Future data challenges (and opportunities) in ICES

The aim of this document is to introduce for discussion and awareness the up and coming challenges that affect directly the way we deal with data in ICES Science and Advice processes. We are aiming to build around these themes within the next ICES strategic plan from a cross-cutting perspective. These considerations will not replace or stop all other data initiatives and work in ICES, but rather shape them for the future.

There are already many and varied challenges of collecting, channelling, quality controlling and making data available within the ICES community. The existing data collections and online portals are in a constant flux, adapting to the changing demands of the ICES community, technological changes and regulatory pressures as well as the constant need for increasing capacity. These activities will benefit from continued collaboration, communication and a community focussed review. The emphasis here is on new challenges that are changing the landscape of data management, and which will ultimately, or are already, having an impact in the ICES community – either through direct technological changes, changes in the expectations and working culture of scientists, or through policy demands.

Data Management and data science are still relatively new disciplines that are typically integrated within other disciplines, rather than stand-alone concepts. It is a rapidly developing area that spans concepts from IT through to business strategic goals. Common for it all, is the drive to use data in many forms for many purposes. Data are the very foundation upon which we build our science and advice, and the ability to do so in a well-structured, transparent, and competent way require consideration of what lies ahead.

We have identified 3 priority future challenges, some of which are already embedded within the ICES community, while some touch more on the very fabric of the way in which we collaborate:

- Machine learning
- Cloud environments and related services
- Open Data[1] and code sharing

Machine Learning

This is a stepping stone towards artificial intelligence, and uses algorithms to learn from (usually a lot of) data without being explicitly programmed. One of the oldest and best known examples is to program computers to learn how to play games[2] such as Chess or Go. Today, there are many branches of applications for machine learning, especially with the exponential growth in data to 'feed' the algorithms.

1 http://sciencecommons.org/about/towards as an example

2 Some Studies in Machine Learning Using the Game of Checkers, Samuel, Arthur (1959) https://doi.org/10.1147/rd.33.0210

In the marine domain, there is a diverse range of applications from modelling and prediction of climate change, fishing patterns from AIS (satellite data), to taxonomic recognition of biological or acoustic samples. Marine science practitioners will see many research and infrastructure projects that will exploit the power of machine learning in the coming years, and this will provide an opportunity to improve the skills and tools available in the ICES community, as well as improving the speed (and arguably quality) of production of data, science, and assessment products.

The key challenge in working with machine learning is that algorithms are often extremely power hungry, requiring high capacity computing as well as large volumes of data to drive the machine learning. The infrastructure requirements and storage capacity may mean that shared virtual or cloud computing spaces are required. Further, there are emerging discussions about the impact and understanding of machine learning in areas, particularly with social or economic impact – e.g. around accountability, review, and transparency of fully automated workflows like these. For example, in previous data analysis, trends and patterns may have been identified and published with a methodology, whereas an automated algorithm encapsulating an entire workflow may be perceived to make a process less transparent. Issues such as these, need to be constantly highlighted and challenged within the community, and this year there will be a workshop on machine learning where experts will explore such challenges and opportunities for ICES (Workshop on Machine Learning in Marine Science WKMLEARN[3]).

Cloud environments and data services

A catalyst of machine learning, and shared analysis, is the wholesale move of computing, data storage and retrieval, and writing and execution of coding over the internet rather than from local hard drive - the cloud[4]. There are too many cloud initiatives and options to mention, but the major research funding programmes on both sides of the Atlantic are pushing into this area alongside traditional commercial cloud systems. Cloud computing, and the virtual modelling/research environments[5] they can provide, offer many opportunities to the marine domain but also some serious challenges.

The main benefits from cloud computing lie in freeing ourselves from hardware and platform restraints, allowing the dynamic addition of computing resource to applications – even on a temporary basis. The cloud offer safeguards against data loss by the duplication of data across data centres to keep it safe, taking away the concerns of researchers carrying unique data on their laptops or hard drives, which may ultimately be lost.

To make the most of the cloud, the ability to work in a much bigger and shared environment is coupled with data services and semantic mapping of data to allow machines to understand the relationships between two datasets that may appear unconnected. Imagine an extended species list that can explain the taxonomic organisation of species (cousins, grandparents etc.) but extend beyond to combine with the genetic make-up, the ecology of species, their geographic distribution etc.

3 http://ices.dk/community/groups/Pages/WKMLEARN.aspx

4 http://uk.pcmag.com/networking-communications-software-products/16824/feature/what-is-cloud-computing

5 See Bluebridge for some marine examples from Europe http://www.bluebridge-vres.eu/

Such mappings can connect data on previously unprecedented scale that can truly allow ecosystem assessments to become integrated.

The very mapping and organisation of data is a massive challenge – even when moved into a scalable place like the cloud. First off – "The cloud" is not a singular space, and so the locations and ways of communicating or obtaining data still needs to be managed. If the data becomes spread across multiple cloud instances, it requires perhaps even more discipline than before to ensure data are obtained from the authoritative resource and is kept up to data. In addition, the mapping and description of the conceptual relationship between the components and parameters that make up marine data is a huge challenge – both in terms of developing and describing that understanding, and the encoding of the logic in marine data. Finally, there are ongoing legal and ownership issues regarding data in the cloud and which national, European or International rules apply to these data, and accountability in the event of an infringement of rights as it can be ambiguous to the data owner/provider where the data are actually stored.

Open data and code sharing

In spite of the advance in cloud environments and ever increasing generation of data, there are still very human challenges in how to deal with intellectual rights, ownership of data, recognition of effort, and the means to make the data and tools available in a way that is understandable and workable for the providers and consumers.

The marine domain is both a good and a bad example of these practises, which is reflected from the individual practitioner all the way through to regional organisations. ICES, for example, is working towards the FAIR[6] principles (Findable, Accessible, Interoperable, and Reusable). This is a long journey that requires work across the whole community in the coming years, and although ICES have come a long way in achieving clarity about data licensing provided into the community, it is a shifting environment with differing rates of change in member countries. Some will push for more openness of detailed data, while others will want to provide access to derived or aggregated data products. ICES is taking a proactive approach to answering this challenge, through its ongoing developments such as Transparent Assessment Framework (TAF[7]), Regional Database and Estimation System (RDBES) and Contaminants assessment tool for the regional sea conventions.

Currently, there many data collections within ICES actively managed and on the website portals. But there are also many data products, overviews, reviews, maps etc. hidden away in reports that could potentially be made more readily accessible. Enabling data access while maintaining and improving the FAIR principles adoption will be a significant challenge in the years to come.

Importantly, such mechanisms are not merely a question of technology to assist, but needs to be incorporated into the strategic planning, and business processes of developing advice, viewpoints, integrated assessments etc.

6 https://www.force11.org/group/fairgroup/fairprinciples

7 http://ices.dk/marine-data/tools/Pages/transparent-assessment-framework.aspx

Concluding remarks

Although we have identified 3 main subjects here, it should be recognised that they can't be isolated entirely. It will be very difficult to adopt machine learning without a scalable cloud environment. Likewise, it will be difficult to run machine learning without access to open, well defined and mapped data. Further, if more data products are openly shared with established links to the detail data, an ecosystem of data will gradually emerge. This ecosystem, provided it is accessible and shared in common and well-defined standards, will in turn enhance analysis that utilise machine learning.

The ICES community are already working on, or developing an understanding of these challenges – but they are also the ones who are likely to place the greatest pressures and expectations on ICES data operations in the coming years, and as such require strategic consideration.