IICWG-XIX: Plenaries 1-2-3 Background

Introduction

An important emerging issue for ice services is that of so-called "big data". There was a good introductory session on this topic at the 2017 meeting in Hobart where we concluded that "big data" is here and has major impact in our community, now and in the years to come. It will affect not only in the way we handle and analyze data, but also on the ice information products and services we provide to mariners.

It is a topic that calls for deeper analysis, understanding, and discussion, including investigations on where it makes sense for us to collaborate. At IICWG-XIX, we will have three plenary sessions to help us focus on these issues.

The 1st Plenary Session sets the stage by forming a mutual understanding of "Big Data" terminology so commonly used today and concludes with a group discussion of key issues confronting all IICWG participants. With these prioritized issues as a backdrop, the 2nd Plenary Session highlights key efforts in data analytics designed to stimulate group discussions to imagine how machine learning and ultimately, artificial intelligence, will shape the role of the ice analyst and help define the space between human and machine. On Wednesday afternoon, the series of three Plenary Sessions culminates with an eye toward the future.

Each session will begin with a few presentations to refresh and focus our minds for the most important part of the session: the discussion! Highlights from session 1 will go into session 2, and similar highlights from session 2 will go into session 3.

Since the "big data" topic deals with many terms and phrases that may not be familiar to all, we have included a "Big Data ABC" as an appendix to this document.

Plenary 1: Identifying the Problem

"So Much Data–So Little Information? The ice navigation dilemma.

Plenary 1 will open with an introduction to the topic of big data, machine learning, machine reasoning, and cloud computing. To help us focus on the most important aspects of this big topic, an on-line exercise was held prior to the meeting asking IICWG participants to rank the priority of the following issues:

- Blending data sources
- Image Analysis
- Roles and Responsibilities of National and Commercial Ice Information Providers
- Visualizing Uncertainty
- Understanding what satellite information is available
- Preserving the continuity of local ice expertise

- Availability and access to sea ice and iceberg climatology
- Standard format
- Educating users

The results of that exercise will be presented and discussed with the participants.

Plenary 2: Solutions

Big Data, Artificial Intelligence, and the Role of the Ice Analyst

Plenary 1 will identify the most critical issues that ice services are currently facing. They should be preparing for the effects of climate change on the Polar Regions where we are seeing increased activity, the availability of more satellite data and new sensors, and a need for more sophisticated and timely routine information provision to end-users. This will require efficient mechanisms to handle large volumes of satellite data with increasingly sophisticated and complex approaches to ice analysis. End users have consistently asked for confidence mapping on ice information products which will become even more essential with increased use of automated tools.

Plenary 2 will begin with a few short presentations to introduce the possibilities. Following that, small breakout groups will be asked to address the issues of big data and machine learning in the ice services by considering the following questions:

Q1. How will ice analysis and information provision change in the future?
- What role will ice analysts performing in 5-10 years?
- What can analysts do to improve the confidence in ice information products? Is this something ice services are responsible for?
- How does/could machine learning contribute to modern ice charts?
- How could AI and machine learning tools be implemented in an ice service?
- Will it be possible to provide ice information multiple times a day?

Q2. Considering the explosion in the volume of data becoming available, what is the best place for data analysis?
- How will analysts develop trust in black-box solutions?
- Must data analysis be done in the ice service to be synergistic with other information (e.g. user reports)?
- What are the benefits and challenges of using cloud based systems for data storage and handling?
- How can we streamline the data toolbox?
- What is the future for Sigrid 3 and standardization for operational ice products?
- Who should analyze iceberg info? Do we expect the ship navigator to know where the icebergs are or is it the responsibility of ice analysts?

Q3. How will ice surveillance improve over time?

- Can ice services integrate information with drones, reconnaissance and other sources?
- The OODA loop is something that is always in place for operators. How do we fit in here from an operations perspective?

Q4. Should ice charts work towards being more targeted to specific activities?
- Whose responsibility is it to provide this type of data for this type of incident?

Q5. Are there, or should there be, boundaries for commercial and government ice information services?
- Are they specific to different countries?
- Are there ways in which government and commercial ice services can work cooperatively for mutual benefit and to best serve end users?
- Is it better for everyone to be independent?

Plenary 3: "Transitioning to Future Satellite/Sensor Conceptions"

Leading up to the 3rd Plenary Session, Wednesday morning's Science Workshop examines current research efforts. These include investigations on combining a variety of data sources with varying spatial/temporal resolutions and frequencies. The science discussed during this workshop could improve our understanding of the changing cryosphere and arm ice information providers with new tools to communicate this dynamic environment to end-users.

The purpose of the third Plenary Session on Wednesday afternoon is to look at future systems that bridge the gaps identified and discussed on Monday and Tuesday. This Session is designed to build on the first two sessions with presentations to provide a sampling of future satellite systems along with a novel application that have particular relevance to the meeting's theme of "Ice Information for Navigating the Sub-Polar Seas". They are not intended to give a comprehensive overview but will try to provide the audience with insight on how future satellite missions (5-10 years) will provide the tools needed by the ice analysts and ice information providers to meet their responsibilities.

The session concludes with a panel discussion led by five members representing operational ice services, the scientific research community and two space agencies, ESA and DLR. As both consumers and producers of ice information, the operational ice services are in a unique position to share both their needs for satellite data and the challenges they face in creating meaningful ice products. The importance of on-going scientific research is essential to developing an in-depth understanding of the benefits derived from new discoveries e.g. blending ice observation data from multiple sources to ultimately improve ice analysts' capability to provide the end-user with the most useful ice information possible. The space agency representatives bring insight to the drivers for missions already planned in the near future and for those not yet conceived beyond a decade. In response to questions from the moderator and the audience, panelists will provide their

unique perspectives on the role that IICWG can play to provide the end-user with the best ice information possible.

It is the Organizing Committee's intention that the results of these first three Plenary Sessions should help the Standing Committees to develop a list of succinct, meaningful and achievable actions that can be pursued collaboratively.

Penny Wagner
Richard Hall
Keld Qvistgaard
Mike Hicks
John Falkingham

September 7, 2018

Appendix - Big Data Dictionary, version 0.1

Modified from https://campus.sagepub.com/blog/glossary-of-big-data-terms

Aggregation – a process of searching, gathering and presenting data

Algorithms – a mathematical formula that can perform certain analyses on data

Analytics – the discovery of insights in data

Anonymization – making data anonymous; removing all data points that could lead to identify a person

API - an application programming interface is a set of subroutine definitions, protocols, and tools for building application software

Application – computer software that enables a computer to perform a certain task

Artificial Intelligence – developing intelligence machines and software that are capable of perceiving the environment and take corresponding action when required and even learn from those actions.

Behavioural Analytics – analytics that informs about the how, why and what instead of just the who and when. It looks at humanized patterns in the da

Big Data - Big data is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data-processing applications.

Brontobyte - A brontobyte is a measure of memory or data storage that is equal to 10 to the 27th power of bytes. There are approximately 1,024 yottabytes in a brontobyte. Approximately 1,024 brontobytes make up a geopbyte.

Classification analysis - a systematic process for obtaining important and relevant information about data, also meta data called; data about data.

Cloud computing – a distributed computing system over a network used for storing data off-premises

Clustering analysis – the process of identifying objects that are similar to each other and cluster them in order to understand the differences as well as the similarities within the data.

Comparative analysis – it ensures a step-by-step procedure of comparisons and calculations to detect patterns within very large data sets.

Complex structured data – data that are composed of two or more complex, complicated, and interrelated parts that cannot be easily interpreted by structured query languages and tools.

Computer generated data – data generated by computers such as log files

Concurrency – performing and executing multiple tasks and processes at the same time

Data aggregation tools - the process of transforming scattered data from numerous sources into a single new one.

Data analyst – someone analysing, modelling, cleaning or processing data

Database – a digital collection of data stored via a certain technique

Database Management System– collecting, storing and providing access of data

Data centre – a physical location that houses the servers for storing data

Data cleansing – the process of reviewing and revising data in order to delete duplicates, correct errors and provide consistency

Data custodian– someone who is responsible for the technical environment necessary for data storage

Data feed – a stream of data such as a Twitter feed or RSS

Data mining – the process of finding certain patterns or information from data sets

Data modelling – the analysis of data objects using data modelling techniques to create insights from the data

Data set – a collection of data

Data virtualization – a data integration process in order to gain more insights. Usually it involves databases, applications, file systems, websites, big data techniques, etc.)

Deep Learning: a subfield of machine learning, deep learning imitates the human brain by building artificial neural networks.

Exploratory analysis – finding patterns within data without standard procedures or methods. It is a means of discovering the data and to find the data sets main characteristics.

Exabytes – approximately 1000 petabytes or 1 billion gigabytes.

Extract, Transform and Load (ETL) – a process in a database and data warehousing meaning extracting the data from various sources, transforming it to fit operational needs and loading it into the database

Failover – switching automatically to a different server or node should one fail

Fault-tolerant design – a system designed to continue working even if certain parts fail Feature - a piece of measurable information about something, for example features you might store about a set of people, are age, gender and income.

Graph Databases – they use graph structures (a finite set of ordered pairs or certain entities), with edges, properties and nodes for data storage. It provides index-free adjacency, meaning that every element is directly linked to its neighbour element.

Grid computing – connecting different computer systems from various location, often via the cloud, to reach a common goal

Hadoop – an open-source framework that is built to enable the process and storage of big data across a distributed file system

HBase – an open source, non-relational, distributed database running in conjunction with Hadoop

HDFS – Hadoop Distributed File System; a distributed file system designed to run on commodity hardware

High-Performance-Computing (HPC) – using supercomputers to solve highly complex and advanced computing problems

Histogram - A graphical representation of the distribution of a set of numeric data, usually a vertical bar graph

In-memory – a database management system stores data on the main memory instead of the disk, resulting is very fast processing, storing and loading of the data

JavaScript - a scripting language designed in the mid-1990s for embedding logic in web pages, but which later evolved into a more general-purpose development language.

Juridical data compliance – relevant when you use cloud solutions and where the data is stored in a different country or continent. Be aware that data stored in a different country has to oblige to the law in that country.

KeyValue Databases – they store data with a primary key, a uniquely identifiable record, which makes easy and fast to look up. The data stored in a KeyValue is normally some kind of primitive of the programming language.

Latency – a measure of time delayed in a system

Load balancing – distributing workload across multiple computers or servers in order to achieve optimal results and utilization of the system

Location data – GPS data describing a geographical location

Log file – a file automatically created by a computer to record events that occur while operational

Machine data – data created by machines via sensors or algorithms

Machine learning – part of artificial intelligence where machines learn from what they are doing and become better over time

Metadata – data about data; gives information about what the data is about.

Multi-Dimensional Databases – a database optimized for data online analytical processing (OLAP) applications and for data warehousing.

MultiValue Databases– they are a type of NoSQL and multidimensional databases that understand 3 dimensional data directly. They are primarily giant strings that are perfect for manipulating HTML and XML strings directly

Natural Language Processing– a field of computer science involved with interactions between computers and human languages

Network analysis– viewing relationships among the nodes in terms of the network or graph theory, meaning analysing connections between nodes in a network and the strength of the ties.

Neural network – A type of artificial intelligence that attempts to imitate the way a human brain works. Rather than using a digital model, in which all computations manipulate zeros and ones, a neural network works by creating connections between processing elements, the computer equivalent of neurons. The organization and weights of the connections determine the output.

Object Databases – they store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.

Object-based Image Analysis – analysing digital images can be performed with data from individual pixels, whereas object-based image analysis uses data from a selection of related pixels, called objects or image objects.

Operational Databases – they carry out regular operations of an organisation and are generally very important to a business. They generally use online transaction processing that allows them to enter, collect and retrieve specific information about the company.

Optimization analysis - the process of optimization during the design cycle of products done by algorithms. It allows companies to virtually design many different variations of a product and to test that product against pre-set variables.

Pattern Recognition – identifying patterns in data via algorithms to make predictions of new data coming from the same source.

Petabytes - approximately 1000 terabytes or 1 million gigabytes. The CERN Large Hydron Collider generates approximately 1 petabyte per second

Predictive analysis – analysis within big data to help predict how someone will behave in the (near) future. It uses a variety of different data sets such as historical, transactional, or social profile data to identify risks and opportunities.

Public data – public information or data sets that were created with public funding

Query – asking for information to answer a certain question

Re-identification – combining several data sets to find a certain person within anonymized data

Regression analysis – to define the dependency between variables. It assumes a one-way causal effect from one variable to the response of another variable.

Real-time data – data that is created, processed, stored, analysed and visualized within milliseconds

Scripting - the use of a computer language where your program, or script, can be run directly with no need to first compile it to binary code. Semi-structured data - a form a structured data that does not have a formal structure like structured data. It does however have tags or other markers to enforce hierarchy of records.

Similarity searches – finding the closest object to a query in a database, where the data object can be of any type of data.

Simulation analysis – a simulation is the imitation of the operation of a real-world process or system. A simulation analysis helps to ensure optimal product performance taking into account many different variables.

Smart grid – refers to using sensors within an energy grid to monitor what is going on in real-time helping to increase efficiency

Spatial analysis – refers to analysing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.

SQL – a programming language for retrieving data from a relational database

Structured data – data that is identifiable as it is organized in structure like rows and columns.

Terabytes – approximately 1000 gigabytes.

Time series analysis - analysing well-defined data obtained through repeated measurements of time. The data has to be well defined and measured at successive points in time spaced at identical time intervals.

Topological Data Analysis – focusing on the shape of complex data and identifying clusters and any statistical significance that is present within that data.

Un-structured data - unstructured data is regarded as data that is in general text heavy, but may also contain dates, numbers and facts.

Variability – it means that the meaning of the data can change (rapidly). In (almost) the same tweets for example a word can have a totally different meaning

Variety – data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data

Velocity – the speed at which the data is created, stored, analysed and visualized

Veracity – ensuring that the data is correct as well as the analyses performed on the data are correct.

Visualization – complex graphs that can include many variables of data while still remaining understandable and readable

Volume – the amount of data, ranging from megabytes to brontobytes

XML Databases – XML Databases allow data to be stored in XML format. The data stored in an XML database can be queried, exported and serialized into any format needed.

Yottabytes – approximately 1000 Zettabytes, or 250 trillion DVD's. The entire digital universe today is 1 Yottabyte and this will double every 18 months.

Zettabytes – approximately 1000 Exabytes or 1 billion terabytes