



HAL
open science

OCEANS OF BIG DATA AND ARTIFICIAL INTELLIGENCE

Ramiro Logares, Josep Alos, Ignacio A Catalán, Ana Crespo Solana, F Javier del Campo, Gemma Ercilla, Ronan Fablet, Antonio Fernandez-Guerra, Marti Gali, Josep M Gasol, et al.

► **To cite this version:**

Ramiro Logares, Josep Alos, Ignacio A Catalán, Ana Crespo Solana, F Javier del Campo, et al.. OCEANS OF BIG DATA AND ARTIFICIAL INTELLIGENCE. Oceans. CSIC scientific challenges towards 2030., pp.163-179, 2021. hal-03372264

HAL Id: hal-03372264

<https://imt-atlantique.hal.science/hal-03372264v1>

Submitted on 10 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/353445844>

OCEANS OF BIG DATA AND ARTIFICIAL INTELLIGENCE

Chapter · July 2021

CITATIONS

0

READS

249

27 authors, including:



Ramiro Logares

Institut de Ciències del Mar

229 PUBLICATIONS 5,957 CITATIONS

[SEE PROFILE](#)



Josep Alós

Spanish National Research Council

133 PUBLICATIONS 2,591 CITATIONS

[SEE PROFILE](#)



Ignacio A. Catalán

Mediterranean Institute for Advanced Studies (IMEDEA)

118 PUBLICATIONS 1,404 CITATIONS

[SEE PROFILE](#)



Javier del Campo

Institute of Evolutionary Biology

156 PUBLICATIONS 3,486 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Using underwater cameras as biological sensors: Deep learning in Marine Ecology (DEEP-ECOMAR) [View project](#)



Expedición de Circunnavegación Malaspina 2010: Cambio Global y Exploración de la Biodiversidad del Océano Global [View project](#)

OCEANS OF BIG DATA AND ARTIFICIAL INTELLIGENCE

Coordinators

Ramiro Logares
(ICM, CSIC)

Josep Alós
(IMEDEA, CSIC - UIB)

Researchers & Centers

Ignacio Catalán
(IMEDEA, CSIC - UIB)

Ana Crespo Solana
(IH-CCHS, CSIC)

Javier del Campo
(IBE, CSIC - UPF)

Gemma Ercilla (ICM, CSIC)

Ronan Fablet (IMT Atlantique)

Antonio Fernández-Guerra
(Lundbeck Foundation
Geogenetics Centre)

Martí Galí
(Barcelona Supercomputing Center)

Josep M. Gasol
(ICM, CSIC)

Ángel F. González
(IIM, CSIC)

Emilio Hernández-García
(IFISC, CSIC - UIB)

Cristóbal López
(IFISC, CSIC - UIB)

Ramon Massana (ICM, CSIC)

Lidia Montiel (ICM, CSIC)

Miquel Palmer
(IMEDEA, CSIC - UIB)

Ananda Pascual
(IMEDEA, CSIC - UIB)

Santiago Pascual (IIM, CSIC)

Fernando Pérez (ICM, CSIC)

Marcos Portabella (ICM, CSIC)

José Javier Ramasco
(IFISC, CSIC - UIB)

Daniel Richter
(IBE, CSIC - UPF)

Valentí Sallarés (ICM, CSIC)

Pablo Sánchez (ICM, CSIC)

Javier Sanllehi
(IMEDEA, CSIC - UIB)

Antonio Turiel (ICM, CSIC)

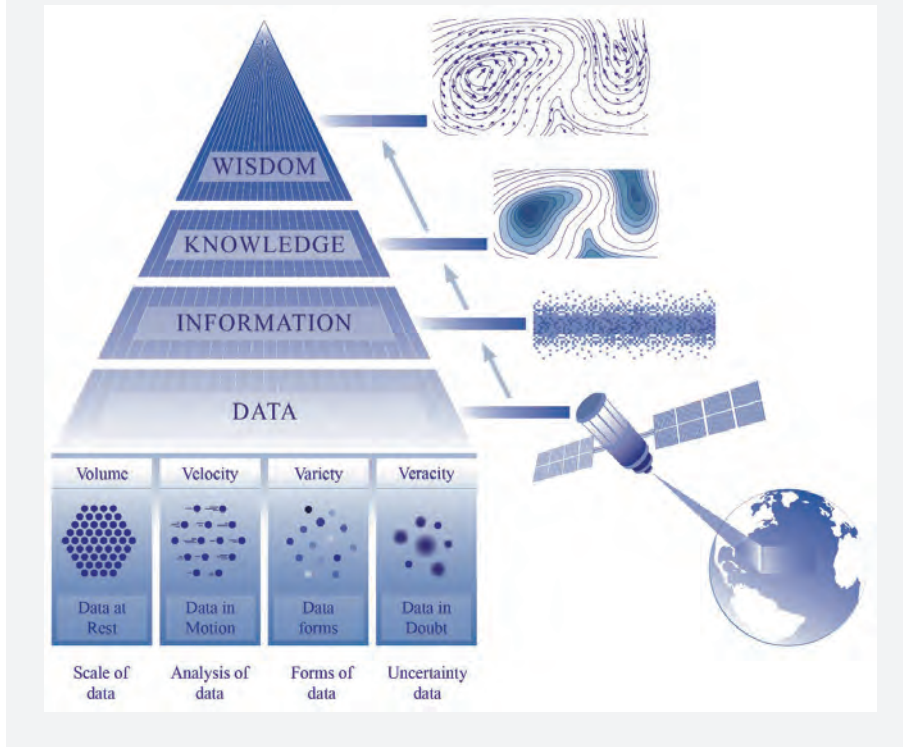
Antonio Villaseñor
(ICM, CSIC)

1. INTRODUCTION AND GENERAL DESCRIPTION

Data are produced when abstracting the world into measures or categories (numbers, characters, images) and are the basis to generate information, knowledge and ultimately wisdom (Kitchin 2014, Fig. 8.1). Data are symbols that represent facts, e.g. temperature records. There is no meaning of data beyond its own existence and can be clean, noisy, structured, unstructured, relevant, or irrelevant. Information can be considered as data that have been processed and that then become useful. In other words, information adds meaning to data. Knowledge can be considered as the application of information and data or the “know-how” that transforms information into instructions. Wisdom is the pinnacle of the knowledge pyramid and refers to being able to apply knowledge (Fig. 8.1).

During the last decades, the capability of humans to generate data has increased exponentially, leading to the so-called Big Data. Even though there is no formal definition of Big Data, it usually is characterised by the **4Vs: Volume, Velocity,**

FIGURE 8.1–The DIKW or Knowledge pyramid, and the characteristics of Big Data applied to ocean science.

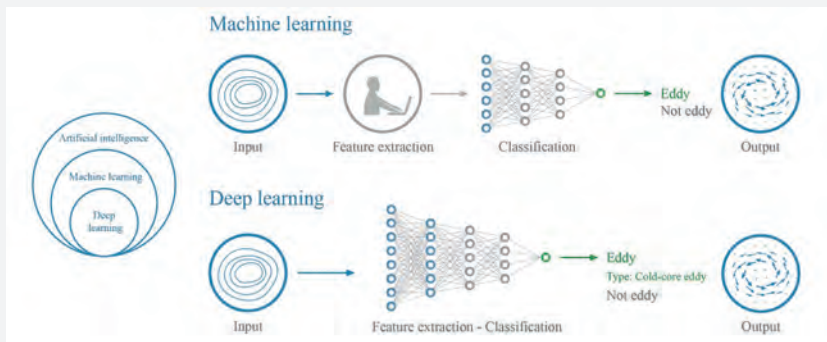


Variety, and Veracity [Fig. 8.1] (Kitchin 2014). Quantifying the **volume** of global data at the moment is not straightforward. According to the International Data Group’s study – “The Digital Universe in 2020” (<https://bit.ly/3b4xggy>), the amount of data in the year 2020 would be ca. 40 trillion gigabytes (or 40 zettabytes). Interestingly, most data has been generated during the last two years and, by 2020, every person was predicted to generate 1.7 Mb per second (<https://bit.ly/3fjEQsH>), or 146,880 GB a day, leading to a production of 165 zettabytes per year by 2025 (<https://bit.ly/3b4xggy>). In particular, ocean sciences have also experienced an explosion of data during the last decade (Brett *et al.* 2020; Guidi *et al.* 2020). Examples are the DNA sequencing of the ocean microbiome, which has produced a few hundred terabytes of raw data since 2010, or the first world’s ocean digital map of seafloor

lithologies based on descriptions of nearly 14,500 samples. Small data differs from Big Data in terms of the **velocity** at which it is generated. Big Data tends to be generated continuously, in many cases, in virtually real-time. For example, satellites continuously stream ocean observation data and weather sensors monitor and transmit weather conditions so their data can be ingested in weather forecasting. Such a continuous stream of data needs continuous management and analysis (Fu *et al.* 2019). In addition, Big Data can display **variety**. That is, it can be a combination of structured, semi-structured or unstructured data, including numbers, text, images, videos and audio, which can be combined. It is widely acknowledged that ca. 80% of Big Data is unstructured. New advances in high-performance computing, database design using Not only Structured Query Language (NoSQL) formats and data mining have allowed to store, manage, process and extract knowledge from unstructured data. Finally, data **veracity** defines, not only how accurate a Big Data set may be, but also how trustworthy the data source, type, and processing is. Removing biases, inconsistencies, duplication, and volatility are just a few accuracy factors of data, which in the context of Big Data becomes a real challenge. Veracity issues in marine sciences arise, for instance, due to the stochastic properties of data-process generation, manual entries, GPS uncertainty, or by model uncertainties in ocean forecasting processes (e.g., hurricanes). Strikingly, most Big Data sets seem to remain unanalysed, with estimates ranging from 97% to 99%. Nevertheless, we need to consider that only a fraction of Big Data may be useful: in 2012, only about 23% of Big Data was considered useful (<https://bit.ly/3b4xggy>). Recently, the Science Brief of the European Marine Board has included the **value** of the data as a new dimension of Big Data in marine sciences. Understanding the costs and benefits of collecting and analyzing data is therefore needed to ensure that its value can be reaped (Guidi *et al.* 2020).

The ocean covers ca. 70% of the surface of the planet and contains ca. 97% of all water on Earth. It plays a central role in regulating the Earth's climate system, and its physical, geological and biological processes play a key role in global biogeochemical cycles (see chapter 2). Due to its importance, a wide array of monitoring efforts have been implemented, including *in-situ* (e.g. gliders, Argo floats, buoys, OBSs (Ocean Bottom Seismometers), Seafloor Observatory Systems) and *ex-situ* (e.g. satellites, drones) sensors covering different spatial and temporal scales (see Chapter 1). Technological advances in sensor technology, autonomous devices and communications allow us to collect Big Data from the ocean in a continuously increasing way. Thus, in

FIGURE 8.2—Artificial Intelligence, Machine Learning and Deep Learning applied to the pattern identification of swirling motion of eddies in the ocean



agreement with the 4 Vs of Big Data, the generated ocean data occupies huge volumes, it is collected continuously in virtually real-time and features variety (i.e. it is unstructured and may consist of images, numbers or DNA sequences). Our capability to generate Big Data from the ocean contrasts with our capacity to analyse them, which has not advanced at the same rate, becoming a bottleneck for the generation of information and knowledge (Malde *et al.* 2019). Recent developments in Artificial Intelligence (AI), in particular Deep Learning (DL) are now allowing processing Big Data and generating new insight (Guidi *et al.* 2020).

AI is broadly defined as “the study of agents that receive percepts from the environment and perform actions” (see White Chapter on Artificial Intelligence). Machine Learning (ML) is a branch of AI (Fig. 8.2) that aims at “iteratively evolve an understanding of a dataset; to automatically learn to recognise complex patterns and construct models that explain and predict such patterns and optimise outcomes”. ML approaches can be supervised (using training data) or unsupervised (using self-organization). Supervised learning involves a model that is trained to match inputs to known outputs, while in unsupervised learning, the model teaches itself to find patterns in the data without the use of training data (Kitchin, 2014). In both cases, a model is generated via a learning process that is modulated by rules and weights. The construction of the model starts simple, and then it evolves into a robust one after changing repeatedly.

Deep Learning (DL) is a subset of ML (Fig. 8.2) that uses multilayer artificial Neural Networks. Traditional ML requires substantial human effort in defining features that represent data, while there is no need to define features in DL, as DL learns the best representation of the data itself in order to produce the most accurate results. DL algorithms require Big Data, and their efficiency improves as more data is added. This contrasts with classic ML approaches that reach a plateau at some point, no matter how much data is added. Another advantage of DL algorithms is that they can represent complex non-linear separating functions, and this is ideal for tasks that require learning complex concepts. Furthermore, feature identification is not required, minimising the chance of human biases. In addition, DL can take advantage of massive parallel processing, as in GPUs, to learn better models.

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

Big Data coupled to AI will revolutionise ocean sciences (Malde *et al.* 2019). Right now, marine sciences are rapidly evolving towards massive data generation from automatic sensors thanks to the increase in computational power and the development of new technologies (Brett *et al.* 2020). High-throughput sequencing, animal and human (e.g. vessel monitoring system) tracking, ocean observing from local stations to satellites, seismic, acoustic, geophysics and sediment data are major examples of how marine sciences are entering into a new Big Data era. How to manage, store, analyse and transform the oceans of Big Data into knowledge is now a fundamental challenge for ocean sciences. This challenge can only be addressed by changing the paradigm in marine sciences, from traditional model-driven representations (e.g. data assimilation in physical and biological models) towards accurate and computationally-efficient data-driven models. AI integration in marine sciences is, without any doubt, the only candidate to bridge this gap. DL is particularly well-positioned to infer data-driven dynamical priors and associated assimilation schemes. However, the integration of AI will need cross-disciplinary expertise at the interface of marine sciences, applied mathematics, and computational sciences to upgrade the current trend of simulation, mapping, forecasting, and assimilation models and technologies towards a novel scientific paradigm bridging the physical, geological and biological paradigms underlying marine sciences and the statistical paradigm on which AI and ML are based. The integration of DL neural networks in marine sciences is in its infancy. However, it will benefit a wide range of, or almost all, oceanographic fields

(actively used so far only in the processing of partial satellite data, animal tracking, classification or measure, and assembly and annotation of high-throughput DNA sequencing data). This novel paradigm will fully benefit from AI-related technological advances to build the next generation of ocean, atmosphere, and climate simulations, mapping, forecasting and reconstruction (assimilation) models. Within biological applications, similar expertise is needed for an effective shift of how field estimates of abundance or species composition are made, as well as for automatization and unsupervised acquisition of long-term time-series of ecological data and processing based on real-time and lagged-time video analysis from underwater sampling devices. Further, for marine conservation (e.g. automatic fish length estimates at the commercial landings, automatic boat detection inside marine reserves), AI-based applications are being developed (Brett *et al.* 2020). Big Data coupled to AI will also revolutionise the field of bioprospecting & blue biotechnology via automated detection of compounds/genes with economic potential (Guidi *et al.* 2020). In addition, early warning systems based on AI and ocean Big Data will likely allow mitigating the effects of ocean events, such as red tides or the outbreak of pathogenic bacteria or viruses. Overall, all fields of life are experiencing a technological revolution due to AI. Oceans, their understanding, function and conservation, have the challenge to incorporate AI in the next decades.

In the smart oceans of the future, it is expected that data that are currently treated independently, such as satellite, genetic, animal tracking, or acoustics, will be jointly analysed using AI by autonomous supercomputers. AI analysis of these massive amounts of data will allow us to discover patterns as well as to provide a detailed and real-time global perspective of the ocean across multiple spatiotemporal scales. Also, AI will help to automatize seafloor mapping and increase the capabilities of the interpreters. Likewise, Big Data analyses together with models and simulations will significantly increase our ability to predict events at multiple scales, from the evolution and spread of a virus and its effects on trophic networks, to the positioning, development and health status of hundreds of millions of fish over time, as well as their relationship with biogeochemical processes or the ocean microbiota. Furthermore, these analyses will generate a renewed holistic insight and understanding of the ocean, being also the base for new conservation policies and applications leading to new products and promoting the Blue Economy in line with EU policies (https://ec.europa.eu/maritimeaffairs/policy/blue_growth_en) and United Nations Sustainable Development Goals (<https://www.undp.org/content/undp/en/home/sustainable-development-goals.html>).

3. KEY CHALLENGING POINTS

Even though AI-based and data-driven frameworks will certainly lead to major breakthroughs in marine science, the state-of-the-art for numerous applications and domains strongly relies on model-driven approaches (e.g., simulation and assimilation frameworks in operational oceanography) using the physical knowledge and associated mathematical representations of geophysical and biological dynamics gained over the past centuries. A major challenge is to bridge the model-driven and AI paradigms to make the most from the current knowledge in physics, biology and geology coupled to the increasing computational efficiency and discovery capability of AI methods as well as the explosion of Big Data. In particular, the advent of Big Data urges researchers for Data Management Plans (DMPs), that will determine, among other things, how data is shared in the scientific community, how it is stored over long periods (e.g. decades) and how it is accessible to the general public, stakeholders and policymakers. With the objective of making research data Findable, Accessible, Interoperable and Re-usable (FAIR), DMPs are key *obligatory* elements for Horizon 2020 EU projects, describing the data management life cycle for the data to be collected, processed and/or generated by marine research projects. Possibly, during the next few years, other funding agencies will require DMPs, such as the Spanish AEI (Agencia Estatal de Investigación).

At this point, we identify three general challenging points that are associated with the generation, management and analysis of ocean's Big Data:

Big Data Generation. Currently, there are a multitude of applications and sensors that are generating Big Data from the ocean in different research areas (Brett *et al.* 2020). For example, satellites, research vessels, buoys, gliders, animal tracking devices (see Chapter 1), AI-processed images, DNA sequencers. Some of these sensors may generate continuous data streams that need to be processed in virtually real-time in order to generate useful outcomes (e.g. ocean forecasting). Other sensors or devices will produce massive amounts of data in a more discrete manner, such as DNA sequencers. As mentioned before, our capabilities to produce Big Data are increasing exponentially, as well as the number of interconnected devices that collect data (the so-called internet of things, IoT). Thus, at the moment and considering the future perspectives, data production *per se* does not seem to be a big challenge. Yet, the challenge is probably related to producing new types of Big Data or datasets that can lead to useful insights. Thus, instead of increasing the data production capacity of current sensors, the development of new sensors or

new data-collection strategies may lead to new types of Big Data. For example, microsensors attached to millions of fish or high-frequency *in-situ* -omics samplers could generate datasets that could provide new knowledge of animal movement or gene function.

Big Data Management. The main challenges are related to data storage, transfer, integration and computing. As for Big Data storage, it must have Petabytes size scale, be highly scalable and flexible due to the need to increase its capacity under demand, and have low latency for real-time access. This storage must be accessible across multiple platforms and systems and be able to handle data from various source systems at the same time. Another aspect that needs to be considered is the Big Data long-term storage and its associated costs: it is becoming evident that storing huge amounts of data over long periods may have substantial costs. Furthermore, Big Data that today has low value could become priceless in the future, therefore, coordinated actions need to be taken in order to reach an agreement on how Big Data will be stored and made available for the next decades. Another specific challenge is related to the real-time transfer and access of Big Data: today, large amounts of data need substantial time to be transferred from one site to another or accessed by different applications, generating a delay in the analyses that could prevent their use in decision-making. Moreover, our capability to efficiently analyse an exponentially increasing amount of Big Data also represents a challenge for the next decades, as the increase in computing power is lagging behind our capabilities to generate Big Data. Thereby, the importance of current research in new computing architectures tailored to the needs of Big Data. Cloud computing technologies can provide suitable scalable solutions (Vance *et al.* 2019), not only to configure *ad-hoc* hardware resources for analysis, but also as data storage. These two characteristics combined may put data and computing in the same place, thus reducing data transfer delays.

Another particular challenge is related to the integration of Big Data from different sources (-omics data, satellites, acoustics, etc.). Currently, datasets from different sources are normally analysed separately (e.g. omics data and satellite observations), thus precluding holistic insights that would emerge from the combined analysis of these datasets. This requires the use of new computational models for the analysis of massive data, such as MapReduce, and new data storage models such as new file systems, NoSQL databases and in-memory Databases.

Big Data Analysis. The massive amounts of unstructured data that are being generated need new methods to analyse them. AI methods, especially neural

networks (DL) are currently the most promising tool for analysing ocean Big Data. At the moment, researchers from different fields are migrating into AI-based analyses and this trend will likely increase dramatically during the next decade. AI-based analyses of Big Data represent a breakthrough in diverse fields, such as marine observatories, early detection systems and image analyses. AI will also be pivotal for new autonomous devices, such as gliders or even ships, where decisions will be taken without human supervision. A fundamental aspect here will be the validation of the decisions taken by the AI, and the potential costs that bad decisions may have.

These three “grand challenges”, that is, Big-Data Generation, Management and Analysis (GMA) are encountered (normally together) in different fields of science. Below, we indicate how Big-Data GMA materialise into three main challenges in marine sciences:

3.1. Observing and understanding the ocean through Big-Data and AI

Remote or *in-situ* ocean observation instruments producing Big Data that is subsequently analysed using AI will likely open a new era in ocean data collection and analysis, contributing substantially to increase our understanding of the ocean at small or large spatiotemporal scales. Some research fields are already transiting through this paradigm change, as is the case of satellite remote sensing. When exploiting remote sensing data, the most usual requirement by end users is to get satellite data interpolated on a high-resolution, gap-free, regular grid. However, the reconstruction of sea surface geophysical fields from partial satellite-derived observations is a challenging, complex task that can be addressed with different strategies. Classical data assimilation is based on simple statistical quantities (e.g. covariance matrix in the case of optimal interpolation) or in the use of an underlying numerical model of the ocean forced with satellite data. Although, the quality, coverage and resolution of ESA's Soil Moisture Ocean Salinity (SMOS)-derived Sea Surface Salinity (SSS) maps and scatterometer-derived stress-equivalent wind products have improved (e.g. Turiel *et al.*, 2008, Fablet *et al.* 2018), new, powerful data assimilations techniques, following the Big Data scheme, have recently emerged, such as the Analog Data assimilation (AnDA) framework, which exploits patch-based analog forecasting operators within a classic Kalman-based data assimilation scheme. AnDa is of particular interest with regards to the upcoming wide-swath surface water and ocean topography (SWOT) mission. Future work will focus on combining these strategies with the AnDA

framework in order to develop useful tools to process real observations from the future SWOT altimetry mission. In this respect, the joint assimilation of SWOT observation gradients and nadir along-track Sea Level Anomalies data should be explored as a possible alternative to deal with the correlated noise sources present in SWOT data. AnDa can be applied to any other oceanographic variable, as Sea Surface Temperature or SSS. Multivariate AnDa is very convenient when multiple variables are assimilated at the same time, although some space-reduction techniques should be applied in order to avoid data scarcity. A different avenue for the applications of Big Data to remote sensing is the use of Random Decision Trees to infer so-far unknown dynamic relations between different variables. This kind of approach has been used for instance, to find relationships between SSS anomalies in particular regions and extreme rainfall over land. Random Decision Trees and similar techniques can be used to group and to validate new physical, chemical and biological processes. In this context, DL models and strategies also arise as promising tools to bridge data-driven and learning-based frameworks to model-driven physical paradigms. This may open new research avenues to embed physical knowledge within data-driven schemes as well as to make the most of state-of-the-art model-driven schemes with the additional flexibility and computational efficiency of learning-based frameworks. The latter may be particularly relevant to address model-data and multimodal synergies.

In the Geosciences the use of ML can be classified into four interconnected categories: automation (e.g. labelling data when the task is difficult or time-consuming for humans), inverse/optimization problems, discovery (extract new patterns, structure, and relationships from data) and forecasting. Despite the availability of large datasets from Earth and Ocean observing systems, often extending over long observation times, many of them remain largely unexplored. Wider adoption by the community of open-science principles such as open source code, open data, and open access would allow taking advantage of the rapid developments that are taking place in ML and AI. Creating an inventory of high-quality datasets, preferably covering large spatial and/or temporal spans that have not been studied using ML and that could immediately benefit from using these approaches (low hanging fruit), represents a sensible course of action. In addition, this field needs to foster collaboration of CSIC groups that have a long history of acquiring large Geoscience datasets with the leading groups in AI/ML research to recognize new potential applications. However, we need to overcome several challenges before working with geoscience datasets. The spatiotemporal structure, the

multi-dimensionality and heterogeneity of the Big Data in geosciences, data noise, incompleteness and error of the data, as well as emerging datasets such as light detection and ranging (LiDAR) point clouds, are among the most relevant challenges.

Biogeosciences lag behind physical oceanography and marine geosciences regarding massive autonomous observation and data collection. The advent of remote sensing of bio-optical variables (chiefly, chlorophyll *a*) in the late 1970s, and its consolidation as an operational technique during the 1990s, represented a major breakthrough in the understanding of upper ocean biogeochemistry and inaugurated the era of Big Data in marine biogeosciences. A similar revolution has occurred since the last decade in the observation of the ocean interior thanks to biogeochemical (bgc-) Argo floats and other autonomous platforms. Fitted with non-invasive chemical and bio-optical sensors and even video cameras, autonomous drifting robots can take measurements of a wide array of variables (chlorophyll and dissolved organic matter fluorescence, particle backscatter, nitrate, oxygen, pH) all year-round between the surface and at least 1,000 m depth at a frequency between 1 and 10 days during several years. The growing swarm of bgc-Argo floats will soon provide a 4D view of variables characterizing the ocean interior biogeochemistry and microbial biomass in near real-time, and efforts are underway to merge this stream of data with remote sensing observations of the upper ocean (e.g. optical satellites, lidar and radar) as well as other *in situ* and *in silico* data streams.

AI techniques are poised to play a key role in the merging of multiscale observations of ocean biogeochemistry, providing end-users with high-quality products including uncertainty estimates, and circumventing the high computational needs of ocean biogeochemistry. Reconstruction of 4D biogeochemical fields from relatively sparse observations using AI will surely yield a leap forward in our predictive capacity, overcoming the limitations of classical climatological approaches based on objective interpolation, which neglected key scales of variability in the temporal (e.g., sub-daily, intraseasonal, interannual) and spatial (e.g., mesoscale) domains, and statistical properties arising from highly nonlinear dynamics. Moreover, AI can be used to infer the underlying processes and to discover unexpected causal links, potentially leading to major advances in process-level understanding and prediction of future system states (Reichstein *et al.*, 2019).

Examples of future applications that will benefit from AI and Big Data include: the accurate estimation of carbonate system and nutrients from hydrological

parameters; improved estimation of the sea-surface distribution and flux of climate-active gases; the fusion of remote and *in situ* bio-optical data to extend high-resolution surface images of microbial plankton and organic carbon stocks to the ocean interior; and the widespread deployment of imaging devices on autonomous platforms (e.g., gliders and Argo floats) to measure the abundance and taxonomy of microplankton as well as severely undersampled metazoans (large zooplankton and micronekton). Some of the main challenges ahead are (1) sustaining and expanding the array of autonomous ocean observation platforms, (2) designing optimized protocols for quality control and data interoperability, (3) ensuring long-term storage and seamless accessibility, (4) merging heterogeneous data sources in formats that make them readily usable across diverse research fields, and (5) moving from purely statistical prediction to process-based models that embody causal relationships.

3.1. Knowing and protecting marine life via Big-Data and AI

In the ocean, the number of microbial genomes and genes have astronomical proportions. It is estimated that 10^{29} prokaryotes, 10^{26} protists and 10^{30} viruses populate the oceans, which may contain 10^{10} prokaryotic lineages alone. Recent estimates indicate that microbes represent two-thirds of the total biomass of marine organisms (Bar-On & Milo, 2019). Addressing this massive gene and taxonomic diversity is now becoming possible thanks to high-throughput DNA sequencers (HTS) (Logares *et al.* 2012), which generate massive amounts of genomics data (TeraBytes per run per machine). Even though we still know only a small fraction of the total diversity of genes and lineages populating the ocean, HTS increased the amount of available genomic data several orders of magnitude during the last 15 years, and given that the sequencing capacity continues increasing, the amount of available data keeps growing. These data need large computing infrastructures to be stored and analysed, and these requirements will increase substantially in the near future. So far, AI has not been widely used for bioinformatics applied to big genomic data, but it is expected that, in the near future, it will become extensively used for applications such as assembly of short or long reads, finding gene homologies, predicting protein function and finding causative links or correlations between changes in a large suite of biotic and abiotic conditions and organismal abundances.

Thus far, DNA (and its actively transcribed gene-coding counterpart, RNA) data have been predominantly used for capturing the genomic information that is present in the ocean in order to understand microbial diversity,

species abundance and metabolic activity, ecological interactions, and also how different lineages have evolved. Yet, during the next 10-20 years, HTS techniques, together with all the acquired knowledge on the ocean metagenome will be used for large scale bioprospecting (blue biotechnology), real-time DNA monitoring (to e.g. detect pathogens that spend part of their life cycle in a free-living form, analyse changes in microbial gene expression or track metazoans via eDNA), as well as laboratory-based or ecosystem-level experiments (e.g. mesocosms or *in situ* ocean work). In addition, an important future challenge will be to integrate DNA data from the ocean with other data types from marine observatories to generate a more comprehensive understanding of the ocean ecosystem. For example, chlorophyll observations from satellite data could be coupled to changes in gene transcription detected by gliders or buoys, that also inform on changes in nutrients and currents. In addition, future genomic observatories aiming at capturing DNA from viruses to metazoans may inform of changes in the architecture of ecological networks and link those to e.g. the appearance of a pathogen or other ecosystem-level disruptions. These genomic observatories will become highly relevant in the context of global change (including ocean warming and acidification; see chapter 4), where the distributions of marine species and genes are expected to respond.

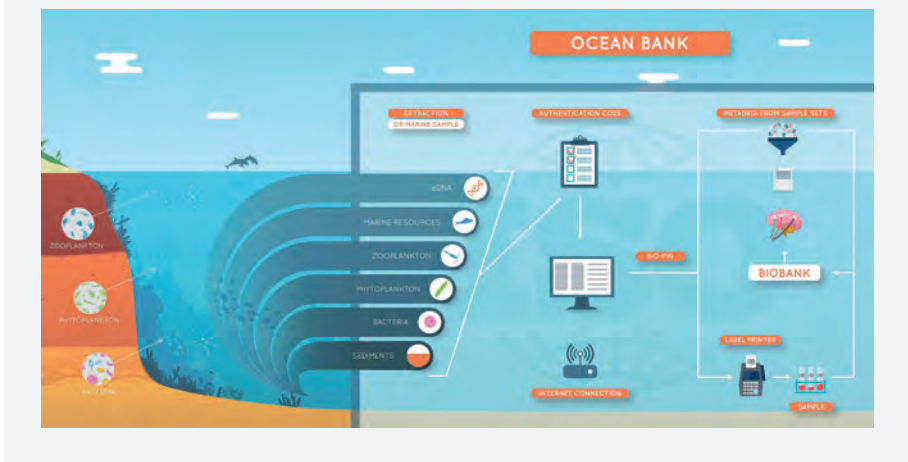
To understand microbial life in the ocean, databases represent a key resource. Genomic information is commonly automatically annotated using databases that are: 1) biased towards certain model organisms, 2) incomplete and, 3) too many times wrong. The result is that these automatically annotated genomes that are poorly annotated (because they differ too much from model organisms), present lots of missing data (because the databases are incomplete) or contain errors (because databases contain errors), end up becoming part of these same databases. So far, the best way to generate reliable reference databases is through manual “human” curation. These curated databases can be used to train AI algorithms to perform, at a larger scale, a similar curation task than that initially performed by humans. Such AI-curated databases represent a future challenge that will contribute to understand ocean genomes.

Big Data and AI will not only affect the way we understand microscopic organisms, but also large counterparts. The collection and analysis of machine-sensed (through the use of electronic tracking devices) data regarding animal social behaviour to model behavioural patterns is deeply changing the

way to study marine animal populations (Krause *et al.* 2013). Animal-tracking technology allows nowadays gathering exceptionally detailed machine-sensed data on the social dynamics of almost entire populations of individuals living in the oceans. High-resolution aquatic tracking is profoundly revolutionizing our views and understanding of ocean functioning, and now we have a powerful tool for studying the *in situ* behavioural variation in hundreds of free-living individuals, in an unprecedented spatiotemporal scale (Sequeira *et al.* 2018). This will enable the creation of experimental platforms to revisit basic and applied unresolved questions of ecology, coastal management, and conservation biology. For instance, the first reality-mining experiment in marine systems where nearly three hundred fish individuals were simultaneously tracked at a high-resolution scale was developed by a CSIC institute (Laboratory of fish ecology, IMEDEA). This experiment has generated in three weeks approximately millions of 2-dimensional positions and behavioural records at a high temporal resolution (5 seconds in average) and high spatial accuracy (1 m) that have changed our views of ocean functioning and animal social networks with conservation implications. The challenge of the reality-mining approach to aquatic social systems is to close the gap between biological and physical patterns and their underlying processes, providing insight into how animal social systems arise and change dynamically over different timescales.

Big Data and the application of AI have also arrived to the field of ocean conservation (Lamba *et al.* 2019). For instance, the recent footprint of fisheries, when 22 billion automatic identification system messages and the >70,000 tracked industrial fishing vessels were combined with DL algorithms, have created a global footprint of fishing effort. This global fisheries map has revealed that fishing activity occurs in more than 55% of the ocean with serious implications for the conservation of wild fish stocks. At a more local scale, the dynamics of fish length distribution is a key input for understanding the fish population dynamics and taking informed management decisions on exploited stocks. Recent applications of AI to fisheries science are opening a promising opportunity for the massive sampling of fish catches. For instance, a deep convolutional network (Mask R-CNN) for unsupervised (i.e. fully automatic) European hake length estimation from images of fish boxes was successfully developed to automatically collect data from landing (Álvarez-Ellacuría *et al.* 2020). The potential applications of DL in ocean conservation are immense and go beyond the classification of visual, spatial, and acoustic data, with their ability to self-learn patterns in large volumes of data (Christin *et al.* 2019).

FIGURE 8.3—The Ocean Bank can biobankise any ocean sample (i.e. zooplankton, DNA, sediments or tissue). The biobanking process can be integrated with AI-assisted analyses of Big Data.



This makes deep artificial neural networks very useful for modelling complex ecological systems, real-time monitoring and surveillance sources (Lamba *et al.* 2019).

Understanding the biology of the ocean requires large sampling efforts that may generate hundreds of thousands of samples per year. Oftentimes, these samples become the basis of new Big Data. Organising these Big-Sample sets so that they are properly catalogued and stored, being also available for the community is a challenge. Such organization of samples require combined efforts at the national and international level. For example, the Ocean Bank network initiative (Fig. 8.3) consists of upgrading the concept of sample sharing under a cession-donor scheme. The long-term objective of the Ocean Bank initiative is to create a network of Big Sample banks. Biobank samples may include marine resources and seafood, gene proteins of plankton, chemicals from seabed sediments, marine biomolecules, or seawater. Biobanks will develop to become a safe warehouse for marine ecosystem samples that will be used for improving research on animal and human diseases, marine ecosystem health, food productivity and safety and development of environmental technologies. Traceable Big Sample sets of *biobanks* will open immense opportunities for Big Data and AI. For example, AI-assisted analyses of Big Data may determine the potential value of specific unanalysed samples for

different research questions, pointing to samples sets, including samples distributed around the world, that may be the most suitable for addressing a specific question. Implementing biobanks among researchers investigating ocean's life and connecting these large samples sets to the produced Big Data in a context of AI analysis represents a challenge for the next decade.

3.3. Comprehending historical interactions between humans and the ocean

The ocean contains important information that bears witness to continuous human interactions (anthropic action) with the sea. These interactions have been the focus of study for quaternary scientists and social scientists who have collected large bodies of data and developed databases and GIS models. A current challenge is the need to develop complex data models and associated integration, visualization, and analysis tools that manage to integrate the study of the relationships between humankind and the ocean from a holistic and multidisciplinary perspective and not separately as has been the case. In particular, an interdisciplinary problem refers to the lack of a data model oriented to integrate, manage, store and analyse all kinds of structured and unstructured Big Data and associated metadata referring to coastal archaeological remains and underwater cultural heritage (including submerged landscapes and settlements, shipwrecks and downed aircraft), and associated historical and intangible cultural information. Furthermore, there is a need to create new tools for the integration and sharing of historical and archaeological information with data from life sciences.

The creation of new models for integrating data from the human and social sciences to identify maritime cultural heritage should meet the standards of Spatial Data Infrastructures (SDI). The management of maritime historical and archaeological Big Data should also contribute to documentation, surveillance, and data monitoring leading to better governance of this heritage. The use of AI on this data may allow discovering patterns that may lead to new archaeological or historical insights. Particularly, the use of AI in maritime cultural heritage may enhance computer-driven information management. Visualization software and GIS tools, will help formulating formal ontologies that express the nature of reality and the relations among entities, and may help to develop an evolutionary GIS model capable of updating multiple data types, which includes multidisciplinary analysis of life and social sciences.

CHALLENGE 8 | REFERENCES

- Álvarez-Ellacuría, Amaya, Miquel Palmer, Ignacio A Catalán And Jose-Luis Lisani. (2019). «Image-Based, Unsupervised Estimation of Fish Size from Commercial Landings Using Deep Learning», *ICES Journal of Marine Science*, fsz216.
- Bar-On, Yinon M, Ron Milo. (2019). «The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet», *Cell*, 179:1451-1454.
- Brett, Annie, Jim Leape, Mark Abbott, Hide Sakaguchi, Ling Cao, Kevin Chand, Yimnang Golbuu, Tara J. Martin, Juan Mayorga & Mari S. Myksovoll. (2020). «Ocean data need a sea change to help navigate the warming world», *Nature*, 582: 181-183.
- Christin, Sylvain, Eric Hervet And Nicolas Lecomte. (2019). «Applications for Deep Learning in Ecology», *Methods in Ecology and Evolution*, 10 : 1632-1644.
- Fablet, Ronan, et al. (2018). «Improving Mesoscale Altimetric Data from a Multitracer Convolutional Processing of Standard Satellite-Derived Products», *IEEE Transactions on Geoscience and Remote Sensing*, 56: 2518-2525.
- Fu, Lee-Lueng, Tong Lee, W Timothy Liu And Ronald Kwok. (2019). «50 Years of Satellite Remote Sensing of the Ocean», *Meteorological Monographs*, 59: 5.1-5.46.
- Guidi, Lionel, et al. (2020). «Big Data in Marine Science». ed. Alexander, Britt, et al. European Marine Board, Ostend, Belgium.
- Kitchin, Rob. (2014). «The Data Revolution. Big Data, Open Data, Data Infrastructures and Their Consequences» London, UK: SAGE Publications Ltd.
- Krause, Jens, Stefan Krause, Robert Arlinghaus, Ioannis Psorakis, Stephen Roberts And Christian Rutz (2013). «Reality Mining of Animal Social Systems», *Trends in Ecology & Evolution*, 28: 541-551.
- Lamba, Aakash, Phillip Cassey, Ramesh Raja Segaran And Lian Pin Koh (2019). «Deep Learning for Environmental Conservation», *Current Biology*, 29 (2019): 977-982.
- Logares, Ramiro, Thomas Ha Haverkamp, Surendra Kumar, Anders Lanzen, Alexander J Nederbragt, Christopher Quince, Havard Kauserud (2012). «Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches». *Journal of Microbiological Methods* 91:106-113.
- Malde, Ketil, Nils Olav Handegard, Line Eikvil And Arnt-Børre Salberg (2019). «Machine Intelligence and the Data-Driven Future of Marine Science», *ICES Journal of Marine Science*, fsz057.
- Reichstein, Markus, et al. (2019). «Deep Learning and Process Understanding for Data-Driven Earth System Science», *Nature*, 566: 195-204.
- Sequeira, Ana M., et al. (2018). «Convergence of Marine Megafauna Movement Patterns in Coastal and Open Oceans», *Proceedings of the National Academy of Sciences*, 115: 3072-3077.
- Turiel, Antonio, Hussein Yahia And Conrad J. Pérez-Vicente (2007). «Microcanonical Multifractal Formalism—a Geometrical Approach to Multifractal Systems: Part I. Singularity Analysis», *Journal of Physics A: Mathematical and Theoretical*, 41: 015501.
- Vance, Tiffany C, et al. (2019). «From the Oceans to the Cloud: Opportunities and Challenges for Data, Models, Computation and Workflows», *Frontiers in Marine Science*, 6: 211.