of the ocean and their interactions with biogeochemistry and biology. European initiatives including SeaDataNet[30], EMODnet [31] and Copernicus Marine Environment Monitoring Service (CMEMS[32]) are crucial in the implementation of big data in marine science and the coordination of marine data collection, management and synthesis (see Box 4 and 6).

[30] https://www.seadatanet.org/

[31] https://www.emodnet.eu/

[32] http://marine.copernicus.eu/

Box 6: Marine data infrastructures and initiatives for collating marine biogeochemical data

SeaDataNet is a pan-European infrastructure for the management, indexing and access to ocean data and data products obtained from research cruises and observations in European coastal waters, regional seas and the global ocean. Physical, geological, chemical, biological and geophysical data from National Oceanographic Data Centres (NODCs) can be accessed and discovered in SeaDataNet. SeaDataNet develops and promotes common data standards, vocabularies for metadata and software tools, and is compliant with the Infrastructure for Spatial Information in Europe (INSPIRE[33] - a European Directive for developing common standards to easily share spatial information between public authorities in Europe through an online portal). This contributes to consistency in data quality, interoperability and FAIRness (Box 3). SeaDataNet standards also overcome the challenge of managing high volumes of high variety data and processing it into harmonized data collections. SeaDataNet has close cooperation with various ocean observing communities, including EuroGOOS and Euro-Argo[34], and other marine data management infrastructures including EMODnet and CMEMS, to enable data exchange.

Marine organic and inorganic carbon data are combined with a wealth of reanalysis and model forecasts, satellite data, and other *in situ* observations as part of CMEMS. Data are standardized and collated into comparable and searchable data sets. These data span decades and are complimented by physical and biogeochemical variables such as temperature, sea-level, chlorophyll concentration, primary productivity, and nutrient concentrations derived from Euro-Argo, EuroGOOS Regional Operational Oceanographic Systems (ROOS[35]), SeaDataNet and several international observation portals and networks. Data are analyzed and transformed into value-added data products including maps, anomalies and other statistical information by CMEMS. This enables monitoring of changes such as harmful algal blooms, which are increasing due to climate change. Data products are useful for end-users in a wide range of applications related to understanding climate change impacts such monitoring and reporting for the Marine Strategy Framework Directive[36] (MSFD) since some indicators are affected by climate change. CMEMS is working towards becoming the first Earth Observation programme that is artificial intelligence-ready by collating vast amounts of satellite data and data products and deploying a cloud-based platform to centralize access to data and analytical tools at scale via the Copernicus Data and Information Access Services (WEkEO DIAS[37]).

For climate related analyses, EMODnetChemistry collates chemical observations from a consortium of organizations. It focuses on eutrophication, contaminants, marine litter and ocean acidification. Data are brought together for seawater, sediment and biota to support the MSFD and that are relevant for global climate change.

The ocean component of the Integrated Carbon Observing System (ICOS-Ocean[38]) is a European initiative specifically for coordination, standardization, and processing of *in situ* ocean greenhouse gas measurements, including for carbon dioxide and associated carbonate chemistry variables. This European research infrastructure coordinates data from 21 stations in seven countries, collected from diverse platforms such as ships of opportunity and fixed buoys in oceanic and coastal waters (Steinhoff, 2019). ICOS-Ocean collaborates with the cloud computing service provider European Grid Infrastructure (EGI[39]) to make carbonate chemistry and other data FAIR and publicly available in near real-time on a dedicated data portal. Data are saved in storage linked to the EGI Data Hub[40], which can subsequently be used to perform near real-time estimates of marine greenhouse gas fluxes.

ICOS-Ocean contributes data and data management support to international synthesis efforts such as the Surface Ocean CO_2 Atlas (SOCAT[41]) and the Global Data Analysis Project (GLODAP[42]). These initiatives create complementary, publicly available databases that synthesize data from multiple sources with regular updates. They have been developed through a bottom-up collaboration among ship-going marine carbon scientists worldwide. SOCAT annually releases publicly available data products which document the increase in the important climate variable global surface ocean partial pressure of CO_2 (fCO_2) (Bakker *et al.*, 2016). GLODAP does the same for the full oceanic water column, and incorporates all ocean carbonate chemistry variables (fCO_2, pH, total alkalinity, and total dissolved inorganic carbon), as well as other physical and biogeochemical measurements including temperature, salinity, oxygen and nutrient concentrations, and chlorofluorocarbon data (Olsen *et al.*, 2019).

[33] https://inspire.ec.europa.eu/
[34] https://www.euro-argo.eu/
[35] http://eurogoos.eu/regional-operational-oceanographic-systems/
[36] https://ec.europa.eu/environment/marine/eu-coast-and-marine-policy/marine-strategy-framework-directive/index_en.htm
[37] https://www.wekeo.eu/
[38] https://otc.icos-cp.eu/
[39] www.egi.eu
[40] https://www.egi.eu/use-cases/research-infrastructures/icos/
[41] www.socat.info
[42] www.glodap.info

Many applications that use SOCAT and GLODAP (see Box 6) data products are based on machine learning. For example, multidisciplinary *in situ* observational data, such as marine carbonate chemistry variables, oxygen, and nutrients are difficult and expensive to collect from remote or inaccessible parts of world oceans. However, these measurements are required for the provision of accurate data on the role of the ocean in climate change. To address this, diverse interpolation methods, such as regression and neural networks, are used to fill the gaps (Rödenbeck *et al.,* 2015). For example, measurements of temperature, salinity, and oxygen have been used to estimate carbonate chemistry variables (Bittig *et al.,* 2018). Gap filling methods for the surface ocean typically determine relationships between sparse *in situ* observations (e.g. fCO_2) and variables from remote sensing and reanalysis products with good spatial and temporal coverage (e.g. sea surface temperature). Neural networks and other gap filling methods using SOCAT data products enable quantification of ocean CO_2 uptake in the Surface Ocean pCO_2 Mapping Intercomparison initiative (SOCOM[43]; Rödenbeck *et al.,* 2015). The resulting mapping products are verified against independent observations, where these are available, and are used to calculate the progression of ocean acidification using machine learning. Machine learning can also be used to identify and prioritize the type and resolution of urgently needed future observations or to help identify signal versus noise ratios in collected data, which gives more confidence as to the robustness of climate change trends inferred from available data.

SOCAT and GLODAP estimates play a central role in informing and feeding into the Global Carbon Budget[44], Intergovernmental Panel on Climate Change (IPCC[45]) reports and climate negotiations at the Conference of the Parties (COP) of the United Nations Framework Convention on Climate Change (UNFCCC[46]). Outcomes of these negotiations are then also considered by international bodies, e.g. the GCOS, who provide guidelines and strategies for future marine climate data collection to the observing communities (Figure 2.2).

The GLODAP and SOCAT synthesis products and associated mapping products are based on well-calibrated and highly accurate observations from ships and moorings. They are used to validate other measurements using machine learning e.g. from biogeochemical sensors on biogeochemical Argo floats and gliders, and for validation of ocean biogeochemical models within the framework of Observations for Model Intercomparisons Project (Obs4MIP[47]), Earth System Model Evaluation Tool (ESMValTool[48]), and the World Climate Research Programme Coupled Model Intercomparison Project (CMIP[49]).

Credit: Christian Harboe-Hansen

The Royal Arctic Line, a ship that has recorded pCO_2, sea surface temperature and sea surface salinity data for the past 16 years on the Denmark-Greenland passage.

[43] http://www.bgc-jena.mpg.de/SOCOM/
[44] https://www.globalcarbonproject.org/
[45] https://www.ipcc.ch/
[46] https://unfccc.int/process/bodies/supreme-bodies/conference-of-the-parties-cop
[47] https://esgf-node.llnl.gov/projects/obs4mips/
[48] https://www.esmvaltool.org/
[49] https://www.wcrp-climate.org/wgcm-cmip

Credits: IPCC, 2013, ©Global Carbon Project 2017, Quéré *et al.*, 2018 (CC BY 4.0) Landschützer *et al.*, 2014, Bakker *et al.*, 2016 (CC BY 4.0), the Pacific Marine Environmental Laboratory of NOAA, Tobias Steinhoff, Laurence Beaumont, Jerry Tjiputra.

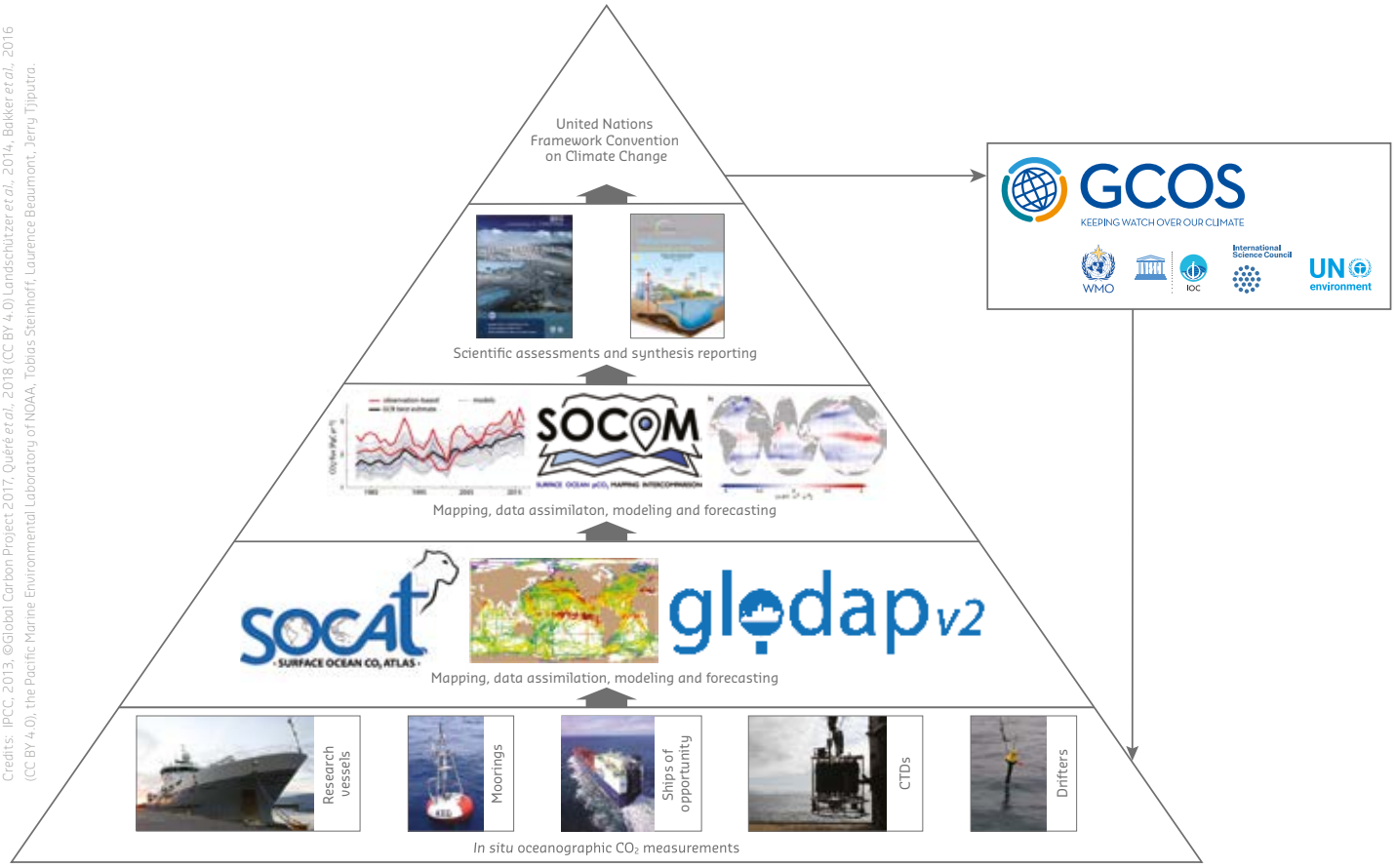


Figure 2.2. The value chain that connects *in situ* oceanographic measurements of carbonate chemistry variables to climate negotiations.

Insights from ocean observations need to be regularly communicated to the marine modelling community since data assimilation into models from both physical and biogeochemical observations has the potential to improve existing model projections. When model predictions are combined with data on marine ecosystem responses to multiple environmental and human pressures, they can provide policymakers and ecosystem managers with more relevant information on which to base key decisions, e.g. best-case fisheries management scenarios under given future climate projections (Serpetti *et al.,* 2017; ICES, 2020). Understanding the complex spatial and temporal variability of the physical and biological dynamics of the ocean depends on numerous region-specific factors. Progress can be made by applying a combination of unsupervised machine learning with different models and observations to identify emergent patterns with spatial and temporal commonalities (Sonnewald *et al.,* 2019). Machine learning can be used to identify and constrain projection biases in multi-model ensembles, which are linearly linked to observable variables (e.g. Goris *et al.,* 2018). This allows reliable models to be identified, unreliable models to be rejected, and eventually reduces climate projection uncertainties. Additionally, machine learning can be applied on a suite of marine biogeochemical and ecosystem data to address complex societally relevant issues, e.g. to predict extreme or harmful marine climate events such as marine heat waves (Liu *et al.,* 2016).

Credit: Dzintra Grinbergs

The seawall in Guernsey

Challenges and recommendations

Big data are important for understanding the role of marine biogeochemistry in climate. Models are becoming increasingly complex and coping with the rapidly expanding variety, volume and complexity of ocean-model outputs is a key challenge. This leads to subsequent challenges for analysis using conventional methods. Climate is a dynamic and non-linear system that requires long-term observations. However, funding for *in situ* marine biogeochemical data collection, processing and synthesis is scattered, unstable and largely dependent on individual research grants. A further challenge for data collection is that valuable *in situ* ocean surface CO_2 measurements are often not collected in nations' Exclusive Economic Zones (EEZs), since gaining permission to carry out these activities is very time consuming.

Without permission, shipboard scientists are required to switch off their instruments when entering countries' EEZs. International efforts are synthesizing marine biogeochemical observations into regional and global data products and into the international climate negotiation process. However, a key bottleneck is that many marine climate observations are not yet included in these syntheses, nor widely available via data repositories due to different data formats, questionable data veracity and a lack of understanding by scientists of the ethical reasons behind data sharing. Many groups have created best practices for data collection, management and storage but these are not yet all in a centralized place (Pearlman *et al.,* 2019). For more information on the challenges and solutions related to sharing marine data see Pendleton *et al.* (2019).

To increase the uptake of big data in marine biogeochemistry and climate we recommend to:

- Adopt global operational data standards based on FAIR principles;
- Coordinate and centralize best practices for ocean observations;
- Motivate and educate data originators and funding agencies on the use of existing marine data infrastructures to enable all marine climate observations to be easily accessible and interoperable. This could be done by the more widespread use of Digital Object Identifiers (DOIs) allowing data to be cited;
- Increase collaborations between marine data management infrastructures and e-infrastructures, such as that between ICOS-Ocean[50] and EGI[51], to provide seamless tools to synthesize, summarize, visualize and analyse a wide variety of data;
- Adopt new analytical workflows such as the 'zero download' paradigm, where scientists process and analyse their data using cloud computing;
- Provide long-term funding for the continued provision of accurate *in situ* biogeochemical observations and their synthesis into climate research and international climate negotiations;
- Implement a high-level agreement for *in situ* surface ocean CO_2 measurements (similar to that of weather observations), to increase their spatial and temporal resoltuion by including measurements in national EEZs;
- Establish official partnerships between marine scientists and industries, e.g. those operating commercial ships hosting scientists with their instruments and sensors, preferably with governmental supports, towards a long-term commitment in high-quality data collections; and
- Increase interdisciplinary collaborations, e.g. among marine ecologists, biogeochemical and physical oceanographers, climate scientists, statisticians, socio-economists, data managers and computer scientists. Such collaborations should integrate diverse data into data products such as Essential Climate Variables (ECVs), Essential Biological Variables (EBVs) and Essential Ocean Variables (EOVs) with higher levels of accuracy and with broader applications.

[50] https://www.icos-cp.eu/observations/ocean/otc

[51] https://www.egi.eu/

3 Habitat mapping for marine conservation

With advances in ocean observatories and the exponential growth in data acquisition, big data are becoming indispensable for marine conservation and the management of human activities. No area of the ocean is untouched by human activity (Jones *et al.*, 2018) and the negative effects of e.g. climate change, unsustainable fishing, shipping and pollution are rapidly increasing (Halpern *et al.*, 2019). Thus, there is an urgent need to ensure the protection of marine ecosystems in coastal and nearshore environments, and in offshore and deep-sea areas, including Areas Beyond National Jurisdiction (ABNJ)[52] especially given the increasing interest from oil, gas, and mining industries to exploit the resources in these areas. To address this problem, ecologically coherent networks of Marine Protected Areas (MPAs) across vast spatial extents have been suggested (Grorud-Colvert *et al.*, 2014), including in the deep-sea, with the International Union for the Conservation of Nature (IUCN) calling for 30% of the ocean to be designated as MPAs by 2030.

To design and create these MPA networks, and to manage specific MPAs, Marine Restricted Areas (MRA) or Fisheries Restricted Areas (FRAs), in-depth knowledge of the marine habitats and an adequately mapped seafloor is needed. Currently, less than 10% of the ocean is mapped in adequate resolution with modern high-resolution technology. Big data will be important for achieving both the proposed networks of MPAs and the aim of the SeaBed 2030[53] initiative: to make a map of the global ocean seafloor by 2030 using all available bathymetric data. This will require strong international cooperation with respect to data acquisition, sharing, assimilation and compilation. The use of the most modern technology for bathymetric data collection needs to be fostered worldwide. It will need to rely on acoustic technologies deployed from surface submerged vessels, ship to shore data transfer, data cloud storage, and new data-processing tools using machine learning. EMODnet bathymetry[54] is the European counterpart for SeaBed 2030, for which it cooperates globally with the International Hydrographic Organization (IHO) and the International Oceanographic Data and Information Exchange (IODE). EMODnet bathymetry are now using a high-performance computing cloud platform for collaborative data analysis and processing. EMODnet bathymetry brings together data from bathymetric surveys into Digital Terrain Models (DTMs) for the European seas. SeaDataNet infrastructure is used to manage and gather the data sets and associated metadata. The maps are openly available for users to view and download.

Big data will be invaluable in creating high-resolution habitat maps that combine bathymetry data with other multidisciplinary, large-scale habitat data. This is demonstrated in the following case study on the Bari Canyon that focuses on creating local-scale habitat maps to inform the planning of a potential site for a new deep-sea MPA.

[52] https://www.unep-wcmc.org/resources-and-data/governance-of-abnj
[53] https://www.gebco.net/about_us/seabed2030_project/
[54] https://www.emodnet-bathymetry.eu

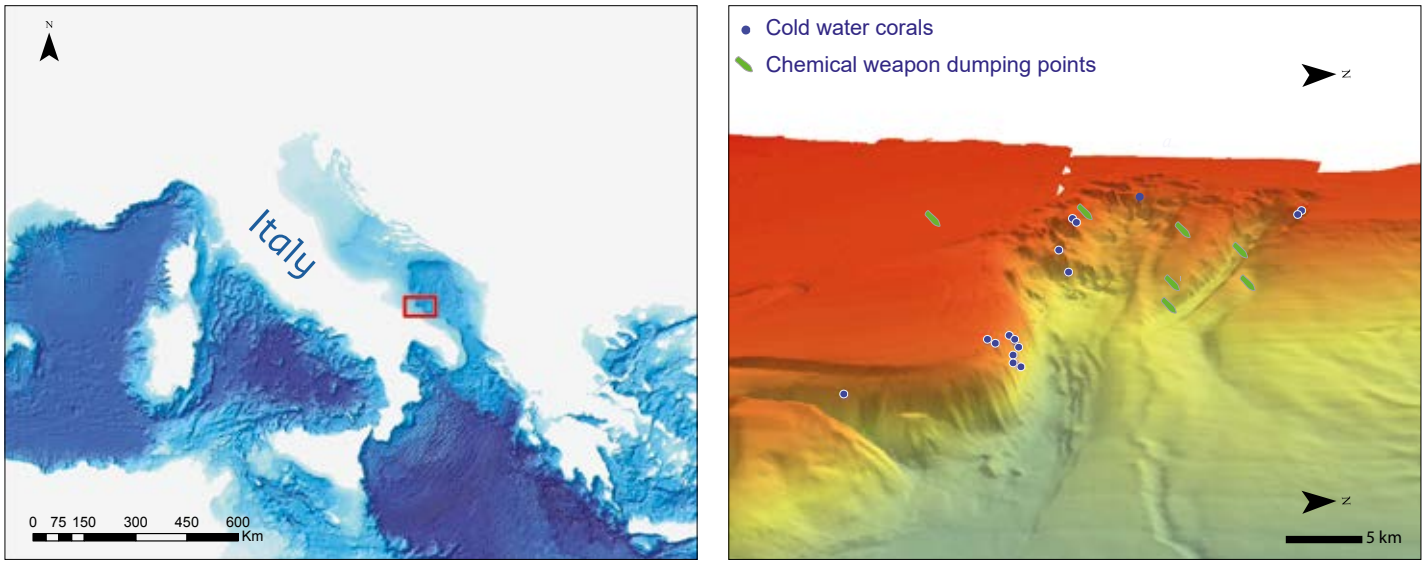
The Bari Canyon case study

The Bari Canyon is a deep-sea ecosystem located in the Southern Adriatic (Mediterranean Sea). It is home to cold-water corals and a hotspot for biodiversity and ecosystem functioning (D'Onghia *et al.,* 2003), but is highly vulnerable and has a low recovery rate from environmental disturbances (Figure 3.1). It is also an essential fish habitat that is vital for reproduction, growth, feeding and shelter of commercial species (Sion *et al.,* 2019). However, it is threatened by marine litter and the impact of fishing activity from bottom trawling and long line fishing. The Bari Canyon has not yet been selected for MPA designation, and there is debate on whether this should happen. Here we use the Bari Canyon as an example of data that can be used to help designate an MPA in this area.

Habitat maps are important to support the design of adequate measures of protection and management for any MPA. Challenges exist for the acquisition of vast amounts of habitat data, the adoption of novel data management approaches to ease analysis and collaboration, the development of automated data analysis, and the creation of novel sensor technologies for near real-time environmental monitoring.

For the Bari Canyon, hundreds of gigabytes of high variety multidisciplinary and heterogeneous oceanographic data have been collected from a multitude of data sources (Figure 3.2). The integration and management of this multidisciplinary, complex data is a key big data challenge because the data are:

- In many heterogeneous data storage formats (e.g. the format for multibeam echosounder (MBES) raw data varies by device manufacturer), leading to difficulties with sharing data and use in downstream applications;
- Collected across different temporal scales (e.g. over a number of hours, days, months, seasons, years) creating complexities in raw data, processed and filtered data, and data products such as Digital Terrain Models; and
- In large volumes (raw data) and require significant data storage capacity and computational power for analysis.



Credit: Federica Foglini, CNR ISMAR

Figure 3.1: Left: Bathymetric map showing the location of the Bari Canyon in (red square). Right: Map showing the location of cold water corals and chemical weapon dumping points in the Bari Canyon.

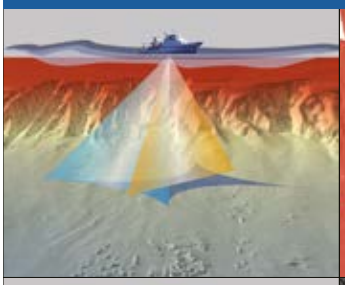
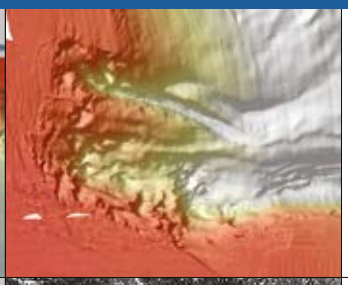
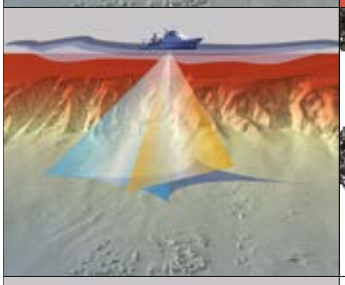
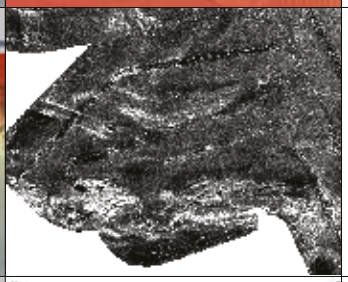
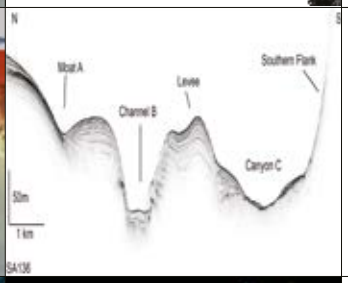
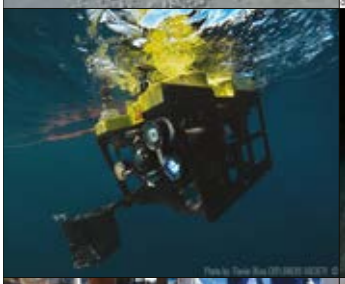
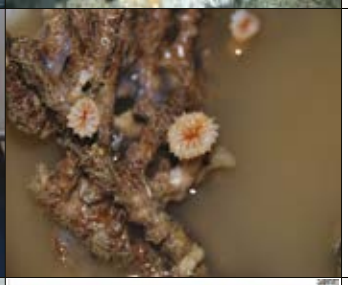
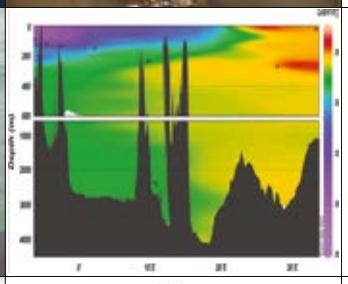
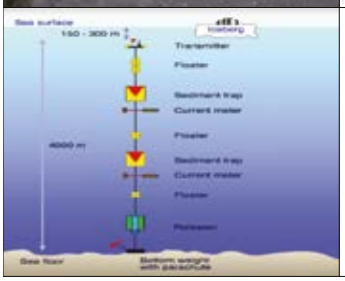
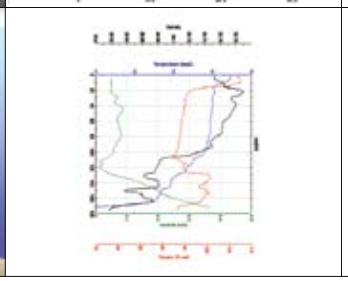
Device	Data collected	Description
		**Multibeam echosounders (MBES)** produce high-resolution data on seafloor bathymetry and morphology with continuous coverage. This is fundamental for habitat identification. Image credits: CNR ISMAR
		**MBES backscatter** measures the acoustic reflectivity of the seafloor and provides data on substrate composition. Image credits: CNR ISMAR
		**Sub bottom profilers** are single-channel systems used for shallow reflection seismic profiling to characterize sediment or rock on the seafloor. Image credits: CNR ISMAR
		**Remotely Operated Vehicles (ROVs)** acquire images and videos that are needed to characterize the megabenthic communities and for habitat mapping. Image credit right: CNR ISMAR
		**Grab, box-corers, robotic arms and push corers** collect seabed and biological samples. This provides data on macro- and micro-benthic epifaunal, infaunal components, and textural and compositional properties of the seabed. Image credits: CNR ISMAR
		**Conductivity, temperature and depth (CTD) casts** collect water samples and provide data on temperature, salinity, carbonate chemistry, nutrient concentrations and isotope composition. Image credit left: Artevelde Hogeschool, right: Paolo Montagna
		**Mooring stations** measure oceanographic variables including water characteristics, vertical particle fluxes, hydrology, suspended matter distribution, temperature, salinity and current speed. Image credits left: Hannes Grobe CC BY-SA 2.5, right: CNR ISMAR

Figure 3.2. Key data sources available for the design of a potential deep-sea MPA for the Bari Canyon.

To overcome challenges in data heterogeneity and fragmentation, and to enable efficient data integration, a database known as 'Spatial Relational Database Management System' (RDBMS: Geodatabase[55]) was created for the Bari Canyon based on adapted INSPIRE[56] guidelines, which includes WebGIS (i.e. a web-based version of a geographical information system) and a metadata catalogue. Integrated WebGIS services, reduce data heterogeneity and fragmentation. This RDBMS is accessible via a Virtual Research Environment (VRE, see Box 7), developed within the framework of the Horizon 2020 EVER-EST project[57] (Figure 3.3). The EVER-EST VRE is a multidisciplinary platform based on research objects, which aggregate information in a form that can be processed by both humans and machines and that follow FAIR data principles. Research objects provide the basis for the development of e-infrastructures for preserving, sharing and reusing scientific data and knowledge within and across communities (Garcia-Silva *et al.,* 2019).

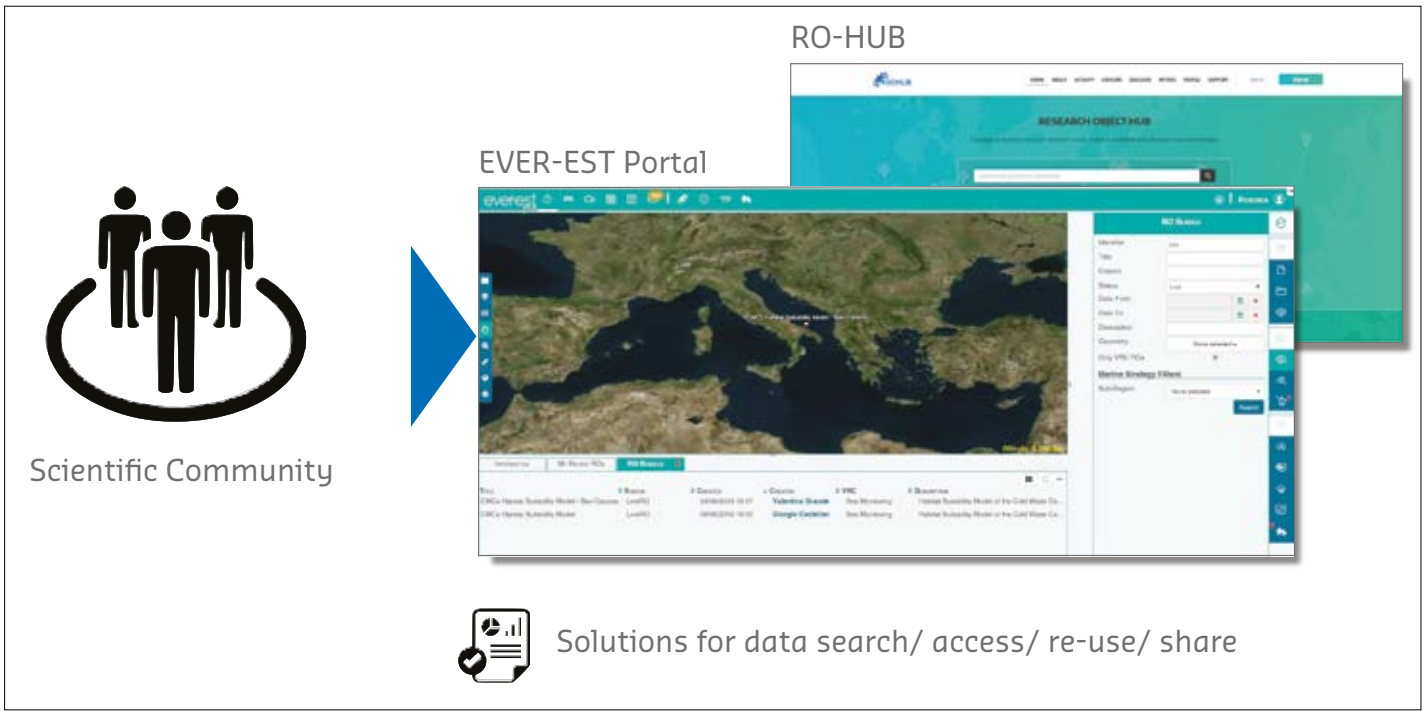
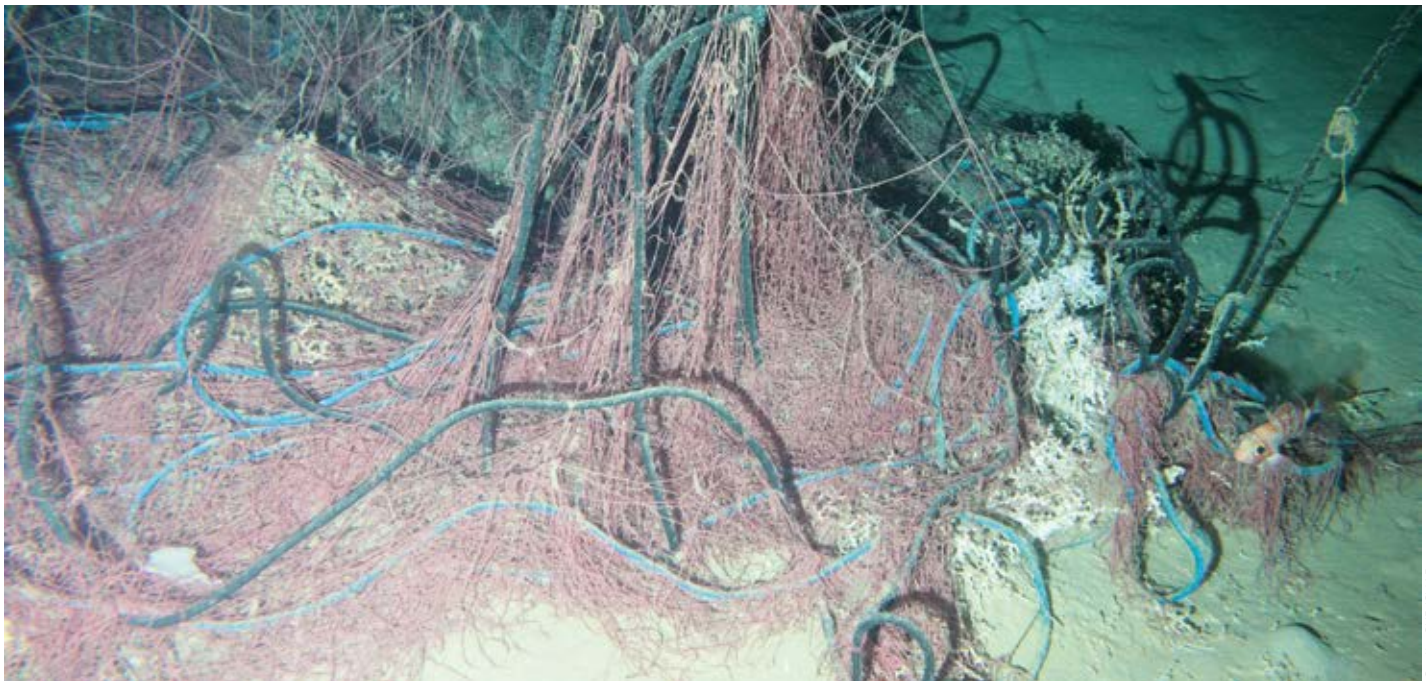


Figure 3.3. EVER-EST Virtual Research Environment (VRE) portal, which enables scientists to access and visualize data, and create, publish and geographically represent research objects. The research object hub (RO-HUB) portal collects and stores all research objects and includes functions to visualize research object quality, history, and metadata. Virtual machines contain data-processing tools and a workflow manager that enables users to run executable e-workflows.



Credit: CNR ISMAR.

Abandoned long line fishing gear in the Bari Canyon.

[55] http://gismarblack.bo.ismar.cnr.it:8080/mokaApp/apps/ismarBoApp/index.html?null

[56] https://inspire.ec.europa.eu/

[57] https://ever-est.eu/

Box 7: Virtual Research Environments (VREs) and the European Open Science Cloud (EOSC)

Virtual Research Environments (VREs) are web-based data-sharing platforms based on cloud computing with integrated analytical tools, which are crucial for creating and analysing big data. They bring together scientific and computer science communities and are a powerful tool to help bridge the gap between marine researchers and European Open Science Cloud[58] (EOSC) service providers i.e. e-infrastructures offering services for cloud computing. VREs are based on FAIR data principles and allow discovery, access, sharing, validation, processing and re-use of heterogeneous data, algorithms, results, and best practices within and between research communities and with the general public. These features overcome barriers for sharing data and information, and facilitate transition to open science. VREs are a key step towards integrating research into EOSC.

EOSC aims to provide a centralized virtual environment for EU researchers to store, manage, analyse and re-use FAIR research data and data products across borders and scientific disciplines. Launched by the European Commission in 2016, it will leverage and federate existing e-infrastructures and scientific data infrastructure, making access to scientific data and outputs easier and more efficient. Several projects are developing parts of the EOSC including projects relevant for the planning of the overall approach and structure such as EOSC Pilot[59], EOSC-hub[60], and EOSC Secretariat[61]. There are striking differences in usage rates of EOSC services between different scientific domains and many more marine scientists could be engaged in building or utilizing potential EOSC services.

The **Blue-Cloud** project[62] is part of 'The Future of Seas and Oceans Flagship Initiative' of the EU and aims to develop a thematic marine EOSC cloud as a pilot to demonstrate the potential of cloud-based open science for ocean sustainability. The project is developing a VRE for the collaborative integration, exchange and analysis of data from existing European marine data infrastructures including SeaDataNet[63], EMODnet[64], Euro Argo[65], EurOBIS[66], ICOS-Ocean[67], EcoTaxa[68], ELIXIR[69], Euro-Bioimaging[70] and European Nucleotide Archive (ENA)[71], which are brought together with e-infrastructures including EUDAT[72], D4Science[73] and WEkEO DIAS[74]. It aims to develop a smart federation of blue data resources, computing facilities and analytical tools to provide researchers with access to large volumes of high variety data from *in situ* and remote sensing observations, data products and outputs of numerical models. This will be piloted through five demonstrators addressing societal grand challenges linked to the UN Decade of Ocean Science for Sustainable Development. The Blue-Cloud project will also develop a roadmap for the expansion and sustainability of these infrastructures by the EU, with input welcome from stakeholders. The implementation of effective handling of big data, data mining and machine learning are key challenges to be addressed.

In addition to the Blue-Cloud project, the marine science community is beginning to use VREs more frequently. VREs are also developed and offered in the PlutoF platform[75] for distributed management of complex biological data (Abarenkov *et al.,* 2010), and is currently being developed to become a service for the EOSC as part of the EOSC-NORDIC[76] project and the Nordic e-Infrastructure Collaboration (NeIC). SeaDataCloud[77] (a collaboration between SeaDataNet and EUDAT) also works towards solutions for VREs and computational infrastructures. It is advancing SeaDataNet services and increasing their usage by adopting cloud and high-performance computing technologies in a collaborative environment with high analytical performance. European Strategy Forum on Research Infrastructures (ESFRI) infrastructures such as EMSO-ERIC[78] and LifeWatch ERIC[79] are also connecting scientific end-users to EOSC resources.

[58] https://ec.europa.eu/digital-single-market/en/european-open-science-cloud
[59] https://eoscpilot.eu/
[60] https://www.eosc-hub.eu/
[61] https://www.eoscsecretariat.eu/
[62] https://www.blue-cloud.org
[63] https://www.seadatanet.org/
[64] https://www.emodnet.eu/
[65] https://www.euro-argo.eu/
[66] https://www.eurobis.org/
[67] https://otc.icos-cp.eu/
[68] https://ecotaxa.obs-vlfr.fr/
[69] https://elixir-europe.org/
[70] https://www.eurobioimaging.eu/
[71] https://www.ebi.ac.uk/ena
[72] https://www.eudat.eu/
[73] https://www.d4science.org/
[74] https://www.wekeo.eu/
[75] http://plutof.ut.ee/
[76] https://www.csc.fi/-/eosc-nordic
[77] https://www.seadatanet.org/About-us/SeaDataCloud
[78] http://emso.eu/
[79] https://www.lifewatch.eu/

The EVER-EST VRE facilitates automated data processing and machine learning by providing the computational e-infrastructure to run algorithms and by encapsulating workflows for data processing within research objects. Approaches for automated analysis of bathymetric data using machine learning are being developed that are beneficial in terms of increasing time efficiency and reducing personnel effort. Examples include:

- Bathymetric data reduction algorithms that minimize the size of data to be stored and thereby increasing the ease and efficiency of data analysis (Wlodarczyk-Sielicka & Stateczny, 2016);
- Machine learning to automatically classify and reject common types of background noise in sonar survey data and dramatically reduce processing time; and
- Novel automated and semi-automated classification techniques to cope with the high volume and high diversity of bathymetry data, validated through physical samples and images. In the Bari Canyon a combined approach uses manual interpretation and automatic classification techniques such as the Remote Sensing Object-Based Image Analysis (RSOBIA) (Lacharité *et al.,* 2018). This reduces the time and human effort required for image segmentation while maintaining expert data analysis, which will always be necessary for accurate habitat mapping.

Increasing volumes of video footage obtained from Remotely Operated Vehicles (ROVs) in the Bari Canyon is currently manually annotated with the support of the ADELIE[80] software from the Institut français de recherche pour l'exploitation de la mer (Ifremer) which georeferences the data. Manual annotation is extremely time intensive and presents a significant bottleneck (i.e. one working week per hour of video footage). This software would benefit from automated image recognition using machine learning (e.g. Piechaud *et al.,* 2019), which requires further development of the algorithms to fully implement automated feature detection given the complexity of benthic habitats and species diversity (Qin *et al.,* 2016).

A further application of machine learning in the Bari Canyon is for the creation of habitat-suitability models using Bayesian analysis (Tantipisanuh *et al.,* 2014). These models couple environmental data, hydrodynamic models, and ROV observations to characterize environmental conditions where potential cold-water coral sites may exist and are used to make decisions on measures for their protection (Bargain *et al.,* 2018).

The Bari Canyon will be one of the test sites for the deployment of new long-endurance Autonomous Underwater Vehicles (AUV) within the framework of the H2020 Project ENDURUNS[81] that will allow continuous spatial and temporal monitoring in the area. The AUV will be capable of operating for up to eight months and will integrate various sensors including a multibeam echosounder (MBES), sidescan sonars, a high-resolution camera, and Conductivity, Temperature and Depth sensors (CTDs). The AUV will be equipped with on-board algorithms, i.e. "Edge AI" that will automatically filter, remap, and compress collected data in order to reduce its volume while maintaining its resolution and content. Data will be categorized and stored wirelessly on SD cards in special pencil-like bubbles, which can be ejected to the surface when required allowing more efficient data transmission. An unmanned surface vehicle (USV) will follow the AUV and transmit data to an onshore control centre. Further improvements in data transfer efficiency through satellite and high-bandwidth connection systems will be a key enabler. Trade-offs should be considered between the amount of data processed on-board a research vessel or onshore vs. on-board an AUV in order to maximize energy available for data collection.

Credit: Ezio Amato

Discarded chemical weapon in the Bari Canyon.

[80] https://www.flotteoceanographique.fr/en/The-Fleet/Shipboard-software/ADELIE

[81] https://enduruns.eu/

Challenges and recommendations

The Bari Canyon case study demonstrates the comprehensive data management and analysis needed for a multidisciplinary observatory system at a local level, with the aim of creating high-resolution habitat maps for planning new marine protected areas using a big data approach. The need for continuous monitoring and subsequent advances in ocean observation equipment leads marine scientists to enter into the big data era with the acquisition of increasing volumes of data. Combining and centralizing large volumes of high variety data is a key challenge. Machine learning has the potential to replace some oceanographic devices with truly autonomous sensors that are able to extract information in real-time. However, it also introduces the risk of artefacts, errors and noise going unchecked with potential consequences in the misinterpretation of bathymetric and other features.

To increase use of big data for local-scale habitat mapping for marine conservation we recommend to:

- Manage the data lifecycle using FAIR principles;
- Integrate and analyse data using VREs based on research objects;
- Increasingly adopt machine learning for data processing, analysis and modelling to reduce human intervention;
- Integrate data acquisition and analyses at each potential MPA site and at larger scales to design networks of MPAs;
- Continue development of novel technology, including satellite and high-bandwidth network connectivity, to increase data transfer efficiency and real-time, or near real-time data transfer; and
- Ensure data robustness and veracity and guide the use of machine learning by humans to minimize risks.

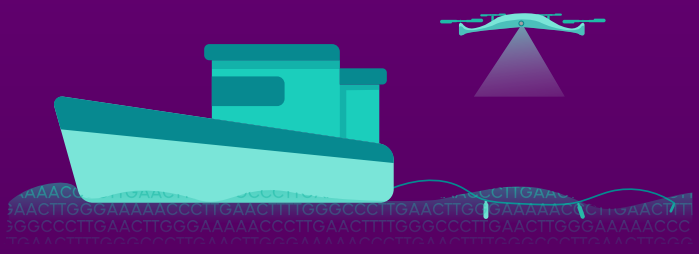
Credit: IMEDEA / CSIC-UIB / © Enrique Vidal Vijande

A Slocum Glider floats on the sea surface transmitting mission and sea state data via satellite to the IMEDEA (Mediterranean Institute for Advanced Studies) data facility.

4 Marine biological observations

Biological observations need to improve radically to serve our understanding of marine ecosystems and biodiversity under long-term global change and multiple stressors (Benedetti-Cecchi *et al.,* 2018). However, this is not trivial as biological properties are more difficult to measure and integrate compared to physico-chemical parameters. The achievement of an extensive, coordinated, and standardized global network of biological observations is a key goal for the next decade and will enable scientifically viable data products relating to Essential Biodiversity Variables (EBVs, Kissling *et al.,* 2018) and Essential Ocean Variables (EOVs, Miloslavich *et al.,* 2018) It will also deliver critical data for descriptors of the Marine Strategy Framework Directive (MSFD) and other legislation for the management of marine biodiversity.

We need a globally connected network of long-term biological observations to create and analyze big data for improved understanding of marine biodiversity under global change.



New biological data sources such as genetic sequences, imagery and hydro acoustic data will be used more frequently in the big data era. These rapidly generate enormous volumes of data and can be combined and analyzed using artificial intelligence.

The focus on complexity and larger spatio-temporal scales in marine biological research is increasing rapidly and is driven by a combination of new observation techniques producing a wealth of biological data, fast method development, increasing availability of high computational capacity, and the increasing adoption of an open science culture. These advances have rejuvenated the field of systems ecology and many other marine biological disciplines. However, the complexity of biological systems constitutes a major challenge for the scientific community in its attempt to structure, manage and link data.

Prominent bottlenecks preventing marine biology and systems ecology from becoming a big data discipline include:

- Lack of semantic concepts, e.g. phenotypes for marine species (Costello *et al.,* 2015);
- Environmental DNA (e-DNA) are typically organized around samples, while conventional biodiversity data are organized around species. It is therefore difficult to integrate e-DNA data with global biodiversity data infrastructures, although some initial resources have been developed (e.g. Deck *et al.,* 2017; Buttigieg *et al.,* 2019); and
- The absence of platforms for archiving and processing large amounts of marine image data, and linking these with global biodiversity data infrastructures.

A deeper understanding of spatio-temporal trends remains limited when relying on conventional measurement methods such as abundance of individuals from sediment cores, dredging, and trawling samples as currently used in most of the national biological monitoring programs. As we enter the big data era, these methods need to be complimented with new biological data sources, including high-throughput imagery, hydro-acoustic data and genetic sequences. This section highlights the data sources and advances needed to move towards regionally and globally integrated data products based on marine biological observations.

Marine genomic observatories

Between 2000 and 2020 a number of microbial planetary inventories took place resulting in massive reference data sets for marine genomics research. Examples include the Global Ocean Sampling Expedition (Rusch *et al.,* 2007) and the Tara Ocean Expedition (Karsenti *et al.,* 2011). These expeditions created large data sets, but they were generated only once. Such genetic inventories are currently complemented by genetic monitoring programs, which collect repeated measurements of the genetic content in a marine ecosystem (Davies *et al.,* 2014). For example, the Ocean Sampling Day campaign began in 2013 and generates a continuously growing global data set on coastal microbial diversity every year on the same day (Kopf *et al.,* 2015).

More recently, genetic monitoring has expanded to measure benthic communities using Artificial Reef Monitoring Structures (ARMS), which may function as biological forecast systems for invasive species in the future[82]. ARMS are set up in proximity to marine research stations and left for two to four months, after which the species composition of settled recruitment organisms are analysed. These early observations inform machine learning algorithms designed to model suitable habitats and predict potential areas for further spreading of invasive species in European coasts. This relies on novel monitoring methods as well as thorough data management, allowing for interoperability between databases for genetic data (Barcode of Life Data Systems, BOLD[83]), species occurrences (Global Biodiversity Information Facility, GBIF[84]) and species taxonomy (World Register of Marine Species, WoRMS[85]), as well as access to computational resources. Linking measurements (i.e. raw and processed genetic data, images, derived species lists, estimates of quantity) with metadata (e.g. field and laboratory protocols) and downstream data products (e.g. species lists derived from the analysis of sequence data) are facilitated by the development of eco-genomic standards (Yilmaz *et al.,* 2011) and eco-genomic databases (Deck *et al.,* 2017). These linkages need to be further developed in the future. A Virtual Research Environment (VRE, PlutoF[86], see Box 7) used in the European ARMS program was essential for creating and publishing very large and consistent data sets from individual and distributed efforts and providing links across the entire data lifecycle.

Credit: Alf Norkko

Artificial Reef Monitoring Structures (ARMS) function as biological forecasts for marine invasive species.

[83] http://www.boldsystems.org/
[84] https://www.gbif.org/
[85] http://www.marinespecies.org/
[86] http://plutof.ut.ee/

Credit: VLIZ CCBY-NC-SA 4.0

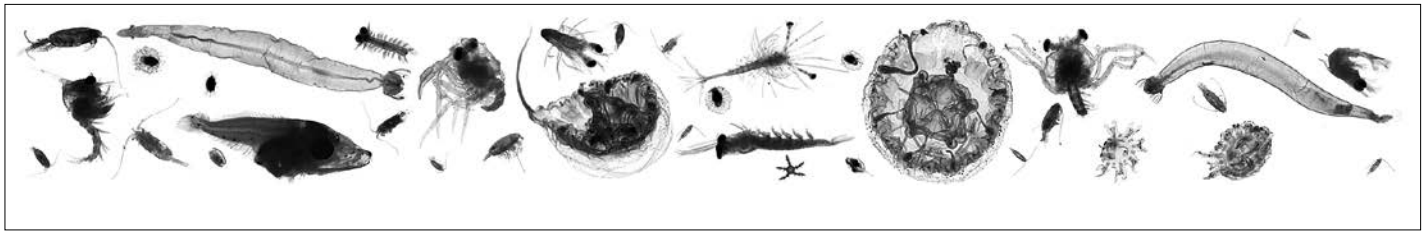
A video plankton recorder, which is a semi-automated underwater microscope that records images of plankton in real-time. Machine learning is being developed to automatically classify the images.

Image and hydro-acoustic observatories

As is described in the Bari Canyon Case study in Chapter 3, marine biologists are increasingly using autonomously operated technologies for data collection, which generate enormous volumes of data at ever faster speeds. This is especially the case for high-definition optical imagery coming from ROV's, AUVs, drop-cameras, video plankton recorders, and drones as well as for hydro-acoustic data coming from passive hydrophones that collect data on underwater soundscapes, and active sonars such as single- and multi- beam echosounders, side-scanners, sub-bottom profilers, and fisheries echosounders and sonars. The scientific potential of these technologies will change the *modus operandi* of the marine biological community in the near future. The combination of hydro-acoustic and camera-based systems allows scientists to extract biological information from marine ecosystems with unprecedented quantity and quality. As image and sonar-based research is revolutionizing the fields of marine biology and biodiversity monitoring, these methods impose completely novel demands on data management and processing. Machine learning methods can be used to increase processing efficiency. One example is the EcoTaxa[87] database that has integrated supervised machine learning algorithms and human curation of plankton images, which are processed at a rate of several thousand per hour therefore allowing processing at scale. This database has been used to validate approximately 40 million occurrences of plankton throughout the world's oceans.

Autonomous optical sensors and cameras will soon be collecting data at increased spatio-temporal scales and a systems architecture will be needed to integrate and analyse data that are geographically widely distributed (e.g. data and resources will need to be integrated regionally and globally from many local-scale habitat mapping projects like the Bari Canyon). This architecture should connect raw data, analysis platforms, and archiving resources. ESFRI programs LifeWatch[88] and the European Marine Biological Resource Centre (EMBRC ERIC[89]) are currently addressing technical challenges that come with such systems architecture design. They are building service-oriented architecture that allows systems to be coupled, data to be exchanged, and links to be made with high-performance computing services, which is also the goal of the European Open Science Cloud (EOSC, see Box 7). In addition, regional e-infrastructures such as the Nordic e-infrastructure Collaboration (NeIC[90]) are supporting marine biologists in the same way. The seamless availability of data and computational resources is needed to allow marine biologists to develop more holistic and interdisciplinary approaches to investigate how entire communities of organisms interact with their physical environment and with human activities. For example, use-oriented simulations, or avatars, of entire social-ecological systems are being developed by the Islands Digital Ecosystems Avatars (IDEA) consortium (Davies *et al.,* 2016).

The Swedish LifeWatch[91] has a pilot system architecture (Figure 4.1) for the management of large volumes of hydro-acoustic and image data collected by various projects. In this case, scientists need to be able to make decisions on which data are valuable and worth archiving, and which are not. This becomes important when dealing with large volumes of data and is not an easy task since knowing which data will be most influential is a challenge. Archiving is easiest when raw data are linked with analysis platforms and archiving resources. Data sets deemed highly valuable will be accessible online (also known as hot storage), while less-used data sets will be stored offline (cold storage). Data management plans,



Credit: EcoTaxa

The EcoTaxa database uses machine learning algorithms to process plankton images at scale.

[87] http://ecotaxa.obs-vlfr.fr
[88] https://www.lifewatch.eu/
[89] http://www.embrc.eu/
[90] https://neic.no/
[91] https://biodiversitydata.se/

e.g. those promoted by Assemble Plus as part of 'Open Research Data Pilot'[92], contain useful information on whether the data are likely to be valuable. Analytical results will show whether the data were useful or not and the frequency of metadata searches will determine data value. All data sets are available for exploration using machine learning algorithms developed by the Ocean Data Factory consortium[93]. The end results are data products on the distribution and abundance of key ecological species, such as corals and sponges as well as signatures of human activities (e.g. trawling tracks and marine litter).

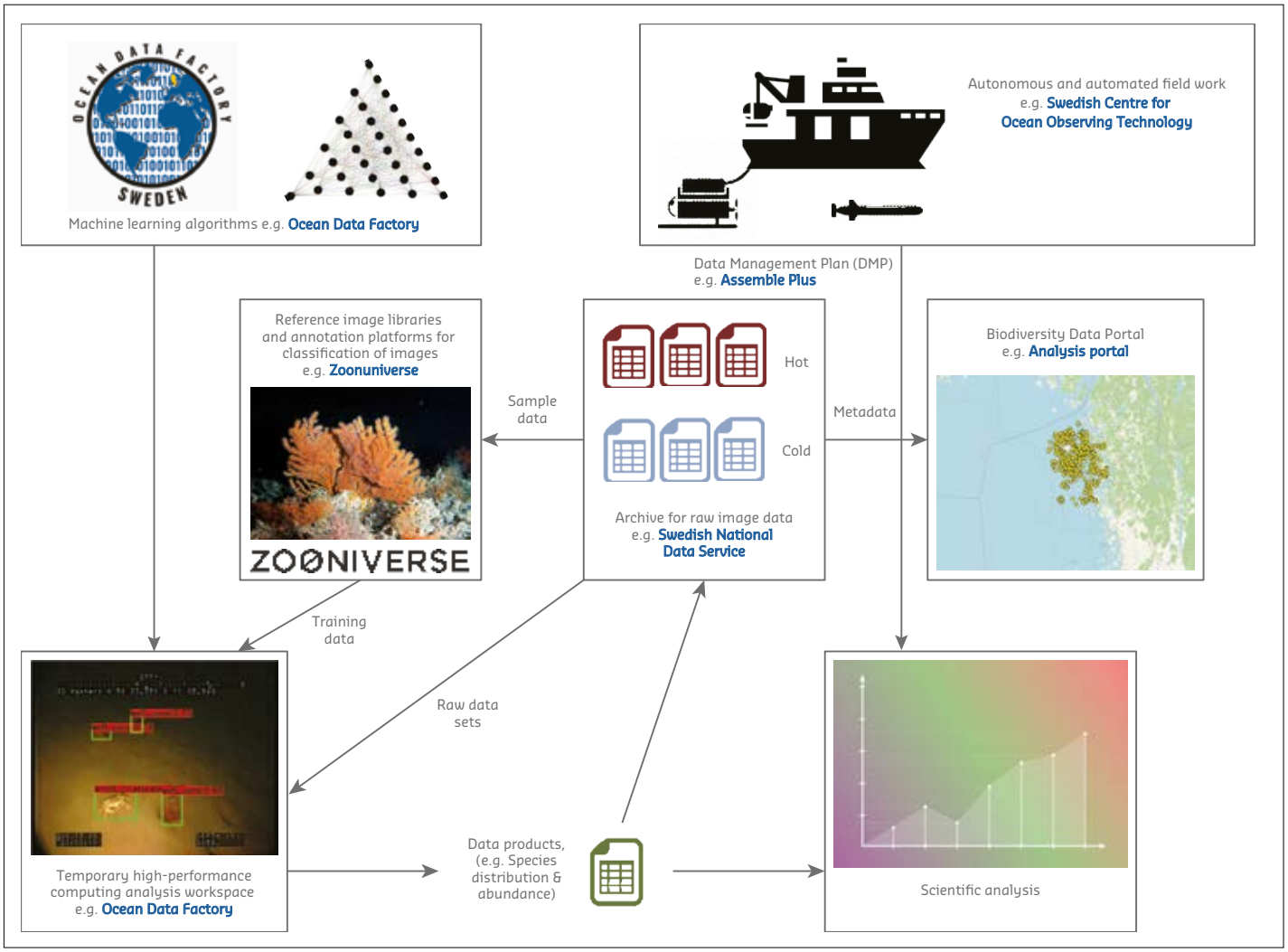


Figure 4.1. Example of a systems architecture for management and processing of large volumes of image and hydro-acoustic data, currently piloted as part of Swedish LifeWatch.

[92] http://www.assembleplus.eu/access/DMP

[93] https://scootech.se/odf/

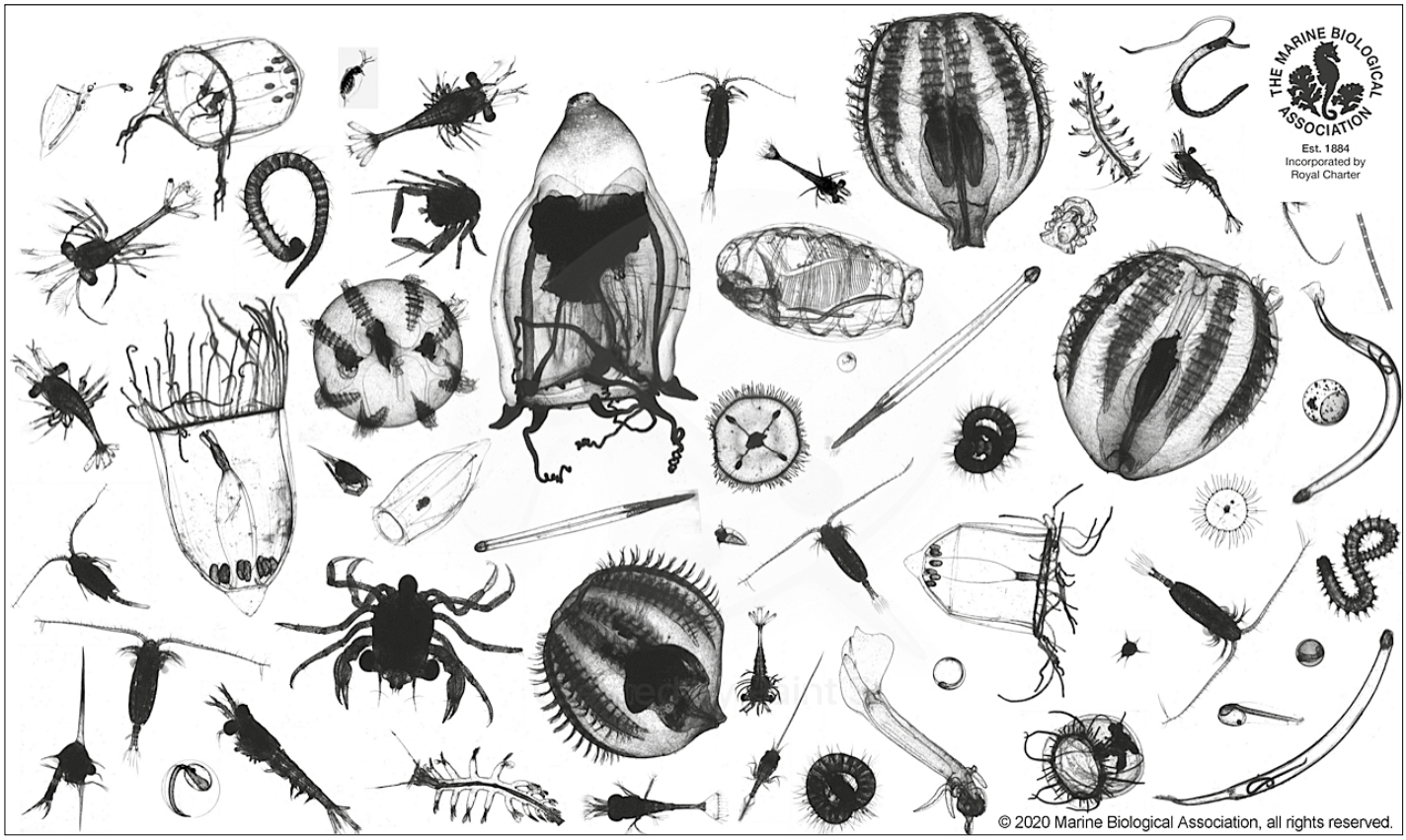
Challenges and recommendations

The transition of marine biological research into a big data driven discipline and the generation of regionally and globally integrated data products for marine biological observations, and wider applications, will first require the creation of truly big data sets. Some marine biologists are already dealing with large data sets e.g. for image processing. However, at present most mainstream marine biologists are not yet facing problems with data management and processing that require e-infrastructures. Dealing with complex data is a more prominent challenge for marine biologists than dealing with large data sets.

There are many advantages of applying big data approaches to biological data, such as ease of scaling-up acquisition of machine-generated data, and reduced human biases in data collection and analysis. However, there are also considerable challenges such as lack of standardized data management and archiving practices, complexities of data processing and incomplete provenance trails preventing reproducibility of their biological interpretation. Another challenge for this community is the lack of taxonomic expertise to ground-truth machine learning algorithms and their results.

To transition marine biological research into a big data driven discipline and to improve marine biological observations we recommend to:

- Store and curate biological data in standardized data formats such as the DarwinCore[94] schema with metadata captured with Ecological Metadata Language (EML[95]);
- Establish a sustainable, globally connected network of long-term biological observatories building on existing biological research infrastructures and scientific networks;
- Stimulate new initiatives from international programs that promote a culture for open science and build relationships of trust among researchers, and can support European data initiatives, such as e.g. the Marine Biodiversity Observatory Network (MBON[96]) from the Group on Earth Observation Biodiversity Observation Network (GEO BON[97]) and the Genomic Standards Consortium (GSC[98]);
- Increase the technical and semantic interoperability of existing marine data infrastructures;
- Increase engagement of the marine science community with the European Open Science Cloud[99] (EOSC) to allow more large-scale, interdisciplinary analyses and societally relevant data products by exploring more big data use cases; and
- Scrutinize the veracity of new biological data sources including imagery, hydro-acoustics, and genetic sequences and train experts in taxonomy to ensure high-quality data feeding into big data applications.



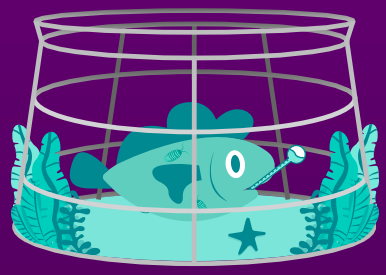
Credit: Marine Biological Association

Plankton community images taken using a CPR Flowcam at the Marine Biological Association.

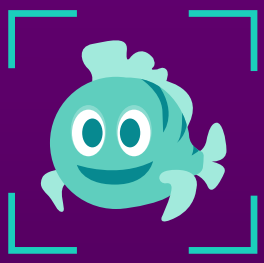
[94] https://obis.org/manual/dataformat/
[95] https://eml.ecoinformatics.org/
[96] https://marinebon.org/
[97] https://geobon.org/
[98] https://gensc.org/
[99] https://ec.europa.eu/digital-single-market/en/european-open-science-cloud

5 Food provision from seas and the ocean

Fisheries and aquaculture are fast growing food sectors that need to be managed sustainably to minimize environmental impacts and to meet both the sustainable development goals (SDGs) on zero hunger (SDG2) and life below water (SDG14). The potential of a big data approach to benefit sustainable seafood production is already being realized in a range of applications in both aquaculture and wild-capture fisheries.



Using shared data from aquaculture farms, artificial intelligence can be used to predict the location and timing of sea-lice outbreaks so they can be treated.



Artificial intelligence can be used to develop facial recognition for fish.

Automated surveillance systems in rivers could be used to sort wild salmon and escaped, farmed salmon using facial recognition.

The most well-known of these is Global Fishing Watch[100] which uses on-board automatic identification systems (AIS) on fishing vessels to track fishing activity in real-time to monitor illegal, unregulated and unreported (IUU) fishing. Research applications of Global Fishing Watch data include investigating the spatial overlap between fishing effort and species of conservation interest, such as sharks, and the detection of global transshipment of catch from one vessel to another. Another good example is the National Oceanographic and Atmospheric Administration's (NOAA) online tool EcoCast[101], which generates daily forecasts of the spatial distribution of migratory species, including sea turtles and blue sharks off the west coast of the United States. Commercial fishing vessels can consult EcoCast maps to identify areas that should be avoided to reduce bycatch[102]. In Europe, automated image analysis is being developed as a reliable and economical alternative to on-board fisheries observers for monitoring catch-at-sea (James *et al.,* 2019). The big data approach also holds considerable promise for the avoidance of protected species, species without any value and nuisance species in commercial fisheries catch. A cloud-based Bycatch Avoidance Tool is currently facilitating data sharing within a Scottish demersal fishery for mapping discards (BATmap)[103]. This data collation and dissemination capability is the first step towards forecasting the risk of bycatch, similar the EcoCast tool.

Aquaculture is one of the fastest-growing food sectors, supplying approximately 50% of fish eaten worldwide[104]. The 2019 EU Fish Market report[105] indicates that EU aquaculture production in 2017 was similar to that of Norway, and that Norway is the main supplier of fish and seafood to the EU. Norway is home to the world's largest salmon farming industry, and salmon is the main aquaculture species imported into the EU[106]. In 2018, Norway produced almost 1.3 million tonnes of salmon[107]. Salmon in fish farms now outnumber wild salmon in Norway by a thousand to one, and a single salmon pen can contain up to 200,000 fish. While indisputably a commercial success, the growth of salmon farming is not without biological and environmental consequences. Among these concerns are pollutants (including organic matter, pharmaceuticals, and other chemicals), the spread of sea-lice and other diseases that can spread

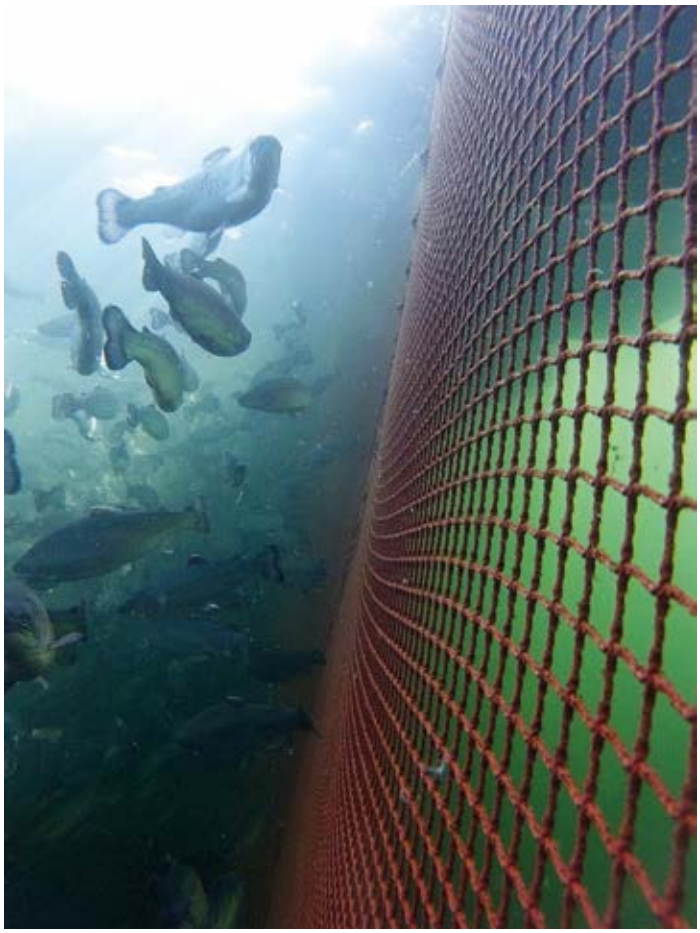
[100] https://globalfishingwatch.org/
[101] https://coastwatch.pfeg.noaa.gov/ecocast/about.html
[102] https://coastwatch.pfeg.noaa.gov/ecocast/
[103] https://batmap.co.uk/rtr/
[104] https://ec.europa.eu/fisheries/cfp/aquaculture_en
[105] https://www.eumofa.eu/market-analysis
[106] https://www.eumofa.eu/documents/20178/157549/EN_The+EU+fish+market_2019.pdf
[107] https://www.ssb.no/en/fiskeoppdrett/

to wild populations, and the risk of salmon escaping from farms and breeding with wild populations thereby reducing their fitness. Big data can be used to mitigate these concerns and in the near future advanced data systems, sensors, and camera-based technology will be used to manage the entire aquaculture value chain from hatchery to consumer. The case studies in this section demonstrate current and future uses of big data for the management of sea-lice outbreaks and escaped salmon populations.

The AquaCloud Platform case study: predicting and managing sea-lice outbreaks

Sea-lice are external parasites that kill juvenile salmon and reduce disease resistance in both juveniles and adults. Infection rates increased as global salmon farming expanded throughout the 1980s and 1990s. The use of biopesticides provided a short-term solution but their efficacy declined due to increasing resistance. Currently, sea-lice pose a significant challenge to the growth of the global salmon farming industry. The scale of the sea-lice problem has spurred research into effective mitigation responses globally (Jackson *et al.,* 2018). Big data are becoming a part of the industry-led solutions for combating sea-lice by taking advantage of the wealth of environmental and production-level data being collected in real-time.

AquaCloud[108] is a digital platform that gathers data from salmon farmers across Norway with the long-term ambition to use machine learning to predict the timing and location of sea-lice outbreaks. This would allow salmon farms to implement more effective prevention measures and treatments. AquaCloud, launched in April 2017 by the Norwegian Centre of Excellence Seafood Innovation Cluster[109], is hosted by the IBM Watson Data Platform[110]. The platform stores cage-level data provided routinely by salmon farming industry partners from 311 locations (2,657 cages; numbers from 2018) including data on the volume of salmon in cages, environmental information, feeding data, mortalities, and treatments. Each farmer has exclusive access to their own data and a say in whether or not they share non-sensitive data. Additional data are available to AquaCloud from BarentsWatch[111], an online portal that aggregates data about Norwegian coastal and marine areas. The BarentsWatch Fish Health module includes monthly data for biomass and number of salmon from all fish farms in Norway. Physical oceanographic data, including water depth, turbidity, salinity, oxygen, temperature, and pH levels, available through online portals such as Copernicus Marine Environmental Monitoring Service (CMEMS)[112] are used for short-term forecasts of sea-lice outbreaks based on physical circulation models. Predictive models of sea-lice abundance in space and time have also been developed by aquaculture researchers (Myksvoll *et al.,* 2018) and may be useful to integrate into future operational forecasting.



Credit: Björgólfur Hávarðsson

Farmed salmon in a salmon pen.

The initial goal of AquaCloud was to develop machine learning algorithms for forecasting sea-lice outbreaks one, two and three weeks in advance. In collaboration with IBM, AquaCloud used a big data approach to forecasting sea-lice infestations, successfully predicting 70% of temporal variation in sea-lice outbreaks[113]. However, this exercise indicated that the underlying data quality limited the forecasting capability of the model and that the data provided by different farms needed to be of a more consistent standard. This insight highlights the need for convergence on ocean best practices throughout the ocean data value chain. Currently AquaCloud is building an open platform for sensor technology, which requires establishing industry standards and best practices for sensors and sensor deployment, etc.

[108] https://www.aquacloud.ai/
[109] http://www.seafoodinnovation.no/
[110] https://dataplatform.cloud.ibm.com/
[111] https://www.barentswatch.no/en/
[112] http://marine.copernicus.eu/
[113] https://www.ibm.com/blogs/cloud-computing/2018/09/17/data-science-norway-fish-farmers/

Challenges and recommendations

A key challenge that needs to be overcome to enable more big data applications in aquaculture is sharing business-critical data among commercial competitors. AquaCloud has shown real innovation in this regard. The incentive for competing aquaculture companies to share data is the common goal of mitigating impacts of sea-lice on salmon production, which imposes significant costs for the industry. Increasing the data quality feeding the predictive models is also a key challenge.

To scale-up the big data approach for managing sea-lice outbreaks we recommend to:

- Develop smart sensors e.g. camera-based sea-lice counters, automated fish welfare monitoring systems and improved automated environmental monitoring systems to increase the temporal resolution of the biological and environmental data thereby allowing improvement in forecasting algorithms;
- Improve connectivity of sensors and data transfer for better monitoring data;
- Use data standards and best practices based on FAIR principles across the whole ocean data value chain;
- Involve digital firms (such as IBM) for cross-industry collaboration and cross-fertilization of technology and expertize;
- Develop effective collaborations across government, industry, universities and the digital sector to deliver real-time operational data analytics for forecasting sea-lice outbreaks; and
- Develop a viable and sustainable business model to maintain and scale-up monitoring networks.

Credit: Björgólfur Hávarðsson

Salmon farm in Norway.

[114] https://www.vitenskapsradet.no/Portals/vitenskapsradet/Pdf/Status%20of%20wild%20Atlantic%20salmon%20in%20Norway%202018.pdf

'Facial' recognition case study: automated surveillance of farmed salmon escapees

Over the last 10 years, average salmon escapees from aquaculture farms in Norway amount to 183,500 annually[114]. Large numbers of escapees pose a risk to wild salmon populations due to cross-breeding, which erodes their genetic diversity and fitness, ultimately leading to reduced wild stock sizes and numbers of salmon available for fisheries. Efficient methods for monitoring and management of escaped salmon are required to reduce these negative impacts.

Like wild salmon, escaped farmed salmon migrate upriver to spawn, and rivers are monitored manually by having multiple people on site to capture, quantify, sort and eliminate them. Escaped salmon are easily identified because they tend to be better fed and are therefore larger and often also have damaged fins from brushing against pens. This monitoring method is costly and labour intensive, and even with significant resources allocated, it is only practical to monitor a few rivers. Monitoring also partly depends on data collected by recreational fishers, the timing and location of which is based on wild salmon abundance. However, escaped farmed salmon do not only occur during the fishing season. The sparsity and bias in sampling based on the timing of the recreational fishing season contribute to statistical inaccuracies and failure to detect many instances of salmon escapees.

The most complete and comprehensive data on escaped salmon have come from the Etne River in western Norway. In 2013, a one-way barrier known as a resistance-board weir trap was placed across the river, forcing all salmon migrating upstream to enter a collection cage. Once in the trap, the salmon are inspected, and the escapees registered and killed. The Etne River trap allows the continuous monitoring and collection of unique information about both wild salmon and escapees including size of individuals and their populations, timing of migration upriver, health status and sexual maturity. The Etne River trap is expensive to maintain because it requires frequent manual intervention to sort the escapees from the wild salmon. In addition, the sorting procedure is invasive for the fish as it involves temporarily trapping and handling.

FishFace is a 'facial recognition' application being developed using a machine learning algorithm to automatically identify fish species based on photographs that can be used as a 'photobooth' for accurate reporting in wild-catch fisheries[115]. Machine learning and 'facial recognition' could also be used to identify escaped salmon and to provide comprehensive automated monitoring systems of salmon populations in a non-invasive and cost-effective way. Salmon display unique patterns of pigmented spots and an algorithm can identify salmon individuals through automated image analysis (Stien *et al.,* 2017). Farmed salmon have more spots than wild salmon (Jørgensen *et al.,* 2018) and deep-learning techniques could be developed to automatically differentiate the two.

Rugged camera equipment, or 'action cams', are routinely used by scientists as they are highly cost effective (e.g. under €500) and there are associated computing and communications equipment with high-resolution imagery capabilities. An automated video surveillance system of the Etne River trap could be constructed at a much lower cost than the current surveillance system, including cameras, computing infrastructure, and data transfer, storage, and analysis. Using the proposed algorithms, current counting and sorting methods could be expanded to continuous monitoring at multiple sites in the river, and to an increase in the number of monitored rivers. Comprehensive, automated monitoring of rivers would significantly change our ability to monitor and manage the salmon farming industry and river ecosystems in general. 'Facial recognition' technology could drive the detection, identification, reporting and subsequent elimination of escaped farmed salmon. Correlating escapees with reports of escape from farms would allow verification of the fidelity of the reporting process, and to chart the behaviour of escaped salmon, helping us to implement more effective management measures. If this system was successful for a valuable resource such as salmon, the proposed surveillance system could also collect information about other aspects of the river e.g. other species of ecological and scientific interest.

Credit: Björgolfur Hávarðsson

Facial recognition and machine learning can be used to recognize individual salmon.

[115] https://blog.nature.org/science/2016/10/17/put-face-vanishing-fish-fishface-fisheries-science-technology-overfishing-data/

Challenges and recommendations

Big data and machine learning could be used for large-scale data-driven management of escaped farmed salmon. Machine learning for automated analysis offers reduced costs and improved efficiency of the surveillance system. To move towards this vision, challenges of data transfer, storage and processing from a few thousand cameras would need to be addressed.

To implement the proposed big data-driven management of escaped, farmed salmon we recommend to:

- Develop automated data collection, storage and processing from many locations using appropriate standards and metadata based on FAIR principles;
- Make data accessible and aggregated with centralized analysis using cloud computing resources;
- Ensure high-quality images for the accuracy of the deep-learning algorithms;
- Use the increasing volumes of structured data to train algorithms and iteratively improve analyses;
- Develop an easy-to-use framework and sequence of pre-trained models for salmon identification, which will require a well-structured repository of algorithms with online tutorials and documentation;
- Integrate data management, cloud computing and machine learning into the aquaculture monitoring and management value chain by engaging key stakeholders; and
- Develop dedicated training programmes for personnel and support for interdisciplinary and cross-disciplinary projects to develop and retain the necessary competences for sustaining the proposed monitoring infrastructure.

Credit: Ø. Poulsen

The Etne River trap, Norway.

6 Recommendations for the future of big data in marine science

This document presents recent advances, challenges and opportunities for big data to support marine science. We highlight that the marine science community has not yet reached the big data revolution. To develop solutions to key societal challenges, there is an increasing need for more complex analyses across traditionally siloed disciplines and sectors. To achieve these goals we need to move towards increased digitalization and the adoption of big data in marine science. We have identified overarching challenges and recommendations within the categories of data acquisition, data handling and management, service interoperability, computing infrastructures and data accessibility, data sharing, big data analytics, training networks and collaboration, which we elaborate on below.

Data acquisition

Marine data will increasingly be collected by machines, rather than humans. This will make marine science a big data driven discipline. Monitoring and observations of the marine environment are rapidly increasing due to an expanded scope from traditional physical and biogeochemical observations to a more interdisciplinary approach with the inclusion of genes, individuals, populations, communities, ecosystems and the biosphere. Physical and biogeochemical ocean data are collected automatically via sensors, autonomous platforms (e.g. moorings, gliders, surface drifting buoys, coordinated Argo profiling floats) and remote sensing, etc. A key challenge is to also enable the automated collection of biological data, i.e. genomics, imagery, and acoustics. A further challenge is the transmission ashore of the deluge of data originating from marine monitoring and observations.

To enhance data acquisition we recommend to:

- Support continued development of automated smart ocean sensors to work towards creating an *Ocean Internet of Things*, particularly for biological parameters (e.g. automated optical sensors), which are typically less well suited for automation. This will enable the collection of truly big data sets for which more complex analyses can be carried out;
- Develop and support infrastructure and training programs for autonomous and automated data collection and downstream processing and archiving;
- Support advances in ocean observation technologies towards real-time, or near real-time, data transmission through improved satellite and high-bandwidth connection systems, such as 'Edge AI' where algorithms are run locally on hardware devices where the data are collected; and
- Ensure adequate, long-term funding for continuing collection of accurate *in situ* ocean observations and their synthesis for essential ocean predictions. This includes collection of physical, biogeochemical and biological observations. We also recommend increased partnerships with and engagement by marine industries, e.g. those operating commercial ships hosting scientists, their instruments and sensors.

Data handling and management

The handling and management of large volumes of heterogeneous data generated from different observation-based infrastructures is a key challenge. Many parameters are measured using different observing platforms and sensors, and there are unique storage and archiving needs for raw data of complex origins.
Users of marine data infrastructures require high-quality data and clear provenance. This is driving the infrastructures to improve FAIRness of data and data products in anticipation of the deluge of big data and the increasing use of artificial intelligence and cloud-based open science.

Credit: AUV Team GEOMAR (CC BY 4.0)

Smart ocean sensors will be important to transition marine science to a big data driven discipline.

To enhance data handling and management we recommend to:

- Adopt community standards and well-designed data management plans for handling and documenting collected data and data processing steps to contribute to long-term data preservation and accessibility. This will ultimately allow data to be FAIR[116] and therefore machine-readable. It will allow heterogeneous data to be shared and integrated from various sources through successful interoperability between data services. More complex analyses can then be performed on big data sets thereby increasing the value of data;
- Ensure transparency in data management, with a complete provenance trail throughout the data lifecycle. Data management plans should be defined by the goals of a research project or observation program and data should also be documented through metadata;
- Include feedback loops in data management plans to converge towards best practices and standards to facilitate the sharing of information by providing identifiers, labels, and controlled vocabulary. The convergence towards community data standards and best practices is needed e.g. the development of ocean best practices for ocean observations. Cooperation between local and global scales is critical for the development and promotion of standards;
- Design and implement data management plans in collaboration with marine data management infrastructures who provide standards and operate tools for submitting data and metadata to link data with upstream protocols (e.g. field/lab protocols) and downstream protocols (e.g. analytical algorithms) to facilitate data archiving;
- Promote widespread adoption of FAIR data principles by data collectors through incentives. Data infrastructures should increasingly adopt DOIs so that data are citable, thereby encouraging data originators to share data with data infrastructures; and
- Further develop and expand the capabilities of marine data infrastructures so they are ready to host large volumes of heterogeneous data originating from different sources and formats i.e. a combination of structured and unstructured data varying in size and complexity.

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

Data service interoperability, computing infrastructures, and data accessibility

The integration and exchange of heterogeneous data sources is key for realizing big data applications for marine science. Infrastructures such as SeaDataNet, EMODnet, EurOBIS and CMEMS (see Box 4) represent very important building blocks for a unified European marine data infrastructure. However, many data originators and funding agencies are not yet making use of existing marine data management infrastructures for sharing data with a larger community of users. A lack of awareness and incentives to share data may be contributing factors. Progress is being made to provide cyber platforms with integrated access to data from multiple sources and high computational capacity for the marine science community and wider research communities, particularly within the framework of EOSC[117] and European initiatives such as ENVRI-FAIR, Blue-Cloud, LifeWatch and Copernicus. These initiatives feature developments for Virtual Research Environments (VREs), which are mostly in pilot stages for marine applications and are promising for wider uptake by marine research communities as steps towards exploring the opportunities offered by EOSC. However, the majority of the marine science community are not yet aware of, and do not use, VREs or other cloud computing services.

To increase interoperability and data accessibility we recommend to:

- Incentivize and promote the use of existing European marine data management infrastructures by the marine science community to increase data accessibility, repurposing, and subsequent use in big data applications;
- Further develop the interoperability of European marine data management infrastructures for handling and exchanging high variety, multidisciplinary data;
- Expand and upgrade services of European marine data management infrastructures in cooperation with e-infrastructures to include more computational infrastructure for cloud computing, data storage and access to big data analytical tools;
- Encourage cross-disciplinary fertilization of technologies between more advanced multimedia sectors and digital sectors with marine science to scale-up cloud computing initiatives for wider transdisciplinary applications;
- Promote increased participation of the marine community in the development and operation of EOSC and its services. This will require raising awareness, highlighting the benefits and encouraging marine scientists to identify use-cases that can benefit from cloud computing infrastructures through open EOSC calls. This will stimulate the creation and deployment of customized VREs that can guide marine scientists to make optimal use of new cyber opportunities;
- Ensure interoperability of VREs to promote interdisciplinary collaboration and accelerate innovation, and provide resources to ensure their long-term sustainability; and
- Further develop cooperation, interoperability and exchange of data and services including computational platforms, between European data infrastructures and international counterparts to facilitate common access to data on wider sea-basin and global scales. This will allow progress towards a 'digital ocean twin' that aligns with objectives of the UN Decade of Ocean Science for Sustainable Development[118].

Data sharing

The importance of sharing data is becoming increasingly recognized as data should not only be understood by individual scientists or research teams, but also need to be open and transparent to the world. However, business-critical data are often confidential, e.g. sharing bycatch data or sea-lice data for fish farming.

To enhance data sharing we recommend to:

- Identify new incentives for data sharing between scientists, industry, and governments to create a sense of community and trust in the provenance of data generated. This could be in the form of social networks or data impact factors;
- Promote and incentivize the widespread use of existing protocols and data management infrastructures for data stewardship and sharing within the marine science community; and
- Develop protocols to help recognize types of data that should be shared (e.g. those with no immediate economic value) and that do not need to be shared. Data collected by industry can be commercially sensitive and therefore policies are needed to incentivize data sharing while at the same time meeting industry requirements for confidentiality.

[117] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[118] https://www.oceandecade.org/
[119] https://www.ices.dk/community/groups/Pages/WGMLEARN.aspx

Big data analytics and data validation

The production of increasingly large and complex data sets in marine science requires machine learning to process and identify emergent patterns, and to improve existing models. As the use of big data analytics becomes increasingly widespread there is also a need to be cautious of the potential introduction of error and biases, since models and algorithms are only as good as the data used to feed them and poor data quality is a significant challenge.

To increase big data analytics and data validation we recommend to:

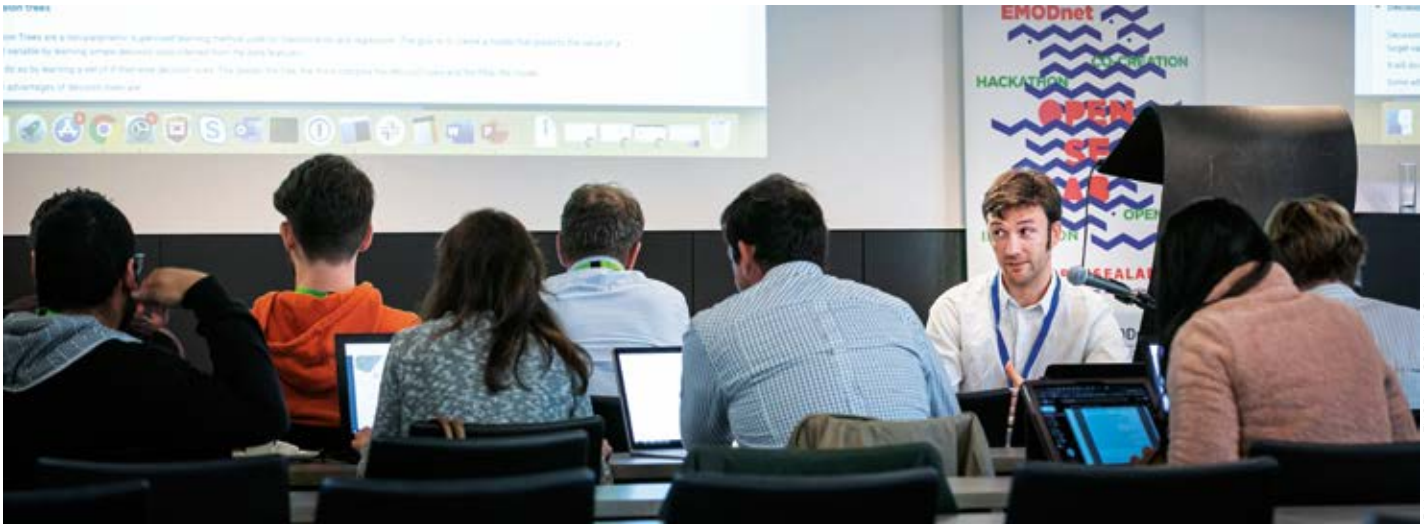
- Develop close collaboration between data scientists and marine scientists to define the limits of big data analytics e.g. via networks proposed by the ICES Working Group on Machine Learning in Marine Science[119];
- Maintain quality assurance of data submitted to marine data management infrastructures and for data that feeds into big data analyses. As the use of big data analytics increases in marine science, humans will play an increasingly important role in generating well-curated data sets. Well-trained data curators are needed to generate data sets for ground truthing and to validate results from big data analyses;
- Develop standardized models and algorithms and use DOIs to incentivize sharing; and
- Develop well-curated, community-maintained data sets that can be used for training and calibration of machine learning algorithms.

Training and collaboration

Big data is a dynamic field, where analytical tools are constantly evolving. This leads to new challenges as these tools are largely unfamiliar to most marine scientists. Many marine scientists do not have in-depth training in data science and programming and may be overwhelmed by the myriad and sophistication of big data methods available as well as the requirements for developing effective data management plans. Decisions on which analytical method is the most appropriate to tackle a certain problem can be difficult and results from machine learning methods are difficult to interpret.

To enhance collaborations and training for big data we recommend to:

- Develop specialized training to empower marine scientists to adopt the use of machine learning in their work. This should encompass lifelong learning as well as training for the next generation of young marine scientists in advanced data science, big data analytics and programming;
- Encourage active collaborations with data scientists, statisticians, computer scientists, and data managers, particularly for early-career scientists. This could be in the form of working groups, multidisciplinary teams and the involvement of data scientists in designing marine research to allow early identification of data, computational and analytical needs. We recommend aligned efforts across these communities to avoid duplication and reduce overheads at organizational and national levels;
- Provide training on the use of VREs for marine scientists. This will foster more strategic partnerships between marine science and the European computer science and data science communities;
- Promote the training of data curators as well as improved training for scientists on the design and implementation of data management plans; and
- Establish and consolidate regional and global marine scientific networks to strengthen the development, training, and communication of big data in marine science.



Dirk Leermans, EMODnet CC BY-NC-SA 4.0

Specialized training programmes are needed to empower marine scientists to use machine learning in their work.

[119] https://www.ices.dk/community/groups/Pages/WGMLEARN.aspx

Acknowledgements

Dorothee Bakker thanks the EU Horizon2020 INFRADEV RINGO project (730944) for enabling her work. The many researchers and funding agencies responsible for data collection and quality control are thanked for their contributions to the Surface Ocean CO_2 Atlas (SOCAT) and the Global Data Analysis Project (GLODAP). Tara Marshall thanks the Marine Alliance for Science and Technology for Scotland (MASTS) for their funding support. Matthias Obst acknowledges supported by Nordic e-Infrastructure Collaboration (NeIC), the Swedish Research Council (LifeWatch program, grant N°. 2017-00634), the Ocean Data Factory program funded by the Swedish Innovation Agency (Grant No. 2019-02256) and the Swedish Agency for Marine and Water Management (Grant N°. 956-19). Jerry Tjiputra thanks funding from the Research Council of Norway (275268). Edward Curry thanks funding from Science Foundation Ireland grant SFI/12/RC/2289_P2 and the European Union's Horizon 2020 research programme Big Data Value ecosystem (BDVe) grant N° 732630.

References

Abarenkov, K., Tedersoo, L., Nilsson, R. H., Vellak, K., Saar, I., Veldre, V., ... Kõljalg, U. (2010). PlutoF—a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. *Evolutionary Bioinformatics Online, 6,* 189–196. https://doi.org/10.4137/EBO.S6271

Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., Brien, K. M. O., Olsen, A., ... Jones, S. D. (2016). A multi-decade record of high-quality f CO 2 data in version 3 of the Surface Ocean CO_2 Atlas (SOCAT). *Earth System Science Data,* 383–413. https://doi.org/10.5194/essd-8-383-2016

Bargain, A., Foglini, F., Pairaud, I., Bonaldo, D., Carniel, S., Angeletti, L., ... Fabri, M. C. (2018). Predictive habitat modeling in two Mediterranean canyons including hydrodynamic variables. *Progress in Oceanography, 169,* 151–168. https://doi.org/10.1016/j.pocean.2018.02.015

Barisits, M., Beermann, T., Berghaus, F., Bockelman, B., Bogado, J., Cameron, D., ... Wegner, T. (2019). Rucio: Scientific Data Management. *Computing and Software for Big Science,* 3(1), 11. https://doi.org/10.1007/s41781-019-0026-3

Benedetti-Cecchi, L., Crowe, T. P., Boehme, L., Boero, F., Christensen, A., Gremare, A., ... Zingone, A. (2018). Strengthening Europe's Capability in Biological Ocean Observations. In A. M. Piniella, P. Kellett, K. Larkin, & S. J. J. Heymans (Eds.), *EMB Future Science Brief 3*. Retrieved from http://www.marineboard.eu/publication/strengthening-europes-capability-biological-ocean-observations-future-science-brief

Bittig, H. C., Steinhoff, T., Claustre, H., Fiedler, B., & Williams, N. L. (2018). An Alternative to Static Climatologies: Robust Estimation of Open Ocean CO 2 Variables and Nutrient Concentrations From T , S , and O 2 Data Using Bayesian Neural Networks. *Frontiers in Marine Science*, (September), 1–29. https://doi.org/10.3389/fmars.2018.00328

Buttigieg, P. L., Janssen, F., Macklin, J., & Pitz, K. (2019). The Global Omics Observatory Network: Shaping standards for long-term molecular observation. *Biodiversity Information Science and Standards, 3*. https://doi.org/10.3897/biss.3.36712

Costello, M. J., Claus, S., Dekeyzer, S., Vandepitte, L., Tuama, É., Lear, D., & Tyler-Walters, H. (2015). Biological and ecological traits of marine species. *PeerJ*, (8), 1–29. https://doi.org/10.7717/peerj.1201

Curry, E. (2016). The Big Data Value Chain : Definitions , Concepts , and Theoretical Approaches. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New Horizons for a Data-Driven Economy*. https://doi.org/10.1007/978-3-319-21569-3

D'Onghia, G., Capezzuto, F., Cardone, F., Carlucci, R., Carluccio, A., Chimienti, G., ... Tursi, A. (2003). Macro- and megafauna recorded in the submarine Bari Canyon. *Mediterranean Marine Science, 4,* 57–66. https://doi.org/https://doi.org/10.12681/mms.1082

Davies, N., Field, D., Amaral-Zettler, L., Clark, M. S., Deck, J., Drummond, A., ... Zingone, A. (2014). The founding charter of the Genomic Observatories Network. *GigaScience, 3*(1). https://doi.org/10.1186/2047-217X-3-2

Davies, N., Field, D., Gavaghan, D., Holbrook, S. J., Planes, S., Troyer, M., ... Consortium, I. (2016). Simulating social-ecological systems: the Island Digital Ecosystem Avatars (IDEA) consortium. *GigaScience, 5*(1), 14. https://doi.org/10.1186/s13742-016-0118-5

Deck, J., Gaither, M. R., Ewing, R., Bird, C. E., Davies, N., Meyer, C., ... Crandall, E. D. (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLOS Biology, 15*(8), e2002925. Retrieved from https://doi.org/10.1371/journal.pbio.2002925

European Marine Board. (2019). Navigating the Future V: Marine Science for a Sustainable Future. *Position Paper 24 of the European Marine Board, Ostend, Belgium ISBN: 9789492043757. ISSN: 0167-9309.* https://doi.org/10.5281/zenodo.2809392

Garcia-Silva, A., Gomez-Perez, J. M., Palma, R., Krystek, M., Mantovani, S., Foglini, F., ... Altintas, I. (2019). Enabling FAIR research in Earth Science through research objects. *Future Generation Computer Systems, 98*(March), 550–564. https://doi.org/10.1016/j.future.2019.03.046

Goris, N., Tjiputra, J., Olsen, A., Schwinger, J., Lauvset, S. K., & Jeansson, E. (2018). Constraining Projection-Based Estimates of the Future North Atlantic Carbon Uptake. *J. Climate, 31*(April), 3959–3978. https://doi.org/10.1175/JCLI-D-17-0564.1

Grorud-Colvert, K., Claudet, J., Tissot, B. N., Caselle, J. E., Carr, M. H., Day, J. C., ... Walsh, W. J. (2014). Marine Protected Area Networks: Assessing Whether the Whole Is Greater than the Sum of Its Parts. *PLoS ONE, 9*(8), e102298. https://doi.org/10.1371/journal.pone.0102298

Halpern, B. S., Frazier, M., Afflerbach, J., Lowndes, J. S., Micheli, F., O'Hara, C., ... Selkoe, K. A. (2019). Recent pace of change in human impact on the world's ocean. *Scientific Reports, 9*(1), 11609. https://doi.org/10.1038/s41598-019-47201-9

ICES. (2020). *Workshop on an Ecosystem Based Approach to Fishery Management for the Irish Sea (WKIRISH6; outputs from 2019)* (Vol. 2). https://doi.org/http://doi.org/10.17895.ices.pub.5551

IPCC. (2013). Climate Change 2013: The Physical Science Basis. In T. F. Stocker, D. Qin, G.-K. Plattner, M. M. B. Tignor, S. K. Allen, J. Boschung, ... P. M. Midgley (Eds.), *Climate Change Working Group 1 on the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Retrieved from Cambridge University Press website: https://www.ipcc.ch/report/ar5/wg1/

IPCC. (2019). *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. [H.-O. Pörtner, D. C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, N. M. Weyner (eds.)] https://www.ipcc.ch/site/assets/uploads/sites/3/2019/12/SROCC_FullReport_FINAL.pdf

Jackson, D., Moberg, O., Djupevag, E. M. S., Kane, F., & Hareide, H. (2018). The drivers of sea lice management policies and how best to integrate them into a risk management strategy : An ecosystem approach to sea lice management. *Journal of Fish Disease, 41,* 927–933. https://doi.org/10.1111/jfd.12705

James, K. M., Campbell, N., Viðarsson, J. R., Vilas, C., Plet-Hansen, K. S., Borges, L., ... Ulrich, C. (2019). Tools and Technologies for the Monitoring, Control and Surveillance of Unwanted Catches. In S. S. Uhlmann, C. Ulrich, & S. J. Kennelly (Eds.), *The European Landing Obligation: Reducing Discards in Complex, Multi-Species and Multi-Jurisdictional Fisheries* (pp. 363–382). https://doi.org/10.1007/978-3-030-03308-8_18

Jones, K. R., Klein, C. J., Halpern, B. S., Friedlander, A. M., Possingham, H. P., Watson, J. E. M., ... Grantham, H. (2018). The Location and Protection Status of Earth's Diminishing Marine Wilderness. *Current Biology, 28,* 2506–2512. https://doi.org/10.1016/j.cub.2018.06.010

Jørgensen, K. M., Solberg, M. F., Besnier, F., Thorsen, A., Fjelldal, P. G., Skaala, Ø., ... Glover, K. A. (2018). Judging a salmon by its spots : environmental variation is the primary determinant of spot patterns in Salmo salar. *BMC Ecology, 18,* 1–13. https://doi.org/10.1186/s12898-018-0170-3

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., de Vargas, C., Raes, J., ... Zingone, A. (2011). A holistic approach to marine Eco-systems biology. *PLoS Biology, 9(10),* 7–11. https://doi.org/10.1371/journal.pbio.1001177

Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernández, N., García, E. A., ... Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews, 93*(1), 600–625. https://doi.org/10.1111/brv.12359

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., ... Glöckner, F. O. (2015). The Ocean Sampling Day Consortium. *GigaScience, 27*(4). https://doi.org/10.1186/s13742-015-0066-5

Lacharité, M., Brown, C. J., & Gazzola, V. (2018). Multisource multibeam backscatter data: developing a strategy for the production of benthic habitat maps using semi-automated seafloor classification methods. *Marine Geophysical Research, 39*(1), 307–322. https://doi.org/10.1007/s11001-017-9331-6

Landschützer, P., Gruber, N., Bakker, D. C. E., & Schuster, U. (2014). Recent Variability of the Global Ocean Carbon Sink. *Global Biogeochemical Cycles*, 927–949. https://doi.org/10.1002/2014GB004853.Received

Liu, Y., Correa, J., Lavers, D., Wehner, M., Kunkel, K., & Collins, W. (2016). Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets. *ABDA'16-International Conference on Advances in Big Dta Analytics,* 81–88. https://doi.org/10.475/123

Miloslavich, P., Bax, N. J., Simmons, S. E., Klein, E., Appeltans, W., Aburto-Oropeza, O., ... Shin, Y.-J. (2018). Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. *Global Change Biology*. https://doi.org/10.1111/gcb.14108

Myksvoll, M. S., Sandvik, A. D., Albretsen, J., Asplin, L., Johnsen, A., Karlsen, Ø., ... Skardhamar, J. (2018). Evaluation of a national operational salmon lice monitoring system — From physics to fish. *PLOS ONE, 13*(7), e0201338. https://doi.org/10.1371/journal.pone.0201338

Olsen, A., Lange, N., Key, R. M., Tanhua, T., Álvarez, M., Becker, S., ... Feely, R. A. (2019). GLODAPv2 . 2019 – an update of GLODAPv2. *Earth System Science Data,* (September), 1437–1461. https://doi.org/https://doi.org/10.5194/essd-11-1437-2019

Pearlman, J., Lewis, M. R., Jenkyns, R., Ep, A., Bensi, M., Pearlman, J., ... Chandler, C. (2019). Evolving and Sustaining Ocean Best Practices and Standards for the Next Decade. *Frontiers in Marine Science, 6*(June), 1–19. https://doi.org/10.3389/fmars.2019.00277

Pendleton, L. H., Beyer, H., Estradivari, Grose, S. O., Hoegh-Guldberg, O., Karcher, D. B., ... Blasiak, R. (2019). Disrupting data sharing for a healthier ocean. *ICES Journal of Marine Science, 76*(6), 1415–1423. https://doi.org/10.1093/icesjms/fsz068

Piechaud, N., Hunt, C., PF, C., & NL, F. (2019). Automated identification of benthic epifauna with computer vision. *Marine Ecology Progress Series, 615,* 15–30. Retrieved from https://www.int-res.com/abstracts/meps/v615/p15-30/

Qin, H., Li, X., Yang, Z., & Shang, M. (2016). When underwater imagery analysis meets deep learning: A solution at the age of big visual data. *OCEANS 2015 - MTS/IEEE Washington.* https://doi.org/10.23919/oceans.2015.7404463

Quéré, C. Le, Andrew, R. M., Friedlingstein, P., Sitch, S., Pongratz, J., Manning, A. C., ... Zhu, D. (2018). Global Carbon Budget 2017. *Earth System Science Data, 10*(1), 405–448. https://doi.org/10.5194/essd-10-405-2018

Rödenbeck, C., Bakker, D., N., G., Iida, Y., Jacobson, A. R., Jones, S., ... Zeng, J. (2015). Data-based estimates of the ocean carbon sink variability – first results of the Surface Ocean p CO 2 Mapping intercomparison (SOCOM). *Biogeosciences, 12,* 7251–7278. https://doi.org/10.5194/bg-12-7251-2015

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., ... Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology, 5*(3), 0398–0431. https://doi.org/10.1371/journal.pbio.0050077

Ryabinin, V., Barbière, J., Haugan, P., Kullenberg, G., Smith, N., Mclean, C., ... Rigaud, J. (2019). The UN Decade of Ocean Science for Sustainable Development. *Frontiers in Marine Science, 6*(July 2019). https://doi.org/10.3389/fmars.2019.00470

Serpetti, N. A., Baudron, A. R., Burrows, M. T., Payne, B. L., Helaouët, P., Fernandes, P. G., & Heymans, J. J. (2017). Impact of ocean warming on sustainable fisheries management informs the Ecosystem Approach to Fisheries. *Nature Scientific Reports, 7*. https://doi.org/10.1038/s41598-017-13220-7

Sion, L., Calculli, C., Capezzuto, F., Carlucci, R., Carluccio, A., Cornacchia, L., ... D'Onghia, G. (2019). Does the Bari Canyon (Central Mediterranean) influence the fish distribution and abundance? *Progress in Oceanography, 170,* 81–92. https://doi.org/https://doi.org/10.1016/j.pocean.2018.10.015

Sonnewald, M., Wunsch, C., & Heimback, P. (2019). Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions. *Earth and Space Science*. https://doi.org/10.1029/2018EA000519

Steinhoff, T. et al. (2019). Constraining the Oceanic Uptake and Fluxes of Greenhouse Gases by Building an Ocean Network of Certified Stations : The Ocean Component of the Integrated Carbon Observation System, ICOS-Oceans. *Frontiers in Marine Science, 6*(September), 1–15. https://doi.org/10.3389/fmars.2019.00544

Stien, L. H., Nilsson, J., Bui, S., & Fosseidengen, J. E. (2017). Consistent melanophore spot patterns allow long-term individual recognition of Atlantic salmon Salmo salar. *Journal of Fish Biology,* (91), 1699–1712. https://doi.org/10.1111/jfb.13491

Tantipisanuh, N., Gale, G., & Pollino, C. (2014). Bayesian Networks for Habitat Suitability Modeling: A Potential Tool for Conservation Planning with Scare Resources. *Ecological Applications, 24*(7), 1705–1718. https://doi.org/10.1890/13-1882.1

Wilson, E. O. (1998). Consilience: *The Unity of Knowledge.* ISBN: 978-0679768678

Wlodarczyk-Sielicka, M., & Stateczny, A. (2016). Clustering Bathymetric Data for Electronic Navigational Charts. *Journal of Navigation, 69*(5), 1143–1153. https://doi.org/10.1017/S0373463316000035

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., ... Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology, 29*(5), 415–420. https://doi.org/10.1038/nbt.1823

Zillner, S., Curry, E., Metzger, A., & Auer, S. (2017). *European Big Data Value Strategic Research and Innovation Agenda.* Retrieved from http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf

Abbreviations and Acronyms

ABNJ	Areas Beyond National Jurisdiction
AI	Artificial Intelligence
AIS	Automatic Identification System
ARMS	Artificial Reef Monitoring Structures
AUV	Autonomous Underwater Vehicle
BATmap	Bycatch Avoidance Tool for mapping discards
BOLD	Barcode of Life Data Systems
CMIP	World Climate Research Programme Coupled Model Intercomparison Project
CMEMS	Copernicus Marine Environmental Monitoring System
CO_2	Carbon Dioxide
COP	Conference of the Parties
CTD	Conductivity, Temperature, Depth sensors
DIAS	Data and Information Access Service
DNA	Deoxyribonucleic Acid
DOI	Digital Object Identifiers
DTM	Digital Terrain Model
EBV	Essential Biological Variable
ECV	Essential Climate Variable
eDNA	environmental DNA
EEZ	Exclusive Economic Zone
EGI	European Grid Infrastructure
EMBRC	European Marine Biological Resource Centre
EML	Ecological Metadata Language
EMODnet	European Marine Observation and Data Network
ENA	European Nucleotide Archive
EOSC	European Open Science Cloud
EOV	Essential Ocean Variable
ERIC	European Research Infrastructure Consortium

ESFRI	European Strategy Forum on Research Infrastructures
ESMValTool	Earth System Model Evaluation Tool
EU	European Union
EurOBIS	European Ocean Biogeographic Information System
EuroGOOS	European Global Ocean Observing System
EVER-EST	European Virtual Environment for Research – Earth Science Themes
FAIR	Findable, Accessible, Interoperable, Reusable
FAO	Food and Agricultural Organization
fCO_2	fugacity of Carbon Dioxide
FRA	Fisheries Restricted Area
GBIF	Global Biodiversity Information Facility
GCOS	Global Climate Observing System
GEO	Group on Earth Observations
GEO BON	Group on Earth Observations Biodiversity Observation Network
GLODAP	Global Data Analysis Project
GOOS	Global Ocean Observing System
GSC	Genomic Standards Consortium
H2020	Horizon 2020
IBM	International Business Machines Corporation
ICES	International Council for the Exploration of the Sea
ICOS	Integrated Carbon Observation System
IDEA	Islands Digital Ecosystems Avatars
Ifremer	Institut français de recherche pour l'exploitation de la mer
IHO	International Hydrographic Organization
INSPIRE	Infrastructure for Spatial Information in Europe
IOC	International Oceanographic Commission
IODE	International Oceanographic Data and Information Exchange
IPCC	Intergovernmental Panel on Climate Change
IUCN	International Union for the Conservation of Nature
IUU	Illegal, unregulated and unreported

MBES	Multibeam Echosounder
MBON	Marine Biodiversity Observatory Network
MPA	Marine Protected Area
MRA	Marine Restricted Area
MSFD	Marine Strategy Framework Directive
NeIC	Nordic e-infrastructure Collaboration
NOAA	National Oceanographic and Atmospheric Administration
NODC	National Oceanographic Data Centre
Obs4MIP	Observations for Model Intercomparisons Project
OOI	Ocean Observatories Initiative
pCO_2	Partial Pressure of Carbon Dioxide
RDBMS	Spatial Relational Database Management System
RNA	Ribonucleic Acid
RO HUB	Research Object Hub
ROOS	Regional Operational Oceanographic Systems
ROV	Remotely Operated Vehicle
RSOBIA	Remote Sensing Object-Based Image Analysis
SD	Secure Digital
SDG	Sustainable Development Goal
SOCAT	Surface Ocean CO2 Atlas
SOCIB	Balearic Islands Coastal Observing and Forecasting System
SOCOM	Surface Ocean pCO_2 Mapping Intercomparison
UNFCCC	United Nations Framework Convention on Climate Change
USV	Unmanned Surface Vehicle
VRE	Virtual Research Environment
WebGIS	Web Geographical Information Systems
WMO	World Meteorological Organization
WoRMS	World Register of Marine Species

Glossary

Autonomous Underwater Vehicle – These are unmanned and autonomous vehicles that are deployed from vessels for survey missions at remote distances from the vessel.

Causality – The generation and determination of one phenomenon by another.

Cloud computing – This is where computation resources, such as data storage or computing power, is available on-demand to users without the user needing to actively manage the resources. An example would be a data centre that is available to many users online

Cold Storage – These are lower performing and less expensive storage environments holding data accessed less frequently, no longer in active use or that might not be needed for months, years, decades, or maybe ever.

Digital Terrain Models – These are 3D computer generated representations of a terrain's surface created from a terrain's elevation data.

Drop-cameras – These are high-resolution, standard or a wide view angle waterproof video recording devices.

E-infrastructure – This is an abbreviation for electronic infrastructure and refers to any computational resource, network, software, support etc. that facilitates collaboration between research communities by sharing resources, data and tools

Echosounder – This is a type of sonar used to map the seafloor and can be either single- or multi-beam.

Eco-genomic – This is making links between specific organisms or genes and their environmental context.

Edge AI – This is where AI algorithms are processed locally on hardware where the data are collected. An Edge AI device is able to operate without external connections and can process data independently.

Environmental DNA – This is genetic material obtained directly from environmental samples (e.g. soil, sediment, water).

Essential Fish Habitats – These are areas that are crucial for fish life stages i.e. areas where they spawn, breed, feed and mature.

Essential Biological Variables – These are derived measurements required to study, report and manage biodiversity change.

Essential Climate Variables – These are physical, chemical or biological variables or groups of linked variables that critically contribute to the characterization of Earth' s climate.

Essential Ocean Variables – These are physical, biogeochemical, biological or ecosystem variables that critically contribute to the characterization of the ocean.

Fisheries Restricted Area – These are areas where fishing activities are banned or restricted to protect marine ecosystems.

Fisheries sonar – A sonar is a system that uses sound waves to detect objects underwater, in this case fish.

Georeferencing – This is where data are associated with a location or physical space.

GIS layers – This is a mechanism used to display geographic datasets from a Geographic Information System (GIS) containing groups of point, line or area (polygon) features representing a particular class or type of real-world entities.

Gliders – These are a type of autonomous underwater vehicle that are deployed from vessels for survey missions at remote distances from the vessel. They typically do not have an engine, and instead use changes in buoyancy to move up and down through the water.

Ground-truthing – This is where *in situ* observation data is used to check the accuracy of e.g. remotely sensed data or results from a machine learning algorithm

Hot Storage – These are storage environments with fast and consistent response times to access data right away.

Invasive species – These are organisms that are not native to an ecosystem.

Machine-readable – This refers to data which are in a form that a computer can process.

Marine Protected Area – This is an area designated and effectively managed to protect marine ecosystems, processes, habitats and species, which can contribute to the restoration and replenishment of resources for social, economic and cultural enrichment.

Marine Restricted Area – Designated and effectively managed area in the marine environment in which one or more human activities (such as fisheries, aggregate extraction, water sports) are restricted.

Metadata – This is a supplementary set of data that describes and gives information about other primary data, e.g. the time and location at which the primary data was collected.

Neural Network – These are a set of algorithms, modelled loosely after the human brain, that are designed to recognize patterns.

Ocean Internet of Things – This refers to a network of smart, interconnected underwater objects that enable monitoring of vast, unexplored areas of the ocean.

Passive hydrophones – These are highly sensitive microphones designed to be used underwater for recording underwater sound.

Phenotype – This is a set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.

Protocol – This is a defined procedure or set of rules.

Real-time data – This are data that are delivered immediately after collection.

Reanalysis product – This refers to the use of forecast models and data assimilation systems to 'reanalyse' archived observations, creating global data sets that describe the recent history of the atmosphere, land surface and oceans (consistent and convenient 'maps without gaps').

Research Objects – These contain and describe scientific data, workflows, methods, papers, documents, scientists involved and other important metadata. They support reliability, reproducibility, and interoperability of data and results.

Remotely Operate Vehicle – This is a tethered underwater surveying, monitoring and/or sampling vehicle that is operated from a vessel

Ships of Opportunity – These are commercial or non-commercial vessels that may voluntarily agree to collect data or operate sampling equipment during their normal operations

Smart sensors – These are devices that take information from the physical environment and use built-in computing power to process that information before sending it to an information receiver.

Soundscape – This is a description or perception of an acoustic environment, in this case of the sounds that can exist underwater.

Spatial Relational Database Management System – This is a program that allows the creation, updating, administration, management and to query a database of spatial objects. It is based on the relational model; an intuitive, straightforward way of representing data in tables.

Sub-bottom profiler – This is a type of acoustic seismic survey equipment used to determine physical seabed properties and characterize subsurface geological information.

Systems architecture – This is a conceptual model that describes the structure and behaviour of a system.

Vulnerable Marine Ecosystems – These are areas of the ocean that are considered hotspots of biodiversity and ecosystem functioning, and are also highly vulnerability to disturbances and have a low recovery potential

Video plankton recorder – This is a semi-automated underwater microscope that records images of plankton in real-time. It also collects conductivity, temperature and depth (CTD) data.

WebGIS services – This is an advanced form of Geographic Information System (GIS) available as a web platform.

Workflow – This is a sequence of tasks that processes a set of data.

Annexes

Annex I: Members of the European Marine Board Working Group on Big Data in Marine Science

NAME	INSTITUTION	COUNTRY
Working Group Chairs		
Lionel Guidi	Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), CNRS	France
Antonio Fernàndez Guerra	Max Planck Institute for Marine Microbiology	Germany
Contributing authors		
Dorothee Bakker	University of East Anglia (UEA)	United Kingdom
Carlos Canchaya	University of Vigo	Spain
Edward Curry	National University of Ireland Galway (NUIG)	Ireland
Federica Foglini	Consiglio Nazionale delle Ricerche (CNR)	Italy
Jean-Olivier Irisson	Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), CNRS	France
Ketil Malde	Institute of Marine Research (IMR)	Norway
C. Tara Marshall	University of Aberdeen	United Kingdom
Matthias Obst	University of Gothenburg	Sweden
Rita P. Ribeiro	University of Porto	Portugal
Jerry Tjiputra	Norwegian Research Centre (NORCE)	Norway

Annex 2: External Reviewers

NAME	INSTITUTION	COUNTRY
Bjorn Backeberg	EGI Foundation	The Netherlands
Ghada El Serafy	Deltares	The Netherlands
Frank Muller-Karger	University of South Florida	USA
David Schoeman	University of the Sunshine Coast	Australia
ADDITIONAL CONTRIBUTIONS		
Dick Schaap	MARIS	The Netherlands

European
MARINE BOARD
Advancing Seas & Ocean Science

European Marine Board IVZW
Belgian Enterprise Number: 0650.608.890

Wandelaarkaai 7 I 8400 Ostend I Belgium
Tel.: +32(0)59 34 01 63 I Fax: +32(0)59 34 01 65
E-mail: info@marineboard.eu
www.marineboard.eu